Automatic detection of hypernasality using phase feature

Abstract

Hypernasality is a speech disorder in cleft palate (CP) children. Hypernasality detection is done by analyzing the vowels present in the speech stimuli. The presence of nasal peak in the vicinity of first formant F_1 is an important acoustic cue for the hypernasality detection. Generally, a two-step process is followed for hypernasality detection where first of all selection of vowel resign from stimuli is done using manually annotated points and then extraction of a feature from the selected vowel region and classification is done automatically. In this work, a method of hypernasality detection is proposed where the selection of vowel region, feature extraction and classification are done automatically. The method uses the vowel onset points (VOPs) for automatic selection of vowel region. The modified group delay feature (MODGDF) is then extracted from the vowel region. MODGDF is derived from the high resolution modified group delay spectrum which can resolve the closely spaced nasal peak and first formant F_1 and hence can capture the nasality evidence in a better way. When the feature is used for the hypernasality detection using support vector machine (SVM) classifier it gives an accuracy of 83.50 % and 93.21 % for /a/ and /i/ vowels respectively.

Index Terms: Hypernasality, vowel onset point, modified group delay spectrum, cleft palate.

1. Introduction

Hypernasality is a resonance disorder where excess nasal resonance is heard in the speech. This happens because extra nasal resonances are added in the vowels due to the coupling of nasal tract with the oral tract during the production of speech [1]. Hypernasality in cleft palate (CP)speech is due to the structural and functional abnormalities in velopharyngeal mechanism and oral cavity [2]. The repairing of structural abnormalities by the plastic surgeons may not lead to the normal speech. The nasal tract coupling happens in the repaired CP also due to the velopharyngeal insufficiency and mis-learning [3]. Hypernasality regards the intelligibility of CP speech. The information provided by the hypernasality evaluation are used by the plastic surgeons and the speech-language pathologists (SLPs) for the follow-up operation and speech therapy respectively. Mostly, the evaluation is done perceptually by the experienced SLPs. However, the perceptual judgment depends on the experience of SLPs and may be affected by the abnormalities in pitch, loudness, voice quality and/or articulation [4]. Hence the instrumental techniques are used to confirm the perceptual decision. The instrumental techniques like X-Ray (Cephalometry), videofluoroscopy, nasendoscopy etc. are direct techniques and techniques like accelerometry and nasometry etc. are indirect techniques. The direct techniques provide direct information about the velopharyngeal port movement whereas the indirect techniques provide information which can be used to understand velopharyngeal activity [5]. The direct techniques may be invasive and may have ionization radiation effect whereas the indirect techniques require addition good quality sensing device. The nasometer is device widely used for the hypernasality evaluation in clinics. But the device cannot be used for the prerecorded speech.

The researchers have also proposed some acoustic cues based on the digital signal processing of speech for hypernasality detection. The acoustic cues based hypernasality is done in two steps: first manually annotate the vowel region in the stimuli and then extract a feature which can effectively capture the nasality evidence from the vowel region. Generally, the stimuli for hypernasality detection consist of consonant-vowel-consonant-vowel (CVCV) word having plosive as a consonant and low vowel /a/ or high vowel /i/ as a vowel. The vowels of hypernasal speech get nasalizes due to nasal coupling, hence the acoustic cues for nasalized vowels are used for hypernasality detection. The important cues for nasalized vowels are the presence of extra-nasal peak in vicinity of first formant F_1, reduction in strength of F_1 and broadening of formants. Based on these cues features based on Teager energy operator [6], linear prediction cepstral coefficient (LPCC) [7], high spectral resolution modified group delay spectrum [4], energy distribution [8], [9], acoustic, noise and cepstral analysis, nonlinear dynamic and entropy feature [10], [11], [12], Teager energy operator combined with Mel frequency cepstral coefficient (MFCC) [13], and zero time windowing [14], [15] are proposed in the literature for hypernasality detection

