NEW METHODS OF ANALYSIS IN SPEECH ACOUSTICS

Hisashi Wakita, Speech Communications Research Laboratory Inc., 806 West Adams Boulevard, Los Angeles, California 90007, U.S.A.

Chairperson: Hans Werner Strube

Introduction

The recent development in digital techniques has brought substantial innovations to methods and techniques for acoustical analysis of speech sounds. The advantages of using digital computers over the conventional analog techniques are that the analysis processes can be repeated precisely and that the control of the parameters is relatively easy. The use of the digital computer also permits the processing of a large amount of data within a relatively short period of time with satisfactory accuracy. Because of the above advantages, digital techniques are playing a more and more important role in speech research. As this tendency becomes stronger, proper care has to be taken when the digital techniques are applied to speech research. This paper, thus, concerns primarily the recent digital techniques in the acoustic analysis of speech, particularly the linear prediction method, with special attention to its advantages and disadvantages, and also to the limitations involved in the technique.

The concept of linear prediction was first applied to speech analysis by Itakura and Saito in Japan (1966) and by Atal and Schroeder in the United States (1967). Since then the linear prediction method has been fairly thoroughly studied theoretically and experimentally (see Makhoul 1975; Markel and Gray 1976; Wakita 1976), and the method is currently being used as a powerful tool for acoustical analysis of speech sounds.

Linear prediction of speech

A very simplistic model of speech production as shown in Figure 1 (a) is assumed in the linear prediction of speech. The excitation source is an impulse and the filter, which mainly represents the vocal tract, has the frequency characteristics of resonances only, without any anti-resonances. The model thus exclusively represents the voiced and non-nasalized sounds.

For an analysis model, an inverse filter is assumed, which maintains the precise inverse relation between the input and the output of the production model, as shown in Figure 1 (b). Thus,

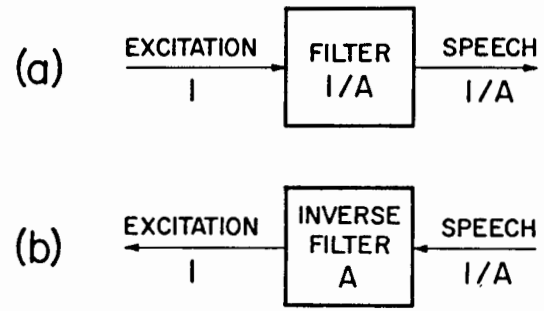


Figure 1. Models for the linear prediction method: (a) Production model; (b) Analysis model.

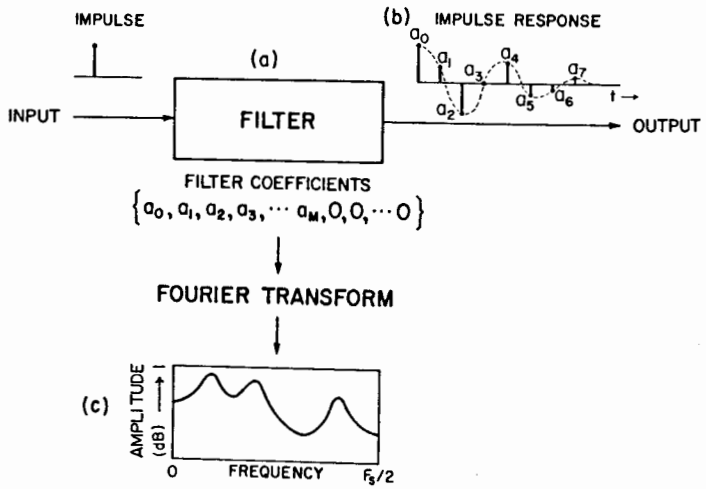


Figure 2. Determination of filter characteristics: (a) a model; (b) discrete impulse response; (c) frequency characteristics (transfer function) of the filter.

the problem in linear prediction analysis is to determine the characteristics of the inverse filter from a given input speech wave.

Since the linear prediction method is a digital technique, all the data, and parameters to specify the filter characteristics, are handled in a discrete sampled format instead of as continuous quantities. The main task of linear prediction is to predict the current speech sample $\hat{x}_n$ in terms of a linear combination of the past M samples. Letting the predicted current sample be $\hat{x}_n$, $\hat{x}_n$ is given by

$$\hat{x}_n = \alpha_1 x_{n-1} + \alpha_2 x_{n-2} + \ldots + \alpha_M x_{n-M} \qquad (1)$$

In equation (1), the α_i's are called predictor coefficients. They play a role of "weighting" the past samples to predict the current one. The problem in the linear prediction method is to determine these predictor coefficients in such a way so as to minimize the error between the current sample and the predicted one, and to relate the predictor coefficients to the parameters of the inverse filter. In this case, the sum of the squared errors over a certain period,

$$E = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2 \qquad (2)$$

is minimized. Because of this, speech samples during this period are assumed to be sufficiently stationary so that the predictor coefficients do not change during this period.

How are the predictor coefficients thus determined related to physically meaningful parameters, that is, to the inverse filter in Figure 1 (b)? In general, the frequency characteristics of a filter can be determined by observing its impulse response when an impulse signal is applied to the filter as shown in Figure 2 (a). In the discrete case, the impulse response of a filter is then given as shown in Figure 2 (b). The amplitude at each sampled point in the impulse response is given by a_i and the period between the two sample points is given by the sampling period T. From this impulse response, the transfer function, A(z), of the filter is given by use of "z-transform" notation as

$$A(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_M z^{-M} \qquad (3)$$

Equation (3) represents not only the transfer function of the filter but also the impulse response in the time domain. The a_i's in equation (3) are called filter coefficients. It is easily seen from Figure 2 (b) that the interpretation of the "z-transform" notation is that z^{-1} represents a unit delay in the time domain in terms of the sampling period T. Thus, the power of z^{-1} in equation (3) denotes the number of time delays.

Since $z=\exp(j2\pi fT)$, where j is the imaginary unit ($j=+\sqrt{-1}$) and f is frequency, equation (3) itself represents the discrete Fourier transform of the impulse response. Thus the frequency domain representation of equation (3) is given by applying the Fourier transform to the filter coefficients. In this case, the impulse response is truncated at t=MT and normally sufficient zeroes (e.g. 256 minus M zeroes) are added to the a_i's to ensure sufficient frequency resolution before the Fourier transform is applied. An example of a power spectrum obtained from the output of the Fourier transform is given in Figure 2 (c). Note that the frequency band is bounded at $F_s/2$ where $F_s=1/T$ is the sampling frequency. Note also that when the amplitude of the frequency components is represented on a logarithmic scale, the frequency characteristics of the inverse filter as shown in Figure 2 (c) become those of the vocal tract filter in Figure 1 (a) just by relabeling the negative sign of the ordinate with a positive sign.

