S11.4

A NONPARAMETRIC METHOD OF FORMANT ESTIMATION USING GROUP DELAY SPECTRA

G. Duncan (1), B. Yegnanarayana (2), and Hema A. Murthy (2)

(1) Nixdorf Computer (France), 14 Avenue des Beguines, 95802 Cergy St. Christophe , France

(2) Indian Institute of Technology, Dept. of Computer Science and Engineering, Madras-600036, India

ABSTRACT

Rule-based methods of speech recognition have a heavy dependence on reliable formant trajectory estimation. The underlying spectral analysis must be able both to demerge closely-coupled formants and detect weak (e.g. nasal) formants. Ideally, detec tion of antiformant activity would also help in phoneme discrimination. The new minimum-phase group delay technique discussed in this paper is a nonparametric spectral analysis method possessing the above properties. It therefore requires no assumptions concerning the underlying nature of the signal, other than it originates from an LTI filter system. The technique provides a level of performance in formant detection normally associated only with larynx-synchronous techniques. Moreover, it is extremely easy to implement, requiring only 3 FFT operations per analysis frame.

1. INTRODUCTION

Formant estimation algorithms have tended to rely in the main on parametric forms of spectral analysis, where explicit assumptions are made regarding the nature of both the vocal tract filter and the glottal source. Typically, the use of such parametric models will necessitate explicit choices of model parameters which may have a direct bearing on the subsequent performance of the spectral estimator. In linear predictive coding (LPC) analysis, for example, the choice of model order determines the level and quality of spectral detail. If the model order is insufficient, then certain formants will not be adequately modeled into the spectrum, particularly where formants merge in the time-frequency plane. Conversely, a model order which is excessive will deteriorate the signal-to-noise performance of the LPC-based spectral estimator and will usually create a profusion of spectral peaks from which it is then difficult to choose formant candidates.

This paper presents a novel approach to formant estimation which makes no assumptions regarding the nature of the speech waveform other than that it conforms to a linear time-invariant filter over short-time observation intervals. This new approach, which is nonparametric, exploits both the additive and high-resolution properties of the group delay (GD) function, which is simply the negative derivative of the Fourier transform phase. Central to the success of this new method is the calculation of a minimum-phase signal, from which the GD function is then derived. The method is shown to give reliable estimation of formants without resorting to any modeling approach for spectral smoothing.

2. DERIVATION OF THE MINIMUM PHASE SIGNAL

The minimum phase signal required in this technique can be easily obtained from the Fourier transform (FT) of a short segment of speech. In particular, the FT phase spectrum of this minimum phase signal has several interesting properties which can be exploited for estimation of formant frequencies of the speech signal. In particular, the additive and high resolution properties of the GD function resolve even closely spaced formants, which is not possible from the FT magnitude spectrum [1]. In the algorithm presented here, a minimum phase signal is firstly derived from the inverse FT applied to the co-phase (zero-phase) magnitude function. The forward Fourier transform is then applied to only the first few milliseconds of this new minimum phase signal, The group delay spectrum is then computed from the new FT phase spectrum, and it is shown that significant peaks in the group delay function correspond to formants.

Standard methods of formant estimation tend to rely on short-time analysis employing some form of magnitude spectrum representation, usually based upon the tenet that phase is of relatively little importance to speech. However, this assumption generally refers to the effects of variations in path length between a sound source and the human ear, whereby relative changes of position between source and receiver produce linear changes in the arrival time (and hence phase) of various frequency components of the signal [2]. Fourier transform phase as a signal-information-bearing entity has quite different properties, however. For example, for a non-complex signal such as speech, all of

CH2673-2/89/0000-0572 $1.00 © 1989 IEEE

the information present in the magnitude spectrum is also present, in a different form, in the Fourier phase spectrum. This information is usually affected, however, by the angular resolution of the FT phase, which is calculable to within +/-180°. As phase values exceed either of these limits, phase may change abruptly from a negative-going value near to, say, -180° to a positive value near to +180°. Thus, the phase-resident information is often obscured by phase-wrapping effects.

