# S10a.1

## ROBUST LPC ANALYSIS AND SYNTHESIS USING THE KL TRANSFORMATION OF ACOUSTIC SUBWORDS SPECTRA

V. Ralph Algazi, Sang Chung*, Michael J. Ready**, Kathy L. Brown

Speech Research Lab
CIPIC, Center for Image Processing and Integrated Computing
University of California, Davis

### ABSTRACT

LPC analysis and synthesis is a classical method for the representation and coding of speech which has received widespread acceptance. While the performance of LPC is often satisfactory in a controlled, noise free environment, it will degrade rapidly in the presence of noise. This paper proposes a new approach to the modeling and estimation of the speech spectral envelope over acoustic sub-words that exhibits robust performance in noise. The technique exploits the underlying signal structure of speech to improve parameter estimates, as well as the perceptual properties of hearing to decrease the computational requirements in a perceptually meaningful way. The new approach provides dramatic speech quality improvement over other methods.

### I. INTRODUCTION

Linear Predictive Coding (LPC) of speech is one of the most successful parametric modeling techniques in speech analysis. Because LPC models speech signals by comparatively fewer parameters than other analysis schemes, it has become the predominant technique for low bit-rate transmission and for speech recognition tasks.

In most practical environments, speech is degraded by additive background noise. The accurate estimation of speech parameters in noisy environments is important not only for the bandwidth compression of noisy speech but also for subsequent enhancement of the speech signal.

However it has been established that LPC performance degrades rapidly in the presence of noise[1, 2]. Based on a new speech signal representation strategy, we have developed a new spectral parameter estimation approach that exploits speech signal characteristics and the perceptual properties of human hearing and exhibits robust performance in noise.

There are at least two conventional approaches to improve LPC performance in noise: 1) direct parameter estimation techniques designed to be robust in noise; and 2) speech enhancement prior to LPC parameter estimation. Direct AR parameter estimation algorithms that have been developed to reduce the effects of noise include Wiener type filtering, Autoregressive Moving Average (ARMA) modeling, and bias subtraction methods. Most of these methods, proposed in the context of system identification rather than for speech analysis/synthesis estimate only the filter coefficients and do not consider gain, an important parameter for speech synthesis. Furthermore, they are computationally intensive nonlinear techniques with stability problems.

The alternative approach to LPC parameter estimation in noise is to use an enhancement algorithm as a preprocessor to LPC. Such an approach is advantageous because it circumvents the instabilities and nonlinearities encountered by the direct parametric estimation schemes. But the speech resulting from such an approach usually retains some musical tones and background noise characteristic of the enhanced speech.

To overcome the poor performance of the classical LPC based algorithms, including the above two approaches, recent work employs information on the noise statistics in addition to the noisy speech statistics. Working principally on a mathematically-based analysis frame (typically 10-20 msec) they track the nonstationarity of speech but do not exploit fully the underlying structure of the speech signal.
As a result, the short time estimation interval (frame) of the previous approaches leads to unsatisfactory statistics of the noisy speech, which, in turn, results in poor performance.

We propose a new spectral parameter estimation approach that exploits both speech signal characteristics and the perceptual properties of human hearing. We have developed a new framework for speech processing, contrasting sharply with conventional approaches, that exploits linguistic knowledge embodied in the speech signal. Speech signals, although generally considered to be nonstationary, have a relatively well defined structure, related to linguistic events, that can be exploited in the development of speech processing tasks. We partition the speech into variable duration subword units that have approximately stationary spectral characteristics. These sub-word units, denoted acoustic subwords are typically much longer than 10-20 msec. By using an acoustic sub-word rather than a frame as the estimation interval, we can obtain more accurate statistics of signals and more accurate estimates of the parameters.

Additionally, we exploit perceptual properties to relax constraints on the estimation of the speech signal. Two perceptual properties of hearing used in this work are the insensitivity to the short time Fourier transform (STFT) phase[3] and the spectral resolution. The spectral resolution of hearing, proportional to the critical bands is used to decrease the dimensionality of the representation space of the speech spectral envelope in a perceptually meaningful way.

