Issues in High Quality LPC Analysis and Synthesis

Melvyn J. Hunt  
Marconi Speech & Information Systems  
Airspeed Road, The Airport  
Portsmouth, PO3 5RE, U.K.

Dariusz A. Zwierzyński  
Dept. of Linguistics  
University of Ottawa  
Ottawa, K1N 6N5, Canada

Raymond C. Carr  
Dept. of Electrical Engineering  
University of Ottawa  
Ottawa, K1N 6N5, Canada

(Work carried out at the National Research Council, Montreal Road, Ottawa, K1A 0R6, Canada.)

Abstract
This paper deals with careful non-real-time LPC analysis. A baseline system is first described. It uses a pitch-synchronous covariance-method analysis with a laryngograph signal providing the pitch synchrony. Work to improve the voicing decision and $F_0$ determination and to find a better voiced excitation waveform is described. Setting a lower limit on the value of $B_0$ is found to be useful. Buzziness in the synthesis of voiced fricatives can be reduced by adding white noise to the excitation, and regions where this should be done can be automatically detected using three parameters: total power, the ratio of high-frequency power to total power, and $B_0$. The location of the analysis frame for covariance-method analysis is found to be unimportant. A modified autocorrelation method with a carefully placed, Hanning-windowed, pitch-synchronous analysis frame is found to give results that are as good as or better than the covariance method. Finally, a method of resynthesizing speech using a modified LPC residual is described. The method allows prosody and formant parameters to be manipulated with minimal degradation.

1. Introduction
Linear predictive coding (LPC) [1] is most commonly used for low bit-rate digital speech communication. For this application, the analysis and resynthesis must necessarily be carried out in real time, thus limiting the complexity of the algorithms used. Also, real-time communications systems must often operate with speech degraded by noise, reverberation and distortions.

There are, however, other applications of LPC for which these limitations need not apply. LPC may be used for formant analysis [2], for generating stimuli for speech perception experiments [3], or in a speech output system, where LPC is used to encode whole phrases or demisyllable or phone subunits from which speech is reconstructed. In all these cases, the analysis need not function in real time, and the acoustic quality of the speech to be analyzed can be carefully controlled. The work described in this paper is relevant to this set of applications. Careful analysis with speech of high acoustic quality also helps to explore the limits imposed by the all-pole model of speech production assumed by LPC.

We have previously described [4] a system intended to incorporate the best possible (laryngograph-aided covariance-method pitch-synchronous) analysis of speech recorded and digitized with extreme care, and others have described similar systems [5]. This paper describes our original system briefly and our attempts to improve it experimentally, the synthesis of voiced fricatives can be reduced by adding white noise, reverberation and distortions.

Evaluation of the performance of high quality LPC analysis and synthesis is not easy. Signal-to-noise ratios, often used for evaluating waveform coding schemes, are not useful here because LPC-encoded waveforms can show differences from the original speech that are not perceptible. Although we have not confirmed this belief experimentally, the intelligibility of our synthesised speech is almost certainly too high to be measured. Methods based on the analysis of synthetic speech [6,7,8] seem circular: LPC naturally works well in analyzing speech generated by the LPC model. Similarly, judgments of accuracy in fitting spectra or in locating formants depend on using synthetic speech or on having some standard of reference for what constitutes “correct” spectra or formant data for real speech. Consistency of formant tracks may give some indication of the reliability of the analysis [2], but it can also be misleading, since some methods, such as those using long, overlapping analysis windows impose continuity on the analyses, while others do not. The normalised error power has sometimes been used [6,7], but in our experience it does not correlate with subjective impression of quality, which is the ultimate criterion for a high quality speech output system.

The most of these assessments were informal, though some formal subjective tests are described in Section 5. We assume that good quality synthesis generally implies a good analysis.

Further details on the topics covered in Sections 2 to 7 can be found elsewhere [9].

