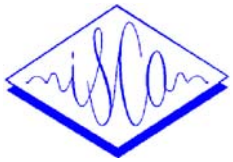

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECH ENHANCEMENT FOR LINEAR-PREDICTIVE-ANALYSIS-BY-SYNTHESIS CODERS

Marcin Kuropatwinski[1], Dieter Leckschat[1], Kristian Kroschel[2], Andrzej Czyzewski[3], Chaz Hales[1]*

[1]SIEMENS AG, Information and Communication Products, Dept. ICP CD TI24, Bocholt, Germany
[2]Institut f. Nachrichtentechnik, Universität Karlsruhe, Karlsruhe, Germany
[3]Sound Engineering Dept., Politechnika Gdanska, Gdansk, Poland

ABSTRACT

Speech coding techniques commonly used in low bit rate analysis-by-synthesis linear predictive coders (LPAS coders) create a model that emphasizes the important features of a speech signal. The utilization of these coding methods for speech enhancement is shown. Specifically, the speech signal will be modeled as the output of a cascade of an adaptive formant filter and an adaptive pitch filter, driven by a white Gaussian process with a time changing variance. The parameter and signal estimation method, which is based on the Expectation Maximization (EM) algorithm and implements this speech signal model, is investigated. The proposed approach yields better performance both in SNR and subjective impression than do speech enhancement methods which use only AR speech signal parameters.

1. INTRODUCTION

Research on speech coding has resulted in a number of solutions to the problem of accurately representing a speech signal digitally at the lowest bit rate possible. These solutions involve achievements in source coding theory, like vector quantization, as well as methods which came from the experimental trials aimed at finding the most appropriate speech signal representation. Knowledge about the nature of the speech signal gained through this research can be utilized in other areas of speech signal processing such as speech enhancement and noise reduction in mobile telephony.

It is a known fact that for noisy speech inputs, the performance of low bit rate coders deteriorates considerably. This is illustrated in Table. 1, which shows the mean opinion scores (MOS), estimated using the method standardized by the ITU [1], for a GSM Enhanced Full Rate (GSM EFR) Coder and a GSM Half Rate (GSM HR) Coder. It can be seen that the reduction of the signal quality, i.e. the reduction of the MOS score of the noisy signal compared to that of the clean signal, is greater as the bit rate decreases.

	Estimated MOS scores		
	Reference signal	GSM EFR 12.2kb/s	GSM HR 5.6kb/s
clean speech	4.03	3.88	3.31
noisy speech 5dB SNR	4.03	3.50	2.62

Table 1.: Estimated MOS scores for two GSM Coders with clean and noisy speech inputs

* Corresponding author: marcin.kuropatwinski@bch.siemens.de

The quality of the transcoded noisy signal vs. bit rate is a particularly important consideration in adaptive mode rate coders. Using speech enhancement with these coders allows the coder to operate at a higher channel rate (lower source rate) in the noisy environment without compromising the transcoded signal quality.

To make the speech coders robust against noise, the vector quantizers of the short term parameters are trained with a training set including noisy data [2]. This implies that more bits must be used to code the STP parameters.

The information theory point of view (communication through an acoustic channel in the presence of noise) adds interesting insights into the speech enhancement problem. The possibility of fully recovering a clean signal, i.e. a signal whose entropy is bounded above by the bit rate of the currently available coders, about 2.4kb/sec for the telephone band speech, exists only if we assume a white Gaussian channel (a general, physically reasonable assumption) with SNR greater than –3dB. Assuming such a channel, the problem is how to make use of this possibility.

The method proposed in this paper of creating a noise reduction system that attempts to achieve this theoretical upper bound is to use the speech signal representation given by the synthesis model of an LPAS coder [3] to model the joint probability density of the speech samples in a given frame. The resulting probability density function is used within a statistical framework to estimate the parameters of the coder in the presence of additive noise. The method of estimating the formant and pitch filter parameters of an LPAS coder, based on the Expectation Maximization (EM) algorithm, is provided. This method allows us to incorporate in the estimation procedure the a'priori distributions of the parameters in the form of a vector quantizer. The proposed method is compared to the method of modeling the speech signal as an AR process.

2. SIGNAL MODEL FORMULATION

The sum of statistically independent speech and noise signals is observed:

$$r(t) = s(t) + n(t)$$

where t is the time index. The noise is assumed to be white Gaussian noise with standard deviation σ_n.

The probability density function of the speech signal is derived from the Code Excited Linear Prediction coder system [4] shown in Fig. 1. For our purposes, we model the

excitation of the CELP system as a zero mean, variance one, white Gaussian noise process. The validity of this model is shown by observing the output of an analysis filter system excited by clean speech signals.

