THE ANALYSIS ON BAND-LIMITED HYPERNASAL SPEECH USING GROUP DELAY BASED FORMANT EXTRACTION TECHNIQUE

P. Vijayalakshmi and M. Ramasubba Reddy

Biomedical Engineering Division, Dept. of Applied Mechanics
Indian Institute of Technology, Madras
pvijayalakshmi@iitm.ac.in, rsreddy@iitm.ac.in

10.21437/Interspeech.2005-192

Abstract

Speakers with defective velopharyngeal mechanism, produce speech with inappropriate nasal resonances across vowel sounds. The acoustic analysis on hypernasal speech and nasalized vowels of normal speech shows that there is an additional frequency introduced in the low frequency region close to the first formant frequency [1]. The conventional formant extraction techniques may fail to resolve closely spaced formants. In this paper, an attempt is made to use the group delay based algorithm [2] for the extraction of formant frequencies from hypernasal speech. Preliminary experiments on synthetic signal with closely spaced formants show that the formants are better resolved in group delay spectrum when compared to conventional methods. But when formants are too close with wider bandwidths, the group delay algorithm also fails to resolve prominently. This is primarily because of the influence of the other resonances in the signal. To extract the additional frequency close to the first formant, the speech signal is low-pass filtered and the formants are extracted using group delay function. Following the satisfactory results on synthetic signal, the above technique is used to extract formants from phonations /a/, /i/, and /u/ uttered by 15 speakers with cleft palate who are expected to produce hypernasal speech. Invariably in all the tests, an additional nasal resonance around 250 Hz and first formant frequency of vowels are resolved properly.

1. Introduction

Speakers with velopharyngeal incompetence produce hypernasal speech across voiced elements [3]. Many researchers have revealed the fact that, hypernasal speech is characterized by increased bandwidth of first formant, intensity decrease of first formant, and appearance of nasal formants and anti-formants. **Glass and Zue** [4] have shown that, the short-time spectra of nasalized vowels often exhibit extra nasal formants or a broadening of the vowel formants typically in the first formant region. Frequency analysis of **Hawkins and stevens** [5] strengthens the above results. They have shown 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 pole-zero pair due to acoustic coupling. The acoustic analysis of hypernasal speech and nasalized vowels by **Vijayalakshmi and Reddy** [1] reveals that, invariably there exists a nasal formant in the low frequency region around 250 Hz for all the phonations /a/, /i/, and /u/.

From the literature, it is noted that the formants are extracted using the conventional methods like Magnitude spectrum derived from Fourier transform, Linear prediction, homomorphic deconvolution technique, etc., from the hypernasal speech. The major advantage of magnitude spectrum, derived from Fourier Transform (FT) is that the underlying characteristics (number of poles and zeros) are not modified. But this cannot be directly used for formant extraction because of the presence of the pitch harmonics. The magnitude spectrum derived from FT of the given signal can be smoothed using homomorphic deconvolution, in other words cepstrum based smoothing technique. But the disadvantage here is due to the smaller size of the cepstral lifter used to smooth the magnitude spectrum. The frequency resolution of the resultant smoothed spectrum is directly proportional to the size of the cepstral lifter. Eventhough, the resultant magnitude spectrum is a smoothed version of the original magnitude spectrum, it may not be able to resolve two closely spaced formant frequencies as in the case of hypernasal speech. Another conventional method is the Linear Prediction (LP) based formant extraction. The major disadvantage of this technique is the vulnerability to the prediction order.

The above mentioned conventional methods of formant extraction can be best utilized for the normal speech as the formants are farther apart. If two formant frequencies are closely spaced then the conventional methods do not resolve the corresponding frequencies because of the following two reasons, (i) poor frequency resolution and (ii) influence of adjacent poles.

The problem of poor frequency resolution can be overcome by a technique called group delay based formant extraction method as explained in [2]. In [2], it is argued that because of the additive property of the group delay

function, the closely spaced formants can be discriminated. But due to the influence of the adjacent poles, the closely spaced formants may not be discriminated very prominently. To avoid the influence of the less important formants on interested spectral region (low-frequency), in the present study, the signal is band-limited by low-pass filtering it. To see the effect of band-limiting in the resolving power of group delay spectrum, synthetic signal with closely spaced formants is used. Based on the results obtained on synthetic signal, the above method is used to extract the closely spaced formants in hypernasal speech of 15 cleft palate patients.

The rest of the paper is organized as follows. In Section 2, the group delay function and its properties are discussed. In Section 3, a comparative study is made on different formant extraction methods applied to a synthetic speech having two closely spaced formants. Section 4, describes the closely spaced formant extraction technique applied on low-pass filtered hypernasal speech and the results are discussed.

