Proceedings of The Institute of Acoustics

THE APPLICATION OF ANALTIC SIGNAL ANALYSIS IN SPEECH PROCESSING

D. A. Seggie

Department of Phonetics and Linguistics,
University College London

ABSTRACT

Through the use of the Hilbert transform, the real speech pressure waveform can be transformed into a complex signal from which *time domain* amplitude envelope and phase can be uniquely specified. The utility of time domain phase in speech processing is assessed. Real data are used to demonstrate the potential of this signal attribute for voicing determination, fricative consonant identification, and conveying information about changes in formant structure

INTRODUCTION

This paper examines the application of analytic signal analysis in speech processing. The basis of the analysis is the treatment of the speech pressure waveform as the real part of a complex or analytic signal. Signal manipulations such as this, where the data are transformed from one form into another, are common in speech signal analysis. For example, in short-time Fourier analysis [1] the Fourier integral is used to transform from the time domain to the Fourier (or frequency) domain. This transformation brings to light average properties of the speech signal which are much less evident in the time domain. The objectives of analytic signal analysis are the same as those in short-time Fourier analysis, i.e. to gain new insight into how information is encoded in the speech pressure waveform, and to uncover relationships not otherwise evident. However, unlike short-time Fourier analysis, the representation of the data in complex form is a transform technique which retains local significance.

Analytic signal analysis allows the relatively straightforward separation of signal amplitude envelope and temporal phase. Of these two time domain signal attributes, the former is perhaps the more familiar since it relates to signal energy. To date, the interpretation of signal time domain phase in speech analysis, (and indeed in the analysis of acoustic data generally), has not been fully explored. Here, real speech data obtained from a normal adult male are used to examine the latent information encoded in the temporal phase or, equivalently, instantaneous frequency (time derivative of phase) of speech signals.

TIME DOMAIN AMPLITUDE, PHASE, AND FREQUENCY

Given a real bandlimited speech pressure waveform, $s(t)$, a complex function may be associated with it. This complex function, $\tilde{s}(t)$, known as the analytic signal, is defined as [2],

$$\tilde{s}(t) = s(t) + jH[s(t)]$$

where $j = \sqrt{-1}$,

and,

$$H[s(t)] = \text{Hilbert transform of } s(t) = \frac{1}{\pi} P \int \frac{s(t)}{t' - t} dt$$

(P denotes Cauchy principal value)

The use of the analytic signal makes it possible to define the following *time domain* signal attributes. Signal amplitude envelope, e(t), is defined as,

$$e(t) = \left|\tilde{s}(t)\right| = [s^2(t) + (H[s(t)])^2]^{1/2} \quad (1)$$

Signal instantaneous phase, $\phi(t)$, is defined as,

$$\phi(t) = \arg\left[\tilde{s}(t)\right] = \arctan(H[s(t)]/s(t)) \quad (2)$$

and a time-dependent frequency known as instantaneous frequency is given by,

$$v(t) = \dot{\phi}(t)/2 \quad (3)$$

where $\dot{}$ denotes derivative with respect to time. It should be stressed that there is no one-to-one correspondence between time domain and Fourier domain frequencies. The former is a signal attribute defined at every instant in time throughout the signal duration. The latter relate to the sinusoidal oscillations, (defined only over the infinite time domain), into which the signal may be decomposed.

Note that signal phase as defined above is modulo 2π , and must therefore be "unwrapped" prior to differentiation to avoid discontinuities in v(t) whenever $\phi(t)$ extends beyond 2π . Here, all derived signal instantaneous frequency functions were obtained via the analytic signal approach, using a standard unwrapping algorithm after Schafer [3].

The quantities defined in expressions (1), (2), and (3) are illustrated in Figs. 1-4. Figure 1 is a 60 ms segment of a real speech pressure waveform for the simple vowel [ɑ:]. The waveform was low-pass filtered, (cut-off frequency = 5 kHz), and digitized at a sampling rate of 20 kHz, to a maximum amplitude resolution of 12 bits. Data acquisition and all subsequent processing were carried out using a Masscomp MC-5500 computer. Figure 2 shows the amplitude envelope of the signal depicted in Fig. 1. Observe that this function corresponds to an intuitive notion of the envelope of the speech waveform. Figure 3 shows the unwrapped phase function for the signal in Fig. 1. Differentiation of the unwrapped phase gives signal instantaneous frequency which, as shown in Fig. 4, highlights the phase fluctuations of the speech pressure waveform.