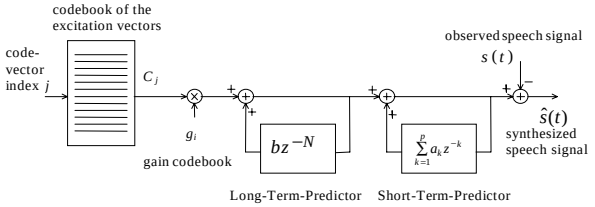

Figure 1.: Synthesis model of the CELP coder

A number of utterances were processed through the cascade of the analysis LTP and STP filters. The variances of the resulting long term residual signal were normalized for each frame, and the probability density functions were estimated for each frame using a kernel density estimator.

Fig.2a shows the estimated distribution of the residual signal samples after removing only the short term correlation from the speech signal. After removing both the short and long term correlation from the speech signal, the resulting residual is approximately normally distributed. This is shown in Fig.2b.

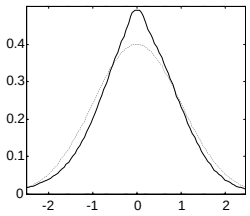

Fig. 2a: Distribution of the short term residual signal samples (solid); Gaussian distribution with the same mean and variance (dashed)

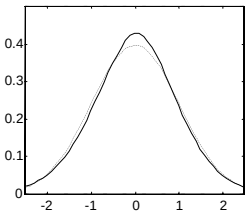

Fig. 2b: Distribution of the long term residual signal samples (solid); Gaussian distribution with the same mean and variance (dashed)

By comparing Figures 2a and 2b it can be seen that adding the long term analysis filter increases Gaussianity of the residual. Near Gaussian distribution of the long term residual allows us to model the speech signal as the output of the cascade of a pitch and a formant filter driven by white Gaussian noise with time changing variance.

Accordingly, the probability distribution of one frame of speech samples, conditional upon the STP and LTP parameters and previous speech samples, is given by:

$$p_s(\mathbf{s}\,|\,\theta,\mathbf{s}_N(0)) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma_s^N}\exp\left\{-\frac{1}{2\sigma_s^{\,2}}\sum_{t=1}^{N}[s(t)+\mathbf{a}^T\mathbf{s}_p(t-1)+bs_{sr}(t-L)]^2\right\}$$

with the following notation used:

N - frame length

$s(t)$ - speech sample at the time instant t

(t = 1 corresponds to the frame beginning)

$s_{sr}(t)$ - short term residual sample

$\theta = [\mathbf{a}^T, b, L, \sigma_s]$ - parameters of the speech signal

$\mathbf{a} = [a_p, a_{p-1}, ..., a_1]^T$ - speech signal AR parameters

b - pitch predictor tap

L - pitch predictor lag

σ_s^2 - variance of the driving term (excitation signal)

p (subscript) - STP synthesis filter order

$\mathbf{s}_n(t) = [s(t)...s(t-n+1)]^T$ - vector of the recent n signal samples

$\mathbf{s} = [s(1)...s(N)]^T$ - vector of the speech samples in the current frame.

3. PARAMETER AND SIGNAL ESTIMATION ALGORITHM

We assume that the speech signal parameters are deterministic, or are uniformly distributed over the parameter space. This assumption suggests using the Maximum Likelihood (ML) method to estimate the parameters. In this method the probability distibution of the observation, in this case the observed noisy frame, conditional upon the speech and noise parameters, is maximized over the parameter space. The maximizing parameter vector is called the ML parameter estimate. For of our assumed signal model, the ML function is equal to:

$$g(\mathbf{r}\,|\,\theta,\sigma_n) = \int_{\Omega_s} p_n(\mathbf{r}-\mathbf{s}\,|\,\sigma_n)\,p_s(\mathbf{s}\,|\,\theta)d\mathbf{s} \qquad (1)$$

where $\mathbf{r} = [r(1)...r(N)]^T$ is the vector containing samples of the observed noisy frame. Because it is difficult to compute the ML estimates by maximizing (1) directly over the parameter set, the estimation will be performed using the Expectation-Maximization (EM) algorithm [5], a frequently used tool for ML estimation problems which cannot be solved analytically.

The EM algorithm is an approach to iterative computation of ML estimates when the observations can be viewed as incomplete data. The algorithm assumes the existence of two sample spaces: X, called complete data, and Y, called incomplete data, with a many to one mapping from X to Y. In the case of speech enhancement, the complete data space is given by the speech and noise vectors, X = [n, s], and the incomplete data consists of the noisy, observed vector, Y = [n+s]. The mapping from the complete data space to the incomplete data space is the summation of the speech and noise signal in the acoustic channel.

