Interspeech 2018
2-6 September 2018, Hyderabad

Pitch-Adaptive Front-end Feature for Hypernasality Detection

Akhilesh Kumar Dubey[1], S. R. Mahadeva Prasanna[1, 2], S. Dandapat[1]

[1]Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, India
[2] Department of Electrical Engineering
Indian Institute of Technology Dharwad, Dharwad, India
(d.akhilesh, prasanna, samaren)@iitg.ernet.in

Abstract

Hypernasality in cleft palate (CP) children is due to the velopharyngeal insufficiency. The vowels get nasalized in hypernasal speech and the nasality evidence are mainly present in low-frequency region around the first formant (F_1) of vowels. The detection of hypernasality using Mel-frequency cepstral coefficient (MFCC) feature may get affected because the feature might not be able to capture the nasality evidence present in the low-frequency region. This is due to the fact that the MFCC feature extracted from high pitched children speech contains the pitch harmonics effect of magnitude spectrum. The pitch harmonics effect results in high variance for the higher dimensions of MFCC coefficients. This problem may increase due to high perturbation in pitch of CP speech. So in this work, a pitch-adaptive MFCC feature is used for hypernasality detection. The feature is derived from the cepstral smooth spectrum instead of magnitude spectrum. A pitch-adaptive low time liftering is done to smooth out the pitch harmonics. This feature when used for the detection of hypernasality using support vector machine (SVM) gives an accuracy of 83.45 %, 88.04 and %, 85.58 % for /a/, /i/ and /u/ vowels respectively, which is better than the accuracy of MFCC feature.

Index Terms: Hypernasality, Pitch adaptive Mel-frequency cepstral coefficient, Cleft palate.

1. Introduction

The cleft palate (CP) is a congenital craniofacial disorder. The speech of cleft palate (CP) children exhibit deviance due to structural abnormalities, inadequate functioning of velopharyngeal port and mis-learning [1]. This speech deviance is universally reported in terms of resonance disorder, nasal air emission and/or turbulence, consonant production errors and voice disorder [2]. Hypernasality in speech is an important resonance disorder where excess nasality is heard during the production of voice sounds, especially vowels. The nasality is heard due to the coupling of nasal tract with the oral tract during the production of speech. The structural abnormalities correction done by the plastic surgeons may not be sufficient to restrict the nasal tract coupling and hence, nasality remains present in repaired CP children. This happens because of velopharyngeal insufficiency and mis-learning [3]. The intelligibility of CP speech gets affected due to hypernasality. The evaluation of hypernasal speech is needed for proper diagnosis of CP children by plastic surgeons and speech-language pathologists (SLPs).

In the clinical environment, evaluation of hypernasality is done perceptually by SLPs and the decision is confirmed by using some instrumental method of evaluation. The confirmation is needed because the perceptual decision may sometimes vary among the SLPs [4]. The variation happens due to the presence of abnormalities in pitch, loudness, voice quality and/or articulation in CP speech in conjunction with the hypernasality [5]. The instrumental method of evaluation may be direct or indirect. In direct method, the instrumental techniques like X-Ray (Cephalometry), videofluoroscopy, nasendoscopy, etc [6] are used to observe the movement of velopharyngeal port. These techniques may have radiation effect or may be invasive which may be painful to the children. In indirect method, the aerodynamics and/or acoustic measurements are done using the techniques like accelerometry and nasometry to infer about velopharyngeal activity. The nasometer is an example of such type of device which is widely used clinically, to measure the nasality in speech in terms of "nasalance" value. Indirect techniques are radiation free, noninvasive but require extra sensing device at the nose besides the microphone at the mouth. Due to aforementioned limitations of both direct and indirect techniques, another indirect technique based on the spectral analysis of speech using digital signal processing is used by the researchers for the evaluation of hypernasality. This technique is objective, noninvasive and simple [4] and requires only a microphone and a computer.