In this work, an automatic method for the hypernasality detection is proposed. The method automatically detects the vowels region from the CVCV word and then extract a feature from detected vowel region. The automatic detection of vowel region is done using the vowel onset points (VOPs) in the stimuli. VOP is the instant at which the onset of vowel takes place. The VOPs are detected using the source, spectral peaks, and modulation spectrum energies as proposed in [16]. From each detected VOP, initial $30ms$ region of the vowel is left and next $100ms$ from the middle portion of the vowel is selected as an automatic detected vowel regions. The modified grope delay feature (MODGDF) proposed in [17] is then extracted from the detected vowel regions. The MODGDF feature is derived from the modified group delay spectrum by converting the spectrum into the cepstra using the discrete cosine transform (DCT). The modified group delay spectrum is a high-resolution spectrum due to the additive property of group delay and hence can resolve the closely spaced nasal peak and the F_1. The MODGDF is different from the earlier proposed group delay function based acoustic measure (GDAM) [4] which is based on the ratio of two dominant peaks corresponding to nasal and F_1 formants in modified group delay spectrum below a cutoff frequency. Since

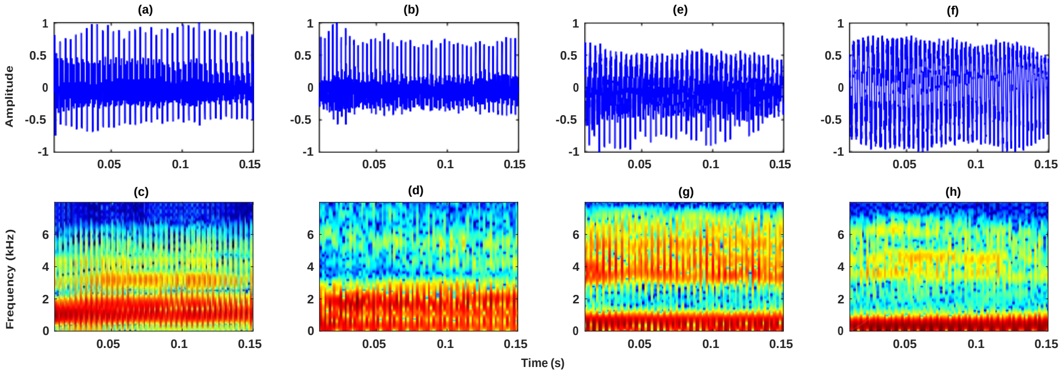

Figure 1: *Illustration of waveforms of speech signal and the corresponding spectrograms. (a)-(d) are the waveform of the normal vowel /a/, hypernasal vowel /a/, normal vowel /i/ and hypernasal vowel /i/ respectively. (e)-(f) are the their corresponding spectrogram*

the shifting of formant happens in hapernasal speech due to addition of extra pole-zero pair, hence exactly finding the peaks corresponding to nasal and F_1 formants may not be accurate because F_1 may shift to higher frequency greater than the cut-off frequency. Hence the MODGDF is used in this work. In the last, MODGDF feature extracted from the detected vowel region is used for hypernasality detection using support vector machine (SVM) classifier. This kind of automatic hypernasality detection method may be useful for the clinical purpose.

The rest of the paper is organized as follows. In Section 2, speech database used in this work is explained. The spectra analysis of hypernasal speech is given in section 3. The section 4 explains the automatic method of vowel region selection from the stimulus word. Section 5 gives the feature extraction method from the vowel region. Section 6 explains the whole automatic method of hypernasality detection with result. Finally section 7 contains the summary and conclusion of the work.

