The Phase-only Version of the LPC Residual in Speech Coding.

Evangelos E. Milios and Alan V. Oppenheim

Research Laboratory of Electronics, Room 36-615
Mass. Institute of Technology
Cambridge, MA 02139

ABSTRACT

Recent work has demonstrated the perceptual importance of the long-time phase of speech. In this paper, the possibility of using the long-time phase of the LPC residual signal in speech synthesis is investigated. The perceptual importance of the various parts of the long-time spectrum of the residual is also suggested.

1. INTRODUCTION

While the short-time Fourier transform phase of speech is relatively unimportant perceptually, recent experiments have demonstrated that the phase associated with the long-time Fourier transform of speech incorporates most of the intelligibility [Opp.1979, Opp.1981]. While a careful theoretical analysis supporting these results is not available, there are a number of informal and useful interpretations. Among these is the notion that in the long-time Fourier transform, the time location of events is represented principally in the phase of the Fourier transform rather than its magnitude. This interpretation suggests the possibility that the concept of phase-only speech could be useful in representing and eventually coding the residual signal associated with linear predictive coding (LPC). Specifically, the residual signal in linear predictive coding is the result of whitening the spectral envelope representing the vocal tract so that the residual signal approximately represents the vocal tract excitation. To a first approximation, this corresponds to a train of pulses during voiced sounds and noise during unvoiced sounds, and for voiced speech it is the time location of these pulses that is important to represent. Many LPC systems, in fact, code the residual by applying pitch detection and a voiced-unvoiced decision [Mar.1974]. There has been considerable interest in developing more robust ways of representing the residual signal [Fla.1979, Mak.1979]. The experiment described in this paper represents an approach to representing the residual signal through its phase-only version. Since phase-only speech has the subjective quality of the original speech with added broad-band background noise, this would suggest that the output speech from an LPC synthesizer driven by the long-time phase only residual would have the quality associated with using the true residual with added noise. As described in the paper, this in fact characterizes the principal degradation apparent in the speech synthesized with the phase-only residual. Appropriate scaling of the residual before the phase-only operation, and incorporation of some spectral magnitude information permits some control over this degradation. Current research is directed toward optimizing the associated analysis-synthesis system, aimed at a medium-band system which will retain the naturalness and quality of the original speech.

2. IMPLEMENTATION

In implementing the basic LPC analysis-synthesis to be used in these experiments, it was important to account for the nonstationarity of the speech signal in a computationally tractable way and to produce a system which reduces to an identity system if no modification of the prediction residual is performed.

To account for the nonstationarity of the speech signal, LPC analysis is performed over overlapping speech frames that are short enough to guarantee that the signal features change very little within each frame. Each frame has an offset with respect to its previous frame. A set of LPC filter coefficients is computed from each frame, and interpolation is used for generating the LPC filter coefficients between consecutive frames. Interpolation assures that the prediction residual does not change abruptly at the frame boundaries, which would be an artifact due solely to the specific analysis scheme. Typical values for speech sampled at 10KHz are: frame length 256 points, frame offset 64 points and LPC filter order 16.

To ensure an overall system that is an identity system when the unmodified prediction residual is used as input to the synthesizer, the synthesis procedure must be the exact inverse of the analysis procedure. A particular prediction residual point is obtained from the relation:

$$e(k) = s(k) - \sum_{l=1}^{p} a_l(k)\, s(k-l) \qquad (1)$$

where $a_1(k), a_2(k), \ldots, a_p(k)$ are the LPC filter coefficients for time k.

The corresponding synthesis equation for s(k) is:

$$s'(k) = e(k) + \sum_{l=1}^{p} a_l\, s(k-l) \qquad (2)$$

If e(k) is obtained from equation (1), and s(k) is reconstructed from equation (2) using the same filter coefficients as in equation (1) and also the same previous speech points s(k-1), s(k-2),..., s(k-p), then it is obvious that

$$s'(k) = s(k) \quad \text{for } all\ k \qquad (3)$$

This implies that the synthesized speech sample at time k is the same as the original speech sample at time k.

Thus if the time-varying synthesis procedure is initialized properly, the original speech will be generated point by point forward in time. The synthesis equation can be rewritten as

This work was partially supported by Bodossakis Foundation, Athens, Greece and by the Advanced Research Projects Agency monitored by ONR under contract N00014-81-K-0742 NR-049-506.

$$s'(k) = e(k) + \sum_{l=1}^{p} a_l(k)\, s'(k-l) \qquad (4)$$

and it was shown that $s'(k)=s(k)$ for all k.

A block diagram for the implemented system is given in Fig. 1.

The LPC analysis system generates a set of time-varying prediction coefficients updated at the frame rate. These coefficients are then used to generate an LPC residual signal through equation (1). This procedure is used to avoid having artifacts in the residual occurring at the frame rate. It should be stressed that this residual signal will not in general correspond to a concatenation or overlapping of the LPC residual obtained as a by-product of the LPC analysis on the individual frames.

The synthesis system has as inputs a prediction residual and a set of predictor coefficients updated at the frame rate (which must be synchronized and consistent with the prediction residual on the same time scale). It produces as an output the synthesized speech utterance. If $e_m(k)=e(k)$ for all k, then the overall system is an identity system, i.e. $s_m(k)=s(k)$ for all k.