2. Group delay function and its properties

The negative derivative of the Fourier transform phase ($\theta(\omega)$) is defined as **group delay**. The group delay function $\tau(\omega)$ can be written as,

$$\tau(\omega) = -\partial\theta(\omega)/\partial\omega \tag{1}$$

The group delay function can be computed directly from the signal as described in [2]. i.e.,

$$\tau_p(K) = \frac{X_R(K).Y_R(K) + X_I(K).Y_I(K)}{|X(K)|^2} \tag{2}$$

$$\text{for } K = 0, 1, \ldots N-1$$

In Equation 2 $X(K)$ and $Y(K)$ are the N-point DFTs of the sequences $x(n)$ and $nx(n)$. The subscripts R and I denote the real and imaginary parts, respectively. To reduce the spiky nature of the group delay function, which is because of the pitch peaks, noise, and windowing effects, Equation 2 is modified as given below [2].

$$\tau_p(K) = sign\left|\frac{X_R(K).Y_R(K) + X_I(K).Y_I(K)}{S(K)^{2\gamma}}\right|^{\alpha} \tag{3}$$

In Equation 3, $S(K)^2$ is a cepstrally smoothed version of $|X(K)|^2$ and the **sign** in Equation 3 is the sign of the original group delay function given in Equation 2.

The group delay function exhibits an additive property unlike the magnitude spectrum. But, at the same time, the influence of the adjacent roots of the system function is not negligible. If

$$H(\omega) = H_1(\omega).H_2(\omega) \tag{4}$$

Then the group delay function $\tau_h(\omega)$ can be written as,

$$\tau_h(\omega) = -\partial(arg(H(\omega)))/\partial\omega \tag{5}$$

$$= \tau_{h1}(\omega) + \tau_{h2}(\omega) \tag{6}$$

From Equations 4 and 6, we see that multiplication in the spectral domain becomes addition in the group delay domain. Because of this additive property, the resolution of the spectrum is improved. But in some cases, the influence of the adjacent poles can not be neglected. The value of the group delay function of a pole k, at its own angular frequency ω_k is influenced by the rest of the roots and the influence $\tau_k^i(\omega_k)$ may be written as [6],

$$\tau_k^i(\omega_k) = \sum_{p=1\&p\neq k}^{P} \tau_p(\omega_k) + \sum_{n=1\&n\neq k}^{N} \tau_n(\omega_k) \tag{7}$$

where, - $\tau_p(\omega_k)$ is the value of the group delay function at the angular frequency $\omega(k)$ because of the pth pole - $\tau_n(\omega_k)$ is the value of the group delay function at the angular frequency $\omega(k)$ because of the nth zero - P is the number of poles - N is the number of zeros

For the present study, the influence of zeros are not studied, which may be taken for future study. Because of the influence of the adjacent poles, the closely spaced frequencies might not be resolved very distinctively. Hence, to resolve the closely spaced formant frequencies, the speech signal is filtered over low frequency band and the corresponding formant frequencies are extracted. Initially, this approach is applied to synthetic speech signal and analyzed as explained in the following Section.

3. Analysis on synthetic speech signal

The significance of band limiting the speech signal and extracting the additional nasal frequency introduced into the vowels using group delay formant extraction technique, is analyzed initially using a synthetic speech signal. The synthetic speech signal is generated using cascade configuration. The cascade configuration $H(Z)$ (for three formants) is given by,

$$H(Z) = H_1(Z).H_2(Z).H_3(Z) \tag{8}$$

$$H_i(Z) = \frac{1}{1 - 2e^{-\pi B_i T}cos(2\pi F_i T)Z^{-1} + e^{-2\pi B_i T}Z^{-2}} \tag{9}$$

$$\text{for } i = 1, 2, 3$$

In Equation 9, F_i are the formant frequencies and B_i are the formant bandwidths. Synthetic speech signals are generated using Equations 8 and 9, with different combinations of closely spaced F_1 and F_2 (for example, F_1 = 900 Hz, F_2 = 1100 Hz with varying bandwidths). Spectra

are computed for the resultant synthetic speech signals using DFT, LP with different orders, and the modified group delay function as described in the previous Section. For the computation of modified group delay function, α and γ are taken as 0.6 and 0.9 respectively.