In spectral analysis method, the vowels /a/, /i/ and /u/ of hypernasal speech are analyzed to find the spectral deviation in these vowels compared to the normal speech vowels. The presence of nasal peak in low-frequency region around the first formant F_1, reduction in strength of F_1 and hence broadening of F_1 are some important spectral cues proposed for nasalized vowels [7], [8] which are used by the researchers for the hypernasality detection. The important works reported in the literature for hypernasality detection are based on Teager energy operator [9], Teager energy operator plus Mel frequency cepstral coefficient (MFCC) [10], linear prediction cepstral coefficient (LPCC) [11], high spectral resolution group delay spectrum [4], set of features based on acoustic, noise and cepstral analysis, nonlinear dynamic and entropy measurements [12], [13], [14], energy distribution [15], [16] and zero time windowing [17], [18].

The detection of hypernasality can be done phoneme wise or frame wise. In most of the above works, it is done phoneme wise. The frame wise hypernasality detection work is done in [10] using the MFCC feature. The accuracy of hypernasality detection using MFCC feature may get affected because the studies [19] [20] shows that the MFCC feature gets affected for high-pitched children speech. This happens because the Mel-filter bank employed is unable to sufficiently smooth out the pitch harmonics present in the magnitude spectrum of the windowed speech signal. Hence, ripples appear in the smoothed spectral envelope corresponding to MFCC feature in the lower frequency region which gives the high variance for the higher coefficients of MFCC feature. The variance may be higher in

10.21437/Interspeech.2018-1251

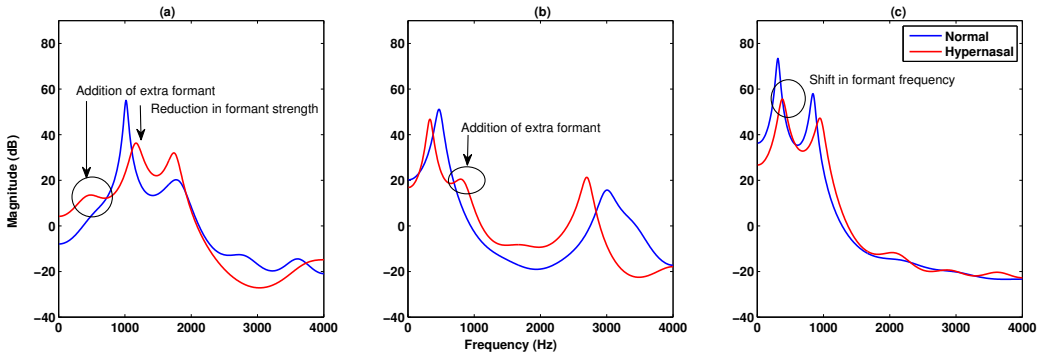

Figure 1: *Linear prediction spectrum of normal and hypernasal vowels. (a) for /a/ vowel (b) for /i/ vowel (c) for /u/ vowel. Figure shows the addition of extra formants, reduction in formant strength and shift in formant frequency for hypernasal vowels.*

CP speech due to high pitch perturbation. Further, since the nasality evidence is mainly present in the low-frequency region around F_1 in hypernasal speech and the ripples in smoothed spectral envelope corresponding to MFCC feature also appears in the low-frequency region, hence MFCC feature may not be able to capture the nasality evidence effectively. This will also affect the accuracy of hypernasality detection.

So in this work, a pitch adaptive Mel-frequency cepstral coefficient (PAMFCC) feature is used for the hypernasality detection. To compute the PAMFCC feature the Mel-filter bank is employed on the cepstral smoothed spectrum rather than magnitude spectrum. A pitch adaptive liftering of cepstral coefficients derived from magnitude spectrum is done to compute the cepstral smoothed spectrum due to high perturbation in children speech, especially CP speech [13]. The PAMFCC feature is free from the pitch harmonics effect and hence can capture the nasality evidence present in low-frequency of hypernasal speech which may enhance the accuracy of hypernasality detection.