The accuracy of formant estimation from the magnitude spectrum alone depends on several factors such as size and shape of the data window, preprocessing, etc. Moreover, it is difficult to clearly identify formants from the short-time FT magnitude spectrum because of the fluctuations due to noise and pitch, besides the problem of dynamic range of the spectrum. Smoothing of the magnitude spectrum through cepstral processing or some similar approach produces peaks which sometimes fail to provide adequate formant feature resolution, and hence closely-spaced formants may appear merged as one single peak. Adoption of a model-based approach, like linear prediction (LP) analysis, produces sharp unambiguous peaks, but it is then difficult to know which peaks correspond to formants. Moreover, the number of peaks in the LP spectrum depends on the nature of the magnitude spectrum and the choice of analysis parameters such as the order of the model used for analysis. Employing the group delay function from LP phase [1] will generally adequately resolve all the peaks in the spectrum, but the problems of spurious peaks and modeling inaccuracies due to the precise choice of model parameters still remain.

The group delay spectrum derived from the FT phase of a signal posesses two important properties, namely, additive and high resolution properties [1,3]. In this paper the possibility of using these properties for formant extraction is explored. Arbitrary placement of a finitely-long observation window on the speech waveform will generally result in a mixed phase signal segment. That is, viewed from a transfer function standpoint, the zero-time reference point of the observation window may not coincide exactly with the start of a (pitch-synchronous) impulse response of the vocal tract. If a pitch-asynchronous segment is under analysis, as will generally be the case, then the equivalent transfer function of the vocal tract, as far as the analysis window is concerned, may not in fact appear to be generated by a minimum-phase filter system. That is, poles and zeroes of the virtual filter system generating the segment are not guaranteed to all lie within the z-plane unit circle, but will be placed both within and outwith the unit circle in no particular order, but in such a way as to produce an appropriate magnitude spectrum. Negative-going changes of phase in the FT phase spectrum may therefore be due to either a pole within the unit circle or a zero outside the unit circle.

Computation of unwrapped phase for a mixed phase signal of this type is difficult. In addition, the FT phase spectrum will contain changes of phase which are directly related to pitch. In this case, post-application of the group delay function, even with a successful unwrapping of phase, would render interpretation of group delay peaks as formants to be a highly non-trivial task. Therefore here, a minimum phase signal for each segment of speech is derived, without recourse to any explicit pitch-synchronous analysis techniques, and it is shown that the FT magnitude and phase of the minimum phase signal are related. It is demonstrated that the phase characteristic contains all the information of the corresponding smoothed magnitude spectrum. Because of the additive and high resolution properties of the negative derivative of phase (group delay) function, the formant peaks stand out clearly compared to the smoothed magnitude spectrum. The study also shows that a specific model order as in LP analysis is not required. The performance of this new method is illustrated for a segment of speech data.

3. PROPERTIES OF THE GROUP DELAY FUNCTION

The group delay function for a signal can be defined in terms of two functions of the Fourier transform [3]. The one, $Tm(w)$, derived from the FT magnitude spectrum and the other, $Tp(w)$, derived from the FT phase spectrum. Using these functions it is possible to explain the relation between the FT magnitude and phase. The important properties of the group delay functions [1,3] with respect to formant detection are summarised here.

1. Minimum phase property:

   For a minimum phase signal $Tp(w)=Tm(w)$.

2. Additive property:

   The group delay function of a cascade of resonators is the sum of the group delay functions of each individual resonator.

3. High resolution property:

   Significant information of a resonator is concentrated around the resonant frequency and the function values are proportional to the squared FT magnitude function.

