A METHOD OF MEASURING FORMANT FREQUENCIES AT HIGH FUNDAMENTAL FREQUENCIES

Hartmut Traunmüller
Dept. of Linguistics
Stockholm University
S-106 91 Stockholm, Sweden
hartmut@ling.su.se

Anders Eriksson
Dept. of Phonetics
Umeå University
S-901 87 Umeå, Sweden
anderse@ling.umu.se

ABSTRACT

Accurate measurement of formant frequencies is important in many studies of speech perception and production. Errors in formant frequency estimation by eye, using a spectrogram, or automatically, using linear prediction, have been reported to be as high as 60 Hz at F_0 < 300 Hz. This exceeds the typical auditory difference limens (DLs) for formant frequencies and is also greater than some of the variation that one would like to study, e.g., the acoustic effects of varying vocal effort. The problem becomes substantially worse when F_0 is as high as 500 to 600 Hz, which is not uncommon in the speech of women and children at high vocal efforts. In comparison with ordinary linear predictive analysis, the method described here drastically reduces measurement errors, given that the formant frequency is not below or only slightly above F_0 (which rarely happens in speech). It thus becomes possible to study formant frequency variation in speech material that hitherto could not be analysed meaningfully since the effects of interest were no larger than the probable errors in measurement.

1. INTRODUCTION

Measuring formant frequencies in speech is technically problematic. At the low fundamental frequency (F_0) that is typical of male speech produced at moderate vocal effort, the problem is less severe, but it increases very noticeably with increasing F_0. This holds not only for methods based on inspection of spectrograms, but also for automatic methods based on linear prediction (LP).

Atal & Schroeder [1] analysed the errors obtained by LP. They showed formant frequency estimates to be affected by a bias towards the partial closest to the centre frequency of the formant, and the estimated bandwidths to be too low when the formant was close to a partial and too high when it was in the middle between two partials. These systematic errors increase with F_0/B, where B is the bandwidth of the formant. Ironically, these kinds of errors are absent in unvoiced segments and in whispered speech, where the signal prediction error tends to be largest (in relative terms).

Monsen & Engebretson [2] compared formant frequency measurements obtained by LP with measurements by inspection of spectrograms. Using LPC, they obtained errors up to about ±60 Hz for F_1, F_2 and F_3 at F_0 < 300 Hz. For F_1 and F_2, similar errors were obtained in measurements made from spectrograms. For F_3, the results obtained by LPC were clearly more accurate than those obtained from spectrograms.

Wood [3] compared spectrogram and LP measurements of formants in naturally produced Bulgarian words. Assuming Monsen & Engebretson's [2] results to be qualitatively correct, Wood estimated average improvements in measurement accuracy, using LP, to be 34 Hz and 26 Hz for F_1 in stressed and unstressed vowels respectively. The corresponding estimates for F_2 were 9 Hz and 21 Hz.

Miller *et al.* [5] tested accuracy and reliability of techniques for acoustic analysis of infant speech, using 64 synthesised tokens covering the range of acoustic variation encountered in natural speech. Three trained phoneticians measured the frequencies of F_1, F_2, and F_3 at signal onset, midpoint and offset, essentially by inspection of FFT-spectra. At high F_0s, this method provided a higher accuracy and reliability for measurements of F_1 and F_2, when compared with both LPC-based and spectrogram based measurements.

Smits [5] investigated the performance of spectrographic and LP techniques in measuring formant frequencies and transition rates in highly dynamic speech with special regard to the assumption of quasistationarity that is inherent in some of these techniques. He showed that wide band spectrograms and LP using an analysis window whose effective length is equal to a glottal period are best suited for the accurate measurement of formant transitions. When the analysis window included several glottal periods, there was a tendency for moving formants to be represented by several simultaneous resonances, each originating in a different one of the glottal periods.

Human auditory difference limens (DLs) for formant frequency were investigated by Flanagan [6], who obtained an average of 3.9% for F_1 and F_2. Nord & Sventelius [7] obtained 3.3% with similar stimuli synthesised at an F_0 of 120 Hz. Mermelstein [8] obtained DLs for isolated vowels with stationary formants and an

F_0 of 120 Hz of 14% for F_1 and 7% for F_2. For vowels in a CVC contexts, lager DLs were obtained for F_2.

