Intelligibility Enhancement of Cleft Lip and Palate Speech

Protima Nomo Sudro

Department of Electronics and Electrical Engineering
Indian institute of Technology Guwahati, Guwahati, India
protima@iitg.ernet.in

1. Introduction

Cleft lip and palate (CLP) is a craniofacial condition which leads to various speech-related disorders. CLP interferes with the speech intelligibility resulting in communication impairments. Even after surgical repair, the speech disorders persists due to the associated velopharyngeal dysfunction (VPD) or oronasal fistula [1, 2, 4]. The speech-related disorders are broadly categorized into hypernasality, hyponasality, nasal air emission, consonant production errors, and voice disorders [3]. Clinically, the main concern is to enhance the intelligibility of CLP speech. Improvement of CLP speech intelligibility is necessary to help the CLP speakers communicate effectively.

1.1. Need for intelligibility enhancement

The clinical treatment for improving the intelligibility of CLP speech involves primary and secondary surgery. The CLP speech correction requires a long period of time and the ratio of speech-language pathologists (SLPs) to the individuals with CLP is very small. To correct functional disorders, speech therapy is mostly recommended. In standard speech therapy technique, an SLP simulates the disordered speech and presents to the individual with CLP along with the correct speech. Sometimes biofeedback (auditory, visual, and tactile) mechanisms are also employed to enhance the speech intelligibility during speech therapies [5]. Several prior studies have reported that modified auditory feedback helps in acquiring correct speech sound production [7, 8, 9, 10, 12]. However, studies like modified auditory feedback mechanism based on signal processing techniques were not attempted for CLP speech till now. Hence, if the disordered speech and its modified version is provided as an auditory feedback to the individuals with CLP along with other standard speech therapy techniques, it may be more effective for the speakers with CLP. Also presenting the individuals with their modified speech will motivate them by giving them a preview of the voice that would be achieved after successful speech therapy.

1.2. Motivation of the study

In literature, several approaches based on signal processing techniques are proposed for improving the intelligibility of various pathological speech. One of these corresponds to dysarthric speech modification based on acoustic transformation [13], Gaussian mixture modeling (GMM) [15], etc. Alaryngeal speech enhancement includes the transformation of speech by enhancing formants, perceptual weighting technique [16, 17]. Methods like frequency lowering system were proposed for enhancing the intelligibility of degraded speech [18]. A few studies also report speech intelligibility enhancement for individuals with articulation disorders using voice conversion technique [19, 20]. The existing pathological speech enhancement studies involve issues of various communication disorders. Despite the potential of the enhanced speech in speech therapy, CLP speech enhancement is not studied abundantly in the literature. The exception is one recent work which involves spectral enhancement of hypernasal speech [22]. In CLP speech due to misarticulated obstruents and vowel nasalization, intelligibility is mostly distorted. Hence, we first try to analyze intelligibility distortions caused by specific obstruent errors like misarticulated fricative /s/ and stop consonants /b,d,g,k,t,T/ and vowel nasalization /a,i,u/. Based on the observed distortions, the errors are modified and recombined with the original speech to observe the impact of the enhancement method.

2. Database Description

The speech materials were collected from the All India Institute of Speech and Hearing (AIISH), Mysuru, India. Native Kannada-speaking children were recruited for the recording. 29 CLP speakers participated in the study which consists of 17 male and 12 female, and the age was 9 ± 2 years (mean $\pm$ standard deviation) during the recording. Not one of the CLP participant bears any history of developmental difficulties. 31 speakers (12 male and 19 female) with normal speech and language characteristics also participated in the study as a control group. The age of the control group was 10 ± 2 years (mean $\pm$ standard deviation). Prior consents were collected from the parents before the recording of speech samples. Speech samples were recorded under clean room condition using the sound level meter (Bruel & Kjær, Nærum, Denmark) at 48 kHz sampling rate and 16-bit resolution. The SLPs assess the speech distortions by transcribing the speech samples and providing deviation scores on a scale of 0 to 3, where, $0 =$ close to normal, $1 =$ mild deviation, $2 =$ moderate deviation, and $3 =$ severe deviation. Subsequently, using PRAAT software [23], waveform, and spectrographic analysis are also performed to decide the category of errors.