2. The Original System
Speech spoken in an anechoic chamber was recorded digitally using a Sony PCM-F1 system together with the signal ($L_m$) from a laryngograph (or electroglottograph). The signals were played back and digitised in our computer at a 20kHz sampling rate. They were then digitally low-pass filtered at 5 kHz and subsampled at 10kHz. We now use an OROS AI device to input the digitised recordings directly and subsample them to 10kHz. This change in procedure has brought no change in quality.

The digitised $L_m$ was differentiated, and the impulses in this signal were taken to correspond to instants of glottal closure [10], and thus provide indications of voicing and fundamental frequency ($F_0$). Speech judged as voiced was preemphasised with a .95 preemphasis factor, while speech considered voiceless was not preemphasised. For voiced speech, a 10th-order covariance-method LPC analysis was then applied to rectangular analysis frames beginning up to ten samples before glottal closure and ending just before the next glottal closure. For voiceless speech, the analysis frame was generally fixed at 10ms, though when bordering on a period of silence it could be extended up to 20ms.

Any instabilities in the filter specified by the covariance-method analysis were corrected by solving the predictor polynomial to obtain the frequencies and bandwidths of the poles, generally corresponding to formants, and reflecting the unstable poles in the unit circle.

For resynthesis, the pole parameters were recomputed to predictor coefficients. The excitation signal for voiced speech consisted of the residual taken from a single glottal cycle and extended with zeros if necessary, and that for voiceless speech was white noise. The filter parameters were updated pitch-synchronously.

Thus, although the analysis was relatively complex, the resynthesis was simple. Only the ten pole parameters, $F_0$, and a power parameter needed to be stored. The changes described in Sections 3 to 8 have not added to the storage requirements or the complexity of the synthesis process apart from the addition of a one-bit parameter needed to specify the mixed excitation described in Section 5.

3. Determination of Voicing and Fundamental Frequency
The original system generated occasional pops and clicks due to errors in $F_0$ estimation or to voiced speech being taken to be voiceless. At the ends of voiced regions, where the speech signal, though weak, is still clearly being periodically excited, $L_m$ sometimes shows no activity (Fig. 1). This occurs even for speakers whose $L_m$ is in...
general particularly strong. The resulting errors can be corrected by using the ratio of the first to the zeroth autocorrelation coefficient of the speech signal to check the spectral balance in each analysis frame judged from $L_x$ to be voiceless. If low-frequency energy predominates, the frame is declared to be voiced, and $F_0$ is determined from the peak in the autocorrelation function.

In voiced /h/ sounds, the differentiated $L_x$ has been observed to produce spikes of opposite polarity to those in normal voicing (Fig. 2), these spikes apparently corresponding to glottal opening rather than closure. The original system tried to track them, but an autocorrelation analysis of the speech signal is now used here too.

To test whether the laryngograph was strictly necessary, we replaced in one sentence $L_x$ by the LPC residual derived from a preliminary pitch-asynchronous LPC analysis. The resulting synthesis was not degraded. $L_x$, when available, however, remains a convenient and accurate method of tracking $F_0$ most of the time.

![Fig 1. Speech waveform and differentiated laryngograph output showing continuation of voiced speech after the laryngograph signal has stopped.](image)

![Fig 2. Differentiated laryngograph signal for "three hundred" showing the large inverted impulse during the voiced /h/.](image)

### 4. Voiced Excitation

The simplest LPC synthesizers use a single impulse for voiced excitation. Our residual from a single glottal cycle gave a slightly better quality, but there remained an overall difference in quality between the original and the synthesized speech. Averaging residuals from many glottal cycles after aligning the main impulse is unsuccessful because the high-frequency components jitter with respect to the impulse and cancel each other out in the averaging process to give a low-pass filtered excitation signal. The problem can be avoided by applying a Fourier transform to each glottal cycle and averaging the amplitude and phase spectra separately before transforming back to the time domain (Fig 3). The result is, however, no better than using a single glottal cycle, and we have failed to improve on our original system in this respect.