The DLs observed with stationary formants in all these studies were of the same order or smaller than the measurement errors reported in the studies on measurement accuracy mentioned above. It is therefore desirable to be able to measure formant frequencies with greater accuracy even at the extremely low F_0 that is typical of adult male speech produced at low to moderate vocal effort. With any traditional method of measurement, the accuracy deteriorates considerably with increasing F_0. We would expect the auditory formant frequency DLs to increase with F_0 as well, but we are not aware of any investigation of such DLs at higher F_0s.

In a study of the acoustic effects of variations in vocal effort [9], the authors of the present study found that F_0s above 500 Hz were not uncommon in the speech of women and children at high vocal efforts. When speakers directed their speech to a person 190 m away in an open field, the average F_0 and its standard deviation in a 9-word utterance was 267±34 Hz for men, 417±83 Hz for women, and 512±97 Hz for 7-year olds. This is above the range looked at by Monsen & Engebretson [2] and by Miller *et al.* [4].

The formant frequencies have been reported to raise with increasing vocal effort ([10]; [11] for references), but the uncertainty in the measurements appears to increase with F_0 even more than the formant frequencies. Although the method described by Miller *et al.* [4] achieves some improvement, its application is quite time consuming and not sufficient. We would, therefore, rather have an automatic method by which formant frequencies could be measured with substantially increased accuracy. Linear prediction is such a method and to the extent that the errors it leads to are predictable, they can certainly be avoided. The method described in the following achieves more than this.

2. METHOD

The method developed in order to solve the problem involved an analysis by synthesis procedure based on linear predictive coding (LPC). The basic idea was that a synthetic signal should be found that produces the same result as the natural signal when subjected to LP analysis. It can then be assumed that the synthetic signal, the values of whose acoustic parameters are known, is very similar to the natural signal. In generating this synthetic signal, the inverse of the LP analysis procedure is to be used. This also involves the use of the type of excitation assumed by the method.

The procedure involves the following steps:

(1) LP analysis of a natural speech signal, including an estimation of formant frequencies and bandwidths.

(2) Synthesis of a signal based on the analysis result of step (1), using spike excitation.

(3) LP analysis of the synthetic signal (2), including an estimation of formant frequencies and bandwidths, with all settings for the LP analysis being the same as in step (1).

(4) For each formant frequency and bandwidth that shall be corrected, calculation of the error resulting from synthesis and analysis, i.e., the deviation of the value obtained by analysis (3) from the synthesis parameter value, which was taken from analysis (1). The working hypothesis is that the unknown errors, inherent in the original analysis are similar in polarity and magnitude to these.

(5) The error obtained in (4) is subsequently taken as an estimate of the error in the original analysis (1) and it has to be decided how the values of the synthesis parameters should be modified in order to minimise this error. A rough guess is based on the assumption that a formant frequency error of +n Hz can be compensated by modifying the frequency of that formant by –n Hz in the synthesis. However, in order for the procedure to converge more rapidly, a more accurate calculation of the necessary compensation is used, as detailed below. As for formant bandwidth, it is assumed that a change by a factor of n can be roughly compensated by modifying the bandwidth in the synthesis by a factor of 1/n.

(6) Synthesis of a signal based on the analysis result of step (1), but now with compensation according to step (5) in the values of the centre frequencies and bandwidths of the formants.

(7) Repetition of the procedure (3...6) until the analysis result of the synthetic signal is satisfactorily similar to that of the natural signal.

The relation between the magnitude of the observed errors and the amount by which the formant frequencies have to be modified in order to compensate for the errors is, inconveniently, far from linear when F_0 is larger than the bandwidth of the formant in question. Therefore, the actions described in step (5) would in certain cases be quite inefficient, and the process would require many repetitions to converge.

This has been improved substantially by employing a more accurate estimate of the necessary compensation for formant frequency errors. It involves a weighting of the formant frequency correction as a function of the change in the bandwidth of the formant. In the first loop, the new formant frequencies are calculated as

$$F'' = F - (F' - F) B/B',$$

where F and B are the original values (1), F' and B' the values obtained in step (3) and F" the value used in step (6). In this way, the formant frequency values obtained after the first loop are already quite close to the final values. Convergence is speeded up by multiplying the error terms by a factor > 1 when the errors in two subsequent loops have the same polarity. If this is not the case a factor < 1 is used to prevent oscillations which may delay or block convergence.