In addition to the AR parameters, pitch plays a key role in the performance of LPC modeling of speech. The accurate and reliable measurement of pitch is exceedingly difficult in clean speech[4] as well as in noisy speech[5]. We have developed a new pitch detection and estimation algorithm, reported in a previous paper[6], that exploits speech signal characteristics and performs robustly in noise.

We assume an additive noise model together with the following assumptions: 1) the speech and noise signals are uncorrelated; 2) only the noisy signal is available for analysis; 3) estimates of the noise characteristics can be obtained during non-speech activity; 4) the noise characteristics are fixed during the speech utterance.

---

* now with AMD, Austin, TX
** now with AST, Inc. Sunnyvale, CA

This research supported in part by the Research Program (MICRO) of the University of California and by Pacific Bell and Hewlett Packard



CH2673-2/89/0000-0468 $1.00 © 1989 IEEE

## II. NEW APPROACH TO SPEECH PROCESSING

It has been established in acoustic-phonetics that speech can be decomposed into a set of linguistic units called phonemes that exhibit distinctive acoustic properties.

Because the shape of the vocal tract varies with the phonetic content of the speech signal, speech signals are characterized as slowly time varying nonstationary signals. As a result, speech processing schemes employ a short-time, 10-20 msec, stationarity model of the speech signal.

However phonetic sounds, characterized by relatively stationary vocal tract configurations, represent higher-level segments of stationarity that are much longer than the 10-20 msec short-analysis frame. Such high-level segments exhibit a high degree of correlation between short-analysis frames. We exploit this linguistically based global structure of speech by decomposing the speech signal into *acoustic sub-words* that represent *acoustically* homogeneous regions in the spectral domain.

## III. NEW APPROACH TO AR PARAMETER ESTIMATION IN NOISE

It is well known that the Short Time Fourier Transform (STFT) magnitude is important to the perceptual properties of hearing while the STFT phase is not[3]. Generally, the short time spectral magnitude of speech can be decomposed into two components: a fine structure due to the excitation source and a gross spectral envelope due to the shape of the vocal tract. The fine structure (pitch) is estimated using the robust UCC pitch detection algorithm reported in a previous paper. In this section, we concentrate on the estimation of AR parameters from the spectral envelope of speech.

The philosophy is to obtain accurate AR parameter estimates by exploiting the acoustic sub-word rather than frame size for the estimation interval and by using a signal dependent Karhunen-Loeve (KL) filter for spectrum estimation. Briefly, our approach provides optimal filtering of the noisy speech spectrum and then computes the AR parameters directly from the estimated speech spectral envelope.

Each acoustic sub-word is a spectrally homogeneous region composed of multiple highly correlated analysis frames. The speech spectral envelope is estimated by applying an estimator, the Optimal KL Filter (OKLF), that takes the estimated speech spectral envelope in the frequency domain. Figure 1 shows a system block diagram of the new approach.

The new approach has 2 major blocks: Spectral Envelope Estimation and AR Parameters Estimation. Spectral Envelope Estimation involves four operational blocks: 1) the ACOUSTIC SUB-WORD SEGMENTATION block parses the speech into acoustic sub-words 2) the ACOUSTIC SUB-WORD ANALYSIS block calculates the OKLF parameters based on characteristics obtained over the whole acoustic sub-word. 3) the ACOUSTIC SUB-WORD BUFFER holds the acoustic sub-word while the OKLF parameters are estimated. 4) the OPTIMUM KL FILTER block estimates the spectral envelope of an acoustic sub-word. AR PARAMETER Estimation extracts the AR parameters from the spectral envelope estimate. Each acoustic sub-word is processed independently because the spectral characteristics between acoustic sub-words, not necessarily correlated, require different OKLF parameters. The system operates on two different internal time scales. The ACOUSTIC SUB-WORD SEGMENTATION and ANALYSIS blocks are event driven by the signal characteristics and operate on an acoustic sub-word basis. The OKLF parameters are calculated independently for each sub-word and remain fixed for the duration of the sub-word. In contrast, the KL FILTER and AR PARAMETERS estimates blocks operate on a fixed frame (25.6 msec) scale determined by the resolution of human hearing in the time domain and are independent of signal characteristics. Each major block is considered in detail below.