3. Methodology and Results

In this thesis, we attempt to analyze and modify obstruent misarticulations and vowel nasalization. At first, we analyzed and modified the fricative /s/ misarticulation. When the original /s/ is replaced with modified /s/ in consonant-vowel-consonant-vowel (CVCV) words, it is observed that due to vowel nasalization, the CVCV word is not perceived as intelligible as it should be. Thus, to obtain word-level intelligibility, the vowels must also be de-nasalized. Therefore, we analyzed three nasalized vowels: /a/, /i/, and /u/ and modified them using temporal and spectral processing. Further, we attempt to modify the compensatory errors produced for stop consonants using nonnegative matrix factorization method.

3.1. Analysis of fricative /s/ misarticulation and its enhancement

This work involves investigating the acoustic characteristics of misarticulated fricatives for automatic segmentation followed by modification of these errors close to normal like /s/. We

analyze three types of misarticulated fricative /s/: palatalized articulation, nasal air emission distorted /s/ and glottal stop substituted /s/ in initial and medial position in fricative-vowel-fricative-vowel (FVFV) structure [3]. We perform automatic segmentation of the misarticulated fricatives using the onset of glottal activity region as an anchoring point. Within 150 ms of the onset point of glottal activity region, we search for fricative evidence using zero crossing rate, band energy ratio and spectral tilt. If fricative evidence is detected, then the category of fricative error is determined using the features namely, spectral centroid, dominant spectral centroid, and maximum normalized spectral slope. Further, an average spectral distance is measured between the detected misarticulated /s/ and normal /s/, where an optimal filter is designed to compensate for the difference using generalized singular value decomposition. In the above-discussed work, only the sustained region of fricative is analyzed and modified. As the transition region between F and V and V and F is also distorted due to co-articulation effect, and since important perceptual cues are embedded in the transition region, it is, therefore, necessary to modify the transition region characteristics for more intelligible speech. Therefore, 2D-DCT based joint spectro-temporal features are exploited for the modification [21].

3.2. De-nasalization of hypernasal vowels using temporal and spectral processing

The nasalization of vowels reduces the clarity of speech, thus making the speech unintelligible [1, 24]. From the acoustic analysis of HN speech, it is observed that the spectral characteristics of vowels (/a/,/i/, and /u/) are deviated due to the introduction of additional formant and anti-formants in the spectrum, broadening of formant bandwidths and spectral flattening [25]. It is also noted that, because of the deviated spectral characteristics, the obtained residual signal consists of scaled and delayed versions (interfering components) of the original speech [26]. While synthesis, the interfering signal components in the residual signal may sometimes introduce unnatural spectral changes which are perceived as distortion in the enhanced speech. Therefore, modification of residual and spectral characteristics are expected to result in intelligible speech. The spectral characteristics of the HN speech signal are modified by transforming the spectral envelope while keeping the fundamental frequency unmodified. The transformation is achieved using GMM based spectral conversion function derived from the source (HN) and target (normal) speakers probabilistic model. The extended linear prediction coefficient (XLPC) cepstrum from both the source and target speakers are used to build the conversion function for training. The transformation consists of mapping the XLPC cepstrum of HN speech towards target XLPC cepstrum by using trained conversion function. A fine weight function is used for deemphasizing the interfering signal components of the XLP residual signal.

3.3. Modification of compensatory errors produced for stop consonants in CLP speech

The presence of compensatory errors in cleft lip and palate (CLP) speech degrades the speech intelligibility severely. This work focuses on the modification of compensatory errors produced for stop consonants in CLP speech using spectral transformation technique. Stop consonants are characterized by dynamically varying spectro-temporal characteristics. Hence, for the modification of compensatory errors, namely, glottal, palatal, and velar stop substitutions produced for the unvoiced stops (/k/, /t/ and /T/) and devoicing of voiced stops (/b/, /d/ and /g/) the spectral transformation should specifically represent the dynamic characteristics rather than the mixture of different spectral components present in the utterance. Therefore, in this work, an event-based approach is proposed for the spectral transformation. First, automatic detection of burst onset and vowel onset events is carried out. Having detected burst and vowel onsets, the region from burst onset to 20 ms transition followed by vowel onset is segmented. The segmented regions of the source (CLP) and target (normal) speech are used for learning the transformation matrix, which is optimized using nonnegative matrix factorization method. The optimized transformation matrix is further used to modify the compensatory errors.

