ESTIMATION OF SHORT-TERM PREDICTOR PARAMETERS FOR CODING AND ENHANCEMENT OF NOISY SPEECH

Sriram Srinivasan, Jonas Samuelsson and W. Bastiaan Kleijn

Dept. of Signals, Sensors and Systems
KTH (Royal Institute of Technology), Stockholm, Sweden
{sriram.srinivasan, jonas.samuelsson, bastiaan.kleijn}@s3.kth.se

ABSTRACT

In this paper, we describe a technique for obtaining estimates of the short-term predictor parameters of speech under noisy conditions. We use a-priori information about speech in the form of a trained codebook of speech linear predictive coefficients. The contribution of this paper is two-fold. First, we provide a framework where the standard vector quantization search to obtain the quantized linear predictive coefficients can be replaced by a maximum likelihood search, given the noisy observation, the speech codebook and an estimate of the noise. This results in an enhancement method that is integrated with parametric coders such as linear predictive analysis-by-synthesis coders. Second, we provide a scheme where the chosen vector is not restricted to be an element of the codebook. An interpolative search between the maximum likelihood estimate and its nearest neighbors in the codebook is used to improve the precision of the estimated parameters. Such a scheme is relevant when enhancement is considered separately from coding. Experimental results show improved performance for the proposed methods.

1. INTRODUCTION

With the increasing spread of mobile communications, enhancing speech subjected to background acoustic noise is a problem that has received much interest. Several techniques have been proposed such as the classic subtractive type method [1], Kalman filter techniques [2] and subspace based methods [3] to name a few. Recently, in [4] and [5], methods that use a-priori information about speech and noise have been proposed. The a-priori information consists of trained codebooks of speech and noise linear predictive (LP) coefficients. From each pair of speech and noise spectral shapes from the codebooks, a model of the noisy spectrum is created. The optimal speech and noise LP parameters and the respective excitation variances are those that minimize a distortion measure between the corresponding model spectrum and the observed noisy power spectrum. A schematic diagram of this method is shown in figure 1. In [5], the speech and noise spectra and excitation variances were used to construct a Wiener filter to enhance the noisy speech.

An important feature of the a-priori information based method is that it provides an enhancement technique that can be easily integrated into parametric coders that require accurate estimates of the spectrum. In this paper, we provide a framework where the standard vector quantization (VQ) search to obtain the quantized LP parameters is replaced using a maximum likelihood (ML) search to obtain the best entry from the speech codebook. A sufficiently large speech codebook is necessary to provide an acceptable accuracy in the parameter description. A 10-bit speech codebook was used in [5], which was found to be sufficient for obtaining the enhanced waveform using a Wiener filter, but is clearly inadequate for direct quantization. Increasing the codebook size without imposing a structure results in unmanageable computational complexity. Here we use a multi-stage VQ to provide a better description at reduced complexity. The indices resulting from the ML search can directly be used in the coder.

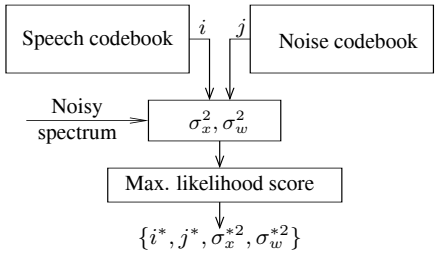

Fig. 1. Estimation of excitation variances and spectral shapes: i^*, j^* are the indices of the selected entries from the speech and noise codebooks and $\sigma_x^{*2}, \sigma_w^{*2}$ are the corresponding excitation variances.

If we look at the method as a 'black-box' enhancement scheme, the emphasis being on producing an enhanced waveform using a-priori information rather than quantization, then it is possible to further improve performance using a simple extension. The improvement arises from the fact that in such a 'black-box' scheme, the speech LP vector is no longer restricted to be an element of the codebook as is the case in direct quantization. A straightforward extension is to interpolate between the ML estimate and its nearest neighbors in the codebook to define a refined search space. The short-term predictor parameters (STP) resulting from the refined search can be used in the enhancement process.