### Speech Spectral Envelope Estimation

The ACOUSTIC SUB-WORD SEGMENTATION block segments the spectrum into approximately stationary acoustic sub-words. A speaker independent segmentation algorithm previously reported in ICASSP 1988[7] automatically decomposes speech based on spectral criteria into multiple quasi-stationary acoustic sub-words.

Under the assumption of sub-word stationarity, the spectral characteristics do not change significantly. Then the OKLF derived from the sub-word as a whole can be used to estimate the spectral envelope from the noisy speech envelope for all the frames in the sub-word. The KL FILTER block estimates the clean speech spectral envelope by applying the OKLF to the noisy speech spectral envelope. The OKLF parameters are fixed for the duration of each sub-word.

The algorithm works as follows[8]; The speech is windowed by a 256 point Hanning window and a 256 point DFT computed. The magnitude and phase are separated giving 128 unique STFT magnitude components, $Y(k,l)$, where $k$ is the discrete frequency index and $l$ is the frame index. The noisy speech magnitude envelope $Y_e(k,l)$ is first estimated for each frame by smoothing $Y(k,l)$ with a 7 point rectangular window.

The smoothed STFT magnitude envelope is approximated by its average value within 22 critical bands between 0 and 4KHz, corresponding to the first 104 samples of $Y_e(k,l)$. This operation results in the vector $Y(m,l)$ where $m$ is the band index. The approximate 22 critical band spectrum, $Y(m,l)$, is denoted the Bank of Filters (BOF) output because it approximates the output of bandpass filters. The analysis window is then advanced 128 points to the next frame and the analysis performed on this new frame.

The OKLF estimator parameters are obtained on an acoustic sub-word basis, from the output of BOF, $Y_e(m,l)$. The spectral envelope for each frame is obtained by a generalized Wiener filtering operation based on the mean squared error criterion:

$$\epsilon = E\left\{ |\hat{\mathbf{S}}_e(l) - \mathbf{S}_e(l)|^2 \right\} \tag{1}$$

where $E$ is the expectation operator and $\mathbf{S}_e(l)$ is a 22 dimensional speech spectral envelope vector:

$$\mathbf{S}_e(l) = [S_e(1,l), S_e(2,l), \ldots\ldots S_e(22,l)]^t .$$

Since the spectral envelope has a nonzero mean, the linear estimator, $\hat{\mathbf{S}}_e(l)$, of the speech spectral envelope is restricted to the form;

$$\hat{\mathbf{S}}_e(l) = \mathbf{A}\mathbf{Y}_e(l) + \mathbf{b} \tag{2}$$

For white noise, the linear estimator, $\hat{\mathbf{S}}_e(l)$, which minimizes the mean squared error, $\epsilon$, of equation (1) is obtained by taking the gradient of equation (1) with respect to $\mathbf{A}$. By substitution of equation (2) into equation (1) under the restriction of an "unbiased estimator," $\hat{\mathbf{S}}_e(l)$ is given by

$$\begin{aligned} \hat{\mathbf{S}}_e(l) &= C_s C_y^{-1} \mathbf{Y}_e(l) + \overline{\mathbf{S}}_e - C_s C_y^{-1} \overline{\mathbf{Y}}_e \\ &\equiv \mathbf{A}_o [\mathbf{Y}_e(l) - \overline{\mathbf{Y}}_e] + \overline{\mathbf{S}}_e \end{aligned} \tag{3}$$

where the optimum filter $\mathbf{A}_o$ is defined as $\mathbf{A}_o \equiv C_s C_y^{-1}$, $\overline{\mathbf{S}}_e$, $\overline{\mathbf{Y}}_e$ denote the average spectral envelopes over the sub-word; and $C_s$, $C_y$ denote the covariance matrices of the spectral envelope of the clean and noisy speech respectively.

