Objective Assessment of Cleft Lip and Palate Speech Intelligibility

Sishir Kalita

Objective Assessment of Cleft Lip and Palate Speech Intelligibility

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Sishir Kalita

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

November 2019

Certificate

This is to certify that the thesis entitled "**OBJECTIVE ASSESSMENT OF CLEFT LIP AND PALATE SPEECH INTELLIGIBILITY**", submitted by **Sishir Kalita** (146102012), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute, and in our opinion, has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

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

Dated:
Guwahati.

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

To

My sweet and loving family,

for their love, affection, encouragement and prayers

Acknowledgements

This thesis would not have been possible without the immense help and support of several people in various measures. I take this opportunity to convey my most sincere acknowledgment to all of them. I would like to express my deepest and most heartfelt gratitude to my supervisors Prof. S. R. Mahadeva Prasanna and Prof. Samarendra Dandapat, for their constant guidance, help, and encouragement throughout my research work. Their insightful feedbacks have helped me much in improving the quality of my thesis. More than guiding my thesis work, they have shaped my personal and professional life through their mentorship. Without them, I would not have been where I am today. Notably, the weekly meetings with Prof. S. R. Mahadeva Prasanna acted as the periodic excitation to my research work and helped me to upgrade as a researcher continuously.

I am grateful to Prof. Rohit Sinha, the Chairman of the Doctoral Committee and Head, Department of Electronics and Electrical Engineering, IIT Guwahati, for providing valuable suggestions on my work throughout the years. I want to express my sincere gratitude to Dr. Prithwijit Guha and Dr. Priyankoo Sarmah, distinguished members of the Doctoral Committee for their constructive criticisms during my progress seminars. During frequent interactions with Dr. Priyankoo Sarmah, I must admit, I could learn many new concepts of acoustic-phonetics. I am indebted to Dr. S. Sundaram, who always encouraged me in every bit of discussion with him. I enjoyed the TA work done with him. I would also like to thank other faculty members of the department for their care and support. A sincere thanks to Dr. Abhishek Shrivastava, Department of Design, for his valuable suggestions and encouragement for my research work.

I would also like to convey my gratitude to all technical and non-technical staffs of the EEE department for their help and support throughout my Ph.D. duration; especially, I acknowledge Mukut Da for timely forwarding of different applications.

This thesis would become highly impossible without the help and support of speech-language pathologists and staff of All Indian Institute of Speech and Hearing, Mysuru, India. Especially, I would like to express my sincere thanks to Prof. M. Pushpavathi and Prof. Ajish K. Abraham for giving me lots 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 support of Dr. Gopi Sankar, Dr. Navya, Mrs. Deepthi, Ms. Nikitha, and Mr. Girish, during data

collection and perceptual evaluation. Further, I would also like to thank the children, parents, and teachers for their cooperation during the data collection process.

I take this opportunity to extend my gratitude towards my seniors Dr. Deepak K.T., Dr. Biswajit Dev Sarma, Dr. Nagaraj Adiga, Dr. Banriskhem K. Khonglah, Dr. Rohan Kumar Das, Dr. Bidisha Sarma, Dr. Syed Shahnawazuddin, Dr. Jiss Nullikuzhy, Dr. Suman Deb, and Dr. Rajib Sarma for various technical discussions and suggestions they provided whenever I needed. I never forget my close friends Vikram, Akhilesh, and Protima, 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 to past/present members of the lab Himakshi, Subhasis, Mawsumi, Bhukya, Abhishek, Sarfaraz, Sikha, Moa, Mrinmoy, Saswati, Sandeep, Sreeram, Nagendra, Sukanya, Prabhakar, Vineeta, Alex, Ato, Samarjit, Tilendra, Shibasis and the rest for their direct/indirect contributions during my stay at IITG. I would like to thank Dr. Luke Horo, Wendy Lalhminghlui of Phonetics and Phonology Lab, IIT Guwahati, and Pamir Gogoi, the University of Florida, for sharing the knowledge of acoustic-phonetics during some collaborative works.

I want to thank the funding agencies, such as ISCA and Microsoft Research India, DST, Indian Speech Communication Association, and IIT Guwahati, for funding travel to attend conferences abroad, which allowed me to meet experienced researchers and explore very nice destinations like Stockholm and Vienna. I would like to thank MHRD, Government of India, for providing me a fellowship to pursue my Ph.D. Also, I would like to acknowledge DBT, Government of India, for providing financial assistance during the data collection.

The stay at the IITG campus would not have been comfortable without the facilities provided by the institute. The campus life remained lively in the presence of my friends, Vikram, Akhilesh, Protima, Kaushik, Sharmila, Vivek, Alex, Matthew, Vijith, Abhishek, Sarfaraz, Aniruddha, Abhijit, Bhaskar Da. Their constant help and support accelerated my life in the last five years of IITG.

I am incredibly fortunate to have a caring and loving family who deserve special mention for providing mental strength and encouragement to explore my potentials and pursue my dreams. With great pleasure, I would like to give special thanks to my wife, Parismita. Her more in-depth support, greater patience, and empathy made me strong and capable of completing this thesis. Finally, I thank God Almighty for the blessings he has bestowed upon me and for giving me the wisdom to achieve this dream.

Sishir Kalita

Abstract

Conventionally, the intelligibility of cleft lip and palate (CLP) speech is evaluated using the auditory-perceptual based subjective method by speech-language pathologists (SLPs). This method is considered as the gold standard in clinical settings; however, by nature, it is a subjective method. The measures based on the speech technology systems may assist the SLPs by providing objective and interpretable results. This thesis aims to propose a set of objective measures for sentence-level intelligibility, and three different potential directions are identified to derive the measures.

Before deriving the intelligibility measures, a subjective analysis of intelligibility is conducted to know which speech disorders have the significant effect in reducing the intelligibility. From the study, it is found that intelligibility of CLP speech is primarily degraded due to the articulation error and hypernasality. Motivated by this observation, in the first work, a composite measure of intelligibility is proposed by combining the information of articulation deficits and hypernasality. The objective measure of hypernasality is computed using the Gaussian mixture model and the Bayesian posterior method. Articulation scores are computed by analyzing the speech regions anchored around vowel onset points and vowel end points. Later, scores derived from these measures are mapped to the intelligibility score using a regression model.

The second work is motivated by the fact that intelligibility of CLP speech is mainly degraded due to the deviations in obstruent production. Therefore, the *g(lottis)–landmarks* which are associated with obstruents to vowel and vowel to obstruent transitions are used to characterize the deviations in obstruents production. The spectral and joint spectro-temporal features computed around the *g–landmarks* are used to derive a measure of speech intelligibility. The measure based on joint spectro-temporal features shows better correlation with respect to perceptual ratings of intelligibility compared to the measure based on spectral features.

Further, methods are developed based on the comparison of posterior sequences to measure the intelligibility. The degree of deviation in the posterior sequence of test speech from the normal speaker's template is used to quantify the intelligibility using two approaches. In the first approach, dynamic time warping is used to quantify the deviation between the posterior sequences. In the second approach, each posterior sequence is transformed into another representation, termed as self-similarity matrix, and comparison is performed in that representation.

Finally, a visual representation, based on a spider plot is proposed for visualizing the intelligibility scores. The polygon in spider plot represents intelligibility space of the speaker, and area of the polygon is computed to derive speaker-level intelligibility score.

The salient contributions of the thesis are summarized as follows:

- A sentence-level CLP speech database is developed for intelligibility assessment.
- The relative impact of hypernasality, articulation error and voice quality on CLP speech intelligibility is analyzed.
- Developed objective measures for predicting articulation deficits and hypernasality levels, and combine these measures to derive correlates of intelligibility.
- The importance of joint spectro-temporal features over the conventional MFCCs in intelligibility estimation is studied.
- The distortion of abrupt landmark's expression in CLP speech is analyzed, and its importance in deriving the acoustic correlates of CLP speech intelligibility is demonstrated.
- The Gaussian posteriorgram and self-similarity matrix representations of speech are explored in intelligibility assessment. The distance between the normal template and test utterance is studied as a correlate of intelligibility.
- A visual representation based on the spider plot is proposed for graphing the intelligibility scores. The polygon in spider plot represents intelligibility space of the speaker.

Keywords: Cleft lip and palate, intelligibility, glottis landmarks, Gaussian posteriorgram, dynamic time warping, self-similarity matrix.

Contents

List of Figures

List of Tables

Glossary

Cleft lip and palate	A congenital disorder of craniofacial region. Depending on the place of cleft, this disorder can be classified as cleft lip, cleft palate, and cleft lip and palate.
Compensatory articulations	They are misarticulation that occur due to anatomical inability to close the velopharyngeal valve. These are learned articulation errors.
Fricative	A sound produced with a narrow supraglottal constriction resulting in a turbulent flow of air at the constriction.
Hypernasality	It is a resonance disorder that occurs due to the coupling of oral and nasal cavities in the production of oral speech sounds..
Intraoral pressure	It is the pressure between the glottis and the mouth opening. When the vocal tract is closed for an oral stop, intraoral pressure builds up behind the occlusion.
Nasal air emission	It is a speech disorder associated with cleft lip and palate individuals. It occurs due to the abrupt air leakage through the nasal cavity during the production of obstruents when there exist velopharyngeal dysfunction or oronasal fistula.
Nasal	A consonant sound produced with complete closure in the oral tract but without velopharyngeal closure, so that the airstream escapes only through the nasal cavity.
Nasal cavity	The large cavity above the roof of the mouth, connected to the upper part of the pharynx at the rear and having the nostrils at the front.

Nasalized vowel	A vowel sound produced without velopharyngeal closure so that air escapes simultaneously through the oral cavity and the nasal cavity. Vowel may be nasalized due to co-articulation and velopharyngeal dysfunction.
Obstruent	A class of sounds produced with either a complete closure or a narrow constriction causing a burst or frication. The obstruents are stops, fricatives, and affricates.
Oral sound	A sound produced with velopharyngeal closure to prevent the nasal escape of air, so that the airstream escapes through the oral cavity alone.
Percentage of consonant correct	It is a measure to evaluate the consonant production ability of individuals with speech disorder. It is measured by the percentage of the correctly produced consonants out of the total number of consonants in a respective paragraph or a set of words.
Sonorant	A sound produced with no turbulence due to a narrow constriction.
Stop	A consonant sound that involves a complete closure in the oral cavity.
Landmarks	Landmarks are defined as the time locations of abrupt acoustic events in a speech signal, and they are correlated with the major articulatory movements
Velopharyngeal valve	Velopharyngeal valve is made up of (i) velum, (ii) lateral pharyngeal walls, and (iii) posterior pharyngeal wall.
Velopharyngeal dysfunction	It is a generic term which describes a set of disorders resulting in the leakage of air into the nasal passages during speech production.
Vowel onset point	It is the instant of time at which vowel region starts in a speech signal.
Vowel end point	It is the instant at which offset of vowel take place in the speech signal.

List of Acronyms

AIISH	All India Institute of Speech and Hearing
ASR	Automatic Speech Recognition
CLP	Cleft Lip and Palate
CV	Consonant-Vowel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
GAD	Glottal Activity Detection
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GP	Gaussian posteriograms
HE	Hilbert Envelope
HELPR	Hilbert Envelope of Linear Prediction Residual
HMM	Hidden Markov Model
IPM	Intelligibility Prediction Model
IPA	International Phonetic Alphabet
LOSO	Leave One Speaker Out
LP	Linear Prediction
LPR	Linear Prediction Residual
MFCC	Mel-Frequency Cepstral Co-efficients
PCC	Percentage of Consonant Correct
SLP	Speech-Language Pathologist
SODA	Substitution, Omission, Distortion, and Addition
SVR	Support Vector Regression

SVM	Support vector machine
SSM	Self-similarity Matrix
SSIM	Structural Similarity
SoE	Strength of Excitation
TFR	Time-Frequency Representation
UBM	Universal Background Model
VC	Vowel-Consonant
VOP	Vowel Onset Point
VEP	Vowel End Point
VPD	Velopharyngeal Dysfunction
VQ	Voice Quality
WER	Word Error Rate
WA	Word Accuracy
ZFF	Zero-Frequency Filter
ZFFS	Zero-Frequency Filtered Signal

1

Introduction

Contents

Objective of the thesis

The objective of this research work is to derive intelligibility measures based on the acoustic analysis of speech for cleft lip and palate (CLP) children. To achieve this objective, the preliminary requirement is to develop a speech database of CLP and controlled normal speakers, and it is done with the collaboration of All India Institute of Speech and Hearing (AIISH), Mysuru, India. In the clinical setting, intelligibility is evaluated using the auditory-perceptual based methods, which are subjective in nature. The measures based on the speech technology systems may assist the speech language-pathologist by providing objective and interpretable results. Moreover, these measures only rely on the acoustic characteristics of the deviant speech; hence, contextual and familiarization biases may not be present. CLP speech intelligibility is primarily degraded due to the presence of articulation error and hypernasality. Therefore, first characterizing the articulation deficits and hypernasality in CLP speech, and then combine them to derive a composite measure of intelligibility is explored in the thesis. In CLP speech, articulation errors occur mostly for the obstruent sounds. Therefore, processing the speech around specific acoustic events to characterize the deviations of obstruent production may be helpful. This work investigates the importance of abrupt consonant landmarks which are associated with the obstruents to characterize the articulation error, and if the deletion and substitution of these landmarks provide the clinically relevant information to analyze the loss of intelligibility. Another direction to derive the measure intelligibility is based on the comparison of normal and CLP speech posterior sequences. The relative deviation from the normal speaker's posterior sequence is considered as the representation of intelligibility loss. The importance of speaker-independent representation based Gaussian posteriorgram and self-similarity matrix in deriving the intelligibility measure is shown.

1.1 Cleft lip and palate speech: An introduction

Cleft lip and palate (CLP) is one of the most common congenital disorders of the craniofacial region. According to the World Health Organization, it occurs in approximately one in every 700 live births worldwide [1]. Children with CLP require surgical intervention to establish the appropriate oral motor skills; however, the speech disorders may persist due to velopharyngeal dysfunction (VPD), dental occlusion, and mislearning even after the surgical repair of cleft [2]. The children with clefts may present with different speech symptoms, such as hypernasality, articulation errors, nasal emission, and voice disorders, factors that can influence the speech intelligibility [2–5]. Hypernasality is the

abnormal nasal resonance caused due to the coupling of oral and nasal cavities in the production of oral speech sounds [2]. With mild hypernasality, the effect is less notable [2, 6]. However, in severe hypernasality, nasal consonants replace obstruents (i.e., stops, fricatives, and affricates), thereby significantly hampering the speech intelligibility [2, 6, 7]. Apart from the hypernasality, intelligibility is also affected by the incorrect articulation patterns in CLP speech [7]. Compensatory errors, such as glottal stops, pharyngeal stops, velar substitution, etc. are some of the primary contributors to diminish the intelligibility [8]. Apart from these, other error, such as the nasalized voiced pressure consonants severely affect the speech intelligibility [2, 6, 9]. After palatal surgery, some children may exhibit normal speech, others may have hypernasality, but no articulation errors and yet others may have compensatory articulation errors [10, 11]. Those who possess any sort of speech disorders require intensive therapeutic intervention for better speech abilities [2]. Another speech problem in children with CLP is the deviant voice characteristics due to the structural or functional problems at the level of the larynx, which can occur secondary to VPD or separately. Voice problems in children with CLP generally include the hoarseness and the soft voice syndrome [8]. However, voice disorder may or may not affect the speech intelligibility in children with CLP. Therapeutic and surgical interventions help in enhancing their speech intelligibility by giving proper treatment guides.

In literature, several researchers have studied the effect of speech disorders in reducing the intelligibility [12, 13]. A single-word intelligibility test performed on thirty-eight children found a high correlation ($r = 0.79$) between perceptual intelligibility ratings and percentage of consonants correct (PCC) [10]. Authors in [14] found that, as the hypernasality increases, intelligibility decreases, and they suggested that hypernasality affects the intelligibility. M. Copeland et al. studied the effect of articulation, nasal resonance, and nasal escape on intelligibility, and they concluded that these speech measures have a direct influence on the degree of intelligibility [15]. However, a statistical correlation between the intelligibility and those speech measures was not performed. Articulation patterns and intelligibility in 54 Vietnamese individuals with un-operated and operated cleft palate in the age range of 3-24 years were evaluated in [16]. Articulation was measured by calculating the PCC, while speech intelligibility was rated on a 10-point scale. The findings showed that PCC and intelligibility were positively correlated. Authors in [17] found that the correlation between intelligibility and nasality is highly context-specific, and they observed a correlation of 0.56 between intelligibility and nasality ratings overall. A study presented in [18] compared ratings of different speech disorders and overall

intelligibility and found a good correlation between articulation ratings and intelligibility ratings. An early study showed a positive correlation between nasality ratings and speech intelligibility scores (0.720) and also between articulation errors and speech intelligibility (0.715) [19]. Therefore, it is clear that articulation error and hypernasality have a significant impact on the intelligibility of CLP individuals.

1.2 Overview of speech intelligibility assessment

Intelligibility is one of the important parameters to evaluate the severity of a speech disorder, and enhancing intelligibility is the primary concern of therapeutic intervention [20]. A means of assessing speech intelligibility is required to determine (i) the overall articulation capability, (ii) the improvement in speech due to therapy, and (iii) the outcomes of other interventions [21,22]. Intelligibility is a basic aspect of speech performance as it provides an estimate of the viability of oral communication [12,23]. Especially in children, intelligible speech is not only crucial for successful communication with the environment but also indirectly crucial for the speech and language development that is influenced by the interaction with other speakers in the society [24]. In clinical settings, speech-language pathologists (SLPs) evaluate speech intelligibility using auditory-perceptual measures, which are the de-facto standard for CLP speech assessment. An SLP makes clinical assessments of intelligibility and plans speech therapy to improve the quality of speech [23,25]. In the perceptual evaluation process, SLPs ask the patients to utter certain speech stimuli, and responses of the patients are recorded. Later, the responses are analyzed to gauge the intelligibility of the patient. To get a reliable estimation of one's intelligibility, it is mandatory to make judgments from several raters and take the average or median value as the final intelligibility score. Another important criteria to follow is that rater should not be too familiar with the patient to whom he/she is going to evaluate, as familiarization creates a positive bias on the evaluation [26].

1.2.1 Measuring the intelligibility in clinical settings

Measuring the speech intelligibility is a routine practice in clinical settings. In general sense, it can be measured as the degree to which the acoustic realization a speaker's speech can be understood [27]. However, this definition of intelligibility is not much clear to provide a quantification of it [28,29]. The process of measuring the intelligibility depends on several factors, and thus a large number of tests and scales are available for measuring different contributing aspects of intelligibility [22,28].

However, defining a reliable and accurate perceptual measure of intelligibility is still an important research direction [22]. Generally, in the perceptual evaluation, two strategies are applied to judge intelligibility (i) scale-based, and (ii) transcription-based. In the first approach, SLPs use equal-appearing rating scale in the evaluation, and several such scales ranging from a 3-point to a 10-point have been proposed [8, 25]. One such rating scale is the 4-point equal-appearing scale, where 0 represents intelligible, and 3 represents unintelligible speech [8]. In the second approach, SLPs are instructed to make an orthographic transcription of what they have heard for some predefined speech stimuli, and the degree of intelligibility is quantified using the fraction of correctly uttered phonemes/words among all the phonemes/words used for the evaluation [6, 7, 23]. For this type of intelligibility measurement, different types of speech stimuli can be used, including isolated words, pseudo-words, or sentences [22]. Transcription-based intelligibility assessment is more reliable as compared to the rating scale [10, 20]; however, it is a very time-consuming method [28, 29]. For both the methods, ratings from multiple raters are required to compensate familiarity biases. Rating scale can be used to measure intelligibility at speaker-level as well as at utterance-level, and it provides a relatively broad impression of intelligibility. However, the transcription-based measure is generally used to assess intelligibility at speaker-level. Intelligibility of the speaker can be measured at several linguistic levels, ranging from phoneme-level to conversational speech intelligibility [22, 28, 29]. One of the primary goals of speech therapy is to improve the consonant production capability of a speaker, as consonant errors have a significant effect on the intelligibility. To this end, SLPs also use the PCC as a marker for perceptual evaluation of intelligibility.

1.3 Issues in perceptual evaluation based methods

Even though perceptual evaluation is considered as the gold standard; however, it is subjective in nature, and may lead to inconsistent, unrepeatable and biased decisions [7, 30–32]. There is always some extent of intra-rater and inter-rater disagreement in perceptual evaluations [33]. Other important issues related to the perceptual evaluation are susceptible to the bias due to contextual information and familiarity biases. Also, perceptual methods are not cost effective and require expert knowledge in the field of speech pathology. Therefore, the reliability of perceptual judgment depends on the many factors, such as expert knowledge, the mood of the listeners, familiarization of the listener with the patient, and usages of speech stimuli. Moreover, the number of individuals living with

CLP is significantly high as compared to expert SLPs in a society. Another issue related to the subjective evaluation is that the process is very time consuming, and become more problematic if the number of speakers to evaluate is significantly high. According to the American Speech-Language-Hearing Association (ASHA), SLPs spend around 67% of their time doing the perceptual evaluation and conducting speech therapy [34]. As the ratio of CLP population and expert SLPs is always increasing, an alternate solution to handle this situation is urgently required. Therefore, an objective method based on the acoustic analysis of speech for assessing speech is required to assist SLPs with their therapeutic and other rehabilitation processes [6, 30, 32]. However, the primary requirements of the computational algorithm based objective measure are: (i) higher correlation with the perceptual ratings, and (ii) provide a reliable evaluation.

1.4 Overview of current objective measures

Recently, speech technology based approaches have been explored for objective assessment of speech intelligibility for different speech pathologies [7,21,31,35,36]. However, there have been relatively fewer studies reported in the literature for CLP compared to other pathologies, such as dysarthria, aphasia, and head and neck surgery. Current proposals on the intelligibility measure for CLP speech mostly concentrated on the application of automatic speech recognition (ASR) techniques. In these systems, the ASR is first trained using the speech data from normal healthy speakers. The test utterances uttered by a CLP speaker are fed to the trained ASR, which gives a transcription of recognized words. The transcription is then compared with the reference text string, and the word/phoneme error rate is used to derive the intelligibility score. This ASR-based method provides a direct way of estimating intangibility, where the ASR acts as a listener to estimate the intelligibility score of a speaker. The ASR-based system was first explored as a means of objectively quantifying the intelligibility of CLP speech in German [7, 9]. A similar system was also developed for Italian CLP speech database [6]. In these approaches, the word error rate (WER) is considered as the *degree of intelligibility*. The WER is highly correlated with the single-word perceptual intelligibility ratings [30]. Prosodic incorrectness is also combined with the WER of ASR to predict the speech intelligibility of laryngectomees and children with CLP [37]. A study showed that the WER of an ASR system differed significantly between children with cleft lips and those with either cleft palates or CLP [30]. Apart from the CLP speech, ASR-based systems have been widely used to derive intelligibility score for other types of

speech pathologies [26, 36, 39–42].

The methods based on ASR require annotated normative data as well as pathological data to build acoustic models, something that is relatively difficult for low-resource scenarios [43]. Different attempts have been made in the literature to overcome the issues associated with the ASR-based intelligibility prediction system. In [43, 44], supervectors generated from the speaker-specific GMMs are used to model the acoustic space of disordered speakers. A mapping function, using the support vector regression, is derived to mapped the supervectors into the intelligibility scores. In addition to this, i-vector based modeling is also explored to derive the intelligibility score at speaker-level for the individual with dysarthria [45–47], and patients with head and neck surgery [48]. Intelligibility measures based on ASR-systems, GMM supervectors, and i-vectors modeling use the mel-frequency cepstral coefficients (MFCCs), and perceptual linear prediction as the acoustic front-end features.

Additionally, acoustic measures related to voice quality, articulation, nasality, and prosody are used to characterize the distorted speech. Combined information of these measures is exploited to derive a composite intelligibility metric for dysarthric speakers [42, 49, 50]. A mapping function using a linear regression model is used to predict intelligibility from these features. Several acoustic features related to phonetic quality, such as zero-crossing rate, spectral centroid, spectral bandwidth, spectral flatness, spectral tilt, and spectral roll-off, are explored in this regard [42]. Prosodic features, such as duration, fundamental frequency, and speech rate are also analyzed to derive measures of intelligibility [49, 51]. Excitation source features derived from linear prediction residual to characterize the phonatory disorders are studied [50]. Alternatively, to model the distortion of the temporal dynamics, auditory-inspired modulation spectral signal representation is explored to predict the dysarthric speech intelligibility [35, 50]. The F1 and F2 formant slopes at the transition regions are found to be a significant factor of intelligibility [52]. Apart from these approaches, objective measures of speech quality using the dynamic time warping (DTW) and the Itakura-Saito distortion measure are proposed in [53, 54].

1.5 Motivation for the present work

Objective intelligibility measure using the speech technology based system is expected to be the upcoming clinical tool for assisting the SLPs in therapy and other rehabilitation processes. Uses of these approaches may contribute significantly to the assessment process in terms of (i) providing

results that are consistent and objective, (ii) allowing speech disorders to be monitored remotely over land-lines, mobile phones, and internet-based systems, and (iii) reducing the cost of care [21, 36, 40]. Another advantage of using these methods is that they rely only on the acoustic characteristics of the signal. Hence, bias due to contextual information (topic and semantic) may not be present in these methods [30]. Signal processing based methods only require a good quality microphone and a computer set-up to do the speech analysis. Therefore, it is expected that speech technology based objective measures may help to reduce the time required for SLPs in assessing intelligibility. Most of the earlier studies, based on speech technology have used word-level stimuli for evaluation [9, 30, 31, 35, 43]. However, the inclusion of sentence-level stimuli or conversational speech data, which reflects real-world communication scenarios better is also important [21]. Earlier works provided intelligibility at speaker-level, i.e., assigning an intelligibility score for each speaker. Giving a score for each sentence stimulus may be helpful to gauge the underlying cause of reduced intelligibility to some extent. Also, speaker-level intelligibility can be estimated from the knowledge of sentence-level scores obtained for the speaker.

Most of the earlier intelligibility assessment methods used ASR systems, and ASR was built using normal adults' speech data and adapted for children's data to determine the WER for intelligibility evaluation. However, the development of a reliable ASR for the normal children itself is a very challenging task, due to the mismatch between the acoustic properties of children's and adults' speech [55]. Also, as mentioned earlier, a reliable ASR-based method needs a large amount of speech as well as text data to build acoustic and language models. However, if the trained models are available beforehand, intelligibility prediction system can directly used those models to easily get the WER. Additionally, these systems are also criticized due to their complexity and unpredictability of recognition for the severe patient's speech [56]. Compared to the ASR-based methods, GMM supervectors and i-vector based measures are relatively easy to design and implement. These ASR-free systems, which exploit the acoustic deviation of the disordered speech, and do not require the transcription of the target speech. As only acoustic properties of the disordered speech are used, this approach is claimed to be language-independent [43, 44]. However, the meaning of these existing feature representation methods may not be intuitive, although they are successfully used for that task.

Apart from the estimation of overall intelligibility, a detailed evaluation is also important to diagnose and to compare the speech outcome of different types of therapy [22]. However, WER of ASR

system may not provide information about how different speech disorders tend to affect the intelligibility. In the clinical environment, SLPs are interested in assessing the intelligibility along with other parameters, such as articulation and hypernasality, to get a complete understanding of speech problems and to plan for the therapy. It is suggested that reliable and valid intelligibility measures are needed to analyze the severity of speech disorders and to provide the cause of intelligibility deficits from an acoustic and articulatory point of view [27]. Therefore, during the development of an objective measure of intelligibility, it is also important to have the capability of the measure to describe the underlying causes of intelligibility impairment, which will be helpful for the SLPs to plan for the treatment in a better way. Measures based on the combination of acoustic cues related to articulation and nasality may help in this regard. In existing methods, acoustic features related to articulation and nasality are derived frame-wise and averaged out for all the speech utterances for a speaker to represent its acoustic space [42,49,50]. Most of the times intelligibility of CLP speech is degraded due to the deviation of consonant production, and not for the vowel. Therefore, these approaches may not be suitable for CLP speech. Thus, averaging the features for all the utterances may not represent the articulation deficits explicitly. However, processing the consonants region may be more helpful to derive the measure of intelligibility of CLP speech. Additionally, no objective measures which give the gradation of hypernasality have been proposed, and such measures may be more reliable to predict the intelligibility as compared to the nasality cues only.

As mentioned earlier, the production of obstruents are mainly affected due to the loss of adequate intra-oral pressure or mislearned compensatory articulation and have a significant effect on the intelligibility degradation in CLP speech. Due to this production error, the important acoustic-phonetic cues related to obstruents, such as transient burst, frication noise, formants dynamics in the transition region of the adjacent sonorant sounds are distorted. Therefore, it is expected that deviations in their production reflected on the acoustic signal may be correlated with the perceived CLP speech intelligibility. Characterizing these deviations using relevant acoustic features are expected to provide correlates of intelligibility. This way of representing the intelligibility score may also be useful to get low-level information about the underlying causes of intelligibility loss. One solution in this direction is to locate the information-rich events in the transition between two sounds, and then, acoustic features are computed by anchoring those events, on which further processing is carried out. However, conventional MFCCs and their derivatives do not explicitly model the dynamic characteristics of the

transition region [57, 58]. As the important perceptual cues of intelligibility are embedded in these transition regions, it is essential to model such regions properly [59, 60]. In CLP speech, intelligibility degrades primarily due to the deviations in the place of articulation and manner of articulation. Acoustic cues related to the place of articulation [58] and manner of articulation [61,62] reside predominantly in the transition regions. Hence, the acoustic features which explicitly captures the spectro-temporal dynamics between two sounds may helpful.

Apart from the ASR-based measures, DTW and Itakura-Saito distortion based distance measures are also explored for objective evaluation of speech. In this case, DTW is used to align the test and the reference speech signal, and then the frame-to-frame distance is computed to evaluate the speech. The deviation from the reference speech is used to evaluate the speech. However, the deviation may not be due to the speech disorder, rather may be due to the speaker or other speech variabilities. Therefore, a speaker-independent representation of speech, which implicitly models the acoustic units in an utterance by compensating the speaker variabilities is required to explore. The comparison between reference and test speech utterances should be performed in the speaker-independent representation.

1.6 Organization of the thesis

To address the issues discussed in the previous section different attempts are made in this thesis. The content of the thesis is organized as follows.

- In Chapter 2, several existing methods for objective assessment of CLP speech intelligibility are reviewed. A few works have been reported in the literature which proposes the objective methods for CLP speech intelligibility. Therefore, a brief discussion about objective measures of intelligibility proposed for other speech disorders is also included. Finally, the limitations of existing methods along with the scope for present work are elaborated.

- In Chapter 3, a detailed description of the database collection and perceptual evaluation are provided. This chapter also provides an analysis to study the relative impact of consonant production error, hypernasality, and voice quality deviations on CLP speech intelligibility. A multiple linear regression model is applied to investigate the relative contributions, and the weights of the regression model provide relative importance of individual speech disorders on overall speech intelligibility.

- In Chapter 4, an objective measure of sentence-level intelligibility by combining the information of articulation deficits and hypernasality is proposed. These two speech disorders represent different aspects of CLP speech and showed a significant impact in reducing the intelligibility. Hence, it is expected that the composite measure based on them may utilize complementary clinical information. Initially, acoustic measures for consonant production error and hypernasality are proposed, and both measures are significantly correlated with the perceptual ratings. The objective scores of consonant production error and hypernasality are used as features to train a regression model, and output of the model is considered as the predicted intelligibility score.

- In Chapter 5, we explore how abrupt landmark's expressions deviate due to misarticulation of obstruents and shows the importance of *g(lottis)–landmarks* in deriving the intelligibility measure. The speech region around the *g–landmark* is used to compute acoustic features. Sentence-specific acoustic models are built using the extracted features from the normal speakers group. The mean log-likelihood score for each test utterance is calculated and tested as the acoustic correlates of intelligibility. Additionally, a roadmap is also provided to analyze the underlying acoustic deviation along with knowledge of intelligibility degradation.

- In Chapter 6, two approaches based on the comparison of posterior sequences are proposed to estimate the intelligibility scores. In the first approach, DTW based temporal matching between the posterior sequence of the normal template and CLP test is performed, and accumulated DTW distance is studied as the representative of intelligibility ratings. In the second approach, a unique speech representation based on self-similarity matrix (SSM) is derived from the posterior sequence, which provides additional speaker-independence. Deviation of CLP speech SSM from that of normal speech SSM is studied as the correlates of the sentence-level intelligibility.

- In Chapter 7, a system is developed for clinical application to assess the intelligibility. We define the range of intelligibility and propose a visual representation based on the spider plot for graphing the intelligibility scores. Also, the system provides the global intelligibility score for each CLP individual. This is done using the combined decision of sentence-level intelligibility scores from the previous studies.

- In Chapter 8, a detailed summary of the present work is reported by highlighting the major contributions made in this thesis. A discussion regarding the issues related to the present work

is also included in this chapter. Finally, the directions of future work related to this thesis are discussed.

2

Objective Intelligibility Assessment Methods: A Review

Contents

2.1 Introduction

Intelligibility is considered the primary measure to evaluate speech outcome of different interventions and to estimate overall speech production capability of disordered speakers [8,9]. In the clinical environment, as mentioned in the previous chapter, intelligibility is measured using perceptual evaluation, which is subjective by nature. By considering the problems associated with perceptual evaluation, researchers have been consistently focusing on the development of objective measures based on the acoustic characterization of speech signal. Several such approaches based on speech technology have been proposed in the literature to quantify and interpret the intelligibility [9,21,30,31,35,36,43]. The objective of this thesis is to develop speech technology-based measures to quantify the utterance-level intelligibility for individuals with CLP. Therefore, a detailed summarization of the state-of-the-art signal processing and machine learning based objective measures proposed for the CLP speech is required to define the motivation of our work appropriately. However, most of the existing approaches concentrate on the intelligibility evaluation of individuals with dysarthria, head and neck surgery, aphasia, Parkinson's disease, and tracheoesophageal voices. Works specific to the CLP speech intelligibility assessment, which is the aim of the present thesis, are very less in the literature. Therefore, it is necessary to study the measures which were proposed for other pathologies as well to get a complete picture of intelligibility assessment methods. In this regard, different attempts which have been made to predict the intelligibility scores are reviewed for all kinds of speech pathologies.

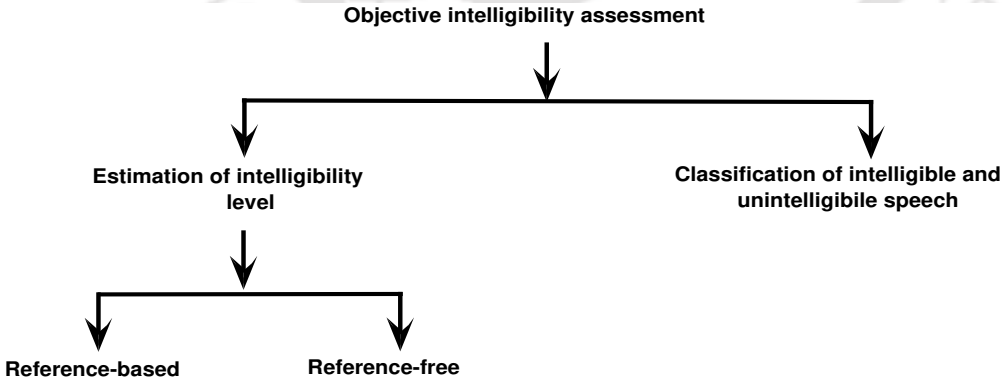

Figure 2.1: Classification of objective intelligibility assessment methods.

Figure 2.1 shows the classification of objective intelligibility assessment methods. As shown in the figure, researchers have been following two ways to evaluate speech intelligibility using computer-based systems. In the first one, an intelligibility score is derived for each speaker/utterance, utilizing the

system, and performance is evaluated by comparing the estimated scores with the perceptual ratings provided by expert raters [29]. Therefore, these systems are called intelligibility prediction systems. The second way to evaluate is to classify the intelligible and unintelligible pathological speech. However, the latter way is not appropriate for clinical applications, as the gradation of intelligibility scores is more important [21]. The objective estimation of intelligibility scores based on signal processing and machine learning techniques are classified into two categories (i) reference-based, and (ii) reference-free methods [29]. In the reference-based approach, a model/template developed from normal speech signals exists, and deviation from the reference model is used to derive the intelligibility score. However, the reference-free approach does not use any reference signal to compare with; instead, this approach identifies one or more acoustic attributes and combines them to provide intelligibility score. The issue related to the reference-based approach is the requirement of knowledge from normal speech, whereas reference-free methods only rely on degraded speech and generate an estimate of intelligibility. However, the reference-free approaches require knowledge of pathology specific characteristics, especially when prosodic cues are employed [29]. Apart from the methods proposed for objective intelligibility assessment, it is also important to review the database each measure used, to evaluate performance.

The remaining part of the chapter is organized as follows: In Section 2.2, different speech databases used in the existing literature are discussed. A detailed discussion about the different approaches to estimate the intelligibility scores using reference-based and reference-free methods is presented in Section 2.3 and Section 2.4, respectively. Section 2.5 describes the methods used to classify the intelligible and unintelligible pathological speech. Section 2.6 presents the overall discussion of the reviewed intelligibility measures and the scope of the present work. This section also provides a brief description of the works planned to address the limitations of existing approaches.

2.2 Speech databases used for intelligibility evaluation

In this section, the databases developed for intelligibility assessment is discussed. Initially, a detailed discussion about the CLP speech databases mentioned in the literature, and later, databases related to other pathologies, such as tracheoesophageal voices, dysarthria, speakers with head and neck surgery are elaborated.

2.2.1 Database of CLP speech

A German database consisting of 35 children and adolescents with CLP in the age range of 3.3–18.5 ($\mu \pm \sigma$: 8.3 $\pm$ 3.6) years was developed for the intelligibility assessment [6, 9, 37, 63]. A group of 45 children (27 girls and 18 boys) with normal speech ability were considered as the control for the study. The age range for the control group is 7.4–10.7 ($\mu \pm \sigma$: 9.5 $\pm$ 0.9) years. Speech samples from the CLP speakers were recorded with a close-talking microphone at a sampling frequency of 16 kHz, and quantized with 16 bit. The recordings were post-processed to remove the instructor voices. The children were asked to name pictures that were shown according to the PLAKSS test [64], which is a standard German speech test and consists of 33 slides which show pictograms of the words to be named. It includes all the possible phonemes of the German language in different positions (beginning, center, and end) of a word. A panel of five experts SLPs were recruited to evaluate the intelligibility of each speaker perceptually. They used a 5-point Likert scale (1 denotes very high, 2 denotes rather high, 3 denotes medium, 4 denotes rather low, 5 denotes very low) to rate the level of intelligibility. Then for each CLP individual, an average of the five raters' scores was considered as intelligibility rating for that individual.

Similar to the German database, another CLP speech database to evaluate the intelligibility was also developed for Italian language in [6, 43]. This database was collected at the Azienda Ospedaliera San Paolo Hospital, Milan, Italy. A set of 10 sentences were used as the speech stimuli for the database, and all the phonemes of the Italian language were contained in the sentences. However, the description of the sentences was not provided. Speech samples were recorded with a headphone set containing a microphone at a sampling frequency of 16 kHz. The database consists of 12 CLP individuals, with an average age of 8 years. An expert SLP who had several years of experience in the evaluation of pathological speech was recruited to assess the intelligibility. For each child, a global intelligibility score was given using a 4-point equal appearing scale of 0–3 (0 denotes within normal limits, 1 denotes mild, 2 denotes moderate, and 3 denotes severe).