The rest of the paper is organized as follows. Section 2 described about the CP speech database. In Section 3 the Pole-zero analysis of hypernasal speech is presented. Section 4 describes the effect of pitch on MFCC feature. Section 5 describes the steps of computing pitch-adaptive MFCC feature. Section 6 gives the experimental result and finally section 7 contains the summary and conclusion of the work.

2. Speech database

There is a great challenge in the collection of CP speech for hypernasality detection research due to the limited subjects, variability in their language accent and design of a stimuli list which can capture the hypernasal speech characteristics. In this work, data is collected from two group of children: 30 normal children having normal speech and 30 repaired CP children having hypernasal speech. Out of 30 children of each group, 18 are boys and 12 are girls. The age range of children lies between 7-12 years. The native language of all children is Kannada, so the data is recorded in the Kannada language which is a Dravidian language spoken in the southern part of India. The stimuli considered here are /papa/, /pipi/ and /pupu/ as suggested in [2] where the vowels /a/, /i/ and /u/ immediately follow the pressure consonant /p/. The vowels are extracted from the stimuli. For that, the manual annotation of vowels is done using Wavesurfer tool [21]. The database consists of total 542 normal, 464 CP phoneme /a/, 516 normal, 452 CP phoneme /i/ and 524 normal, 484 CP phoneme /u/. The data is recorded in the sound-treated room of All Indian Institute of Speech and Hearing (AIISH), Mysore, India [22] using Bruel & Kjaer sound level meter (SLM) microphone. During the time of recording, the instructor first utters the word and then the child repeats the same. The recording is done at sampling frequency 44.1 kHz, 16 bps in .WAV format, which is down-samples at 16 kHz for the analysis in this work. Perceptually each recorded sound is judged separately by three SLPs for normal and hypernasal speech classification. The recordings having common decision from three SLPs are considered for the database. The perceptual judgment is considered as a ground truth for normal and hypernasal speech classification.

3. Pole-zero analysis of hypernasal speech

In the hypernasal speech, the vowel spectrum gets affected due to the addition of extra formant and anti-formant pairs [4]. The natural frequencies of the nasal tract and the sinuses inside the nasal tract decide the range of frequency in which addition of formants and anti-formants happens. The range of natural frequencies of nasal tract are in 450 to 650 Hz and 1800 to 2400 Hz [23] whereas it is around 400 Hz and 1300 Hz for the sinuses [24]. The extra formats give peaks in the spectrum and the anti-formants reduces the strength of vowel formats and shift them into the higher frequencies. The low-frequency nasal formants around 300 Hz and 1000 Hz have greater strength than the other formants in higher frequencies, hence these two formants in low-frequency region are generally used as the nasality evidence in hypernasal speech [4]. Fig. 1(a)-(c) shows the linear prediction (LP) spectrum of normal and hypernasal vowels /a/, /i/ and /u/ respectively. In the low-frequency region the nasality evidence in the form of extra format, reduction, and shifting of F_1 is shown in Fig. 1. The proper capturing of these low-frequency nasality evidence is required for hypernasality detection.

4. Effect of pitch on MFCC

For the extraction of MFCC feature from the speech signal first the framing the signal using overlapping Hamming/Hanning windows is done and then magnitude spectrum of each frame is computed using short-time Fourier transform (STFT). Next, Mel-scale warping of magnitude spectrum is done using triangular filters having nonuniform bandwidth. The discrete cosine transform (DCT) of the log-energies obtained as an output of