![Fig 3. Log amplitude spectra of residuals averaged pitch-synchronously over many glottal cycles of voiced speech. The continuous line shows the result of averaging the individual amplitude spectra and the dotted line that of averaging in the time domain. Note the loss of high frequencies in the latter case.](image)

To test the importance of the temporal structure of the excitation, we generated a sequence of Gaussian white noise. Exciting with this sequence (repeated at the fundamental frequency) gave noticeably worse synthesis than an impulse-like sequence having the same amplitude spectrum but with its phase spectrum set to zero. This latter sequence, which looked like a residual, gave synthesis of similar quality to that obtained with a residual.

### 5. Treatment of Voiced Fricatives

The most salient remaining problem with the synthesis was a "buzziness" in prolonged voiced fricatives. We suspected that it could be resolved by adding a special excitation for such sounds, and tested the idea by identifying the offending portions manually and trying various excitations. Changing the excitation to the LPC residual from one glottal cycle of a voiced fricative did not help, but adding white noise to the periodic excitation caused the buzziness to go away. We conclude that the impression of buzziness results not from the spectrum of the excitation but from its high correlation from glottal cycle to glottal cycle. A fixed proportion of noise worked well for all voiced fricatives, so the excitation does not need to be continuously variable.

To maintain our goal of a fully automatic analysis, we need an automatic method of discriminating between voiced fricatives and other voiced sounds. To this end, many parameters were examined, including, for example, measures of correlation between consecutive glottal cycles and attempts to detect noise in the LPC residual. Three parameters (see Fig. 4) proved to be particularly useful. They were:

i) the total power

ii) the ratio of power above 3 kHz to total power

iii) the bandwidth of the first formant ($B_1$)

![Fig 4. (a) Waveform of "five seven" showing the automatic segmentation into silence (S), unvoiced speech (U), voiced speech (V), and speech with mixed excitation (M). (b) The bandwidth of the first formant. (c) The ratio of high-frequency power to total power. The low power in the voiced fricatives is evident in the waveform.](image)
A standard approach to combining the parameters is to construct a quadratic classifier. This assumes that the parameters have a multivariate normal distribution. The parameter distributions here are far from normal, so, not surprisingly, the quadratic classifier proved unreliable.

We then tried training a multi-layer-perceptron [11] with various numbers of hidden nodes in one or two layers on the discrimination task. This worked better, though there were cases where voiced glottal cycles without frication were erroneously judged to contain friction, and these errors were noticeable in the resynthesis.

Finally, we took a more empirical approach. Histograms (Fig. 5) of the parameter distributions were fitted either with simple functions or with a sequence of straight lines. The emphasis in choosing the functions was on their simplicity and their behaviour in regions where data was sparse rather than on their closeness of fit to the histograms. The parameters were treated as though they were independent, and a glottal cycle was classified as belonging to a voiced fricative or not depending on the relative probabilities obtained by multiplying together the three probabilities from the smoothed histograms for the two classes. The results were similar to those obtained with the multi-layer-perceptron. However, it proved possible to eliminate the salient errors by submitting glottal cycles judged to contain friction to a simple additional test using only the relative energy above 3 kHz. This filters out the perceptually salient errors. It also removes a few glottal cycles that had previously been correctly identified as belonging to voiced fricatives, but since they did not have much high-frequency energy, they do not sound buzzy when excited by the standard voiced excitation signal.

The method was tested on speech from a new speaker consisting of five sentences chosen to contain a high proportion of voiced fricatives and including examples of all the voiced fricatives found in English. No major errors could be detected either by listening or by examining the locations of the regions declared to contain voiced frication.

