Extracting Formants from Short Segments of Speech using Group Delay Functions

Anand Joseph M., Guruprasad S., Yegnanarayana B.

Department of Computer Science & Engg.
Indian Institute of Technology Madras, Chennai-600 036, India
{anand,guru,yegna}@cs.iitm.ernet.in

Abstract

Speech is a non-stationary signal, with the shape of the vocal tract changing over several pitch periods, and also within the open and closed glottis phases. The effect of these changes is reflected in the locations of the formants which correspond to the resonant frequencies of the vocal tract. To observe these changes, the analysis window should be small enough (relative to a pitch period), and appropriately anchored. A non-model based method is proposed in this paper to accurately determine formants from short segments (less than a pitch period) of speech signals. It makes use of high resolution properties of group delay function to estimate formants from segments of duration less than a pitch period. The main advantage of this method is its lack of dependence on the parameters of a model. Analysis segments are synchronised with instants of glottal closure, to increase the robustness of formant extraction. Since continuity or additional acoustic-phonetic knowledge are not used, this method is fairly reliable and robust.

Index Terms: formant extraction, group delay function, short segment, speech analysis.

1. Introduction

Speech is the output of a dynamic vocal tract system, which is excited by a time varying excitation. Resonances of the vocal tract system which mainly reflect the vocal tract shape are called formant frequencies, or simply formants. Formants typically correspond to those frequencies that pass most acoustical energy from the source to the output [1]. Hence their locations are robust to degradations in a speech signal. Methods of formant estimation can be either model based (eg., linear prediction), or non-model based. Most non-model based approaches estimate formants from the magnitude spectrum, while ignoring the phase spectrum completely. This is primarily because the phase spectrum is difficult to analyse for discrete-time signals due to the associated problem of phase wrapping. However various attempts [3]-[5] have been made to demonstrate the significance of the phase spectrum in estimating formant frequencies. These methods typically involve the use of group delay function. Linear prediction model assumes a particular order of prediction, which in turn determines the number of peaks in the spectrum envelope, rather than the actual number of peaks that correspond to formants. Model and non-model based methods typically use analysis windows of 20-25 ms. Hence they are not useful for observing changes that occur in short intervals (less than a pitch period). Formant estimation using magnitude spectrum from short (< 5 ms) segments of speech are affected by the poor resolution caused by the size of the window. Even the group delay functions for short segments show several spurious peaks which mask the peaks due to formants. Model based linear prediction analysis is not suitable either, due to the necessity of choosing the order of prediction. In this paper a new method is proposed to estimate formants from short segments (less than a pitch period) using group delay functions.

Section 2 briefly discusses the group delay function, some of its properties, issues involved in using the group delay function for formant extraction, and some methods used to address these issues. Section 3 discusses the proposed method, and the motivation behind it. Sections 4 and 5 address some issues in the proposed method. Examples using real speech signals are considered to illustrate the effectiveness of the proposed method for formant extraction.

2. The group delay function

The group delay function is defined as the negative derivative of the Fourier transform phase of a signal [2]. For a minimum phase signal, the group delay computed from the magnitude spectrum of the Fourier transform is equal to that computed from the phase spectrum [8]. It has been shown that group delay functions of a cascade of linear systems are additive [2]. It has also been shown that the group delay function of any single resonator is approximately proportional to the square of the magnitude spectrum, around the resonant frequency [9].