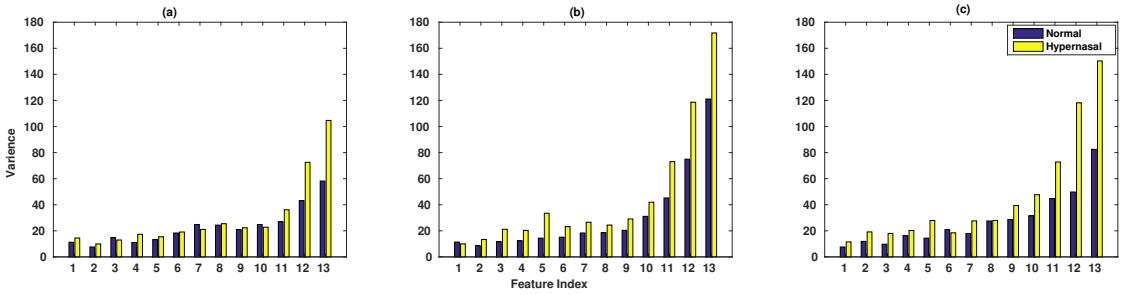

Figure 3: *Plot showing the variance (in bar) for each coefficients of 13-dimensional MFCC feature extracted for normal and hypernasal vowels from entire database. (a) for vowel /a/, (b) for vowel /i/ and (c) for vowel /u/ vowels*

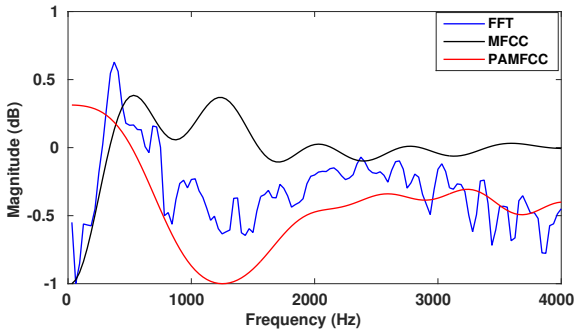

Figure 2: *Plots of smooth spectra corresponding to MFCC and PAMFCC feature along with STFT magnitude spectrum for the vowel /i/ of high pitch CP children speech*

the Mel-filterbank, gives the MFCC feature. The MFCC feature model the magnitude spectrum and gives the smooth spectral envelope representing the vocal tract characteristics. Hence it is expected that the feature is free from the pitch harmonics effect present in the magnitude spectrum. But the studies [19] [20] shows that for high pitch signals like children speech, the Mel-filter bank employed in the MFCC feature extraction is unable to sufficiently smooth out the pitch harmonics present in the magnitude spectrum and hence the MFCC feature get affected for high pitch signal. The smoothed spectral envelope corresponding to affected MFCC feature contains ripples in the lower frequency region. The ripples give the high variance for the higher coefficients of MFCC feature. Fig.2 shows the magnitude spectrum and the smoothed spectra corresponding to MFCC feature for vowel /i/ of CP children speech. The smooth spectra corresponding to the MFCC feature is derived by taking the 128-point inverse discrete cosine transform (IDCT) of 13-dimensional MFCC feature. Ripples in smoothed spectra corresponding to MFCC feature in the lower frequency region can be observed in Fig.2. The effect of ripples on the variance of the 13-dimensional MFCC feature for normal and hypernasal vowels /a/, /i/ and /u/ is shown in Fig.3 (a)-(c) in the form of bar plot. It can be observed from the bar plot that the variance in higher for higher coefficients (11-13 coefficients) of MFCC feature. Further, it can also be observed that the variance is more for the hypernasal vowels compared to the normal vowels. The reason may be high pitch perturbation in CP speech [13] which can be proved by measuring the mean and standard deviation of the pitch for all three vowels /a/, /i/ and /u/ from entire database. Table 1 shows the mean and standard deviation (std) of the pitch for normal and hypernasal vowels from the entire database. It can be observed that the standard deviation (std) in pitch (which shows the pitch variation) is high for both normal and hypernasal vowels, but it is higher for hypernasal vowels compared to the normal vowels. The pitch is measured using the method proposed in [25]. The ripples in low-frequency smooth spectra corresponding to the MFCC feature and high variance in higher coefficients of MFCC feature may affect the classification accuracy of the normal and hypernasal speech.

