BRITISH ACOUSTICAL SOCIETY

"SPRING MEETING" at Chelsea College, London, S.W.3 on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'B': Speech Analysis and Transmission.

Paper No:

73SHB5

Formant Frequency Measurement by Waveform Matching During Closed-Glottis Periods

J N Holmes and E M Thornber

Joint Speech Research Unit

Abstract Within the last few years there have been a number of publications of methods of speech analysis in the time domain which, either explicitly or implicitly, assume that voiced speech signals can be regarded as sums of expotential functions where some or all of the complex exponents are closely related to the frequencies and bandwidths of the formants. This assumption is justified during the closed-glottis periods, but when the glottis is open the excitation caused by the air flow and the modification of formant parameters by sub-glottal coupling significantly disturb the derived formant frequencies in a way dependent on vocal cord activity. For applications where formant measurements are required to represent articulatory behaviour this dependence is undesirable. It can be avoided by restricting the analysis to the closed glottis periods, say about 3 ms for adult male talkers. This paper demonstrates that formant measurements based on such short waveform segments normally show less variation between successive glottal cycles than measurements using longer periods which include part of the open-glottis phase. This result supports the hypothesis that segments related to the closed-glottis period are preferable for representing vocal tract behaviour.

Introduction Speech waveforms during voiced sounds are the result of convolving the volume-velocity waveform at the glottis with the impulse response of the vocal tract. Since glottal pulses are normally clearly separated by closed-glottis periods their Laplace transforms can only contain zeros /1/. For non-nasalised vowels the vocal tract transfer function contains only poles /2/. The poles, which lie near the $j\omega$ axis in the complex frequency plane, normally cause significant peaks (or formants) in the speech spectrum, and it is therefore natural to assume that the spectral peaks are simply related to the frequencies of the poles of this transfer function.

However, the above argument assumes that the glottal impedance is so high that the poles of the complete vocal system are the same as those of the supra-glottal vocal tract. In fact, sub-glottal coupling is quite significant /3/, particularly in the first-formant region, and thus the system poles vary periodically at the fundamental larynx frequency. This fact needs to be considered when attempting to define what is meant by the term "formant frequency".

Many of the difficulties of spectrum-based formant tracking methods /4/ can be overcome by using a direct time-domain approach. The difficulties of sub-glottal coupling and the effects on the speech waveforms of excitation during the glottal pulse can be avoided by analysing only such parts of the speech waveform as may reasonably be assumed to correspond to

closed-glottis periods. Under such conditions the waveform should be able to be represented very closely by the sum of a few damped sinusoids of suitably chosen frequency, rate of decay, amplitude and phase. One would therefore expect such a model, if its parameters were optimised using a least-squared error criterion, to give good measures of formant frequencies. This idea is not new; indeed such a scheme was published by Pinson /5/ in 1963, and the present work may be considered as an extension of Pinson's study.

Method It is essential to the method described here to choose portions of the speech waveform which are assumed to represent the force-free response of the closed-glottis vocal tract. Associated with each such portion a second signal is generated by combining a number of damped sinusoids, whose parameters are adjusted to minimize, in the least squares sense, the difference between the original and generated waveforms. Assuming the vocal tract to be stationary and truly force free for the analysis period, the generated signal should fit the speech signal closely, provided that the number of sinusoids is sufficient. The frequencies and damping factors of the sinsusoids are related to the imaginary and real coordinates of the pole positions, and the amplitude and phase are determined by the initial conditions. To achieve a unique set of optimised parameters it is essential that the number of independent ordinates in the input signal should be at least equal to the number of degrees of freedom in the generated signal. This condition requires a minimum section length of $2n/B$ where n is the number of pole pairs of the generated signal and B is the speech signal bandwidth, assuming four degrees of freedom per pole pair.

The efficiency of the optimisation process was not considered important in the present study. The main concern was to determine whether the optimised parameters would give sensible values for the formant frequencies for a wide variety of speech sounds. For this reason a slow iterative method was formulated which incorporates various rules to constrain the results to represent speech-like signals. These rules are still being developed and will not be described here.