2.2.2 Database of patients with cancer of oral cavity

A database consisting of 46 patients (13 female and 33 male) with cancer of the oral cavity in the age range of 34–80 ($\mu \pm \sigma$: 60 $\pm$ 10) years was developed for the work presented in [40, 65]. All the patients had surgical treatment of the oral cavity before the data recording was done. All the

speakers were asked to read the German version of the text "The North Wind and the Sun", and responses were recorded using a close-talking microphone at a sampling rate of 16 kHz. This passage is a phonetically rich text with 108 words, commonly used in German speech therapy. Four experts SLPs listened to the speech samples of each speaker and rated each speaker's intelligibility using the five-point Likert scale. Later, the average of all the raters scores was considered as intelligibility score for a patient.

2.2.3 Database of patients with tracheoesophageal voices

A work presented in [40], developed a speech database consisting of 41 patients (2 female and 39 male) with tracheoesophageal substitute voices. The age range of the patients was 56-70 ($\mu \pm \sigma$: 62 $\pm$ 8) years. The tracheoesophageal substitute voice is a treatment to restore the speaking ability after laryngectomy is done. All the patients were asked to read the German version of the text "The North Wind and the Sun", as mentioned in [65]. Speech samples were recorded at a sampling rate of 16 kHz and 16 bit quantization. 5-point Likert scale was used to rate the intelligibility by a group of five SLPs.

2.2.4 Database of patients with oral squamous cell carcinoma

The database consisting of 46 patients with oral carcinomas and an age-matched control healthy group of forty speakers without oral diseases was developed in [39]. All the participants were native German speakers. Forty speakers without any diseases speaking the same language served as the control group. The speech stimuli used for the databases discussed in [40,65] was considered. Moreover, the recording procedure and the perceptual evaluation of the speech samples were also similar as described in [40, 65].

2.2.5 Database of dysarthric speech

A database of 60 dysarthric speakers speech was developed for the work presented in [26,41]. The Dutch Intelligibility Assessment test was used, and this test was specifically designed to measure the intelligibility at phoneme level for the Dutch language. Each speaker was asked to read 50 consonant-vowel-consonant words, and responses were recorded with a microphone in a quiet environment. The intelligibility score is calculated as the percentage of correctly identified phonemes, and one experienced SLP rated all speech samples. It was found that the intelligibility ratings of the dysarthric speakers ranged from 36% to 100% with mean intelligibility of 79.1%.

A publicly available database for intelligibility evaluation is the Universal Access Speech database [66]. The database consists of 19 subjects diagnosed with cerebral palsy of age range 18–58 years. All the participants were native speakers of English. Each of the speakers participated in the recording procedure was asked to read isolated words displayed on a computer screen. A total of 765 isolated-word utterances were recorded per participant. Speech samples are recorded at a sampling frequency of 16 kHz and a 16-bit resolution. Transcription-based intelligibility assessment is performed to judge the intelligibility of each speaker. Five listeners were recruited to judge the intelligibility, and they were instructed to give the orthographic transcription of the recorded speech samples. The transcription was then analyzed and the mean percentage of correct responses averaged across the five listeners, was calculated to obtain the subjective intelligibility score of each dysarthric speaker. Then each speaker intelligibility was classified into one of the four categories, viz., very low (0–25%), low (26–50%), mid (51–75%), and high (76–100%).

Another dysarthric speech database for intelligibility assessment was developed for the work presented in [67]. Speech data from 174 native Korean speakers with dysarthria were considered, while 30 non-dysarthric speakers (20 males and 10 females) served as the control. All speakers were asked to utter the Assessment of Phonology and Articulation for Children words, which are commonly used for assessing articulation disorders in Korea. Speech samples were recorded with a Shure SM12A head-worn microphone at 16 kHz sampling rate in a mono-channel. Later, each dysarthric speaker was evaluated by an expert SLP using the transcription-based method.

2.2.6 Database of patients with head and neck surgery

A widely used speech database for intelligibility assessment is the NKI CCRT Speech Corpus [68], which contains utterance-level speech audio and its perceptual intelligibility score. The database consists of read speech samples of 17 Dutch sentences spoken by 55 head and neck cancer patients. All 55 patients underwent concomitant chemo-radiation treatment. This database was recorded at the Department of Head and Neck Oncology and Surgery, Netherlands Cancer Institute, Netherlands. Each utterance of the database was evaluated by a group of thirteen native Dutch-speaking SLPs. A seven-point scale ranging from very poor intelligibility (1) to very good intelligibility (7) was used. For each utterance, the average of 13 SLPs ratings was used to provide intelligibility scores ranging from 2.0 to 6.7.

2.3 Intelligibility level estimation using reference-based approaches

Here, the basic principle is to first derive the quantity of interest, such as a phonetic transcription of utterances spoken by the disordered speakers and posterior probabilities for each phone class, and then compare that with the reference text. In pathological speech intelligibility estimation, ASR technique is often used for the reference-based approach. Also, the distance-based measure is also explored in this context to quantify the intelligibility.

2.3.1 ASR-based approach for intelligibility prediction

ASR system based on Hidden Markov model (HMM) is widely used for the assessment of pathological speech intelligibility. It is considered as the unbiased listener to evaluate intelligibility, though the reliability of these systems is yet to improve much to be used in the clinical settings. In these systems, test speech data from disordered speakers is input to an already built ASR using normal speech, and the lexical decoding of input speech is performed. Later, the percentage of correctly decoded phonemes/word is used as the representation of intelligibility. A very low recognition rate represents degraded intelligibility, whereas a high recognition rate represents highly intelligible speech. Currently, researchers have explored the Google speech to text conversion API to get the transcription of disordered speech, and these automatically obtained transcription is compared with the reference text to derive the measure of intelligibility [69].

CLP speech intelligibility assessment:

The first attempt to predict CLP speech intelligibility using the ASR system was made in [70]. Authors considered 12-dimensional MFCCs with 12 delta coefficients as acoustic front-end for the speech recognizer. The semi-continuous HMM-based acoustic modeling was used to build the ASR, where the codebook contains 500 Gaussian densities and all HMM states shared these Gaussians. Authors used polyphones as the elementary recognition units in their work. The unigram language model was used to estimate the prior probability of a word sequence. The acoustic models were trained using normal children's data and to make the ASR system more robust adults' data was adapted with the vocal tract length normalization technique. For each CLP individual, word accuracy (WA) was computed, and it was considered as the proxy of intelligibility. The WA (%) is computed as,

$$\text{WA} = \frac{C - I}{R} \times 100 \tag{2.1}$$

where, 'C' is the number of correctly recognized words by the recognizer and 'R' is the total number of words present in the reference paragraph. 'I' is the number of wrongly inserted words. Authors found that correlation between the WA and the perceptual intelligibility was more when adaption was performed. Further, the authors extended the work to study the effect of different cleft types on WA in [9]. They found that WA was higher in the case of isolated cleft lip and less for the individual with bilateral CLP. This is because, a few sounds are distorted in case of the isolated cleft lip; however, if the cleft occurs in the palate as well as lip most of the oral pressure consonants are distorted. For both the studies, single-word speech stimuli were used, and the description of the database used in those works are discussed in Section 2.2.1. Further, word recognition (WR) accuracy was also studied as the biomarker for intelligibility in [63] using the similar system proposed in [9, 70]. The WR (%) is computed using the following equation,

$$\text{WR} = C/R \times 100 \tag{2.2}$$

Authors in [6] reported the suitability of ASR-based intelligibility estimation method proposed in [9,70] for Italian language. HMM was used with 39-dimensional MFCCs as acoustic front-end to build the Italian ASR system. For both training and testing phases, 39-dimensional acoustic features were normalized on a speaker-by-speaker basis to compensate speaker, channel, and environmental variability. Acoustic models for recognition were tied-state, cross-word triphone HMMs.

To improve the performance of intelligibility prediction, authors in [37] incorporated the prosodic cues derived from the fundamental frequency (F_0) and the energy of signal along with WA and WR and fed them to a regression model. The WA and WR for each CLP individual were computed from the ASR developed in [9, 70]. For each word, 37 prosodic features were computed, and the mean, maximum, minimum, and variance of these features were calculated for each speaker, thereby prosody of the entire speech utterances of a speaker was characterized by a 148-dimensional vector. Finally, the resultant 150-dimensional (WA + WR + 148-dimensional prosodic feature) vector was mapped to the intelligibility score using a regression model. A work presented in [32] developed an isolated-word recognition system to predict the intelligibility levels of CLP speech in a database provided by the American Cleft Palate-Craniofacial Association. MFCCs were considered as the front-end, and the GMM was used to build the isolated-word recognition system. They found that as intelligibility degrades, the WA of recognition system also decreases. However, the authors did not specify how the

intelligibility evaluation was done. Moreover, the development of the word recognizer was also not very clear.

Intelligibility assessment of other pathologies:

Apart from the CLP speech, ASR technology has been explored for assessing the intelligibility of other speech pathologies. In one of the works, ASR was used to derive the measure of intelligibility for adult patients with cancer of oral cavity [65]. The database discussed in Section 2.2.2 was used in their work. Authors found a correlation of 0.92 between the estimated WR and the corresponding expert listeners scores. It is observed from their study that automatic evaluation provided results with less variance as compared to expert listeners decision. Authors have pointed out that automatic evaluation methods are independent of biases due to the contextual information which influences the perceptual ratings. However, it is important that other factors which influence the word recognition systems should be minimized, and speaker's acoustic space information should be the only primary factor to influence. Other factors, such as speech stimuli and input medium can be minimized using the standard text as stimulus and a stable setting. The system, as proposed in [65] was also explored for the objective assessment of speech intelligibility of patients with oral carcinomas [39]. The primary motivation was to assess the improvement of postoperative speech outcome for these individuals. Authors used the database discussed in Section 2.2.4. A correlation value of 0.93 was achieved between the expert listener's evaluation and objective scores.

The importance of automatic evaluation of pathological speech over the telephone line was studied in [40]. The usages of the telephone channel may reduce the patient's burden to visit the clinic, and the cost of care regularly. However, the problem may arise as the telephone channel has an adverse effect on the signal quality due to bandwidth and channel distortion, which reduces the perceptual quality of the speech signal. Due to this problem, the automatic speech evaluation over the telephone may be less correlated with the perceptual ratings. To overcome this problem, the authors combined the knowledge of two speech recognizers. The motivation for combining two speech recognizers was that independent recognizer produces different errors, and the combination of them may reduce the inaccuracies caused by recognition errors. Authors adapted two approaches to combine the speech recognizers, namely, the direct combination of output generated from the two recognizers, and fed WA and WR as the input of an SVR model. The output of the model was used as a predicted intelligibility score. The authors in [40] found that the combined system outperforms the single speech recognizer

performance in evaluating the pathological speech.

Authors in [41, 71] also used the ASR-based technique, but a slightly different way without using WA or WR as the marker for intelligibility. The recorded pathological speech was forced aligned using a reference model, which was a whole-word HMM trained from normal speakers' speech [71]. The likelihood score computed for a dysarthric speech utterance from the reference model is considered as the measure of intelligibility. The ASR system was also used to estimate the phoneme intelligibility of dysarthric speakers in [41]. The ASR was developed based on the context-independent acoustic models of phonemic units. Based on this speech alignment, a set of phonemic and phonological features were extracted for the all utterances of a speaker, and this features set characterized the respective speaker. The phonological model was trained on multiple phonemes which share the same features, and authors expected that it might better extrapolate the acoustic space of a disordered person. Then all the extracted features were fed to a regression model to predict the intelligibility score of the speaker. The highest correlation (0.943) between the perceptual and the objective intelligibility scores were found for models combining phonemic and phonological features. However, the method proposed in [41] incorporated both the acoustic and phonological feature spaces to predict the intelligibility, and made the system more complicated. Thus, in [26] authors modified the method and showed that only phonological features were sufficient to obtain good performance. They showed that good results could be obtained using a simple ASR system that comprises of 55 context-dependent acoustic models. They used a forward selection algorithm to select the best phonological features to predict the intelligibility.

Feature representation and selection from the higher dimensional data is very important in predicting the speech intelligibility of disordered speech [42]. A detailed comparison of different feature selection methods, e.g., forward selection and maximal relevance selection methods was provided in [42]. However, they lacked in systematic selection strategy or the selection was performed through greedy wrapper-based feature selection, such as forward selection which is likely to incur disadvantages from over-fitting and performance limitation [67]. A work reported in [67] addressed these issues and proposed a method for feature representation and selection. An ASR was used to decode the phone sequence, and the output phone sequence was aligned with the canonical phone sequence from a pronunciation dictionary using a weighted finite-state transducer to capture the pronunciation mappings. The histograms of the pronunciation mappings on a pre-defined word set were used as the features. In

the prediction step, a structured sparse linear model incorporated with phonological knowledge that simultaneously addressed phonologically structured sparse feature selection was proposed. Authors found a root mean square error of 8.14 while comparing the estimated scores with the perceptual ratings.

The ASR-based methods are also used to automatically assess three aspects of aphasic speech intelligibility: clarity, fluidity, and prosody of persons with aphasia in [72]. Authors used forced-alignment-based techniques for automatic transcript generation that perform well on aphasic speech in spite of limited data and atypical speech input. The deep neural network based acoustic modeling and its out-of-domain adaptation were performed. Different features based on transcript, pronunciation, rhythm, and intonation were extracted to classify the intelligibility.

2.3.2 Distance measure based intelligibility prediction

Authors in [53, 54] proposed objective measures of speech quality assessment using the DTW technique. The proposed algorithm was evaluated on the speech of patients with Parkinson's disease. The speech signal from a healthy adult speaker was considered as the reference template with which the comparison was made. Initially, the feature vector sequence of the reference utterance and the test utterance was aligned using the DTW algorithm, and then several distance measures, such as Itakura-Saito and log-likelihood ratio were used to calculate the frame-to-frame distance between the test and the reference speech. The resultant scores were studied as the measure of speech quality, and correlation analysis was performed between these proposed objective scores and the subjective ratings. Authors found that Itakura-Saito is strongly correlated with the subjective scores compared to LLR scores.

2.4 Intelligibility level estimation using reference-free approaches

In the reference-free approach, alternative feature representation for the acoustic space of disordered speech is investigated besides the ASR techniques. These representations may be based on GMM supervectors, and total variability subspace modeling and they are called as transform domain features. Additionally, acoustic measures related to voice quality, articulation, nasality, and prosody can also be used to characterize the distorted speech. Later, using a feature selection method best features to predict the intelligibility are chosen for each speaker. A mapping function is derived between the resultant feature vectors and the corresponding perceptual intelligibility ratings using the

regression model, and the regression model is used to predict the intelligibility of a test feature vector. Generally, linear regression and support vector regression models are used in these approaches.

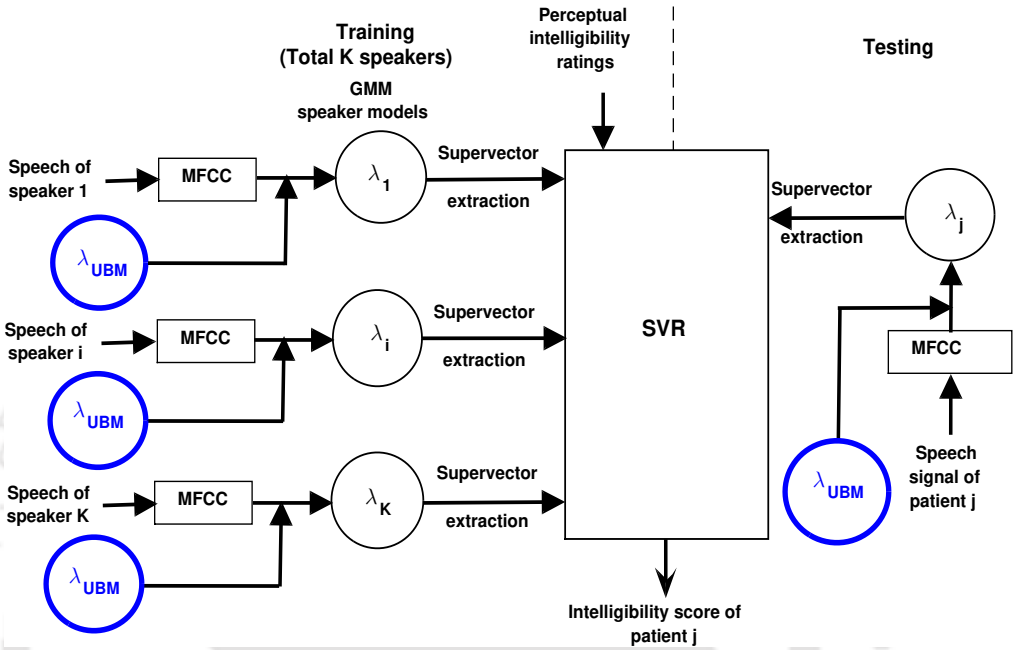

Figure 2.2: Block diagram of GMM supervector based intelligibility assessment method [43].

2.4.1 Intelligibility measure based on transform domain features

GMM supervector based approach:

Supervectors generated from speaker-specific GMMs were used to quantify the speech intelligibility in [43]. Figure 2.2 shows the block diagram of GMM supervector and SVR based approach for intelligibility assessment [43]. For each speaker of the training set, a GMM λ was created. The GMM for each speaker was adapted from a universal background model (λ_{UBM}) using the MAP adaptation. Then, the different realization of GMM supervectors generated from mean, weights, and covariance matrix were considered. These supervectors were labeled with the expert's intelligibility score and used as input vectors for an SVR, and output of the SVR was considered as the metric for intelligibility. Since the system is solely based on the acoustic properties of the speaker, it is most likely that the system is language independent. To show the language independent nature of the system, the authors evaluated the system on German and Italian languages (see Section 2.2.1).

It was found that the intelligibility evaluation of speech with partial laryngectomy using the ASR-based method provides significantly low correlation [44]. The reason behind the less correlation may be due to the higher and more uniform voice quality of partially laryngectomized persons, in which

case word accuracy alone may not be a sufficient feature to achieve high correlations to the perceptual intelligibility ratings [44]. To overcome this issue, authors in [44] explored GMM supervector based acoustic characterization of the disordered speaker as discussed in [43]. The system achieved a correlation up to 0.83 for covariance matrix based supervector with PLP as an acoustic feature. The GMM supervector based system proposed in [44] was also extended for the intelligibility evaluation for patients with tracheoesophageal substitute voices after total laryngectomy in [73].

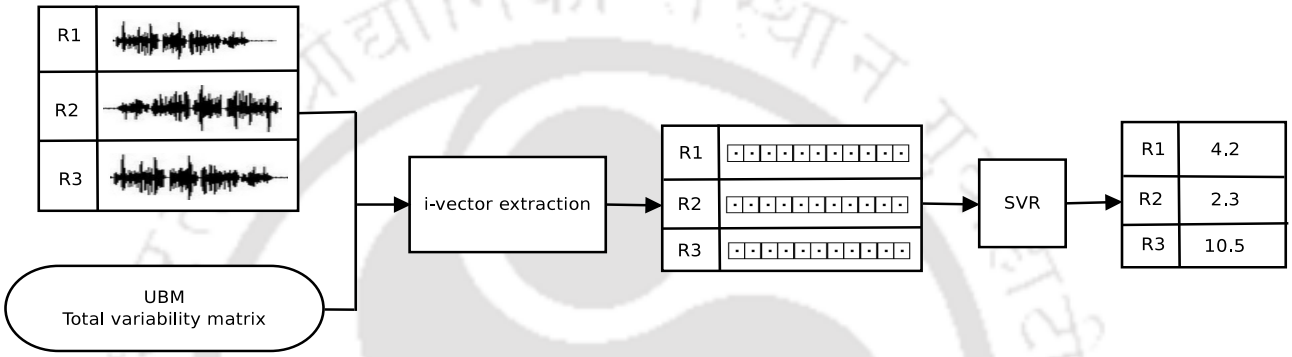

Figure 2.3: Block diagram of i-vector based intelligibility assessment [48].

i-vector based approach:

Apart from the GMM-based supervectors, the i-vector based modeling of speech acoustics has been also explored for intelligibility prediction. The i-vector is a data-driven representation of speech acoustics, where the sequence of feature vectors for a given utterance mapped into a low-dimensional vector space, i.e., total variability space, based on a factor analysis technique [74]. It was found that i-vector can represent the speaker characteristics very well. Currently, the i-vector representation, which compresses all the acoustic information present in an utterance into a reduced number of features, has gained research importance for objective evaluation of pathological speech intelligibility. A block diagram describing the i-vector based intelligibility prediction algorithm is shown in Figure 2.3. Initially, speech utterances of a speaker are parametrized by cepstral features and represented speech utterance in the total variability subspace. Then, a regression model is used to map the i-vector subspace to the intelligibility score, which is one dimensional. Authors in [75] used i-vector representation along with a large set of acoustic, syllable-level, and phonotactic features to predict the UPDRS ratings of Parkinson's Disease patients. Several researchers used the i-vector representation of the utterances produced by dysarthric speakers to predict the intelligibility [45–47]. Authors in [48] used the i-vector

paradigm to automatically predict the speech intelligibility of head and neck cancer patients, and they obtained a high correlation with the perceptual intelligibility.

Phonological features based approach:

The methods proposed in [26,41] used two ASR systems to predict phoneme intelligibility (PI) using the word-level data. However, the PI is moderately correlated with the speech ability in a more realistic situation, where the running speech is a mode of communication. It is also important to derive a measure for the running speech intelligibility, which reflects the natural communication scenario more. However, using an ASR-based system may not be very reliable as it might be required to handle difficulties arises for the out-of-vocabulary condition due to the errors. To overcome this situation, an ASR-free system was proposed in [76]. Authors used eight phonetically rich sentences of German as the speech stimuli. The proposed system relies on phonological feature detectors that were trained once on a sufficiently large database of normal speech. The artificial neural network-based phonological feature extractor was trained on a corpus of read speech of 174 normal speakers. Authors used the MFCCs as acoustic front-end, and each MFCC vector was mapped to the phonological feature vector. Then, statistical analysis of each component of phonological feature vector was performed to derive one feature vector per speaker. Linear regression and SVR based intelligibility prediction modules take the phonological feature vectors of all the speakers and their corresponding intelligibility ratings and train the regression models. During the testing, the phonological feature of a speaker was fed to the regression model, and the output of the model was considered as the intelligibility score. Results showed that the proposed ASR-free system outperforms the earlier ASR-based system [26, 41].

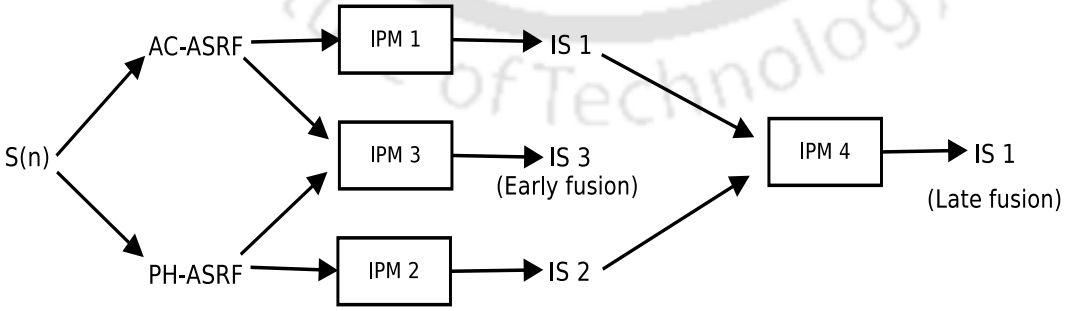

Figure 2.4: Block diagram combination of PC-ASRF and PH-ASRF based intelligibility assessment method [77].

The ASR-free methods for intelligibility prediction based on GMM supervectors [44,73] and phonological feature detectors [26,41] are gaining more importance primarily due to their language indepen-

dence nature and simple working principle. Since both the methods capture different characteristics of the patient's speech signal, it is important to investigate whether combining them really outperforms the individuals. In [77], authors showed the effectiveness of intelligibility prediction for two databases of different languages and pathologies. They used German partial laryngectomees database [73] and Flemish pathological speech [26, 41]. For each speaker, GMM-based supervector, termed as acoustical ASR-free features (AC-ASRF) and phonological feature vector, termed as phonological ASR-free features (PH-ASRF) were extracted as discussed in [44, 73] and [26, 41], respectively. As shown in Figure 2.4, four different intelligibility prediction models (IPMs) per dataset were created to predict the intelligibility score (IS). IPM 1 used the AC-ASRF feature, while IPM 2 used the PH-ASRF feature, which is considered as the baseline. The two others (IPM 3 and IPM 4) employ combinations of both feature sets using late and early fusion based techniques, respectively. Authors used two machine learning models to build each IPMs, one based on ensemble linear regression with feature selection, and another one was based on the SVR. Results showed that combining the two feature sets in one system was beneficial.

2.4.2 Intelligibility measure based on acoustic features

Features related to the phonetic quality, which characterize the distorted spectral and temporal information of pathological speech were also explored to predict the intelligibility in [42]. Authors used several acoustic features, such as zero-crossing rate, spectral centroid, spectral bandwidth, spectral flatness, spectral tilt, spectral roll-off, spectral flux, and second-order spectral flux, which did not need the acoustic model to derive. These features are computed at frame-level; therefore, to consider the temporal behavior of each feature, statistical parameters of them were computed for each utterance. Then, for each speaker, the average value for all the utterances was considered, and each speaker was characterized using a high dimensional feature vector. Since all the dimensions may not be important to describe the intelligibility of a speaker, a feature selection strategy was applied to select the best features. All selected speaker-level features need to be converted into an objective intelligibility score for a speaker, and linear regression was applied for that, and the estimated intelligibility score was compared with the perceptual ratings.

Generally, the dysarthric speech can be characterized by voice quality, articulation, nasality, and prosody, and each speech dimension effectively provides complementary information in assessing the intelligibility [49]. Thus, it is important to consider all the speech dimensions while deriving a measure

of dysarthric speech intelligibility. Method proposed in [42] only explored one dimension to assess the intelligibility. However, the authors did not use the articulation features explicitly, and phonetic features are used instead. Considering the importance of other speech dimensions authors extended their work to include the prosody and voice quality along with phonetic quality for intelligibility score prediction [51]. The prosody features related to the duration, fundamental frequency, and speech rate was computed. The LP residual was considered as the voice source excitation signal, and different features, such as mean, variance, skewness, and kurtosis of both the LP residual signal and the ZCR of LP residual signal in utterance-level were computed. The utterance-level features were averaged across all utterances of each speaker, and then discriminative features were selected from that. The speaker-level feature vector was mapped into an objective intelligibility score for a speaker using the SVR. All experiments were performed in leave-one-out cross-validation. Results showed significant improvement in performance over the work presented in [42].

Imprecise articulation is a primary characteristic of the dysarthric speech, and it often causes the perturbations in temporal dynamics of speech. Due to this distortion, intelligibility is significantly reduced for individuals with dysarthria. Therefore, it is expected that acoustic cues related to the temporal dynamics of speech may be used to derive the objective measure of intelligibility. Motivated by this information, authors in [50] explored the short and long-term temporal dynamics measures, which were computed based on the log-energy temporal dynamics information and auditory-inspired modulation spectral signal representation, respectively. In addition to these measures, the total duration of voiced segments within the uttered word was also used. Finally, a composite metric of intelligibility was derived by linearly combining all the explored features. Apart from the distortion of temporal dynamics, the dysarthric speech can be characterized by the atypical voice source, nasality, and the distortion of prosodic cues. Authors in [35] extended the work presented in [50] to derive a composite measure which incorporates all the information. Atypical voice source was characterized by the kurtosis of the linear prediction residual. The first two formant frequencies and their bandwidths were used as the nasality measures, and variation in fundamental frequency was considered as the prosody measure. Similar to [50], a regression model is used to predict the intelligibility using all the derived features.

Intelligibility assessment based on the combined knowledge of different speech subsystems, such as articulatory, resonatory, and laryngeal are gaining more research interest. A work proposed in [78]

extracted different acoustic features which properly represent these subsystems fed to a regression model, and the output of the model predicts the intelligible score of the speaker. Authors carried out the experiments on the speech dataset of individuals with cerebral palsy. Temporal duration, vowel spectral cues, nasality cues, and voice measures were selected as speech acoustic measure to describe the speech subsystems. To represent the articulatory subsystem, F1 and F2 formants of the vowel, duration of the vowel, and F2 formant's slope in the transition region are considered. Since the resonatory subsystem is mainly dependent on the functionality of velopharyngeal valve, acoustic cues of nasalization such as the difference between the amplitudes of the first formant and the extra peak (A1-P1) introduced by oro-nasal coupling are considered. To represent the laryngeal subsystem, the average value of fundamental frequency and signal-to-noise ratio are considered. Authors found that the articulatory subsystems much better explain the variation in intelligibility than the other subsystems.

A work reported in [52] studied the features derived from statistics of first and second formant trajectories and their first and second derivatives to analyze the intelligibility of persons with amyotrophic lateral sclerosis. Features were also computed from the lowpass and high pass versions of the formant trajectories. Apart from the above-mentioned acoustic features, the importance of cepstral peak prominence for predicting the intelligibility loss in dysphonic speech was shown in [79]. Authors found that cepstral peak prominence is moderately correlated with the perceptive intelligibility, and they suggested that cepstral peak prominence can not alone describe the dysphonic speech intelligibility.

2.4.3 Acoustic landmark analysis for intelligibility prediction

Acoustic landmarks are defined as the time locations of abrupt acoustic events in a speech signal, which are correlated with the major articulatory movements [80–83]. Authors in [81] exploit the landmarks to characterize the dysarthric speech intelligibility. They observed that landmark analysis provides information about how a set of acoustic cues are not able to produce and insert unnecessarily by the dysarthric speakers. It was found that the rate of landmark detection, deletion, substitution, and insertion correlated with the perceived intelligibility ratings. These event rates were mapped to the intelligibility score using a multiple linear regression model.

2.5 Classification of intelligible and unintelligible speech

A work to classify the intelligible and unintelligible speech was presented in [84] using the NKI CCRT Speech Corpus (see Section 2.2.6). In the work, authors used different features, such as voice quality, spectral and harmonicity, and hierarchical features to characterize the pathological voice. The spectral feature comprises the MFCCs, RASTA style filtered auditory spectra and their delta features. The harmonicity feature attempts to capture the various form of periodicity disturbances in the acoustic signal, and this feature is found important to characterize the pathological voice. The hierarchical features provide the temporal modeling of the acoustic feature contours, and they used functionals applied to low-level descriptors. The intelligibility detection is based on the fusion of the individual linear dimensionality reductions, such as asymmetric sparse partial least squares (ASPLS) trained by different sets of normalized features. Experimental results showed that the proposed method achieves the accuracy of 71.88% on the unweighted recall value on the test set.

The importance of auditory-inspired spectro-temporal modulations as a front-end acoustic feature for the intelligibility assessment is also studied in the literature. The motivation for using these features is the importance of spectro-temporal modulations in perceived intelligibility, and distortion in those modulations results in loss of intelligibility. A work presented in [85] explored these spectro-temporal modulations in classifying the intelligible speech from the unintelligible for the NKI CCRT Speech Corpus [68]. Authors found that the unintelligible speech tends to have its modulation amplitude peaks shift towards a smaller rate and scale. Since the feature dimension was very large, the tensor principal component analysis was applied to perform dimensionality reduction of the features. The support vector machine (SVM) and GMM were used as the classifier in their work. Experimental results showed an accuracy of 68.3% on the unweighted recall value on the test set.

It is important to consider the long-term rhythm disturbances in the signal while designing the intelligibility classification algorithm, especially for the dysarthric speech. A work proposed in [86] explored the acoustic cues at different time scales that capture short-term voicing and long-term rhythm distortions. Authors extracted important features at different time scales (resolution of phonetic, segmental, and suprasegmental information), using the phoneme-level, vowel/consonant level, and the sentence-level. Table 2.1 shows the features extracted at different linguistic levels. Nonlinear classifiers were trained at each time scale, and a final intelligibility decision was made using ensemble learning method. The classifiers were trained on a training set and tested on a development set, and found

Table 2.1: Perceptual cues to correlate to intelligibility and their respective feature set at different levels.

Levels	Perceptual cues	Feature set
Sentence	Rate, rhythm, and prosody	Envelope modulation spectrum (EMS)
	Nasality, breathiness, loudness variation	Long-term average spectrum (LTAS)
Vowel/ Consonant	Unnatural pitch/formant contours	Basic speech descriptors
	Vowel space reduction	Formant structure statistics
	Articulatory Imprecision	Vocal tract statistics
	Vocal quality (nasality and breathiness)	Spectral energy distribution
Phoneme	Distortions/substitutions	Silence statistics, spectral statistics
	Rate and rhythm	Duration

unweighted recall of 94.4%, which outperforms the results presented in [87]. All results are presented for the NKI CCRT Speech Corpus as mentioned earlier.

A feature extraction method based on the multi-resolution sinusoidal transform coding framework was proposed for intelligibility assessment, and effectiveness of the features was studied using the NKI CCRT Speech Corpus. The motivation for using this framework for its high quality and accuracy in representing spectral properties including high frequencies that are often ignored in other representations [88]. One of the early symptoms of neck or laryngeal cancer is degradation in voice quality. This degradation is due to a decrease of higher harmonics in the source spectrum; therefore, it is essential to characterize the higher harmonics in the spectrum [88]. The classification of intelligible vs. unintelligible groups was performed as discussed in [84–86]. The experimental results clearly showed that the features derived from the multi-resolution spectral domains were useful for speech intelligibility assessment.

Authors in [21] classify the intelligible and unintelligible speech using the combined evidence from prosodic, voice quality, and pronunciation subsystems. While doing the classification, it may so happen that enough data to cover the wide variability of pathological speech during the training phase is not present. Also, there can be a speaker-related mismatch between the train and the development databases. To overcome these situations, authors also proposed a post-classification smoothing scheme that makes a final decision on a test sample based on the likelihood score of both the test speech samples itself and other samples in the test set. Each subsystem initially extracts sentence-level features to capture the distorted prosodic, phonation, and articulation behavior from the speaker. Then, the authors performed feature-level fusions and subsystem decision fusion for arriving at a final intelligibility decision. Different classifiers, namely, linear discriminant analysis,

k-nearest neighbor and SVM were explored, and best results were found using the SVM classifier.

2.6 Summary and discussion

In the previous sections, a detailed review of different approaches proposed for objective intelligibility assessment of pathological speech is presented. The review also focuses on the speech databases widely used in literature for the intelligibility prediction. In this section, the advantages and disadvantages of the explored methods are discussed, and the scope for the present work is highlighted. The intelligibility of pathological speech can be evaluated in two ways using the computer-based systems: (1) by predicting the intelligibility scores and compare the predicted scores with the perceptual ratings provided by the expert raters, and (2) by providing the binary classification of intelligible vs. unintelligible speech. However, the earlier one is more relevant to the clinical applications [21], and the objective of the present thesis is also to predict the intelligibility scores.

Generally, HMM-based ASR systems are widely used for the estimation of intelligibility scores. In these systems, the phone/word error rate was considered as the measure of intelligibility. The phone/word error rate showed significant correlation with the perceptual intelligibility rating. Speech is parametrized using the MFCCs and PLP features. Methods based on the ASR system require a large amount of annotated normative data to build acoustic models, something that is relatively difficult for low-resource scenarios [43], unless trained models are available. Another important criticism regarding these systems is that they are trained only on non-pathological speech and the result may not be predictable for a very severely degraded speaker [47]. Moreover, it is difficult to port one system to another language, as lots of training data and the corresponding transcriptions are needed. The atypical speech behavior of CLP individual may also be challenging for the unconstrained ASR system. In most of the previous work, ASR was built with adult data and then adapted for child speech to determine word accuracy for intelligibility evaluation. The acoustic mismatch between adult and child speech presents a considerable challenge for child ASR trained on adult data [89]. Therefore, using those approaches to assess CLP speech may be unreliable. In ASR-based systems, WER gives a global view of intelligibility for each CLP individual. Moreover, WER does not provide information about how different speech disorders tend to affect intelligibility. In some of the investigations, researchers have employed phonological information in the feature representation by averaging the posterior probabilities of phones which are associated with identical phonological attributes and then mapped the

feature representation to the intelligibility scores. However, it was found that this kind of approach requires a set of phonetically diverse utterances from a single speaker; therefore, it may be challenging to derive intelligibility score at the utterance-level using this approach.

Apart from the ASR-based measure, DTW and Itakura-Saito distortion based distance measures are also explored as the reference-based measures. The DTW is used to align the test and reference speech signal, and then the frame-to-frame distance is computed to evaluate the speech. The deviation from the reference speech is used to evaluate the disordered speech. However, the deviation may not be due to the acoustic deviation, instead may be due to the speaker or other speech variabilities. Moreover, the same speaker's speech is impossible to obtain as the template in the case of pathological speech analysis. Therefore, a speaker-independent representation, which implicitly models the acoustic units of the utterance by compensating the speaker variabilities is required. Authors in **??** overcome this situation by utilizing the phoneme posterior probability sequences for the intelligibility assessment of text-to-speech systems. The works proposed in [26, 41] used the phonological posterior vector to predict the dysarthric speech intelligibility. However, they computed the posterior vectors in a supervised way, which requires a huge amount of annotated data to build the models. Moreover, to calculate the phonological features, additional effort needs to be given to categorize the phonemes into voicing, place of articulation, turbulence, nasality, etc. Thus, posterior vectors computed in an unsupervised way, such as using the GMM, may be explored.

Other approaches for intelligibility prediction does not use a reference model, rather one or more features which represent the deviant acoustic characteristics of the pathological speakers are extracted and are combined them to predict the intelligibility using a regression model. The degradation of intelligibility primarily occurs due to the distortion of several speech dimensions, such as articulation, phonation, prosody, and nasality. Therefore, the acoustic correlates of these perceptual dimensions can be combined to derive a composite measure of intelligibility. One primary advantage of the approach is that it does not require *a priori* knowledge about the signal characteristics of the target word being uttered. Additionally, using these approaches, clinicians may able to know which dimensions are most affected, and it will help the clinician to plan for the proper therapy. Unlike the ASR-based methods, these measures explore the explicit knowledge of speech production and perception to assess intelligibility. It was found that articulation errors and hypernasality have the strongest impact on intelligibility loss in case of CLP speech [13,20]. Hence, it is hypothesized that combining the knowledge

of both articulation error and hypernasality may serve as an estimate of CLP speech intelligibility. Moreover, both the speech disorders represent different aspects of CLP speech production, thereby composite measure based on them may utilize complementary clinical information. However, no such composite measures have been derived for the speech assessment of CLP speech. Alternate to these acoustic features, transform domain representations of the acoustic space of each speaker, such as GMM-based supervectors, and i-vectors are also explored. Compared to the ASR based methods, these approaches are relatively easy to design and implement. As only the acoustic properties of the disordered speech are used, this approach is claimed to be language-independent. However, the meaning of these existing feature representation methods may not be intuitive, although they were successfully used for that task. Therefore, it is difficult to analyze further and interpret prediction result [67].

To address the above-mentioned issues, different attempts are made in this thesis to quantify the speech intelligibility of individuals with CLP. Along with the intelligibility score, this thesis also tries to provide some insight into the underlying causes of intelligibility degradation. Most of the earlier objective measures provide intelligibility scores at the speaker level, but a few studies also explored the intelligibility prediction at utterance-level. Therefore, there is a scope to work for the prediction of the utterance-level intelligibility of CLP speech. Moreover, no work has been attempted in this direction for CLP speech. Since no database for the CLP speech is available in the public domain, one such database will be created by collaborating with AIISH, Mysuru, India.