One of the important features of the linear prediction method is that the predictor coefficients in linear prediction of speech can be shown to be identical to the filter coefficients with $a_0=1$. Consequently, minimizing the overall error in linear prediction is equivalent to finding the transfer function of the inverse filter of the analysis model in Figure 1 (b).

Analysis condition

Proper analysis conditions for the linear prediction method are important to ensure satisfactory results. The analysis conditions to be noted are (1) sampling frequency, (2) the number of coefficients, (3) time window and length, (4) window shift, and (5) preemphasis. The sampling frequency determines the frequency range of interest. The frequency range must be less than or equal to half the sampling frequency (normally the latter is chosen). The number of coefficients is dependent on the frequency range to be chosen. When the frequency range is exactly half the sampling frequency (F_s kHz), a good rule of thumb for the number of filter coefficients is from F_s+2 to F_s+4. The reason for this appears to be that there will be about $F_s/2$ resonances in the frequency band limited by $F_s/2$, provided that F_s is given in units of 1 kHz. Each resonance requires 2 coefficients for its representation, and so about F_s coefficients will be needed to account for the expected resonances in the analysis band. In addition, 2-4 coefficients are normally used for approximating the spectral slope due to the excitation source.

The analysis conditions (3) and (4) vary depending upon which of two different methods of linear prediction is used, the autocorrelation method or covariance method (e.g. Markel and Gray 1976). The two methods use different definitions for computing the coefficients from sampled speech. The autocorrelation method requires a window length of at least 1.5 pitch periods and a Hamming window is recommended to suppress the spectral disturbances in the high frequency region due to the edge effect of the time window. The covariance method, on the other hand, does not require any particular time window, and the window length can be less than a pitch period. Thus this method can be used for pitch-synchronous analysis of speech sounds. When a window length of less than a pitch period is chosen, care must be taken since the analysis results vary depending upon what portion of the pitch period is chosen for analysis. This method is particularly useful for extracting the true vocal tract characteristics by choosing the glottis-closed portion of the speech waves. The major disadvantage of the covariance method is that there is theoretically no guarantee for obtaining a stable transfer function for the inverse filter, and thus a more sophisticated algorithm is required to automatically process the cases of instability. Also a more sophisticated algorithm is needed for automatically windowing the speech wave into pitch-synchronous intervals.

The window shift in the covariance method, thus, involves a more complicated procedure than it does in the autocorrelation method. In the latter method, the window shift is rather arbitrary, depending upon the speech samples to be analyzed. The shift can be greater than the window length for steady-state sounds, whereas, for speech sounds in which the formant frequencies are rapidly changing, a smaller window shift will be better for obtaining the smooth contour of the formant frequencies.

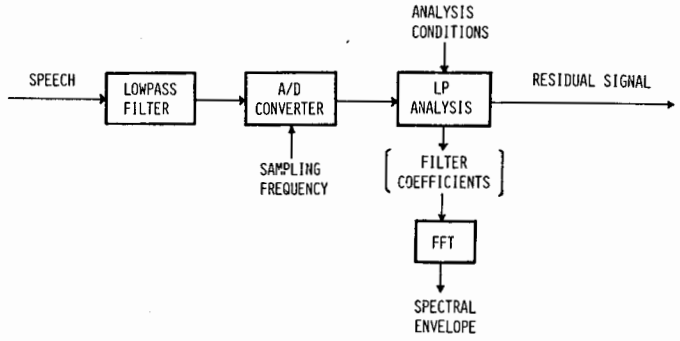


Figure 3. A block diagram to compute the smooth spectral envelopes of speech sounds by the linear prediction method.

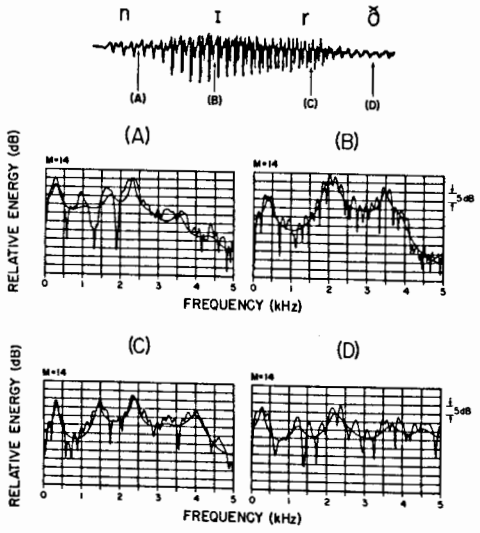


Figure 4. An example of linear prediction analysis. (Sampling frequency 10 kHz; number of coefficients 14; window size 20ms with a Hamming window and +6dB/octave preemphasis.)

A 6 dB/octave preemphasis is recommended for formant analysis. This is accomplished by taking the backward differencing of the sampled speech. The purpose of the preemphasis is to enhance the spectral peaks in the high frequency region. The 6 dB/octave preemphasis also roughly compensates the -12 dB/octave glottal source characteristics and the +6 dB/octave lip radiation characteristics.

Estimation of formant frequencies

As mentioned before, the Fourier transform of the predictor coefficients gives the frequency characteristics of the inverse filter, the inverse of which are the frequency characteristics of the vocal tract filter. Thus the procedure for obtaining the smooth spectral envelope by use of the linear prediction method is given by the block diagram shown in Figure 3. The speech signal is first digitized at some sampling frequency after being passed through a lowpass filter to limit the frequency band according to the sampling frequency. Linear prediction analysis is then performed using predetermined analysis conditions, and resulting in a set of filter coefficients for each speech segment analyzed. Smooth spectral envelopes are computed from the output of the Fourier transform of the filter coefficients with added zeroes. As a result of linear prediction analysis, the residual signal, which is an error signal given by equation (2) is saved for detecting pitch periods as will be described later.