2. MAXIMUM LIKELIHOOD ESTIMATION

In this section, we describe how the ML estimation technique can be applied in the quantization of speech LP parameters under noisy conditions using a multi-stage VQ. This is followed by an extension to speech enhancement using an interpolative search. We first introduce the notation used in the paper.

Assume an additive noise model where speech and noise are

This work was supported by the European Commission under the ANITA project (IST-2001-34327)

independent:

$$y(n) = x(n) + w(n), \quad (1)$$

where $y(n), x(n)$ and $w(n)$ represent the noisy speech, clean speech and noise respectively. For each frame, the noisy spectrum is modelled by a combination of speech and noise spectral shapes from the respective codebooks, together with their excitation variances. The modelled spectrum can be written as

$$\hat{P}_y(\omega) = \frac{\sigma_x^2}{|a_x(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2}, \quad (2)$$

where σ_x^2 and σ_w^2 are the excitation variances of the clean speech and the noise respectively and

$$a_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j\omega k}, \;\; a_w(\omega) = \sum_{k=0}^{q} a_{w_k} e^{-j\omega k},$$

where a_{x_k}, a_{w_k} are the LP coefficients of clean speech and noise with p, q being the respective prediction orders.

2.1. Quantization using multi-stage VQ with ML search

In the absence of background noise, under Gaussianity assumptions, the probability density of the speech samples given the LP parameters can be written as

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x) = \frac{1}{(2\pi)^{N/2}|\mathbf{R}_x|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{R}_x^{-1}\mathbf{x}), \quad (3)$$

where $\mathbf{x} = [x(0)x(1)\ldots x(N\text{-}1)]^T$, $\mathbf{a}_x = [1 a_{x_1} a_{x_2} \ldots a_{x_p}]^T$ and $\mathbf{R}_x = \sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$, where $\mathbf{A}_x$ is the $N \times N$ lower triangular Toeplitz matrix with $[1 a_{x_1} a_{x_2} \ldots a_{x_p} 0 \ldots 0]^T$ as the first column. If we let the frame length approach infinity, then the log-likelihood can be simplified to [6]

$$L = C - \int_0^{2\pi} \big(\log \big(\frac{\sigma_x^2}{|a(\omega)|^2}\big) + \frac{P_x(\omega)|a(\omega)|^2}{\sigma_x^2}\big) d\omega, \quad (4)$$

where $P_x(\omega)$ is the power spectrum of $\mathbf{x}$ and C is a constant. The Itakura-Saito distortion is defined as [7]

$$d_{\text{IS}}(P(\omega), \hat{P}(\omega)) = \frac{1}{2\pi}\int_0^{2\pi} \big(\frac{P(\omega)}{\hat{P}(\omega)} - \log\frac{P(\omega)}{\hat{P}(\omega)} - 1\big)\, d\omega. \quad (5)$$

Comparing (5) and (4) leads to the well-known result that maximizing the log-likelihood is equivalent to minimizing the Itakura-Saito distortion between the observed and modelled spectra. Thus the selected codebook index can be written as $i^* = \arg\max_i p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x^i)$, where $\mathbf{a}_x^i$ is the i^{th} vector from the speech codebook.

When there is background acoustic noise, then in the absence of any speech enhancement system, the index is simply chosen as $\arg\max_i p_{\mathbf{x}}(\mathbf{y}|\mathbf{a}_x^i)$. If an estimate of the noise LP vector $\mathbf{a}_w$ is available, then the ML estimate of the clean speech LP vector given the noisy observation and the estimated noise LP vector can be written as

$$i^* = \arg\max_i p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x^i, \mathbf{a}_w; \sigma_x^2, \sigma_w^2), \quad (6)$$

where the dependence on the excitation variances is explicitly shown. The ML estimate of the speech codebook entry i^* can be equivalently written as

$$i^* = \arg\min_i \big\{ \min_{\sigma_x^2, \sigma_w^2} d_{\text{IS}}(P_y(\omega), \frac{\sigma_x^2}{|a_x^i(\omega)|^2} + \frac{\sigma_w^2}{|a_w(\omega)|^2})\big\}. \quad (7)$$

