6th International Conference on Computer and Knowledge Engineering (ICCKE 2016), October 20-21 2016, Ferdowsi University of Mashhad

Detection of Hypernasality from Speech Signal Using Group Delay and Wavelet Transform

Atefeh Mirzaei, Mansour Vali
Department of Electrical and Computer Engineering
K.N Toosi University of Technology
Tehran, Iran
a.mirzaei@ee.kntu.ac.ir, mansour.vali@eetd.kntu.ac.ir

Abstract— **One of the most common disorders in children with cleft palate is hypernasality that survives also after operation. To solve this problem, it is required to set many speech therapy sessions. Therefore, assessment of hypernasality is fundamental for speech therapists and could be done either by a nasometer equipment or an expert speech therapist. Recently speech processing methods are introduced as an efficient alternative tool. In this study, vowels (/a/) extracted from 392 utterances of disyllables (/pamap/) that were uttered by 22 normal subjects and 13 subjects with cleft palate have been used and are recorded by nasal and oral microphones. Some analyses are performed on Group Delay parameters as well as features of wavelet transform. The results show that extracted parameters from Group Delay spectrum of second (/a/) in (/pamap/) context, obtained from both nasal and oral signals, are better than that of the first (/a/), and in the best outcomes an accuracy of 94.1% is achieved. In wavelet transform, statistical features are calculated from 5 sub-bands of Daubechies4 coefficients of two (/a/) vowels and their transients. In the best results an accuracy of 97.1% for transient (/ma/) from combination of nasal and oral features is obtained.**

Keywords: hypernasality, Group Delay, Wavelet Transform

INTRODUCTION

Hypernasality occurs frequently in children with cleft palate due to excessive nasal resonance. The term cleft palate refers to a malformation affecting the soft and/or hard palate, and is usually congenital. This anatomical defect may result in speech with reduced quality of speech, making the speech unintelligible. Hence, an assessment of hypernasality is important to decide whether to take a surgical intervention and/or the kind of speech therapy to be provided. Approaches for the assessment of hypernasality classified into two categories of invasive and non-invasive techniques.

Nasendoscopy and videofluoroscopy are two invasive intrusions that may cause pain and discomfort to the patients, especially to young children. Noninvasive assessment of hypernasality is accomplished with the nasometer. The nasometer is a PC-based device that determines the percentage of "Nasalance"; which is defined as the ratio of nasal acoustic energy to oral-plus-nasal acoustic energy. This is measured during speech production, and is used clinically for the assessment of hypernasality. This method uses a head set with a plate containing microphones attached to the top and bottom of the plate. The nasometer is widely used in the clinical environment, but it is an expensive device.

hypernasality perception can also be done by speech therapists. The obtained results by an expert trained speech therapist may be even more accurate than the results of nasometer [1]. Signal processing-based technique is the other noninvasive approach that has been used for detecting hypernasality in recent years. By analyzing speech signal in time domain, it has been reported that longer nasalization durations in children with cleft palate in comparison to children without cleft palate, show the delayed or deviant temporal patterns in cases with cleft palate [2]. Frequency analysis of Hawkins et al. [3] and Glass et al. [4] shows that the main features of nasalization are changes in the low-frequency regions of the speech spectrum, where there is a very low-frequency peak with wide bandwidth along with the presence of a pole-zero pair due to the acoustic coupling. Researchers claimed that nasalization increases the first formant bandwidth and intensity and also introduce nasal formants and anti-formants [5]. They compared the output of a low pass filter with cut-off frequency between the first and second formant and a band pass filter that just filters the first formant that both applied to speech samples and found a distinctive difference for nasalized vowels, whereas the normal vowels do not show any remarkable difference. Vijayalakshmi et al. found that introduction of nasal resonances around 250 Hz plays a significant role in the nasalization of vowels. They carried out an acoustic analysis on hypernasal speech, produced by cleft palate and lip speakers, using Group Delay function. It was mentioned that LP-based formant extraction technique cannot be used for the detection of hypernasality due to the vulnerability to the prediction order [6]. In our previous work it was claimed that since an autoregressive (AR) model for frequency response of the vocal tract system of these patients is not accurate, the hypernasality was estimated by comparing the distance between the sequences of cepstrum coefficients of AR model and cepstrum coefficients of autoregressive moving average

978-1-5090-3586-1/16/$31.00 ©2016 IEEE

(ARMA) model [7]. It has been proved that when the formants are very close as in hypernasal speech, Group Delay spectrum, estimates the locations of the formants better than power density spectrum [8]. In present research we used modified Group Delay function and wavelet transform in addition to energy ratio of oral and nasal speech, to extract the best features for hypernasality detection. In the next sections, modified Group Delay and wavelet transform will be discussed.