Researchers have been found that articulation error and hypernasality have a significant effect on the degradation of CLP speech intelligibility. Motivated by this finding, an objective measure of sentence-level intelligibility can be proposed by combining the information of articulation deficits and hypernasality. The objective scores of consonant production error and hypernasality will be used as the features to train a regression model, and the output of the model is considered as the predicted intelligibility score. Since the intelligibility in CLP speech is mainly affected due to problems in obstruents production; therefore, acoustic events which are associated with the obstruents to sonorant sounds, or vice-versa should be more useful. Therefore, the *g(lottis)–landmarks* can be used as the anchored points to extract relevant features. Also, the presence and absence of these landmarks may provide some insight into the cause of intelligibility degradation. The acoustic characteristics of articulatory deviations near *g–landmarks* will be used to derive the correlates of cleft lip and palate

speech intelligibility. Sentence-specific acoustic models will be built using these features extracted from the normal speakers' group. The mean log-likelihood score for each test utterance will be tested as the acoustic correlate of intelligibility. A visual representation based on the two-dimensional plot of the mean log-likelihood scores vs. the number of detected *g–landmarks* for an utterance will be explored to infer some insight about the underlying cause of intelligibility loss.

Apart from the above measures, two comparison based approaches will be explored to estimate the intelligibility scores. In the first approach, acoustic features computed from an utterance are mapped to a speaker-independent representation using Gaussian posteriorgrams, then DTW based temporal matching technique between the normal template and CLP test is applied and accumulated DTW distance is studied as the representative of intelligibility ratings. In the second approach, a unique speech representation based on the self-similarity matrix (SSM) is derived from the feature sequence, which provides additional speaker-independence. Deviation of CLP speech SSM from that of normal speech SSM is studied as the correlates of the sentence-level intelligibility. The intelligibility scores should be represented visually so that it can be readily interpreted by the SLPs. Spider plot-based visual representation will be proposed for graphing intelligibility scores. Apart from these, the subject-specific intelligibility scores will be predicted using the sentence-level scores. Since ten sentence-level stimuli are used in this thesis, evidence from all the sentences is used to derive the speaker-specific intelligibility.

3

CLP Speech Database Development and Subjective Analysis of Intelligibility

Publications

- **Sishir Kalita**, Pushpavathi M, Ajish K Abraham, Girish K S, S. R. M. Prasanna, S. Dandapat, "Relative contribution of hypernasality, consonant production errors, and voice disorders on the intelligibility of cleft lip and palate speech", in *Proc. Workshop on Speech Processing for Voice, Speech and Hearing Disorders (WSPD)*, Mysuru, India, September 2018.
- **Sishir Kalita**, "Objective assessment of cleft lip and palate speech intelligibility", 4^{th} *Doctoral consortium, Interspeech 2018*, Hyderabad, India, September 2018.

Contents

Objective

Present chapter discusses the development of CLP speech database, which to be used in the subsequent chapters for the objective intelligibility assessment. The procedure to conduct the perceptual evaluation of speech recordings are also mentioned. Prior to derive the objective intelligibility measures, it is important to study the impact of different speech disorders, such as hypernasality, articulation error, and voice disorder in reducing the CLP speech intelligibility. Therefore, an experiment is performed to study the relative contributions of these speech disorders on the intelligibility deficits. Results showed that the articulation error has the highest impact in degrading the intelligibility, while the voice disorder has significantly less contribution.

3.1 Introduction

Speech disorders due to VPD or mislearning lead to reduce intelligibility, and it requires a detailed assessment by SLPs [12]. It is also essential to know how different speech disorders affect intelligibility. In literature, as mentioned in the introduction chapter, several researchers have studied the effect of speech disorders on reducing intelligibility [10, 12–14]. These studies analyze the impact of individual speech disorders separately on the degradation of intelligibility. It is important to study the degree of the relative contribution of each speech disorder on intelligibility. It will provide the information whether all the speech disorders are important to derive the objective measure of intelligibility. No such study has been reported for the CLP speech database of Kannada language. Motivated by the factors mentioned above, present work studies the contribution of different speech disorders on the intelligibility of CLP speech. It is hypothesized that the intelligibility of CLP speech can be defined as the linear combination of hypernasality, articulation errors, and voice quality by using the multiple linear regression model.

The rest of the chapter is organized as follows: Section 3.2 provides a detailed description of the database development, while details of the perceptual evaluation are discussed in Section 3.3. The methodology to study the relative contribution of speech disorders on intelligibility deficits is discussed in Section 3.4. Section 3.5 gives a detailed discussion about the results, and finally, Section 3.6 summarizes the works present in the chapter.

3.2 Database development

Up to the best of our knowledge of from the existing literature, no database of CLP speech with corresponding intelligibility ratings is available in the public domain. Therefore, one such database is created by collaborating with the SLPs of AIISH, Mysuru, India.

3.2.1 Speakers details

Each participant had either a repaired CLP or a repaired cleft palate and was a native speaker of Kannada aged between 7-12 years. None of the CLP individuals had any history of hearing impairment or other congenital syndromes or developmental difficulties. Additionally, each CLP individual had adequate language abilities. Children with normal speech and language characteristics who were matched for age and gender served as controls for the study. The details of speakers in both groups are shown in Table 3.1. Before the recording, ethical consent was obtained from the parents/caregivers of each speaker. Parents/caregivers were provided with information about the aim, objective and approximate duration of the testing procedure. The present study was conducted with clearance from the AIISH Bio-behavioral ethical committee.

Table 3.1: Description of CLP and normal speakers

	CLP	Normal
Total number	42	42
Number of females, males	18, 24	22, 20
Age ($\mu \pm \sigma$)	8.79 ± 1.94	9.8 ± 1.42

3.2.2 Speech stimuli

In this thesis, two different sets of sentences are used. Ten sentences with rich in obstruent consonants are used for the performance evaluation of proposed intelligibility measure and this set of sentences are referred to as *oral sentences*. These sentences are designed by SLPs of AIISH, Mysore for intelligibility and hypernasality assessment of Kannada CLP individuals. Another set includes

Table 3.2: Description of oral sentence stimuli (Written in IPA).

O1 kaːge kaːlu kappu, **O2** giːʈa beːga hoːgu, **O3** ɖana ɖaːri ʈappiʈu, **O4** baːlu ʈabala baːrisu, **O5** beːɖa kaːɖige oːɖiɖa, **O6** sariʈa kaʈʈari ʈaː, **O7** ʃivana uːru kaːʃi **O8** ʧaːʧaː ʧapaːʈi koɖu, **O9** paʈa paʈa b^{h}aːvuʈa, **O10** ʈaːʈa ʈabala ʈaː

Table 3.3: Description of nasal sentence stimuli (Written in IPA).

N1 manu aːnejannu noḍiḍ̃a, **N2** naviːna maneji nd̃a band̃anu,
N3 naːnu aːnejannu noḍiḍ̃e, **N4** manga maneja meːliḍ̃e,
N5 maːmaː mand͡ʒad̃inda band̃anu, **N6** maːmana mane manga͡ɭuːrinalliḍ̃e,
N7 miːnaː͡ɭige negad̠ibanḍiḍ̃e, **N8** nari neladinḍa negejɪt̪u

Table 3.4: Correlations of the individual SLPs (raters) to the mean of the other SLPs

SLP (Rater)	Mean to other raters	
	ρ	κ
Rater 1	0.80	0.61
Rater 2	0.79	0.60
Rater 3	0.81	0.63

only sentences which are rich in nasal consonants and nasalized vowels, and termed as *nasal sentences*. Present work only uses the *nasal sentences* to build acoustic models for hypernasality detection module in Chapter 4, and this set of sentences are not used for the intelligibility assessment. Table 3.2, and Table 3.3 list the *oral sentences* and *nasal sentences*, respectively. From now onwards, sentence-level stimuli will only represent the *oral sentences*, unless specified.

Speech samples are recorded in a sound-proof room using a Bruel & Kjaer unidirectional microphone with sampling frequency of 44 kHz and 16-bit resolution on a mono channel. The microphone was kept at a distance of 15 cm from each child while recording. For each sentence, 2-3 sessions of recording are conducted for the normal groups. Thus, we collected approximately 120 utterances from all the normal speakers for each sentence-level stimulus. For the CLP group, a total of 420 (10 sentences/speaker×42 speakers) sentences are recorded and perceptual evaluation of these 420 sentences are performed. In the present study, the recording environment was same for both the normal and CLP groups.

Table 3.5: Details of speakers in each intelligibility level for each sentence-level stimuli

# sentence / **Intelligibility rating**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
0	8	9	8	7	8	7	7	6	7	7
1	11	11	10	13	12	8	11	11	9	11
2	15	15	14	13	14	15	13	13	14	15
3	7	6	9	8	7	11	10	12	11	7

3.3 Speech assessment

Three SLPs were recruited to conduct the perceptual evaluation of each speech utterance for all the CLP individuals. Each SLP had around five years of experience in the field of CLP speech evaluation and they were instructed to assess the sentence-level intelligibility, hypernasality, articulation, and voice quality by the auditory-perceptual based method. The SLPs are not familiar with the considered CLP speakers; hence, the speaker bias will be minimized. The evaluation was conducted in a sound-proof room and all the three SLPs used the same computer set-up to listen to the samples. SLPs were allowed to listen to one sample as many times as needed before making a decision. All three SLPs rated every utterance in the database. The order of the samples presented to SLPs for evaluation were randomized. The evaluation was performed in three sessions and accomplished in two conjugative days. In this work, we considered a 4-point equal-appearing-interval (EPI) scale ranged from 0 to 3 [8, 25]. Here, 0 indicates that speech is highly intelligible, and 3 represents that speech is not intelligible at all. This scale describes intelligibility deficits using sentence-level stimuli. Table 3.5 provides a detailed description of the number of CLP individual belongs to each intelligibility level for respective sentence-level stimulus. The perceptual evaluation of hypernasality for the oral sentences is carried out by three expert SLPs using Henningsson 4-point EPI scale as defined in [25]. The 4-point hypernasality rating scale consists of values in the range of 0–3, where 0 represents normal, 1 represents mild, 2 represents moderate, and 3 represents severe [8,25]. The articulation was measured by calculating the percentage of consonants correct (PCC) for each sentence. To do this, SLPs listened to each sentence and wrote the transcription in the international phonetic alphabet (IPA). Later, the percentage of the correctly produced consonants out of the total number of consonants in a respective sentence was considered as the PCC measure. From now onwards, in this chapter, the articulation score of an utterance represents the PCC score. The voice quality was also assessed using a 4-point rating scale to measure the overall severity. To compare the rating agreement, we computed Spearman's rank correlation coefficient (ρ) and Cohen's kappa (κ) between the score of an individual rater and the mean of the other two raters. From Table 3.4, it can be seen that the ratings are sufficiently reliable to be considered as the ground truth. Also, we have computed Fleiss' kappa, which is used to assess the agreement between more than two raters. We found $\kappa = 0.62$ ($p < 0.01$) with confidence interval (95%; 0.59, 0.66), which is quite reliable to consider as the ground truth. Herein, we considered the median score of the three raters as the ground truth.

The intelligibility, hypernasality, and articulation ratings for each CLP speaker are computed by combining the individual sentence-level scores of 10 stimuli. The summation of 10 sentence-level scores is computed, and that score is normalized to [0, 1] by dividing the maximum attainable rating. The maximum attainable rating for intelligibility for speaker is 30 (maximum intelligibility rating (3) $\times$ number of sentence stimuli (10)). This is the same for the hypernasality and voice quality. If the summation of all the ten sentence-level intelligibility ratings is A_i, then the intelligibility rating for that speaker will be,

$$G_i = A_i/30. \tag{3.1}$$

Similar to intelligibility, the summation of all sentence-level hypernasality ratings (A_h) is mapped to [0, 1] using the following equation,

$$G_h = A_h/30. \tag{3.2}$$

The consonant production error is measured by calculating the percentage of consonants correct (PCC) for each sentence. In this case, CLP individual with all the consonants error has PCC score of 100, while PCC score is 0 for the individual with no consonant production errors. Thus, the summation of all the sentences ratings (A_p) should be divided by 1000 (maximum PCC score (100) $\times$ number of sentence stimuli (10)) to mapped the ratings [0, 1] as shown in the following equation,

$$G_p = A_p/1000. \tag{3.3}$$

Similar to intelligibility and hypernasality, the summation of all sentences ratings (A_v) is mapped to [0, 1] using the following equation,

$$G_v = A_v/30. \tag{3.4}$$

3.4 Relative contribution of speech disorders on intelligibility deficits

To study the relative impact of speech disorders on intelligibility deficits, we have considered randomly selected 75% data of total data, i.e., 32 speakers speech data out of 42 CLP speakers.

3.4.1 Method

Initially, we have investigated the strength of perceptual rating of each speech disorder to describe the overall intelligibility loss. Ratings of one speech disorder are considered as the independent variable and intelligibility ratings as the dependent variable to build a linear regression model. Leave-one-speaker-out cross-validation (LOSO-CV) is applied to the evaluation process of each speech dimension. For each fold of LOSO-CV, except for one CLP speaker ratings, all the individuals' ratings are used to build the linear regression model. Then the derived regression model is used to estimate the intelligibility rating of the remaining one. This process is repeated for 32 times, as speech data of 32 CLP individuals are considered, and there will be 32 linear regression models. Pearson correlation coefficient between estimated scores and perceptual intelligibility ratings for all the 32 folds are computed. The averaged R^2, F-value, and p-value of 32 linear regression models are used for the statistical analysis. The parameters are used to interpret linear regression output statistics. The R^2 measures how well a regression model predicts the dependent variable, which is intelligibility in our case. The value of R^2 falls between 0 and 1, and the higher value of R^2 signifies the better model. The R^2 is the ratio of the sum of squares for error (SSE) and corrected sum of squares for the model (SSM), and can be calculated using the following equation.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSM}} = 1 - \frac{\sum_{i=1}^{N}(I_i - \hat{I}_i)^2}{\sum_{i=1}^{N}(I_i - \bar{I})^2}, \tag{3.5}$$

where I and $\hat{I}$ represent the ground truth intelligibility scores and the estimated intelligibility scores, respectively. $\bar{I}$ is the mean value of I. N is the number of points, in our case it the number of speakers, i.e., 32. The F-value in a regression model determines whether the model fits significantly better than a model consisting of only an intercept. The F-value can be computed as the ratio of the mean of squares for the model (MSM) and the mean of squares for error (MSE). The MSM and MSE are calculated as follows,

$$\text{MSM} = \frac{\text{SSM}}{\text{Corrected degrees of freedom for model (DFM)}} = \frac{\sum_{i=1}^{N}(I_i - \bar{I})^2}{p-1}, \tag{3.6}$$

$$\text{MSE} = \frac{\text{SSE}}{\text{Degrees of freedom for error (DFE)}} = \frac{\sum_{i=1}^{N}(I_i - \hat{I}_i)^2}{N-p}, \tag{3.7}$$

where p is the number of parameters in the regression model. Finally, the significance in correlation difference between each speech dimension is analyzed using Williams pairwise statistical significance test [90, 91], and the test is performed for each pair of speech dimensions.

The primary objective of this work is to analyze whether the deficits of CLP speech intelligibility can be expressed as the weighted linear combination of all the mentioned speech disorders, as given below,

$$\text{Intelligibility} = \beta_1 \times \text{PCC} + \beta_2 \times \text{HN} + \beta_3 \times \text{VQ}, \tag{3.8}$$

where HN, PCC, and VQ represent the perceptual ratings of hypernasality, articulation, and voice quality. β_1, β_2, and β_3 are the regression coefficients. A multiple linear regression model is built to investigate the degree of the relative contribution of each speech dimension on the overall intelligibility. Dependent (intelligibility ratings) and independent (PCC, HN, and VQ ratings) variables of the model are normalized to the zero mean and unit variance. LOSO-CV is used to derive the weights of the regression model. The averaged absolute weights of the 32 regression models are considered as the final weights of the model. The averaged absolute weights provide the degree of contribution of each dimension on CLP speech intelligibility. Average values of R^2, F-score, and p-value of 32 regression models are used for the statistical analysis. At each fold, the multiple linear regression model is used to estimate the intelligibility of the test speech samples. The Pearson correlation coefficient is computed between the predicted intelligibility scores of the 32 folds and the corresponding perceptual ratings of intelligibility.

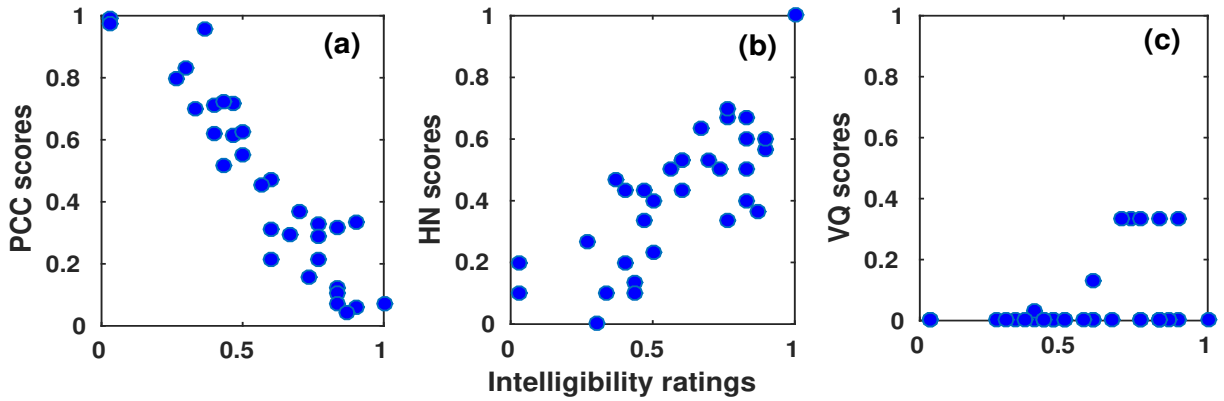

Figure 3.1: Scatter plots of intelligibility ratings with respect to (a) PCC scores, (b) HN scores, and (c) VQ scores.

3.5 Results and discussion

Table 3.6 lists the results of correlation and regression analysis for the strength of each speech disorder to describe the overall CLP speech intelligibility. Pearson correlation coefficient (r) is computed

Table 3.6: Correlation and regression analysis between perceptual intelligibility scores and other measures of speech disorder. In the table, absolute of correlation values is mentioned.

	Correlation analysis		Regression analysis		
Measures	r	p-value	R^2	F-score	p-value
PCC	-0.919	<0.001	0.867	190.0	< 0.001
HN	0.726	<0.001	0.576	39.53	< 0.001
VQ	0.244	0.178	0.124	4.147	0.053

Table 3.7: p-value of the Williams pairwise significance test between perceptual intelligibility and PCC, HN, and VQ.

-	0.01	4.90×10^{-5}	PCC
-	-	0.01	HN
-	-	-	VQ
PCC	HN	VQ	

between PCC, HN, and VQ scores and intelligibility ratings individually. From the table, it can be seen that PCC has the highest correlation (r = 0.919) with CLP speech intelligibility. PCC measures the articulation capability of the CLP individual to produce the consonants. Literature also suggests that consonant production error have the most significant effect on reducing the intelligibility of CLP individuals, which support the results observed in this work. The correlation between hypernasality and intelligibility ratings is r = 0.726 and the relationship is less statistically significant (R^2 = 0.576, F = 39.53) compared to the relationship of PCC and intelligibility ratings (R^2 = 0.867, F = 190.0). It is found that the correlation between intelligibility ratings and VQ is lower and not statistically significant at $p < 0.001$. Table 3.7 gives the results of Williams test for different pairs of speech dimensions. In the table, each p-value inside a cell (i, j) represents that measure i (named on the right side of the table) correlates significantly higher with intelligibility ratings than measure j (named on the bottom of the table). From the table, it can be noted that the increased correlation with intelligibility rating in the case of PCC and HN that of VQ is statistically significant at $p < 0.05$. Also, the increased correlation in case of PCC than that of HN is statistically significant. The contour

Table 3.8: Correlation and regression analysis between perceptual intelligibility scores and other measures of speech disorder

	Correlation analysis		Regression analysis		
Measures	r	p-value	R^2	F-score	p-value
PCC+HN+VQ	0.925	<0.001	0.885	71.97	<0.001
PCC+HN	0.929	<0.001	0.887	111.23	<0.001

Table 3.9: The β coefficients and intervals with 95% of confidence of the linear regression analysis

	PCC	HN	VQ
β	-0.662	0.232	0.0491
Intervals	-0.830 ± 0.504	0.020 ± 0.446	-0.237 ± 0.335

plots of the intelligibility ratings vs PCC, HN, and VQ scores are shown in Figure 3.1 (a), (b), and (c), respectively. From the contour plot, it is evident that there exists no correlation between the perceived intelligibility and voice quality (Figure 3.1(c)).

The results of multiple linear regression are shown in Table 3.8. While combining the three speech dimensions to estimate the intelligibility a correlation of 0.925 is noted. The significance of only PCC and HN to describe the intelligibility by excluding the effect of VQ is also shown in the table. From Table 3.8, it can be observed that the correlation value is similar for both the PCC + HN + VQ and PCC + HN models. However, a higher F-score (111.23) is observed if we exclude voice quality ratings from the predictor variables. The β weights and intervals with 95% of confidence in the multiple regression model are shown in Table 3.9. The CLP speech intelligibility can be expressed as the linear combination of PCC, HN, and VQ using the β weights as follows,

$$\text{Intelligibility} = -0.662 \times \text{PCC} + 0.232 \times \text{HN} + 0.0491 \times \text{VQ}. \tag{3.9}$$

Equation 3.9 demonstrated the relative contribution of the different speech dimensions on intelligibility. The effect of VQ on intelligibility loss is statistically insignificant; therefore, Equation 3.9 can be rewritten by excluding VQ, as shown in Equation 3.10.

$$\text{Intelligibility} = -0.662 \times \text{PCC} + 0.232 \times \text{HN}. \tag{3.10}$$

From the equation, it can be observed that the PCC has the highest impact on CLP speech intelligibility, whereas VQ has the lowest impact. These findings seem to support the observation noted from the correlation between the ratings of speech dimensions and intelligibility.

3.6 Summary

This chapter presents the procedure for database development and the protocol used for perceptual assessment of the intelligibility. Additionally, the present work also studies the contribution of different speech disorders, such as articulation errors, hypernasality, and voice disorder on the CLP speech

intelligibility. From the analysis, it is found that articulation error errors, measured by a parameter called PCC, have the highest correlation with the intelligibility ratings. The least or no correlation is observed for voice quality rating with the intelligibility. Multiple linear regression shows that PCC has the highest contribution (-0.662) to the overall speech intelligibility, while hypernasality has a comparatively low contribution (0.232) on the CLP speech intelligibility. Voice quality has no contribution (0.0491) on the loss of intelligibility. The relative impact of each speech dimension on the speech intelligibility may provide some insight in deriving the objective measure of intelligibility in the next chapters.

4

Intelligibility Measure Based on the Articulation and Hypernasality Information

Publications

- **Sishir Kalita**, Girish K S, Pushpavathi M, S. R. M. Prasanna, S. Dandapat, "Objective assessment of cleft lip and palate speech intelligibility using articulation and hypernasality measures", *The Journal of the Acoustical Society of America*, 146(2), August (2019), PMID: 31472592.

Contents

Objective

The objective of this chapter is to develop a measure of sentence-level intelligibility by combining the information of articulation deficits and hypernasality. These two speech disorders represent different aspects of CLP speech. Hence, it is expected that the composite measure based on them may utilize complementary clinical information. Initially, a measure for articulation deficits using the acoustic features extracted around vowel onset and vowel end points is proposed. Processing of the speech by anchoring around these two events is motivated by the perceptual importance of transition region for consonants. The joint spectro-temporal based features from the overlapping patches of time-frequency representation for better characterization of spectral and temporal modulations in the transition region is explored. In this work, articulation deficits represent the error in consonant production for CLP speech. An algorithm to objectively quantify the hypernasality level is proposed, which is motivated from the working principle of the Nasometer instrument. Later, the objective scores of articulation and hypernasality are used as the features to train a regression model, and output of the model is considered as the predicted intelligibility score. This predicted score is studied as the acoustic correlate of perceptual intelligibility rating, and Spearman's correlation coefficient based analysis shows a significant correlation between the predicted and perceptual intelligibility scores.

4.1 Introduction

The speech of individuals with CLP is characterized by different speech-related disorders, such as hypernasality, articulation errors, and voice disorder, which affect the subject's speech intelligibility [4, 38]. In the previous chapter, we have shown the impact of different speech disorders on the CLP speech intelligibility. Moreover, several researchers have studied the effect of speech disorders in reducing intelligibility [10, 12–14]. Articulation error and hypernasality are found to be the primary contributors of the intelligibility degradation [20, 92]. The PCC score, which is a representative of the consonant production capability of CLP speakers showed a significant correlation with the intelligibility. Therefore, it is hypothesized that combined knowledge of both articulation error and hypernasality may serve as an estimate of CLP speech intelligibility. Both the speech disorders represent different aspects of CLP speech production, thereby composite measure based on them may utilize complementary clinical information. In this chapter, a measure of intelligibility degradation is proposed for CLP speech by combining the evidence from consonant production errors and hy-

pernasality. Researchers have shown the significance of combining the different speech aspects, viz., phonation, articulation, nasality, and prosody to monitor the disease progression and to evaluate the speech intelligibility [21, 35, 93]. In one such study, phonation features (jitter and shimmer), prosody features (pitch and energy temporal variation), and articulation features were used for GMM supervector and i-vector based individual speaker modeling [93]. Then distances corresponding to the phonation, prosody, and articulation aspects were combined to derived a single measure using the linear regression method and used this measure to evaluate neurological state and dysarthria level of the Parkinson disease patients. The possibility of deriving an estimator of dysarthric word-intelligibility by linearly combining the glottal source excitation, temporal dynamics and prosody features was shown in [35]. They found that the composite measure is significantly correlated with the subjective scores. Authors in [21] classify the intelligible and unintelligible using the combined evidence from prosodic, voice quality, and pronunciation subsystems. However, no such composite measure has been derived for the CLP speech, and correlation of acoustic measure of articulation and hypernasality with perceived intelligibility scores have not been studied. Prior deriving the metric of intelligibility using the articulation error and hypernasality information, a brief overview of the existing objective measures based on acoustic analysis to evaluate these two speech disorders is provided.

4.1.1 Existing methods to evaluate articulation error and hypernasality

Articulation error evaluation

Articulation error is associated with the deviation of position, stress, and shape of the speech articulators involved in the speech production process from the normal condition [94]. Due to this, speech sounds are not properly articulated, and sometimes replaced by other sounds or entirely deleted the target sound. In the clinical environment, improving articulation is considered as one of the primary goals of treatment, as increasing the articulation capability of phonemes enhance the intelligibility of speech. In the literature, different methods have been explored to quantify the articulation capability and to analyze the SODA (substitution, omission, distortion, and addition). However, specifically for CLP speech, a few works have been reported. ASR is widely used to evaluate the pronunciation score of each phoneme and SODA analysis of individual with speech disorders [95–98]. A most widely used measure to derive the articulation score is the goodness of pronunciation [99, 100]. In the ASR-based approach, phone-level log-likelihood scores and decoded phoneme sequence for a test utterance are used to evaluate the articulation scores. In the case of CLP speech, ASR-based method for the eval-

uation of different articulation error, such as pharyngeal backing, laryngeal replacement, nasalized obstruents using automatic and semi-automatic methods was proposed in [101]. Authors used the automatic phone segmentation using the forced-alignment, and acoustic features were derived from the phone segments to evaluate the articulation. The primary issue related to the ASR-based approach is it requires a considerable amount of normal speakers speech data to build the acoustic models. Also, while computing the articulation scores, the automatic phone segmentation may not be proper for the test utterance of disordered speech due to the occurrence out-of-vocabulary sounds unit.

Apart from the ASR-based approaches, several researchers explored the explicit knowledge of speech production to characterize the deviation in discorded speech. Authors in [93, 102] extracted features from the transition region of unvoiced to voiced sounds and vice-versa. The GMM supervectors and i-vectors based representation were used to estimate the articulation impairment of patient with Parkinson disease. Authors in [103] proposed a method to detect the omission of the initial consonant of a syllable by identifying the acoustic differentiations between initial consonants and finals for Mandarin CLP speakers. Automatic detection of glottal stop misarticulation in CLP speech was proposed in [104], where authors used the MFCCs, formants, features based on Gammatone filtering energy and wavelet packet energy, and Shannon entropy. To represent the utterance-level pronunciation variation formants cepstral mean normalized MFCCs and phone duration, and their statistics are used in [21]. Earlier methods mainly analyzed whether uttered speech possesses any articulation disorders and provide information about SODA. However, no objective measure to quantify the sentence-level consonant production error has been proposed for CLP speech.

Hypernasality evaluation

Hypernasality refers to the perception of excessive nasal resonances on vowels and voiced consonants [2, 105]. It is considered as an important parameter during the evaluation of the outcome of the surgery and the speech therapy of individuals with CLP. Currently, perceptual evaluation and instrumental based methods are used to assess the hypernasality. A detailed review of different instrumental methods for the assessment of hypernasality can be found in [106]. Among the different instrumental based hypernasality assessment techniques, Nasometer is used widely in the clinical and research applications [106]. It operates on the real-time speech data, and provide the impression of nasality in terms of nasalance score. However, Nasometer cannot be operated on stored speech data and also, requires the subject's cooperation and technically trained persons to handle it.

Apart from perceptual and instrumental methods, speech processing based techniques have been proposed for the hypernasality evaluation. These measures do not require complex hardware and give objective evaluation results [106, 107]. The presence of extra-nasal formants around 250 Hz and 1000 Hz in vowel spectrum, increase in the first formant bandwidth, reduction in second formant strength, and an increase in the spectral flatness are considered as the important acoustic cues of hypernasality [108–110]. The Teager energy operator [111], voice low-tone to high-tone ratio [112], MFCCs, glottal source related features (jitter and shimmer), and wavelet transform based features have been extensively used for the hypernasality evaluation [103, 113, 114]. The GMM and SVM classifiers have been used to classify the hypernasal speech from the normal for hypernasality detection. Authors in [115] explored the vowel space area in combination with MFCCs to detect the hypernasality. Also, automatic classification of speech into normal, mild, moderate, and severe levels of hypernasality is proposed in [103, 113], where GMMs are explicitly trained for these classes. In most of the earlier methods, isolated vowels /a/, /i/, and /u/ or vowels are taken from word and sentence-level data are considered as the stimuli to evaluate hypernasality.

However, these methods have the limitation to use for the clinical applications, as they only classify the hypernasal speech from the normal. The results are not in the form of continuous scores like Nasometer instrument. The estimation of the degree of hypernasality is less attempted in the literature, and the algorithm needs speech data from each class of hypernasality level [103, 113]. Hypernasality is continuous and difficult to obtain from these algorithms. The estimation of the severity of hypernasality is most essential to verify the effect of speech therapy and surgery. Additionally, vowels are only considered as the stimuli to evaluate the hypernasality. However, the inclusion of other voiced sounds may provide a comprehensive evaluation of the hypernasality.

4.1.2 Contributions

Present work proposes a metric of intelligibility based on the combined information of articulation error and hypernasality. Since no utterance-level objective measure to quantify the articulation error has been proposed for CLP speech, one such measure is proposed beforehand. Articulation error of CLP speech primarily represents the misarticulation of the production of consonant, more specifically the misarticulation of pressure consonants. It may be due to malocclusion, compensatory errors, obligatory errors, and nasal air emission. Therefore, the proposed objective measure for articulation deficits is the acoustic correlate of the consonant production error. Features extracted from the speech

regions around the vowel onset points (VOPs) and vowel end points (VEPs) are used to derive the acoustic measure of the consonant production error. Since the transition region between two sounds is used to extract the features, thus we explore the joint spectro-temporal feature derived from the time-frequency representation (TFR) of speech [57]. For the estimation of the degree of hypernasality, an algorithm is proposed. In this algorithm, speech signals representing two extremely opposite cases of nasality are used to develop the acoustic models, where oral sentences (rich in vowels, stops, and fricatives) of normal speakers and nasal sentences (rich in nasals and nasalized vowels) of moderate-severe hypernasal speakers represent the groups with minimum and maximum attainable degrees of nasality, respectively. The acoustic features derived from glottal activity regions are used to model the maximum and the minimum nasality classes using the GMM. Finally, a composite measure is derived by fusing both the articulation and hypernasality scores using a regression model. The derived measure is tested as the acoustic correlates of CLP speech intelligibility. The proposed composite measure offers quantification of utterance-level intelligibility. Thus, the major contributions of the present chapter are as follows.

- Derive an objective measure to evaluate the consonant production error and study its correlation with perceived intelligibility ratings.
- Proposal of a novel algorithm for the estimation of the degree of hypernasality.
- Analyze the correlation between objective hypernasality scores and perceived intelligibility ratings.
- Derive a composite measure of intelligibility based on the relative contribution of both articulation and hypernasality scores.
- Analyzing the importance of joint spectro-temporal features for deriving intelligibility measure.

The rest of the chapter is organized as follows: The procedure to compute a composite measure of intelligibility from the objective scores for articulation and hypernasality is discussed in Section 4.2. Results and discussion of the proposed method are included in Section 4.3. Finally, Section 4.4 concludes the chapter by summarizing the work.

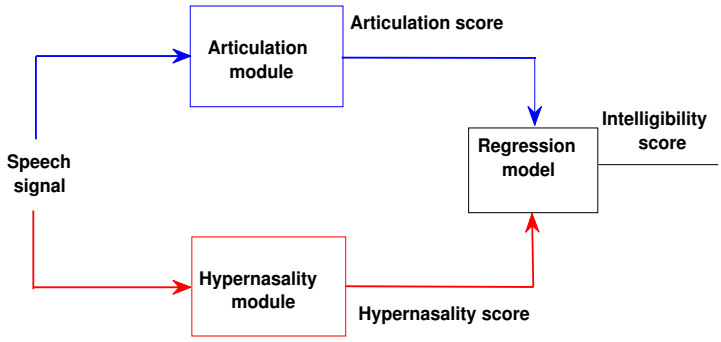

Figure 4.1: Block diagram for deriving a composite measure of intelligibility based on articulation and hypernasality scores.

4.2 Development of composite measure of intelligibility

This section describes the development of the proposed composite intelligibility measure. Figure 4.1 shows the block diagram for deriving the composite measure based on the articulation and hypernasality measures. A test speech signal is passed through the articulation and hypernasality modules, and articulation and hypernasality scores are computed for that speech signal. Then, those two scores are mapped to a single quantity using a trained regression model. This single quantity is considered as the composite intelligibility measure in this work and tested as the correlate of perceptual intelligibility. The regression model is trained by considering the perceptual ratings and both articulation and hypernasality scores as the dependent and independent variables, respectively. Initially, in this section, we discuss the procedure to derive an objective measure for the articulation. Later, a detailed discussion about the estimation of hypernasality scores from the sentence-level stimuli is provided. Finally, a composite measure is derived using both the acoustic measures for intelligibility assessment. All the speech signals are normalized by the maximum of the absolute value of the signal to keep the amplitude of speech signal in the interval [-1, 1]. To reduce the computation complexity speech signals are down-sampled to 16 kHz before performing the experiments.

4.2.1 Objective measure for articulation

An objective measure is proposed to characterize the consonants production capability of CLP speakers. This measure is analyzed as the correlate of articulation/PCC scores provided by the SLPs. It has been well studied that the consonant-vowel (CV) and vowel-consonant (VC) transition regions are the salient regions in speech signal to perceive the consonant production [59, 116]. Such regions

contain several important acoustic cues of obstruent production, such as burst, frication, aspiration, formant transition, and play an important role in the speech recognition [61,62]. These acoustic cues define the place and manner of articulation of obstruents. In CLP speech, the articulation errors mainly exhibit for the obstruents consonants due to the inadequate build-up of intra-oral pressure and mislearning [2,8]. When articulation error, such as compensation of laryngeal/pharyngeal consonants, weakening of obstruents, velar and nasal substitutions occur in the speech signals, the acoustic characteristics of the target sounds are entirely distorted. Due to this distortion, the most essential static and dynamic acoustic cues of consonant production present in the CV and VC transition regions may deviate. These transition regions are anchored around the VOPs and VEPs, and these speech events mark the boundary between consonant and vowel, and vice-versa. It is hypothesized that features extracted from the region anchored around these events may convey the information about the articulation error of consonants. Researchers have shown the importance of VOPs in the evaluation of articulatory disorders in Parkinson's disease [117] and the detection of the place of articulation of stop consonants [58]. The performance of the CV unit recognition system improves significantly if the acoustic features are extracted anchored around VOPs [116]. The slope of F1 and F2 formants in the CV transition region was to be found very significant for the assessment of intelligibility for speakers with cerebral palsy [78]. Moreover, Kent et al. in [118], reported F2 formant as an important predictor of speech intelligibility. Therefore, processing the speech around VOPs and VEPs may be helpful to characterize the acoustic deviations due to articulation error.

Motivated from the above discussion, present work proposes an objective measure of articulation by characterizing the acoustic deviations around the VOPs and VEPs. For each sentence, an acoustic model is built using the features extracted from transition regions around these events from the normal speaker's speech. The deviation of test utterance from the normal model is quantified, and it is considered as the measure of overall consonant articulation error for that utterance. The steps to derive the objective measure of articulation are given below.

- Detection of VOPs and VEPs in the speech signal.
- Consider a speech region of 80 ms around each detected VOPs and VEPs.
- Extract acoustic features from the segmented 80 ms region.
- Features extracted around VOPs and VEPs are grouped.

- Using the extracted features, sentence-specific GMMs are built.
- During the testing phase, for each test utterance features are extracted around the VOPs and VEPs, and log-likelihood scores are computed from the respective sentence-specific GMM.
- The mean of log-likelihood scores for each utterance is considered as the acoustic correlate of articulation for that utterance.

Detection of VOPs and VEPs

The VOP and VEP are two important events around which acoustic cues related to the production of consonants are preserved. In this work, VOPs and VEPs are detected based on the algorithm proposed in [119]. In the algorithm, zero-frequency filtered signal (ZFFS) and Hilbert envelope of the linear prediction residual (HELPR) are used to locate the VOPs and VEPs in a speech signal [119]. A detailed description of the computation of VOPs and VEPs is given below.

Let us consider a speech signal $s[n]$. To compute the ZFFS, $s[n]$ is first passed through a cascade of two zero-frequency resonators (Equation 4.1), which results an exponential growing or decaying signal $y[n]$.

$$y[n] = -\sum_{k=1}^{4} a_k y[n-k] + s[n], \tag{4.1}$$

where a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1. Then, the trend of $y[n]$ is removed by using a local mean subtraction process (Equation 4.2).

$$\hat{y}[n] = y[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y[n+m]. \tag{4.2}$$

Here, $2N+1$ corresponds to the number of samples in the average pitch period. The resultant signal is termed as the ZFFS ($\hat{y}[n]$). The steps to derive the evidence for locating VOPs and VEPs from ZFFS are given below,

- Compute second-order difference of the ZFFS.
- Convolve the resultant difference signal with a 100 ms long first-order Gaussian differentiator (FOGD) window having a standard deviation of one-sixth of the window.
- Convolved output is called the VOP evidence using ZFFS.
- Further, VEP evidence is obtained by doing the convolution operation from right to left.