In the proposed analysis-synthesis system, the residual signal e(n) is replaced by its phase-only version $e_m(n)$. The phase-only version is taken as follows: The long-time residual of 2 seconds of continuous speech sampled at 10KHz is padded with zeros up to the next higher power of 2 and the Discrete Fourier Transform is computed using the FFT algorithm. The magnitude of the Discrete Fourier Transform is set to unity, while the correct phase is retained. Then the Inverse Discrete Fourier Transform is computed and the resulting signal is truncated to the length of the original long-time residual. This truncated signal is used as input to the LPC synthesizer block together with the time-varying predictor coefficient set obtained from the analysis block.

3. RESULTS

Subjectively, the synthesized speech with the phase-only residual is extremely close in quality to the original except for the presence of a slight hoarseness. Since phase-only speech sounds very close to the original with the addition of some background noise, this suggests that the hoarseness using the phase-only residual is due principally to noise present in the phase-only residual due to the phase-only operation. To test this hypothesis, white Gaussian noise was added to the correct residual signal and this was then used as the excitation to the LPC synthesizer along with the exact predictor coefficients. The resulting synthesized speech subjectively had exactly the kind and quality of degradation as was apparent in the experiment with the phase-only residual. Adding noise to the residual is, of course, equivalent to adding whispered speech to the original. Consequently, as a further verification speech generated by exciting the synthesizer with only white Gaussian noise was added to the original, and as expected the quality and degradation were identical to that of the speech with the phase-only residual.

To reduce the hoarseness, some experiments were carried out, retaining some long-time Fourier transform magnitude information in the residual signal. Specifically, a magnitude which is a smooth version of the original long-time magnitude was combined with the correct long-time phase. While this reduced the hoarseness slightly, it was not totally removed. Finally, the correct long-time magnitude at low frequencies (for example between 0 and 600 Hz) and smoothed version of the magnitude over the remainder of the frequency band was combined with the long-time phase of the residual. This resulted in the elimination of the hoarseness.

4. CONCLUSIONS

The potential for synthesizing high-quality speech using the phase-only residual has been demonstrated. In particular, we propose coding the long-time phase together with either the smoothed long-time magnitude alone or the smoothed long-time magnitude augmented by some more accurate low frequency magnitude information.

REFERENCES.

[Fla.1979]: J. Flanagan, M. Schroeder, B. Atal, R. Crochiere, N. Jayant, J. Tribolet: "Speech Coding", *IEEE Trans. on Comm.*, April 1979.

[Mak.1979]: J. Makhoul, M. Berouti: "Adaptive Noise Spectral Shaping and Entropy Coding in Predictive Coding of Speech", *IEEE Trans. on ASSP*, Febr. 1979.

[Mar.1974]: J. Markel, A. Gray: "A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method", *IEEE Trans. on ASSP*, April 1974.

[Opp.1979]: A. Oppenheim, Jae Lim, G. Kopec, S. Pohlig: "Phase in Speech and Pictures", *Proc. of the 1979 ICASSP*, April 1979.

[Opp.1981]: A. Oppenheim, Jae Lim: "The Importance of Phase in Signals", *Proc. of the IEEE*, May 1981.

[Pear.1978]: W. Pearlman, R. Gray: "Source Coding of the Discrete Fourier Transform", *IEEE Trans. on Information Theory*, Nov. 1978.

[Poh.1980]: S. Pohlig: "Signal Duration and the Fourier Transform", *Proc. of the IEEE*, May 1980.

[Sing.1967]: R. Singleton: "A Method for computing the Fast Fourier Transform with Auxiliary Memory and Limited High-Speed Storage", *IEEE Trans. on Audio and Electroacoustics*, June 1967.

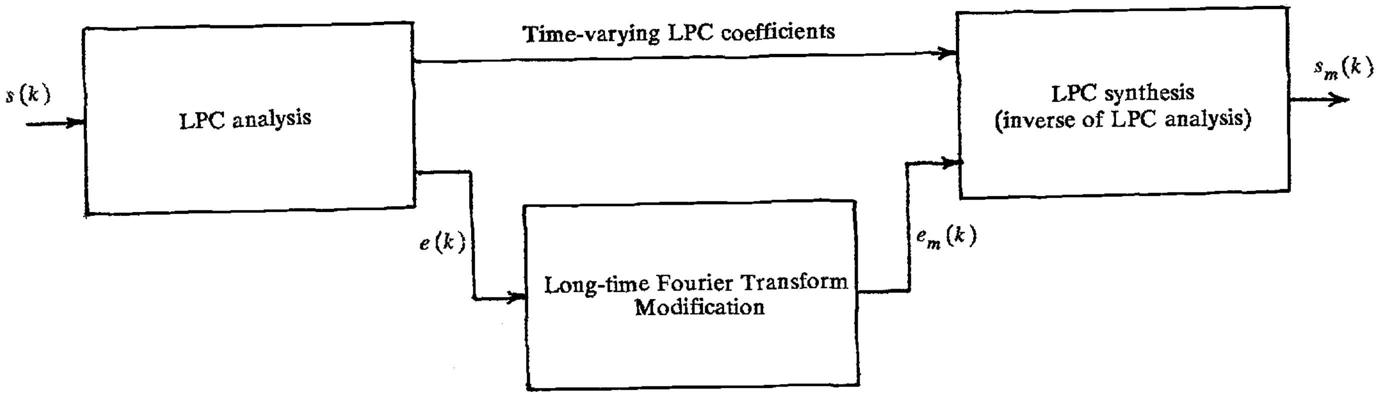
s(k)
LPC analysis
Time-varying LPC coefficients
LPC synthesis
(inverse of LPC analysis)
s_m(k)
e(k)
Long-time Fourier Transform
Modification
e_m(k)

Figure 1