An example of analysis results is shown in Figure 4. This example is a part of a sentence "Near the boat ..." and the spectral envelope estimation for /n/, /ɪ/, /r/, and /ð/ are shown in the figure together with the direct Fourier transform of the corresponding speech waves. It is seen that spectral peaks are well approximated by the extracted spectral envelope. However, the spectral dips due to anti-resonances as in the sound /n/ are ignored in the linear prediction method, in which the nasal tract is not considered. It should be noted that the linear prediction method was developed as a method for efficient speech analysis-synthesis telephony on the basis of the fact that the human ear is insensitive to spectral dips. Thus ignorance of spectral dips is not a major problem as far as analysis-synthesis telephony is concerned. However, if one is interested in more accurate estimation of spectral dips as well as peaks, a new model has to be developed, which is currently being investigated by some researchers.

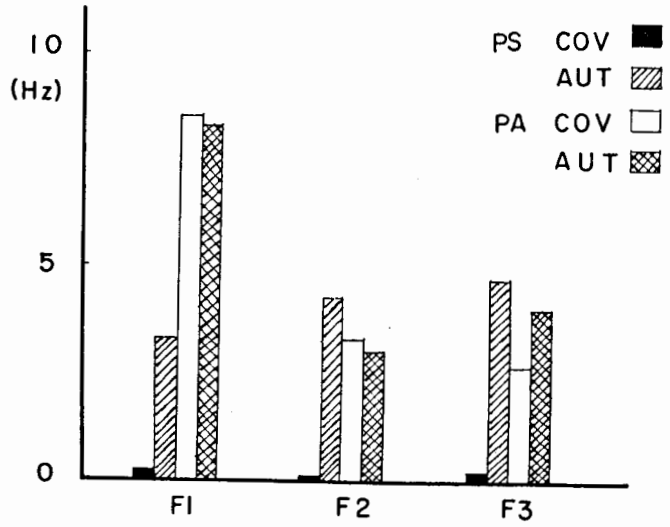


Figure 5. Evaluation of formant frequency estimation by autocorrelation and covariance methods for pitch-synchronous and pitch-asynchronous cases.

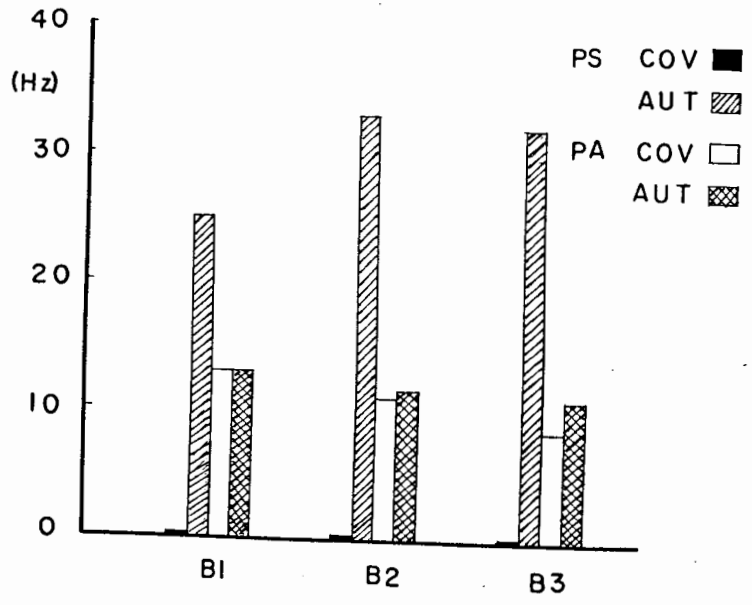


Figure 6. Evaluation of formant bandwidth estimation by autocorrelation and covariance methods for pitch-synchronous and pitch-asynchronous cases.

The formant frequencies are estimated from the smooth spectral envelope by finding the locations of the spectral peaks by a peak-picking method. Although this method is simple and worthwhile, it presents problems when two peaks are close together or merged into a broad peak. Another method is to compute the exact locations of the peaks by solving for the roots of the transfer function, A(z), of the inverse filter. In both methods, the spectral peaks do not always correspond to the formant frequencies, and thus a certain algorithm to automatically select formant peaks has to be designed (e.g. McCandless 1974). For both methods, a careful inspection of the analysis results is recommended before further processing of the formant frequencies is initiated.

Accuracy of formant estimation

It is rather difficult to determine the accuracy of formant estimation for natural utterances, since there is no way of accurately measuring the vocal tract configuration to compute its resonances while a sound is being produced. Chandra and Lin (1974) made an evaluation of the autocorrelation and covariance methods of linear prediction by using synthetic vowels. In their study, vowels in the 'h-d' context were synthesized by a simulated formant synthesizer, and the two linear prediction methods were applied to analyze those synthetic vowels. As analysis conditions in this case, the sampling frequency was 10 kHz and the number of coefficients was 12. The results of their study are shown in Figures 5 and 6. Figure 5 shows the estimation error (in Hz) of the first three formant frequencies for both methods applied pitch-synchronously and pitch-asynchronously. For the pitch-synchronous case, the window length coincided with the segment position between the two pitch pulses. For the pitch-asynchronous case, the window length of 24 ms was arbitrarily chosen on the speech waves. The results indicate that the pitch-synchronous covariance method gives better accuracy than the others. In the pitch-asynchronous case, when the window length becomes greater than one and a half pitch period, the two methods give similar accuracy. The pitch-synchronous autocorrelation method resulted in the worst accuracy. This is more so in estimating formant bandwidths as shown in Figure 6.

For natural utterances, it is anticipated that the accuracy of estimating formant frequencies and bandwidths becomes worse

than for the synthetic sounds. Especially, it is anticipated that the result of the pitch-synchronous case will become worse, because the condition at the glottis varies during one pitch period for natural utterances, whereas the glottal condition for this particular synthesizer was constant. When the glottal condition varies during a chosen analysis segment, the resulting formant frequencies will probably be the average of the instantaneous formant frequencies. The result obtained by Chandra and Lin (1974) indicate that the pitch-synchronous covariance method gives more accurate estimates of formant frequencies and their bandwidths than the pitch-asynchronous autocorrelation method. Although the estimation accuracy of the formant bandwidths is not well known, it is known that the bandwidth estimates are sometimes too narrow or too broad. If the bandwidth information is needed, it has to be carefully checked against the direct Fourier transform of the corresponding sampled speech.

