# Objective assessment of cleft lip and palate speech intelligibility

*Sishir Kalita*

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, India-781039
sishir@iitg.ac.in

## 1. Introduction

Cleft lip and palate (CLP) is one of the most common congenital disorders of the craniofacial region. Children with CLP require surgical intervention to establish the appropriate oral motor skills; however, the speech disorders persist due to velopharyngeal dysfunction (VPD) even after the surgical repair of the cleft [1]. The primary speech-related disorders CLP individuals exhibit are (i) hypernasality, (ii) articulation errors, (iii) nasal air emission, and (iv) voice disorders, all of which affect the overall intelligibility of the speech [1, 2, 3]. Improving the intelligibility is the primary concern in clinical care. A means of assessing speech intelligibility is required to determine (i) overall articulation capability, (ii) the improvement in speech due to therapy, and (iii) the outcomes of other interventions [9].

### 1.1. Clinical intelligibility assessment and need for objective measure

In the clinical environment, speech-language pathologists (SLP) perceptually assessed the intelligibility using (i) various rating scales, (ii) through the transcription, and (iii) different protocols [2, 10, 4]. By its nature, perceptual evaluation is subjective and can lead to inaccurate and biased decisions, but it is considered nevertheless to be the gold standard [12, 5]. Even so, there is always some extent of intra-rater and inter-rater disagreement in perceptual evaluation [8]. Therefore, an objective method for assessing speech intelligibility is urgently required to assist SLPs with their therapeutic and other rehabilitation processes [5, 23]. The application of these approaches may significantly contribute to the assessment process in terms of (i) consistent objective results, (ii) remote monitoring of speech disorders, and (iii) reduction in the cost of care [9]. Since objective measures only rely on the acoustic characteristics, bias due to the contextual information may not be present in these methods [5].

### 1.2. Previous works

Currently, researchers have shown the significance of automatic speech recognition (ASR) techniques to quantify the intelligibility of CLP speech [5, 11, 12, 23]. In these approaches, word error rate (WER) is considered to quantify the speech intelligibility, and a significant correlation is observed with respect to SLPs perceptual scores. Although ASR based systems provide a high degree of correlation with perceptual ratings, a large amount of annotated data is needed to build the acoustic and language models. This is relatively difficult for a *low resource* scenario like CLP speech analysis [8]. To overcome the requirement of a large amount of annotated data in the ASR-based system, supervectors generated from speaker-specific Gaussian mixture models (GMMs) are used as the input for support vector regression (SVR), and the SVR output is used to quantify the speech intelligibility [8].

### 1.3. Motivations

In the ASR based system, WER gives a global view of the intelligibility for each CLP individual. However, a sentence-level intelligibility score is difficult to obtain from the WER. Moreover, the WER does not provide the information about how different speech disorders are impacting the speech intelligibility. The relative impact of each speech disorder on the CLP speech intelligibility is essential for the SLPs. Furthermore, in the previous works, ASR is built on adult data and adapted for children's data to determine the word accuracy for intelligibility evaluation. Most of the ASR systems use MFCCs to model the vocal tract characteristics, which requires a smooth vocal tract spectrum. However, in the case of high pitch, particularly for children's speech, harmonics are widely spaced which results in inadequate smoothing of the spectral envelope [14, 15]. The inadequate smoothing of the spectral envelope may lead to the improper representation of the sound units [14]. Also, the acoustic mismatch between adults' and children's speech is a great challenge for the children's ASR trained on adults' data [16]. Hence, assessment of intelligibility using these approaches may not be very reliable. Conventional MFCCs and their derivatives do not explicitly model the dynamic characteristics present in the transition region between two sounds [17]. Since the important perceptual cues of intelligibility are embedded in the transition region, proper modeling of this region is essential [18, 19]. In CLP speech, the intelligibility mainly degrades due to the deviations in the place of articulations (PoA) and the manner of articulations (MoA). Acoustic cues related to PoA [20] and MoA [21, 22] are predominantly preserved in the transition regions. Hence, acoustic features which give the explicit representation of temporal dynamics between two sounds are required.

