

FORMANT ESTIMATION BASED ON TEMPORAL SYNCHRONOUS ANALYSIS

X.D. Huang*, M.A. Jack, and G. Duncan

ABSTRACT

The accuracy of formant frequency estimation on voiced speech in frame-based linear predictive analysis is affected by the position of the analysis frame relative to the instant of onset of vocal tract excitation. An automatic waveform-dependent point-wise analysis which employs a weighted least-square lattice (WLSL) algorithm to minimise these errors is described here. Experiments on both synthetic speech and real speech are included to show that the algorithm offers improved accuracy in comparison to the frame-based method.

INTRODUCTION

Resonant frequencies of the vocal tract (formants) are very important phonetic features both for recognition and synthesis of speech. The problem of automatic formant analysis of speech has received considerable attention over recent years (ref 1-2) and a variety of approaches have been explored. Three primary effects limit the accuracy with which formant frequencies, can be estimated from the spectrum of the speech signal. These are: (1) the effect of the periodic vocal cord excitation, (2) the effect of time averaging over several excitation cycles in the analysis when the vocal cords are repeatedly in open-phase and closed-phase conditions; (3) the effect of the excitation-spectrum envelope.

Because of the periodicity of the excitation, the spectrum derived from short time Fourier analysis of the speech signal consists of lines at the fundamental frequency and its harmonics. The centre frequency of any formant is difficult to estimate if it is located between two such lines, since little information about the spectrum between pitch harmonics is available and the frequencies of the spectral peaks must be interpolated between these lines. Pitch-synchronous analysis can reduce this kind of error, but the exact position of the start of a pitch period is not easy to determine automatically. Linear predictive coding (LPC) based techniques attempt to model the vocal tract, reducing the effect of source periodicity and can provide a good estimate of the envelope of the speech spectrum. However, frame based LPC analysis, as with all the other frame-based short time analysis methods, still cannot reduce the effect of time averaging over several excitation cycles. The quest for truly accurate LPC analysis of voiced speech has led to the proposal (ref 3) that LPC analysis should be performed using only speech samples corresponding to the closed glottal condition, and it has been shown (ref 4) that closed-phase analysis has improved formant tracking properties. Here however, analysis frame boundaries must be carefully positioned with the aid of an electroglottograph. The principal difficulty in closed-phase analysis lies in the accurate determination of glottal state directly from the speech signal itself.

The prediction analysis error is related to the position of the analysis frame, model order and frame length (ref 5), with a large value of residual taken as an indication of excitation onset. In order to improve analysis performance, a waveform-dependent method is proposed here to detect possible excitations in the signal and to de-weight those excitation intervals during the analysis. For every new data sample, an adaptive weighted recursive least-square lattice (WLSL) algorithm is used to generate a new set of reflection coefficients. The likelihood variable derived from WLSL is employed as a measure of the changes in the nature of the observed process. An optimum position for the analysis frame is thus determined according to the local minimum of residual and likelihood variable, which can be expected to occur near the end of the interval of glottal closure. Based on this waveform-dependent analysis, LPC coefficients can then be derived from reflection coefficients at the optimum position. Final formant estimation is based here on a pole enhancement technique (ref 2), which offers high noise-tolerance and good spectral resolution.

LIMITATIONS OF FRAME-BASED LINEAR PREDICTION ANALYSIS

It can be generally assumed that the speech signal is generated by an all-pole model of order p

Centre for Speech Technology Research, Edinburgh University, 80, South Bridge, Edinburgh EH1 1HN, U.K.