Problems in formant estimation

Since the estimation of formant frequencies is made from the envelope estimation of speech spectra, the accuracy of estimation is highly dependent on harmonic density. The more sparse the harmonic density becomes as pitch goes up, the more difficult the estimation of formant frequencies becomes. This is a rather inherent problem in the estimation of vocal tract resonances from given speech waves, irrespective of method. In many cases, the linear prediction method works well for speech sounds with fundamental frequencies of up to approximately 250 Hz. For female speakers and children with fundamental frequencies higher than 250 Hz, difficult cases of formant estimation are frequently observed. Formant estimation becomes impossible as the pitch becomes extremely high, in which case harmonics are picked up as spectral peaks.

In case the exact vocal tract resonances need to be known, some other methods may have to be used. One approach to this is to use external excitation with a low fundamental frequency such as an artificial larynx buzzer. One such example is shown in Figure 7 (a). This example is a female vowel /a/ with a fundamental frequency of 250 Hz. The linear prediction spectral envelope has one broad peak in the low frequency region instead of the first two formant frequencies. The peak-picking method de-

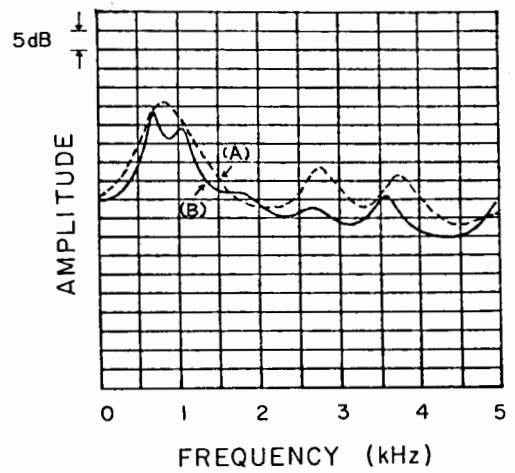


Figure 7. An example of difficult case of formant estimation. (a) Linear prediction spectral envelope for the vowel /a/ by a female speaker with a fundamental frequency of 250 Hz (sampling frequency 10kHz; number of coefficients 12; window size 25.6ms with a Hamming window and +6dB/octave preemphasis). (b) Linear prediction spectral envelope for the vowel /a/ by the same speaker excited by an external buzzer with a fundamental frequency of 80Hz (analysis conditions are the same as in (a)).

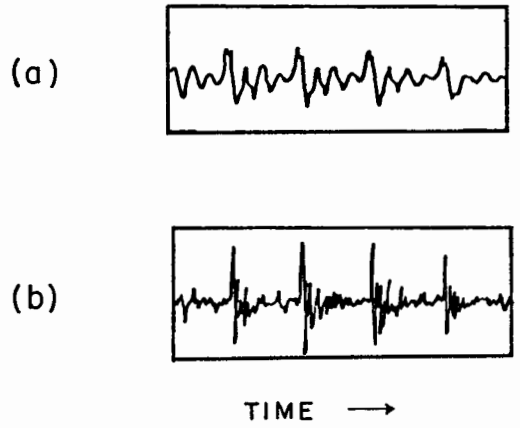


Figure 8. (a) Speech waves; (b) the residual signal after linear prediction analysis.

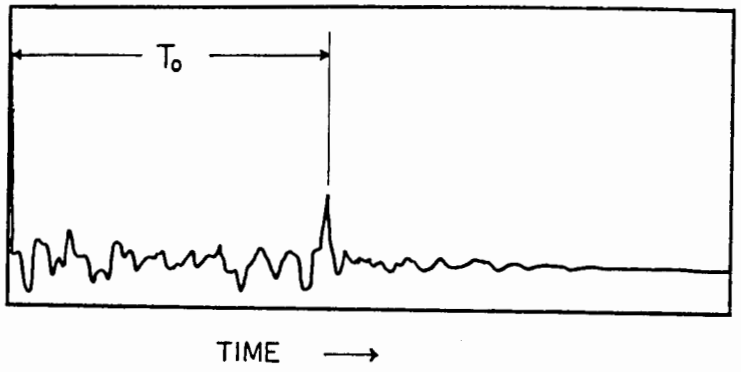
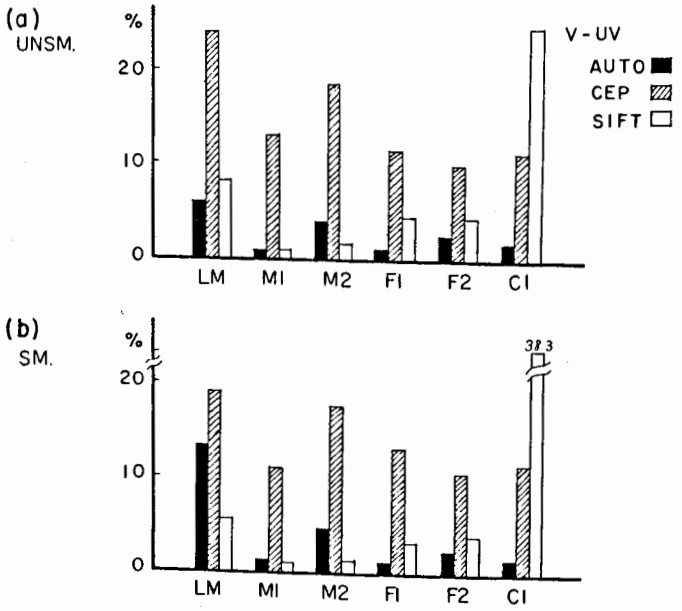


Figure 9. Autocorrelation function of the residual signal in Figure 8.



Figur 10. Voiced-to-unvoiced errors for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM:low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the percentage error rate against total number of voiced intervals.

finitely fails to detect two peaks for F_1 and F_2. Instead it will detect the broad peak as the first formant frequency.

The root-solving method will give two roots to approximate the broad peak. It has not been ascertained, however, that the two roots obtained by the root-solving method for such cases as above correspond accurately to the first two formant frequencies. For the above case, the use of a commercial artificial larynx buzzer with a low fundamental frequency gives a good resolution for the formant frequencies as shown in Figure 7 (b), which is for the same vowel and the same speaker as in Figure 7 (a). In this case, the buzzer had undesirable sharp peaks in its own frequency characteristics. The monotonous frequency characteristics of a buzzer are desirable for this purpose.