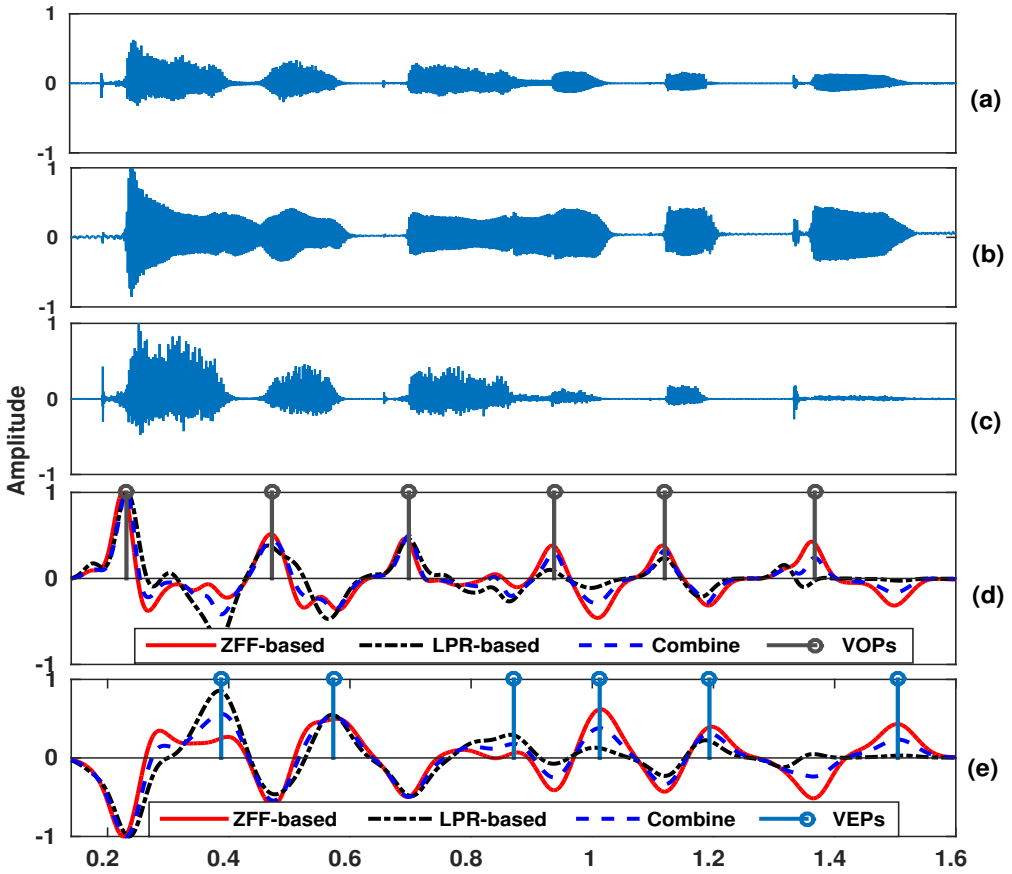

Figure 4.2: Illustration to compute the VOPs and VEPs using ZFFS and LP residual. (a) Speech signal, (b) ZFFS, (c) LP residual, (d) VOP evidence and detected VOPs, and (e) VEP evidence and detected VEPs.

If $e[n]$ be the LP residual of the speech signal $s[n]$, then Hilbert envelope ($h_e[n]$) of $e[n]$ can be computed using Equation 4.3.

$$h_e[n] = \sqrt{e^2[n] + e_h^2[n]}, \tag{4.3}$$

where $e_h^2[n]$ is the Hilbert transform of $e[n]$. The steps to derive the evidence from HELPR to locate VOPs and VEPs are given below,

- Smooth representation of HELPR signal is obtained by taking the maximum value of HELPR for every 5 ms frame with one sample shift.
- Smoothed HELPR signal is convolved with a FOGD window of length 100 ms and a standard deviation of one-sixth of the window.
- VEP evidence is derived by doing the convolution operation from right to left instead of left to

right as in the case of VOP.

Finally, ZFFS-based and HELPR-based evidences are combined and normalized by the maximum value of the sum. The locations of positive peaks in the combined evidence whose amplitudes are larger than a predefined threshold are considered as the detected VOPs or VEPs of the speech signal. An illustration to compute the VOPs and VEPs using ZFFS and HELPR signal is shown in Figure 4.2. Figure 4.2(a), (b), (c), and (d) represent the speech signal, ZFFS, LP residual, VOP evidence and detected VOPs, and VEP evidence and detected VEPs, respectively.

Speech signal of different intelligibility levels (ILs) and corresponding spectrograms along with the detected VOPs and VEPs are shown in Figure 4.3. It can be seen from the spectrograms in Figure 4.3 that acoustic characteristics near the VOPs and VEPs are distorted as the intelligibility degrades. The burst evidence and formant transitions around the VOPs and VEPs are deviated due to the misarticulations, especially for Figure 4.3 (h) and (j), in which case pressure consonants are replaced by glottal stops and nasal consonants, respectively. It is expected that features computed from the speech region around the VOPs and VEPs may provide the information about articulation error. Around the detected VOPs and VEPs, 80 ms transition region (40 ms each side of VOP and VEP) is considered for the feature extraction. Later, all the features extracted around VOPs and VEPs are combined and used for acoustic modeling.

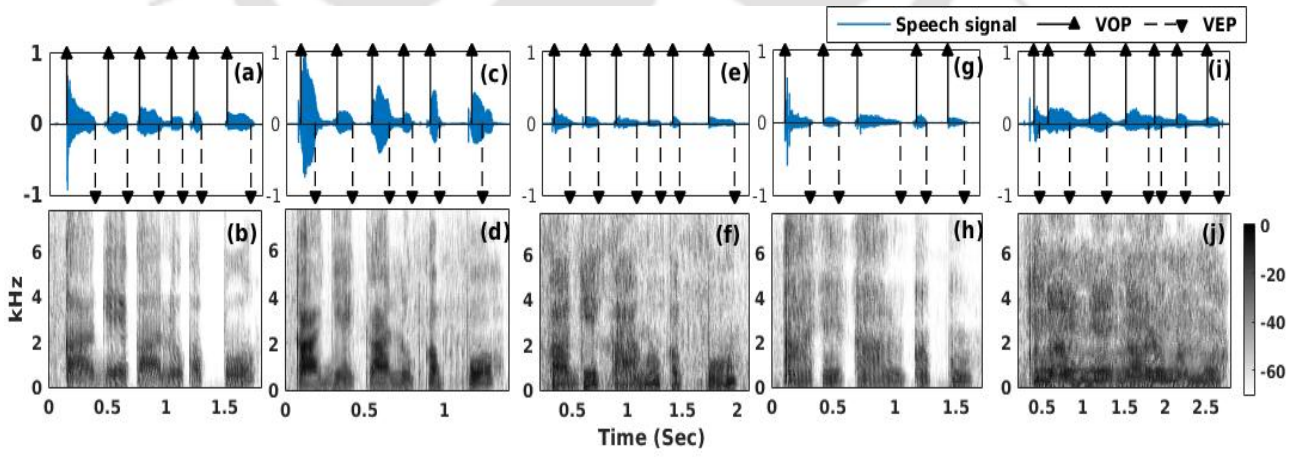

Figure 4.3: Time waveforms with VOPs and VEPs and spectrograms of target sentence O1 (**ka:ge ka:lu kappu**) for normal (a, and b), CLP Intelligibility Level (IL)-0 (c, and d), CLP IL-1 (e, and f), CLP IL-2 (g, and h), and CLP IL-3 (i, and j), respectively. Upper solid arrows and down dashed arrows represent the VOPs and VEPs, respectively.

Feature extraction

A feature which can properly characterize the acoustic deviations of the speech region around VOPs

and VEPs are important in this study. We explore two features, namely, MFCCs and two-dimensional discrete cosine transform (2D-DCT) based joint spectro-temporal (JST) features in the present study.

MFCCs: To compute the MFCCs, initially, speech signals are pre-emphasized by a factor of 0.97. The pre-emphasized speech signals are short-term processed by a 15 ms hamming window with a shift of 5 ms, and Fourier analysis is performed for each windowed speech region. The Fourier magnitude spectrum of the windowed speech signal is passed through a Mel-filter bank of 40 filters, and Mel energies are derived. The perception mechanism of human ear motivates the design of the Mel-frequency filter, and filters are an approximation of the frequency response of inner ear [120]. Mel-frequency filter is used to reduce the spectral resolution of the spectrum, and all frequency components are converted to be placed according to the Mel-scale [121]. Then, the discrete cosine transform (DCT) is applied to the logarithm of Mel energies, and low-order 13 DCT coefficients are termed as the MFCCs. The long-term dynamics of MFCC features, such as delta (Δ) MFCC and delta-delta ($\Delta\Delta$) MFCC are computed by taking first and second derivatives of the base 13-dimensional MFCC features, respectively, and augmented with the base MFCC features [122]. The resultant 39-dimensional MFCC (13 base + 13 Δ + 13 $\Delta\Delta$) feature vector is used further to develop the GMM.

Joint spectro-temporal (JST) features: The JST features are used to model spectro-temporal modulations explicitly from TFR patches. The classical MFCCs can properly model only the spectral dynamics, and delta and acceleration features of the MFCCs are used to model the temporal information. However, it has been shown that augmenting delta and the acceleration features with base MFCCs does not model the joint spectro-temporal modulations explicitly [123]. Spectro-temporal modulations in the transition regions are very important acoustic cues to differentiate among consonants, especially the obstruents [59, 124]. Therefore, it is not effective to model the obstruents using MFCCs, which may not capture the dynamic characteristics of these sounds [58]. In CLP speech, the obstruents are misarticulated most frequently, resulting in reduced speech intelligibility. Therefore, it is very important to model the acoustic cues of these sounds properly when evaluating the intelligibility of CLP speech. Several studies of recognizing stop consonants have explored joint spectro-temporal features and found a substantial improvement over classical MFCCs [58, 125]. This motivates us to explore the joint spectro-temporal features in the intelligibility assessment of CLP speech.

In the present study, we use a 2D-DCT of TFR to extract the joint spectro-temporal features. The effectiveness of joint spectro-temporal features based on 2D-DCT has been demonstrated for (i)

detecting the place of articulation [58,126] and (ii) studying the acoustic characteristics and goodness of production of error-ridden /t/ and /k/ in typical and misarticulated child speech [124]. The 2D-DCT based JST features are computed from the overlapping patches of TFR. To derive the TFR, Fourier magnitude spectrum of short-term processed speech signals are passed through the Mel-filter bank. The Mel-log energies of each frame are stacked in a matrix and the resultant representation is termed as the Mel-log TFR. Overlapping 2D spectro-temporal patches are extracted from the Mel-log TFR and then projected to a 2D cosine basis [58]. Let P_{NM} be a spectro-temporal 2D patch of size $N \times M$, where N and M represent the spectral and temporal extents of the 2D patch, respectively. Then, the 2D-DCT $C(k,l)$ of the patch is given by,

$$\frac{2\omega(k)\omega(l)}{\sqrt{NM}} \sum_{a=0}^{M-1} \sum_{b=0}^{N-1} P_{NM} \cos\frac{\pi l(2a+1)}{2M} \cos\frac{\pi k(2b+1)}{2N}, \tag{4.4}$$

where $k = 0, 1, \ldots, N-1$, $l = 0, 1, \ldots, M-1$, and

$$\omega(k) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0, \\ 1 & \text{if } k \neq 0, \end{cases}$$

$$\omega(l) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } l = 0, \\ 1 & \text{if } l \neq 0. \end{cases}$$

The horizontal DCT coefficients of matrix C correspond to the formant's temporal characteristics, whereas the vertical coefficients correspond to the spectral envelope. Later, we consider low-order 2D-DCT coefficients, which provide a compact representation of the spectro-temporal modulations contained in 2D patch. The low-order DCT coefficients are found to be very sensitive to dominant spectro-temporal modulations of the transition region of two phonemes [123].

In the present work, a spectral extent N is the number of Mel-filter banks used to compute Mel-log-TFR and temporal extent M is the number of frames, which includes the speech region of 50 ms. We use a 2 ms patch rate to capture sudden transitions properly. For each overlapping patch, we apply a 2D-DCT to compute a set of DCT coefficients. Each column of the truncated low-order DCT-coefficient matrix is then flattened to form a single column vector that is used as the feature vector. The corresponding feature is referred to as Mel 2D-DCT (M2DDCT) feature. We considered 13 truncated vertical DCT coefficients and 3 horizontal DCT coefficients, i.e., 39 (13×3) dimensional 2D-DCT feature, as in the case of MFCC features. The dimension of both the features is kept similar

to show the effectiveness of joint spectro-temporal feature than MFCCs.

Development of sentence-specific GMM

The phonetic composition of each sentence-level stimulus used in this work is different. Hence, for each sentence-level stimulus, one GMM is needed to characterize the acoustic space of transition regions around the VOPs and VEPs. Approximately 120 speech utterances are used for the feature extraction from 42 normal speaker's data to build GMM for each sentence stimulus. All the dimensions of extracted features are normalized to have zero mean and unit variance. The expectation maximization algorithm is used to learn the parameters of GMM in 10 iteration steps. Since 10 sentence-level stimuli exist in our database; therefore, 10 multivariate speaker-independent GMMs (λ_j, where $j = 1, 2, \ldots, 10$) are built. Here, each GMM represents the acoustic space of the CV and VC transition regions of each sentence stimulus, built using only the normative data. A GMM of M component Gaussians is derived by,

$$p(\boldsymbol{x}|\lambda_j) = \sum_{i=1}^{M} \omega_{ij} p_{ij}(\boldsymbol{x}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}), \tag{4.5}$$

where ω_{ij}, $\boldsymbol{\mu}_{ij}$, and $\boldsymbol{\Sigma}_{ij}$, $i = 1, 2, \ldots, M$, correspond to the weights, mean vectors, and covariance matrices, respectively, of the different mixtures for GMM model λ_j.

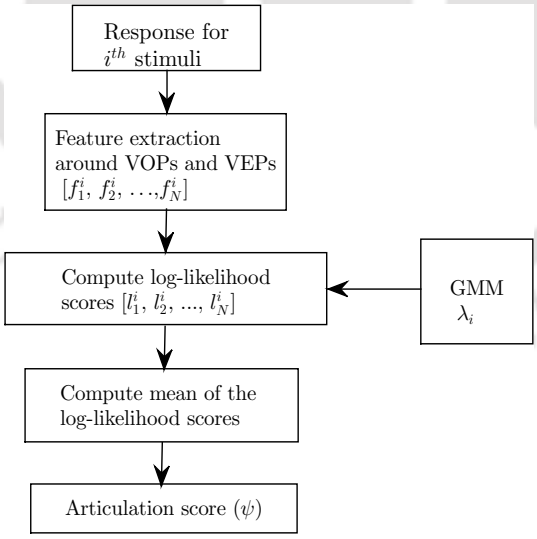

Figure 4.4: Block diagram to compute the articulation score.

Articulation score computation

To compute the articulation score for each test utterance (CLP/normal), initially, VOPs and VEPs are detected. A speech segment of size 80 ms (40 ms of each side of) is considered around each VOP and VEP, and features are extracted from those segments. Then, features extracted for both the regions

are concatenated and compute the log-likelihood scores from the respective GMM. A block diagram to compute the articulation score is shown in Figure 4.4. Let us consider, one test utterance of i^{th}, where $1 \leq i \leq 10$, sentence stimulus, and feature vectors computed around the speech regions of VOPs and VEPs are $[\boldsymbol{f}_1^i, \boldsymbol{f}_2^i, \ldots, \boldsymbol{f}_N^i]$. N is the number of frames in the utterance. The log-likelihood score of the feature vectors are computed using the λ_i GMM, and the log-likelihood scores are $[l_1^i, l_2^i, \ldots, l_N^i]$. The articulation score (ψ) is computed as the average value of log-likelihood scores.

$$\psi = \frac{1}{N} \sum_{r=1}^{N} l_r^i. \tag{4.6}$$

Since the number of component Gaussians needed to build the GMM of a particular sentence is not known in advance, its estimation is essential for better derivation of articulation scores. Since the database is relatively small, leave-one-speaker-out cross-validation is used to estimate the number of Gaussians needed for each sentence accurately.

4.2.2 Objective measure of hypernasality

Working principle

Existing speech processing based approaches concentrate only on the classification of speech into normal or hypernasal, which do not give the degree of hypernasality in terms of continuous values like Nasometer. In this work, we propose an algorithm to derive the degree of hypernasality in CLP speech. This algorithm is mainly motivated by the functionality of the Nasometer, a widely used instrument in the clinical environment for hypernasality assessment [106]. Nasometer uses two microphones to acquire the acoustic signals from the oral and the nasal cavities and provides the percentage of nasalance (N_s). Nasalance score may corroborate a perception of hypernasality; however, it is not a measure of the degree of hypernasality. Mathematically, the nasalance score (N_s) can be defined using Equation 4.7.

$$N_s = \frac{E_N}{E_N + E_O} \times 100. \tag{4.7}$$

Here, E_N and E_O are the acoustic energies captured by nasal and oral microphones, respectively. The nasal and oral sounds give the maximum, and minimum nasalance score, respectively, and both the categories represent two opposite cases of nasality. Motivated from the functionality of Nasometer, in the algorithm, speech signals representing two extremely opposite cases of nasality are used to develop the acoustic models. The extremely opposite cases of nasality considered are the oral sentences of

normal speakers (Table 3.2) (rich in vowels, approximants, stops, and fricatives) and nasal sentences of moderate-severe hypernasal speakers (Table 3.3) (rich in nasals and nasalized vowels), which are referred as the oral and nasal classes, respectively. It is obvious that the oral sentence produced by normal speakers should have very low nasality. However, nasal sentences produced by moderate-severe hypernasal speakers should have high nasality. We have also analyzed the nasalance scores acquired by Nasometer for normal and different categories of hypernasality. The level of hypernasality was judged by perceptual evaluation, not by nasalance scores. Then, we checked whether the mean nasalance score is also higher for nasal sentences produced by moderate-severe hypernasal speakers than that of oral sentences produced by normal speakers. Bar plots in Figure 4.5 show the nasalance values of nasal and oral sentences for the normal and different category of hypernasal speakers. Where x-axis corresponds to the perceptual hypernasality ratings. From the bar graph, the following observations can be drawn:

(i) The nasalance scores are high for nasal sentences when compared to oral sentences.

(ii) The discrimination of nasalance scores between the group of nasal and oral sentences is greater for normal speakers and reduces from mild to moderate and moderate to severe hypernasal speakers. This is because, in moderate-to-severe hypernasal speakers, most of the oral consonants (/b/, /p/) are replaced by nasals (/m/, n/) and vowels get severely nasalized [105].

(iii) The oral sentences of the normal group exhibit minimum nasality, whereas the nasal sentences of the severe group show maximum nasality.

Thus, the mean nasalance value of moderate-severe CLP speakers is the highest, while the mean nasalance score of the oral sentence is the lowest. These two extremely opposite nasality cases represent the group with minimum and maximum attainable degree of nasality.

Motivated by the behavior of nasalance values for target oral and nasal sentences as a function of the degree of hypernasality, the current work proposes an approach for the estimation of hypernasality scores. In this work, a two-class classifier is developed for the oral class: using sentences with minimum nasality (oral sentences of normal speakers) and nasal class: using sentences with maximum attainable nasality (nasal sentences of the moderate-severe hypernasal group). During the testing, features derived from the response of the speaker for the target oral sentence is given for the classifier. The posterior probabilities derived for the nasal class are considered as hypernasality scores.

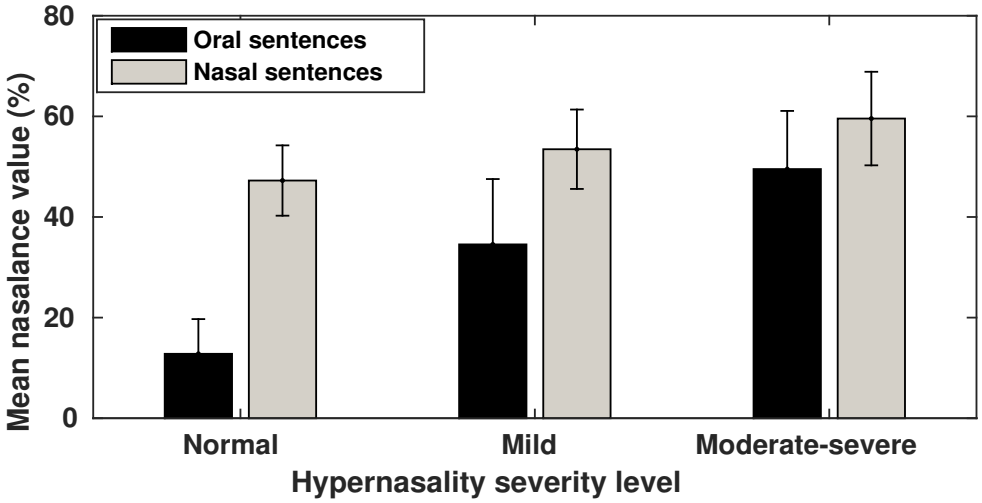

Figure 4.5: Bar plots of the nasalance scores for oral and nasal sentences of normal, mild and moderate-severe hypernasal speakers.

Hypernasality score computation

Figure 4.6 shows the block diagram representation of hypernasality score estimation module. Initially, the glottal activity (GA) region of the speech signal is detected. GA refers to the excitation of the vocal tract system by the vibration of vocal folds during the speech production [127]. The glottal activity detection (GAD) is based on the characterization of the excitation source signal [127]. Here, ZFFS-based approach is used to identify the glottal activity. The positive to negative zero-crossing of ZFFS corresponds to the glottal closure instant (GCI), and the slope of ZFFS around the GCIs is termed as the strength of excitation. Later, speech regions which correspond to the strength of excitation value greater than a predefined threshold are considered as GA regions [127]. In this work, a threshold of 0.4 times the average value of the strength of excitation of the entire signal is considered. The detected GA region of the speech signal is used for the feature extraction. Present work uses MFCC features for hypernasality estimation, and a detailed description to compute the MFCCs is given in Section 4.2.1.

The acoustic features derived from the glottal activity regions are used to build GMM-based acoustic models for oral and nasal classes. Let, $\lambda_O = \{\omega_i^O, \boldsymbol{\mu}_i^O, \boldsymbol{\Sigma}_i^O\}_{i=1}^{M_O}$, represent GMM model for the class of oral sentences with M_O number of mixtures. Similarly, $\lambda_N = \{\omega_i^N, \boldsymbol{\mu}_i^N, \boldsymbol{\Sigma}_i^N\}_{i=1}^{M_N}$, where M_N is number of component Gaussian, represent GMM for the class of nasal sentences (λ_N) belonging to the group of moderate-severe hypernasality. The parameters ω_i^O, $\boldsymbol{\mu}_i^O$, and $\boldsymbol{\Sigma}_i^O$ represent weight,

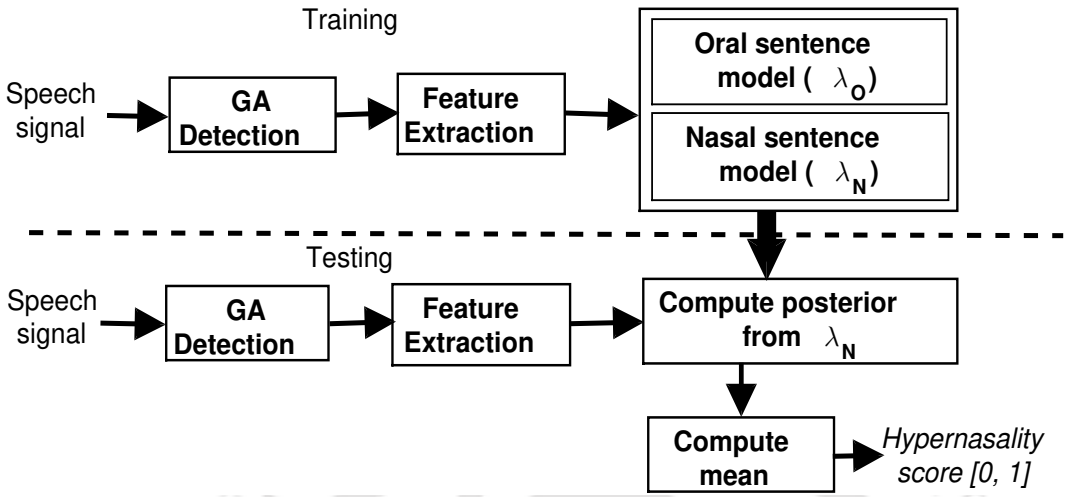

Figure 4.6: Block diagram of hypernasality score estimation module.

mean, and covariance matrix of the i^{th} Gaussian for the λ_O GMM, respectively. While, the parameters ω_i^N, $\boldsymbol{\mu}_i^N$, and $\boldsymbol{\Sigma}_i^N$ represent weight, mean, and covariance matrix of the i^{th} Gaussian for λ_N GMM, respectively.

During the testing, the target oral sentences are used, and posterior probabilities for nasal acoustic model λ_N are computed for each feature vector. Let us consider an oral sentence S_o, which has K feature vectors and can be represented symbolically as $S_O = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots \boldsymbol{f}_K]$. The posterior probability of model λ_N for the feature vector $\boldsymbol{f}_i$, $i = 1, 2, \ldots, K$ is computed by Equation 4.8.

$$p(\lambda_N|f_i) = \frac{p(\boldsymbol{f}_i|\lambda_N)p(\lambda_N)}{p(\boldsymbol{f}_i|\lambda_N)p(\lambda_N) + p(\boldsymbol{f}_i|\lambda_O)p(\lambda_O)}, \tag{4.8}$$

where class prior probabilities $p(\lambda_N)$ and $p(\lambda_O)$ are considered as equal, i.e. 0.5. $p(\boldsymbol{f}_i|\lambda_N)$ and $p(\boldsymbol{f}_i|\lambda_O)$ are the likelihood values estimated from GMMs λ_N and λ_O, respectively.

The likelihood of nasal class $p(\boldsymbol{f}_i|\lambda_N)$ and oral class $p(\boldsymbol{f}_i|\lambda_O)$ in Equation 4.8 are proportional to the nasal and oral sound characteristics present in the speech signal, respectively. The terms E_N and E_O in Equation 4.7 are proportional to the nasal and oral sound energies, respectively. Since the nasalance value in Equation 4.7 is proportional to the amount of nasal energy present in the speech signal. Similarly, the posterior probability value $p(\lambda_N|\boldsymbol{f}_i)$ in Equation 4.8 is proportional to the presence of nasal sound characteristics in the speech signal. Therefore, due to the existence of the similarity between the computation of nasalance score and the posterior probability of nasal class (Equation 4.7 and Equation 4.8), the posterior probability scores are expected to give a measure of

nasality. Thus, the hypernasality score (η) of the oral sentence S_o is computed as,

$$\eta = \frac{1}{K}\sum_{i=1}^{K} p(\lambda_N|\boldsymbol{f}_i), \tag{4.9}$$

The models λ_O and λ_N are trained using EM algorithm with diagonal covariance matrix. Different

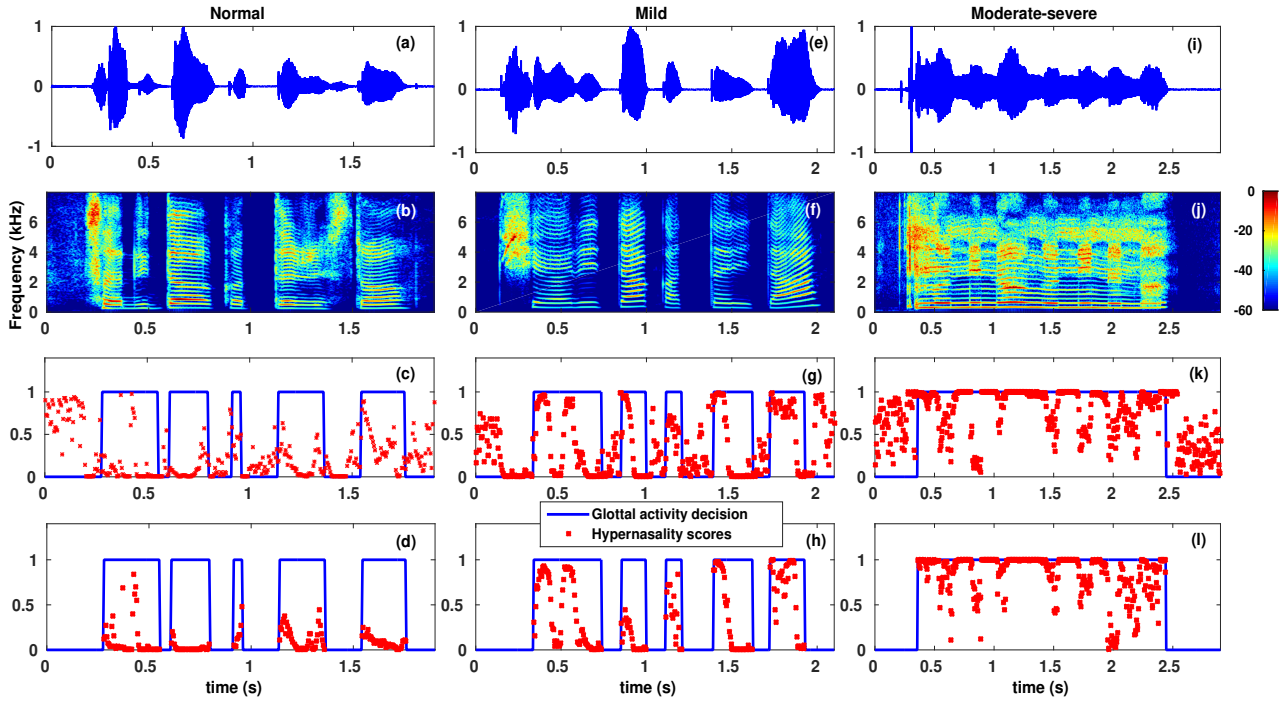

Figure 4.7: Illustration of the significance of GAD in the computation of hypernasality scores for the sentence " sariṉa kaṯṯari ṯa:". (a)-(d), (e)-(h), and (i)-(l) represent the speech signal, spectrogram, contour of posterior probabilities scores for hypernasal class without glottal activity and with glottal activity detection, respectively for normal, mild and moderate-severe hypernasal speech. In (d), (h), and (l) dotted lines indicate the posterior probability values and solid lines indicate the detected glottal activity regions. Application of GAD reduces the spurious scores resulting from unvoiced regions.

Gaussian mixtures, such as 32, 64, and 128 are experimentally analyzed, and found GMM with 64 component Gaussians is optimum for the work. Posterior probabilities $p(\lambda_N|\boldsymbol{f}_i)$ computed for the target oral sentence "*sariṉa kaṯṯari ṯa:*" for different levels of hypernasality are illustrated in Figure 4.7. Figure 4.7(a)-(d) show the waveform of normal speaker for the given target sentence, spectrogram, contours of $p(\lambda_N|\boldsymbol{f}_i)$ without the inclusion of GAD pre-processing stage, and with the inclusion of GAD, respectively. Here, without GAD pre-processing stage refers to training and testing of models are carried out without the inclusion of GAD preprocessing stage. The contour of $p(\lambda_N|\boldsymbol{f}_i)$ for normal speech (Figure 4.7(c)) shows that the exclusion of GAD pre-processing stage results in the

spurious values at unvoiced regions like fricatives, stop gaps and silence regions (example: around 0.8 seconds corresponding to stop gap of /k/). This will result in an increased hypernasality score for the normal speaker and increases the false alarm rate. Similarly, Figure 4.7(e)-(h) and (i)-(l) show the speech waveform, spectrogram, contours of $p(\lambda_N|\boldsymbol{f}_i)$ without and with inclusion of GAD for mild and moderate-severe hypernasal groups, respectively. The $p(\lambda_N|\boldsymbol{f}_i)$ values are increased for the mild group than the normal, severe group than that of mild. This shows that the $p(\lambda_N|\boldsymbol{f}_i)$ values vary in proportional to the severity of perceived hypernasality. Glottal activity region based processing significantly reduces the spurious hypernasality scores resulted from the unvoiced/silence regions and hence, reduces false alarm rate.

4.2.3 Computation of intelligibility score

After estimating the articulation and hypernasality scores for each utterance, they are used to predict the utterance-level intelligibility score. The regression model used in this scenario predicts the intelligibility scores of a given speech utterance based on the information obtained from the previously evaluated speech utterances by SLPs. In this work, we explore two different regression models, namely, linear regression and support vector regression (ϵ-SVR) to find the mapping function between the predictors, i.e., articulation and hypernasality scores, and estimated intelligibility scores. The linear function is used as the kernel function in SVR. SVR estimates intelligibility score that deviates at most ϵ from the perceptual intelligibility values for all training data, subject to being as flat as possible at the same time. In SVR, features are transformed into a higher dimensional space where linear separability can be achieved, and we are expecting that the SVR-based model will give better result than linear regression.

4.3 Results and discussion

In this section, performance of the proposed intelligibility measure is discussed. Initially, the performance of the objective measures of articulation and hypernasality is evaluated. Later, the results of the proposed composite measure of intelligibility based on estimated articulation and hypernasality scores are discussed. Spearman's rank correlation coefficient (ρ) between the estimated scores and the perceptual scores is considered as the parameter for performance evaluation. Present work demonstrates the intelligibility prediction at the utterance-level for the proposed objective measures. The evaluation of all the measures is performed using the leave-one-speaker-out cross-validation (LOSO-

Table 4.1: Mean (μ) and standard deviation (σ) of the absolute correlation coefficients computed between proposed articulation scores (ψ) and articulation ratings given by the SLPs for all the explored features.

Measures	$\mu \pm \sigma$	p-value
ψ_{MFCC}	0.68 ± 0.0411	<0.001
ψ_{M2DDCT}	0.70 ± 0.0375	<0.001
$\psi_{\text{MFCC}} + \psi_{\text{M2DDCT}}$	0.72 ± 0.0384	<0.001

CV) criteria.

4.3.1 Performance evaluation of the proposed articulation measure

To compute the articulation score for a test utterance its frame wise log-likelihood scores are calculated from the respective sentence-specific GMM. Then, the mean log-likelihood values of the utterance and its corresponding PCC score are considered as independent and dependent variables to build a regression model, respectively. To evaluate the performance, in each fold of LOSO-CV one person's speech data is used as test data, and remaining speakers data is used to build the model. Since we have 42 CLP speakers in our database, this process is repeated for all the 42 CLP speakers, and articulation score is estimated for each test utterance at every fold. The prediction accuracy is evaluated in terms of correlation coefficient between the estimated articulation scores for the test samples of all 42 folds and subjectively rated articulation scores. We found that both the variables, i.e., perceptual ratings and estimated articulation scores are non-normally distributed. And hence, Spearman's rank correlation coefficient will be appropriate [128]. The number of Gaussians for which the best correlation is achieved serves as the optimum for the respective sentence stimulus. The overall performance of the algorithm is evaluated by taking the mean and standard deviation of the absolute correlation values of ten sentences, and it is shown in Table 4.1. It can be observed from the table that a significant correlation is achieved in predicting the articulation scores objectively. M2DDCT feature gives a higher correlation value of $\rho = 0.70$ than MFCCs ($\rho = 0.68$), which justifies the usage of this feature in predicting sentence-level articulation scores. Since the features are extracted from the transition region between two sounds, the M2DDCT feature perhaps properly characterize the spectro-temporal dynamics embedded in that region. Apart from the capability of individual features to describe the articulation, the combination of both features is also explored. For all the sentence stimuli, the performance is improved when both the features are combined.

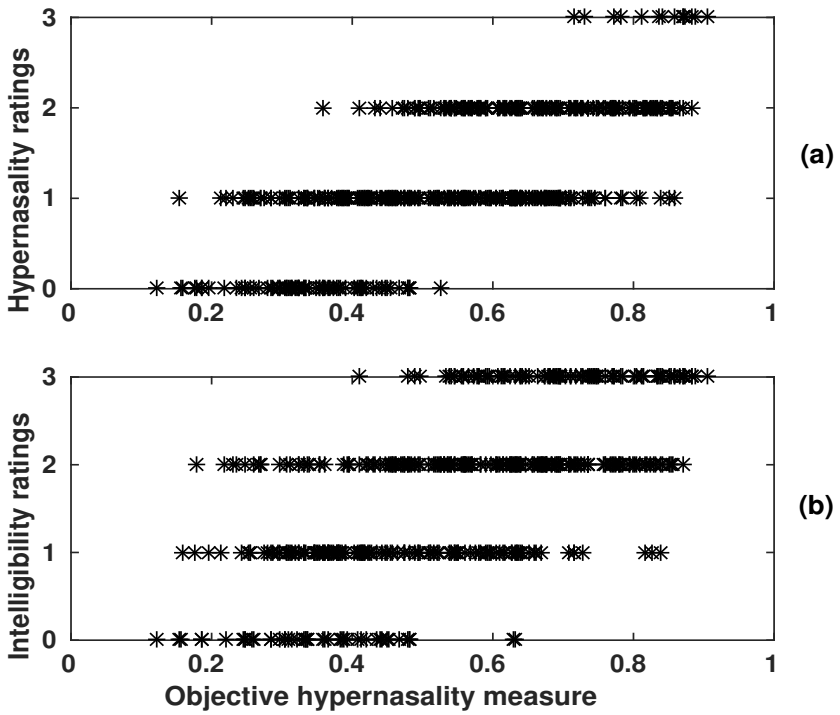

Figure 4.8: Scatter plots of objective hypernasality measure (η) vs. perceptual ratings of (a) hypernasality and (b) intelligibility for all the utterances.

4.3.2 Performance evaluation of the hypernasality measure

The mean posterior score (η) computed for each oral sentence is correlated with the respective perceptual rating of hypernasality. To evaluate the performance, LOSO-CV is performed as discussed in the case of articulation score evaluation. Spearman's rank correlation coefficient between the utterance-level hypernasality scores and perceptual ratings is computed. And a correlation value of $\rho = 0.68$ is obtained. We have shown the scatter plot of utterance-level posterior scores with respect to perceptual hypernasality ratings in Figure 4.8(a). It can be seen that as the hypernasality increases, the posterior probability score also increases. Since the posterior value is estimated from λ_N, the higher value of posterior score represents higher nasality present in the utterance. It can be seen from Figure 4.8(a) that there exists significant discrimination between the hypernasality level (HL)-0 and HL-1, and HL-2 and HL-3. However, the discrimination between HL-1 and HL-2 is insignificant. We have also shown the scatter plot of the utterance-level posterior scores with respect to perceptual intelligibility ratings in Figure 4.8(b). As the loss of intelligibility increases, the hypernasality score also increases. It can be seen from the scatter plots, the number of utterances falling in the category of

Table 4.2: Mean (μ) and standard deviation (σ) of the absolute correlation coefficients computed between the estimated articulation scores and perceived intelligibility ratings given by the SLPs for all the features.

Measures	$\mu \pm \sigma$	p-value
ψ_{MFCC}	0.67 ± 0.0459	<0.001
ψ_{M2DDCT}	0.69 ± 0.0442	<0.001
$\psi_{\text{MFCC}} + \psi_{\text{M2DDCT}}$	0.72 ± 0.0478	<0.001

HL-3 is significantly less as compared to the number of utterances in the category of IL-3. A possible reason for this observation is that the speaker with mild hypernasality level may have associated with articulation errors which reduce their speech intelligibility. We have found that in IL-3 most of the speakers possess maladaptive articulation, and some of the speakers have the nasal substitution of obstruents due to severe hypernasality.

4.3.3 Performance evaluation of the composite measure of intelligibility

First, we have analyzed how the objective measures of articulation (ψ) and hypernasality (η) individually correlated with the perceived intelligibility ratings. The correlation coefficient between the proposed measure of articulation and perceived intelligibility for each sentence stimulus are computed. Then, the average value of utterance-level correlation is computed and listed in Table 4.2. A correlation value of $\rho = 0.67$ is obtained for MFCC based measure (ψ_{MFCC}), whereas 0.02 increment of correlation is obtained in the case of M2DDCT feature based measure (ψ_{M2DDCT}). The combination of both features further increases the average correlation value up to $\rho = 0.72$. The MFCC based hypernasality measure (η_{MFCC}) is also tested as an acoustic correlate of the perceived intelligibility, and a correlation value of 0.62 is obtained between them. For one sentence, we have shown the scatter plots of articulation (ψ_{M2DDCT}) and hypernasality (η_{MFCC}) based measures *versus* perceptual ratings of intelligibility in Figure 4.9(a) and (b), respectively. It can be seen from the figures that articulation based scores are more correlated with the intelligibility than hypernasality. Articulation based scores properly distinguish the speech with IL-0 and IL-1 or IL-1 and IL-2; however, it is not observed in the case of hypernasality based measure. From the above analysis, we observed that the correlation values are higher for articulation measure than that of hypernasality measure. These results also support the study done on the subjective evaluation (Chapter 3), where it was found that perceived intelligibility ratings are less correlated with the hypernasality ratings than the articulation error.