Assuming small distortion and using a series expansion for $\log(x)$ up to second order terms, it was shown in [5] that the excitation variances minimizing the distortion for a given pair of speech and noise spectral shapes can be obtained as

$$\begin{bmatrix} \| \frac{1}{P_y^2(\omega)|a_x(\omega)|^4} \| & \| \frac{1}{P_y^2(\omega)|a_x(\omega)|^2|a_w(\omega)|^2} \| \\ \| \frac{1}{P_y^2(\omega)|a_x(\omega)|^2|a_w(\omega)|^2} \| & \| \frac{1}{P_y^2(\omega)|a_w(\omega)|^4} \| \end{bmatrix} \begin{bmatrix} \sigma_x^2 \\ \sigma_w^2 \end{bmatrix} = \begin{bmatrix} \| \frac{1}{P_y(\omega)|a_x(\omega)|^2} \| \\ \| \frac{1}{P_y(\omega)|a_w(\omega)|^2} \| \end{bmatrix}, \quad (8)$$

where $\|f(\omega)\| = \int |f(\omega)| d\omega$.

For simplicity, we assume that the noise codebook contains a single entry (with its spectrum denoted by $a_w(\omega)$), which can be obtained using any noise estimation technique such as the minimum statistics method [8] or quantile based estimation [9] for example. For each entry from the speech codebook, the excitation variances are calculated using (8) and the distortion is evaluated. Codebook entries that result in a negative value for either the speech or noise excitation variance are discarded since they are infeasible due to the non-negativity constraints on the variances. The speech spectrum minimizing the distortion (equivalently, maximizing the likelihood) is determined. The selected speech spectrum, $a_w(\omega)$ and the corresponding excitation variances together represent the combination most likely to have generated the observed noisy spectrum.

This approach can easily be extended to multi-stage codebooks using a greedy approach. At each stage, we choose the codebook entry that results in the highest likelihood. A configuration with a two stage speech codebook is shown in figure 2. The ML search using the first stage results in the selection of a single codebook entry as the ML estimate. The second stage codebook forms an additive refinement to this codebook entry, producing a refined codebook. The ML search is repeated with the refined codebook. The two resulting indices (one for each stage) can be transmitted to the decoder. The search can be generalized in a straightforward manner to more than two stages.

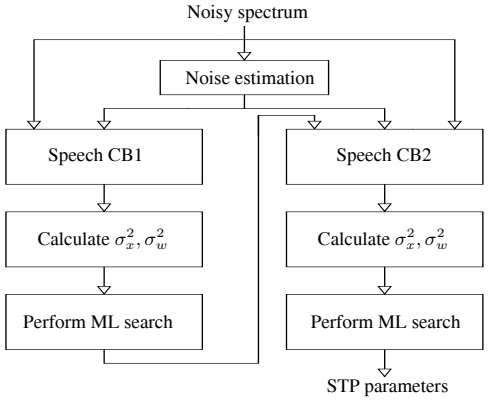

Fig. 2. VQ search under noisy conditions using a two stage speech codebook. The second codebook forms an additive refinement to the speech LP vector resulting from the ML search of the first stage.

2.2. Interpolative search

An important feature of the proposed method using a-priori information is that the estimated clean speech LP vector is produced from a codebook trained with speech data and is hence guaranteed to posses speech-like properties. When used in speech coding, the method has the added advantage that it can easily be integrated into coders since the method outputs a codebook index (indices if multiple stages are used) that can be transmitted to the decoder. If the application is noise reduction alone, performance can be improved if, while retaining the advantage due to a-priori information, we can reduce the errors resulting from the limited precision of the codebooks.

Performing an interpolative search is a natural way to achieve improved performance. Given two centroids from the codebook, we generate a set of points between the centroids and search for the point in the set that maximizes the likelihood. The starting point is the centroid that was selected as the ML estimate. Performing an interpolation along the lines between the ML centroid and each of its neighbors ensures that we reach a point that results in a likelihood not smaller than the codebook-based ML value. Figure

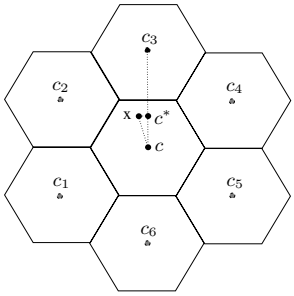

Fig. 3. Interpolative search in two dimensions. The figure shows the Voronoi region containing the ML estimate c and the neighboring Voronoi regions. x is the true data point.