This latter property is illustrated in Fig. 1, which compares the frequency domain characteristics of a 1kHz decaying tone as obtained from the FT short-time squared magnitude spectrum and the GD function derived from the same FT phase spectrum. All amplitudes have been normalised to respective peak values to allow comparison. Note that the difference in centre frequency between the resonance characteristics arises

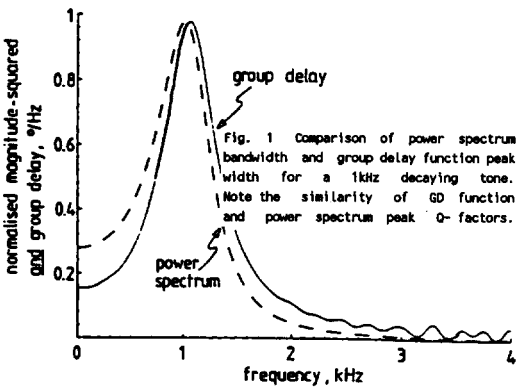


Fig. 1 Comparison of power spectrum bandwidth and group delay function peak width for a 1kHz decaying tone. Note the similarity of GD function and power spectrum peak Q- factors.

since the centre frequency of the power spectrum peak will depend directly on the damping factor associated with the resonance, as well as on its undamped natural frequency. That is, it will depend on both the angular and radial position of the pole pair characterising the resonance in the digital-domain z-plane. The GD-based centre frequency will, however, depend principally on the undamped natural frequency of the resonance, i.e. on the angular position of the pole pair in the z-plane.

4. FORMANT ESTIMATION FROM GROUP DELAY

In this section it is demonstrated that the peaks in the group delay function Tp(w) of a minimum phase signal can be used for estimating the formant frequencies. Firstly, a segment (25.6ms, sampling frequency 16kHz) of speech (Fig.2a) is taken at any random position on the signal waveform. This is then multiplied with a Hamming window and a 1024 point discrete Fourier transform is computed so as to obtain magnitude and phase functions. The inverse FT of the (zero-phase) magnitude function (Fig.2c) gives a minimum phase signal (Fig. 2b). That is, in an analogous manner to the autocorrelation function derived from the (zero-phase) power spectrum, any signal periodicity is exactly referred back to time T=0 in the signal derived from the inverse zero-phase FT. Thus, if the filter system which generated the signal segment is a stable, minimum-phase, then the new signal around time T=0 is itself guaranteed to be minimum phase. This new signal is subsequently multiplied with a half-Hamming window to select only the first p samples (Fig.2b). The criterion determining window size is that p should be taken as large as possible to obtain sufficient resolution of the formants, but should be less than the pitch period. This condition is imposed in order to maintain a smooth frequency response, devoid of fluctuations which would otherwise be manifest if the analysis window were to contain more than one pitch period. It is usually difficult to see the relationship between magnitude and (mixed-) phase spectra for the original speech segment, which may be non-minimum phase in nature depending on the exact positioning of the data window with respect to the start of a pitch period. However, their relationship is obvious for the minimum phase signal of Fig.2b, to which a second FT is applied to produce new FT magnitude and phase functions (Figs.3a-b). The details of spectral shape show up as rapid fluctuations in the phase. Moreover, since the segment is now guaranteed to be minimum-phase, negative-going changes of phase can be identified with resonances (formants), and positive-going changes of phase with antiresonances (zeros). Finally, applying the GD function converts the negative-going phase characteristic associated with formants into positive-valued peaks.

The details can be more easily seen in the group delay function (Fig.3c) when compared against the FT magnitude function (Fig.2c). Also shown as an overlay on Fig.2c is an 18th-order LPC smooth spectrum. It is interesting to note that the group delay function brings out the details of F3 which are not evident either in the original FT magnitude spectrum or in the LPC spectrum. This is mainly due to both the additive and the high resolution properties of the group delay function. The most prominent peaks in the group delay spectrum generally correspond to formants. Moreover, the height of the peak in the group delay function can be shown to be inversely proportional to the bandwidth of the formant [1]. The resolution of the peaks in the group delay function is significantly high, although here, a data window (4ms) has been employed which is less than a pitch period. It has been found that closely spaced formants are resolved even with a small (minimum-phase) window length, which is a direct consequence of the high resolution property of the group delay function. This feature is potentially very useful in the analysis of high-pitched sounds such as female and childrens' voices.