Later, the objective scores of articulation and hypernasality are combined to predict the composite

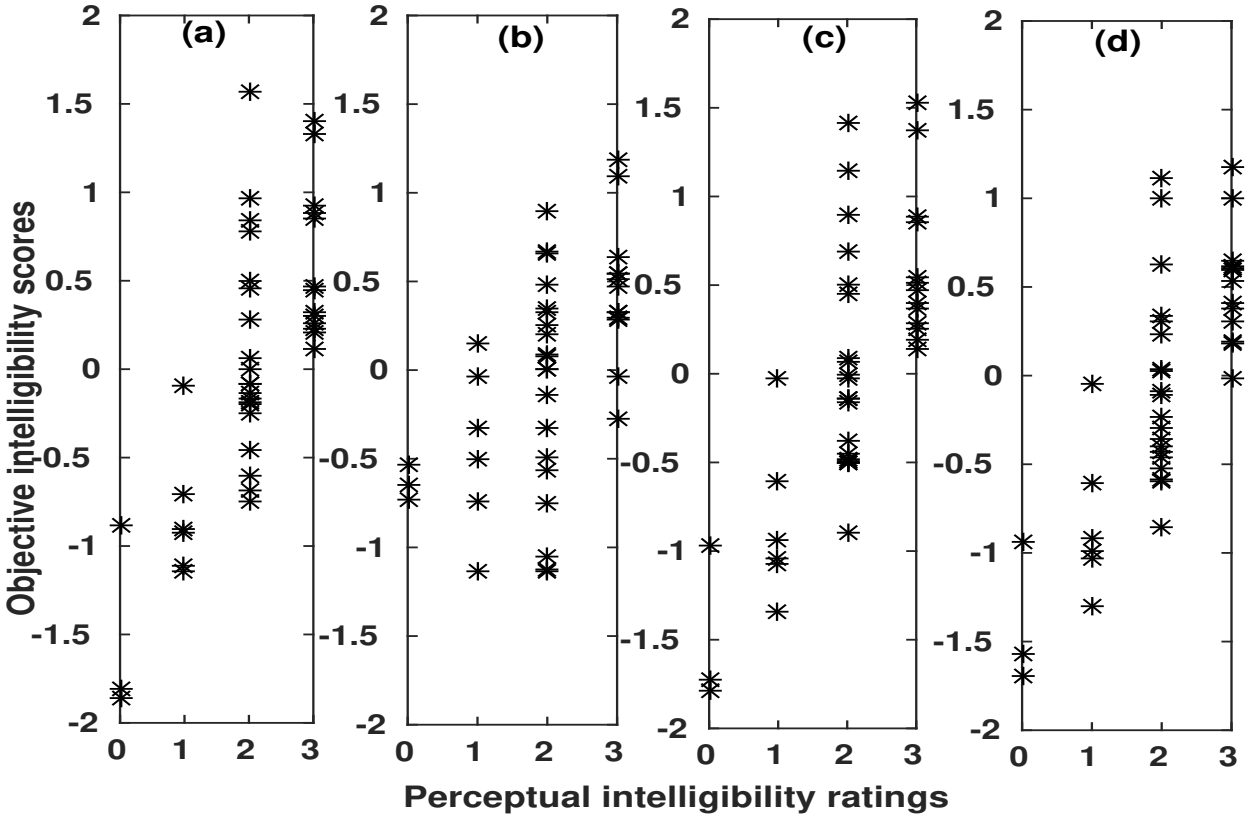

Figure 4.9: Scatter plots of predicted intelligibility scores vs. perceptual ratings of intelligibility for one sentence stimulus. (a) based on articulation scores, (b) based on hypernasality scores, (c) linear regression based composite scores, and (d) SVR-based composite scores.

intelligibility score. To estimate the intelligibility scores, linear regression and ϵ-SVR models are used. For both cases, LOSO-CV is used for performance evaluation. The dependent (perceptual intelligibility ratings) and independent (objective measures of articulation and hypernasality) variables of the model are normalized to have zero mean and unit variance before training the models. Initially, we have shown the scatter plots of composite measure ($\eta_{\text{MFCC}} + \psi_{\text{M2DDCT}}$) based on linear regression and SVR *versus* perceptual ratings of intelligibility for one sentence stimulus in Figure 4.9(c) and (d), respectively. From the scatter plots, it can be seen that the predicted scores are linearly increasing with respect to the perceived intelligibility ratings. The mean of sentence-level correlation values is studied for the overall performance evaluation for all the combinations of acoustic measures. Table 4.3 lists the results of correlation analysis, and it can be observed that the correlation is improved as compared to individual measures of articulation and hypernasality. For both linear regression and SVR, a combination of ψ_{M2DDCT} and η_{MFCC} based composite measure outperforms the ψ_{MFCC} and

Table 4.3: Mean (μ) and standard deviation (σ) of the absolute correlation coefficients computed between composite measures and perceived intelligibility ratings.

Composite measure	Linear Regression		ε–SVR	
	$\mu \pm \sigma$	p-value	$\mu \pm \sigma$	p-value
$\eta_{\text{MFCC}} + \psi_{\text{MFCC}}$	0.70 ± 0.0523	< 0.001	0.72 ± 0.0434	< 0.001
$\eta_{\text{MFCC}} + \psi_{\text{M2DDCT}}$	0.72 ± 0.0519	< 0.001	0.75 ± 0.0415	< 0.001
$\eta_{\text{MFCC}} + \psi_{\text{MFCCs}} + \psi_{\text{M2DDCT}}$	0.74 ± 0.0509	< 0.001	0.77 ± 0.0432	< 0.001

η_{MFCC} based measure. While combining the ψ_{M2DDCT} and ψ_{MFCC} with η_{MFCC}, correlation is further improved. The SVR gives consistently better performance as compared to the linear regression method. The composite metric based on ψ_{M2DDCT}, ψ_{MFCC} and η_{MFCC} gives correlation values of $\rho = 0.74$ and $\rho = 0.77$ for linear regression and SVR, respectively. However, a careful investigation of individual sentence-specific correlation values, we found that the correlation values are not consistent for all the sentences. This inconsistency is reflected in the high standard deviation of 10 correlation values. One reason for this inconsistency may be due to the inaccurate detection of VOPs and VEPs in some of the utterances for articulation score estimation. The accurate detection of VOPs and VEPs is important to properly capture the acoustic features from CV and VC transition regions.

Thus, it is indeed possible to predict CLP speech intelligibility by combining the objective measures of articulation error and hypernasality. The potentiality of VOP and VEP based speech analysis in deriving a measure of consonants production errors in CLP speech is also shown. This measure of articulation is tested as the biomarker of intelligibility and found a significant correlation with the perceived intelligibility ratings. The method is primarily based on the accurate detection of VOPs and VEPs, which automates the process of assessment. The temporal accuracy of VOP and VEP detection is important to characterize the transition region properly. However, the method explored in this work uses a smoothing operation on the evidence to detect the VOPs and VEPs. This smoothing operation may reduce the temporal accuracy of the VOP and VEP detection. Therefore, a detailed analysis is needed to observe how the deviation of VOPs and VEPs from the original location affect the performance of intelligibility assessment, and future work is planned in this direction. The features are extracted from the speech regions around which abrupt spectral changes occur; therefore, acoustic features which can capture these sharp spectral discontinuity are required. The 2D-DCT based JST features can provide a better representation of the spectro-temporal dynamics present in the transition region between two sounds as compared to MFCCs, and results observed in the current work also justify

this hypothesis. Since for each sentence-level stimulus, separate GMM is needed, this may lead to the complexity of the proposed algorithm. Apart from this, the hypernasality estimation algorithm is dependent on the speech data from moderate-severe hypernasality, which may not be feasible in all the times. This demands further exploration of the possibility to use normal or mild hypernasality speaker's nasal sentences to build an acoustic model for the hypernasality estimation.

The proposed composite measure may assist the speech-language pathologists during the intelligibility assessment by providing the information whether articulation or hypernasality more impacting the intelligibility. It may help to analyze the CLP speech intelligibility by decomposing into two different aspects, such as articulation and hypernasality. During the therapy or other interventions, this information is essential to provide proper feedback to the CLP speaker. The measure may give insight into how articulation error is affecting the intelligibility. We have provided sentence-level intelligibility scores for each sentence stimulus, and each stimulus represents one or two particular pressure consonants. Hence, it may be easier for SLPs to observe which pressure consonants are more responsible for intelligibility deficits. This may help SLPs to give more concentration to those particular pressure consonants to improved intelligibility during therapy. In some of the children, hypernasality may present with no articulation error where the proposed measure may help to analyze what extent nasality is affecting the intelligibility.

4.4 Summary and conclusions

In this chapter, an effort is made to derive a measure of CLP speech intelligibility at utterance-level by combining the information of articulation deficits and hypernasality. An objective measure to quantify articulation error is developed by utilizing the acoustic features extracted from the speech regions around VOPs and VEPs. MFCC and 2D-DCT based JST features are explored to show the effectiveness of the proposed measure. Results show a good correlation between the proposed measure of articulation and the PCC scores given by the SLPs. A novel method to gauge the degree of hypernasality is proposed. Then, the objective measures for articulation and hypernasality are combined using the linear regression and ϵ-SVR models to predict the intelligibility scores. The predicted scores are tested as the correlate of subjective intelligibility ratings, and Spearman's correlation coefficient is studied as the performance measure. The proposed composite measure provides a significant correlation with the perceptual ratings. For all the sentence-level stimuli 2D-DCT based JST feature

shows the improvement over the MFCC feature. SVR-based intelligibility predictor gives superior performance over the linear regression.

From this chapter, it is found that objective articulation scores are more correlated with the perceptual intelligibility ratings than hypernasality scores. Distortion of consonant production is characterized using the acoustics features computed from the CV and VC transition regions around the VOP and VEP events, respectively. In CLP speech, obstruents are primarily distorted due to VPD and mislearning; hence, events which are associated with the obstruent to sonorant sounds or vice-versa need to be explored. In the next chapter, articulation deficits associated with obstruents production is only characterized to derive a measure of intelligibility, and *g-landmraks* are considered to segment the required speech regions.

5

Exploring Glottis Landmarks for Intelligibility Assessment

Publications

- **Sishir Kalita**, S. R. M. Prasanna, S. Dandapat, "Importance of glottis landmark for the assessment of cleft lip and palate speech intelligibility", *The Journal of the Acoustical Society of America*, 144(5), October (2018), PMID: 30404473.

Contents

Objective

The primary objective of this chapter is to analyze if the acoustic information extracted from the speech region around g(lottis) landmarks can be used to derive the correlates of CLP speech intelligibility. Initially, we have provided a glimpse of earlier studies on utilizing the deviations of landmark expression to analyze the articulatory deviations and intelligibility of pathological speech. Then, an investigation is made to analyze how spectral characteristics are distorted around the consonant landmarks due to articulation errors. We explore the acoustic characteristics of articulatory deviations near g(lottis)–landmarks to derive a measure of speech intelligibility. To derive the measure the acoustic feature is computed at the vicinity of two g–landmarks. Two separate sentence-specific Gaussian Mixture Models (GMMs) are built for each sentence stimulus, using the features extracted around +g–landmark and -g–landmark, respectively. Speech utterances from the normal speaker's group are used to train GMMs. The GMMs developed for each sentence stimulus are used to compute the log-likelihood scores of the respective test utterance. The average value of log-likelihood scores for the features is calculated. Since two separate GMMs are built for +g–landmark and -g–landmark for each sentence stimulus, two average values of log-likelihood scores are obtained for each test utterance. Both average scores are studied as the acoustic correlates of CLP speech intelligibility.

5.1 Introduction

Acoustic landmarks are defined as the time locations of abrupt acoustic events in a speech signal and are correlated with major articulatory movements [80–83]. According to the landmark theory, humans perceive phonemes in response to the acoustic cues that are anchored temporally around the landmarks [129]. Liu defined landmarks for the consonants, vowels, and glides [82]. Among all the landmarks, consonant landmarks correspond to abrupt discontinuities in the spectrum, and a significant amount of phonetic information is concentrated near these landmarks [60]. Since intelligibility of CLP speech is primarily degraded due to the problems in consonant production, it is expected that analyzing the speech regions around consonant landmarks may be more helpful. Three types of landmarks are defined for consonants: *g(lottis)*, *b(urst)*, and *s(onorant)* [82]. Liu mentioned in [82] that *g–landmarks* pinpoint the start or stop of the vocal folds free vibration. The vocal folds vibration may completely stop or get suppressed due to the increase intraoral pressure [60, 82]. The examples of suppressed vocal-folds vibration are the voice bar after the closure of a voiced stop consonant and

the voicing in the production of a voiced fricative consonant [60]. The *g–landmarks* are denoted as +g and -g, which represent the starting and ending locations of vocal folds' vibration, respectively [82]. The *g–landmarks* distinguish obstruent consonants from the vowels or sonorant consonants, and the vocalic transition from obstruent to these sounds and vice-versa are associated with *+g–landmark* and *-g–landmark*, respectively [60]. Such abrupt vocalic transition regions contain important perceptual cues to identify the obstruents [60]. The *b–landmarks* are associated with the existence of frication noise and burst in the production of obstruents. The adequate build-up of intra-oral pressure is very important to fire a *b–landmark*. A *b–landmark* marks the boundary between a silent region and a frication noise or burst. While an *s–landmark* generally represents the opening or closing event of the velopharyngeal valve in the production of nasal sounds. However, it also marks the boundary between a vowel or glide and a sonorant consonant. Depending on the energy increase or decrease at the vicinity of a consonant landmark, they are classified as '+' and '-', respectively.

In the production of consonant landmarks, abrupt spectral changes occur which are associated with the transition region from one manner of articulation to another. From the perceptual experiment, it has been found that humans more concentrate on these abrupt discontinuities [60]. Any distortion in the spectral abruptness may change the perceptual judgment of consonant identification. In the speech production system, one of the most important aspects responsible for abrupt transition is the proper oro-nasal coupling in the production of oral sounds. For example, the proper closure of the velopharyngeal port is essential for the production of obstruent sounds, where the adequate intra-oral pressure build-up is necessary for the sudden release of frication noise. However, in the presence of CLP condition, adequate build-up of intra-oral pressure may not be possible due to VPD or oro-nasal fistula. Though, there is no direct relation of movement of the velopharyngeal valve in determining the burst and glottis landmarks. However, dysfunction in the velum movement may lead to the changes of acoustic characteristics near the glottis landmarks and sometimes delete the landmarks. The F1 and F2 formants dynamics in the sonorant transition region near the *g–landmarks* provides important acoustic cues to perceive the obstruents. Due to the reduction of intra-oral pressure in the production or obstruents, the phonetic distinctiveness is lost for these sounds [109]. Also, in some of the cases, CLP speakers try to omit the obstruent consonants from their speech. This chapter explores the potentiality of the consonant landmark in deriving the CLP speech intelligibility. However, this work only explores the *g–landmarks* to show that they can be utilized to derive measures for CLP speech intelligibility and

to provide clinically relevant information. The *b–landmarks*, which signify an affricate or aspirated stop burst and the offset of frication or aspiration noise due to a stop closure, may not be fired all the time in CLP speech due to the inadequate build up of intra-oral pressure. In the case of weak obstruent condition, the energy changes associated with the occurrence of *b–landmarks* may not be sufficient to occur an energy discontinuity. Since the *b–landmarks* are always associated with the *g–landmarks*, information derived around the *g–landmarks* may be able to characterize the deviations around *b–landmarks*. Moreover, in the case of nasal substitution for obstruent, burst landmark may not be present. The detection of the *s–landmarks* is more complicated, and detection rate is abysmal as compared to the *g–landmarks* [82]. Moreover, the production of nasals and approximants not affected in the CLP speech. A brief review of the landmark processing based approaches in analyzing the pathological speech is provided in the next subsection.

5.1.1 Exploration of landmark-based pathological speech analysis

Recently, landmark-based speech analysis is gaining research interest to evaluate the pathological speech [80, 81, 130]. Researchers have shown the potentiality of landmark-based speech analysis to derive the biomarker for speech intelligibility by characterizing the expression of landmarks [80, 81]. The importance of landmarks in estimating the speech rate for the assessment of dysarthric speech is demonstrated in [131], and it is found that the proposed algorithm is sufficiently robust in the detection of slow as well as variable speech rates of disorder speech. In [132], acoustic landmarks are considered as the anchor points in the speech signal, and several acoustic features extracted from the speech region around landmarks are explored to detect the presence of Parkinson's disease. A pronunciation evaluation method for substitution error problem in L2 learners of English is proposed based on the combination of landmarks and support vector machine (SVM) classifier in [133]. Authors showed that in contrast to the confidence score based method, landmark-based SVM gave better performance in estimating the pronunciation scores, and SVM-based approach can provide the acoustic characteristics of the mispronounced phone. Similar work is also proposed in the case of pronunciation error detection for L2 learners of Chinese language [134]. A software package called *SpeechMark* is developed in [135, 136] for the automatic analysis of several acoustic parameters, such as speech articulation and syllable cluster rate of normal and pathological speech using the landmark-based analysis. Authors showed the potential of the software in analyzing the syllabic cluster rate for Parkinson's disease patients, who are receiving deep brain stimulation in rested vs. sleep-deprived conditions. A study based on

the *SpeechMark* tool showed that landmark-based syllabic cluster analysis gives a rough measure of Parkinson patient's ability to repeat speech utterance with a certain level of articulatory precision at a particular speech rate [137]. A system based on the landmark is proposed to study the infant's vocalization for syllable complexity to detect the early speech-related disorders [138,139]. Landmarks are also explored to count the number of total and canonical syllables in infant vocalization [140]. Authors in [81] exploit the landmarks to characterize dysarthric speech. They observed that dysarthric speakers have difficulty in producing certain landmarks and insert unnecessary landmarks. Authors also showed that the detection, substitution, and insertion of landmarks significantly correlated with the perceived intelligibility ratings. A study presented in [80] analyzed the landmarks expression in normal speech and found an increased number of landmarks signifies the greater intelligibility, and it is gender dependent. However, the authors suggested that the analysis of a large-scale database should be done to confirm the results. In [141], non-word repetition tasks are used to assess the modifications produced by children with different speech disorders, where they study the modification of acoustic landmark pattern as a cue to provide a more accurate and insightful diagnosis. Despite of encouraging findings, limited works have been reported in the literature in this direction. Besides, as per the knowledge of existing literature, no attempts have been made to analyze the CLP speech using the landmarks.

5.1.2 Motivation for the work

From the earlier chapters, it is evident that CLP speech intelligibility is primarily degraded due to the presence of articulation errors, and these errors occur for the obstruent consonants due to inadequate intra-oral pressure. Hence, the characterization of articulatory deviations only for obstruent consonants may be more effective in deriving the measure of intelligibility. In Chapter 4, we have shown that the objective measure of consonant production error is highly correlated with the perceptual intelligibility ratings. In that case, features are extracted from the vowel to consonant transition region and vice-versa, and VOP and VEP events are considered as the anchor points to analyze the transition regions. One important criterion of this method is that VOP and VEP should be detected more precisely, and temporal accuracy of the detection should be high. The temporal accuracy is important to segment the transition region accurately. If the transition region is not properly segmented, there may get a chance to degrade the performance of the algorithm. Moreover, in CLP speech nasal, semi-vowels, and approximants are not misarticulated enough to contribute to the loss

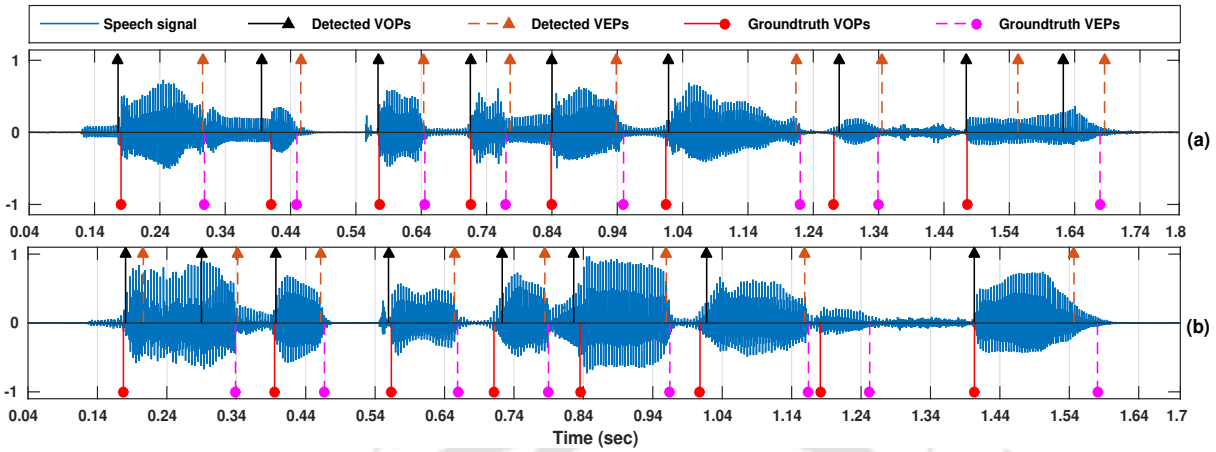

Figure 5.1: Illustration of inaccurate detection of VOPs and VEPs in the (a) normal speech, and (b) CLP speech.

of intelligibility. Hence, features extracted around the transition region of these sounds to vowel may not provide significant information in intelligibility assessment. Some of the times those features may likely provide degrading information in deriving the objective measure. This is because detection of VOPs and VEPs in the case of a vowel to these sonorant sounds and vice-versa is not correctly using the explored method.

Figure 5.1 (a) and (b) show two speech signals with automatically detected and manually marked ground-truth VOPs are VEPs uttered by a normal speaker and a CLP speaker with IL-0, respectively. From the figure it can be seen that VOPs and VEPs are detected almost accurately within certain temporal tolerance, except a few instances where the deviation is significant, e.g., the VOP just before 0.44 sec (sonorant consonant to vowel transition), VEP just after 0.74 sec (vowel to sonorant sound transition), VEP near 0.94 sec (vowel to voice consonant transition), and VOP after 1.24 sec (/r/ to vowel transition). Additionally, a set of extra VOP and VEP are substituted near 1.54 sec and 1.64 sec. Similar to the normal speech, in CLP speech also some of the VOPs and VEPs are detected inaccurately and some of them are not detected (VOP just before and VEP just after 1.24 sec), though most of them are detected almost accurately. Most of the inaccurate detection occurs for sonorant consonant to vowel and vice-versa; inaccurate detection for VOPs and VEPs associate with these sounds may severely degrade the performance. Moreover, the insertion of VOPs and VEPs also likely to degrade the intelligibility prediction performance. The perceptual rating for the utterance used in Figure 5.1 (b) is 0; however, the algorithm provides the intelligibility score near the IL-2. Therefore, if we can detect the location where the transition from obstruent to sonorant sound occurs, or vice-

versa, the acoustic feature around that location may be helpful to characterize the deviations. The *g–landmarks* serve the purpose in this case and can be used as the acoustic loci to extract useful acoustic features. We expect that analysis of speech region anchored around *g–landmarks* may anticipate the degree of intelligibility loss in CLP speech. Apart from this, the present and absence of the landmarks expression may provide some insight into the underlying causes of low intelligibility, which can not be obtained from the VOP and VEP detection based approach.

The rest of the chapter is organized in the following order: Section 5.2 describes how the evidence of abrupt acoustic landmarks are distorted in CLP speech due to articulation problems and motivation for using the *g–landmarks* for intelligibility assessment. The proposed framework for intelligibility assessment based on the characterization of spectral deviations near the *g–landmarks* is discussed in Section 5.3. In Section 5.4, a detailed analysis of the acoustic deviations near *g–landmarks* and its relationship to the intelligibility ratings is provided. Section 5.5 provides experimental results and discussion. Finally, the contribution of the chapter is summarized in Section 5.6.

5.2 Deviation of consonant landmark's evidence in CLP speech

In this section, we analyze how acoustic deviation due to the articulation error distorts or eliminates the landmarks evidence in CLP speech. Additionally, we also discuss why considering *g–landmarks* may be more useful compared to other abrupt landmarks. The acoustic characteristics of the consonants are dependent on place, manner and voicing attributes. If there are any changes in these aspects, the acoustic characteristics near the consonant landmarks may change. Moreover, any deformities in the speech production system also lead to the distortion in landmarks expression. The distortion may be reflected as changes in abrupt discontinuities in the spectrum, which is generally used to derive evidences for landmark detection. The evidence to detect landmarks are extracted from the wide-band spectrogram, which is computed by processing the speech signal with a window size of 6 ms and shift of 1 ms [82]. Smaller frame size and higher frame rate are used to monitor abrupt changes of the speech. Then, spectrogram is divided into six frequency bands (Band-1 (B1): 0 - 0.4 kHz, Band-2 (B2): 0.8-1.5 kHz, Band-3 (B3): 1.2-2.0 kHz, Band-4 (B4): 2.0-3.5 kHz, Band-5 (B5): 3.5-5.0 kHz, and Band-6 (B6): 5.0-8.0 kHz) [82]. The energy in each of these six bands is computed by averaging the square magnitude of spectrogram over the corresponding frequency band. The rate of each energy increase and decrease is used to detect different consonant landmarks. Band-1 is used to detect the

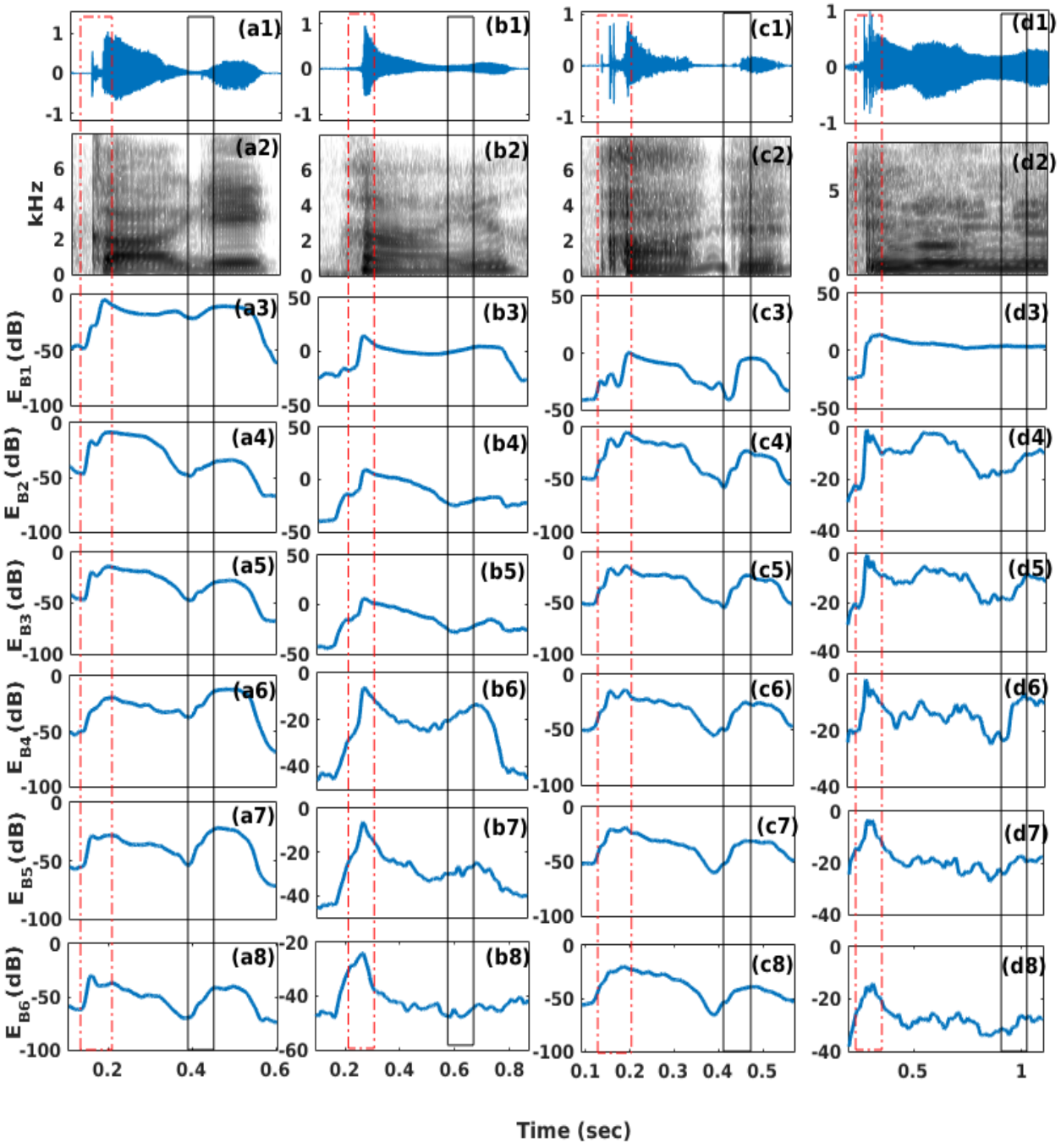

Figure 5.2: Illustration of the changes in landmark's evidence due to articulation errors in CLP speech for the target word /kage/. Speech signal, spectrograms, and smoothed band energies are shown for normal production (a1-a8), and CLP speech with different articulation errors, such as glottal fricative for /k/ and /y/ for /g/ sounds (b1-b8), glottal stops substitution for /k/ and /g/ (c1-c8), and nasal substitution for /k/ and /g/ (d1-d8), respectively. Red dotted rectangles represent the target /k/ transition region, while black solid rectangles represent the /g/ transition region.

TH-2142_146102012

g–landmarks, while B2-B5 are used to detect the *b–landmarks* and *s–landmarks*.

Figure 5.2 shows all the smoothed band energies of speech utterances for the target word /kage/ produced by normal healthy speaker (a1-a8), and CLP speakers with different articulation errors, such as glottal fricative for /k/ and /y/ for /g/ sounds (b1-b8), glottal stop substitutions for /k/ and /g/ (c1-c8), and nasal substitutions for /k/ and /g/ (d1-d8), respectively. From the figure, it can be observed that B1 energy (E_{B1}) waveform corresponds to glottal vibration, and energy increase and decrease correspond to the start and end of the glottal vibration, respectively. The dip in E_{B1} waveform is observed during the obstruents production in case of column-1, column-2, and column-3; however, no such dip is observed in column-4, since nasal sound replaced both the obstruents (/k/ and /g/). The evidence of burst (/k/) can be found in the normal production (B2-B5 within the red rectangle (a4-a7)); however, no such evidences are found in the column-2, column-3, and column-4 (B2-B5 ($E_{B2} - E_{B5}$) within red rectangles). This is also evident from the spectrograms of the corresponding signals. Since /k/ and /g/ in column-2 replaced by glottal fricative and a sonorant sound, no acoustic cues related to burst are present. Moreover, the formant transition in adjacent vowels of these sounds are also distorted. In column-3, evidence of *b–landmark* is present due to the production of glottal stop sound, and it is not due to the supraglottal constriction (B2-B5 ($E_{B2} - E_{B5}$) within red rectangles); however, change in the acoustic characteristics due to the glottal stop production can also be captured using the nearby *g–landmark*. Moreover, in column-4, burst characteristics are deviated due to the nasal replacement. Therefore, the *b–landmarks* will be difficult to use as the anchor points to extract features, since the existence of this landmark is questionable due to the reduced oral pressure. The number of deletions in *b–landmarks* may signify the loss of intelligibility, as discussed in [81] for dysarthric speech. Proper exploration is needed for CLP speech in this direction. The acoustic characteristics of obstruents always associated with a *+g–landmark* and/or a *-g–landmark*. Thus, the *g–landmarks* will be useful to extract the information of articulatory deviations of obstruents. From the figure also, it is evident that energy increase and decrease in E_{B1} is associated with the obstruents production. If the obstruents are replaced by nasal consonants, such as in the case of column-4, then the evidence of *g–landmarks* may not be present (E_{B1} in column-4). However, an utterance always starts with a *+g–landmark*, and end with a *-g–landmark*. Thus, there must be one pair of *g–landmarks* in every utterance. This pair of landmarks can be used to extract the feature in that situation. Moreover, the deletion of *g–landmarks* directly signifies the compensation of

Table 5.1: Counts of the *g–landmarks* in normal and different intelligibility levels of CLP speech in case of sentence O1

	Normal	CLP 0	CLP 1	CLP 2	CLP 3
g–landmark (+g and -g)	58 (29, 29)	56 (28, 28)	60 (30, 30)	62 (31, 31)	56 (28, 28)

sonorant consonants for obstruents, which may provide some insight to the SLP about the articulation error.

5.3 *g–landmarks* based intelligibility assessment

In this section, we discuss the methodology to derive intelligibility measure by extracting acoustic features around *g–landmarks*. Initially, a description of the detection of *g–landmarks* is provided, and following that feature extraction procedure around the detected landmarks is discussed. Then the development of sentence-specific GMM using the derived features are discussed, and finally, computation of log-likelihood score based intelligibility measure is discussed.

5.3.1 Detection of *g–landmarks*

As mentioned in Section 5.2, the frequency band of range 0-400 Hz is used to detect the *g–landmarks*, and it is hypothesized that information about the presence or absence of glottal vibration is embedded in this band [82]. The energy in this band is computed by averaging the square magnitude of short-term Fourier transform over the corresponding frequency band. The computed band energy is passed through a two-pass system: *fine* and *coarse* to avoid the noise and to get the high time-resolution [82]. The smoothing parameter and the temporal duration required to smooth the band energy and to compute the rate of rising (ROR) of band energy are different for both the processes, respectively. A window of 20 ms is used for smoothing the band energy and ROR is computed using a 50 ms time step in the *coarse* processing. However, in the *fine* processing, a 10 ms smoothing interval is used instead of 20 ms and a time step of 26 ms is used instead of 50 ms for the calculation of ROR. The *+g–landmark* correspond to abrupt (6 dB or more) energy increase in the band energy, while *-g–landmark* corresponds to the sharp energy decrease in the same band [82]. The detected *g–landmarks* are considered as the anchored points in the speech signal, around which acoustic features are extracted.

We have shown how counts of *+g–landmark* and *-g–landmark* vary with respect to intelligibility

Table 5.2: Counts of the *g–landmarks* in normal and CLP speech for all the sentence stimuli

		O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
g–landmark	Normal	348	310	354	300	298	452	254	328	508	448
	CLP	368	310	372	340	304	410	264	314	498	462

ratings. The counts of *g–landmarks* in normal and different intelligibility level of CLP speech in case of sentence O1 are tabulated in Table 5.1. However, from this analysis, no information can be derived for loss of intelligibility. Apart from this, we have also studied the difference in overall counts of *g–landmarks* between normal and CLP speech. The counts of *g–landmarks* in normal and CLP speech for the all the sentence-level stimuli is noted in Table 5.2. From the table, it can be seen that the number of *g–landmarks* is increased in the case of CLP speech. In this experiment, utterances of 40 healthy and 40 CLP speakers are considered.

5.3.2 Feature extraction

Similar to the earlier chapter, this work also explores MFCC and M2DDCT features, and a detailed description of the features extraction procedure is provided in Chapter 4. For each utterance, *+g–landmarks* and *-g–landmarks* are detected, and around each landmark, a region of 80 ms (40 ms before and 40 ms after *-g–landmarks*) is considered to derive the features.

5.3.3 Development of sentence-specific GMMs

For each sentence-level stimulus, features derived from the regions around *+g–landmarks* and *-g–landmarks* are used to build two separate GMMs. The motivation to build separate GMM for both the *g–landmarks* is to see how acoustic deviations around them individually correlate with the intelligibility ratings. These two GMMs represent the acoustic space of the corresponding sentence stimulus. Hence, for 10 sentences, 20 GMMs (2 GMMs/sentence $\times$ 10 sentences) are developed.

5.3.4 Intelligibility score computation

For a test utterance, features are extracted from the speech region around *+g–landmarks* and *-g–landmarks*. Then, the log-likelihood scores are obtained for the extracted features from the respective GMM. Let us consider, the feature extracted from the speech segments around the *+g–landmarks* of test O1 sentence is denoted as F_{+g}^{O1}. Then log-likelihood scores for the feature F_{+g}^{O1} are computed from the respective GMM (λ_{+g}^{O1}), i.e., GMM built using the features extracted around the *+g–landmark*

from the normal version of O1 sentences. A similar process is applied to compute the log-likelihood scores for the features computed around the *-g–landmark*. For each test utterance, two average values of likelihood scores are computed, one for the features around *+g–landmark* and another for *-g–landmark*. These two mean likelihood scores are considered as the acoustic correlates of the CLP speech intelligibility. The process of log-likelihood score computation is similar for all the ten sentence-level stimuli used herein. Apart from predicting the intelligibility using individual measure based on the features extracted around *+g–landmark* and *-g–landmark*, information of both the landmarks can be combined to derive the intelligibility measure. Since both the *g–landmarks* represent the totally different acoustic characteristic of the speech signal, combining them may provide the complementary information which improves the performance. Further, the linear regression model is used to map log-likelihood scores from *+g–landmark* and *-g–landmark* based models to the intelligibility score.