As suggested by Fig. 3, the unwrapped phase of speech signals may be separated out into two components. The first is a ramp-like component, $\omega(t)t$, with slowly varying (or constant) gradient. The second, $\theta(t)$, corresponds to the small, more localized fluctuations

ANALYTIC SIGNAL ANALYSIS

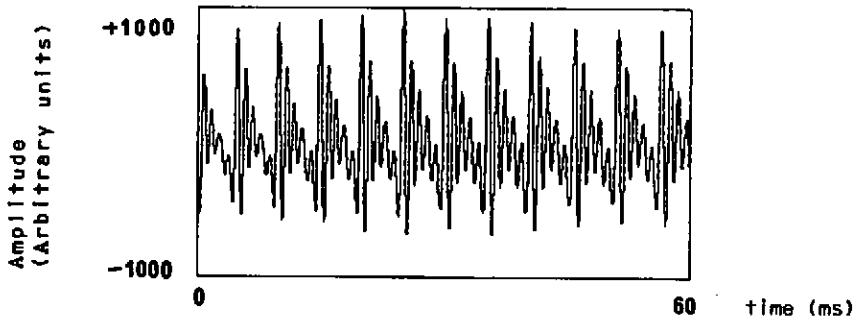

Fig. 1 Segment of speech pressure waveform for [a:].

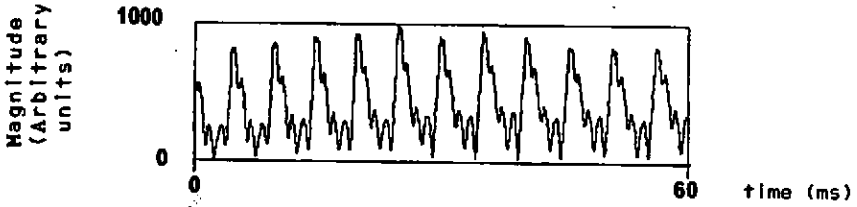

Fig. 2 Amplitude envelope of signal in Fig. 1

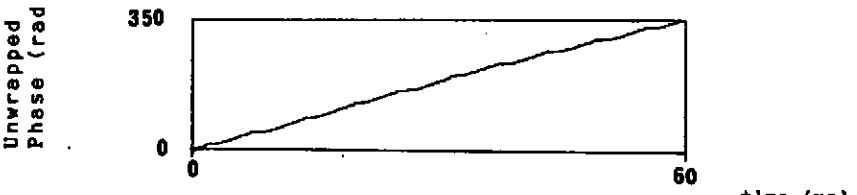

Fig. 3 Unwrapped phase of signal in Fig. 1

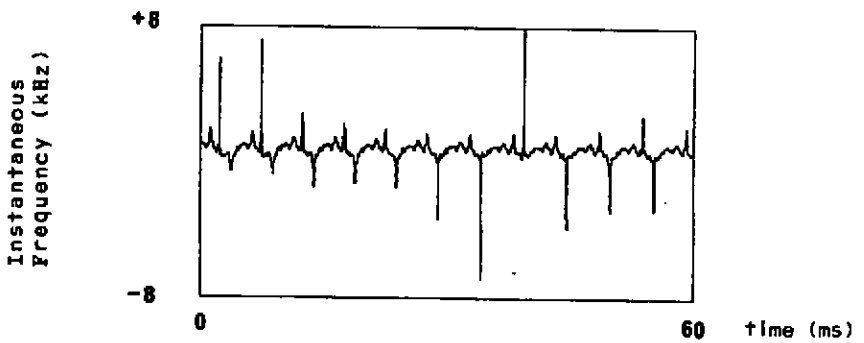

Fig. 4 Instantaneous frequency of signal in Fig. 1

from the almost linear-with-time behaviour of the first component. $\phi(t)$ can thus be written as,

$$\phi(t) = \omega(t)t + \theta(t)$$

and,

$$\dot{\phi}(t) = \omega(t) + \dot{\omega}(t)t + \dot{\theta}(t)$$

Provided that $\omega(t)$ is a sufficiently slowly varying function, such that for all times of interest, $|\dot{\omega}(t)t| \ll |\dot{\theta}(t)|$, then the instantaneous frequency can be approximated to,