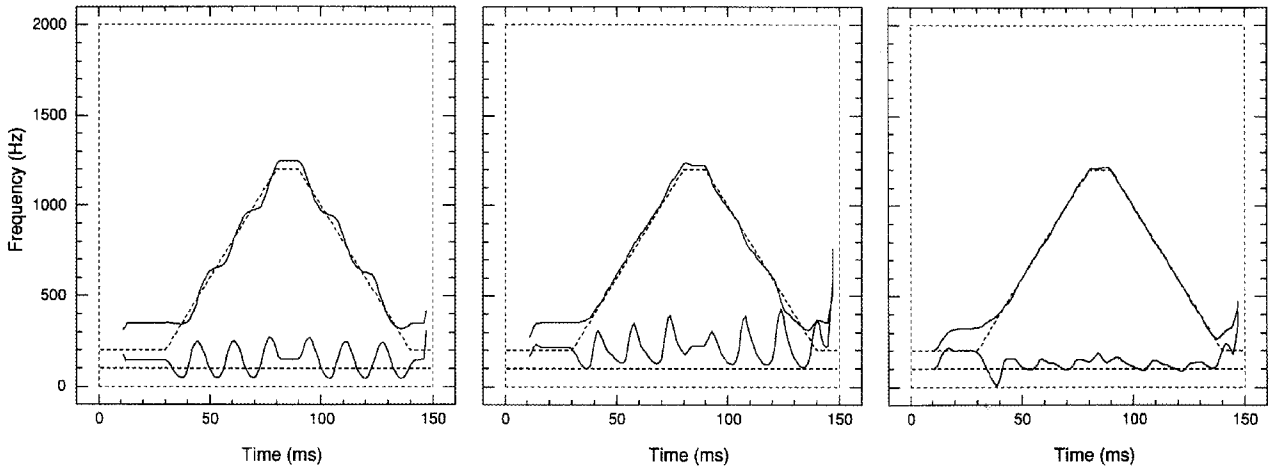

Figure 1. The left panel shows the result of an LP analysis of a synthetic signal (F_0 = 320 Hz) and its synthesis parameters, B_1 = 100 Hz, F_1 = 200 ... 1200 ... 200 Hz (dotted lines). The middle panel shows the results after the first correction and the rightmost panel after six iterations.

Since the main objective was obtaining more accurate values of the formant frequencies, no effort has been made to optimise the bandwidth estimation. However, at very high F_0s, the original bandwidth values tend to be drastically too low (B < 10 Hz) when the formant is close to a partial. In order to avoid such unrealistically small values, a limitation of the type

$$B' = (B^2 + 40^2)^{1/2}$$

has been introduced in the initial loop.

In principle, the method allows correction of any errors that conserve their polarity and are amplified by iterative LPC, although the method is subject to the limitations given by the not quite realistic assumptions of a spike-shaped excitation pulse for each glottal period and the absence of spectral zeros, which are inherent in the basic type of LPC.

3. TEST RESULTS AND DISCUSSION

The method was tested with various synthetic signals that were generated using LPC technique including one with F_0 = 320 Hz and a first formant that varied between 200 and 1200 Hz. Second and higher formants were kept constant at 1650, 2750, 3850 Hz etc. The signal was sampled at 16 kHz with 16 bit/sample. A Hamming window with a length of 20 ms was used and this was moved forward in steps of 5 ms. This test signal illustrates several crucial points, high fundamental frequency, formant frequencies at or below the fundamental, and a rapidly changing first formant.

The F_1 values used in the synthesis, the result of the initial LP analysis, and that after the first and the sixth correction loops are shown in Figure 1. The second formant was also subject to correction, but this is not shown in the figure, since the correction required was much smaller and not as telling as that in F_1.

It can be seen that the accuracy in formant frequency estimation is increased substantially already after the first loop of the procedure, while the subsequent improvements are smaller. The bandwidth estimates do not improve so fast. It can also be seen that the method fails when the formant frequency is below or slightly above the first partial. At higher F_0s, the method becomes increasingly sensitive to small fluctuations in the amplitude of partials for formants that are far from the closest one.

Figure 1 illustrates two of the problems with this method.

1) One of these is the necessity to keep the various synthetic versions precisely in synchrony with the original signal. Failure to do so will result in sub-optimal functioning where the formant is moving. In Figure 1, the result after the first correction shows some clear signs of asynchrony. Where F_1 was moving up in frequency, this accidentally improved the frequency correction, while the opposite effect can be observed, where F_1 was moving down. The asynchrony also caused some inefficiency of the bandwidth correction.