5.4 Analysis of the acoustic deviations near *g–landmarks* and its relation to the intelligibility loss

In this section, we analyzed the expression of *g–landmarks* in CLP speech and studied how the acoustic characteristics near these landmarks deviate. Also, we have shown the potential of landmark-based analysis to provide clinically relevant information. Figure 5.3 (a, b), (d, e), (g, h), (i, k), and (m, n) show speech signal corresponding to the target O1 sentence (see Table 3.2) with *+g–landmark* and *-g–landmark* and spectrograms for normal, CLP intelligibility level(IL)-0, CLP IL-1, CLP IL-2, and CLP IL-3, respectively. In the normal speech signal, five *+g–landmarks* and five *-g–landmarks* are detected, and all the *+g–landmarks* and *-g–landmarks* are associated with the obstruent to vowel and the vowel to obstruent transition region, respectively. However, the number of *g–landmarks* detected for the speech signal of CLP speaker with IL-0 is reduced, as phoneme /g/ is heavily voiced (red dashed rectangle in Figure 5.3 (d), around 0.2 sec). The acoustic characteristics near *+g–landmarks* and *-g–landmarks* of CLP IL-0 are almost similar to normal, which can be seen from the respective spectrogram in Figure 5.3 (e). Though the number of *g–landmarks* in CLP IL-1 and CLP IL-2 speech is the same as normal, the acoustic characteristics near landmarks deviate from that of normal. In CLP IL-1 speech, phoneme /g/ is replaced by a unvoiced sound (red dashed rectangle in Figure 5.3 (g), around 0.3 sec), while in CLP IL-2, it is compensated by a glottal stop (red dashed rectangle in Figure 5.3 (j), around 0.3 sec), which distorts the voice bar and formant transitions in the adjacent sonorant sound. In the speech signal of CLP IL-3, all the obstruents are replaced by nasal consonants,

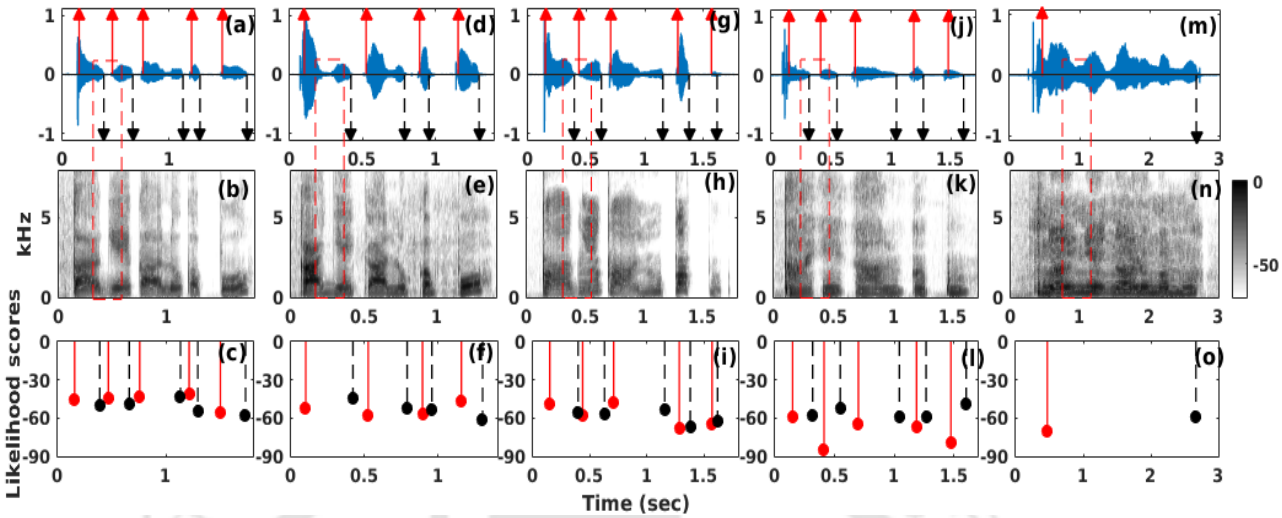

Figure 5.3: Time waveforms with detected *+g–landmarks* and *-g–landmarks*, spectrograms, and log-likelihood scores at *+g–landmarks* and *-g–landmarks* of target sentence O1 (*kage kalu kapːu*) for normal (a, b, and c), CLP Intelligibility Level (IL)-0 (d, e, and f), CLP IL-1 (g, h, and i), CLP IL-2 (j, k, and l), and CLP IL-3 (m, n, and o), respectively. All vowels and lateral liquids of CLP speech are nasalized. In (a), (d), (g), (j), and (m), upper solid (red) and down dashed (black) arrows represent the +g landmarks and -g landmarks, respectively. In (c), (f), (i), (l), and (o) red solid down arrows and black dashed down arrows represent the log-likelihood scores computed at *+g–landmarks* and *-g–landmarks*, respectively. Dashed rectangles represent the locations of target /g/ phoneme

which results in only one *+g–landmarks* and one *-g–landmarks* (Figure 5.3 (m)). Thus, apart from the deviations in the landmarks expression, the acoustic correlates of stop production, such as burst energy and formant transitions in the consonant-vowel (CV) and vowel-consonant (VC) transition regions are also distorted. Therefore, the analysis of acoustic features around *+g–landmarks* and *-g–landmarks* may provide the degree of intelligibility loss. Figure 5.3 (c), (f), (i), (l), and (o) show the log-likelihood scores around the *+g–landmarks* and *-g–landmarks* in case of sentence O1 for normal, CLP IL-0, CLP IL-1, CLP IL-2, and CLP IL-3, respectively. Log-likelihood scores are computed from the M2DDCT-based GMMs. As the intelligibility degrades from CLP IL-0 to CLP IL-3, the log-likelihood scores decrease accordingly. In the next paragraph, a detailed description is provided on how likelihood scores and number of detected *g–landmarks* are related to the intelligibility. Figure 5.4 shows a 2D plot of the mean log-likelihood scores and number of detected *g–landmarks* for normal, CLP IL-0, CLP IL-1, CLP IL-2, and CLP IL-3 in case of sentence stimulus O1. From the figure two important points can be noticed (i) the deviations of log-likelihood scores from the normal average value, and (ii) the reduction of *g–landmark's* count some specific CLP speakers. It provides a visual representation of how numbers of *g–landmarks* and likelihood scores give information about the loss of intelligibility.

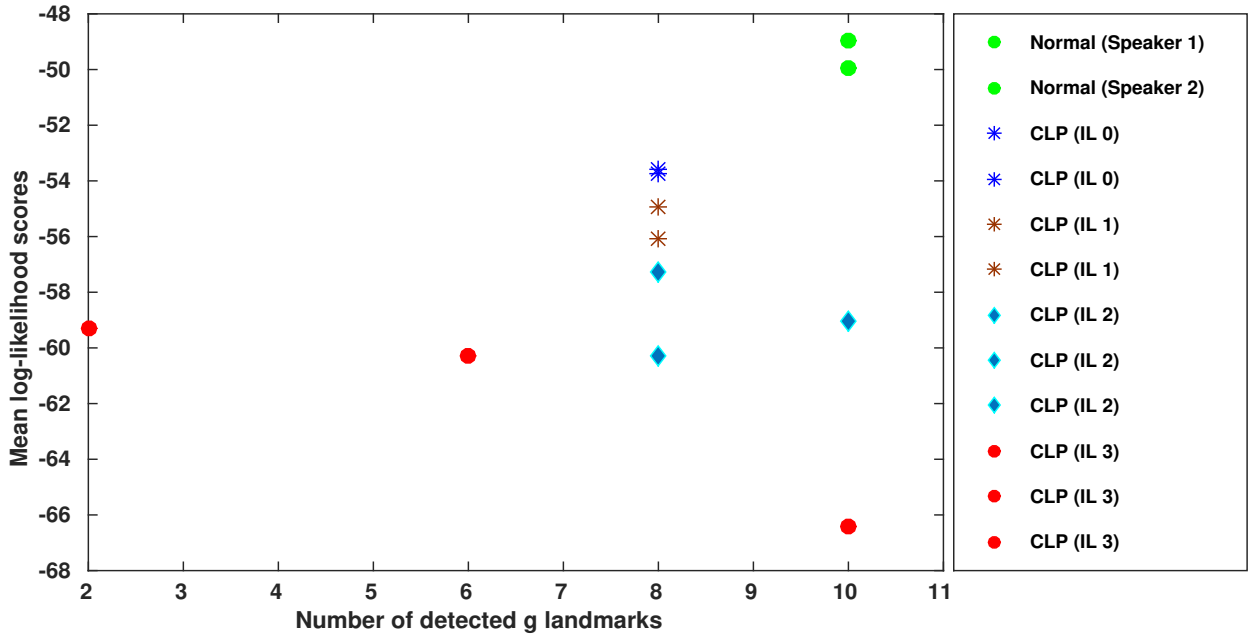

Figure 5.4: Number of *g–landmarks* vs. mean log-likelihood scores for different normal and CLP speakers in case of O1 sentence.

The estimation of voicing is linked to *g–landmarks*; therefore, the reduction of *g–landmarks* count represents the replacement of the obstruents by sonorant sounds. A similar plot is shown in Figure 5.5 for the sentence stimulus O8, where the number of *g–landmarks* for a normal speaker is 14. In the clinical setting, this representation may be helpful for the SLPs for better planning of the therapy. If for one speaker, all the hypothesized landmarks are intact, but the likelihood scores deviate from the normal speaker, then there may be a chance of velar substitution or laryngeal compensation. This situation is observed for one of the CLP speakers with IL-3 in Figure 5.4, whose number of detected *g–landmarks* are same as that of the hypothesized number, i.e., 10. However, the mean log-likelihood score much deviates that of the normal. It is found that this speaker compensated all the obstruents by the glottal stops, due to which the number of *g–landmarks* is intact, whereas the log-likelihood score decreases. However, if the number of *g–landmarks* is reduced, there may be a chance of occurring the nasal replacement for the obstruents. This situation is observed for another speaker with IL-3 and number of detected *g–landmarks* is 2, who have severe hypernasality and nasal consonants replace all the obstruents, and the log-likelihood score is also very low. Moreover, this representation may also have a scope to act as a self-evaluation tool for CLP children. In this case, the children can set a goal to achieve the log-likelihood score near the normal speech with the proper

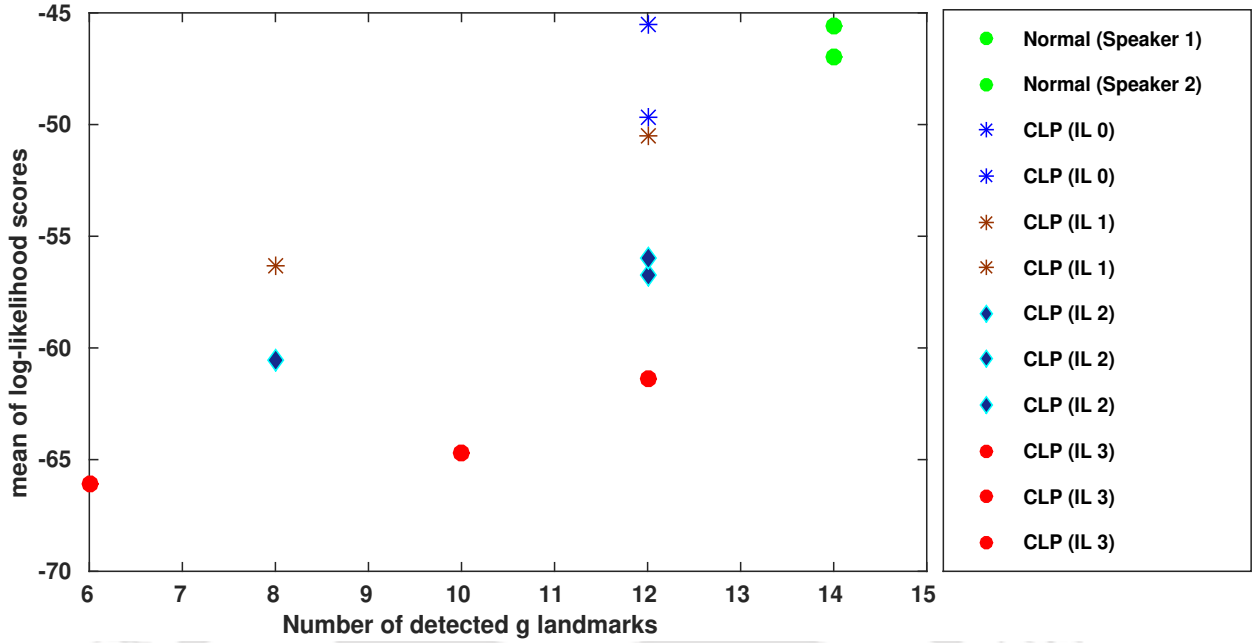

Figure 5.5: Number of *g–landmarks* vs. mean log-likelihood scores for different normal and CLP speakers in case of O8 sentence.

number of *g–landmarks*. However, further exploration is needed to analyze its effectiveness to derive clinically relevant information useful for the SLPs.

A careful observation of the log-likelihood scores around the *g–landmarks* of CLP speech in Figure 5.3 provides some information about the localized articulatory deviation near them. However, current work only exploits the global deviations of the log-likelihood scores for each *g–landmark* to interpret the loss of intelligibility. We also show the importance of landmark-based processing in deriving the localized acoustic characteristic deviations around the *+g–landmarks* and *-g–landmarks*. Figure 5.3 (c), (f), (i), (l), and (o) show the log-likelihood scores around *+g–landmarks* and *-g–landmarks* in case of sentence O1 for normal, CLP IL-0, CLP IL-1, CLP IL-2, and CLP IL-3, respectively. The O1 sentence has five *+g–landmarks* and the same number of *+g–landmarks* (Figure 5.3 (c)), and for properly articulated speech utterance, the log-likelihood score in the respective locations should be similar. However, for CLP speech with CLP IL-0 four +g and four -g landmarks are detected with approximately similar log-likelihood scores as in case of normal (Figure 5.3 (f)). In CLP IL-1, same number of*+g–landmarks* and *-g–landmarks* are detected; however, the log-likelihood scores around them are deviated than that of normal (Figure 5.3 (i)). In CLP IL-2 also same number of *+g–landmarks* and *-g–landmarks* are detected (Figure 5.3(l)). However, the log-likelihood scores are

very less and especially for the *+g–landmarks*, as glottal stops replace all the obstruents. In the case of CLP IL-3, all the obstruents are replaced as nasal consonants, and only two *g–landmarks* is detected (Figure 5.3 (o)). The reduced localized log-likelihood scores may provide the extent of articulatory deviations in the respective landmark positions. Hence, this localized information may be helpful during the therapy, where feedback for the number of *+g–landmarks* and *-g–landmarks* are required to produce for a corresponding sentence and provide the score of precision to articulate each syllable in terms of likelihood value.

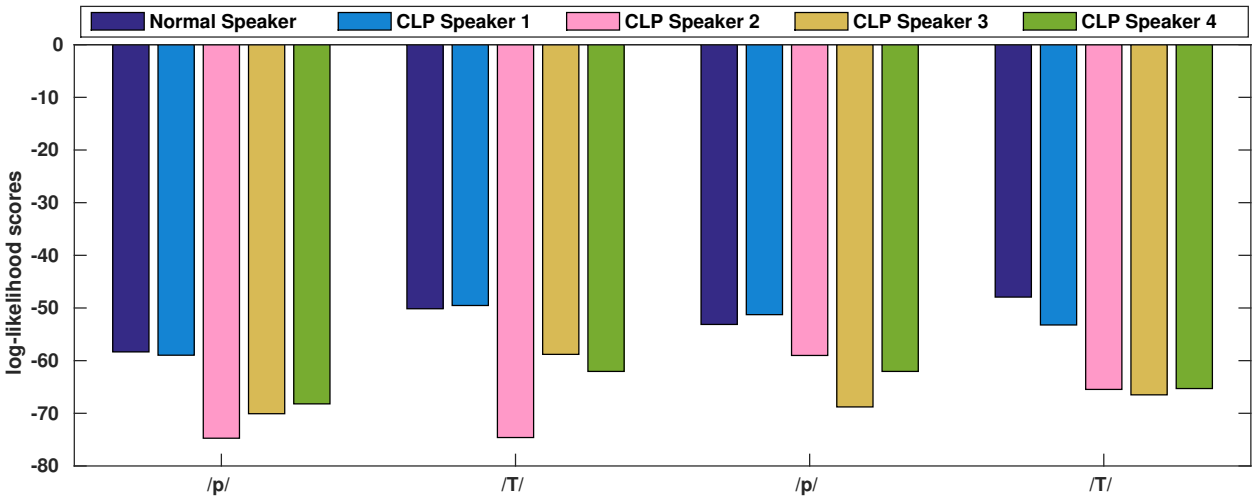

Figure 5.6: Bar plots of the log-likelihood scores for normal speaker, CLP speaker with correct production (CLP Speaker 1), and three CLP speakers with misarticulation (CLP Speakers 2-4) for the target sounds /p/ and /ʈ/ in the context of /paʈa/. Here, /T/ represents the /ʈ/.

To show the clinical importance of localized log-likelihood scores, let us consider the sentence-level stimulus O17 and a target word /paʈa/ of this sentence for the case study. Here, /p/ and /ʈ/ sounds are considered as the target consonants to be evaluated. Bar plots of the log-likelihood scores for one normal speaker, one CLP speaker with correct production, and three CLP speakers with misarticulation for the target sounds /p/ and /ʈ/ in the context of /paʈa/ are shown in Figure 5.6. From the bar plots, it can be seen that mean log-likelihood scores around the *g–landmarks* near /p/ and /ʈ/ for normal and CLP speaker 1 are all most similar. This is because CLP speaker 1 produces these two sounds correctly. However, the log-likelihood scores are decreased for the case of other CLP speakers. CLP Speaker 2 produces the unvoiced velar stops for the /p/ and /ʈ/. While, both /p/ and /ʈ/ are nasalized for the case of CLP speaker 3. Additionally, the CLP Speaker 4 replace /p/ sound by glottal fricative and /ʈ/ by glottal stop. The detailed analysis to check the effectiveness of localized

log-likelihood scores to describe the articulation problem is out of the scope of present thesis. Future exploration is required to investigate the localized log-likelihood scores, and its clinical importance in analyzing the articulation errors.

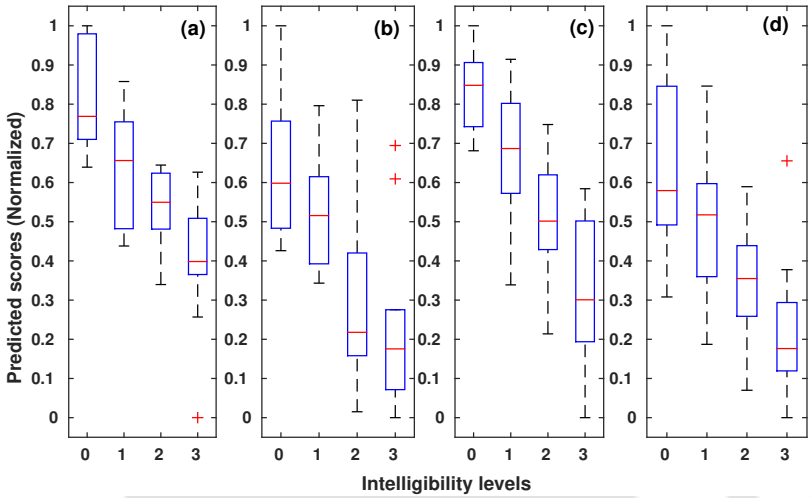

Figure 5.7: Box plots of the log-likelihood scores for different level of intelligibility in case of O1 sentence. (a) $MFCC_{+g}$, (b) $MFCC_{-g}$, (c) $M2DDCT_{+g}$, and (d) $M2DDCT_{-g}$.

5.5 Results and discussion

In this section, experimental results and performance evaluation of the proposed *g–landmarks* based intelligibility measures are discussed. Initially, we have studied how predicted intelligibility scores are distributed with respect to the perceptual ratings. The log-likelihood scores of different intelligibility groups are shown in Figure 5.7. It is shown for all the measures in case of sentence-level stimulus O1. All the scores are normalized to the [0, 1] interval. As intelligibility degrades, the mean log-likelihood scores also decrease for all the features used herein. The deviation of predicted scores for each group of intelligibility is more for the measures based on *-g–landmark* (Figure 5.7 (b) and (d)). However, it is comparatively less for the *+g–landmark* based measures. The mean and standard deviation based analysis is performed to quantify the inter-group difference for each measure. Table 5.3 shows that the mean values are almost equally spaced for $MFCC_{+g}$ and $M2DDCT_{+g}$ based measures, as compared to $MFCC_{-g}$ and $M2DDCT_{-g}$. Also, the standard deviation of the predicted scores for each group is significantly high in the case of $MFCC_{-g}$ and $M2DDCT_{-g}$ based measures. Thus, intelligibility scores derived using the features computed around *-g–landmark* is not consistent, which may lower the performance of those measures.

Table 5.3: Results of mean (μ) and standard deviation (σ) based statistical analysis of intelligibility scores for O1 sentence stimulus.

Groups	Scores (normalized)			
	$\mu \pm \sigma$			
	$MFCC_{+g}$	$MFCC_{-g}$	$M2DDCT_{+g}$	$M2DDCT_{-g}$
IL0	0.83 ± 0.14	0.64 ± 0.19	0.83 ± 0.10	0.64 ± 0.22
IL1	0.65 ± 0.15	0.52 ± 0.15	0.67 ± 0.19	0.49 ± 0.18
IL2	0.53 ± 0.09	0.29 ± 0.21	0.51 ± 0.14	0.34 ± 0.15
IL3	0.39 ± 0.17	0.24 ± 0.23	0.31 ± 0.20	0.22 ± 0.19

Table 5.4: Mean (μ) and standard deviation (σ) of 10 individual sentence-level correlations for overall performance evaluation

Measures	ρ ($\mu \pm \sigma$)	p-value
$MFCC_{+g}$	0.68 ± 0.062	< 0.001
$MFCC_{-g}$	0.62 ± 0.069	< 0.001
$MFCC_{+g}$ + $MFCC_{-g}$	0.72 ± 0.064	< 0.001
$M2DDCT_{+g}$	0.70 ± 0.055	< 0.001
$M2DDCT_{-g}$	0.65 ± 0.061	< 0.001
$M2DDCT_{+g}$ + $M2DDCT_{-g}$	0.74 ± 0.059	< 0.001

The Spearman's rank correlation coefficient (ρ) between the estimated scores and the perceptual rating is considered as the performance measure. Both the variables, i.e., perceptual grades and log-likelihood scores are non-normally distributed; therefore, Spearman's rank correlation coefficient will be appropriate [128]. Leave-one-speaker-out cross-validation (LOSO-CV) is carried out for performance evaluation. The acoustic-phonetic composition of each sentence stimulus is different; hence, the number of Gaussians which can effectively model the acoustic space of each sentence stimulus will be different. Therefore, we have experimented for the different number of component Gaussians to build the GMM, and the number of Gaussian with the best result is considered for evaluation. For each fold of LOSO-CV, all the CLP children's utterances except one speaker's speech utterances are used to build a linear regression model as discussed in Chapter 4. This cross-validation process is applied separately for each sentence stimulus used herein.

For all the 10 sentence-level stimuli, correlations between the objective intelligibility scores and the perceptual ratings are calculated individually. Then, the average of 10 individual sentence-level correlations is considered to study the overall performance of the system. Table 5.4 shows the average correlation for all the sentence-level stimuli. It can be clearly observed that the correlation values are relatively high for the +g model than -g model for both MFCC (ρ=0.68, 0.62) and M2DDCT

Table 5.5: Results of Williams significance test conducted between pairs of acoustic correlates of intelligibility. (p-value of a given pair of measures is computed whose absolute Spareman rank correlation with perceptual rating is higher than that of the other in the pair. p-value less than 0.05 is considered as the statistically significant). Here, * represents a p-value greater than 0.05, and † represents a p-value less than 0.05

–	*	†	†	†	†	M2DDCT$_{+g}$ + M2DDCT$_{-g}$
–	–	†	†	†	†	MFCC$_{+g}$ + MFCC$_{-g}$
–	–	–	*	†	†	M2DDCT$_{+g}$
–	–	–	–	†	†	MFCC$_{+g}$
–	–	–	–	–	†	M2DDCT$_{-g}$
–	–	–	–	–	–	MFCC$_{-g}$
M2DDCT$_{+g}$ + M2DDCT$_{-g}$	MFCC$_{+g}$ + MFCC$_{-g}$	M2DDCT$_{+g}$	MFCC$_{+g}$	M2DDCT$_{-g}$	MFCC$_{-g}$	

(ρ=0.70, 0.65) features, respectively. This high correlation value in case of +g model is justified, as it captures the characteristics of transition regions and preceding obstruent regions in most of the times. Least correlation value is observed in case of MFCC$_{-g}$ (ρ=0.62). For both *+g–landmark* and *-g–landmark* based measures, M2DDCT gives comparatively high correlation than MFCCs. The high correlation using M2DDCT show the significance of JST features in better representing the acoustic characteristics near the landmarks. Later, the combination of *+g–landmark* and *-g–landmark* based measures outperforms the individual measures. It can be seen that compared to MFCC$_{+g}$ and MFCC$_{-g}$ measure, their combination gives an increment in correlation value of 0.04 and 0.1, respectively. Also, as compared to M2DDCT$_{+g}$ and M2DDCT$_{-g}$ measures, their combination gives an increment in correlation value of 0.04 and 0.09, respectively. The measure based on the combination of M2DDCT$_{+g}$ and M2DDCT$_{-g}$ gives improvement over the combination of MFCC$_{+g}$ and MFCC$_{-g}$ measures. The combination of scores from *+g–landmark* and *-g–landmark* model utilize the acoustic characteristics of both the onset and offset transition region; hence, improves the performance.

Although we see significant improvement in correlation over the MFCCs, we must test for statistical significance. To study the statistical significance of the difference in correlation values, Williams pair-wise significance test [90] is performed. In this case, we perform the significance test for each pair of acoustic correlates of CLP speech intelligibility. Table 5.5 lists the outcomes of the *+g–landmark* and *-g–landmark* test. In the table, each p-value inside a cell (i, j) indicates whether measure i (named in the rightmost column of the table) is correlated significantly higher with the perceptual ratings than is measure j (named in the bottom of the table). From the table, it can be seen that the better correlation using M2DDCT$_{+g}$-based measure as compared to M2DDCT$_{-g}$ and MFCC$_{-g}$ based measures is statistically significant at $p < 0.05$. Similarly, increased ρ value for MFCC$_{+g}$-based

measure relative to $MFCC_{-g}$ is statistically significant. However, increased ρ value for $M2DDCT_{+g}$ relative to $MFCC_{+g}$ is not statistically significant at $p<0.05$. The increased ρ value for the combination of *+g–landmark* and *-g–landmark* based measures is statistically significant for both the features, as compared to the individual measures.

We have shown that it is indeed possible to derive the acoustic correlates of the CLP speech intelligibility by extracting acoustic features in the vicinity of the *g–landmarks*. The potentiality of both the *g–landmarks* in estimating the intelligibility is shown. Since the features are extracted in the vicinity of the abrupt spectral change, acoustic features which can capture these sharp spectral discontinuity are required. For both the landmarks, M2DDCT feature outperforms MFCCs, which signifies the importance of JSTFs in deriving the acoustic correlates of intelligibility. The JSTFs may retain the critical discriminatory information in the time-frequency plane about the articulatory deviations in the CLP speech, which helps to better discriminate among the groups. The loss of intelligibility is the cumulative effect of the malformed acoustic cues present around the *+g–landmark* and *-g–landmark*. Therefore, the combination of scores from the *g–landmark* significantly improves the performance. Since for each sentence-level stimuli separate GMM is needed, this may lead to the complexity of the proposed algorithm. Even so, if normal speakers produce heavily voiced obstruent consonants, the detection of the *g–landmark* may not be proper in their speech. Further, refinement of the *g–landmark* detection will be needed in the case of heavily voiced obstruent consonants.

5.6 Summary and conclusions

In this work, the importance of *g–landmark* in deriving the acoustic correlates of CLP speech intelligibility is studied. Two acoustic features, namely, MFCCs and M2DDCT features are extracted in the vicinity of the *g–landmark*, which characterize the acoustic deviation near those landmarks. For each sentence stimulus, two separate sentence-specific GMMs are built for *+g–landmark* and *-g–landmark* using the extracted features. While testing, utterance wise mean log-likelihood scores are computed from the respective GMMs, which is considered as the proposed acoustic correlates of CLP speech intelligibility. Results show that M2DDCT based +g model gives the highest correlation ($\rho = 0.70$), while the MFCCs based -g model gives the lowest correlation ($\rho = 0.62$) with the perceptual ratings. The combination of both log-likelihood scores obtained from +g model and -g model is also studied as the intelligibility measure. A 2D-plot based on likelihood score vs. the number of detected

g–landmarks is proposed to get some information about the loss of intelligibility. We have also shown the possibility to study a particular target sound which may be responsible for the intelligibility deviations. However, a detailed analysis is needed to explore the potentiality of landmarks to reveal the underlying factors of reduced intelligibility.

Intelligibility measures proposed in the present and previous chapters are dependent on the detection of certain events of the speech signal. Therefore, accurate detection of these events is important to derive reliable intelligibility measures. In the next chapter, a framework to estimate CLP speech intelligibility is proposed based on the comparison of posterior sequences of normal and CLP speech.

6

Posterior Sequence based Intelligibility Assessment

Publications

- **Sishir Kalita**, S. R. M. Prasanna, S. Dandapat, "Intelligibility assessment of cleft lip and palate speech using joint spectro-temporal features based gaussian posteriogram", *The Journal of the Acoustical Society of America*, 144(4), October (2018), PMID: 30522275.
- **Sishir Kalita**, S. R. M. Prasanna, S. Dandapat, "Self-similarity matrix based intelligibility assessment of cleft lip and palate speech", in *Proc. Interspeech 2018*, Hyderabad, India, September 2018.
- **Sishir Kalita**, "Objective assessment of cleft lip and palate speech intelligibility", 4^{th} *Doctoral consortium, Interspeech 2018*, Hyderabad, India, September 2018.

Contents

Objective:

The present chapter proposes two comparison based frameworks using dynamic time warping (DTW) and matching of self-similarity matrices (SSMs) to quantify the intelligibility of CLP speech. Initially, acoustic features are computed for an utterance and mapped into Gaussian posteriorgrams (GPs). The motivation for using GP is that it provides speaker independent acoustic segment representation, and can be derived completely in an unsupervised manner. GP representation of the distorted unintelligible speech of CLP children will be distinctly different from the normal children's speech. Then in the first approach, DTW is used to compute the deviation of GP of test speech from the normal speaker's template, and DTW distance is studied as a correlate of intelligibility. However, in SSM-based approach, the posterior sequence of the speech signal is transformed into another representation, termed as self-similarity matrix. The SSM of a posterior sequence is a square matrix, which encodes the acoustic-phonetic composition of the underlying speech signal. Deviations in the acoustic characteristics of underlying sound units due to the degradation of intelligibility may deviate the CLP speech's SSM structure from that of normal speech. The degree of deviations is quantified using the structural similarity (SSIM) index, which is considered as the representative of objective intelligibility score. Since the SSM-based representation is robust against the speaker variabilities; therefore, SSM computation from the GP may provide better performance as compared to the DTW-based measure. Additionally, to show the importance of SSM-based measure over DTW in deriving intelligibility measure, these systems are also developed using the raw feature representation without using GPs. Spearman's rank correlation coefficient between the objective intelligibility scores and the perceptual intelligibility rating is studied.

6.1 Motivation for using Gaussian posteriorgrams

In Chapters 4 and 5, we have proposed intelligibility measures based on the characterization of acoustic deviations around specific events or glottal activity region of the speech signal. Those approaches are mainly motivated by the fact that intelligibility of the CLP speech is primarily degraded due to the effect of articulation error and hypernasality. Although, intelligibility measures based on both the approaches have shown significant correlation with perceptual ratings; however, they are dependent on the detection of certain acoustic events of the speech signal. Therefore, the performance highly relies on the accurate detection of those events. Moreover, the detection may be more difficult

for the case of CLP speech due to deviant speech characteristics. Also, it is found that correlation values are not highly consistent, which can be inferred from the high standard deviation of sentence-level correlations in Table 4.3 and Table 5.4. This high standard deviation may be due to the fact that for some sentences the event detection is not accurate, which leads to less correlation for those sentences. Whereas for other sentences, the event detection is proper, and correlation is high for those sentences. However, this situation does not seem to be desirable for an intelligibility prediction algorithm. To overcome these issues, in this chapter, methods are proposed based on the comparison of posterior sequences to measure the intelligibility. Here, speech utterance from the normal group is considered as the template with which CLP speaker utterance is compared, and deviation from the normal template is considered as the degree of intelligibility loss. The approach is also motivated by the normal speech intelligibility assessment, where the knowledge of clean/undistorted speech is needed [142]. These measures typically compare the auditory spectro-temporal representations of the test signal to those of the undistorted reference signal. The comparison of acoustic properties requires both signals be from the same speaker [142]. However, these measures may not be suitable to assess the intelligibility of pathological speech, where a more intelligible version of the recording from the same speaker is not available. To solve this problem in intelligibility assessment of synthesized speech DTW-based comparison between phoneme class-conditional probability sequences of original and synthesized speech is proposed [142]. Posterior features, which provide speaker-independent phonetic representation may compensate the speaker variabilities. Therefore, deviations in acoustic properties can be captured by comparing two posterior sequences. However, to compute the phonetic posteriorgram a phonetic classifier for a specific language is required [121]. A large amount of transcribed data is required to develop the phonetic classifier, which may not be available for under-resourced scenarios. Therefore, posterior features which can be derived from unsupervised modeling of the acoustic units may be more helpful. Other efforts have focused on developing distance measures between healthy and pathological speech [53, 54], where speech uttered by a normal healthy speaker is considered as the reference for measuring the speech quality of Parkinson speech. Initially, the authors use DTW to align the speech signal, and then several distance measures, such as Itakura-Saito, log-likelihood ratio, and cepstral distance are used to compare the test and reference speech. A similar approach has also been explored to estimate the tracheoesophageal speech quality objectively; however, the authors used perceptually relevant features computed from the auditory model in their method [143]. Apart from the intelligi-

bility and quality assessment, several other studies also reported the effectiveness of comparison-based frameworks for (i) objectively evaluate the mispronunciations of non-native speakers [144, 145], (ii) assessing intelligibility deficits in patients with Parkinson's disease [146], and (iii) scoring shadowing speech automatically [147]. Thus, a comparison based framework which utilizes the knowledge about the acoustic-phonetic composition of underlying speech stimulus may be helpful in this regard. The attractiveness of comparison based approach, such as DTW is that they do not make any assumption about the underlying linguistic information [121]. Moreover, the generalization of these systems while porting from one language to another for speech assessment may not be difficult in the case of low-resource scenario as well.

6.2 Contributions

Two comparison-based frameworks using DTW and matching of SSMs are applied to compute the deviation of test CLP speech from the normal template. The relative deviation from the normal speaker's template is considered as the representation of intelligibility loss. However, a speaker-independent representation of speech is required to characterize each acoustic-phonetic unit before comparing two speech signals. Since raw features, such as MFCCs may possess the inter- or intra-speaker variabilities; thus, comparison between reference and test may not give the deviations due to loss of intelligibility but may be due to the speaker variabilities. Therefore, we map the features into Gaussian posteriorgrams (GPs) [121], which are derived from a sentence-specific GMM built using normal speaker's speech. The beauty of GP is that it can be derived in a completely unsupervised manner, without requiring any annotated speech data as in the case of phonetic or phonological posteriorgrams. It provides a speaker-independent representation of the underlying acoustic segments that are present in the respective utterance. GP representation of the distorted unintelligible speech of CLP children will differ distinctly from that of normal child speech. In the case of DTW-based method, DTW distance between the GP representation of test CLP speech and the normal speech template is considered as the sentence-level intelligibility score. The hypothesis is that a lower accumulated DTW distance represents higher CLP speech intelligibility. However, in SSM-based approach, the posterior sequences of speech signals to be compared are transformed individually to another representation termed as self-similarity matrix, and comparison between them is performed in that representation. The SSM of any feature sequence is a square matrix, which encodes the acoustic-phonetic composition

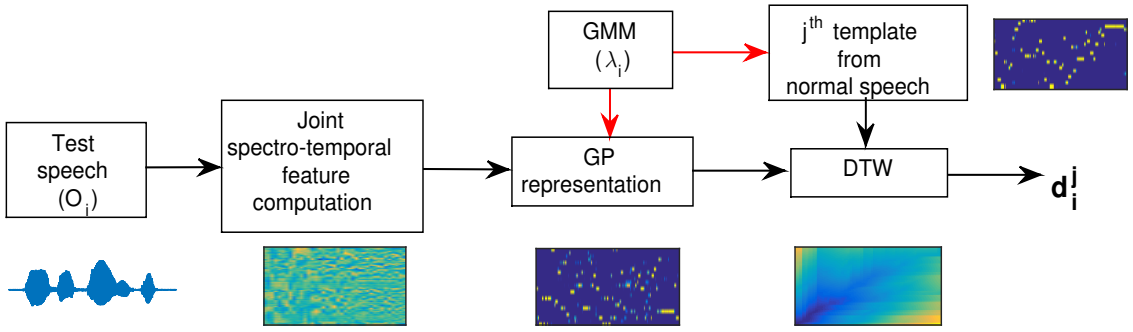

Figure 6.1: Block diagram of DTW-based intelligibility measure.

of the underlying speech signal. Deviations in the acoustic characteristics of underlying sound units due to degradation of intelligibility will deviate CLP speech's SSM structure from that of normal speech. This degree of deviations in CLP speech's SSM from the corresponding normal speech's SSM may provide information about the severity profile of speech intelligibility. The degree of deviations is quantified using the structural similarity (SSIM) index [148], which is considered as the representative of the objective intelligibility score. Since SSM-based representation is robust against the speaker variabilities; therefore, SSM computation from the GP may provide better performance as compared to the DTW-based method. Thus, the salient contributions reported in this chapter are summarized as follows.

- Analysis of the importance of GP-based speech representations for intelligibility assessment.
- Assessment of intelligibility using a DTW-based framework with GP.
- Exploration of the SSM representation of speech for intelligibility assessment.

The rest of the chapter is organized as follows: In Section 6.3, we describe the methodology of proposed work, illustrating the computation GP-based speech representation and its importance in analyzing the intelligibility degradation of CLP speech. We follow that by describing the DTW-based and SSM-based frameworks. In Section 6.4, a detailed description of the experimental results of the proposed methods is provided. Finally, the works presented in the chapter are summarized in Section 6.5.

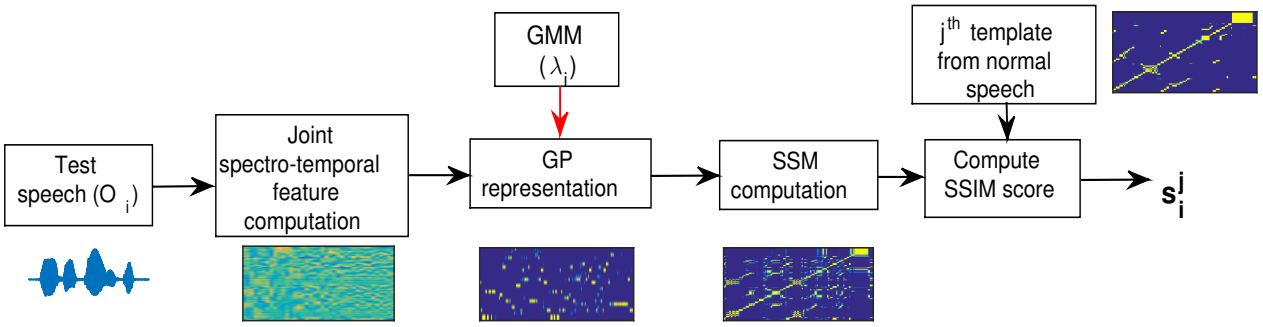

Figure 6.2: Block diagram of SSM-based intelligibility measure.

6.3 Comparison-based frameworks for intelligibility assessment

In this section, we discuss the methodology of proposed approaches for intelligibility assessment. The block diagrams for both DTW-based and SSM-based methods are shown in Figure 6.1 and Figure 6.2, respectively. Acoustic features extraction and GP representation are the same for both the methods; however, the subsequent blocks are different for an individual method. Similar to the earlier chapters, MFCC and M2DDCT features are considered in the present chapter. However, unlike the earlier chapters, those features are not directly used, but GPs are derived from those acoustic features to get the speaker-independent representations. The GP derived from the M2DDCT feature is termed as $\text{GP}_{\text{M2DDCT}}$, whereas, the one derived from the MFCCs is termed as GP_{MFCC}. Therefore, in this work, we are exploring four features, where two features are based on the unsupervised modeling of phonetic segments, i.e., $\text{GP}_{\text{M2DDCT}}$, GP_{MFCC}, and others are raw acoustic features, namely, M2DDCT and MFCCs. All the features are used in both DTW- and SSM-based approaches. The DTW-based measures using raw MFCCs and M2DDCT are considered as the baseline measures with which proposed measures are compared. Initially, the feature extraction procedure is discussed, and followed by a detailed description to build the sentence-specific GMM and mapping of the computed features to GPs is given. Then, the description of DTW-based and SSM-based methods to derive the intelligibility measures is provided.

In each experiment, the speech signals are first normalized by the l_2-norm of the signal and down-sampled to a sampling rate of 16 kHz. Short-term energy based speech activity detection is used to mark the speech regions. Because all the recordings were made under sound-proof conditions, we find that energy-based speech activity detection is effective for the present work. Features extracted be-

tween the detected beginning and end points are considered for further analysis. Before the processing, we pre-emphasized the speech signal with a pre-emphasis factor of 0.97.