Fundamental frequency estimation

In inverse filtering in the linear prediction method, most of the vocal tract characteristics are filtered out into the predictor coefficients. The residual signal, the output of the inverse filter, still contains the information on the excitation source. A typical residual signal is shown in Figure 8. It is seen that large errors synchronous with pitch periods occur. A typical approach to computing the periodicity from this kind of waveform is to compute the autocorrelation function as shown in Figure 9. Two conspicuous spikes are found in the autocorrelation function, one at the origin and one at a distance of one pitch period from the origin. The fundamental frequency is then given by the reciprocal of the pitch period.

Problems in fundamental frequency estimation

It has been shown that the linear prediction method is quite efficient and effective for estimating the formant frequencies. However, how accurate and reliable the extraction of fundamental frequency is is an intriguing question, since there are many other techniques for estimating the fundamental frequency. Rabiner et al. (1976), in their study of the comparative performance of several pitch detection algorithms, point out the following major problems in detecting the fundamental frequency: (1) glottal excitation is not perfectly periodic; (2) defining the exact beginning and end of each period is difficult; (3) the distinction between unvoiced portions and low level voiced portions is difficult; (4) there is an interaction between the vocal tract and the glottal excitation.

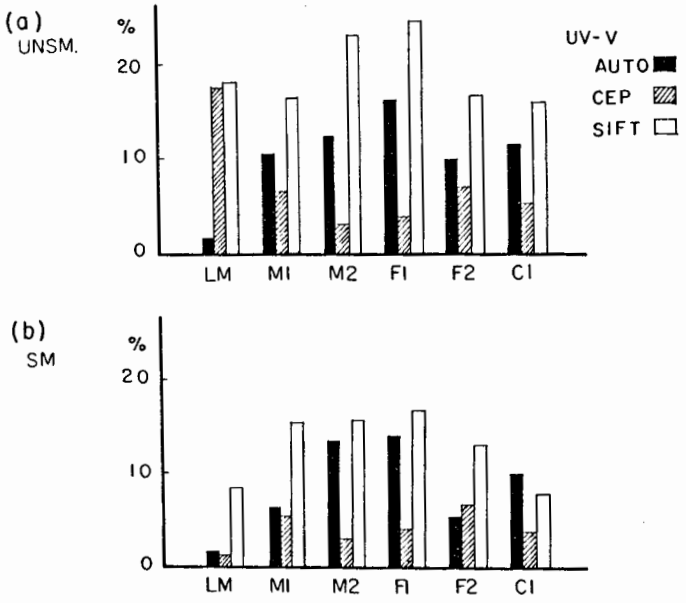


Figure 11. Unvoiced-to-voiced error for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM: low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the percentage error rate against total number of unvoiced intervals.

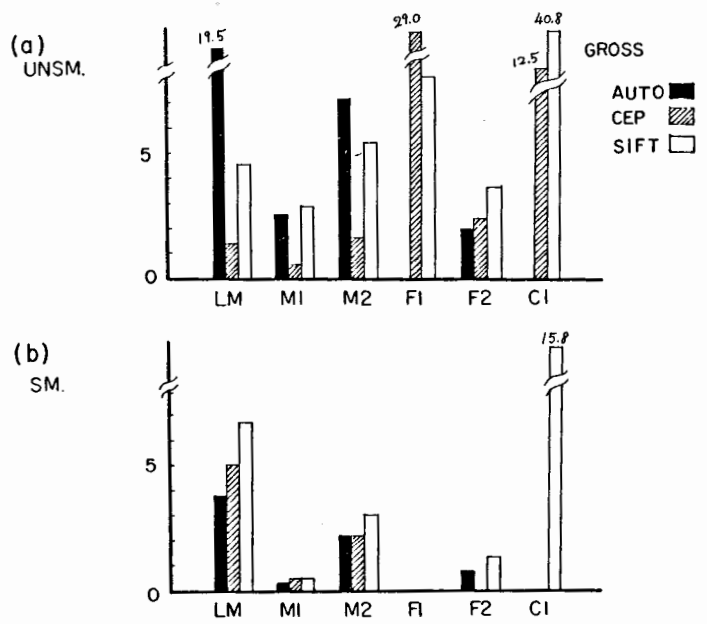


Figure 12. Gross errors for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM: low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the average number of samples.

The above problems are intrinsic in any of the pitch detection methods. However, evaluation of several pitch detection methods indicates some differences in their performance.

Accuracy in fundamental frequency estimation

Let us take the following pitch detection methods from the study by Rabiner et al. (1975): (1) autocorrelation method with clipping (time domain method); (2) cepstrum method (frequency domain method); and (3) linear prediction 'SIFT'[1] method (time-frequency method). The types of errors can be categorized into (a) voiced-to-unvoiced error, (b) unvoiced-to-voiced error, (c) gross error in which the error in detecting the pitch period is greater than a certain threshold; and (d) fine error in which the error in detected pitch period is less than the threshold.

The above three methods were tested against six speakers (3 males, 2 females, and a child) by using four monosyllabic nonsense words and four sentences. The analysis results were compared with the standard pitch contours which were carefully measured by using a semi-automatic pitch detector. The results for the first three types of errors are shown in Figures 10, 11, and 12. The results are shown both for unsmoothed (raw data) and smoothed cases. In the smoothed case a nonlinear smoothing technique was applied to the raw data (Rabiner et al., 1975). It is seen that the nonlinear smoothing generally improves the accuracy: particularly, the gross errors are substantially improved. It is also seen that all three methods are somewhat speaker dependent. For the voiced-to-unvoiced errors, the error rate of the cepstrum method is much higher than the others except for the child speaker. For the unvoiced-to-voiced errors, on the other hand, the error rate of the cepstrum method is better than the others except for one of the female speakers for the smoothed case. In overall performance evaluation, there seems to be not much difference between the performance of the autocorrelation and linear prediction methods, except that the linear prediction method resulted in an exceedingly poor performance for the child speaker for the unvoiced-to-voiced and gross errors.

Other related topics

The filter box in the linear prediction model in Figure 1 contains the contribution from the glottal characteristics and the radiation effect at the lips as well as the vocal tract

1) Simplified Inverse Filter Tracking

characteristics. Since the model assumes a linear system, those factors can be separated and changed in order as shown in Figure 13. If the glottal and radiation characteristics can be eliminated by a proper preprocessing of the speech, the true vocal tract characteristics can be obtained by the linear prediction method. One of the important features of the linear prediction method is that in computing the prediction coefficients, another parameter which is called "reflection coefficient" (or "k-parameter", or "PARCOR coefficient") is obtained. A set of reflection coefficients obtained for a given speech segment gives an acoustic tube shape which has a frequency characteristic identical to the vocal tract characteristics extracted from this speech segment. In this case, the acoustic tube is represented by a concatenation of cylindrical sections of different cross-sectional areas. A reflection coefficient is defined at the boundary between two neighboring sections. Consequently, if the analysis conditions are properly chosen after preprocessing sampled speech to eliminate the glottal and radiation characteristics, the acoustic tube representation thus obtained is expected to be a good approximation to the vocal tract area function which denotes the cross-sectional areas along the vocal tract from the glottis to the lips (Wakita, 1973, 1979).