The signal processing in these experiments was carried out by a general pupose computer. The signals were stored in sampled-data form at 10,000 samples per second after having been low-pass filtered to 4kHz. Since speech normally has most power at low frequencies the stored waveform was differenced to compensate for the general spectrum trend. This operation should make the error contribution from all formants of comparable magnitude, and so ease the optimisation process.

The regions chosen for analysis have so far been selected visually from plotted waveforms. However the criterion for selection has been to choose local power maxima, and so automatic operation of the selection process should be simple to achieve.

Results and Conclusions In the experimental work it was necessary to choose a suitable length of section for waveform matching. If too short a section were chosen the model parameters would be insufficiently defined; too long a section would entail a danger of disturbance by glottal opening. Previous work has shown typical closed-glottis periods for male talkers of the order of 3 ms, so preliminary tests were made with this value. The program had no difficulty in optimising the parameters during vowels, and the power of the error signal was typically between 20 and 25 dB below the input level. A range of section lengths up to 10 ms was then tested (restricting the longest lengths to low-pitched sounds). The results for 3 ms and 7 ms are illustrated by the graphs in Fig 1, which represent six glottal periods of a strongly nasalised /ɑ/ sound, at a time of very little articulatory

movement. The vowel was chosen to be one in which the vocal cord vibration was very irregular, so that any disturbance of movement caused by glottal opening should show up as a random variation of formant frequencies. The graphs show results for F_1, F_2 and the main resonance caused by nasal coupling (F_N). It can be seen that the 3 ms values show a smooth small movement for F_1 and F_2, and that these movements are very similar in shape for both formants, thus suggesting that they are both the result of the same small articulatory change. The 7 ms graphs show more irregular behaviour for all formants. The error power was around 20 dB below the signal level for the short sections, and 10 dB for the longer sections. The results for these formants seem to be a clear indication that 3 ms is the more suitable period for analysing this speech material. The F_N results are generally more variable than those for the higher formants, but the same conclusions apply.

The sinusoid decay rates, which should be determined from formant bandwidths, were not optimised for the measurements shown in Fig 1, but were fixed at known typical values. Assuming the chosen bandwidth values were correct, this had the advantage of reducing the number of degrees of freedom to three per formant, thus improving the potential accuracy of the other measurements. Some more experimental runs were made with different formant bandwidth values over a range of up to 2:1. These produced only slight differences to the measured frequencies of F_1 and F_2, but the extremes of the range of F_1 bandwidth made significant differences (around 50Hz) to F_N results.

The decision as to whether to allow bandwidths to be optimised for each speech section or to use pre-set typical values depends on which method is likely to give the greater bandwidth error. Analysis was performed using many 3 ms sections allowing the bandwidth to be optimised. For most vowels the measured bandwidths were in the normally accepted range. For rapidly changing sounds and voiceless sounds, bandwidth values tended to vary wildly, often going to the extremes allowed by the program (0 and 200Hz). These results led to the tentative conclusion that there was no overall advantage in allowing the bandwidth to be optimised. Indeed, conditions that gave the most difficulty in measurement of formant frequencies were likely to be aggravated by erroneous bandwidth measurements.

The degree of waveform matching normally achieved can be seen from the examples given in Fig 2. The first two examples are for normal voiced sounds, and for these the error power was 19 and 30 dB below the signal power. The third example was included to show the extent to which a match may be obtained during continuous random excitation. In this case the error power is only 11 dB below the signal. Even for this sound, the formant frequencies are in good agreement with estimates derived by inspection of spectrograms.

In conclusion, it seems that the method of formant measurement described in this paper is capable of accurate measurements of the resonant frequencies of the vocal tract, particularly when the initial assumption of closed glottis is justified. The method should be more accurate than other approaches to formant measurement which do not specifically remove source effects resulting from the glottal pulse shape, but further work is needed to test this conclusion for a greater variety of speech material. The question as to whether the optimisation process could reasonably be implemented for any application of formant measurement in real time has not yet been studied, but it is obviously not a simple task.