6.3.1 Acoustic feature extraction

The procedure to compute the 2DDCT-based joint spectro-temporal (M2DDCT) and MFCC features are the same as discussed in Chapter 4. Similar to the earlier chapters, 39-dimensional MFCCs are computed from the short-term processed by a 15 ms hamming window with a shift of 5 ms. The joint spectro-temporal features are computed using 2D-DCT, and 39-dimensional M2DDCT feature is computed from the time-frequency representation (TFR) of the speech signal. To extract M2DDCT, we analyzed the speech signal with a Hamming window of size 15 ms and a shift of 2 ms. Both the calculated features are used to derive GP-based sentence representation.

6.3.2 Gaussian-posteriogram-based sentence representation

The GP is an unsupervised model-based posterior vector proposed for discovering acoustic patterns. It provides speaker-independent feature representation of speech signals [121, 149, 150]. In GP-based speech representation, each frame is represented by a posterior probability vector of Gaussian components in a GMM. In the present work, posteriorgrams are computed from the sentence-specific GMM that is built using a set of sentence-level utterances collected from normal speakers. The expectation maximization algorithm is used to learn the parameters of the GMM in 10 iteration steps. Since 10 sentence-level stimuli are used in the database, 10 multivariate speaker-independent GMMs (λ_j, where $j = 1, 2, \ldots, 10$) are built. A GMM of M component Gaussians is derived by,

$$p(\boldsymbol{x}|\lambda_j) = \sum_{i=1}^{M} \omega_{ij} p_{ij}(\boldsymbol{x}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}), \tag{6.1}$$

where ω_{ij}, $\boldsymbol{\mu}_{ij}$, and $\boldsymbol{\Sigma}_{ij}$, $i = 1, 2, \ldots, M$, correspond to the weights, mean vectors, and covariance matrices, respectively, of the different mixtures for GMM model λ_j. GMM models the distribution of sounds present in the particular sentence across a variety of speakers. Let us consider a sentence O1, which has K feature vectors and can be represented symbolically as $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K]$. The GP for the sentence O1 can be defined as $\text{GP}_{\text{O1}} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_K]$, and each $\boldsymbol{q}_i$, $i = 1, 2, \ldots, K$, is calculated as,

$$\boldsymbol{q}_i = [P(G_1|\boldsymbol{f}_i), P(G_2|\boldsymbol{f}_i), ..., P(G_M|\boldsymbol{f}_i)], \tag{6.2}$$

where G_m, $m = 1, 2, \ldots, M$, corresponds to the m^{th} component Gaussian of sentence-specific GMM λ_1 of sentence O1, and M is the total number of component Gaussians. Dimensions with minimal probability values are set to zero with a pre-defined threshold to avoid an approximation error. Discounting-based smoothing of the posterior vector is then performed to pass a small fraction of probability from non-zero dimensions to zero dimensions [121, 149]. Although the GMM does not explicitly model the phoneme units present in the sentence, each Gaussian component approximates a phoneme-like class [121]. Thus, the mean of the component Gaussians spans an acoustic space that characterizes a particular sentence. The dimension of a posteriogram is equal to the number of component Gaussians used for the GMM.

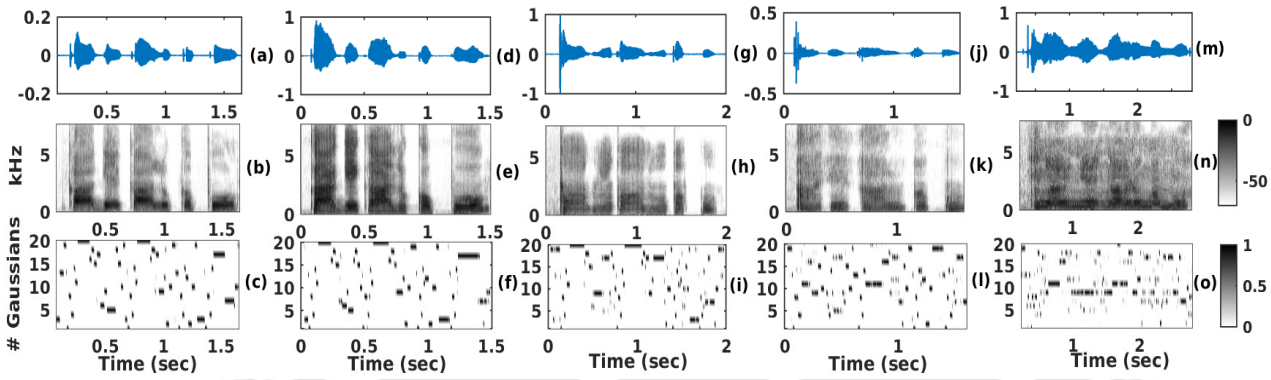

Figure 6.3: Time waveforms, spectrograms, and Gaussian posteriorgrams of speech signals for target sentence stimulus O1 for normal (a, b, and c), CLP IL-0 (d, e, and f), CLP IL-1 (g, h, and i), CLP IL-2 (j, k, and l), and CLP IL-3 (m, n, and o), respectively.

From the spectrograms (Figure 6.3(b), (e), (h), (k), and (n)), it can be seen that the characteristics of the formant structure differ with respect to intelligibility loss. The dynamic formant characteristics in the consonant-vowel and the vowel-consonant transition regions are distorted by misarticulation. The spectral characteristics also deviate in the steady region of vowels for the case of IL-1, IL-2, and IL-3 due to the hypernasality effect. The Gaussian posteriorgram representation of the speech signals for target sentence stimulus O1 uttered by a normal speaker and CLP speakers with intelligibility level (IL)-0, IL-1, IL-2, and IL-3 are shown in Figure 6.3(c), (f), (i), (l), and (o), respectively. The sentence-specific GMM (λ_1) for sentence O1 is used to generate the GPs. Dark and light pixels denote the highest and lowest posterior probabilities, respectively. From the posteriorgram of the normal speech (Figure 6.3(b)), we see that phonetically similar speech segments are modeled by similar Gaussians. For example, we see that the final /u/ of the sentence is probably represented by Gaussian 16, while Gaussian 20 models the /a/ phonemes. It is expected that for the properly articulated version of

sentence O1, whose intelligibility level is near to normal, the GP representation should be similar to that of the normal speech's GP representation (Figure 6.3(f)). However, as the intelligibility degrades, this representation differs increasingly from normal. For example, deviations in the GPs for IL-1 are observed in Figure 6.3(i). In this case, the consonant /g/ is nasalized (/g^{n}/), the characteristics of the initial /k/ deviate, and the final /p/ has weak articulation (/p̚/) (Figure 6.3(h)). Figure 6.3(j) shows CLP speech utterance where glottal stops and pharyngeal stops compensate most of the pressure consonants, and GP deviates from the normal. Also, Figure 6.3(m) shows a CLP speech utterance in which all of the obstruents are replaced by nasal consonants, which is a voiced sound (the intelligibility level is 3); the GP representations now deviate completely from normal. In this case, the hypernasality is severe, and thus, none of the vowels present in the sentence correspond to the Gaussians that modeled the vowels in the normal case (Figure 6.3(c)). Therefore, to quantify the extent of intelligibility, we require a distance-based measure to capture this deviation properly.

6.3.3 DTW-based intelligibility assessment

This section provides a detailed description to compute the intelligibility scores based on the DTW technique. First, we discuss the feature extraction procedure and mapping of the raw features to the GPs. Later, DTW-based matching is done to compute the intelligibility measure.

DTW-based temporal matching

DTW is a method to estimate the optimal match between two feature sequences by using dynamic programing [151]. Let $\boldsymbol{F}_N = (\boldsymbol{f}_{n1}, \boldsymbol{f}_{n1}, \ldots, \boldsymbol{f}_{nK})$ and $\boldsymbol{F}_C = (\boldsymbol{f}_{c1}, \boldsymbol{f}_{c1}, \ldots, \boldsymbol{f}_{cM})$ represent the feature sequences of normal and CLP speech, respectively. Where, K and M corresponds to the number of frames of normal and CLP speech, respectively. The DTW distance matrix $D_{K\times M}$ is computed using the following equation,

$$D_{K\times M}(i,j) = d(\boldsymbol{f}_{n_i}, \boldsymbol{f}_{c_j}), \tag{6.3}$$

where d corresponds to any dissimilarity measure between normal speech feature vector $\boldsymbol{f}_{n_i}$ and CLP speech feature vector $\boldsymbol{f}_{c_j}$. The best path in the distance matrix ($D_{K\times M}$) is searched starting from $(1,1)$ and ending at (K,M) using the dynamic programming method, which provides minimal accumulated distance.

In GP, the sum of all the components of a posterior vector is 1. Therefore, distance metrics that are defined for probability distributions can be adopted in this context. If X and Y are the two Gaussian

posterior vectors, then the common distance metrics that can be used are,

$$ED(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{6.4}$$

$$D_L(X,Y) = -log(X.Y), \tag{6.5}$$

where N is the dimension of the feature vector. Equation (6.4) represents the Euclidean distance between two vectors, which does not require the vectors to be probability distributions. In Equation (6.5), the dot product provides the probability of drawing the posterior vectors from the same underlying distribution [149].

Computation of intelligibility score

As mentioned in Chapter 3, 10 different sentence-level stimuli are used for the perceptual evaluation. Therefore, training templates are required for each sentence stimulus. Let us consider the training-template generation of sentence stimulus O1 (see Table 3.2 of Chapter 3). To obtain a GP representation of this sentence stimulus, we developed a speaker-independent GMM (λ_1) using the M2DDCT features. Features are extracted from approximately 120 different utterances of sentence O1 recorded from 42 normal children. All the feature dimensions are normalized to have zero mean and unit variance. Once the sentence-specific GMM is built, the features extracted from the test normal/CLP utterances are mapped into the GPs with the help of Equation 6.2. For each set, 10 normal templates are constructed from the GP representation of the utterances, which are properly articulated and perceptually verified by SLPs. In the present work, training templates are collected from five male and five female speakers, and one utterance from each speaker. Then, for each CLP sentence, its GP representation is constructed from the respective sentence-specific GMM. For example, for each test utterance of target sentence O1, the GP representation is computed from λ_1. The DTW distances are then computed between each normal training template and the respective CLP utterance. Because 10 normal training templates are considered for the matching, there are 10 distance values for each test utterance. The method of rejecting outliers based on the mean and standard deviation is then used to ignore distances that are far from the mean distance. In this case, a confidence level of $\mu \pm 1.96\sigma$ is used to reject outlier distances, where μ and σ are the mean and standard deviation of the distances, respectively. Then, the mean of remaining distances is taken as the representative intelligibility score for that particular sentence. The same approach is applied for each sentence stimulus in the database to compute the sentence-level intelligibility scores. The above procedure for computing the intelligibility

score for each tested CLP utterance is applied for GP_{MFCC}, M2DDCT, and MFCCs features.

All the DTW distance metrics mentioned in Section 6.3.3 are analyzed experimentally, and the distance metric that provided the best performance is considered herein. Because our database is relatively small, we use leave-one-speaker-out cross-validation. This is done for sentence O1 only, and the distance measure that provides the best results for O1 is then used for the rest of the sentences. For each fold of leave-one-speaker-out cross-validation, all the speech data except those from one CLP children are used to build linear regression models to ensure no speaker overlap between training and testing. A linear regression model is built by considering the perceptual ratings as the independent variables and the DTW distances as the dependent variables at each fold. The root means square errors of the estimated and true DTW distances are computed for the tested CLP speech. This process is repeated for all the CLP children. Because the database includes 42 CLP children, there are 42 folds, and corresponding 42 root mean square errors, of which we compute the mean. The distance measure that gives the minimum root mean square error is considered for the remaining analysis.

6.3.4 SSM-based intelligibility assessment

The motivation for using SSM to derive measures for intelligibility is that it gives a speaker-independent representation of underlying acoustic-phonetics units of an utterance. As depicted in Figure 6.2, after mapping the raw features to Gaussian posterior vectors, they are transformed into an alternate representation called SSM. Then, the SSM of test speech is compared with the template SSM generated from the normal speech. The SSM representation of the normal speech utterance is considered as the training template shown in Figure 6.2, and several such templates are considered. The deviation between the template and test SSM is computed using a distance measure, and several such distance value will be computed. Then, the averaged value of distances is studied as the correlate of intelligibility.

Self-similarity matrix based comparison:

SSM-based comparison is a template matching technique, which was proposed for the word discovery problem [152]. The SSM (Φ_F) of a given frame sequence $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n]$ is a square symmetric matrix, which is computed as,

$$\Phi_F(i,j) = d(\boldsymbol{f}_i, \boldsymbol{f}_j), \tag{6.6}$$

or,

$$\Phi_F = \begin{bmatrix} d(\boldsymbol{f}_1, \boldsymbol{f}_1) & d(\boldsymbol{f}_1, \boldsymbol{f}_2) & d(\boldsymbol{f}_1, \boldsymbol{f}_3) & \dots & d(\boldsymbol{f}_1, \boldsymbol{f}_n) \\ d(\boldsymbol{f}_2, \boldsymbol{f}_1) & d(\boldsymbol{f}_2, \boldsymbol{f}_2) & d(\boldsymbol{f}_2, \boldsymbol{f}_3) & \dots & d(\boldsymbol{f}_2, \boldsymbol{f}_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(\boldsymbol{f}_n, \boldsymbol{f}_1) & d(\boldsymbol{f}_n, \boldsymbol{f}_1) & d(\boldsymbol{f}_n, \boldsymbol{f}_1) & \dots & d(\boldsymbol{f}_n, \boldsymbol{f}_n), \end{bmatrix}$$

where d is any similarity or dissimilarity metric between two frames f_i and f_j [153]. It is obvious that the diagonal elements of the SSMs are zero, i.e., $\Phi_X(i,i) = 0$, if euclidean distance measure is used. Generally, for MFCC feature euclidean distance based dissimilarity measure is used, while for GP based representations $-log(\boldsymbol{g}_1.\boldsymbol{q}_2)$ is used. To get rid of zeros while computing the log, a discounting based smoothing strategy is applied as discussed in [149, 152]. The structure of the SSM of an utterance is completely dependent on its underlying sequence of acoustic-phonetic units. Such information can be exploited for recognizing whether the compared acoustic speech segments share the same lexical identity. The structure of SSM gives robust representation of speech against different speech variabilities, such as, noise and speaker [152,153]. Thus, the SSM representation itself provides a speaker-independent representation of the speech signal, which can not be obtained using the DTW-based method. Moreover, unlike DTW, the SSM based comparison method can encode high information variability among compared patterns by capturing the interaction between all parts of the utterance [152,153].

The two-dimensional pattern of SSM generated by computing distances among mutual parts of a feature sequence is unique for a particular sentence. The consistent similarities of SSMs for sentence O1 across two normal children (female and male) are shown in Figure 6.4 (a) and (b), respectively. A distinct resemblance of shape patterns and local edges of both the SSMs are observed, which are totally dependent on the composition of the acoustic units in sentence O1. Thus, the SSM represents the underlying acoustic phonetic segments of the sentence, not the speaker-related variabilities. However, any distortion in the acoustic-phonetic characteristics of the sound units due to deviations in the articulatory precision or maladaptive compensation will lead to change in SSM's structure. In this work, the deformation of SSM in CLP speech is intended to capture by comparing SSM of the normal speech. The information of dissimilarity may reveal the degree of intelligibility loss in CLP speech. Figure 6.4 (a-b), (c), (d), (e), and (f) show the SSMs of normal and four CLP speech with IL-0, IL-1, IL-2, and IL-3, respectively. In this case, GPs are used to generate the SSMs of sentence O1, where inner product based similarity measure is used to compute the SSM for better visualization.

It can be seen from the figure that structure of the SSMs of CLP speech utterance deviates more as the intelligibility degrades, due to the deviations in the underlying acoustic-phonetic structure of the utterance. To capture the dissimilarity among reference SSM and test SSM, initially, warping

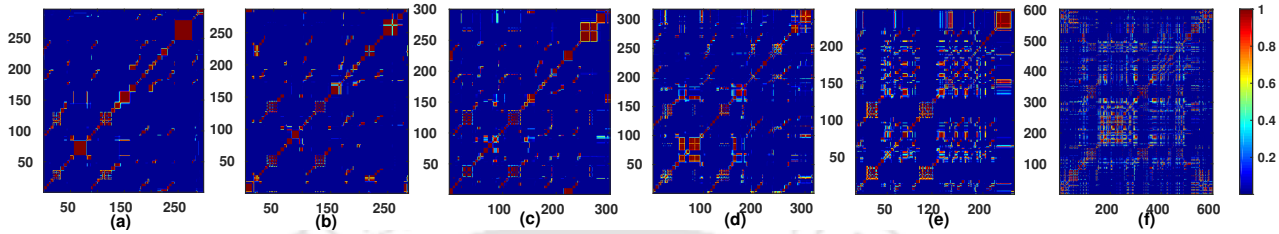

Figure 6.4: Illustration of SSM structure's deviation with respect to the loss of intelligibility for the utterances with different levels of intelligibility. SSMs structures of the speech signals for target sentence stimulus O1 are considered. (a) normal 1 (female), (b) normal 2 (male), (c) CLP IL-0, (d) CLP IL-1, (e) CLP IL-2, and (f) CLP IL-3.

path ($W(P^*)$) between the two frame sequences ($\boldsymbol{F}_R$ and $\boldsymbol{F}_T$) is computed using DTW method. This $W(P^*)$ is used to warp $\boldsymbol{F}_R$ and $\boldsymbol{F}_T$ to $\boldsymbol{F}_R^{'}$ and $\boldsymbol{F}_T^{'}$ to obtain SSMs of same sizes. SSIM index based measure is applied to compare the two SSMs, considering them as the grey scale images.

Apart from the SSM computation from GPs, it is also computed from the raw features (MFCCs and M2DDCT) directly, and raw feature based systems are also explored for the intelligibility assessment. This is done to show the importance of SSM-based method in deriving the reliable intelligibility scores without using the GP representation, and also to show its effectiveness over the DTW-based system which is developed using the raw features.

Computation of intelligibility score

For each sentence level stimulus, we have considered 10 properly articulated reference utterances as mentioned in Section 6.3.3. Thus, 10 SSMs comprise the reference templates for each stimulus. Let us consider, $[X_1^j, X_2^j, .., X_r^j, ..., X_{10}^j]$, where $1 \leq j \leq 10$, be the reference SSMs of the j^{th} stimulus for feature F. For a test SSM (Y^j) of a normal or CLP speech utterance corresponding to j^{th} stimulus, SSIM scores are computed with respect to 10 reference SSMs, i.e. $\{X_r^j\}_{r=1}^{10}$. Therefore, 10 scores ($\{s_r^j\}_{r=1}^{10}$) are generated for that test utterance of j^{th} stimulus. Average value of 10 scores is considered as the estimated intelligibility score (I^j) of corresponding utterance. Thus, the representative intelligibility score (I^j) of the test utterance for j^{th} target stimulus is computed as follows,

$$I^j = \frac{1}{10} \sum_{r=1}^{10} s_r^j. \tag{6.7}$$

Table 6.1: Mean (μ) values of DTW-based intelligibility scores each level for a specific sentence-level stimulus O1.

Groups	DTW distance (normalized)			
	μ			
	MFCC	M2DDCT	GP_{MFCC}	GP_{M2DDCT}
IL0	0.12	0.09	0.09	0.08
IL1	0.15	0.16	0.17	0.13
IL2	0.22	0.18	0.29	0.29
IL3	0.35	0.38	0.43	0.44

Table 6.2: Mean (μ) values of SSM-based intelligibility scores each level for a specific sentence-level stimulus O1.

Groups	SSIM score (normalized)			
	μ			
	MFCC	M2DDCT	GP_{MFCC}	GP_{M2DDCT}
IL0	0.79	0.81	0.77	0.70
IL1	0.65	0.74	0.61	0.67
IL2	0.34	0.52	0.41	0.44
IL3	0.23	0.42	0.27	0.19

6.4 Performance evaluation

In this section, performance evaluation of the proposed intelligibility measures is carried out. Initially, we have analyzed how predicted intelligibility scores are distributed with respect to perceptual ratings. The qualitative discrimination among the different intelligibility groups for all DTW-based and SSM-based measures are shown using box plots in Figure 6.5. For each measure, the predicted scores for all the groups are normalized by min-max normalization to map them between 0 and 1 to maintain the uniformity. From the figure, it can be observed that as the intelligibility degrades from 0 to 3, the DTW-based scores (Figure 6.5 (a)-(d)) are increasing, whereas the SSM-based scores (Figure 6.5 (e)-(h)) are decreasing for all the features used herein. This increasing and decreasing trends in the DTW- and SSM-based measures are due to the usages of dissimilarity and similarity measures, respectively. We notice from the figure that the discrimination between the intelligibility groups for DTW-based measures using MFCC and M2DDCT features is very poor. However, the inter-group discrimination increases in case of DTW-based measures using GP features. The SSM-based measures using MFCC and M2DDCT features provide better discrimination among the groups as compared to the their DTW-based counterparts. From the figure, a consistent discrimination among the groups in cases of SSM-based measure using GP_{MFCC} and GP_{M2DDCT} features and for DTW-

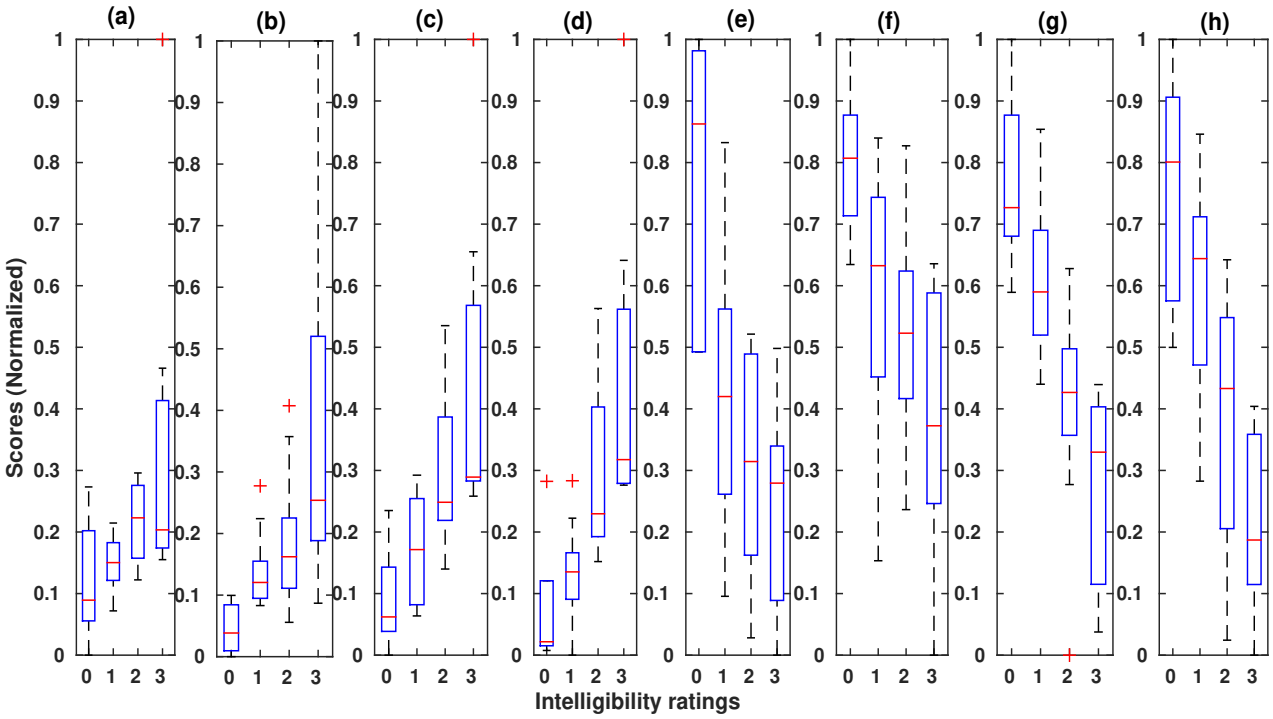

Figure 6.5: Box plots of DTW-based (a-d) and SSM-based (e-h) measures for the explored features. (a, b) and DTW based measure (c, d) for different level of intelligibility in case of sentence O1 for MFCCs and GP based features, respectively.

based measure using GP_{M2DDCT} can be noticed. A similar observation in terms of discrimination of mean values using statistical quantitative analysis is shown in Table 6.1 and Table 6.2 for DTW- and SSM-based measures, respectively. From Table 6.1 it can be noted that the mean values are almost equally spaced for the DTW-based measures using GP_{MFCC} and GP_{M2DDCT} features, as compared to MFCC and M2DDCT features.

The performance of proposed objective measures is evaluated using the correlation analysis with respect to the perceptual intelligibility ratings. For this purpose, Spearman's rank correlation coefficient (ρ) between the SLPs' perceptual intelligibility ratings and the objective intelligibility scores is computed as earlier chapters. The statistical significance of the correlation difference among the measures is studied using William's pairwise significance test [90]. Because the number of Gaussians needed to build GMM for each sentence to compute GP is not known in advance, we must estimate it experimentally for each sentence.

Table 6.3: Mean (μ) and standard deviation (σ) of the absolute correlation values between the objective measure and perceived intelligibility ratings given by the SLPs.

Features	DTW-based (D)		SSM-based (S)	
	ρ ($\mu \pm \sigma$)	p-value	ρ ($\mu \pm \sigma$)	p-value
MFCC	0.51 ± 0.071	<0.001	0.67 ± 0.050	<0.001
M2DDCT	0.54 ± 0.069	<0.001	0.64 ± 0.052	<0.001
GP_{MFCC}	0.71 ± 0.041	<0.001	0.74 ± 0.031	<0.001
GP_{M2DDCT}	0.73 ± 0.040	<0.001	0.73 ± 0.044	<0.001

6.4.1 Performance evaluation of DTW-based measure

The average of 10 individual sentence-level correlations is considered for the performance evaluation. Table 6.3 gives the mean and standard deviation of correlation values between the subjective intelligibility ratings and the DTW-based estimated scores for all the features. It can be seen from Table 6.3 that a maximum average correlation of 0.73 is achieved with a confidence value of $p < 0.001$ using GP_{M2DDCT}; however, M2DDCT feature provides a correlation value of 0.54. Thus, an improvement of 0.19 is obtained using the GP-based representation. The lowest performance among the features is observed for MFCCs, with a correlation value of 0.51. However, its GP counterpart provide significantly higher correlation value, i.e. 0.71. It is clearly observed that raw feature based measures give very less correlation values as compared to their GP counterparts. This improvement over the raw features justify the importance of GP representation in the intelligibility assessment. Also, as compared to GP_{MFCC}, GP_{M2DDCT} feature gives an improvement of 0.02 in correlation. This improvement of correlation over GP_{MFCC} feature clearly reveals the importance of the joint spectro-temporal features. However, a statistical test is conducted to conclude whether the improvement is statistically significant or not. Another observation can be made is that standard deviation of the correlation values is much high for raw acoustic features compared to the GP-based features. This high standard deviation signifies the inconsistency in predicting intelligibility score for individual sentence stimulus.

6.4.2 Performance evaluation of SSM-based measure

The correlation between SSM-based intelligibility scores and subjective intelligibility ratings for all the features are shown in Table 6.3. It can be observed that SSM-based measure using MFCC and M2DDCT features provide correlation values of 0.67 and 0.64, respectively. An improvement of 0.07 correlation is observed for GP_{MFCC}, as compared to MFCC feature. Compared to M2DDCT feature, its GP version shows 0.09 improvement in correlation. Highest correlation is observed for

Table 6.4: p-value of Williams significance test between pairs of intelligibility measures. (p-value of a given pair of measures is computed whose absolute Spareman rank correlation with perceptual rating is higher than that of the other in the pair. p-value less than 0.05 is considered as the statistically significant). Here, * represents a p-value greater than 0.05, and † represents a p-value less than 0.05

–	*	*	†	†	†	†	†	$S_{GP_{MFCC}}$
–	–	*	*	†	†	†	†	$S_{GP_{M2DDCT}}$
–	–	–	*	†	†	†	†	$D_{GP_{M2DDCT}}$
–	–	–	–	†	†	†	†	$D_{GP_{MFCC}}$
–	–	–	–	–	†	†	†	S_{MFCC}
–	–	–	–	–	–	†	†	S_{M2DDCT}
–	–	–	–	–	–	–	*	D_{M2DDCT}
–	–	–	–	–	–	–	–	D_{MFCC}
$S_{GP_{MFCC}}$	$S_{GP_{M2DDCT}}$	$D_{GP_{M2DDCT}}$	$D_{GP_{MFCC}}$	S_{MFCC}	S_{M2DDCT}	D_{M2DDCT}	D_{MFCC}	

GP_{MFCC} feature (0.74), while lowest correlation is observed for M2DDCT feature (0.64). Thus, in this approach as well GP-based representation provide better correlation than their raw version. Here, another observation can be made is that MFCC feature outperforms the M2DDCT feature, for both the raw acoustic and GP-based features. However, the correlation difference is not very significant.

6.4.3 Comparison between DTW- and SSM-based approaches

Initially, the performance of DTW- and SSM-based measures is discussed individually. Now, a comparison of both the approaches are made in this subsection. From Table 6.3 it can be clearly observed that relatively better performance is achieved for SSM-based measures as compared to DTW-based measures, except the GP_{M2DDCT} feature where similar performance is observed for both the methods. Very less correlation values are noted in case of DTW-based measures using MFCCs and M2DDCT features. Hence, DTW-based measure using these features is not reliable, which may be due to the speaker variabilities embedded in those raw features. One way to overcome the problem is the usage of GP-based features, and an improvement using these features are also observed for DTW-based system. To derive the GP-based features, unsupervised statistical modeling of data is required. However, SSM provides a robust unique representation of the underlying phonetic content of the utterance. Also, it

compensate the inherent speech variabilities embedded in MFCC or M2DDCT feature. Since SSM structure is robust against speaker variabilities, it captures the distortions related to acoustic-phonetic segments due to intelligibility degradation and improves the performance. Due to these characteristics, significant improvement is achieved using only MFCCs and M2DDCT features in SSM-based system as compared to DTW-based system. Since no statistical modeling of the acoustic is required; thus, SSM-based measure using raw acoustic features may be helpful in low-resource applications. An improvement of 0.16 is observed for MFCC-based SSM measure as compared to MFCC-based DTW measure, and 0.10 correlation improvement is observed for M2DDCT based SSM measure as compared to M2DDCT based DTW measure. These improvement show the importance of SSM-based representation in the intelligibility assessment. GP-based SSM measure further improves the correlation by adding more robustness against speaker variabilities in a statistical sense, while retaining the phonetic information. Results show higher correlation in case of GP-based SSM method with a correlation coefficient of 0.74 than DTW-based method. The advantage of SSM based approaches is that it captures the dissimilarity among mutual parts of the feature sequence which provides a unique pattern in SSMs for underlying acoustic-phonetic composition. Unlike DTW based approach, SSM based comparison method can encode high information variability among compared patterns by capturing the interaction between all parts of the utterances [152, 153].

To verify the improvement of proposed intelligibility measures over the MFCC and M2DDCT baseline, we perform the Williams significance test [90, 91] for each pair of measure. Table 6.4 lists the outcomes of these test. In the table, each p-value inside a cell (i, j) indicates whether measure i (named in the rightmost column of the table) is correlated significantly higher with the perceptual ratings than measure j (named in the bottom row of the table). We have shown in each cell whether the correlation improvement if statistically significant at $p<0.05$. The results show that increased in correlation values for SSM-based measures using MFCCs (S_{MFCC}), M2DDCT (S_{M2DDCT}), GP_{MFCC} ($S_{GP_{MFCC}}$), and GP_{M2DDCT} ($S_{GP_{M2DDCT}}$) with respect to DTW-based measures using MFCCs (D_{MFCC}), M2DDCT (D_{M2DDCT}). Increased correlation values for $D_{GP_{MFCC}}$, and $D_{GP_{M2DDCT}}$ compared to D_{MFCC}, M2DDCT (D_{M2DDCT}) measures are statistically significant. However, the increased correlation value for $D_{GP_{M2DDCT}}$ compared to $D_{GP_{MFCC}}$ is not statistically significant at $p<0.05$. Also, correlation improvement in $S_{GP_{MFCC}}$ over $S_{GP_{M2DDCT}}$ is not statistically significant.

6.5 Summary and conclusions

In this chapter, GP-based speech representation with DTW distance and SSM comparison are explored to assess the intelligibility of child CLP speech objectively. The estimation of sentence-level intelligibility scores and the comparison of those scores with perceptual ratings in terms of correlation analysis are the primary objectives of this work. Motivated by the perceptual importance of transition regions, especially in the case of obstruents, and the significance of joint spectro-temporal features in characterizing these regions, low-order 2D-DCT coefficients are extracted from overlapping patches of TFR of speech. In DTW-based method, the deviations of reference speech's GP and test speech's GP is quantified using DTW accumulated distance, and it is studied as the acoustic correlate of intelligibility. However, in SSM-based method, another intermediate representation of speech, i.e., SSM is derived from the GPs, which is expected to provide another level of speaker independence. Then, reference speech's SSM and test speech's SSM is compared with the help of a similarity measure, and the resultant score is studied as an intelligibility score. Results show that for both the approaches, GP-based representation outperforms the raw acoustic features. SSM-based measures using MFCC and M2DDCT features give a significantly high correlation as compared to the DTW-based measure using these features. The Williams significance test showed that the increase correlation is statistically significant for the GP based measures when they are compared with the MFCC and M2DDCT.

The works presented till now in the thesis are focused towards the estimation of intelligibility at utterance-level. The proposed intelligibility measures are highly correlated with the perceptual ratings. However, the scores provided by the measures are not in the form to be easily interpreted by the SLPs. Moreover, it is also important to define the range in the scores to be considered as intelligible speech. In the next chapter, these issues are addressed to make the intelligibility measures applicable to clinical applications, such as therapy and disease monitoring.

7

System for Clinical Applications

Contents

Objective

The objective of this chapter is to develop a system for the clinical application to assess intelligibility. While doing this, the present chapter defines the range of intelligibility, and a visual representation based on the spider plot is proposed for graphing the intelligibility scores. The sentence-level intelligibility scores are mapped to [0, 1] range, and mean of the normal speaker's scores is considered as the range of intelligibility. In the earlier chapters, measures for the sentence-level intelligibility are proposed; however, the estimation of intelligibility score at speaker-level is not attempted. The space enclosed by a spider plot is considered as the representative of a speaker's intelligibility space, and corresponding area under the plot is studied as the subject-specific intelligibility score.

7.1 Introduction

In the earlier chapters, we have proposed several measures of CLP speech intelligibility by characterizing the acoustic deviations around certain events and comparison-based approaches. The measures are acoustic correlates of sentence-level intelligibility ratings and provided significant correlation with the perceptual ratings. However, we have not defined the range for intelligible speech for the proposed measures, i.e., up to which objective score of one particular test utterance will be considered as intelligible. This quantification of intelligibility is important for the computer-based assessment algorithm, as it will provide a basis for the SLP upon which intelligibility assessment can be done. Also, it is important to represent the estimated intelligibility scores visually for each speaker with respect to the normal range. Since ten sentence-level stimuli are used to evaluate the intelligibility, thus, ten sentence-level scores are provided to evaluate the overall intelligibility of the speaker. Therefore, a visual representation based on the spider plot may be helpful while doing the clinical assessment using the proposed measures. Spider plot is a very effective tool for describing the multivariate data. It is a circular graphical representation and has a series of rays which are projecting from the center point. Each ray represents one of the variables; the length of the ray represents the value of the variable [154].

The subject-specific intelligibility, which provides the global view of intelligibility for a particular CLP speaker, was not studied in the earlier chapters. Generally, SLPs evaluate the intelligibility of one CLP speaker by perceptually analyzing a set of word/sentence stimuli. The subject-specific intelligibility is important to evaluate the overall articulation capability of the speaker. To estimate the subject-specific intelligibility the sentence-level scores proposed in the earlier chapters are ex-

plored. From Chapter 4, it is found that M2DDCT-based articulation score (ψ_{M2DDCT}) combined with MFCC-based hypernasality score (η_{MFCC}) provides highest correlation with perceptual ratings. This composite intelligibility metric is represented by C_i in this chapter. In Chapter 5, the combination of *+g-landmark* and *-g-landmark* based measure using M2DDCT feature ($\text{M2DDCT}_{+\text{g}}$ + $\text{M2DDCT}_{-\text{g}}$) provides best correlation, and this measure is represented by Γ_i in this chapter. Finally, SSM-based measure derived using the GP-based speech representation gives the best correlation in Chapter 6, and this measure is represented by S_i. Effectiveness of these measures to provide the subject-specific intelligibility is discussed in this chapter. The ground-truth perceptual score for each CLP individual is obtained from Section 3.3 of Chapter 3, and it can be computed as,

$$G_i = A_i/30, \tag{7.1}$$

where A_i represents the summation of all the ten sentence-level perceptual intelligibility ratings, and G_i is the global intelligibility for that CLP individual. The division by 30 (maximum intelligibility rating (3) $\times$ number of sentence stimuli (10)), which is the maximum attainable rating for a CLP speaker, mapped the ratings to [0, 1].

The rest of the chapter is organized as follows: In Section 7.2, we define the range for intelligible speech for the proposed objective measures and describe the procedure to derive a visual representation for intelligibility assessment. In Section 7.3, a detailed description to estimate the subject-specific intelligibility is provided. The experimental results are presented in Section 7.3. Finally, the chapter is concluded in Section 7.4 by summarizing the presented work.

7.2 Defining the range of intelligibility

Defining the range of intelligibility is important for the practical implementation of the system. This can be done by passing the speech utterances from normal speakers through the proposed algorithm and compute the intelligibility score. From these normal speakers scores, we can derive the range of intelligible speech. In this experiment, we have considered 12 normal children's data (6 females and 6 males) of age group 6-12. To demonstrate this, measure which gives the best performance in each chapter are only considered as discussed above. For each measure, sentence-level scores are mapped to [0, 1]. This normalization will help to derive the intelligibility measure such that the lower value near 0 represents the highly intelligible and near 1 represents the severely degraded intelligibility. Since we

Table 7.1: Perceptual intelligibility ratings for five CLP speakers

CLP Speaker	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
1	0	0	0	0	0	1	0	1	0	0
2	1	2	1	1	1	2	2	2	2	1
3	1	2	2	1	1	2	1	2	2	2
4	3	3	2	3	3	3	3	3	2	2
5	3	3	3	3	3	3	3	3	3	3

have used ten sentence-level stimuli, for each sentence, a range is defined for the intelligible speech.