Table 1: *Mean and standard deviation (std) of pitch in normal and hypernasal vowels present in entire speech database*

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

5. Pitch-adaptive MFCC feature

The pitch-adaptive MFCC feature is originally proposed for robust children's automatic speech recognition (ASR) in [26]. The feature is free from the pitch harmonics effect and also deals with the high pitch perturbation in CP speech because pitch adaptive low time liftering of cepstral coefficient derived from the magnitude spectrum is done while computing the feature. The block diagram for the extraction of pitch-adaptive MFCC feature is shown in Fig. 4. The procedure for deriving the pitch-adaptive MFCC feature are as follows:

- Compute the log magnitude spectrum of each frame of the speech signal using the short time Fourier transform (STFT) with a fixed duration hamming window.
- Obtain the cepstral representation through the inverse discrete Fourier transform (IDFT) of the magnitude spectrum.
- Apply a pitch adaptive low time liftering on the cepstral representation because it retains the periodicity of the speech excitation. The pitch adaptive liftering smooth the pitch harmonics. Take the duration of low-time lifter $L = \frac{F_s}{F}$, where F_s is the sampling frequency and F is the average pitch value for the whole utterance. In this work, the pitch of the utterance is detected using the zero frequency filtered signal as proposed in [25].
- Take the discrete Fourier transform of liftered cepstrum to obtain the smoothed cepstral spectrum.

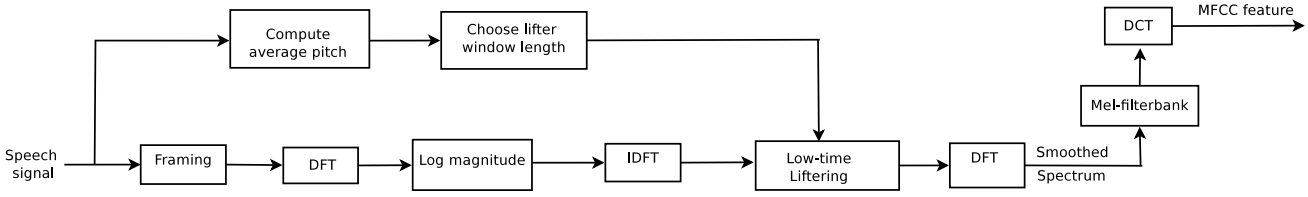

Figure 4: *Block diagram for the extraction of the pitch-adaptive MFCC feature by applying adaptive-liftering for spectral smoothening.*

- Employ Mel-filter bank on the smoothed cepstral spectrum and compute the log-energies for each filter.
- Take the discrete cosine transform (DCT) of the log-energies to find the cepstral coefficients.
- The lower coefficients are pitch-adaptive MFCC feature.

Fig.2 shows the smoothed spectra corresponding to PAMFCC feature. It can be observed from the Fig.2 that the low-frequency ripples are smoothed out in the PAMFCC features hence, the feature can capture the low-frequency nasality evidence in a better way and the variance for higher coefficients in the feature also get reduced. The detection of hypernasality using PAMFCC feature may give better classification accuracy.

6. Hypernasality detction using pitch adaptive MFCC feature

In this section the hypernasality detection is performed using the pitch-adaptive MFCC feature and the result is compared with the result got from the MFCC feature. The results are presented in terms of the overall accuracy, specificity and sensitivity.

6.1. Experimental setup

The 13-dimensional MFCC and PAMFCC features are extracted for each frame of speech. The frame size of $20ms$ and frame shift of $10ms$ is used for the framing of the speech. The SVM classifier with RBF kernel is used for the classification. The 5-fold cross validation of entire train database is done to find the optimum value of the kernel parameters c and γ. The training of the SVM Classifier is done with the 24 normal and 24 hypernasal children data and testing is done with the remaining 6 normal and 6 hypernasal children data for each vowel.