1. MATERIALS AND METHODS

A. Group Delay

Formants properties consist of information about speech sounds. Identification of formants requires estimating the frequency response or transferring function of articulation. In the past, researchers tried to estimate the articulation parameters using Linear Predictive Analysis [9], but this model was unable to detect formant with short amplitude near to a formant with high amplitude. The Group delay has a higher resolving compared to the magnitude spectrum or LPC analysis. We compared the Group Delay (GD) spectrum with Power Spectrum Density (PSD) of original signal to prove that GD could distinct the peaks of pitch harmonics better. The results for a normal and a hyper-nasal speech are shown in Fig.1 and Fig.2, respectively.

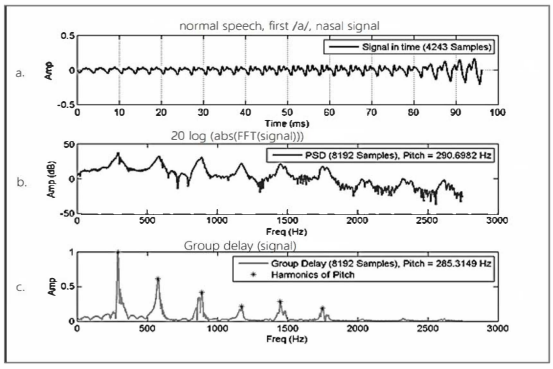

Fig.1*: detection of 5 initial peaks of vowel /a/, a) Time domain signal, b) PSD spectrum and c) GD spectrum, of normal speech.*

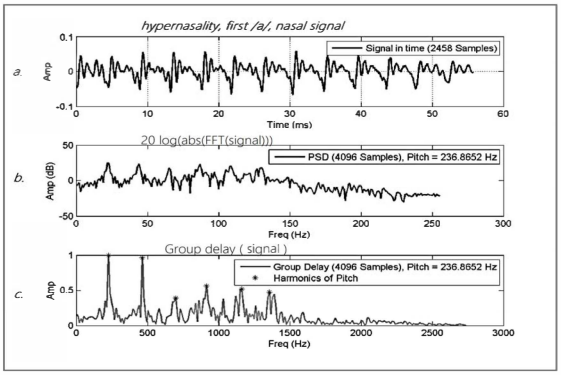

Fig.2*: detection of 5 initial peaks of vowel /a/, a) Time domain signal, b) PSD spectrum and c) Group Delay spectrum of hyper-nasal speech.*

Group Delay function can computed from the speech signal [10]:

$$\tau_x(\omega) = - \mathrm{Im}(d(\log(X(K))) / d\omega) \quad (1)$$

$$\tau_x(\omega) = |(X_R(K).Y_R(K)+X_I(K).Y_I(K)) / S(K)^{2\gamma}|^{\alpha} \quad (2)$$

where X(k) is Fourier Transform of speech signal x(n) and Y(k) is Fourier Transform of nx(n). Subscripts R and I denote the real and imaginary parts of Fourier Transform, respectively. S(k) is the smoothed version of |X(k)|. The introduced α and $\boldsymbol{\gamma}$ vary from 0 to 1 where $0< \alpha <1$ and $0< \boldsymbol{\gamma} <1$ [11].

B. Discrete Wavelet Transform

Discrete wavelet transform (DWT) is an efficient tool for speech analysis. It involves filtering and down-sampling the input signal at each decomposition level. In fact, DWT is the approach based on filter banks and decomposition of main signal at different frequencies. High pass and low pass filter are used in the first level of decomposition, that the outputs of the low pass filter are approximation coefficients (cA) and the outputs of high pass filter are detailed coefficients (cD) [16,17]. First level of decomposition of wavelet is shown in Fig.3. If the sampling frequency is f_s, then cA and cD for the first level of decomposition consist of frequency range of 0-f_s/4 and f_s/4-f_s/2, respectively. Continuing the decomposition levels, it can be revealed narrower frequency bands.

C. Doubechies wavelet

Higher-order Doubechies family wavelets can be used to improve frequency resolution and reduced aliasing introduced in each level of decomposition. Doubechies families are usually written by dbN, where N is the order and 2N is the impulse filter length [16,17]. The first level of decomposition is shown in Fig.3.