2. Speech database

The database used in this work is collected from two group of children: control normal group and repaired CP group . Each group consists of 30 children (18 boys and 12 are girls) in the age range of 7-12 years, so the age and gender distribution of both the group are kept same. All the children have Kannada as their native language. Kannada language is a Dravidian language spoken in Karnataka, a state located in the southern part of India. The children considered for this database is not having any other sensory and neurological impairment. The whole database is recorded in a sound-treated voice recording room of All Indian Institute of Speech and Hearing (AIISH), Mysore, India [18]. The recording is done using Bruel & Kjaer sound level meter (SLM) microphone placed at a distance of 15 cm from the child. The speech stimuli consist of non-meaningful CVCV words /papa/ and /pipi/. The database consists of total 271 normal, 232 CP /papa/ and 258 normal, 226 CP /pipi/ words. The manual annotation of VOP of each vowel of the stimuli is done using Wavesurfer tool [19]. The words are first uttered by the instructor and then repeated by the child. The child repaired sound is recorded. Each recorded sounds from each group is perceptually judged separately by three SLPs to rate it as normal or hypernasal speech. Only the recordings having common ratings from all three SLPs are included in the database. The speech recording is done at sampling frequency 44.1 kHz, 16 bps in .WAV format, which is down-samples at 16 kHz for the hypernasality detection.

3. Spectral analysis of hypernasal speech

The hypernasal speech is characterized by the presence of additional pole-zero pair in vowels due to the coupling of the nasal tract. The addition of pole-zero pairs happens at the natural frequency (resonance frequency) of the nasal tract and the sinuses attached to it. The resonance frequency of the nasal tract lies in the ranges of 450 to 650 Hz and 1800 to 2400 Hz [20] whereas the natural frequency of sinuses lies in the ranges of around 400 Hz or around 1300 Hz [21]. The lower frequency poles around 400 Hz and in the range of 450 to 650 Hz below 1000 Hz mainly characterize the hypernasal speech because they are of high strength. The poles give nasal formant in the vowel spectrum and the zeros are responsible for the reduction in strength of vowel formants, mainly first formant F_1. To show the effect of nasal pole-zero pairs the speech waveform and corresponding spectrograms of normal and hypernasal vowels /a/ and /i/ are compared in Fig. 3 (a)-(f). It can be observed from Fig. 3 (c) that the spectral energy in normal vowel /a/ is mainly in the frequency range of around 700 Hz to 2000 Hz. This is due to its first two formants in this range. The energy in lower frequency below 500 Hz is low. But the energy gets shifted to the lower frequency below 500 Hz also for hypernasal vowel /a/ due to the addition of nasal pole at the frequency around 400 Hz as shown in Fig. 3 (d). The presence of zeros around the F_1 can also be observed in Fig. 3 (d). For the vowel /i/ the F_1 is around 300 Hz so the energy is mainly in lower frequency as shown in Fig. 3 (g). The energy in F_1 region further increases in hypernasal vowel /i/ due to nasal formants which can be observed in Fig. 3 (f).

4. Selection of vowel region in CVCV word

In this section, the automatic method of vowel region selection from the stimuli words /papa/ and /pipi/ is explained. The method is based on VOPs detection, which is the instant at which the onset of vowel takes place. Here VOP detection is done using the source, spectral peaks, and modulation spectrum energies as proposed in [16]. These energies are associated with the different aspect of speech production, and hence contains complementary information about the VOP. The energy of the excitation source gives the changes in its energy levels, the energy of spectral peaks gives the changes in the vo-

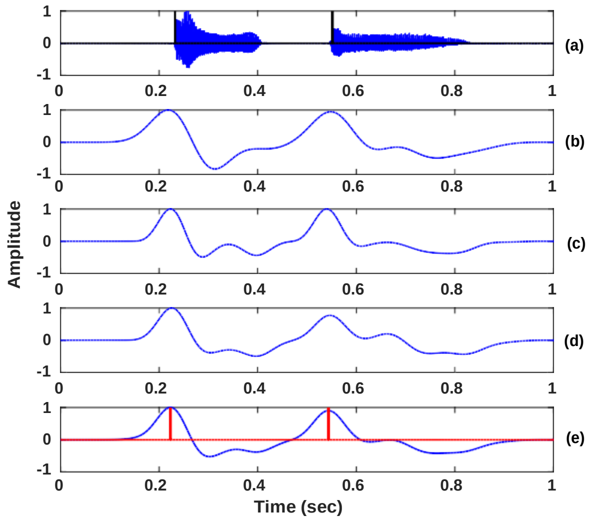

Figure 2: *Plots of smooth spectra corresponding to MFCC and PAMFCC feature along with STFT magnitude spectrum for the vowel /i/ of high pitch CP children speech*