Computation of the group delay function of a real signal is difficult due to various reasons. The most important one is due to the wrapping of the phase function. This is because the phase function of a discrete time signal, results in discontinuities in multiples of $\pm\pi$. This problem may be overcome by computing the group delay function ($\tau_g(\omega)$) directly from the signal $x[n]$ as follows [2]:

$$\begin{aligned} logX(\omega) &= log|X(\omega)| - j\theta(\omega) \\ \frac{d}{d\omega}logX(\omega) &= \frac{d}{d\omega}log|X(\omega)| - j\frac{d}{d\omega}\theta(\omega) \\ \frac{X'(\omega)}{X(\omega)} &= \frac{d}{d\omega}log|X(\omega)| - j\theta'(\omega) \\ \tau_g(\omega) &= -\theta'(\omega) \\ &= -\mathrm{Imag}(\frac{X'(\omega)}{X(\omega)}) \\ &= \frac{X_i(\omega)X'_r(\omega) - X_r(\omega)X'_i(\omega)}{X_r(\omega)^2 + X_i(\omega)^2} \end{aligned} \quad (1)$$

where $X(\omega) = X_r(\omega) + jX_i(\omega)$ is the Fourier transform of the discrete-time signal $x[n]$, and $X'(\omega) = X'_r(\omega) + jX'_i(\omega)$ is

its derivative.

Due to the inevitable truncation of the signal, truncation effects dominate in the computed group delay function. Pitch periodicity and noise in the signal also contribute to the problem. Various methods have been proposed [3]-[5] to overcome some of these problems. Some of these methods involve cepstral smoothing of the spectrum [3] prior to computing the group delay, or by computing the group delay of the autocorrelation sequence obtained from the signal [4]. In [5] the group delay is computed for a z-transform evaluated outside the unit circle. Formants are estimated from this group delay spectrum. However these methods are not suitable for estimating formants from segments of less than a pitch period.

3. Proposed method for estimating formant frequencies

We propose a new method for computing formants from short segments (less than a pitch period) of the speech signal. Most problems associated with the computation of the group delay function, are primarily due to zeros present in the denominator term of (1), which corresponds to the magnitude spectrum. It has been shown that the group delay function of a signal *around the resonant frequency* is proportional to the square of the magnitude of the Fourier transform [9]. That is,

$$\tau_g(\omega) \propto |X(\omega)|^2. \quad (2)$$

Because of the additive property of the group delay function, $\tau_g(\omega)$ due to a cascade of resonators will be proportional to the sum of the spectra $|X(\omega)|^2$ around each resonance frequency. It is the additive nature, and dependence of $\tau_g(\omega)$ on $|X(\omega)|^2$ that gives the high resolution property to the group delay function. If we consider (1) for short segments of data, the denominator term is smooth except for the effect of zeros in the frequency domain. Moreover the denominator term corresponds to the spectrum of the signal, which is typically large near the formant locations and hence it reduces the value of the numerator around the resonance peaks. If we ignore the denominator term, and consider the numerator term alone in the speech signal, we get

$$\begin{aligned} g(\omega) &= X_r(\omega)X_i'(\omega) - X_i(\omega)X_r'(\omega) \\ &= \tau_g(\omega)|X(\omega)|^2 \\ &\propto |X(\omega)|^4. \end{aligned} \quad (3)$$

In other words $g(\omega)$ shows sharper peaks *near the resonances* than $\tau_g(\omega)$. The $g(\omega)$ has been used previously for automatic speech recognition in [7] by computing the mel-frequency cepstral coefficients, where it was termed "product spectrum". However the term "product spectrum" is used to refer to the product of magnitude spectra each of whose frequency scales have been compressed by integer factors [11]. As an alternative, the term "resonance" or "formant group delay" is proposed. Figure 1 shows the linear prediction (LP) spectrum, the group delay function of the LP spectrum, and the numerator ($g(\omega)$) of the group delay for a signal segment of 5 ms. Note that the group delay function of the LP spectrum is that of the all-pole model derived from the signal, and not of the signal itself. The $g(\omega)$ merely helps to determine the location of formant peaks, but otherwise it has no interpretation in terms of signal characteristics. In particular, it is not possible to interpret the amplitude of peaks in terms of the spectrum or group delay. For large amplitude voiced regions, locations of strong peaks can be interpreted as formants. But for low amplitude segments and for noise, they hold little or no significance.