7.2.1 Visualization for clinical applications

The estimated scores of all the sentence stimuli for a speaker can be used to derive a visual representation which will be helpful for the clinical purpose. A visualization tool based on the spider plot by integrating all the derived intelligibility scores in a single representation is proposed. Such plot allows visual comparisons of the scores differences of normal and CLP speakers. Figure 7.1 shows the visual tool to represent all the sentence-level intelligibility scores in a single plot. We have shown the case study of five CLP speakers, and estimated intelligibility scores of each sentence are plotted and form a polygon for each speaker. The perceptual intelligibility ratings for the five CLP speakers are listed in Table 7.1. The average of normal speakers scores, corresponding to the green polygon in Figure 7.1 represents the range of intelligible speech. If the sentence-level scores are coinciding or inside the green polygon, it means that those sentences are highly intelligible. For a polygon which represents the 10 sentence-level intelligibility score, if any of the corners does not match with the green polygon, it suggests that corners corresponding to those sentences, intelligibility level is deviated that of normal. The more it deviates the more unintelligible those sentences are. Let us consider CLP speaker-1 (C1) (Figure 7.1(a)), the estimated scores of all the sentences except O6 and O8 are inside or near the green polygon. It can be seen from Table 7.1 that for C1, O6 and O8 sentences have intelligibility ratings 1, while 0 for other sentences. If we consider CLP speaker-3 (C3), the estimated score for O3 is deviated from the normal range; however, around middle of the ray. This scores corresponds to the intelligibility rating 2. For CLP speaker-5 (C5), all the sentences perceptual ratings is 3; also from the spider plot it can be seen that the estimated scores deviated too far from the normal range. Similar observation can be made for the CLP speaker-2 and CLP speaker-4. Also, as the intelligibility scores are deviated from the normal range, the corresponding area under the polygon increases as shown

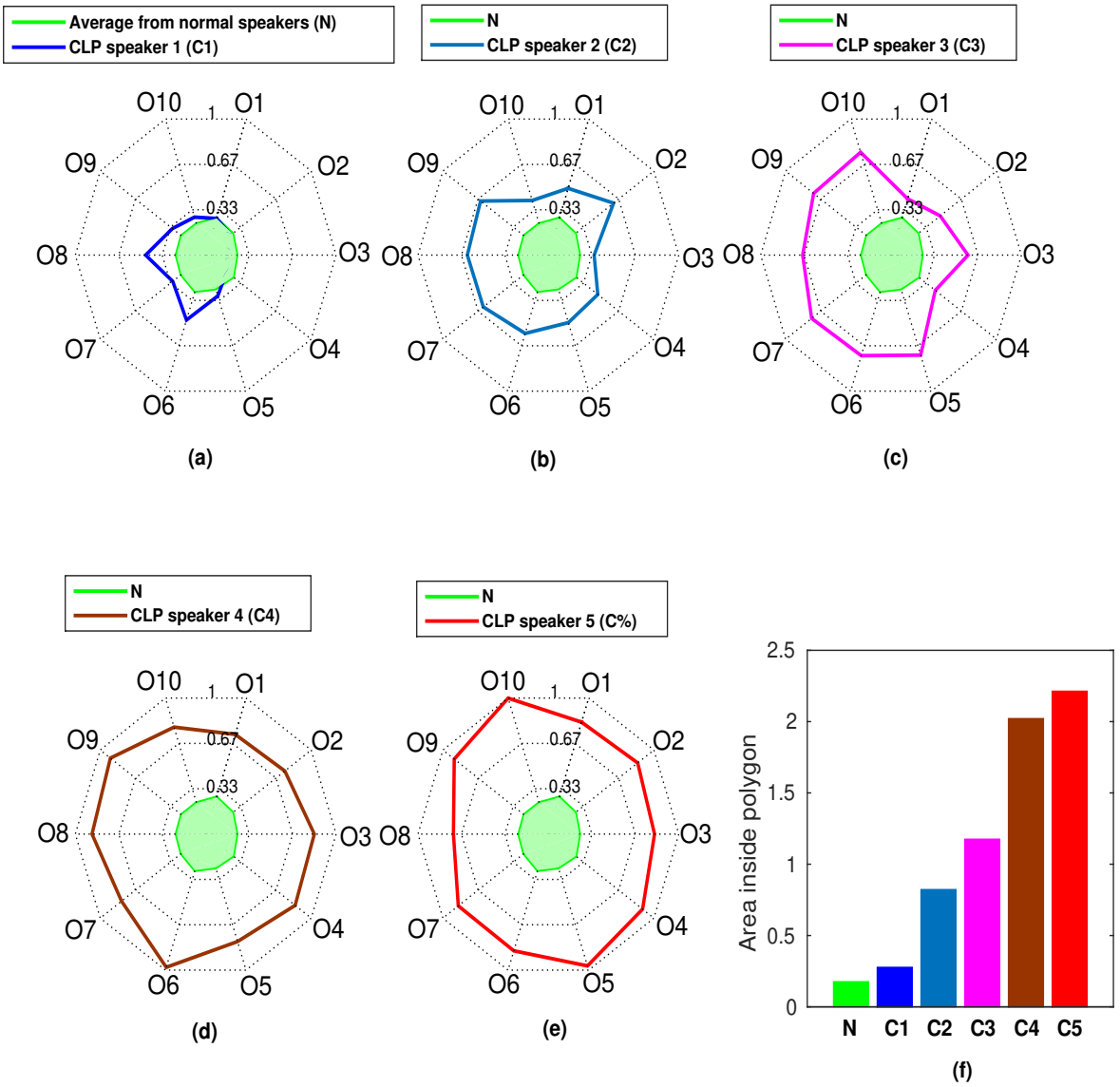

Figure 7.1: Visualization of sentence-level intelligibility scores in the spider plots ((a) - (e)) and the area under the polygons (f) for 5 CLP speakers and average scores from normal speakers.

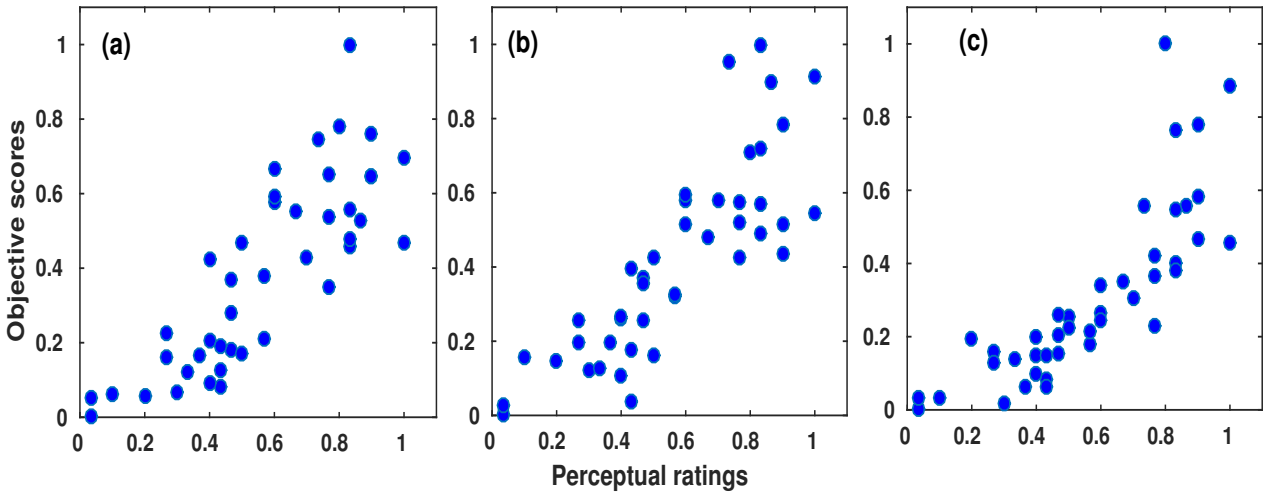

Figure 7.2: Scatter plots of the estimated intelligibility scores for different measures with respect to perceptual scores. (a) C_i, (b) Γ_i, and (c) S_i

in the bar plots. Thus, this representation with the bar diagram as a whole provides SLPs with an informative impression of intelligibility of the speaker. Moreover, each sentence stimulus loaded with different category of obstruents. For example, O6 and O8 are mainly loaded with /t̪/ and /ʧ/ sounds, respectively. Since estimated scores are deviated from the normal range there may be a chance that articulation of these two sounds have some problem. This tool may also be used for self assessment and monitoring the disease progression over the therapy. During the therapy, the SLP may use the graphical representation to check how much intelligibility is deviated from the normal one. And also, they can interpret whether the intelligibility is improving from the last therapy session by looking at the graphical representation.

7.3 Estimation of subject-specific intelligibility scores

As it can be seen from the spider plot based visual representation that area enclosed by the polygon for normal speech sentences is smallest. As the intelligibility degrades the area under the polygon also increases (Figure 7.1 (f)). Each polygon of the spider plot represents intelligibility space of the speaker. Thus, the enclosed area is expected to represent the overall intelligibility of the CLP patient. Figure 7.2 (a), (b), and (c) show the scatter plots of area under the polygon for all the 42 speakers corresponding to perceptual ratings for C_i, Γ_i, and S_i, respectively. From these plots it can be clearly seen that as the intelligibility reduces from 0 to 1 as the area increases.

Table 7.2: Spearman's rank correlation coefficients between the speaker-specific intelligibility measures and perceptual ratings

	Correlation analysis		Regression analysis		
Measures	ρ	p-value	R^2	F-score	p-value
C_i	0.83	<0.001	0.69	83.65	< 0.001
Γ_i	0.84	<0.001	0.70	85.64	< 0.001
S_i	0.88	<0.001	0.68	79.67	0.053

7.4 Experimental results and discussion

We have investigated the strength of each measure to describe the overall intelligibility loss. To estimate the intelligibility scores linear regression model is applied, and the evaluation is done using leave-one-speaker-out cross-validation as described in the earlier chapter. Since 42 speakers are there, 42 linear regression models are built for all the folds. Estimated intelligibility scores for all the folds are correlated with the perceptual ratings. We have also studied the averaged R^2, F-value, and p-value of 42 linear regression models for the statistical analysis. The results of correlation and regression analysis are listed in Table 7.2. From the table, it can be noted that SSM-based measure gives the highest correlation values as compared to composite measure based and landmark-based measures. However, the mean R^2 and F-value for the SSM-based measure is less as compared to the other two measures.

7.5 Summary and conclusions

In this chapter, we define the range of intelligible speech for proposed algorithms. A spider plot-based visual representation is proposed by integrating all the sentence-level intelligibility scores. Each polygon of the spider plot represents intelligibility space of the speaker. This visual tool makes the results of the objective intelligibility measure more interpretable and helpful for clinical purpose. The area inside the polygon is found to be correlated with the subject-specific intelligibility scores. From the experimental results, it is found that SSM-based measure provides the best performance in describing the subjective-specific intelligibility, as compared to the composite based and landmark-based measures. Next chapter provides the summary of works presented in this thesis and a few directions of future research.

8

Summary and Conclusions

Contents

Overview

This chapter provides the summary and conclusions of the works presented in this thesis towards deriving objective measure for CLP speech intelligibility. Based on the contributions and different investigations, we also discuss a few possible directions for future research.

8.1 Summary of the work

Present thesis focuses on the development of objective measures to evaluate the CLP speech intelligibility. The primary objective of the thesis is to provide intelligibility scores at sentence-level, and later these sentence-level scores are combined to predict the subject-specific intelligibility. To achieve these objectives, the preliminary requirement is to develop a speech database of CLP and controlled normal speakers, and it is developed with the collaboration of AIISH, Mysuru, India. Initially, a subjective analysis is conducted to investigate the impact of different speech disorders related to CLP on intelligibility. In the thesis, three attempts have been made to objectively quantify the intelligibility loss. In the first two attempts, measures based on combination of acoustic cues related to articulation and nasality are proposed. The proposal of these measures are motivated by the fact that intelligibility of CLP speech is primarily degraded due to the articulation errors and hypernasality. Further, the comparison of posterior sequences based on DTW and SSM is proposed to derive the intelligibility measures. Finally, a system is developed to represent the intelligibility scores visually and obtain the subject-specific intelligibility scores. The current study may help to define a set of acoustic measures correlated with intelligibility and that can be used as the biomarker for speech progression during therapy. Unlike ASR based methods, the proposed methods explore only the acoustic information of deviant speech, no linguistics information is explored. The contributions incorporated in this thesis are summarized below.

1. **Database development and subjective analysis of intelligibility:** Initially, a detailed discussion about the development of CLP speech database and the procedure to conduct perceptual evaluation is provided. Before deriving the objective intelligibility measures, it is important to study the impact of different speech disorders, such as hypernasality, articulation error, and voice disorder in reducing the CLP speech intelligibility. Therefore, an experiment is performed to study the relative contributions of these speech disorders on the intelligibility deficits. Multiple

linear regression based analysis shows that PCC has the highest contribution (-0.662) to overall speech intelligibility, while hypernasality has a comparatively low contribution (0.232) on CLP speech intelligibility. However, voice quality has a significantly less contribution (0.049) to the loss of intelligibility.

2. **Intelligibility measure based on the articulation and hypernasality information:** In the first work, a measure of sentence-level intelligibility is proposed by combining the information of articulation deficits and hypernasality. Initially, a measure for articulation deficits using the acoustic features extracted around VOPs and VEPs is proposed. The joint spectro-temporal (JST) based feature from the overlapping patches of time-frequency representation is explored for the better characterization of spectral and temporal modulations in the transition region. Next, an algorithm is proposed to quantify the hypernasality level objectively, and this algorithm is motivated by the working principle of the Nasometer instrument. The objective scores of articulation deficits and hypernasality are used as the feature set to train a regression model. The output of the model is considered as predicted intelligibility score. The Spearman's correlation coefficient based analysis shows the correlation values of $\rho = 0.72$ and $\rho = 0.75$ between the composite measure based scores and perceptual intelligibility ratings for linear regression and support vector regression models, respectively. For all the sentence-level stimuli, JST feature shows improvement in performance over the MFCC feature.

3. **Exploring glottis landmarks for intelligibility assessment:** In this work, the importance of *g–landmarks* in deriving the acoustic correlates of CLP speech intelligibility is studied. An investigation is made to analyze how spectral characteristics are distorted around consonant landmarks due to articulation errors. From the analysis, it is found that *g–landmarks* are more suitable events to extract the features. Two separate sentence-specific GMMs are built for each sentence stimulus using the features extracted around the two *g–landmarks* (+g and -g) from the normal speech. The GMMs derived for each sentence-level stimulus are used to compute the log-likelihood scores of the respective test utterance. The average value of the log-likelihood scores of features extracted from the speech region around *g–landmarks* is calculated. Since two separate GMMs are built for both *+g–landmarks* and *-g–landmarks*, two average values of log-likelihood scores are obtained for each test utterance. Both average scores are studied as the acoustic correlate of CLP speech intelligibility. Derived intelligibility measure

shows a significant correlation with the perceptual ratings. A visual representation based on the two-dimensional plot of mean log-likelihood scores vs. number of detected *g–landmarks* for an utterance is proposed. This visual representation may be useful for the clinical application to get some information about the loss of intelligibility.

4. **Intelligibility assessment based on the comparison of posterior sequences:** The intelligibility measures proposed in this work are based on the comparison of posterior sequences of normal templates and CLP test utterance. Two comparison based frameworks are proposed using DTW and matching of SSMs to quantify the intelligibility. Initially, acoustic features are computed for an utterance and mapped into GPs. The motivation for using GP is that it provides speaker independent acoustic segment representation, and can be derived in a completely unsupervised manner. GP representation of the distorted unintelligible speech of CLP children may be distinctly different from the normal children's speech. In the first approach, DTW is used to compute the deviation of GP of test speech from the normal speaker's template, and DTW distance is studied as a correlate of intelligibility. However, in SSM-based approach, the posterior sequence of a speech signal is transformed into another representation, termed as self-similarity matrix. The SSM of a feature sequence is a square matrix, and it encodes the acoustic-phonetic composition of the underlying speech signal. Deviations in the acoustic characteristics of underlying sound units due to degradation of intelligibility deviate the CLP speech's SSM structure from that of normal speech. The degree of deviations is quantified using the SSIM index, which is considered as the representative of objective intelligibility score. Spearman's rank correlation coefficient between the objective intelligibility scores and the perceptual intelligibility rating is studied. Results show that the SSM-based measure outperforms the DTW-based measure. Also, the Williams pairwise significance test confirms that the increased correlation is statistically significant.

5. **System for clinical application:** Finally, a system is developed for clinical application to assess the intelligibility. While developing it, the present work defines the range of intelligibility, and a visual representation based on spider plot is proposed for graphing the intelligibility scores. Each polygon of the spider plot represents intelligibility space of the speaker, and the area inside the polygon is correlated with the subject-specific intelligibility scores. From the experimental results, it is found that SSM-based measure provides the best performance in

describing subjective-specific intelligibility, as compared to composite-based and landmark-based measures.

8.2 Contributions of the thesis

Following are the contributions of this thesis towards deriving the objective measures of CLP speech intelligibility.

- A sentence-level CLP speech database is developed for intelligibility assessment.
- The relative impact of hypernasality, articulation error, and voice quality on CLP speech intelligibility is analyzed.
- Objective measures for predicting articulation deficits and hypernasality levels, and combine these measures to derive correlates of intelligibility are proposed.
- The importance of joint spectro-temporal features over the conventional MFCCs in intelligibility estimation is studied.
- The distortion of abrupt landmark's expression in CLP speech is analyzed, and its importance in deriving the acoustic correlates of CLP speech intelligibility is demonstrated.
- The GP and SSM representations of speech are explored in intelligibility assessment. The distance between the normal template and test utterance is studied as a correlate of intelligibility.
- A visual representation based on the spider plot is proposed for graphing the intelligibility scores. The polygon in a spider plot represents intelligibility space of the speaker.

8.3 Directions for future work

Based on the outcome of this thesis work, this section provides some of the possible future directions for research.

(i) The effect of nasal air emission (NAE) in CLP speech intelligibility is not studied in the present thesis due to the unavailability of ground truth perceptual ratings for NAE. The effect of NAE in degrading intelligibility needs to be analyzed in the future study. Acoustic deviation related

to NAE can be incorporated as another dimension along with articulation and nasality in deriving composite measure. It is expected that the inclusion of this dimension may increase the correlation between the composite measure and the perceptual ratings.

(ii) The hypernasality estimation algorithm is dependent on speech data from moderate-severe hypernasality. The speech utterances used to build the acoustic model for maximum attainable nasality class, i.e., nasal sentences uttered by moderate-severe CLP speakers may not always be practically feasible to obtain. To address this problem, we have analyzed the nasalance score provided by Nasometer of nasal sentences of normal speakers and different category of CLP speakers. Bar graphs in Figure 4.5 of Chapter 4 show the nasalance values of nasal sentences of normal speakers and different category of CLP speakers. From the bar graphs, it can be seen that the mean nasalance value of normal speaker's nasal sentences is almost comparable to that of moderate-severe CLP speakers. Hence, it may also be possible to use the nasal sentences from normal speakers to build the acoustic model for the maximum attainable nasality class. The primary motivation is to develop a hypernasality score estimation method which is not dependent on the CLP speech. Since it is easier to collect the data of nasal sentences from the controlled normal speakers than from CLP speakers, this approach may be more feasible for practical applications. Future exploration can be made in this direction.

(iii) In the hypernasality level estimation algorithm, we have used speech signals from two extreme levels of nasality classes, and GMM is used for acoustic modeling. We obtained a correlation value of 0.69 between the predicted hypernasality scores and perceptual ratings; therefore, there is a scope for the investigation to improve the correlation value. Considering this, future work is planned to learn the nasality characteristics using the convolution neural network.

(iv) From Chapter 5, it is found that the reduced localized log-likelihood scores around *g–landmarks* may provide the extent of articulatory deviations in the respective landmark positions. Hence, this localized information can be helpful during the therapy, where feedback for the number of *+g–landmarks* and *-g–landmarks* are required to produce for a corresponding sentence and provide the score of precision to articulate each syllable in terms of likelihood value. A brief discussion about the possibility of analyzing the articulation error is given; however, a detailed analysis is needed in this direction. Future work is planned to explore the usefulness of derived

log-likelihood scores around the *g–landmarks* to study the correlation of different articulation errors with intelligibility degradation.

(v) In Chapter 6, intelligibility measure is derived based on the comparison of the posteriorgrams of normal and CLP speech utterances. In this case, the posterior vectors for an utterance are given equal importance. However, in the CLP speech, distortion in obstruents production has more effect on the intelligibility loss than the vowel. Therefore, giving more importance to the posterior vectors corresponding to speech regions around *g–landmarks* may be more beneficial. Re-weighting the posterior vectors around landmarks before the comparison can be performed, and future work is planned in this direction.

(vi) Recent works show the potentiality of deep belief network (DBN) to improve the discriminability of the posteriorgrams on phones with a very less amount of annotated data. The DBN can be trained using a semi-supervised manner, i.e., during the pre-training step, no annotation of the data is required; however, to fine-tune the pre-trained generative model speech data with annotation is needed. The effectiveness of the DBN-based posteriorgrams was shown in the mispronunciation detection; and found better results as compared to the GMM-based posteriorgrams. Therefore, DBN-based posteriorgrams may be explored in future for better modeling of acoustics units to derive the measure of intelligibility.

(vii) I-vector based speaker modeling has been found very useful in defining the measure of intelligibility at the speaker-level. In the CLP speech, an event-based i-Vector, which is derived using the features computed around specific events, such as VOPs, VEPs, and landmarks, may be more useful.

(viii) The intelligibility measures proposed in this thesis are explicitly studied for one particular pathology. The usefulness of these measures can be extended for the intelligibility assessment of other speech pathologies, such as dysarthria, hearing impaired.

Other related publications during thesis work

- **Journals**

1. **Sishir Kalita**, Akhilesh Kumar Dubey, C. M. Vikram, Protima Nomo Sudra, S. R. M. Prasanna, and S. Dandapat, "Excitation source based analysis of cleft lip and palate speech", [In review, *Speech Communication*].
2. Upashana Goswami. R. Nirmala, C. M. Vikram, **Sishir Kalita**, S. R. M. Prasanna, "Analysis of Articulation Errors in Dysarthric Speech", *Journal of Psycholinguistic Research*, pp: 1-12, 28 October 2019, doi: doi=10.1007/s10936-019-09676-5.
3. C M Vikram, Ayush Tripathy, **Sishir Kalita**, S. R. M. Prasanna, "Acoustic analysis of misarticulated trills in cleft lip and palate children", *J. Acoust. Soc. Am.*, 143(6), EL474-EL480, June 2018, doi: 10.1121/1.5042339.

- **Conferences**

1. **Sishir Kalita**, Protima Nomo Sudro, S. R. M. Prasanna, and S. Dandapat, "Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech", in *Proc. Interspeech 2019*, Austria, September 2019.
2. Protima Nomo Sudro, **Sishir Kalita**, S. R. M. Prasanna, "Processing Transition Regions of Glottal Stop substituted /s/ for Intelligibility Enhancement of Cleft Palate Speech", in *Proc. Interspeech 2018*, Hyderabad, India, September 2018.
3. C. M. Vikram, Ayush Tripathi, **Sishir Kalita**, S. R. M. Prasanna, "Estimation of Hypernasality Scores from Cleft Lip and Palate Speech", in *Proc. Interspeech 2018*, Hyderabad, India, September 2018.
4. Sarfaraz Jelil, **Sishir Kalita**, S. R. M. Prasanna and Rohit Sinha, "Exploration of Compressed ILPR Features for Replay Attack Detection", in *Proc. Interspeech 2018*, Hyderabad, India, September 2018.
5. Pamir Gogoi, **Sishir Kalita**, Parismita Gogoi, Ratree Wayland, Priyankoo Sarmah, S. R. M. Prasanna, "Analysis of Breathiness in Contextual Vowel of Voiceless Nasals in Mizo", in *Proc. Interspeech 2018*, Hyderabad, India, September 2018.

6. K Nikitha, **Sishir Kalita**, C. M. Vikram, M Pushpavathi, S. R. M. Prasanna, "Hypernasality Severity Analysis in Cleft Lip and Palate Speech Using Vowel Space Area", in *Proc. Interspeech 2017*, Stockholm, Sweden, August 2017.
7. **Sishir Kalita**, W Lalhminghlui, Luke Horo, Priyankoo Sarmah, S. R. M. Prasanna, S. Dandapat, "Acoustic Characterization of Word-Final Glottal Stops in Mizo and Assam Sora", in *Proc. Interspeech 2017*, Stockholm, Sweden, August 2017.
8. **Sishir Kalita**, Luke Horo, Priyankoo Sarmah, S. R. M. Prasanna, S. Dandapat, "Analysis of Glottal Stop in Assam Sora Language", in *Proc. Interspeech 2016*, San Francisco, USA, September 2016.
9. **Sishir Kalita**, S. R. M. Prasanna, S. Dandapat, "Analysis of glottal stops using pitch synchronous integrated linear prediction residual", in *Proc. National Conference on Communications (NCC)*, Guwahati, Assam, India, March 2016.

Bibliography

[1] P. A. Mossey, E. E. Catilla *et al.*, "Global registry and database on craniofacial anomalies: Report of a who registry meeting on craniofacial anomalies," 2003.

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

[3] D. J. Zajac and L. D. Vallino, *Evaluation and Management of Cleft Lip and Palate: A Developmental Perspective*. Plural Publishing, 2017.

[4] J. Stengelhofen, *Cleft palate: The nature and remediation of communication problems*. Churchill Livingstone, 1989.

[5] 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, pMID: 14697067.

[6] M. Scipioni, M. Gerosa, D. Giuliani, E. Noth, and A. Maier, "Intelligibility assessment in children with cleft lip and palate in italian and german," in *Interspeech 2009*, 2009.

[7] 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 *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006, pp. 274–277.

[8] S. J. Peterson-Falzone, M. A. Hardin-Jones, and M. P. Karnell, *Cleft palate speech*. Mosby St. Louis, 2001.

[9] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Noth, "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.

[10] 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.

[11] T. L. Whitehill and C. H.-F. Chau, "Single-word intelligibility in speakers with repaired cleft palate," *Clinical linguistics & phonetics*, vol. 18, no. 4-5, pp. 341–355, 2004.

[12] 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, pMID: 11772170.

[13] 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.

[14] 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.

[15] M. Copeland, "The effects of very early palatal repair on speech," *British journal of plastic surgery*, vol. 43, no. 6, pp. 676–682, 1990.

[16] P. Landis and P. Thi-Thu-Cuc, "Articulation patterns and speech intelligibility of 54 vietnamese children with unoperated oral clefts: clinical observations and impressions." *The Cleft palate journal*, vol. 12, pp. 234–243, 1975.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer Speech and Language*, vol. 29, no. 1, pp. 132–144, 2015.

[22] M. Ganzeboom, M. Bakker, C. Cucchiarini, and H. Strik, "Intelligibility of disordered speech: Global and detailed scores," 2016.

[23] 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, pMID: 14697067.

[24] E. M. Konst, H. Weersink-Braks, T. Rietveld, and H. Peters, "An intelligibility assessment of toddlers with cleft lip and palate who received and did not receive presurgical infant orthopedic treatment," *Journal of communication disorders*, vol. 33, no. 6, pp. 483–501, 2000.

[25] 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.

[26] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 629030, 2009.

[27] R. D. Kent, *Intelligibility in speech disorders: Theory, measurement and management*. John Benjamins Publishing, 1992, vol. 1.

[28] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, 2012.

[29] R. Hummel, "Objective estimation of dysarthric speech intelligibility," *Queens University Kingston, Ontario, Canada*, 2011.

[30] M. Schuster, A. Maier, T. Bocklet, E. Nkenke, A. Holst, U. Eysholdt, and F. Stelzle, "Automatically evaluated degree of intelligibility of children with different cleft type from preschool and elementary school measured by automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 76, no. 3, pp. 362–369, 2012.

[31] V. Berisha, J. Liss, S. Sandoval, R. Utianski, and A. Spanias, "Modeling pathological speech perception from data with similarity labels," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 915–919.

[32] L. He, J. Zhang, Q. Liu, H. Yin, and M. Lech, "Automatic evaluation of hypernasality and speech intelligibility for children with cleft palate," in *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on*. IEEE, 2013, pp. 220–223.

[33] T. Bocklet, K. Riedhammer, E. Nth, U. Eysholdt, and T. Haderlein, "Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling," *Journal of Voice*, vol. 26, no. 3, pp. 390–397, 2012.

[34] A. S.-L.-H. Association *et al.*, "Slp health care survey: Caseload characteristics," *Rockville, MD: Author. Find this author on*, 2011.

[35] T. H. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, 2012.

[36] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2187–2199, Nov 2016.

[37] A. Maier, T. Haderlein, M. Schuster, E. Nkenke, and E. Nöth, "Intelligibility is more than a single word: Quantification of speech intelligibility by asr and prosody," in *International Conference on Text, Speech and Dialogue*, 2007, pp. 278–285.

[38] B. Vogt, A. Maier, A. Batliner, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Numeric quantification of intelligibility in schoolchildren with isolated and combined cleft palate," *HNO*, vol. 55, no. 11, pp. 891–898, 2007.

[39] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–156, 2008.

[40] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Noth, and A. Maier, "Towards robust automatic evaluation of pathologic telephone speech," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.* IEEE, 2007, pp. 717–722.

[41] G. Van Nuffelen, C. Middag, M. De Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International journal of language & communication disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[42] M. J. Kim and H. Kim, "Automatic assessment of dysarthric speech intelligibility based on selected phonetic quality features," in *International Conference on Computers for Handicapped Persons.* Springer, 2012, pp. 447–450.

[43] T. Bocklet, A. Maier, K. Riedhammer, and E. Noth, "Towards a language-independent intelligibility assessment of children with cleft lip and palate," in *In Proc. WOCCI 2009*, November 2009, pp. 4366–4369.

[44] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, "Evaluation and assessment of speech intelligibility on pathologic voices based upon acoustic speaker models," in *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop.* Citeseer, 2009, pp. 89–92.

[45] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 3, p. 10, 2015.

[46] D. Martınez, P. Green, and H. Christensen, "Dysarthria intelligibility assessment in a factor analysis total variability space," in *Proceedings of Interspeech*, 2013.

[47] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Interspeech*, 2017.

[48] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers," in *Interspeech*. ISCA, 2018, pp. 2943–2947.

[49] M. S. De Bodt, M. E. H.-D. Huici, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of communication disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[50] T. H. Falk, R. Hummel, and W.-Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4480–4483.

[51] M. J. Kim and H. Kim, "Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[52] R. L. Horwitz-Martin, T. F. Quatieri, A. C. Lammert, J. R. Williamson, Y. Yunusova, E. Godoy, D. D. Mehta, and J. R. Green, "Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis." in *INTERSPEECH*, 2016, pp. 1205–1209.

[53] L. Gu, J. G. Harris, R. Shrivastav, and C. Sapienza, "Disordered speech assessment using automatic methods based on quantitative measures," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, p. 768125, Jun 2005.

[54] ——, "Disordered speech evaluation using objective quality measures," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–321.

[55] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for childrens speech recognition," *Computer Speech & Language*, vol. 48, pp. 103–121, 2018.

[56] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6405–6409.

[57] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4733–4736.

[58] 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.

[59] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000.

[60] C. Park, "Consonant landmark detection for speech recognition," Ph.D. dissertation, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2008.

[61] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *The Journal of the Acoustical Society of America*, vol. 55, no. 3, pp. 653–659, 1974.

[62] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants." *Psychological Monographs: General and Applied*, vol. 68, no. 8, p. 1, 1954.

[63] A. Maier, E. Nöth, E. Nkenke, and M. Schuster, "Automatic assessment of childrens speech with cleft lip and palate," in *Proc. of the 5th Slovenian and 1st International Conference on Language Technologies (IS-LTC 2006)*, 2006, pp. 31–35.

[64] A. V. Fox, *PLAKSS: psycholinguistische Analyse kindlicher Sprechstörungen.* Harcourt Test Services, 2007.

[65] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[66] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[67] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[68] L. van der Molen, M. A. van Rossum, A. H. Ackerstaff, L. E. Smeele, C. R. Rasch, and F. J. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views," *BMC Ear, Nose and Throat Disorders*, vol. 9, no. 1, p. 10, 2009.

[69] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinsons disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, 2017.

[70] 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 *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 274–277.

[71] P. Green and J. Carmichael, "Revisiting dysarthria assessment intelligibility metrics," in *Eighth International Conference on Spoken Language Processing*, 2004.

[72] D. Le, K. Licata, C. Persad, E. M. Provost, D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2187–2199, 2016.

[73] T. Bocklet, K. Riedhammer, E. Nöth, U. Eysholdt, and T. Haderlein, "Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling," *Journal of Voice*, vol. 26, no. 3, pp. 390–397, 2012.

[74] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[75] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[76] C. Middag, Y. Saeys, and J.-P. Martens, "Towards an asr-free objective analysis of pathological speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[77] C. Middag, T. Bocklet, J.-P. Martens, and E. Nöth, "Combining phonological and acoustic asr-free features for pathological speech intelligibility assessment," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[78] J. Lee, K. C. Hustad, and G. Weismer, "Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 5, pp. 1666–1678, 2014.

[79] K. Ishikawa, A. de Alarcon, S. Khosla, L. Kelchner, N. Silbert, and S. Boyce, "Predicting intelligibility deficit in dysphonic speech with cepstral peak prominence," *Annals of Otology, Rhinology & Laryngology*, vol. 127, no. 2, pp. 69–78, 2018.

[80] K. Ishikawa, J. MacAuslan, and S. Boyce, "Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. EL441–EL447, 2017.

[81] T. M. DiCicco and R. Patel, "Automatic landmark analysis of dysarthric speech," *Journal of Medical Speech-Language Pathology*, vol. 16, no. 4, pp. 213–220, 2008.

[82] 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.

[83] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[84] D.-Y. Huang, Y. Zhu, D. Wu, and R. Yu, "Detecting intelligibility by linear dimensionality reduction and normalized voice quality hierarchical features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[85] X. Zhou, D. Garcia-Romero, N. Mesgarani, M. Stone, C. Espy-Wilson, and S. Shamma, "Automatic intelligibility assessment of pathologic speech in head and neck cancer based on auditory-inspired spectro-temporal modulations," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[86] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 2825–2828.

[87] J. Lee and M. Hahn, "Automatic assessment of pathological voice quality using higher-order statistics in the lpc residual domain," *EURASIP J. Adv. Signal Process*, pp. 1–9, 2009.

[88] 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.

[89] A. Potamianos and S. Narayanan, "Robust recognition of childrens speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 603–616, November 2003.

[90] E. J. Williams, *Regression analysis.* wiley, 1959, vol. 14.

[91] Y. Graham and T. Baldwin, "Testing for significance of increased correlation with human judgment," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 172–176.

[92] S. Kalita, P. M, 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)*, 2018.

[93] T. Arias-Vergara, J. Vásquez-Correa, J. Orozco-Arroyave, and E. Nöth, "Speaker models for monitoring parkinsons disease progression considering different communication channels and acoustic conditions," *Speech Communication*, 2018.

[94] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinsons disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[95] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[96] I. Laaridh, C. Fredouille, and C. Meunier, "Evaluation of a phone-based anomaly detection approach for dysarthric speech." in *INTERSPEECH*, 2016, pp. 223–227.

[97] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer speech & language*, vol. 50, pp. 62–84, 2018.

[98] C. Bhat, B. Vachhani, and S. Kopparapu, "Automatic assessment of articulation errors in hindi speech at phone level," in *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, 2015, pp. 1–4.

[99] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[100] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49–64, 2015.

[101] A. Maier, F. Hnig, T. Bocklet, E. Nth, 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.

[102] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.

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

[104] 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.

[105] 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.

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

[107] 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.

[108] 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.

[109] 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.

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

[111] 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.

[112] 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.

[113] L. He, J. Zhang, Q. Liu, H. Yin, M. Lech, and Y. Huang, "Automatic evaluation of hypernasality based on a cleft palate speech database," *Journal of medical systems*, vol. 39, no. 5, p. 61, 2015.

[114] 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.

[115] A. K. Dubey, A. Tripathi, S. 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.

[116] S. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Proc. of int. conf. signal processing and communications*. Citeseer, 2001, pp. 81–88.

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

[118] R. Kent, J. Kent, G. Weismer, R. Martin, R. Sufit, B. Brooks, and J. Rosenbek, "Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects," *Clinical Linguistics & Phonetics*, vol. 3, no. 4, pp. 347–358, 1989.

[119] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, April 2013.

[120] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[121] Y. Zhang, "Unsupervised speech processing with applications to query-by-example spoken term detection," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[122] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 1.

[123] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4733–4736.

[124] S. Strmbergsson, G. Salvi, and D. House, "Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children's speech," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3422–3435, 2015.

[125] C. Y. Lin and H. C. Wang, "Burst onset landmark detection and its application to speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1253–1264, July 2011.

[126] Z. B. Nossair and S. A. Zahorian, "Dynamic spectral shape features as acoustic correlates for initial stop consonants," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2978–2991, 1991.

[127] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal processing letters*, vol. 16, no. 6, pp. 469–472, 2009.

[128] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Medical Journal*, vol. 24, no. 3, pp. 69–71, 2012.

[129] D. He, B. P. Lim, X. Yang, M. Hasegawa-Johnson, and D. Chen, "Acoustic landmarks contain more information about the phone string than other frames for automatic speech recognition with deep neural network acoustic model," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3207–3219, 2018.

[130] K. Chenausky, J. MacAuslan, and R. Goldhor, "Acoustic analysis of pd speech," *Parkinsons Disease*, vol. 2011, 2011.

[131] H.-D. Huici, H. A. Kairuz, H. Martens, G. Van Nuffelen, and M. De Bodt, "Speech rate estimation in disordered speech based on spectral landmark detection," *Biomedical Signal Processing and Control*, vol. 27, pp. 1–6, 2016.

[132] L. Moro-Velazquez, J. Godino-Llorente, J. Gómez-García, J. Villalba, S. Shattuck-Hufnagel, and N. Dehak, "Use of acoustic landmarks and gmm-ubm blend in the automatic detection of parkinsons disease," in *Models and Analysis of Vocal Emissions for Biomedical Applications: 10th International Workshop, december, 13-15, 2017*, vol. 117. Firenze University Press, 2017, p. 73.

[133] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based svm," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[134] X. Yang, X. Kong, M. Hasegawa-Johnson, and Y. Xie, "Landmark-based pronunciation error identification on chinese learning," *submitted in Speech Prosody*, 2016.

[135] S. Boyce, H. Fell, and J. MacAuslan, "Speechmark: Landmark detection tool for speech analysis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[136] S. Boyce, M. Speights, K. Ishikawa, and J. MacAuslan, "Speechmark acoustic landmark tool: application to voice pathology." in *INTERSPEECH*, 2013, pp. 2672–2674.

[137] J. MacAuslan, H. Fell, S. Boyce, and L. Wilde, "Automated tools for identifying syllabic lndmark clusters that reflect changes in articulation," *Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 63–66, 2011.

[138] H. J. Fell, J. MacAuslan, L. J. Ferrier, and K. Chenausky, "Automatic babble recognition for early detection of speech related disorders," *Behaviour & Information Technology*, vol. 18, no. 1, pp. 56–63, 1999.

[139] H. J. Fell and J. MacAuslan, "Vocalization analysis tools," in *Fourth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2005.

[140] A. S. Warlaumont and H. L. Ramsdell-Hudock, "Detection of total syllables and canonical syllables in infant vocalizations." in *INTERSPEECH*, 2016, pp. 2676–2680.

[141] T. Talkar, "Design of tool for analysis of speech development disorders using landmarks and other acoustic cues," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[142] R. Ullmann, M. M. Doss, and H. Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4924–4928.

[143] R. McDonald, V. Parsa, and P. C. Doyle, "Objective estimation of tracheoesophageal speech ratings using an auditory model," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1032–1041, 2010.

[144] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, May 2012, pp. 382–387.

[145] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8227–8231.

[146] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, and E. Noth, "Word accuracy and dynamic time warping to assess intelligibility deficits in patients with parkinsons disease," in *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, Aug 2016, pp. 1–5.

[147] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, "Automatic scoring of shadowing speech based on dnn posteriors and their dtw," in *2017 Interspeech*, Aug 2017, pp. 1422–1426.

[148] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[149] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, Nov 2009, pp. 398–403.

[150] ——, "Towards multi-speaker unsupervised speech pattern discovery," in *In ICASSP 2010*, March 2010, pp. 4366–4369.

[151] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, ser. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993.

[152] A. Muscariello, G. Gravier, and F. Bimbot, "Towards robust word discovery by self-similarity matrix comparison," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5640–5643.

[153] ——, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2031–2044, Sept 2012.

[154] M. J. Saary, "Radar plots: a useful way for presenting multivariate health care data," *Journal of clinical epidemiology*, vol. 61, no. 4, pp. 311–317, 2008.