6.2. Result

Table 2: *Hypernasality detection accuracy for vowel* $/a/$

Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)
MFCC	77.75	78.64	77.49
PAMFCC	83.45	80.39	85.57

Table 2, Table 3 and Table 4 show the values of accuracy, sensitivity and specificity in percentage for vowels /a/, /i/ and /u/

Table 3: *Hypernasality detection accuracy for vowel* $/i/$

Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)
MFCC	84.21	82.15	86.80
PAMFCC	88.04	88.07	88.02

Table 4: *Hypernasality detection accuracy for vowel* $/u/$

Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)
MFCC	82.89	83.42	82.46
PAMFCC	85.58	82.80	87.91

respectively. The individual accuracies for MFCC and PAMFCC feature are shown in each tables. It can be observed that PAMFCC feature gives the accuracy of 83.45 %, 88.04 % and 85.58% for the vowels /a/, /i/ and /u/ respectively which is better than the accuracy obtained from the MFCC feature.

7. Summary and Future scope

In this work, a pitch-adaptive MFCC feature named as PAMFCC is used for hypernasality detection. This features is computed from the cepstral smoothed spectrum of magnitude spectrum instead of magnitude spectrum. A pitch adaptive way of choosing the size of low time liftering window is used which insures the cepstral smooth spectrum free from the pitch harmonics effect and also deals with the high pitch perturbation in CP speech. In comparison to MFCC feature, the PAMFCC feature capture the low-frequency nasality evidence present in hypernasal children speech in a better way. The feature when used for hypernasality detection using SVM classifier, gives the better accuracy comparision to the MFCC feature.

8. Acknowledgements

The authors would like to thank Prof. M. Pushpavathi and Prof. Ajish K. Abraham of AIISH Mysore for the mentoring and also sharing speech of CLP cases. This work is in part supported by the project grants, for the projects entitled NASOSPEECH: Development of Diagnostic system for Severity Assessment of the Disordered Speech funded by the Department of Biotechnology (DBT), Govt. of India and ARTICULATE +: A system for automated assessment and rehabilitation of persons with articulation disorders funded by the Ministry of Human Resource Development (MHRD), Govt. of India.

9. References

[1] D. Sell, A. Harding, and P. Grunwell, "A screening assessment of cleft palate speech (great ormond street speech assessment)," *International Journal of Language & Communication Disorders*, vol. 29, no. 1, pp. 1–15, 1994.

[2] 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.

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

[4] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 4, pp. 621–629, 2007.

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

[6] 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.

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

[8] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels," *J. Acoust. Soc. Am.*, vol. 77, no. 4, pp. 1560–1574, Apr 1985.

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

[10] 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.

[11] D. K. Rah, Y. I. Ko, C. Lee, and D. W. Kim, "A noninvasive estimation of hypernasality using a linear predictive model," *Annals of biomedical Engineering*, vol. 29, no. 7, pp. 587–594, 2001.

[12] J. R. Orozco-Arroyave, S. M. Rendón, A. M. Álvarez-Meza, J. D. Arias-Londoño, E. Delgado-Trejos, J. F. V. Bonilla, and C. G. Castellanos-Domínguez, "Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection." in *Interspeech*. Citeseer, 2011, pp. 529–532.

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

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

[15] 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.

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

[17] A. K. Dubey, S. M. Prasanna, and S. Dandapat, "Zero time windowing analysis of hypernasality in speech of cleft lip and palate children," in *IEEE Twenty Second National Conference on communication (NCC)*, 2016, pp. 1–6.

[18] A. Dubey, S. M. Prasanna, and S. Dandapat, "Zero time windowing based severity analysis of hypernasal speech," in *IEEE Region 10 Conference (TENCON)*, 2016, pp. 970–974.

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

[20] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

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

[22] AIISH, "All india institute of speech and hearing, mysore, india." [Online]. Available: web- site: http : //www.aiishmysore.in

[23] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

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

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

[26] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's asr." in *INTERSPEECH*, 2016, pp. 3459–3463.