$$v(t) \approx \{ \omega(t) + \dot{\theta}(t) \}/2\pi$$

Hence, $v(t)$ can be thought of as the sum of two components. One a slowly varying mean frequency (or carrier frequency), $\omega(t)$, about which signal instantaneous frequency is approximately centred. The other, a more rapidly fluctuating component, $\dot{\theta}(t)$, which would for example describe the very large instantaneous frequency fluctuations at discrete locations in the speech pressure waveform, (see Fig. 4). Such large, time-localized fluctuations have been shown [4,5] to be an intrinsic feature of the broad class of signals to which speech pressure waveforms belong, (i.e. bandlimited, simultaneously amplitude/phase modulated signals). The physical significance of these large excursions in speech signal instantaneous frequency is currently under investigation. In this communication attention is focused on the information encoded by features of the more slowly varying instantaneous frequency component, $\omega(t)$, from here on referred to as the carrier frequency.

APPLICATIONS OF CARRIER FREQUENCY ESTIMATION

Voicing Determination

The utility of speech signal instantaneous frequency estimation is illustrated by Figs. 5-7. Figure 5 shows a speech pressure waveform for the token [ɑ:sɑ:]. The corresponding signal instantaneous frequency is depicted in Fig. 6. Observe that both the instantaneous frequency structure and the carrier frequency value in particular, demarcate the noisy and quasi-periodic portions of the speech waveform associated with voiceless and voiced segments, respectively. An estimate of the relatively slowly varying carrier frequency was obtained by weighting the richly structured speech signal instantaneous frequency, by the corresponding amplitude envelope squared. That is, $\omega(t)$ was estimated by computing the quantity,

$$\int v(t)e^2(t)dt \; / \; \int e^2(t)dt \qquad (4)$$

The result obtained, using a 20 ms rectangular window, is given in Fig. 7. (It can be shown that the estimate of $\omega(t)$ obtained using the above expression, is related to the centre of the speech

ANALYTIC SIGNAL ANALYSIS

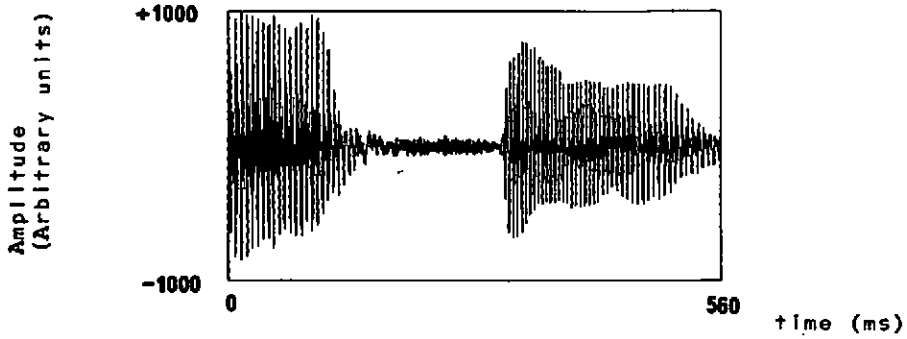

Fig. 5 Speech pressure waveform for the token [a:Sa:] .

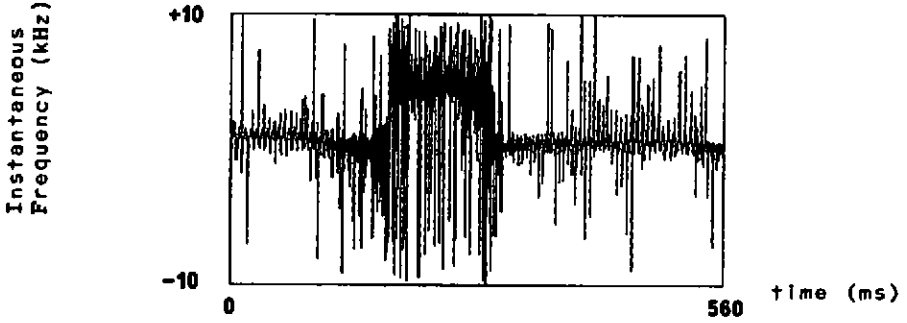

Fig. 6 Instantaneous frequency for signal in Fig. 5 .