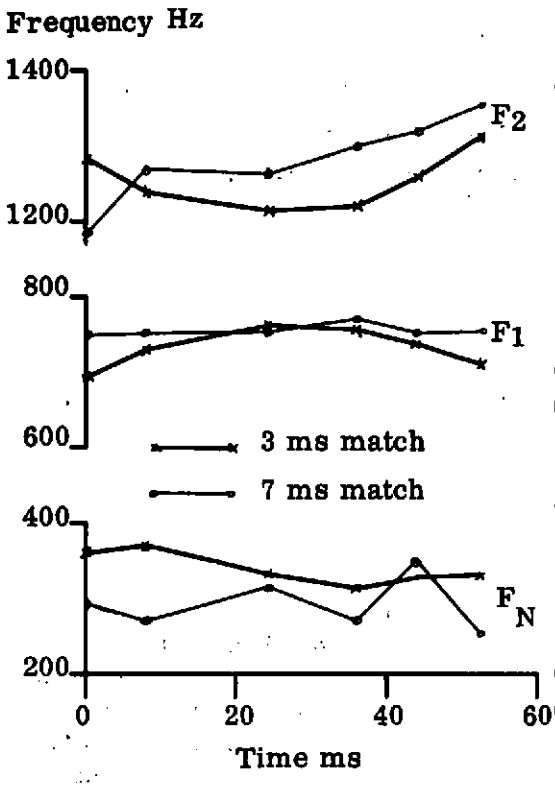

Fig 1 Variation of first, second and nasal formants for consecutive glottal pulses, showing the effect of different matching periods

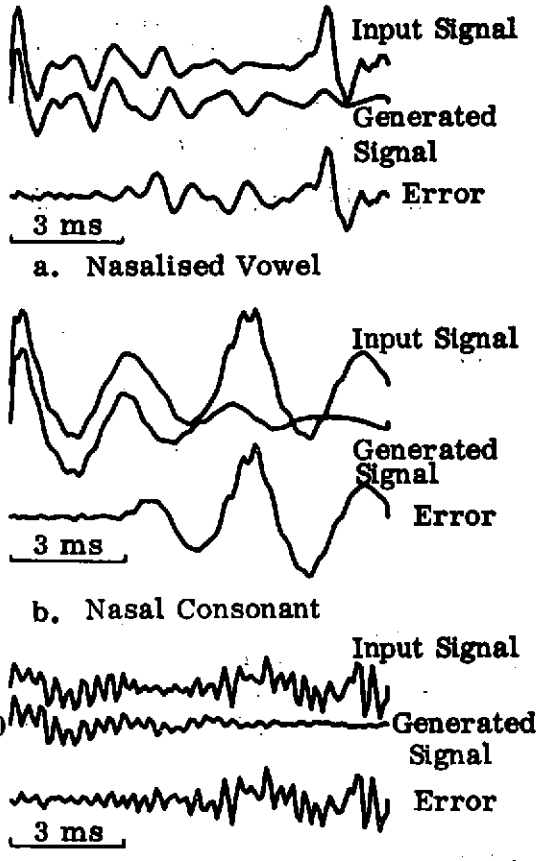

Fig 2 Typical sections of speech waveform, shown with the synthetic approximations chosen for optimum matching over the first 3 ms.

References

1. M.V. Mathews, J.E. Miller and E.E. David Jnr., "Pitch Synchronous Analysis of Voiced Sounds". J. Acoust. Soc. Am. Vol. 33, p. 179 (1961).

2. C.G.M. Fant. "Acoustic Theory of Speech Production". s'Gravenhage Mouton and Company (1960).

3. J.L. Flanagan, "Speech Analysis, Synthesis and Perception". Berlin: Springer - Verlag 2nd Edition (1972) p. 65.

4. Ibid., p.165

5. E.N. Pinson, "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", J. Acoust. Soc. Am. Vol. 35, p. 1264 (1963).