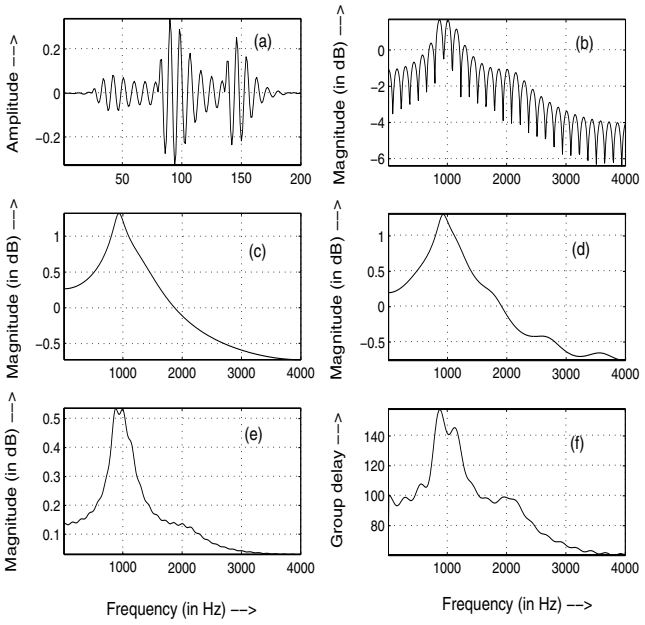

Figure 1: Comparison of different formant extraction techniques on synthetic signal. (a) Synthetic speech signal (Hanning windowed), (b) Magnitude spectrum, (c) LP spectrum (order 8), (d) LP spectrum (order 12), (e) Cepstrum smoothed spectrum, (f) Group delay spectrum

Figure 1 shows a comparison of conventional formant extraction techniques with group delay based technique applied on a synthetic speech signal. The synthetic signal is generated using a configuration of, F_1 = 900 Hz; F_2 = 1100 Hz; and F_3 = 2100 Hz; F_s the sampling frequency = 8000 Hz and $T = \frac{1}{F_s}$. The corresponding bandwidths are taken as, $B_1 = \frac{F_1}{10}$; $B_2 = \frac{F_2}{5}$; and $B_3 = \frac{F_3}{5}$;

From the Figure 1, it is observed that, because of the presence of pitch harmonics, FT spectra (refer Figure 1(b)) cannot be utilized directly for formant extraction. For the LP based formant extraction approach, for both lower (Figure 1(c)), and higher order (Figure 1(d)) spectra, the closely spaced formants are not resolved. Due to the poor frequency resolution, the cepstrum based smoothing technique does not clearly distinguish the closely spaced formants (Figure 1(e)). But compared to the above mentioned conventional methods, group delay based formant extraction technique resolves the formants, but not very distinctly (Figure 1(f)). This may be because of the influence of the adjacent poles.

To make the closely spaced formant frequencies more distinct, low-pass FIR filter with a cut-off frequency 1200 Hz (to discriminate 900 and 1100 Hz) is applied and the group delay function is recalculated for this band and formants are extracted. For comparison, group delay function derived for all-pass signal is also taken (Figure 2(c)). Figure 2(d) shows that low-pass filtering and computing group delay resolves the two closely spaced frequencies very distinctively. Hence the above approach can be utilized for the extraction of two closely spaced formant frequencies as in the hypernasal speech and is explained in the following section.

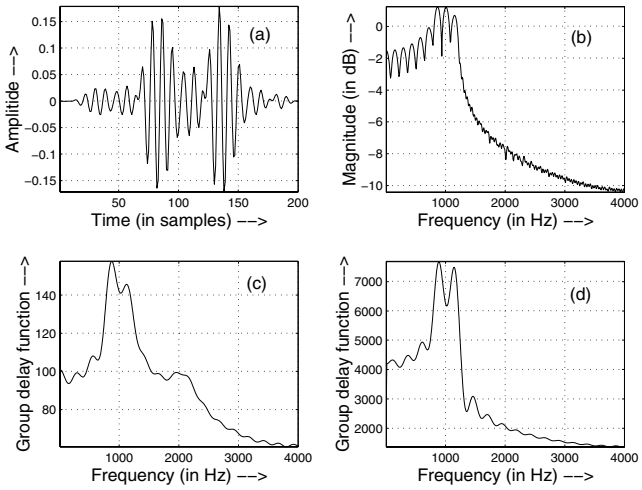

Figure 2: Effect of band-limiting (a) Synthetic speech signal (Hanning windowed) (b) Magnitude spectrum of low-pass filtered signal, (c) Group delay spectrum of all-pass filtered signal (d) Group delay spectrum of low-pass filtered signal

4. Analysis on hypernasal speech

The speech data analyzed in this study is collected from 15 patients with cleft palate who are expected to produce hypernasal speech. The acoustic analysis [1] on hypernasal speech (using LP spectra with varying order) for the phonations /a/, /i/, and /u/ revealed that an additional formant around 250 Hz is introduced due to oral-nasal coupling. This additional formant is close to the first formant of the phonations. In many cases, especially for the phonations /i/ and /u/, resolving this additional nasal formant and the first formant was difficult. Since low-pass filtering and extracting the formant frequencies using group delay function applied on synthetic speech is found to resolve two closely spaced formant frequencies (refer Section 3), the same technique is extended to hypernasal speech also. In this work, a low-pass filter is constructed with cut-off frequency of 800 Hz (F1 of /a/) to have a commonality between all the 3 phonations. The speech signal is passed through this filter to accommodate only the lower formants. The group delay function is computed for this filtered speech signal (see Figure 3(d)) and it is compared with the group delay function com-

puted for the all-pass filtered speech signal (refer Figure 3(c)).