4. Conclusion and Future work

This thesis aims to develop an approach to improve the intelligibility of CLP speech. The impact of intelligibility distortions on specific-phonemes is analyzed and based on the deviated characteristics an approach is proposed to compensate the errors. The following phoneme-specific modifications are attempted, specifically, fricative /s/ in vowel context /a/, vowels and, consonants in vowel context /a/. Based on the current study, future work is planned as the following:

- An attempt is made to improve the word-level intelligibility compared to isolated phoneme enhancement. As an illustration, three severely nasalized words such as /baba/, /dada/ and /gaga/ are studied. The intelligibility is improved using GMM-based spectral conversion. The perceptual evaluation revealed that the enhanced speech sounds muffled. Therefore, we further process the enhanced speech signal with an adaptive filtering method. Prior to further processing, the vowels and voiced consonants are segregated using the knowledge of the glottal activity. The vowel deviations are modified by reducing the formant bandwidths and scaling the formant amplitudes and the corresponding valleys. To modify the voiced stop consonants, the spectral peaks are optimized with respect to the normal speakers stop consonant's spectral peaks.
- Being a preliminary work, in this thesis, only specific phonemes in CVCV structures are analyzed and modified. Moreover, the CVCV structures are analyzed in one vowel contexts only. However, to meet the real-time scenarios further exploitation of the proposed approach is yet to be done for different vowel contexts, meaningful words, and spontaneous speech.
- Additionally, the speech data is collected in a clean room condition. However, with changing environment, background, and reverberation noise may also degrade the segmentation accuracy of the proposed approach, and hence it may affect the enhancement as well. This issue must also be explored in future work.

5. Acknowledgements

I would like to express my sincere gratitude towards my thesis supervisor Prof. S. R. Mahadeva Prasanna under whose guidance, the current work is performed. I would also like to thank Dr. M.Pushpavathi and Dr. Ajish Abraham, AIISH, for their support during data collection and perceptual evaluation. This work is in part supported by a project entitled NASOSPEECH: Development of Diagnostic system for Severity Assessment of the Disordered Speech funded by the Department of Biotechnology (DBT), Govt. of India.

6. References

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

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

[3] 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.

[4] J. H. N. Pinto, G. S. Dalben, and M. I. Pegoraro-Krook, "Speech intelligibility of patients with cleft lip and palate after placement of speech prosthesis," *The Cleft palate-craniofacial journal*, vol. 44, no. 6, pp. 635–641, 2007.

[5] A. W. Kummer, "Speech therapy for errors secondary to cleft palate and velopharyngeal dysfunction," in *Seminars in speech and language*, vol. 32, no. 2, 2011, p. 191.

[6] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech i: Compensation and adaptation," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 2, pp. 295–310, 2002.

[7] V. M. Villacorta, J. S. Perkell, and F. H. Guenther, "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2306–2319, 2007.

[8] D. G. Jamieson and S. Rvachew, "Remediating speech production errors with sound identification training," *Journal of Speech-Language Pathology and Audiology*, vol. 16, no. 3, pp. 201–210, 1992.

[9] D. M. Shiller, S. Rvachew, and F. Brosseau-Lapré, "Importance of the auditory perceptual target to the achievement of speech production accuracy." *Canadian Journal of Speech-Language Pathology & Audiology*, vol. 34, no. 3, 2010.

[10] D. M. Shiller and M.-L. Rochon, "Auditory-perceptual learning improves speech motor adaptation in children." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 4, p. 1308, 2014.

[11] D. M. Shiller, M. Sato, V. L. Gracco, and S. R. Baum, "Perceptual recalibration of speech sounds following speech motor learning," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1103–1113, 2009.

[12] E. D. Casserly, "Speaker compensation for local perturbation of fricative acoustic feedback," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2181–2190, 2011.

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

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

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

[16] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE transactions on speech and audio processing*, vol. 5, no. 2, pp. 97–105, 1997.

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

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

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

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

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

[22] C. Vikram, N. Adiga, and S. M. Prasanna, "Spectral enhancement of cleft lip and palate speech." in *INTERSPEECH*, 2016, pp. 117–121.

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

[24] J. Trost-Cardamone, "Diagnosis of specific cleft palate speech error patterns for planning therapy or physical management needs," *Communicative disorders related to cleft lip and palate*, 1997.

[25] 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.

[26] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE transactions on acoustics speech and signal processing*, vol. 27, no. 4, pp. 309–319, 1979.