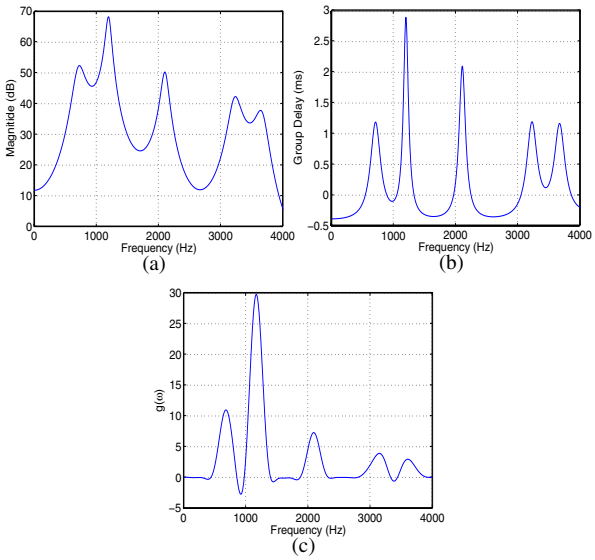

Figure 1: *(a) LP spectrum (16^{th} order), (b) its corresponding group delay function and (c) numerator $g(\omega)$ of group delay function computed from the signal, for a segment of voiced speech of* 5 *ms duration. The sampling rate of the signal was* 16 *kHz.*

One method of distinguishing formant peaks from other peaks is by using visible contour constraints. Figure 2 shows peaks obtained using the $g(\omega)$ function for an utterance of a speech signal sampled at 8 kHz, and the same signal sampled at 44.1 kHz. In the figure spectrograms have been plotted up to 4 kHz only. The differences in these two plots can be explained as follows. A 5 ms segment at 8 kHz corresponds to 40 samples, while that at 44.1 kHz corresponds to 220 samples. The larger number of samples obtained from the 44.1 kHz sampled segment helps mainly in reducing the window effect (since a larger window size in number of samples is used). This also provides a larger number of samples for each frequency component that may be present in the segment. This may help in reducing the effect of spurious peaks, as can be observed in the $g(\omega)$ plots in Figure 3 for a segment sampled at 8 kHz and 44.1 kHz. Note that the large number of samples in analysis window does not improve the frequency resolution, as the duration of the segment in seconds is same in both the cases.

In Figure 3, the peak around 3.7 kHz is relatively stronger in the $g(\omega)$ computed from the segment sampled at 44.1 kHz, than computed from the segment sampled at 8 kHz. Another advantage of using a higher sampling rate is that most of the spurious peaks picked by a simple peak picking algorithm are spread over frequencies higher than those which are significant for speech signals. Fluctuations in formant locations as observed in Figures 2(a) and 2(b), are primarily due to variations that occur as the analysis window shifts from the closed glottis region to the open glottis region. The position of the analysis window also affects the location of peaks in the $g(\omega)$ function, resulting in some random fluctuations of the formant locations around their true value.

4. Reducing the effect of window locations

One way of reducing the effect of the position of the analysis window, is to use the autocorrelation sequence computed from the signal, instead of the signal itself for formant extraction. However for short segments ($\leq$ 5 ms), the values of the autocorrelation se-