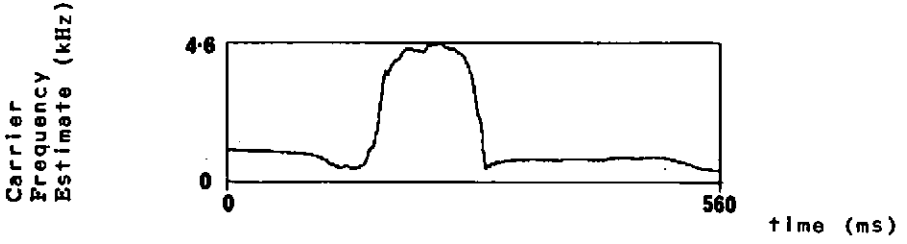

Fig. 7 Estimate of carrier frequency for signal in Fig. 5 .

signal's spectral distribution, [5, 6]). Figure 7 shows that the effect of this weighting is to minimize the influence of large instantaneous frequency fluctuations, and to emphasize the markedly different carrier frequency values associated with voiced and voiceless segments. Initial results, of which Figs. 5-7 are representative examples, suggest that a carrier frequency value (as calculated via expression (4)) of 1 kHz or less, is indicative of voiced excitation. Both voiceless and mixed excitation segments exhibit carrier frequency values greater than 1 kHz. Carrier frequency estimation would therefore appear to provide a straightforward method for voicing determination.

Fricative Consonant Identification

Initial findings also point to the interesting possibility of identifying fricative consonants via carrier frequency estimation. Figure 8 shows a speech pressure waveform for the token [ɑːʃɑː]. The corresponding carrier frequency, as estimated using expression (4), is depicted in Fig. 9. (As before, a 20 ms rectangular window was used in generating this function). Note that the carrier frequency value for the palato-alveolar fricative segment is approximately

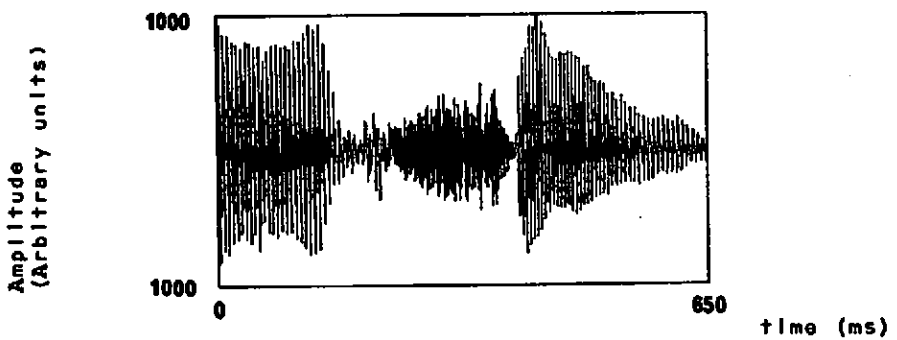

Fig. 8 Speech pressure waveform for the token [ɑːʃɑː] .

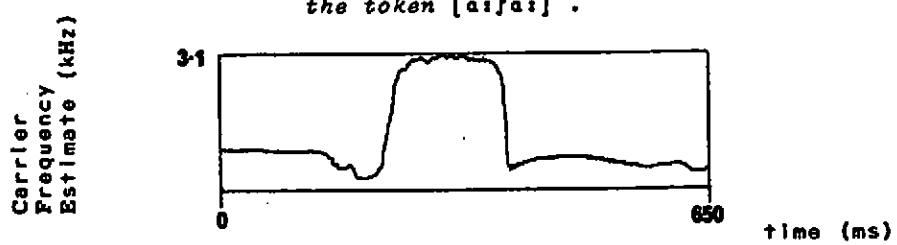

Fig. 9 Carrier frequency estimate for signal in Fig. 8 .

3 kHz, compared with a value of around 4.6 kHz for the alveolar fricative segment seen earlier, (Figs. 5-7).