D. Energy ratio

By analyzing the data of this article with cool edit pro software, we found that, output energy of nasal signal is more than that of oral signal in Hyper-nasal speech in comparison to a normal speech. Therefore, we decided to calculate the energy of nasal signal to energy of oral signal plus energy of nasal signal ratio of a person as below:

$$E_{pow2}=[\sum_n (X_{nasal}(n))^2] / [\sum_n (X_{oral}(n))^2 + \sum_n (X_{nasal}(n))^2], \quad n=1,\ldots,N \quad (3)$$

Where X_{oral} and X_{nasal} are the output signals of oral and nasal cavities that are recorded by two separate microphones. N is the number of signal samples.

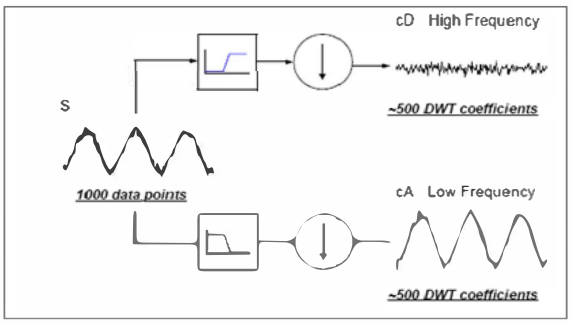

Fig.3*: first level of decomposition in discrete wavelet transform*

II. SPEECH DATA

Oral consonants require velopharyngeal closure for the separation of the oral and nasal cavities. In contrast, nasal consonants involve velopharyngeal opening that allows the. propagation of the sound energy into the nasal cavities[13]. In children with cleft palate early onset and delayed offset of velar movement occurs before and after the oral cavity occlusion, causes the vowel preceding and following nasal consonants to be nasalized [13]. Therefore, in this study, context /pamap/ was uttered by 22 normal subjects and 13 subjects with cleft palate. Then vowels /a/ and four transition parts, /pa/, /am/, /ma/ and /ap/ were extracted from 392 utterances and 196 utterances, respectively. We used cool edit pro.2.1 to separate the vowels and transitions as well. A series of /p/ and /m/ before and after vowels in the text /pamap/ requires velopharyngeal closing and opening movements. Because both, oral phoneme /p/ and nasal phoneme /m/ were produced at labial place of articulation, the influence of the change of articulation position on nasal resonance could be controlled [14]. The age range of the subjects for this study was 4-12 years. Children with cleft palate had the palate repaired through primary surgical correction and also they exhibited moderate or severe hyper-nasality. The sampling rate was 16 kHz with 16 bits of resolution.

III. FEATURE EXTRACTION

We compared all the results of (3) for both normal and Hyper-nasal speech and came to the conclusion that, it would be a proper feature to consider in feature vector. Therefore, energy ratio, E_{pow2}, was calculated for both the vowels and the transition parts. On the other hand, as mentioned, a characteristic of hyper-nasal speech is the nasal formant that would be specified better in Group delay spectrum in comparison to the frequency spectrum [8]. In addition, by comparing the spectrum of children in this study, with that for adult in previous study [2], we found that pitch harmonics of children's speech spectrum are existed far from each other, so that we couldn't estimate the nasal formant for the data in this study. So, we assumed amplitude values of the first five harmonics in the GD spectrum, then we applied (4)-(6) to these 5 selected values. It should be noticed that, these five values covered approximately, the band frequency of 2 kHz, where the nasal formant was estimated by previous studies [2].

The equations mentioned above, are calculated as below:

$$P_1=\text{peak}(i) / \text{peak}(i-1) , i=2,\ldots,5 \quad (4)$$

$$P_1=\text{peak}(i) / (\text{peak}(i-1)+\text{peak}(i)) , i=2,\ldots,5 \quad (5)$$

$$P_1=\text{peak}(i) / \sum_i \text{peak}(i) , i=1,\ldots,5 \quad (6)$$

Where, i denote the number of peaks of GD spectrum and peak(i) indicate the absolute value of GD for the i^{-th} peak.

The feature vector of GD was calculated by using both oral and nasal signals for each data. As can be inferred from (4)-(5), for indexes p_1 and p_2, a feature vector which is of dimension 8 can be calculated, which will apparently consist of both oral and nasal features accordingly. Additionally, the same condition was considered for index p_2 and therefore, a feature vector of dimension 10 was achieved as a result.

In addition, the context, /pamap/, contains 4 transition parts as: /pa/, /am/, /ma/ and /ap/. We assumed that these transition parts would carry information of Hyper-nasality. After studying, we applied Wavelet Transform, which could detect dynamics changes better. The higher-order Daubechies family wavelets can be used to improve frequency resolution and reduce the aliasing introduced in each decomposing level. Daubechies'-4 with the composition level of 4, and 5 sub-bands, was used in this study. Then, we calculated four statistic features as: Variance, zero-crossing rate, kurtosis and skewness, for each 5 sub-bands.