cal tract transfer function and the energy of modulation spectrum gives the changes in the slowly varying temporal envelope. The VOP detection method can be explained with help of Fig. 2 which shows the manually annotated point along with the speech waveform of word /papa/ in Fig. 2 (a) and VOP evidence plots for spectral peaks, modulation spectrum and excitation source enrgies are shown in Fig. 2(b)-(d), respectively. The change in the nature of VOP evidence plots indicate different VOP evidences available in each case. The combined VOP evidence plots with the detected VOPs are shown in Fig. 2 (e). From the Fig. 2, it can be observed that the automatic VOP detection method gives detected VOP very close to the manually annotated VOP.The selection of vowel region is now done using automatically detected VOP. From each detected VOP, initial $30ms$ region of the vowel is left and next $100ms$ from the middle portion of the vowel is selected as an automatic detected vowel regions for feature extraction. Table 1 and Table 2 shows the VOP detection accuracy, miss alarm rate (MAR) and false alarm rate (FAR) for normal and hypernasal vowels /a/ and /i/ respectively with the tolerance window 30 ms evaluated on entire database. The accuracy lies between 96.97 % to 98.89 %.

Table 1: *Vowel onset point (VOP) detection accuracy for vowel /a/ with tolerance window of 30 ms*

Speech type	Accuracy (%)	MAR (%)	FAR (%)
Normal	98.89	1.11	0.92
Hypernasal	96.97	3.03	3.04

5. Modified group delay feature for hypernasality detection

Table 2: *Vowel onset point (VOP) detection accuracy for vowel /i/ with tolerance window of 30 ms*

Speech type	Accuracy (%)	MAR (%)	FAR (%)
Normal	97.93	2.07	3.09
Hypernasal	96.93	3.37	5.04

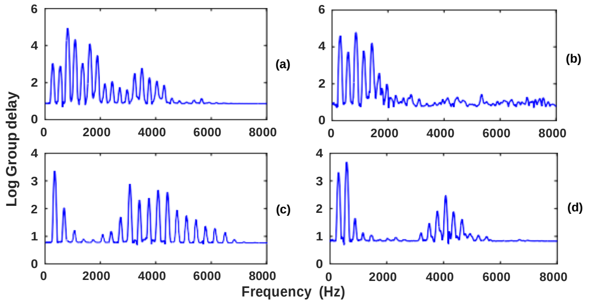

Figure 3: *Plots of smooth spectra corresponding to MFCC and PAMFCC feature along with STFT magnitude spectrum for the vowel /i/ of high pitch CP children speech*

The hypernasal vowel spectrum contains the nasal formants in vicinity of F_1 as explained in section 3. To capture the nasality evidence in hypernasal vowels a high resolution spectrum is need which can resolve the two closely spaced nasal and vowel formats. In literature modified group delay spectrum (MODGDS)has been used to resolve the nasal and first formant in [4]. The high resolution of MODGDS is due to additive property of the group delay function [17]. Hence a parameter called group delay function based acoustic measure (GDAM) extracted from MODGDFS of band-limited speech signal is used in [4] for the hypernasality detection. The parameter is the ratio of strength of nasal peak to the strength of F_1 peak. The peaks are chosen as two most prominent peaks below a cutoff frequency. But, for automatic method of hypernasality detection the choice of cutoff frequency may create problem because due to addition of pole zero pair sometimes the F_1 may shift to the higher frequency than the cutoff frequency. Further the detection of nasal peak for mild hypernasal speech signal may not be prominent. To avoid these problem the cepstral coefficients are extracted from the MODGDS to capture its envelop. The feature is called modified group delay feature (MODGDF) and can be extracted automatically .