Further analysis, using different realizations of the same tokens, revealed that the carrier frequency values associated with [s] and [ʃ] are consistently separated by at least 1.4 kHz. Hence for this speaker, knowledge of the speech carrier frequency value alone, is sufficient to differentiate [s] and [ʃ]. The feasibility of carrier frequency estimation for fricative consonant identification generally, is currently being investigated.

Changes in Formant Structure

Variations in carrier frequency can be used to indicate the presence of changes in the underlying formant structure of speech signals. To illustrate this point, a speech pressure waveform and the corresponding carrier frequency estimate for the diphthong [aɪ] are shown in Figs. 10 and 11, respectively. (Again, a 20 ms rectangular window was used in generating an estimate of $\omega(t)$). Comparison of these two figures shows that the change in formant structure as

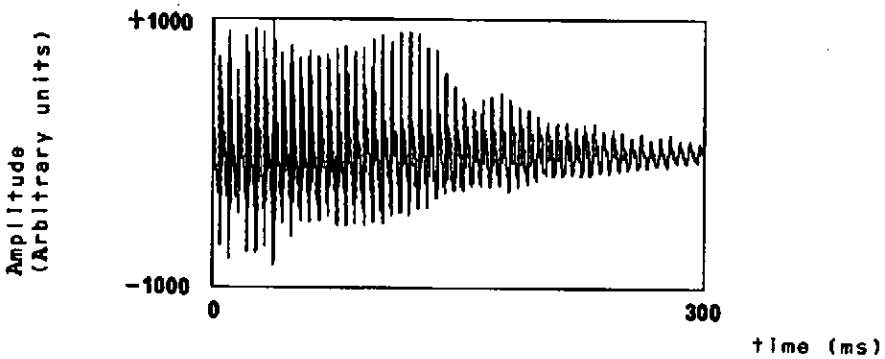

Fig. 10 Speech pressure waveform for the diphthong [aɪ] .

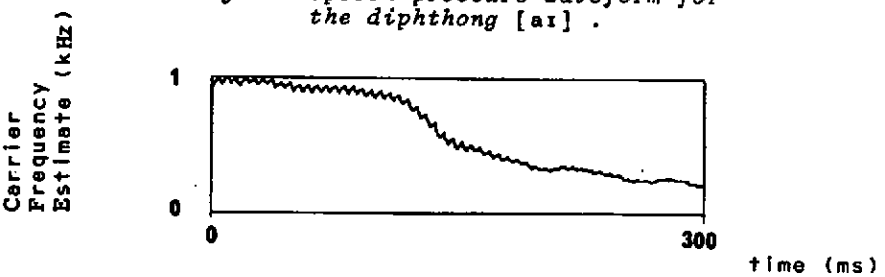

Fig. 11 Carrier frequency estimate for signal in Fig. 10 .

time increases, is reflected by a decrease in the carrier frequency from approximately 1 kHz to around 0.2 kHz. While the information conveyed by Fig. 11 is clearly not sufficient to unravel the formant structure underlined in the speech signal, such carrier frequency variations may well prove to be a useful source of a priori knowledge for formant tracking algorithms.

CONCLUSION

This paper represents only a crude first step in the utilization of speech signal temporal phase. Assuming the promise of these initial results is fufilled, further analysis of temporal envelope and phase fluctuations can be expected to provide valuable insights into how information is encoded in the speech pressure waveform itself, and hence lead to the development of novel *time domain* processing techniques.

ACKNOWLEDGEMENTS

I would like to thank Ian Howard for providing the display software used to generate all eleven figures. This work is part of the programme of the speech pattern algorithmic representations (SPAR) group, financed by SERC-Alvey grant MMI/056.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals", Prentice-Hall, 250-344, (1978).

[2] D. Gabor, "Theory of communication", J. Inst. Elect. Eng., vol. 93, no. 3, 429-441, (1946).

[3] A. V. Oppenheim and R. W. Schafer, "Digital signal processing", Prentice-Hall, 507-509, (1975).

[4] H. B. Voelcker, "Toward a unified theory of modulation", Proc. IEEE, vol. 54, no. 3, 340-353, (1966).

[5] D. A. Seggie, "Digital processing of acoustic pulse-echo data", PhD Thesis, University of London, (1986).

[6] L. Mandel, "Interpretation of instantaneous frequencies", Am. J. Physics, vol. 42, 117-125, (1974).