

ANALYSIS OF SPEECH SIGNAL ENVELOPE-FREQUENCY RELATIONSHIPS

David A. Seggie *

ABSTRACT

It is shown that the complex zero representation of bandlimited signals provides an appropriate mathematical basis for the interpretation of speech signal time-domain amplitude envelope and frequency. The representation is used to explain the relationship between characteristic features of both time-domain signal attributes.

INTRODUCTION

Recent work in speech analysis has explored the possibility that the temporal frequency characteristics of speech signals encode useful information. In an investigation of frequency vs time signal representations, Berthomier (ref 1) proposed switching from the usual concept of frequency in the sense of Fourier's integral theorem, to the concept of frequency as the time derivative of temporal phase. The time derivative of temporal phase is a quantity referred to as instantaneous frequency. Berthomier suggested that the distribution of signal energy on an instantaneous frequency axis is characteristic of the signal, and that such a distribution may be useful in speech recognition. This suggestion appears to have been verified by Demars and Gauvain (ref 2) in their assessment of the application of instantaneous frequency estimation to speaker dependent isolated word recognition. The relationship between the mean instantaneous frequency of a signal and its spectral distribution has also been examined, and possible applications in speech segmentation identified (ref 3). These studies have been restricted to an analysis of instantaneous frequency data averaged either over the entire utterance or over short segments of the utterance. The question of whether useful information may be extracted from the local variations in this time-domain signal attribute has not been addressed. Furthermore, none of the above studies has presented a mathematical basis for interpreting the observed features of speech signal instantaneous frequency functions.

Instantaneous frequency may be defined via an analytic signal approach (ref 4). Given a real bandlimited speech pressure waveform, $s(t)$, the associated analytic signal is ,

$$a(t) = s(t) + j\tilde{s}(t)$$

where $j = \sqrt{-1}$, and $\tilde{s}(t)$ denotes the Hilbert transform of $s(t)$. Signal temporal phase is given by,

$$\phi(t) = \arg[a(t)] = \arctan[\tilde{s}(t)/s(t)] \quad (1)$$

$\phi(t)$ obtained in this way is modulo 2π and must therefore be unwrapped to obtain the desired continuous phase function (ref 5). Instantaneous frequency, $\phi'(t)$, is the time derivative of the unwrapped version of $\phi(t)$. A third time-domain attribute which is of interest is signal amplitude envelope, $e(t) = \text{mod}[a(t)]$.

*Dept. of Phonetics & Linguistics, University College London

TIME-DOMAIN FREQUENCY AND ENVELOPE FLUCTUATIONS

A marked feature of voiced speech is the presence of large, time-localized fluctuations in $\phi'(t)$ which often form a well-defined and structured pattern. This is illustrated by Figs. 1-3. Figure 1 shows a 60 ms segment of a speech pressure waveform for the vowel [a] produced by an adult female with a low-fall intonation contour. Figures 2 & 3 depict the amplitude envelope and instantaneous frequency functions for the signal in Fig. 1. Note that large excursions in $\phi'(t)$ occur when $e(t)$ exhibits a local minimum. However, it should not be thought that these $\phi'(t)$ excursions are simply the result of computational inaccuracies due to low values of $e(t)$, as might be suggested from an analysis of equation (1) for example. Empirical evidence indicates that the values of $e(t)$ coincident with these large $\phi'(t)$ excursions are well above the level at which computational noise is problematic. Such $\phi'(t)$ excursions would therefore appear to be an intrinsic feature of voiced speech. A mathematical basis for both describing the structure of speech instantaneous frequency functions and explaining the relationship between fluctuations in $\phi'(t)$ and $e(t)$, is provided by the complex zero representation of bandlimited signals (ref 6).

Briefly, the analytic signal defined above may be represented by inverse Fourier transformation as,

$$a(t) = \int_0^{\infty} A(f) \exp(2\pi jft) df$$

where $A(f)$ is the Fourier transform of $a(t)$. $a(t)$ can be continued into the complex time domain by replacing t in the above expression by the complex variable $z=u+jv$, giving

$$a(z) = \int_0^{\infty} [A(f) \exp(-2\pi fv)] \exp(2\pi jfu) df$$

It can be shown that $a(z)$ may be written as an expansion product and, in a manner analogous to the factorization of algebraic or trigonometric polynomials, represented unambiguously by its roots or zeros. This in turn means that the original speech pressure waveform, $s(t)$, can be specified completely, (to within a multiplicative constant), by a knowledge of the real and complex zeros of $a(z)$. If these zeros are distinct, then in the region of one such zero of order n_i , located at $z=z_i$, $a(z)$ may be written as,