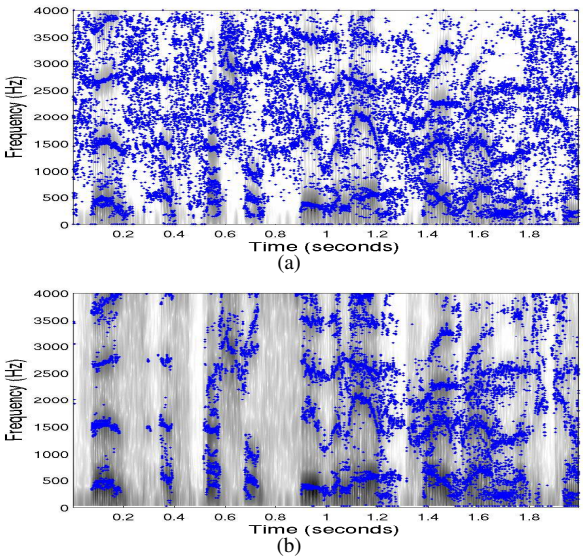

Figure 2: *Formant contour obtained for a* 2 *second segment of speech by picking the best* 5 *peaks in* $g(\omega)$ *function, for speech sampled at (a)* 8 *kHz and (b)* 44.1 *kHz. The size of analysis window used is* 5 *ms, and the window shift is* 0.5 *ms.*

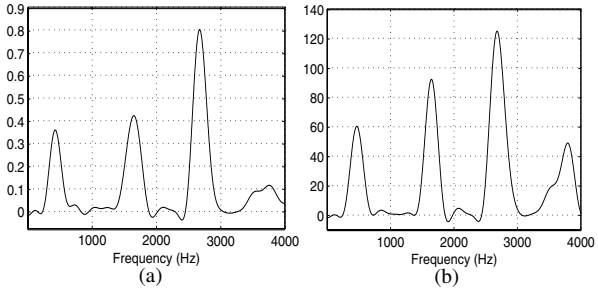

Figure 3: $g(\omega)$ *for a segment of voiced speech of* 5 *ms duration, for speech sampled at (a)* 8 *kHz and (b)* 44.1 *kHz.*

quence for larger lags ($>$ 1 ms) may not be accurate due to fewer number of samples of the signal being used in the computation. To overcome this problem, one can compute the normalised covariance value for different lags. The normalised covariance coefficient $\phi[m]$ for N samples of a speech signal $x[n]$ up to a lag m is given by

$$\phi[m] = \frac{\sum_{n=0}^{N-1} x[n]x[n+m]}{\sqrt{\sum_{n=0}^{N-1} x^2[n]}\sqrt{\sum_{n=0}^{N-1} x^2[n+m]}}. \qquad (4)$$

The computation of the covariance sequence uses $2N-1$ samples of the signal, whereas the computation of the autocorrelation function involves only N samples of the signal. For this purpose, the analysis segment of the speech segment and its covariance sequence can be of the same length. Figure 4 shows formant contours obtained when $g(\omega)$ is computed using the covariance function of analysis segments of 5 ms and 25 ms durations. It can be observed that the number of spurious peaks in Figure 4(a) is less compared to that in Figure 2(b). The effect of using a larger window size can also be observed in Figure 4(b), which has a very smooth formant contour. However the use of the larger window size only shows an average behaviour over the entire segment.

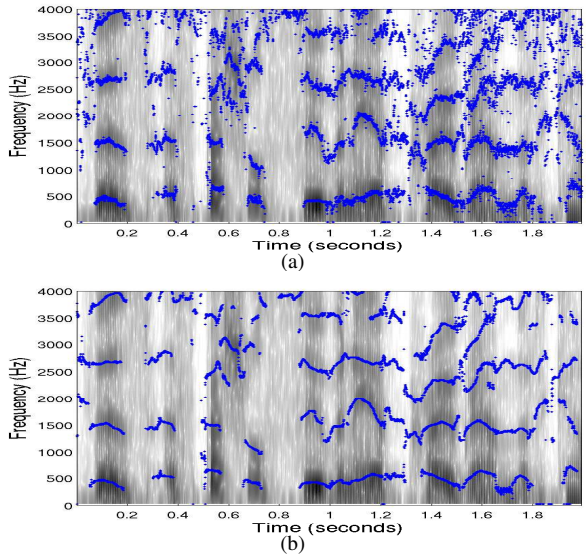

Figure 4: *Formant contour obtained by picking the best* 5 *peaks in* $g(\omega)$ *function computed from covariance sequence* $\phi[n]$ *of the signal. The speech signal was sampled at* 44.1 *kHz. The window size of the signal used to compute* $\phi[n]$ *is (a)* 5 *ms (*220 *samples) and (b)* 25 *ms (*1100 *samples). In both cases, the window shift is* 0.45 *ms (*20 *samples). In both cases, only* 2.3 *ms (*100 *samples) of* $\phi[n]$ *is used to compute* $g(\omega)$.