3 depicts the search in two dimensions assuming an ideal vector quantizer. x represents the true data point and c is its codebook-based ML estimate. If we perform an interpolative search along the line joining c and c_3, we obtain a better estimate c^* of x. While this is a simplified example, it is important to note that in our case, the true data point need not be in the Voronoi region containing the ML estimate.

In higher dimensions, the number of neighbors becomes very large and this is the case with a speech codebook. As an approximation, the search can be performed using K nearest neighbors of the codebook-based ML estimate where K is related to the intrinsic dimensionality of the speech data and can be determined empirically. The search is described in table 1.

1. *Obtain index using codebook-based ML search.*
2. *Form interpolative codebook containing NK vectors (using N interpolation steps and K nearest neighbors of the index chosen in step 1).*
3. *Repeat ML search with interpolative codebook to obtain final estimate.*

Table 1. Interpolative search

The estimate from the interpolative search can be used in an enhancement system such as a Wiener filter as in [5]. In a two stage setting, the second stage forms an additive refinement to the result of the interpolative search of the first stage.

Since the interpolation is always between two vectors from the speech codebook, the search in the p-dimensional vector space is constrained to be along the line that corresponds to speech LP vectors. In this respect, an interpolation search is better than an unconstrained gradient descent, which, while possibly resulting in a lower distortion between the modelled and observed noisy spectra, need not necessarily result in an estimate of the speech LP vector that is speech-like. We observe that the interpolative search only guarantees a better likelihood score (i.e. a lower distortion between observed and modelled noisy spectra), which we expect within the assumptions of our model to also result in a better estimate of the clean speech LP vector.

3. EXPERIMENTS

To evaluate the performance of the proposed methods, experiments were conducted using utterances from the TIMIT database. A two stage codebook of dimension 10 with 10 bits in each stage was trained using the generalized Lloyd algorithm with 10 minutes of speech from the TIMIT database using the weighted euclidean distance measure in the line spectral frequency (LSF) domain. The weighting function used was the inverse harmonic mean function [10]. In an actual coder, 25-30 bits would be required for transparent quality, which can be obtained using a generalization of the method presented here, by using a third stage for example. The test set consisted of ten utterances, five male and five female, from the TIMIT database. A frame length of 240 samples with 50% overlap, with a Hann window was used in the codebook training. The LPC orders were 10, 6 and 16 for clean speech, noise and noisy speech respectively and the coefficients were obtained using the autocorrelation method. As in [5], since the model produces an envelope and the variance estimation uses the assumption of small errors, the noisy spectrum P_y is chosen as the spectral envelope instead of the periodogram.

Experiments were conducted for noisy speech at 10 dB input SNR for highway noise (obtained by recording noise on a freeway as perceived by a pedestrian standing at a fixed point), subway noise (obtained by recording noise as perceived by a passenger standing at a fixed point on the platform) and wind noise. As a reference, we use the noise suppression system of the enhanced variable rate codec (EVRC-NS) [11] for comparisons. The codec employs the noise suppression system as a pre-processing module, prior to encoding.

The LP coefficients were extracted frame-by-frame from the output of EVRC-NS and quantized using the two stage LSF codebook. These are the coefficients that would be quantized in the EVRC system. Comparing the resulting log-spectral distortion (SD) [7] values to those resulting from the proposed multi-stage ML search provides a meaningful measure of performance. Table 2 compares the SD values for output of the EVRC-NS quantized with the two stage codebook, SD values for the output of the multi-stage ML search and finally SD values obtained using the interpolative search. It can be seen from table 2 that in some cases, EVRC-NS results in a higher SD compared to that obtained by just quantizing the noisy speech. The proposed ML search results in a lower SD in all cases. As expected, the interpolative search further reduces SD.