$$a(z) = (z-z_i)^{n_i} s(z)$$

where $s(z)$ is some zero-free function. It follows that,

$$d/dz(\ln a(z)) = n_i/(z-z_i) + d/dz(\ln s(z))$$

That is, each isolated zero of $a(z)$ contributes a 1st order pole to $d/dz(\ln a(z))$. Now by considering the integral

$$\oint_c dz [d/dz(\ln a(z))] / z(z-t)$$

where c is a closed contour containing the points $z=0$, $z=t$, and may be

thought of as an infinite (in the limit) circle centred on the origin, it may be shown that (ref 7),

$$\phi'(t) = k_1 + \sum_i [n_i v_i / ((t-u_i)^2 + v_i^2)] \quad (2)$$

$$d/dt(\ln e(t)) = k_2 + \sum_i [n_i (t-u_i) / ((t-u_i)^2 + v_i^2)] \quad (3)$$

k_1 and k_2 are real constants. The above equations describe fluctuations in signal amplitude envelope and instantaneous frequency in terms of the complex zero locations of $a(z)$. The utility of the equations can be demonstrated by considering the contribution to $\phi'(t)$ and $d/dt(\ln e(t))$ from an isolated zero located at $z=u_i+jv_i$. Equation (2) shows that as time goes from $t < u_i$, to $t > u_i$ the contribution of the zero to $\phi'(t)$ goes through a maximum at $t=u_i$. The zero therefore encodes a local excursion in $\phi'(t)$ at $t=u_i$; the excursion is positive if the zero lies in the upper half of the z -plane, negative if it lies in the lower half plane. Equation (3) shows that the zero encodes a minimum in $e(t)$ at $t=u_i$, regardless of whether the zero is in the upper or lower half plane. Hence, irrespective of the absolute signal amplitude envelope level, minima in $e(t)$ are accompanied by excursions, (both negative and positive), in $\phi'(t)$. Observe however, that this relationship between fluctuations in $\phi'(t)$ and $e(t)$ does not hold for a complex conjugate zero pair; then an envelope minima may be encoded at $t=u_i$ without any excursion in $\phi'(t)$. The complex zero representation thus provides an explanation of the correspondence between large excursions in $\phi'(t)$ and local minima in $e(t)$. The representation shows that the large, time-localized fluctuations in instantaneous frequency (seen in Fig. 3 for example) are not, as has been implied by others, the expression of computational inaccuracies when the amplitude envelope function is small.

SUMMARY

Having established an appropriate representation for interpreting fluctuations in the time-domain attributes of speech signals, the following questions arise. Does this representation, (and the time-domain features it describes) reflect salient physical characteristics of the speech waveform? In particular, can the fact that this representation retains local significance be exploited by extracting useful information from the local variations in $\phi'(t)$ and $e(t)$? Preliminary results suggest that these questions can be answered affirmatively. For example, Fig. 4 shows fundamental frequency contours, (both before and after median smoothing), derived from an analysis of the instantaneous frequency function for the entire speech pressure waveform shown in part in Fig. 1. The fundamental frequency estimates were obtained by computing the inverse of the time intervals measured between adjacent large fluctuations in $\phi'(t)$. The data depicted in Fig. 4 are consistent with estimates produced by standard pitch detection methods.