5. Effect of synchronised windowing

As mentioned in Section 3, the $g(\omega)$ function merely helps in determining formant locations. For an arbitrarily located window, in addition to formant peaks, there may be some spurious peaks with values higher than that of formant peaks. This can be seen in Figure 3. One of the methods for reducing the effect of spurious peaks would be to use a signal at a higher sampling rate. While a higher sampling rate helps to some extent, it does not significantly reduce the number of spurious peaks as can be seen in Figures 2(b) and 4(a). Alternatively, formants could be estimated from the covariance sequence of a larger segment. However this defeats the purpose of observing the dynamic characteristics of the vocal tract system through formant plots.

Recall that formants correspond to frequencies that pass most of the acoustical energy from the source to the output. This energy is not spread uniformly over the entire duration of the signal, but is typically concentrated around the instants of glottal closure (at which significant energy is delivered to the vocal tract system). In every pitch cycle, high SNR regions of the speech correspond to regions around glottal closure instants (GCI). It is in these regions that values of formant peaks are higher than those of spurious ones. Hence if analysis windows are synchronised with GCI, most spurious peaks can be eliminated. Formants can therefore be obtained using the $g(\omega)$ function from analysis windows anchored at GCI, rather than at arbitrary locations. For estimating the locations of GCI in a given speech signal, the method proposed in [10] was used. Using these locations, formants were estimated for a sentence sampled at 16 kHz from the TIMIT database. The size of analysis segments used was 5 ms, and best 5 peaks were picked from the $g(\omega)$ function computed from them. Figure 5(a) shows formant peaks plotted on the spectrogram. It can be seen that most formants are estimated reasonably well. However some spurious peaks are also picked up by the peak picking algorithm. The steps

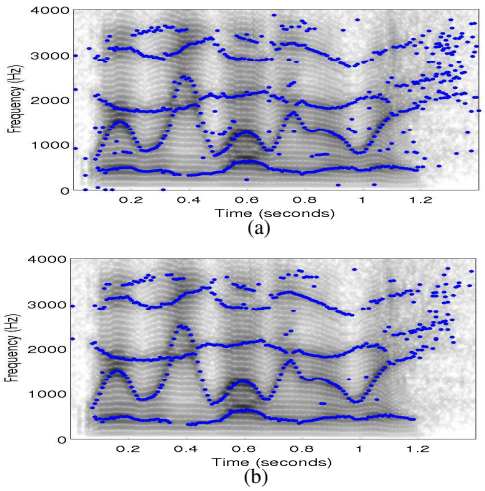

Figure 5: *Formant contour obtained by picking* 5 *peaks in* $g(\omega)$ *function of a TIMIT Sentence "where were you while we were away" from the (a) signal (b) from three levels of differenced speech, with GCI synchronised windows. The window size used was* 5 *ms in both cases.*

involved have been summarized in Table 1.

Spurious peaks are picked up because of relatively lower peak values of higher formants compared to those of spurious ones. To reduce the probability that a spurious peak is picked up instead of a genuine formant, higher formant peak values need to be enhanced, while at the same time keeping spurious ones low. A simple method to enhance high frequency formants is to difference the speech signal. This differencing however acts like a high pass filter, deemphasising lower frequency formants. Hence two or three $g(\omega)$ functions can be computed for each segment for each order of differencing. These $g(\omega)$ functions are then added up to obtain a single $g(\omega)$ function, from which formants could be estimated. This process in effect is equivalent to some kind of weighting of the $g(\omega)$ function to enhance formants while reducing spurious peaks. Formant peaks estimated using this method are plotted on the spectrogram in Figure 5(b). It can be seen in the figure the spurious peaks are significantly reduced by employing this method.