* and also Man-Machine Speech Communication Laboratory, Dept. of Computer Science & Technology, Tsinghua University, Beijing, CHINA

represented by the following equation:

$$x_n = \sum_{k=1}^p \alpha_k x_{n-k} + G e_n \quad (1)$$

where x_n denotes the n -th sample of speech signal, e_n is driving source, and α_k is called predictor coefficient. In frame-based LPC analysis, the prediction coefficients are determined by minimizing the sum of squares of the prediction residual over a finite frame interval, leading to the equation:

$$\sum_{k=1}^{k=p} \varphi_{jk} \alpha_k = \varphi_{j0}, j=1, 2, \dots, p. \quad (2)$$

where $\varphi_{jk} = E[x_{n-j} x_{n-k}]$. In Equation (1), the model assumes that the excitation e_n is white. This is not true for periodic signals. The solution presented in Equation (2) is true only if the correlations $E[e_n x_{n-i}]$ ($1 \leq i \leq p$) are zero, requiring that the vocal tract response dies out within a pitch period. However, the vocal tract response often extends over several pitch periods. When the analysis frame contains several vocal tract responses, the frame position relative to the vocal tract responses will affect the accuracy of the analysis, although this can be mitigated through windowing. It has been shown that use of short duration analysis windows which exclude the open-phase or the vocal cord excitation interval can provide improved spectral estimation (ref 3-4). However, exact determination of closure phase (excitation-free) interval is not easy to obtain automatically, and short duration analysis often brings in gross errors due to the sensitivity of the frame position. Use of longer analysis frames can improve frequency resolution but fails to extract rapid changes in the spectrum and the final smooth result is merely the result of time averaging of several inaccurate damped vocal tract responses.

WAVEFORM-DEPENDENT LINEAR PREDICTION

Sequential adaptive analysis methods offer an attractive alternative processing strategy since they overcome many of the compromises of frame-based analysis. Several adaptive algorithms which are applicable to the linear prediction of speech have been proposed (ref 6). The exact Least-Square Lattice (LSL) algorithm (ref 7) allows the exact solution to the least-squares problem to be updated for every newly observed data sample in contrast to the gradient estimation algorithm. One of the most important features of the LSL algorithm is that at every time step, the gains which update the partial correlation are adjusted so that the normal equation for the least-squares problem is exactly satisfied. The likelihood variable $\gamma_{p,T}$ can be computed recursively during the LSL recursion, and has an important interpretation as being a log-likelihood variable related to the process $\{x_i\}$. If the speech signal $\{x_i\}$ is assumed to be a zero mean Gaussian process, a log-likelihood function can be expressed as follows:

$$\log - \text{likelihood} = \ln R_0 + \sum_{i=1}^p \ln(1 - K_i^2) + X_{|T:T-p|}^T R_p^{-1} X_{|T:T-p|} \quad (3)$$

where R_p is the covariance matrix of the process. $X_{|T:T-p|} = [x_T, x_{T-1}, \dots, x_{T-p}]^T$, and K_i is the reflection coefficient. The value of $\gamma_{p,T}$ obtained in LSL recursions can be interpreted as the sample estimate of the third term in Equation (3). The definition of $\gamma_{p,T}$ uses the sample estimate of the covariance matrix $R_{p,T}$, instead of the known covariance matrix, R_p . Thus $\gamma_{p,T}$ is a measure of the likelihood that the $1+p$ most recent data samples $\{x_T \dots x_{T-p}\}$ are derived from a Gaussian process with sample covariance $R_{p,T}$ determined from all past observations. A small value of $\gamma_{p,T}$ indicates that the recent data samples are likely observations from such a Gaussian process. A value of $\gamma_{p,T}$ near unity implies that given the current Gaussian process assumption, the observations are unexpected, and that either the new observations come from a different Gaussian process due to a time-varying nature of the physical process, or that there is a non-Gaussian component in the observation.

It has been shown that time-domain selection linear prediction analysis (ref 3,8) can improve the source and transfer function separation. In such cases, only those speech intervals in which the value of prediction residual is below some threshold are selected for LPC analysis, and thus an almost excitation-free speech signal is fed to final frame-based LPC. In WLSL, both the residual and likelihood variable are used to detect possible excitations. When they are above the threshold, the function $(1-\gamma_{i,T})$ is used as a weighting factor for the original speech signal values to minimise the effects of the excitation on the spectrum. An exponential window with a short time constant is used to track rapid changes in the spectrum. This is similar to the very short window analysis in frame-based LPC. However, attempting to track transient formant events in the speech signal also tracks the pitch excitation signal. When the position is inappropriately placed, a short window may contain only the excitation and cause the final analysis result to be perturbed due to the excitation. To avoid this, the approximate end of closed-phase is determined according to the local minimum of likelihood variable and residual. Experiments show that in those intervals, formant frequency estimates are almost consistent. Fig. 1 illustrates a schematic

diagram of the waveform-dependent algorithm.