2) The other problem occurs when F_1 is very close to or below F_0. In this situation, the method works less well or fails, since the basic working hypothesis does not necessarily hold. This kind of error is due to the absence of spectral components below the first partial, which leads to an overestimation of the formant frequency. After the sixth 'correction', in this example, the value of B_1 approaches zero when F_1 is just above F_0. It can, however, also be seen that this did not happen towards the end of the signal, where F_1 crosses F_0 in the other direction.

A further problem is the risk that the association of reflection coefficients and formants might change

between subsequent analyses. This can be made less likely to happen by leaving the formants with a more favorable value of F_0/B (above F_2 or F_3) without correction.

At the juncture between voiced and voiceless segments and at places with very rapid formant transitions, the procedure is not 'well behaved' due to severe problems with both synchrony and formant association. The method does not either solve the problem concerning the tendency for rapidly moving formants to be represented by several simultaneous resonances when the analysis window includes several glottal periods [5].

According to Harris (1978), the relation between frequency resolution ΔF, defined as the -6 dB bandwidth, and the total length D of a Hanning window is given by $\Delta F \cdot D = 2$. Thus, if we want to resolve individual formants as separate spectral peaks down to a distance of 100 Hz between their centre frequencies, we need a window length of at least 20 ms.

In order to adequately represent rapid formant movements, D should not exceed 2.2 pitch periods [5]. However, if we use this criterion, the first mentioned requirement is only satisfied for speech at $F_0 < 110$ Hz, frequency resolution being limited to $F_0/1.1$. If we skip the requirement that each formant should be represented by a separate peak in the spectrum, which is not required by the LPC method, the situation is not quite as hopeless, since a shorter time window can be used. The optimum effective length of the time window is probably colse to one pitch period (this corresponds to D = 2 with a Hanning window).

4. CONCLUSIONS

At present, the method works satisfactorily only with fully voiced slices of speech that do not include any drastic changes in voicing and formant frequencies. In such slices of speech, it is not unrealistic to expect an error reduction by a factor of ten in formant frequency measurements within the range above 1.5 F_0.

The problems that still remain with synchronisation, voicing onsets and offsets, and rapid formant transitions could all be solved by applying the method exactly period by period. However, this requires a reliable detection of the pitch periods prior to the application of the procedure described here.

5. REFERENCES

[1] B.S. Atal & M.R. Schroeder (1974) Recent advances in predictive coding – applications to speech synthesis, *Preprints of the Speech Communication Seminar Stockholm Aug 1–3, 1974,* Vol. 2, 27–31.

[2] Monsen, R.B. & A.M. Engebretson. (1983) The Accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction, *J. Speech Hear. Res.*, **26**, 89–97.

[3] Wood, S. (1989) The precision of formant frequency measurement from spectrograms and by linear prediction. *STL-QPSR*, **1989:1**, 91–93. Dep. of Speech, Music and Hearing, Royal Inst. of Technology, Stockholm.

[4] Miller, C.J., N. Roussel, R. Daniloff & P. Hoffman. (1991) Estimation of formant frequency in synthetic infant CV tokens. *Clinical Linguistics and Phonetics*, **5**, 283–296.

[5] Smits, R. (1994) Accuracy of quasistationary analysis of highly dynamic speech signals. *J. Acoust. Soc. Am.*, **96**, 3401–3415.

[6] Flanagan, J. (1955) A Difference limen for vowel formant frequency. *J. Acoust. Soc. Am.*, **27**, 613–617.

[7] Nord, L. & E. Sventelius. (1979) Analysis and prediction of difference limen data for formant frequencies. *PERILUS, Report I,* 24–37. Dep. of Linguistics, Stockholm University.

[8] Mermelstein, P. (1978) Difference limens for formant frequencies of steady-state and consonant-bound vowels. *J. Acoust. Soc. Am.*, **63**, 572–580.

[9] Andersson, A., A. Eriksson & H. Traunmüller. (1996) Cries and whispers: Acoustic effects of variations in vocal effort. *TMH-QPSR*, **1996:2**, 127–130. Dep. of Speech, Music and Hearing, Royal Inst. of Technology, Stockholm.

[10] Traunmüller, H. (1997) Perception of speaker sex, age, and vocal effort. In *Phonum*, **4**, 183–186. Dep. of Phonetics, Umeå University.

[11] Traunmüller, H. (1988) Paralinguistic variation and invariance in the characteristic frequencies of vowels, *Phonetica,* **45**, 1–29.

[12] Harris, F. J. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, **66**, 51–83.