6. Conclusion

A new method for extraction of formants from short segments of speech signals has been proposed to capture the information in rapid movements of the vocal tract system. This method exploits the high resolution property of the group delay function, and defines a function $g(\omega)$ called the formant group delay function, whose peaks give reliable and robust estimates of the formants. This method also distinguishes between voiced and unvoiced regions due to the presence or absence of formant peaks in the lower frequency regions. The number of spurious peaks can be reduced by using the autocorrelation or the covariance sequence computed from the segment of windowed speech. By synchronising the analysis windows with the instants of glottal closure, one can further reduce the number of spurious peaks. If locations of GCI are accurately known, one can use this method to compare formants in the open and closed glottis regions. Formants estimated using this method could be used for formant tracking. Note that the proposed method of formant estimation from the peaks in the 'formant group delay' is superior to the method based on the roots of LPC polynomial, as the latter requires us to specify the order of the polynomial, whereas the proposed method gives whatever peaks that are available in the specified frequency range (say, 4 kHz).

Table 1: *Proposed method for formant extraction*

Step 1.	Window a segment of the signal using a half Hanning window of length less than a pitch period.
Step 2.	$g(\omega)$ function is computed directly from the segment or from the autocorrelation sequence or from the covariance sequence of the segment.
Step 3.	Pick the largest N number of peaks in the computed $g(\omega)$ function.
Step 4.	Shift the window by 1 ms or place the window at instants of glottal closure, if they are available.
Step 5.	Repeat steps $1-4$ for successive window segments.

7. References

[1] Rabiner, L. and Juang, B. H., Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] Oppenheim, A. V. and Schafer, R. W., Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[3] Murthy, H. A. and Yegnanarayana, B., "Formant extraction from group delay function", Speech Communication, Vol. 10, no. 3, pp. 209-221, Aug. 1991.

[4] Duncan, G., Yegnanarayana, B., and Murthy, H. A., "A nonparametric method of formant estimation using group delay spectra", Proc. ICASSP, Glasgow, May 1989, pp. 572-575.

[5] Bozkurt, B., Doval, B., d'Alessandro, C., and Dutoit, T., "Improved differential phase spectrum processing for formant tracking", Proc. ICSLP, Jeju Island, Korea, Oct. 2004.

[6] Bozkurt, B., Doval, B., d'Alessandro, C., and Dutoit, T., "Appropriate windowing for group delay analysis and roots of Z-transform of speech signals", Proc. EUSIPCO, Vienna, Austria, Sep. 2004.

[7] Zhu, D., and Paliwal, K. K, "Product of power spectrum and group delay function for speech recognition", Proc. ICASSP, Montreal, May 2004, pp. 572-575.

[8] Yegnanarayana, B., Saikia, D. K., and Krishnan, T. R., "Significance of group delay functions in signal reconstruction from spectral magnitude or phase", IEEE Trans. on Acoustics Speech and Signal Proc., Vol. 32, no. 3, pp. 610-623, Jun. 1984.

[9] Yegnanarayana, B., "Formant extraction from linear prediction phase spectra", J. Acoust. Soc. Amer., Vol. 63, no. 5, pp. 1638-1640, May 1978.

[10] Smits, R., and Yegnanarayana, B., "Determination of instants of significant excitation in speech using group delay function", IEEE Trans. Speech and Audio Proc., Vol. 3, pp. 325-333, Sep. 1995.

[11] Schroeder, M.R., "Period histogram and product spectrum: new methods for fundamental frequency measurement", J. Acoust. Soc. Amer., Vol. 43, no. 4, pp. 829–834, Apr. 1968.