The optimal filter $\mathbf{A}_o$ is computed for each subword segment . The filtering is implemented in KL space. Thus the linear estimator, using the optimal filter $\tilde{\mathbf{A}}_o$ in the KL domain, is given by

$$\hat{\mathbf{S}}_e(l) = \mathbf{T}^t \tilde{\mathbf{A}}_o \mathbf{T} [\mathbf{Y}_e(l) - \overline{\mathbf{Y}}_e] + \overline{\mathbf{S}}_e \tag{4}$$

where $\mathbf{T}$ is the KLT matrix formed by the eigenvectors of $\mathbf{A}_o$ and the optimum filter in KL space, $\tilde{\mathbf{A}}_o$, is a diagonal matrix.

**Estimation of the Optimum KL Filter Parameters**

We now obtain the signal dependent optimal KL filter parameters, $\mathbf{T}$, $\tilde{\mathbf{A}}_o$, $\overline{\mathbf{Y}}_e$, and $\overline{\mathbf{S}}_e$, in the spectral domain over the whole sub-word.

$$\overline{\mathbf{Y}}_e = \frac{1}{L}\sum_{l=1}^{L}\mathbf{Y}_e(l) \tag{5}$$

$$C_y(p,p) = \frac{1}{L}\sum_{l=1}^{L}\mathbf{Y}_e(l)\mathbf{Y}_e(l)' \tag{6}$$

where $L$ is the number of the frames in the acoustic sub-word and $\mathbf{Y}_e(l)$ is the BOF estimate of the noisy speech envelope for each frame.

For white noise, the eigenvectors of $\mathbf{A}_o$ are the same as the eigenvectors of the noisy speech covariance matrix $C_y$ and that of the clean speech matrix $C_s$ The optimal filter parameters, unitary transform matrix, $\mathbf{T}$, and optimal filter, $\tilde{\mathbf{A}}_o$, can be obtained based on eigenvalue analysis of $C_y$. $\tilde{C}_y$ is a KL domain version of $C_y$ and is a diagonal matrix with elements corresponding to the eigenvalues of $C_y$ denoted $\lambda(p)$ where p is the index in KL space. Since the noise is assumed to be independent of speech and white, *i.e.*, $C_y = C_s + C_d$, $\tilde{C}_s$ is also diagonal and can be approximated by

$$\tilde{C}_s(p,p) = \max\left\{\tilde{C}_y(p,p) - \tilde{C}_d(p,p)\,,\,0\right\} \tag{7}$$

$$= \max\left\{\lambda(p) - \tilde{C}_d(p,p)\,,\,0\right\} \tag{8}$$

where $\tilde{C}_d(p,p)$ are the diagonal elements of the KL space version of $C_d$. Next, the $\tilde{C}_s(p,p)$ is thresholded at zero because the covariance matrix should be non negative definite. The optimal filter, $\tilde{\mathbf{A}}_o$, is given by

$$\tilde{\mathbf{A}}_o = \tilde{C}_s\tilde{C}_y^{-1} = \frac{\max(\lambda(p) - \tilde{C}_d(p,p)\,,\,0)}{\lambda(p)} \tag{9}$$

This filtering operation is thus similar to spectral subtraction , but it is carried out on acoustic subword segments and in the signal dependent KL domain.

**Spectral Linear Modeling- AR Parameters Estimation.**

LPC models spectra as filter bank spectra, by an all pole spectrum.