IV. CLASSIFICATION

Support vector machines (SVM) were used to classify two groups of normal and hyper-nasal speech, in this study. It was necessary to find an optimal line to separate two groups, so we used RBF-kernel with gamma value of 3 which was selected after analyzing the other kernels. Also, for evaluating the classifier results', we employed Leave-one-out cross-validation, so that, we used Leave-one-out cross-validation for all 392 and 196 utterances for vowels and transitions, respectively. In our study, as mentioned before, we consider all children's repetition of saying /pamap/. So, once, SVM was applied to all the utterances, and then, the accuracy was calculated for all the utterances. This accuracy couldn't help to detect if subject's speech is normal or hyper-nasal. So, after detecting the class of each utterances, we estimated the class of each subject, so that if half and more than half of utterances, belonged to first class (hyper-nasality), that subject was known as hyper-nasality, otherwise it would be known as the subject with normal speech. After that, we calculated classification accuracy of 35 subjects. To achieve accurate results, at each repetition of classifier on 35 subjects, one out of 35 subjects with all of its utterances were considered as test data, while the other remained subjects were included as train data.

Table 1. *Classification accuracy of energy ratio feature for utterances and subjects.*

	Accuracy for utterances (%)	Accuracy for subjects (%)
first /a/	54.4	57.2
second /a/	60.6	61.4
Transition /pa/	50.9	60.7
Transition /am/	64.2	68.6
Transition /ma/	70.1	77.5
Transition /ap/	59.8	61.4

Table 2. *Classification accuracy of nasal and oral features extracted from Group Delay spectrum for utterances and subjects.*

	Accuracy for utterances (%)	Accuracy for subjects (%)
p1 features for first /a/	67	75.5
p1 features for second /a/	71	76.3
P2 features for first /a/	72	78.5
P2 features for second /a/	75	80.3
P3 features for first /a/	76	82.8
P3 features for second /a/	84	94.1

Table 3. C*lassification accuracy of nasal and oral features extracted from wavelet coefficients for utterances and subjects.*

	Accuracy for utterances (%)	Accuracy for subjects (%)
first /a/	87.8	90.2
second /a/	89.6	91.4
Transition /pa/	81.3	85.7
Transition /am/	87.0	92.5
Transition /ma/	90.4	97.1
Transition /ap/	75.1	78.3

Table 4. C*lassification accuracy of combination of the best features for utterances and subjects*

	Accuracy for utterances (%)	Accuracy for subjects (%)
Wavelet+energy	93.1	94.1
GD+energy	87.2	94.2
GD+wavelet	92.3	97.8
GD+wavelet+energy	96.9	98.2

V. Results and Discussion

From tables 1-3 it is obvious that for the vowels /a/ of the context, /pamap/, we achieved classification accuracy of 89.6% at second vowel /a/ (the vowel placed after /m/ in the context of /pamap/) and the maximum classification accuracy of 90.4% for the transition part, /ma/, by using wavelet features for utterances as result. As can be seen from tables 1-3, we calculated the classification accuracy for both utterances and subjects. Generally we should know if the subject suffer from hyper-nasal speech? So, we achieved the accuracy of 77.5% for the transition part /ma/ by using energy ratio feature, the accuracy of 94.1% for the second vowel /a/ in the context /pamap/, by using introduced index p_3 of Group delay feature, and also the accuracy of 97.1% for transition part /ma/ by using wavelet transform feature as the best results.

The second /a/ of the context, /pamap/, as an oral vowel requires both application of opening oral cavity and closing nasal cavity, is just pronounced after a nasal phoneme /m/, where oral cavity needs to be closed. This changing state should be done quickly without any delay, while children with cleft palate show delayed temporal pattern [2]. Generally, as tables 1-3 show, second /a/ after /m/ in the context, /pamap/, and also transition part /ma/ carry more information of hyper-nasality. After finding the best accuracy for each feature, we decided to combine them together to achieve high accuracy. Table 4 show the results of classifying best features obtained from tables 1-3 for both utterances and subjects. Although we tested all different combination of features, we found out that the combination of energy ratio feature, wavelet features of transition part, /ma/ with p_3 index features of Group Delay for second /a/ in context /pamap/, would have high accuracy, so that we achieved the accuracy of 98.2% for subjects as result.