Another interesting topic is the use of the linear prediction parameters for speech synthesis. The synthesizer could be the synthesis part of the linear prediction analysis-synthesis telephony (see Markel and Gray 1976; Wakita 1976). Since the formant frequencies and bandwidths constitute the roots of the inverse filter transfer function, they can be related to the filter coefficients. The reflection coefficients, which give an acoustic tube representation of the vocal tract, are also related to the filter coefficients in the mathematical formulation of linear prediction. Thus, those parameters mentioned above are interchangeable for each other, and any of these parameters can be used for the linear prediction synthesizer.

Application examples

The linear prediction method has mainly been used in the area of analysis-synthesis telephony. The method is particularly effective for low bit-rate speech coding. However, the technique is equally useful for acoustical analysis of speech. In concluding this tutorial paper, several examples taken from the author's past studies will be given below.

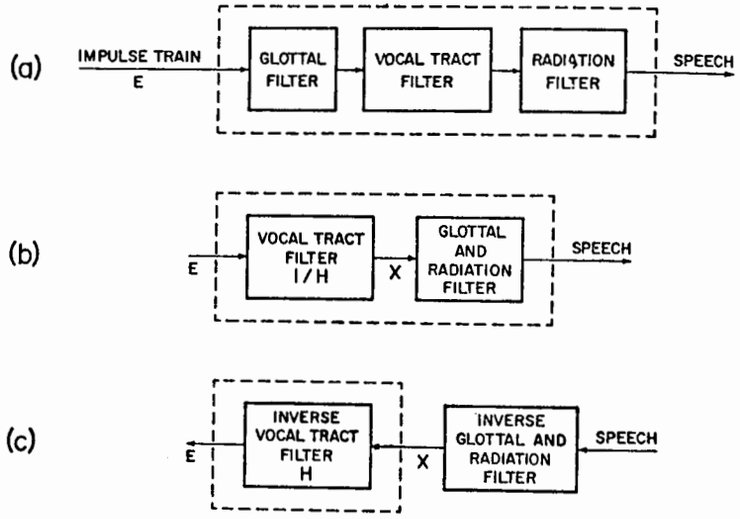


Figure 13. Block diagrams to obtain the vocal tract characteristics by eliminating the glottal and radiation effects.

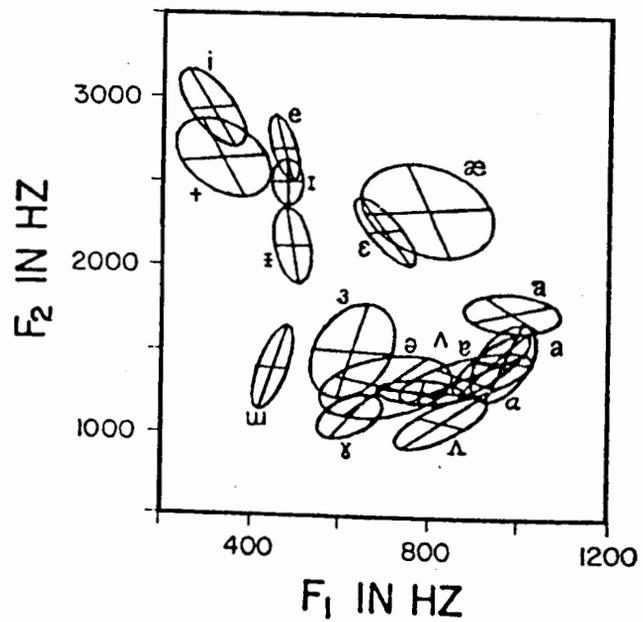


Figure 14. F_1-F_2 distribution of 17 unrounded vowel types produced by a female speaker. Ellipses represent two standard deviations.

Example 1 (Broad and Wakita, 1978).

Figure 14 shows the F_1-F_2 distribution for 17 unrounded vowel types produced by a female phonetician in order to study the variability of formant frequencies. In this study, 30 repetitions of 30 different isolated vowel utterances (900 in total) were analyzed by the linear prediction autocorrelation method (sampling frequency 10 kHz; number of coefficients 12; window size 25 ms with Hamming window and +6 dB/octave preemphasis) and formant frequencies were estimated by using the root-solving method. The analysis results were carefully inspected by displaying the results vowel by vowel on the display terminal. Approximately 5% of apparently wild data were excluded for further processing.

Example 2 (Wakita, 1977)

The example in Figure 15 shows the F_1-F_2 distribution of nine American English vowels spoken by 26 speakers (14 males and 12 females) in order to study the variability of formant frequencies among male and female speakers. Vowels were produced in the context of 'h-d' and the linear prediction autocorrelation method was applied to analyze the vowel portions (sampling frequency 10 kHz; number of coefficients 12; window size 25 ms with Hamming window and +6 dB/octave preemphasis). The formant frequencies were estimated by use of the root-solving method. The formant frequencies which were averaged over the most steady-state five frames were used to represent each vowel.

Example 3 (Kasuya and Wakita, 1979)

Figure 16 is an example in which the linear prediction area functions were used to automatically segment speech into vowel-like and nonvowel-like intervals. The linear prediction area functions, combined with the speech energy function (root mean square of sampled speech), give sufficient cues for the first stage of segmentation without obtaining spectral information which is more time consuming. In this case, the autocorrelation method was also used for analysis. The sampling frequency was 10 kHz, the number of coefficients 14, and window size 15 ms with a Hamming window and +6 dB/octave preemphasis. The relatively short analysis window length was used in this study for detecting the bursts of plosives, and the window shift was 12.8 ms.