EXPERIMENTAL RESULTS

Fig. 2 (a) shows results obtained from point-by-point formant analysis based on LSL on a 30 ms segment extracted from real speech for the word "allow", sampled rate is 16 kHz. It can be seen that the formant position is perturbed at the onset of each excitation. Fig. 2 (b) and (c) show the corresponding likelihood variable and forward residual, which give a measure of the inaccuracy in estimation of the speech spectrum. Comparison with Fig. 2 (a) demonstrates that rapid changes in likelihood variable correspond to perturbations in formant estimation due to the onset of excitation. When the likelihood variable is small and relatively constant in value, the corresponding formant estimation results are smooth. The larger the likelihood variable, the more inaccurate is the estimation of formant frequencies. Fig. 2 shows that frame position is crucial to post formant estimation.

Synthetic speech signals with the same formant parameters but different fundamental frequency were used here to investigate the pitch effect on the formant estimation. The test data are similar to that employed in earlier experiments (ref 9). The synthesis parameters are $F_1=400\text{Hz}$, $B_1=50\text{Hz}$, $F_2=1800\text{Hz}$, $B_2=140\text{Hz}$, $F_3=2900\text{Hz}$, $B_3=240\text{Hz}$, with F_0 varying from 133Hz to 200Hz. The sampling frequency is 10 kHz, the data window used is a 25.6ms Hamming window, and 14-pole standard LPC and waveform-dependent WLSL are used. The results are shown in Table 1, which shows that for LPC, the averaging error rate of F_1 is 4.6%, but the error rate of WLSL is only 2.6%.

Table 1. First formant frequency estimation results.

Pitch	F1	LPC	LSL	WLSL
200	400	400	400	400
189	400	380	380	410
179	400	360	360	385
169	400	350	355	390
160	400	370	380	405
152	400	405	395	420
145	400	415	415	425
139	400	410	410	410
133	400	400	400	400
ave. error rate		4.6%	4.5%	2.6%

Fig. 3 shows the formant estimation results on real speech for the sentence "Our lawyer will allow your rule.". The sample rate for the signal is 16 kHz. Fig. 3 (a) shows the standard LPC method with model order 18, Hamming window of 25.6 ms duration and 5 ms window shift. Fig. 3 (b) is waveform-dependent WLSL with model order 18, exponential weighting factor of 0.995, and excitation threshold of 0.2. In order to compare with frame-based LPC, an optimum position is chosen here within every 5 ms. Experiments on real speech data also indicate that the algorithm performs well both in extracting formants undergoing rapid transitions in frequency and in estimating global signal structure information necessary to facilitate estimation of formant trajectories.

CONCLUSIONS

An automatic waveform-based WLSL method for formant estimation has been presented which is based on de-weighting waveform regions corresponded to excitation onset and choosing the optimum position for the LPC analysis frame according to waveform-dependent likelihood and prediction variables. The influence of voicing periodicity on formant estimation is reduced based on these techniques. The advantages of this algorithm lie in its simplicity (computation is of the same order as for frame-based LPC), its ability to track rapid changes, its estimation accuracy performance, and its flexibility to vary the frame shift. The strategy introduced here can also be applied to any front-end feature extractor for speech recognition systems.

REFERENCES

1. S S McCandless, IEEE Trans. ASSP, Vol.22,p.135,1974
2. G Duncan, M A Jack, Electronics Letters, Vol. 22,p.1213,1986
3. K Steiglitz, B Dickinson, IEEE Trans. ASSP, Vol.25, p.34,1977.
4. A K Krishnamurthy, D G Childers, IEEE Trans. ASSP, Vol.34, p.730, 1986
5. L R Rabiner, B S Atal, M R Sambur, IEEE Trans. ASSP, Vol.25, p. 434, 1977
6. S T Alexander, IEEE ASSP magazine, p.18, 1986
7. D T L Lee, Ph.D. dissertation, Dept. of E.E., Stanford University, 1980
8. Y Miyoshi, et al., ICASSP, p.1245, 1986
9. D H Klatt, Montreal Symposium on Speech Recognition, p. 5,1986