The modified group delay spectrum of the speech signal $x(n)$ is given by

$$\tau_m(\omega) = (\frac{\tau(\omega)}{|\tau(\omega)|})(|\tau(\omega)|)^{\alpha} \qquad (1)$$

where

$$\tau(\omega) = (\frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}) \qquad (2)$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the N -point DFT of the signal $x(n)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ is the N -point DFT of the signal $y(n) = nx(n)$. $S(\omega)$ is the cepstrally smoothed version of $|X(\omega)|$ The value of parameters α and γ, vary from 0 to 1, are taken 0.6 and 0.9 respectively. The cepsrtrum coefficients are extracted from the logarithmic modified group delay spectrum using the discrete cosine transform (DCT II) as

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k)cos(n(2k+1)\frac{\pi}{N_f}) \qquad (3)$$

where N_f is the discrete Fourier transform (DFT) order and $\tau_m(k)$ is the modified group delay spectrum. The MODGDF is earlier used for the speech recognition in [22]

6. Automatic detection of hypernasality using Modified group delay feature

The automatic detection of hypernasality detection proposed in this work consist of two stages. In first stage the automatic selection of a portion of vowel is done by the method explained in section 4 and in the second stage the MODGDF feature as explained in section 5 is extracted from the selected vowel region and support vector machine (SVM) classifier is trained with this feature. So in this section experiment setup for feature extraction and SVM training is explained along with the classification result. The result is presented in terms of the overall accuracy, specificity and sensitivity and is compared with the result of MFCC feature.

6.1. Experimental setup

The 13-dimensional MODGDF and MFCC features are extracted for each frame of automatic selected vowel portion. The framing of speech signal is done with the frame window of size of $20ms$ and frame shift of $10ms$. The SVM classifier with RBF kernel is used for the classification. The 5-fold cross validation of entire train database is dome 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 3: *Hypernasality detection accuracy for vowel /a/ using MODGDF*

Feature	Accuracy (%)	Sensitivity ( )	Specificity (%)
MFCC	79.67	74.80	83.10
MODGDF	83.50	74.14	90.08

Table 4: *Hypernasality detection accuracy for vowel /i/ using MODGDF*

Feature	Accuracy (%)	Sensitivity ( )	Specificity (%)
MFCC	87.80	84.99	91.05
MODGDF	93.21	92.74	93.75

Table 3 and Table 4 show the values of accuracy, sensitivity and specificity in percentage for vowels /a/ and /i/ respectively. The individual accuracies for MODGDF and MFCC features are shown in each tables. The MODGDF gives an accuracy of 83.50 % for vowel /a/ and 93.21 % for vowel /i/. The accuracy of MODGDF is higher than the accuracy of MFCC feature for both the vowels.

7. Summary and Future scope

In this work a method of automatic detection of hypernasality is proposed for the vowels /a/ and /i/. The method first automatically detects the vowels region from the stimuli word using the VOPs. For the select the vowel region automatically beginning from each detected VOP, initial $30ms$ region of the vowel is left and next $100ms$ from the middle portion is taken. The feature MODGDF is now extracted from the automatically detected vowel region. Feature is derived from the high-resolution modified group delay spectrum which can resolve the nasal peak and F_1. Hence the feature captures the nasality evidence resent in the hypernasal speech. At the last, method uses the support vector machine (SVM) classifier for the hypernasality detection using the feature MODGDF. As a part of future work, the vowel end point (VEP) can also be used for robust selection of vowel region.

8. References

[1] 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.

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

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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-Dom´ınguez, “Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection.” in *Interspeech*. Citeseer, 2011, pp. 529–532.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, “Vowel onset point detection using source, spectral peaks, and modulation spectrum energies,” *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[17] H. A. Murthy and V. Gadde, “The modified group delay function and its application to phoneme recognition,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–68.

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

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

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

[21] 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.

[22] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, “Significance of the modified group delay feature in speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.