## 2. Database

Since there is no publicly available database to experiment, I have developed a database for CLP speech. Speech samples were collected in collaboration with the All Indian Institute of Speech and Hearing (AIISH), Mysuru, India. All the children with cleft have undergone primary surgery and do not have other congenital disorders and developmental problems. Only CLP children with adequate language abilities were considered for the study. Before the recording, consent was obtained from the children's parents.

Fifteen phonetically balanced sentences and fifty meaningful words in the Kannada language, rich in obstruent consonants, were used as the speech stimuli. These stimuli were designed by SLPs of AIISH, Mysuru for intelligibility assessment of Kannada CLP individuals. Speech samples were recorded in a sound-proof room using a directional microphone (Bruel & Kjaer) with a sampling frequency of 44 kHz and $16-$bit resolution on a mono channel.

Three SLPs of AIISH, Mysuru with around five years of experience in the field of CLP speech evaluation assessed the intelligibility by a perceptual evaluation method. The SLPs provided sentence-level and word-level intelligibility scores on a scale from 0 to 3, where, $0 =$ near to normal, $1 =$ mild, $2 =$ moderate, and $3 =$ severe. Apart from the sentence-level and word-level intelligibility score, SLPs also provided a global intelligibility score of each CLP individual in the same scale. We computed Spearman's rank correlation coefficient ($\rho$) and Cohen's kappa ($\kappa$) between the score of an individual rater and the mean of the other two raters and found the intelligibility ratings sufficiently reliable. Hence, the median value of the three raters scores was considered as the ground truth for the current work.

## 3. Methods and Results

### 3.1. Analysis of the relative contribution of speech disorder on intelligibility

The relative contribution of different speech disorders, such as hypernasality, articulation errors, nasal air emission, and voice disorders on the CLP speech intelligibility is studied. Also, it is evaluated whether all the speech disorders influence the assessment of CLP speech intelligibility. To investigate this, the perceptual ratings of these speech disorders are used to build a regression model. Using this model, we show that the CLP speech intelligibility can be expressed as the weighted linear combination of the aforementioned speech disorders. The weights provide the relative impact of individual speech disorders on overall speech intelligibility. The articulation error shows the highest impact on the intelligibility, while the voice disorder has significantly less or no contribution on the intelligibility.

Further, the above knowledge is used to develop an objective measure. Here, separate acoustic models are developed for nasality, articulation, nasal air emission, and voice quality. The combined decision from the models is explored as the objective measure of speech intelligibility. The weights derived from the perceptual analysis are used to rank the scores derived from the models. Results show a significant correlation between the proposed objective measure and the perceptual intelligibility ratings.

### 3.2. Exploration of pitch-normalized joint spectro-temporal feature based Gaussian posteriograms for intelligibility assessment

This work studies the role of spectral smoothing and explores the pitch-normalized features in intelligibility assessment of CLP children's speech. Motivated by the perceptual importance of the transition region, we computed the joint spectro-temporal based features from the overlapping patches of time-frequency representation for better characterization of spectral and temporal modulations [17, 20]. Since speaker independent acoustic segment representation is very essential, derived joint spectro-temporal features are mapped into Gaussian posteriograms (GPs). The GP for an utterance is derived from a sentence specific GMM which is built on normal data, and it provides a speaker independent representation of the underlying acoustic segments present in the respective utterance. GP representation of the distorted unintelligible speech of CLP children will be distinctly different from the normal children's speech.

Two comparison based frameworks using dynamic time warping (DTW) and matching of self-similarity matrices (SSMs) are applied to compute the deviation of the GP representation of the test CLP speech from that of the normal template. The relative deviation from the normal speaker's template is considered as the representation of intelligibility. In case of the DTW based method, DTW distance between the GP representation of test CLP speech and the normal speech template is considered the sentence-level intelligibility score. 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 [24, 25] and it is robust against the speaker variabilities. SSM of a feature 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 will deviate from the 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, which is considered as the representative of the objective intelligibility score. Compared to the DTW based system, the SSM based method shows better performance for all the explored features.