Parameter estimation using the EM algorithm is performed in two iteratively repeated steps:

1. *expectation step* (E-step): which comprises computation of the function: $Q(\theta,\theta_{(k)}) = E[\ln p(\mathbf{s},\mathbf{n}\,|\,\theta,\sigma_n)\,|\,\mathbf{r},\theta_{(k)}]$
2. *maximization step* (M-step): choosing $\theta_{(k+1)}$ to be a value of $\theta \in \Omega$ which maximizes $Q(\theta,\theta_{(k)})$, where Ω is a convex parameter set.

These steps are repeated until the required convergence is achieved.

E-step:

Since **s** and **n** are statistically independent, the log-likelihood function of the pdf of the complete data is equal to:

$$\ln p(\mathbf{s},\mathbf{n}\,|\,\theta,\sigma_n) = \ln p_s(\mathbf{s}\,|\,\theta) + \ln p_n(\mathbf{n}\,|\,\sigma_n) \qquad (2)$$

with:

$$\ln p_s(\mathbf{s}\,|\,\theta) = C - N\ln(\sigma_s) - \frac{1}{2\sigma_s^2}\sum_{t=1}^{N}[s(t) + \mathbf{a}^T\mathbf{s}_p(t-1) + bs_{sr}(t-L)]^2$$

and:

$$\ln p_n(\mathbf{n}\,|\,\sigma_n) = C - N\ln(\sigma_s) - \frac{1}{2\sigma_n^2}\sum_{t=1}^{N} n^2(t)$$

To compute the function $Q(\theta, \theta_{(k)})$, we take the conditional expectation of the log-likelihood function, given the observed vector **r** and the parameter values in the k-th iteration:

$$E[\ln p_s(\mathbf{s}\,|\,\theta)\,|\,\mathbf{r},\theta_{(k)}] = C - N\ln\sigma_s - \frac{1}{2\sigma_s^2}\sum_{t=1}^{N} E\{[s(t) + \mathbf{a}^T\mathbf{s}_p(t-1) + bs_{sr}(t-L)]^2\,|\,\mathbf{r},\theta_{(k)}\} \qquad (3)$$

Taking the conditional expectation of the log-liklihood function of the noise signal is analogous to the above operation for the speech signal.

The a'posteriori second order statistics, which are needed to compute (3), are defined as:

$$\mathbf{R}(t\,|\,t) = E[\mathbf{x}(t)\mathbf{x}^T(t)\,|\,\mathbf{r},\theta_{(k)}] \qquad (4)$$

and can be obtained using a Kalman filter [6]. From Kalman filter theory, the value $\mathbf{R}(t\,|\,t)$ can be expressed in terms of the error covariance matrix $\mathbf{K}(t\,|\,t)$ and the system state $\mathbf{x}(t)$:

$$\mathbf{R}(t\,|\,t) = \mathbf{K}(t\,|\,t) + \mathbf{x}(t)\mathbf{x}^T(t) \qquad (5)$$

To use the Kalman filter, we put the introduced signal model into the state space form given by:

$$\mathbf{x}(t) = \mathbf{F}(t\,|\,t-1)\mathbf{x}(t) + \mathbf{v}_1(t) \qquad \text{(process equation)} \qquad (6)$$

$$r(t) = \mathbf{C}\mathbf{s}(t) + v_2(t) \qquad \text{(measurement equation)} \qquad (7)$$

The corresponding Kalman filter matrices can be found as follows. Samples of the speech signal in terms of the short and long term residuals can be expressed as:

$$s(t) = \sum_{k=1}^{p} a_k s(t-k) + s_{sr}(t) \qquad (8)$$

where the short term residual signal given by:

$$s_{sr}(t) = bs_{sr}(t-L) + s_{lr}(t) \qquad (9)$$

As shown above, the long term residual signal can be represented as white Gaussian noise with time changing variance.

For the case where the *state vector* of the Kalman filter is defined as:

$$\mathbf{x}(t-1) = \begin{bmatrix} \mathbf{s}_{sr\,N}(t) \\ \hline \mathbf{s}_{p+1}(t-1) \end{bmatrix}, \qquad (10)$$

the *state transition matrix* based on (8) and (9) is given by:

$$\mathbf{F}(t\,|\,t-1) = \begin{bmatrix} \mathbf{B} & \mathbf{0}_{N\times(p+1)} \\ \hline \mathbf{0}_{(p+1)\times(N-1)} & \mathbf{A} \end{bmatrix} \qquad (11)$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{0}_{(N-1)\times 1} & \mathbf{I}_{(N-1)\times(N-1)} \\ \hline 0 & \mathbf{0}\;\; b\;\; \mathbf{0}_{1\times(N-L-2)} \end{bmatrix} \qquad (12)$$

and

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{p\times 1} & I_{p\times p} \\ \hline 1 & 0\;\; \mathbf{a}^T \end{bmatrix} \qquad (13)$$

In these equations, $\mathbf{0}_{m\times n}$ is the $m\times n$ zero matrix, and $\mathbf{I}_{m\times n}$ is the $m\times n$ identity matrix.