![Fig. 5. Histograms of values of parameters used in detecting voiced fricatives. Histograms from fully voiced data are shown with continuous lines and from voiced fricatives with dotted lines.](image)

**6. Adjustment of Bandwidth of First Formant**

Certain synthesized vowels had an unpleasant resonant quality, which was found to be associated with particularly low values of $B_1$. It is possible that the problem is caused during synthesis by excessive ringing from glottal cycles to the next. Such ringing would be damped out by the glotal open phase in real speech production. Whatever the cause, the problem can be cured by preventing the bandwidth from falling below 40 Hz using the formula

$$B_1' = \sqrt{B_1^2 + 40^2}$$

where $B_1$ is the original bandwidth and $B_1'$ its adjusted value.

The adjustment proved to be much more important in the pitch-asynchronous tests described in the next section.

7. The Need for Pitch Synchrony

It is widely believed that the best LPC needs a covariance-method analysis with the analysis period carefully synchronized with the glottal excitation. We tested the sensitivity to location of the analysis period by analyzing the same sentence eight times with the location of the analysis period progressively advanced by 1 ms with respect to the glottal excitation, so that the last analysis was advanced by 7 ms relative to the first one. The formant parameter values were essentially identical and the synthesized sentences were indistinguishable.

We then tried analyses with non-overlapping analysis periods of fixed duration: that is, completely pitch-asynchronous analyses. Provided the analysis period was of similar length to a glottal cycle, and with the bandwidth adjustment described in the previous section, the results were comparable with the pitch synchronous analyses. The power of the synthesized speech did not match that of the original speech, as it did in the pitch-synchronous case, but this has little perceptual importance. When the analysis period exceeded about 16 ms, the synthesized speech sounded clearly degraded and the excursions in formant tracks were reduced. When the analysis period was shorter than about 7 ms the synthesized speech was again degraded and the formant tracks showed short-term ripples, presumably related to alternations between predominantly open- and closed-phase analyses.

8. Windowing and the Autocorrelation Method

In agreement with many other reports [6,7,12], though unlike Markel and Gray [1], we can find no advantage from applying a window to the analysis frame when the covariance method is used.

Typical real-time systems use the autocorrelation method with overlapping analysis periods of around 20 ms with raised-cosine windows and an analysis order of 12 at 10 kHz or 10 at 8 kHz. We confirmed that this combination gives much worse results on our carefully recorded material than the covariance method with short, unwindowed analysis periods and 10th-order analyses at 10 kHz.

We then tested the use of the autocorrelation method with the analysis period extending from one instant of glottal closure to the next. Consistent with Palivew and Rao [7], but not with Makhou and Wolf [12] or Chandra and Lin [6], the application of a raised-cosine window produced a better result than the unwindowed equivalent, though both were judged somewhat worse than that obtained with the covariance method. Formant bandwidths with the unwindowed autocorrelation analysis tended to be larger than those with the covariance method, while those with the windowed autocorrelation analysis were smaller.

Unlike window selection for Fourier transforms, there seems to be no theory to guide the selection of window shapes for LPC analyses. We conclude from our own tests that windows for autocorrelation-method analyses need to be symmetric and tapered at each end, but the exact form is not critical. A Hanning window (a simple raised-cosine window with no pedestal) may give the best results.

When the start of the windowed analysis period was advanced to about 4 ms before the instant of glottal closure with the end remaining at the next instant of glottal closure (see Fig. 6), the quality of the synthesis equaled or exceeded that obtained with the covariance method, and formant bandwidths were similar.

![Fig. 6. Placement of Hanning window with respect to glottal closures, shown as impulses, found to give best results with autocorrelation-method analyses.](image)
The autocorrelation method makes a biased estimate of the autocorrelation function of the signal because fewer non-zero terms are summed in the higher-order coefficients than in the lower ones. The covariance method avoids this bias by restricting the range of the signal over which the sample products are summed. We introduced this range restriction into the autocorrelation method while retaining the Toeplitz property of the covariance matrix. The resulting modified pitch-synchronous autocorrelation method with the window placement just described gave us the best synthesis so far.