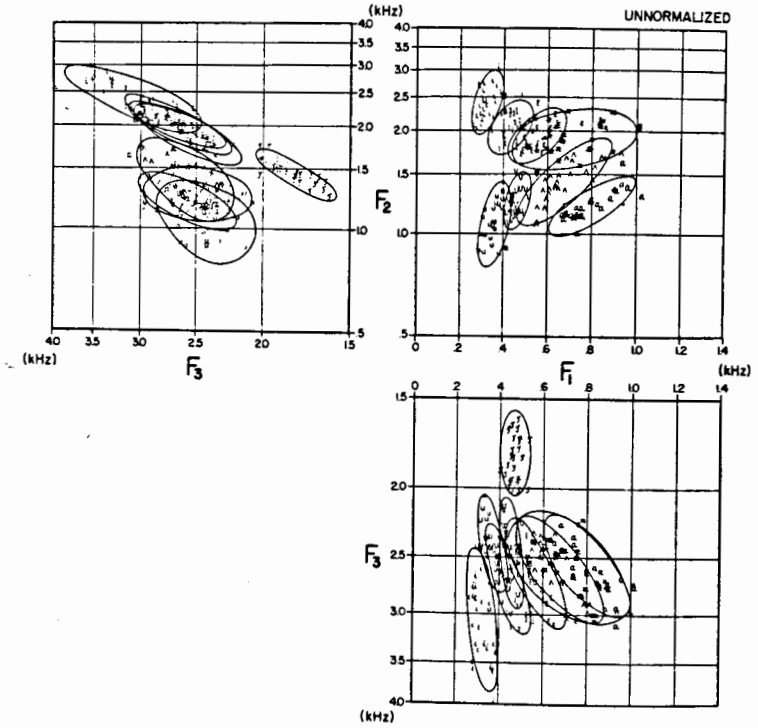


Figure 15. Distribution of formant frequencies projected onto the F_1-F_2, F_1-F_3, and F_2-F_3 planes for 26 speakers (14 males and 12 females). Ellipses represent two standard deviations.

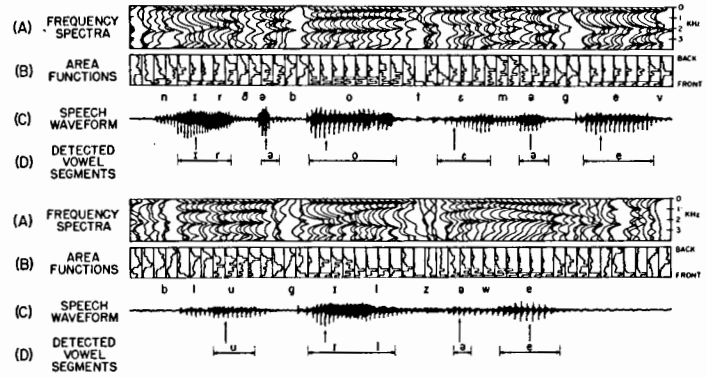


Figure 16. An example of segmenting the vowel-like intervals for the sentence " Near the boat, Emma gave blue-gills away."

Conclusion

The concept and evaluation of the linear prediction method were described in this paper. Because of its tutorial nature, the descriptions in some cases may be inadequate from the theoretical point of view. Readers interested in more advanced knowledge are encouraged to read the original papers or other materials listed in the references.

Acknowledgement

The author would like to thank Dr. P.-A. Benguerel, The Phonetics Laboratory, University of British Columbia, Canada, for his collaboration in investigating the use of an artificial larynx buzzer.

References

Atal, B. and M.R. Schroeder (1967): "Predictive coding of speech, Proc. 1967 Conf. Commun. and Process., 360-361.

Broad, D.J. and H. Wakita (1978): "A phonetic approach to automatic vowel recognition", in Bolc Speech communication with computers, 52-92, London: Macmillan.

Chandra, S. and W. Lin (1974): "Experimental comparison between stationary and nonstationary formulation of linear prediction applied to voiced speech analysis", IEEE Trans. ASSP-22, 403-415.

Itakura, F. and S. Saito (1966): "A statistical method for estimating speech spectrum", Technical Report 3107, Electrical Commun. Res. Lab., NTT.

Kasuya, H. and H. Wakita (1979): "An approach to segmenting speech into vowel- and nonvowel-like intervals", IEEE Trans. ASSP-27, 319-327.

Makhoul, J. (1975): "Linear prediction: a tutorial review", Proc. of IEEE vol. 63, 561-580.

Markel, J.D. and A.H. Gray (1976): Linear prediction of speech, New York: Springer.

McCandless, S.S. (1974): "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. ASSP-22, 135-141.

Rabiner, L.R., M.R. Sambur and C.E. Schmidt (1975): "Applications of a nonlinear smoothing algorithm to speech processing", IEEE Trans. ASSP-23, 552-557.

Rabiner, L.R., M.J. Cheng, A.E. Rosenberg and C.A. McGonegal (1976): "A comparative performance study of several pitch detection algorithms", IEEE Trans. ASSP-24, 399-418.

Wakita, H. (1973): "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. AU-21, 417-427.

Wakita, H. (1976): "Instrumentation for the study of speech acoustics", in Lass (ed.) Contemporary issues in experimental phonetics, 3-40, New York: Academic Press.

Wakita, H. (1977): "Normalization of vowels by vocal-tract length and its application to vowel identification", IEEE Trans. ASSP-25, 183-192.

Wakita, H. (1979): "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art", IEEE Trans. ASSP-27, 281-285.

DISCUSSION

Gunnar Fant, Wiktor Jassem and René Carré opened the discussion.

Gunnar Fant: I think that at the moment LPC analysis is more useful for communication engineering purposes, but it is certainly gaining importance in phonetic analysis: the fact that you can re-synthesize speech with rather good quality with LPC methods is a great advantage in synthesis, and LPC also makes it possible to manipulate e.g. fundamental frequency, independently of other parameters, which makes it well suited for prosodic investigations.

Formant frequencies and bandwidths describe the vocal filter, but what about the vocal source? In LPC analysis, it is treated as a constant function, more or less, but in the future we should pay more attention to the time dynamics of the source, to obtain valuable information for prosody studies. We should make dynamical matches not just to formants but also to source characteristics. (This we can do at present by carefully scrutinizing period after period of the signal, extracting presumed vocal source characteristics.) The fact that LPC is confined to an on/off, or voiced/voiceless, distinction creates some undesirable compensation effects: to compensate for a more steeply falling voice source spectrum, like we get e.g. in open syllables, the system will increase the bandwidths somewhat, which can give a consonantal effect.

Another critical problem is assessing formant frequencies with high pitched voices and in cases where Fo and F1 are close together, which is problematic in any kind of analysis.

