The precision of formant frequency measurement from spectrograms and by linear prediction

Wood, S.

journal: STL-QPSR
volume: 30
number: 1
year: 1989
pages: 091-094

http://www.speech.kth.se/qpsr
THE PRECISION OF FORMANT FREQUENCY MEASUREMENT FROM SPECTROGRAMS AND BY LINEAR PREDICTION

Sidney Wood
Dept. of Phonetics, Umed University

INTRODUCTION
The difficulties involved in measuring formant frequencies have been well known since the early days of the spectrograph, and involve errors related to (i) the ambiguous definition of the object to be measured, (ii) spectral features of the speech wave, (iii) intermodulation distortion, (iv) the spectrographic record, and (v) the measuring procedure:

- A formant is seen both as a spectral prominence in the speech wave and as a filter property of the vocal tract; a definition comprising both components contradicts itself; a definition embracing just the first component presupposes that the relevant information for speech perception is immediately available in the speech wave; a definition based on the second part alone is production oriented and sees the true formant value as a vocal tract pole frequency that is being measured from its (sometimes poor) reflection in the speech wave.

- The resolution of the spectral envelope depends on the interval between the partials, which is equal to the fundamental frequency; a spectral peak may be asymmetrical within the formant band; individual spectral peaks become less well defined as they approach each other or as their bandwidths increase.

- The spectral properties of the original signal may be distorted by tape recorders, amplifiers and spectrograph.

- The spectrogram itself has very diffuse contours.

- Parallax is introduced when reading scales; a pencil mark may span 25 Hz on a spectrogram.

And yet, right from the beginning, one can find claimed accuracies of 5 or 10 Hz in measurements from spectrograms. My own experience is that a 25 Hz interval is too optimistic since most measurements landed in the 50 and 100 Hz intervals and fewer in the 25 and 75 Hz intervals. This skewness disappeared when a 33 Hz interval was used. That the accuracy is no better than this is confirmed by Lindblom (1962), who reported a standard deviation of 40 Hz for errors. That study, where 5 experienced investigators measured spectrograms of synthetic vowels, also confirmed that the error is proportional to the fundamental frequency. Lindblom doubted whether it would be possible to find a general formula that would automatically give the formant frequency and recommended a spectral matching technique where the investigator’s knowledge and experience guided the selection of a suitable standard envelope for comparison with an actual spectrum.

Since then, the linear predictive method (LP) has come into general use and Monsen & Engebretson (1983) have reported a comparison of formant measurements of synthetic vowels by both methods, that confirms Lindblom’s trial. They report similar errors by both methods for F1 and F2, but LP was definitely superior for F3. LP was more easily deceived by proximate peaks. Both methods consistently underestimated F1 and overestimated F2 and F3. LP was less sensitive to fundamental frequency from 100 to 300 Hz. Monsen and Engebretson’s results can also be expressed as follows: the relative error was about 10% for F1 and about 3% for F2, while for F3 it was about 2% by LP
FONETIK -89

and 4% by spectrograms. This comparison was done on spectra produced by parallel synthesis, whereas the LP method presupposes a serial model of speech. When they repeated the LP trial on spectra produced by serial synthesis, the mean errors were halved to 31 Hz (F1), 40 Hz (F2) and 26 Hz (F3), suggesting that LP should be superior on natural speech on all formants.

FORMANTS IN NATURAL SPEECH

Recordings of some 60 words, originally used for an investigation of vowel reduction in Bulgarian (Wood & Pettersson, 1988), were analysed both from spectrograms and by linear prediction, and the two analyses compared.

Unlike the cited trials on synthetic speech by Lindblom and by Monsen and Engebretson, there is no correct answer known in advance in natural speech. However, Monsen and Engebretson’s results (M&E) can be used as a yardstick.

For F1, the M&E results show that spectrogram analysis underestimates the true frequency by about 10%; LP on natural speech should do better, i.e. F1 computed by LP should be higher than the corresponding measurements from spectrograms. The mean difference was indeed LP higher, by 34 Hz on all stressed vowels and 26 Hz on all unstressed vowels. The magnitude of this difference compares well with the 50% improvement obtained by M&E on serial synthesis. LP has presumably underestimated F1 in the Bulgarian vowels by about 5%.

For F2, the M&E results show that spectrogram analysis overestimates the true frequency by about 3%; LP on natural speech should do better, i.e. LP results should be somewhat lower. The mean difference was LP lower by 9 Hz on stressed vowels and 21 Hz on unstressed vowels, i.e. an improvement over spectrogram analysis and closer to the expected true value.

It was not meaningful to measure F3 on the spectrograms as it was not sufficiently well defined on a number of vowels, especially when unstressed. LP gave a solution for virtually all vowels. These results can probably be taken with confidence since LP did so well on F3 in the M&E trial.

The programme used for LP analysis was the ILS API routine, using 10,000 Hz sampling frequency, 5 ms data frames and a 20 ms analysis window. The order of analysis was 16 for 6 poles in the 5000 Hz frequency range. The usual recommendation is an order equal to twice the number of poles plus a small margin, but Momsen and Engebretson found that an order of at least 20 gave the best results.

The poor performance of LP on proximate formants, as in grave vowels like [u,o], improved considerably when the programme was required to look for a reasonable number of poles in the frequency range rather than the default 4 poles. It rarely found all 6 but it usually succeeded in finding 5; when asked to find 4 poles it frequently only identified the two most prominent.

The fine temporal resolution offered by 5 ms LP frames forces the investigator to think more carefully about the moment in time where the spectrum is to be sampled. Every frame contains a solution that is different to its neighbour’s. A spectrogram is usually inspected and sampled where the formant transitions show least influence from surrounding consonants. When using LP it is necessary to be able to justify why one frame rather than another is selected. The procedure adopted for the Bulgarian material was based on the experience of frame by frame analysis of x-ray motion films, where it was found that the most complete configuration (where all relevant manoeuvres had arrived on station) coincided with maximum mandible depression and maximum F1 frequency and was immediately followed by elevation of the mandible and a fall in F1.
The speech frame selected in the LP analysis was the F1 maximum before the offset fall.

Lindblom's advice is thus still valid today. It is still necessary to apply one's knowledge and experience of speech production and expected envelope shapes to the problem of how to select samples to measure and where to look for spectral peaks.

References