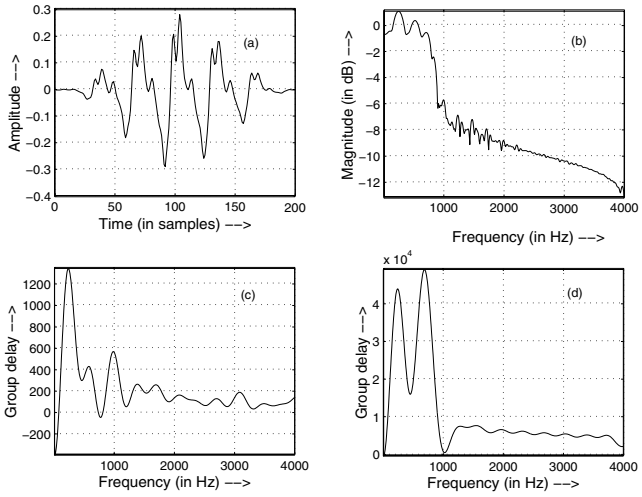

Figure 3: Effect of band-limiting on hypernasal speech (a) Hypernasal speech (windowed), (b) Magnitude spectrum of low-pass filtered signal, (c) Group delay spectrum of all-pass filtered signal and (d) Group delay spectrum of low-pass filtered signal

Figure 3(d) clearly shows that the low-pass filtering approach better discriminates the closely spaced frequencies for hypernasal speech. It is found that invariably for all the hypernasal speakers, the extra nasal formant around 250 Hz introduced due to hypernasality and the first formant frequency corresponding to the vowel are resolved properly (refer Table 1).

As mentioned earlier, the introduction of a nasal formant around 250 Hz in addition to first formant of the vowel is one of major cues for the detection of hypernasality. In the present study, specific interest was shown on resolving only the first two closely spaced formant frequencies (nasal formant and the first formant of the vowel). In fact, different group delay functions can be computed for each of the interested sub-bands to get information about all the formants. This can be taken up as a future study.

5. Conclusions

In this work, the group delay based formant extraction technique is compared with the other conventional techniques, for the synthetic speech signal. The resolving power of group delay technique is found to be consistently better. But, when the formants are closely spaced, even the group delay spectrum fails to resolve prominently. This is due to the influence of the other resonances. To avoid this influence, speech signal is band limited by low-pass filtering and it is observed that this technique clearly resolves closely spaced formants. The same technique is applied to hypernasal speech also. Since,

Table 1: Resolving power of group delay based formant extraction technique applied on low-pass filtered hypernasal speech produced by 15 speakers with cleft palate

speaker	no. of frames	resolved	perf. %
sp1	245	237	96.73
sp2	424	420	99.05
sp3	70	68	97.14
sp4	261	253	96.93
sp5	343	341	99.41
sp6	267	260	97.37
sp7	270	253	93.70
sp8	355	348	98.02
sp9	259	253	97.68
sp10	45	38	84.44
sp11	159	158	99.37
sp12	35	33	94.28
sp13	599	595	99.33
sp14	254	247	97.24
sp15	350	342	97.71

introduction of an additional formant around 250 Hz is an important cue for the detection of hypernasality, the group delay based formant extraction technique applied on band limited speech signal can be utilized to detect it.

6. References

[1] Vijayalakshmi, P., and M. Ramasubba Reddy, "Analysis of hypernasality by synthesis," in *Proceedings of Int. Conf. Spoken Language Processing*, Jeju island, South Korea, Oct. 2004, pp. 525–528.

[2] Murthy, H. A., Venkata Gadde, "The modified group delay function and its application to phoneme recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 2003, pp. 68–71.

[3] Cairns, D.A., J.H.L. Hansen and J.E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Trans. Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, Jan. 1996.

[4] Glass, J.R., and V.W. Zue, "Detection of nasalized vowels in american english," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1985, pp. 1569–1572.

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

[6] T Nagarajan, "Private communication," INRS-EMT, University of Quebec, 2005.