Hisashi Wakita: mentioned a comprehensive LPC analysis of 900 vowels by a female speaker (30 vowels x 30 repetitions) where (50) unlikely analysis items were discarded by visual inspection of the vowels in F1-F2, and F1-F3 plots [see "Application Examples", Example 1 in Hisashi Wakita's paper], but admitted that we do not yet have valid data that tell us how accurately we can estimate formant frequencies, especially when Fo and F1, or two formants, are close together.

If we analyse a little more than one pitch period, using a very small time window and the covariance method we can, from the error signal, determine that point where the interaction between sub- and supraglottal systems is minimum (corresponding to the

closed glottis portion), and if the signal has been carefully recorded, directly from the microphone into the computer storage, so as to avoid phase distortion, we can fairly well recover the glottal wave shape from this portion.

Wiktor Jassem: What is the perspective for phonetics of these methods? First, there is the segmentation problem which can probably be solved, as suggested by professor Fant and others, by determining the maximum rate of change of the spectrum and of the time function. Secondly, there is the extraction of parameters: those extracted for automatic analysis need not be identical to those used by a human being. Thirdly, there is the problem of normalizing for individual speaker characteristics. The fourth problem is concerned with the identification of entities, which is an intricate one, because we do not know how many entities there are. The theory is that they should be sufficient to specify the output in such a way that synthesizing it we would get a normal native accent. The perceptual experiments needed to settle the question are not simple, because the adults' responses will be heavily influenced by phonemic considerations, and with very young children there will be great psychological problems. Fortunately, mathematical methods are developing that will allow us to determine, given a number of data, how many objects or entities we are dealing with. What I want to point out is that if we can get the computers to do phonetic transcriptions they will be better than transcriptions by a human being because they will be more objective.

René Carré: There are two kinds of work in speech analysis. One is the analysis of a small number of speech sounds. Formant frequencies are no problem, but to determine bandwidths we need to consider pre-emphasis, the order of the predictors, the analysis window, and the magnitude of the prediction error. All these operations take time, and such a procedure cannot be adopted in the other kind of study, of a large corpus, where a (semi-)automatic procedure has to be set up. It seems that in that case the procedure must be normalized. Is the autocorrelation method accurate enough for bandwidth measurements? Must we change (automatically or not) the order of the predictor to adapt the system to the speech sound under analysis, e.g. to nasalized vowels? What sampling rate shall we choose? How many frames should be analyzed? And so on. Finally, among the set of pole values we have to choose (automatically or not) the right formants.

Hisashi Wakita: The RMS-function is generally not sufficient to segment a chain into vowel-like and non-vowel-like sounds. But from the pseudo vocal tract area function, generated by the LPC analysis, we can calculate the ratio of the volume of the back (pharyngeal) cavity to the total volume of the vocal tract and this will generally tell us whether a segment is vowel-like or not. It will detect nasal consonants which is difficult to do from the waveform: LPC does not assume any nasal tract, but does produce a sort of equivalent acoustic tube representation, and nasal segments are fairly well detected from the back-to-total ratio of that tube.

We have also worked on the elimination of inter-speaker variability, which is of interest not just to automatic speech recognition, but also in acoustic phonetic studies of e.g. the vowel systems of languages. With LPC we can estimate the vocal tract length for each speaker and each vowel category (tract length is not constant over different vowel qualities), and then normalize to a certain length, e.g. 17 cm, a normalization which reduces the overlap in F1-F2, and F1-F3 plots and results in compact vowel distributions.

Adrian Fourcin: The LPC system represents the complexities of the vocal tract and its excitation by an exceedingly simple model: a vocal tract with no side-branches and a sharp impulse for an excitation, and yet it produces speech of very high quality. When we synthesize we have to pay attention to the zeros introduced by nasality, and the time dependence of the excitation function is also apparent if we have a standard model of the vocal tract. Is there something that we can learn from this with regard to how we hear speech?

If we knew when the point of excitation occurred and for how long a time the glottis is closed, to what extent would you be able then to improve the phonetic utility of the LPC analysis?

Hisashi Wakita: The ear is insensitive to spectral zeros, and a model which has poles and zeros in it (which is much more complicated computationally) does not perceptibly improve the quality of the speech. I have run an experiment, where various musical instruments as well as speech were passed through an artificially generated pole-zero system, and it turned out that the ear was insensitive to dips in the spectrum as large as 35 dB

(a fact which explains why HiFi loudspeakers may have even very sharp dips).

If we can determine that segment of speech where the glottis is closed, i.e. the force-free oscillations, we can apply the covariance method, which assumes that the speech waves can be approximated by a combination of damped sinusoids, and thus compute the exact vocal tract characteristics.

Gunnar Fant: A reply to Dr. Fourcin is that LPC speech sounds good because it resembles natural speech, although its source and transfer functions do not resemble those of real speech. The source function is stylized, but then there is a compensation in terms of the transfer function chosen to get the overall result correct (something which invalidates the data we get on formant frequencies and bandwidths).

Another characteristic of LPC analysis is that all the losses are concentrated at the glottal end of the system. How much does that invalidate the bandwidth data?

Hisashi Wakita: It is true that the LPC method approximates the spectral envelope, without any regard to formant frequencies and bandwidths. All the energy losses are lumped into one single resistance at the glottis end. By means of this single resistance we represent all the bandwidths of the spectrum. If we want to relate it to a particular speech production model, in terms of formant frequencies and bandwidths, it is quite useless, I think, so either we have to build more realistic models, both production and inverse transform models, or we can try to relate the simple LPC model to a more realistic, complicated model.

John Clark: There seems to be no great difference in the intelligibility levels quoted in the recent literature for predictor coded and formant coded speech. For formant coded speech, some of its phonetic weakness appears (when tested with CV-nonsense syllables) in the fricatives. Is this also the case for predictor coded speech, and what sort of evaluation have you done of the perceptual weaknesses of the system as a means of synthesizing speech?

Hisashi Wakita: Normally, with the LPC analysis-synthesis we use the extracted coefficients as they are, but we replace the residual signal with a pulse train which makes the voiced/unvoiced decision very critical, and missing just one frame can be perceptible. We can, however, restore the original signal by using the residual signal for excitation. For phonetic evaluation purposes I think we have to choose the excitation source carefully, - maybe not the residual signal itself, but one with which we do not loose too much information about the source.