VI. Conclusion

In previous studies, researchers tried to detect hyper-nasality by analyzing the speech of subjects with cleft palate or lip, synthesized hyper-nasal speech or nasalized vowels of normal speech. In signal processing-based techniques for hyper-nasality detection, the assessment is usually carried out by finding the deviation of the spectrum of hyper-nasal speech from the normal speech [6]. In this study, Group Delay function, wavelet transform coefficients and energy ratio were used to extract the best features of detecting hyper-nasality in children with cleft palate. Speech data was 392 utterances consisted of disyllables (/pamap/) that uttered by 22 normal subjects and 13 subjects with cleft palate that were recorded by oral and nasal microphones. The results show that extracted parameters from Group Delay spectrum of second (/a/) in the context, obtained from both nasal and oral signals, are better than that of the first (/a/) and in the best outcomes an accuracy of 94.1% is achieved. In wavelet transform, statistical features are calculated from 5 sub-bands of Daubechies4 coefficients of two (/a/) vowels and their transients and in the best results an accuracy of 97.1% for transient (/ma/) from combination of nasal and oral features is obtained. By combining the best features of energy ratio, wavelet features and Group Delay, accuracy of 98.2% for the subjects was achieved.

Acknowledgment

The authors would like to thank Dr. Negin Moradi, faculty member at Ahvaz Jundishapur University of Medical Sciences for cooperation in data gathering.

REFERENCES

[1] F. R. Larangeria, J. Cassia Rillo Dutka, M. E. Whitaker, O. M. Vieira de Souza, J. R. P. Lauris, M. J. Da Silva and M. I. Pegoraro -Krook, "Speech nasality and nasometry in cleft lip and palate," Brazilian Journal of Ortorhinolaryngology, In Press, 2015.

[2] K. Baghban, F. Torabizade, N. Moradi, F. Asadollahpour, N. Ahmadi and N. Mardani, "Temporal characteristics of nasalization in Persian speaker children with and without cleft palate," Int. Journal of Pediatric Otorhinolaryngology, vol. 79, no. 4, pp. 546-552, 2015.

[3] S. Stevens and K. N. Hawkins, "Acoustic and perceptual correlates of the non-nasalnasal," Journal of Acoustic Society of America, vol. 77, no. 4, pp. 1560-1575, 1985.

[4] J. Zue and, V. Glass., "Detection of nasalized vowels in American English," IEEE International Conference on ICASSP'85, vol. 10, pp. 1569–1572, 1985.

[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," International Journal of Pediatric Otorhinolaryngology, vol. 76, no. 3, pp. 362–369, 2012.

[6] P. Vijayalakshmi, T. Nagarajan, and V. Jayanthan Ra, "Selective pole modification based technique for the analysis and detection of hypernasality," TENCON IEEE Conferenc, pp. 1-5, 2009.

[7] E. Akafi, M. Vali, N. Moradi and K. Baghban, "Assessment of Hypernasality for Children with Cleft Palate Based on Cepstrum Analysis," Journal of Medical Signals and Sensors, vol. 3, no. 4,pp. 209-215, 2013.

[8] P. Vijalayakshmi, M. Ramasubba Reddy and D. O'Shaughnessy, "Acoustic analysis and detection of Hypernasality using a group delay function," IEEE Transaction on Biomedical Engineering, vol. 54, no. 4, 2007.

[9] R. Hedge, H. A. Murthy and V. R. Rao Gadde, " Significants of the modified group delay features in speech recognition," IEEE Transaction on Audio, Speech and Language Processing, vol. 15, no. 1, pp. 190-202, 2007.

[10] H. A. Murthy, B. Yegnanarayana, "Group delay function and it's applications in speech technology," Indian academy of sciences, vol.36, no. 5, pp. 745-782, 2011.

[11] 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.

[12] M. Y. Chen, "Acoustic parameters of nasalized vowel in hearing impaired and normal hearing speakers," Journal of Acoustical Society of America, vol. 98, no. 5, pp. 2443-2453, 1995.

[13] D. Telejod, S. Sepulveda, C. Dominguez and G. Robustness "Improvement of hypernasal speech detection by acoustic analysis and the redemcher complexity model," Journal of Acoustical Society of America , vol. 80, no. 5, pp. 159-162, 2009.

[14] 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," Cleft Palate Craniofac Journal, vol. 4, no. 5, pp. 535-543, 2004.

[15] D. A. Cairns, J. H. L. 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, pp. 35-45, 1996.

[16] B. Kim, H. Jeong, H. Kim, B. Han, "Exploring wavelet application in civil engineering", Springer, KSCE Journal of Civil engineering, Vol. 00, No.0, pp. 1-11, 2016.

[17] A. Sharma, "Efficient use of biorthogonal wavelet transform for caridac signals", IJCSNS international journal of Computer science and network security, Vol. 16, No. 2, 2016.