The AR model spectrum corresponds to a transfer function, $\hat{S}(z)$ given by

$$\hat{S}(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{i=1}^{p} a_i z^{-i}} \tag{10}$$

where $\hat{S}(z)$ is the $z$ transform of $\hat{s}(n)$ and $G$ is the gain parameter of the model. Then its power spectrum $\hat{P}(\omega)$ is given by

$$\hat{P}(\omega) = \frac{G^2}{|1 + \sum_{i=1}^{p} a_i e^{-j\omega i}|^2} = \frac{G^2}{|A(\omega)|^2} \tag{11}$$

The filter coefficients $\{a_i\}$ are obtained by minimizing the total squared prediction error $\mathbf{E}$ with respect to $a_i$, $i = 1,2,\ldots,p$ where

$$\mathbf{E} = \frac{1}{N}\sum_{n=0}^{N-1}|E(\omega_n)|^2 \tag{12}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}P(\omega_n)(1 + \sum_{i=1}^{p}a_i e^{-i\omega n})(1 + \sum_{i=1}^{p}a_i e^{i\omega n}) \tag{13}$$

The filter coefficients are obtained from

$$R(i) = \sum_{k=1}^{p} a_k R(i-k) \qquad i = 1,2,\ldots\ldots,p \tag{14}$$

where

$$R(i) = \frac{1}{N}\sum_{n=0}^{N-1}P(\omega_n)\,cos(i\omega_n) = Re\left\{F^{-1}P(\omega_n)\right\} \tag{15}$$

Gain is obtained from the minimum prediction error energy $E_{\min}$ and is given by[9]

$$G^2 = R(0) + \sum_{i=1}^{p} a_i\,R(i) \tag{16}$$

Eq.(14) and Eq.(16) are the LPC normal equations.

We obtain the AR parameters by applying the LPC Spectral Analysis to our specific BOF spectra. Since the 22 BOF spectra are not equally spaced in the frequency domain, the FFT can not be used directly in the analysis of the BOF spectra, $S_B(\omega_l)$. **Linear interpolation** is performed to convert the 22 nonequally spaced filter bank spectra $S_B(\omega_l)$ into a new 128 point equally spaced spectrum. As a result of the linear interpolation, 256 equally spaced STFT magnitude points, $S(\omega_j)$, are obtained in the frequency domain.

Since the linearly interpolated $S(\omega_j)$ are still the STFT magnitude, a squaring operation is performed.

$$P(\omega_j) = S^2(\omega_j), \qquad j=1,2,\ldots.256 \tag{17}$$

so that LPC techniques can be applied to obtain the AR parameters.

Given a spectrum $P(\omega_j)$ and a desired model order, $p$, of the all pole filter, a 256 point cosine transform is applied to the 256 point spectrum $P(\omega_j)$ and $p+1$ number of autocorrelation coefficients, $R(i)$, $i=0,1,2,\ldots\ldots,p$, are computed. The all pole filter coefficients $\{a_i\}$, i = 1,2,.....,p, are obtained by using the Levinson's recursive algorithm to solve $p$ linear equations of the form of Eq.(14). The gain parameter, $G$, can be obtained by substituting the filter coefficient and autocorrelation coefficients into Eq.(16).

**IV. Performance Evaluation of the New Approach**

We have compared LPC the performance of our new algorithm, denoted SBLPC, for Segment Based LPC, with a conventional system (using SIFT for pitch extraction ), and two preprocessing approaches: spectral subtraction (SS) and Boll's residual noise reduction (RNR). Figure 2 shows the performance of these approaches as a function of peak signal to noise ratio (SNR). The performance of these systems has been evaluated for various noise levels using Schroeder's objective measure based on perceptual properties[10]. In the figure, Ls is the loudness or perceived intensity of the speech (masking the noise), and Ln is the loudness of the noise in the presence of the masking speech. Ln takes into account the masked and absolute thresholds of perception, in each of the critical bands and for each of the frames. When D=1, the noise is approximately as loud as the speech. For D=0, the noise is inaudible, principally because of masking.[10]

As shown the new SBLPC approach produces better quality speech than conventional LPC or preprocessing with either SS or RNR.