## 4. Conclusion and Future work

In my PhD work, the primary goal is to develop an objective measure for CLP speech intelligibility. Initially, we have analyzed the relative impact of different speech disorders on the CLP speech intelligibility and used this knowledge to derive an objective measure using the combined evidence of different speech disorders. The importance of pitch-normalized and joint spectro-temporal feature is shown for intelligibility assessment of CLP children. Such features are used in two comparison based frameworks to derive the objective intelligibility scores. The future work is motivated by the current results that have been discussed. Some of our plans are listed below.

- In CLP speech, articulation error has the highest impact on the speech intelligibility, and articulation problems mostly occur due to the improver production of the obstruents. This deviant characteristic results in the distortion of spectro-temporal dynamics at the transition region from obstruent to vowel, and vice-versa. Most of the abrupt landmarks are associated with such vocalic transitions, where important perceptual cues are embedded [26]. The objectives to investigate are (i) how landmark expression is distorted in CLP speech, and (ii) whether an acoustic correlate of CLP speech intelligibility can be derived by analyzing the speech anchored around the abrupt consonant landmarks.
- We are planning to derive the global intelligibility score for each CLP individual using two approaches: (i) the combined decision of sentence-level intelligibility scores from the previous studies, and (ii) the i-vector modeling of the CLP individuals' speech.

## 5. Acknowledgement

I express sincere gratitude to my thesis supervisors Prof. S. R. Mahadeva Prasanna and Prof. S. Dandapat, under whom generous supervision this work has been performed. Also, I would like to acknowledge Prof. M. Pushpavathi and Prof. Ajish Abraham of All India Institute of Speech and Hearing (AIISH), Mysuru for their valuable contribution to the perceptual evaluation of speech. This work is in part supported by the project grants, for the projects entitled ”NASOSPEECH” funded by the Department of Biotechnology (DBT), Govt. of India and ”ARTICULATE +” funded by the Ministry of Human Resource Development (MHRD), Govt. of India.

# 6. References

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

[2] 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 **41**(1), 64–70 (2004), pMID: 14697067.

[3] 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 **126**(5), 2589–2602 (2009).

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

[5] 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 **76**(3), 362–369 (2012).

[6] 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 **33**(6), 483–501 (2000).

[7] B. J. McWilliams, “Some factors in the intelligibility of cleft-palate speech,” Journal of Speech and Hearing Disorders **19**(4), 524–527 (1954).

[8] 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* (2009), pp. 4366–4369.

[9] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, “Automatic intelligibility classification of sentence-level pathological speech,” Computer Speech and Language **29**(1), 132–144 (2015).

[10] 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 **45**(1), 1–17 (2008)

[11] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nth, “Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition,” International Journal of Pediatric Otorhinolaryngology **70**(10), 1741–1747 (2006).

[12] 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)* (2006), Vol. 4, pp. 274–277,

[13] 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 **39**(5), 61 (2015).

[14] S. Ghai and R. Sinha, “Exploring the role of spectral smoothing in context of childrens speech recognition,” in *INTERSPEECH 2009* (2009), pp. 1607–1610.

[15] S. Ghai and R. Sinha, “On the use of pitch normalization for improving childrens speech recognition,” in *INTERSPEECH 2009* (2009), pp. 568–571.

[16] A. Potamianos and S. Narayanan, “Robust recognition of childrens speech,” IEEE Trans. on Speech and Audio Proc. **11**(6), 603–616 (2003).

[17] 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* (2008), pp. 4733–4736.

[18] K. N. Stevens, *Acoustic phonetics* (MIT press, 2000).

[19] C. Park, “Consonant landmark detection for speech recognition,” Ph.D. thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2008.

[20] V. Karjigi and P. Rao, “Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling,” Speech Communication **54**(10), 1104–1120 (2012).

[21] 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 **55**(3), 653–659 (1974).

[22] 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 **68**(8), 1 (1954).

[23] 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).

[24] 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.

[25] A. Muscariello, G. Gravier, and F. Bimbot, “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.

[26] S. A. Liu, “Landmark detection for distinctive feature-based speech recognition,” The Journal of the Acoustical Society of America **100**(5), 3417–3430 (1996).