The covariance matrix of the *process noise vector*, $\mathbf{v}_1(t)$ in (6), is given by:

$$\mathbf{Q}_1 = \sigma_s^2\mathbf{g}\mathbf{g}^T \qquad \mathbf{g} = [\mathbf{0}_{1\times(N-1)}\;\vdots\;1\;\vdots\;\mathbf{0}_{1\times(p+1)}]^T \qquad (14)$$

The additive noise $n(t)$ is represented in the state space equations as the *measurement noise*, $v_2(t)$ in (7), and the corresponding covariance matrix is:

$$\mathbf{Q}_2 = \sigma_n^2\,. \qquad (15)$$

The *measurement matrix* is given by $\mathbf{C} = [0_{1\times(N+p)}\;\vdots\;1]$.

The Kalman filtering operation is accomplished for $t = 1...N$ by the following steps:

1. Compute the predicted state error covariance matrix:
   $\mathbf{K}(t\,|\,t-1) = \mathbf{F}(t\,|\,t-1)\,\mathbf{K}(t-1\,|\,t-1)\,\mathbf{F}^{\mathrm{T}}(t\,|\,t-1) + \mathbf{Q}_1$
2. Compute the Kalman gain matrix:
   $\mathbf{G} = \mathbf{K}(t\,|\,t-1)\mathbf{C}^T[\mathbf{C}\mathbf{K}(t\,|\,t-1)\mathbf{C}^T + \mathbf{Q}_2]^{-1}$
3. Compute the state estimate:
   $\hat{\mathbf{x}}(t) = \mathbf{F}(t\,|\,t-1)\mathbf{x}(t-1) + \mathbf{G}\,[r(t) - \mathbf{C}\,\mathbf{F}(t\,|\,t-1)\,\mathbf{x}(t-1)]$
4. Compute the filtered state error covariance matrix:
   $\mathbf{K}(t\,|\,t) = \mathbf{K}(t\,|\,t-1) - \mathbf{G}\,\mathbf{C}\,\mathbf{K}(t\,|\,t-1)$

The initial values for the given frame are taken from the frame processed previously.

M-step:

Using the results from the *E-step* of the EM algorithm the parameters maximizing the $Q(\theta,\theta_{(k)})$ function are computed according to the following set of formulas: (16)

$$\mathbf{a}_{(k+1)} = -\left(\sum_{t=1}^{N} E[\mathbf{s}_p(t-1)\mathbf{s}_p^T(t-1)\,|\,\mathbf{r},\theta_{(k)}]\right)^{-1}\sum_{t=1}^{N} E[\mathbf{s}_p(t-1)\mathbf{s}(t)\,|\,\mathbf{r},\theta_{(k)}]$$

The computation of the formant filter parameter is derived under the assumption that the interaction between the speech

samples at the time instant *t* and the past short residual signal is negligible. As shown in [7], this is a reasonable assumption which considerably simplifies the M-step, which is otherwise an iterative procedure.

The remaining speech signal parameters are computed according to:

$$L_{(k+1)} = \arg\max_{L}\left([\mathbf{a}^T_{(k+1)} : 1] \sum_{t=1}^{N} E[\mathbf{s}_{p+1}(t) s_{sr}(t-L) \,|\, \mathbf{r}, \theta_{(k)}] \right) \quad (17)$$

$$b_{(k+1)} = -\frac{[\mathbf{a}^T_{(k+1)} : 1] \sum_{t=1}^{N} E[\mathbf{s}_{p+1}(t) s_{sr}(t-L_{(k+1)}) \,|\, \mathbf{r}, \theta_{(k)}]}{\sum_{t=1}^{N} E[s^2_{sr}(t-L_{(k+1)}) \,|\, \mathbf{r}, \theta_{(k)}]} \quad (18)$$

$$\sigma^2_{s\,(k+1)} = \frac{1}{N}[b_{(k+1)} : \mathbf{a}^T_{(k+1)} : 1] \times$$

$$\left(\sum_{t=1}^{N} E\{[s_{sr}(t-L_{(k+1)}) : \mathbf{s}^T_{p+1}(t)]^T [s_{sr}(t-L_{(k+1)}) : \mathbf{s}^T_{p+1}(t)] \,|\, \mathbf{r}, \theta_{(k)}\} \right) \quad (19)$$

$$\times [b_{(k+1)} : \mathbf{a}^T_{(k+1)} : 1]^T$$

The formant filter parameters for the first iteration of the EM algorithm were derived from a correlation sequence computed in the frequency domain, which was found using the spectral subtraction algorithm. The starting parameter of the long term predictor was computed using the standard open-loop LTP parameter estimation procedure running on the noisy observation.