The synthetic speech from the new approach sounds very similar to the speech synthesized from the clean speech at all noise levels. The speech from Spectral Subtraction and Boll's retain some artifacts present in the enhanced speech (some residual background noise and musical tone noise). For high noise levels, the first notable degradation of SBLPC is the loss of very low energy acoustic subwords. Thus the quality and intelligibility will degrade together for a noise threshold value of 20DB. Note that this noise level is also the performance of the UCC pitch extraction algorithm also degrades significantly.

Perceptually, the performance of the new approach is substantially better than for any other AR parameter estimation methods.

## V. Discussion and Conclusion.

The method presented makes use of the least number of parameters to characterize globally the spectral envelope. Thus, the spectral envelope is characterized over an acoustic subword which encompasses a variable number of frames, ten or more is common, to provide a robust estimate of the envelope for each frame. Thus, it is, in that sense an extension of the residual noise reduction technique of Boll to a large number of frames, but with a prior segmentation into spectrally homogeneous subwords .

The use of the KL transformation over each segment is very effective in noise removal, but requires computation of eigenvalues and eigenvectors of the covariance matrices of dimension [22 by 22]. Coupled with the UCC pitch extraction technique, the approach results in very clean synthetic speech for quite noisy conditions. It thus appear as an promising approach to speech enhancement as well as for analysis synthesis in noise.

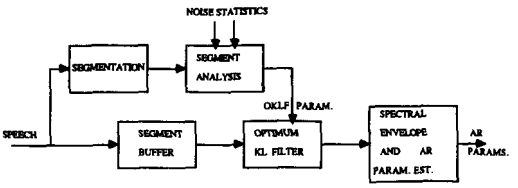


**Figure 1. KL Estimation of LPC Parameters**

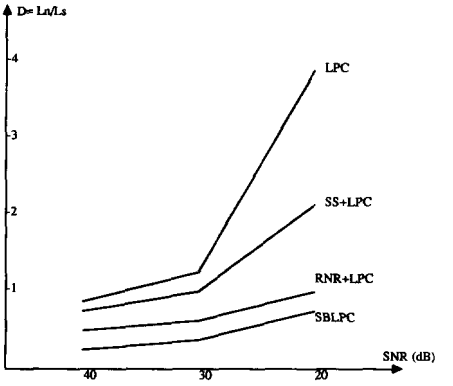


**Figure 2.Comparative Performance of SBLPC by Perceptual Measure [10]**

## References


1. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantization Noise or Additive Noise," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. ASSP-24, pp. 488-494, Dec. 1976.
2. S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-27, no. 5, pp. 478-485, Oct. 1979.
3. D. L. Wang and J. S. Lim, "The Unimportance of Phase in Speech Enhancement," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. ASSP-30, no. 4, pp. 679-681, Aug. 1982.
4. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. Mcgoneal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-24, no. 5, pp. 399-418, Oct. 1976.
5. K. A. Oh and C. K. Un, "A Performance Comparison of Pitch Extraction Algorithms for Noisy Speech," *IEEE Proc. ICASSP*, pp. 18B.4.1-18B.4.4, San Diego, Ca., Mar. 1984.
6. S. Chung and V. R. Algazi, "Improved Pitch Detection Algorithm for Noisy Speech," *IEEE Proc. ICASSP*, pp. 407-410, Tempa, Fla., Mar. 1985.
7. V. R. Algazi and K. L. Brown, "Automatic Speech Recognition Using Acoustic Sub-words and No Time Alignment," *IEEE ICASSP*, New York, New York, April 1988.
8. M. J. Ready, "Enhancement of Noisy Speech Based on Speech Production and Perceptual Models," *Ph. D. thesis*, p. University, Davis, Ca, Dec. 1984.
9. J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. ASSP-21, no. 3, pp. 140-148, June 1973.
10. M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, pp. 1647-1652, Dec. 1979.