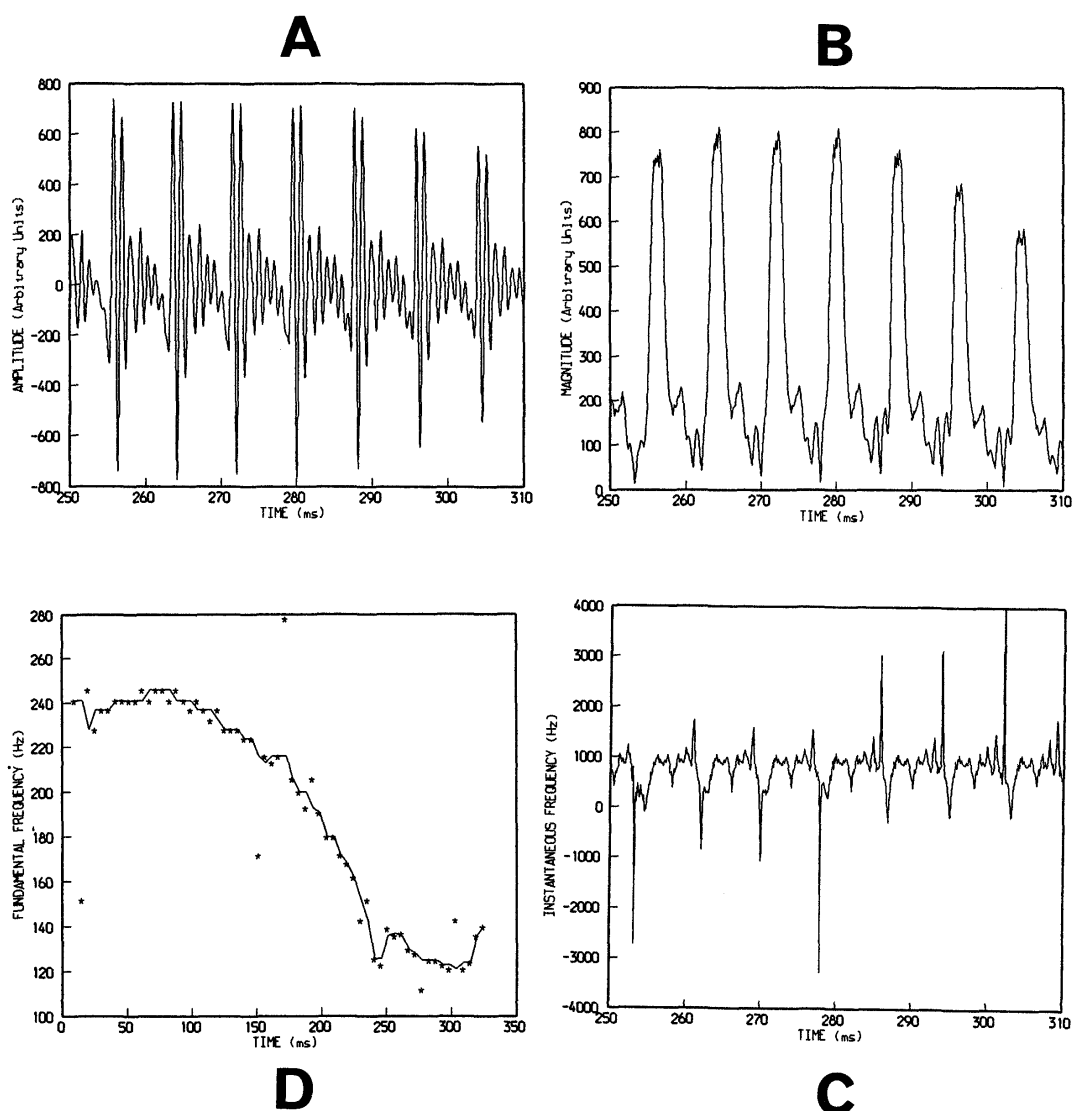
It is tempting to view such initial results, (albeit with a very straightforward speech signal), as indicative of the potential of a detailed analysis of speech signal time-domain envelope/frequency functions. This paper indicates that the complex zero representation and the time-domain signal attributes it describes, certainly merit further investigation.

ACKNOWLEDGEMENTS

This work was supported by SERC-Alvey grant MMI/056.

REFERENCES

1. C Berthomier, *Signal Processing*, 5, (1983), p 31.
2. C Demars and J L Gauvain, *Actes des XIVth Journees de la Parole*, (1985).
3. D A Seggie, *Proc. XIth Int. Congress of Phonetic Sciences*, (1987).
4. D Gabor, *J. Inst. Elect. Eng.*, 93, (1946), p 429.
5. A V Oppenheim and R W Schaffer, *Digital Signal Processing*, (Prentice-Hall, NJ, 1975), p 508.
6. A A G Requicha, *Proc. IEEE*, 68(3), (1980), p 308.
7. H B Voelcker, *Proc. IEEE*, 54(3), (1966), p 340.



- A: Fig. 1 60 ms segment of speech pressure waveform for the vowel [a] .
 B: Fig. 2 Amplitude envelope function for signal in Fig. 1 .
 C: Fig. 3 Instantaneous frequency function for signal in Fig. 1 .
 D: Fig. 4 Fundamental frequency data for entire speech waveform part of which is shown in Fig. 1. "Raw" fundamental frequency estimates denoted by '*'; effect of median smoothing shown by full line.