This straightforward implementation of the proposed algorithm is computationally very complex, however by exploiting the fact that the Kalman filter matrices are sparse, further research can be done to find a more efficient way to program the algorithm.

4. EXPERIMENTS

Computer simulations were carried out to compare the performance of the proposed method with the established speech signal estimation method using only STP parameters [8]. The utterance of duration ca. 3 seconds was contaminated with synthetic noise at three different signal to noise ratios (SNR) to assess the proposed method. In Tables 2 and 3, the SNR results for the conventional Kalman filtering method, STP only, and the new method, both STP and LTP, are shown.

The spectrograms, in Figures 3a and b, show the effect of the new noise reduction method. These spectrograms correspond respectively to the speech signal contaminated with 0dB SNR noise, and the processed signal resulting from our noise reduction method. The periodicity of the speech signal can be seen in the harmonic structures of the spectrograms.

5. DISCUSSION

It has been shown that the application of speech coding techniques for speech enhancement can result in significantly better quality of the enhanced speech. A new method of estimating the speech signal and the parameters of the typical LPAS coder in the presence of noise has been proposed. Further research will be focused on STP- and LTP-parameter estimation in the presence of noise and also vector quantization of the STP parameters.

6. REFERENCES

[1] ITU-T Rec. P.861, Geneva 1996

[2] R.P.Ramachandran, et. al., A combined vector and scalar codebook for roboust quantization of LPC parameters, SPIE, vol. 2277, pp. 167-178

[3] W.B. Kleijn, K.K.Paliwal, *Speech Coding and Synthesis*, Elsevier, 1995

[4] Manfred R. Schroeder, Bishnu S. Atal, "Code-Excited Linear Prediction (CELP) High-Quality Speech at Very Low Bit Rates", Proc. ICASSP'85, pp. 937-940

[5] T.K. Moon, "The Expectation-Maximization algorithm", IEEE SP Magazine, pp. 47-60 (Nov. 1996)

[6] K.Kroschel, *Statistische Nachrichtentheorie*, 3rd ed. Springer, 1996

[7] R.V.Ramachandran, P.Kabal, "Pitch prediction filters in speech coding", IEEE Trans. ASSP, vol. 37, pp. 467-476 (1989)

[8] J.D. Gibson, B. Koo, S.D. Gray, "Filtering of colored noise for speech enhancement and coding", IEEE Trans. SP, vol. 39, pp. 1732-1741 (1991)

Iteration no.	0dB	5dB	10dB
1	4.79	8.53	12.38
2	5.06	8.72	12.45
3	5.22	8.89	12.52
4	5.33	8.88	12.73

Table 2: SNR values of speech signal for each iteration of the enhancement algorithm. STP only. Initial SNR values are shown.

Iteration no.	0dB	5dB	10dB
1	6.95	10.36	13.98
2	7.40	10.71	14.13
3	7.49	10.74	14.15
4	7.50	10.72	14.17

Table 3: SNR values of speech signal for each iteration of the enhancement algorithm. STP and LTP. Initial SNR values are shown.

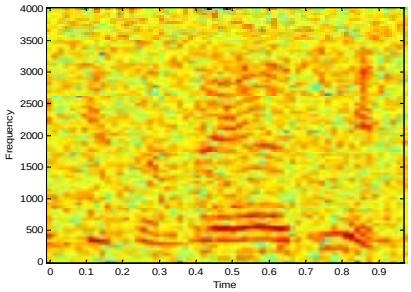

Fig. 3a: Speech signal contaminated with SNR 0dB noise.

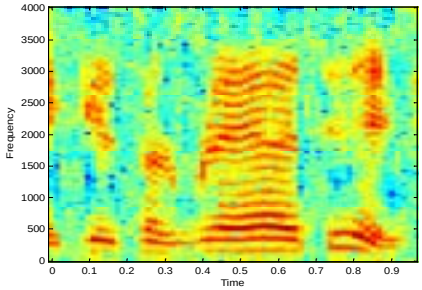

Fig. 3b: Speech signal enhanced with STP and LTP parameters.