5. CONCLUSIONS

A new method has been proposed for extraction of formant information from the speech signal. It has been demonstrated that the additive and the high resolution properties of the group delay functions offer the capability of resolving even closely spaced and weak low-amplitude formants. The method is superior to peak-picking from the smoothed magnitude spectrum or even from a linear prediction model spectrum. In particular, the GD-based method does not depend on any parametric model and hence the data obtained should be a better representation of the underlying nature of the signal than that obtained from model-based techniques.

6. ACKNOWLEDGEMENTS

The authors wish to extend their grateful thanks to the Professor J. Laver and to Professor M. Jack, Centre for

Speech Technology Research, University of Edinburgh, Scotland, where this research was carried out. This collaborative international research effort has been promoted by the British Council.

6. REFERENCES

[1] B. Yegnanarayana, "Formant extraction from linear prediction phase spectrum", Journal of the Acoustical Society of America Vol.63(1), pp.1638-1640:1978

[2] E. de Boer, "A note on phase distortion in hearing", Acustica, Vol.11, pg.182:1961

[3] B. Yegnanarayana, D. K. Saikia and T. R. Krishnan, " Significance of group delay functions in signal reconstruction from spectral magnitude or phase", Trans. IEEE, Vol.ASSP-32(3), pp.610 - 623:June 1984

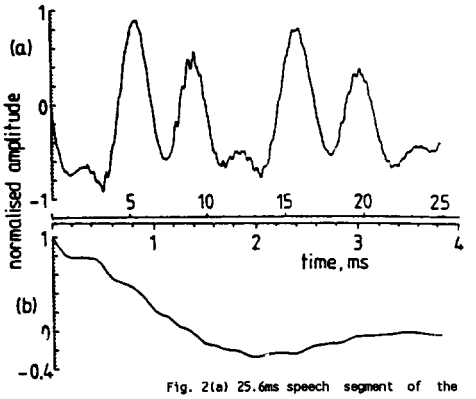
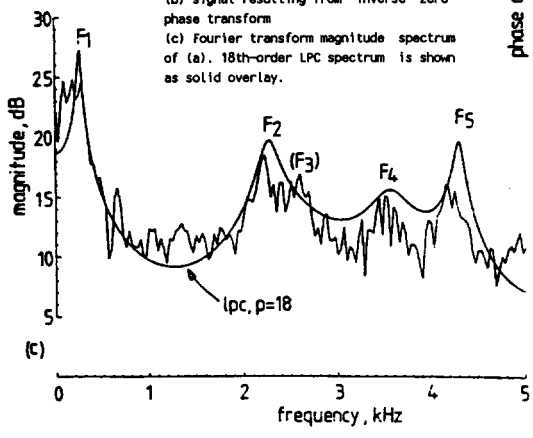
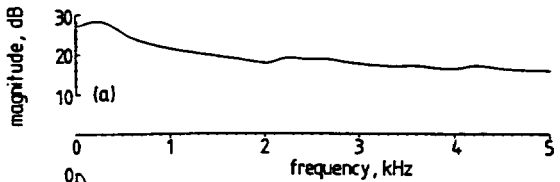


Fig. 2(a) 25.6ms speech segment of the vowel /ee/;
(b) signal resulting from inverse zero-phase transform
(c) Fourier transform magnitude spectrum of (a). 18th-order LPC spectrum is shown as solid overlay.





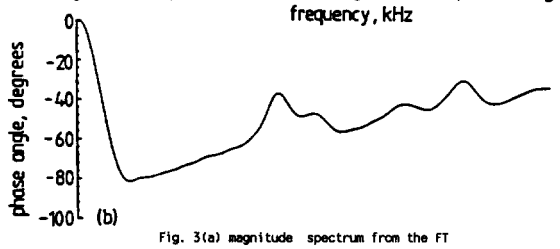


Fig. 3(a) magnitude spectrum from the FT analysis of the waveform of Fig. 2(b).
(b) Corresponding phase spectrum.
(c) Group delay function with formants identified.