9. Excitation with a Modified Residual

If the time varying filter produced by an LPC analysis is excited by the corresponding residual, the synthesized waveform is in principle identical to the original waveform. In our careful pitch-synchronous analyses, the residual consists generally of an impulse at glottal closure. The remaining weak activity is least about 80% into the glottal cycle. We have tested the effect of changing \( F_0 \) by exciting with a modified residual in which samples were deleted or zeros were added at the 80% point. (With large deletions, more samples must be taken from before the 80% point than from after it to avoid encroaching on the next closure point.) Over a range of changes we could hear no degradation in the speech. With increases in \( F_0 \) of more than about 25%, some noise was audible, but \( F_0 \) could be decreased by a much greater proportion before any degradation became apparent.

Modifying the residual in this way also changes the duration of the speech, and the resulting change in speaking rate is more evident than the change in \( F_0 \). However, duration can be easily and undetectably adjusted by periodically adding or deleting a complete glottal cycle. We thus have a complete automatic method of manipulating the prosody of the speech signal without introducing audible degradations.

It is also possible to change formant frequencies and re-excite the residual without noticeable degradations.

Modified-residual synthesis offers no advantage in encoding speech, but it may be interesting for generating stimuli in speech perception experiments and for high quality speech output systems with controllable prosody and voice quality. Moreover, the reduction in size and cost of random-access storage makes it feasible to store residuals for a complete inventory of diphones or demesylatables for a text-to-speech system. The ability to manipulate formant frequencies as well as prosodic features means that smoothing across the boundaries of concatenated units would be possible and multiple voices could be generated from the same inventory of units.

To evaluate the modified-residual LPC we carried out listening tests with three naive subjects. However, space limitations allow only a brief account of them. The tests first established that the subjects were able to discriminate easily between the modified and unmodified sentences and between the sentences synthesized with a fixed excitation waveform and those synthesized with a modified residual. We then presented them with sentences in random order and asked them to note any in which they noticed distortions. Some 13% of the unmodified sentences were said to be distorted, exactly the same proportion as was found with sentences with a 33% rate increase, though with a corresponding rate decrease 22% were judged distorted. Among sentences in which \( F_0 \) was lowered by 20% only 7% were judged to be distorted, though a corresponding increase of 20% resulted in 51% being judged as distorted. With an increase in \( F_0 \) declination across the sentence, 16% were judged distorted, but with a corresponding decrease in declination the figure was 40%. Sentences generated with a fixed excitation waveform were judged much worse, with 65% and 100% of the examples being said to be distorted, depending on the modification applied.

The results indicate that at least some modifications produce no audible distortions, and that modified-residual excitation consistently produces less audible distortion than excitation with a fixed waveform. However, subjects avoided this bias by difficult to discount the value of \( B_0 \) from falling too low. The temporal structure of the voice excitation signal is important, and — at least for voiced fricatives — a degree of aperiodicity is needed between adjacent glottal cycles. The occurrence of voiced fricatives can be detected automatically.

With the covariance method, there is little or no advantage from using a pitch-synchronous analysis, though with the autocorrelation method there is. A suitably windowed pitch-synchronous analysis using the autocorrelation method is as good as a covariance-method analysis, and the modified autocorrelation method described here may be better.

Synthesis with a modified residual following a pitch-synchronous LPC analysis offers a method of generating speech with adjustable prosody with little or no perceptible degradation. Unlike alternative time-domain methods [13,14], formant parameters can also be manipulated. It is in some ways simpler than prosody-modification systems using multipulse LPC [15], and it is likely to have a better formant analysis, and therefore permit more effective formant manipulation, than systems using pitch-asynchronous multipulse LPC.

Acknowledgments

This work benefited from funding from DND DCIEM, NSERC operating Grant No. C48 6661 issued to MJH through the University of Ottawa, Dept. of Electrical Engineering, and NRC IRAP funding issued through the Neil Squire Foundation. We are grateful to the NRC for allowing us to use facilities in the NAE Flight Research Laboratory for this work.

References