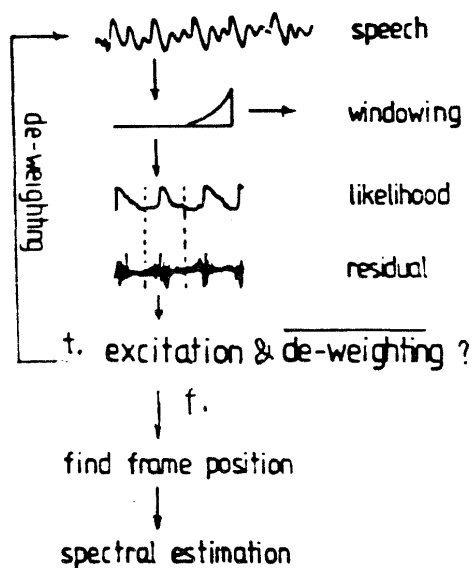


Fig. 1. Schematic diagram of waveform-dependent WLSL.

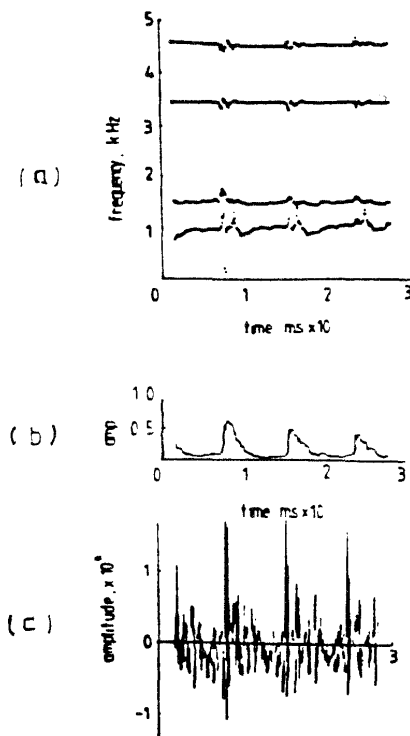


Fig. 2. (a) Formant estimation results obtained on a point-by-point basis; (b) likelihood variable corresponding to the above speech data; (c) residual signal corresponding to (a).

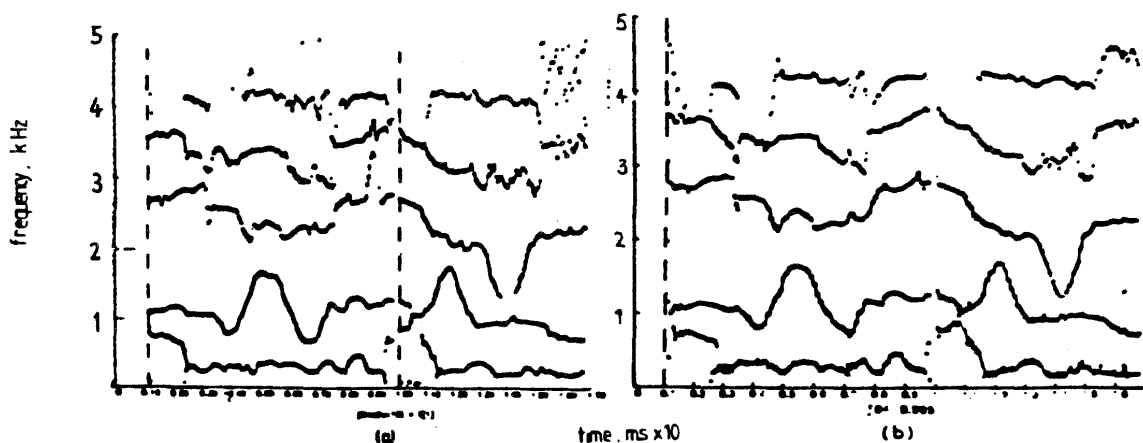


Fig. 3. (a) Standard frame-based LPC on real speech data; (b) waveform-dependent WLSL on the same data.