AB listening tests were conducted with ten listeners to eval-

Noise Type	Noisy	EVRC-NS	ML	Interpolation
Highway	4.90	4.53	4.49	4.18
Subway	4.89	5.03	4.49	4.19
Wind	4.62	4.81	4.11	3.79

Table 2. Spectral distortion values at 10 dB input SNR using a two stage codebook. The ML search was performed using the same two stage codebook as was used to quantize the noisy speech and the output of EVRC-NS.

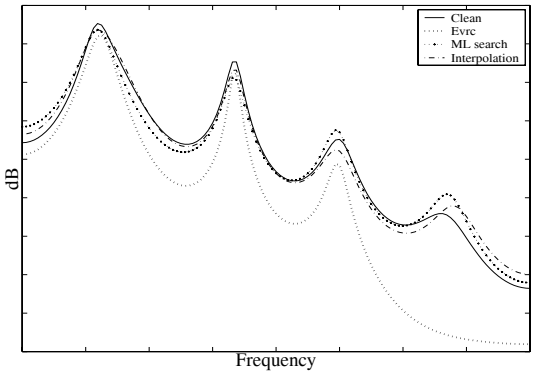

Fig. 4. Comparison of the spectral envelope for the different methods. The figure corresponds to a 30 ms windowed segment.

uate subjective performance. To analyze the effect due to the LP parameters alone, the utterances used in the listening tests were synthesized using the clean residual. One set of utterances was generated by using the quantized values of the LP parameters extracted from the output of the EVRC-NS. The other set was generated using the LP coefficients selected by the two stage ML search (without interpolation). The clean residual was used in both cases. This setup evaluates the performance of EVRC-NS and the ML search in estimating the quantized LP coefficients. The length of the analysis window was 30 ms and the update length was 15 ms. The synthesis frame was further divided into sub-frames of length 5 ms each. Within the sub-frames, LP coefficients were obtained by interpolation. In very low energy speech regions (typically speech pauses), it is possible for the ML search to select random entries from the speech codebook since it assigns a very low excitation variance. When the clean residual is used in the synthesis, the random selection results in musical noise in these low energy regions. By using the quantized noisy LP coefficients in the synthesis whenever the ratio between speech and noise excitation variances fell below a threshold, this problem was avoided. The threshold was experimentally determined to be 0.4.

A second listening test was performed to evaluate the gain due to the interpolative search by generating utterances using LP coefficients obtained from the ML search both with and without interpolation. Again, the clean residual was used in the synthesis to focus on the improvement arising due to the LP coefficients alone. In both the tests, utterances were presented in random order. It can be seen from table 3 that the ML search performs better than EVRC-NS (for use in coding) and that interpolation improves performance (for use in enhancement).

4. CONCLUSIONS

A method to obtain enhanced short-term predictor parameter estimates using a-priori information has been presented. When used in coding of noisy speech, the standard VQ search for obtaining the quantized LP coefficients is replaced by a maximum likelihood search. The method has the potential for being easily integrated into speech coders. Spectral distortion values are lower than when quantizing the output of the EVRC noise suppression system. For applications in noise suppression, the interpolative search provides improved precision and leads to better results as confirmed by listening tests. In this work, we focussed on enhancing the LP parameters. Future work could focus on enhancement of the spectral fine structure in conjunction with the proposed scheme.

	Male	Female
ML search (vs. EVRC-NS)	64	66
Interpolation (vs. ML search)	64	70

Table 3. Preference scores (%) from the listening test. The first row shows the preference for the ML search over EVRC-NS. The second row shows the preference for the interpolation search over ML search.

5. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug 1991.

[3] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[4] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 1, 2001, pp. 669–672.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, September 2003, pp. 1405–1408.

[6] U. Grenander and G. Szego, *Toeplitz forms and their applications*, 2nd ed. New York: Chelsea, 1984.

[7] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 28, no. 4, pp. 367–376, Aug 1980.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.

[9] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 3, 2000, pp. 1875–1878.

[10] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 1991, pp. 641–644.

[11] TIA document, IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, July 1996.