

A NOVEL METHOD FOR WAVELET QUANTIZATION OF NOISY SPEECH

A. S. Madhukumar*, A. B. Premkumar** and H. Abut***

* Centre for Wireless Communications, 20 Science Park Road, #02-34/37, TeleTech Park, Singapore Science Park II, Singapore 117674

** School of Applied Science, Nanyang Technological University, Singapore 639798

*** Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA 92182-1309

Abstract

This paper proposes an architecture for low bit rate coding of noisy speech. The input noisy speech is decomposed into multi-resolution signal components using wavelet transform. An iterative Wiener filtering is used at each level of wavelet analysis to enhance speech. The system model that evolves during enhancement is processed further to get optimal parameters for the quantization. A multistage vector quantizer is used for compression of decomposed speech. The enhanced speech is reconstructed at the receiving end by a VQ decoder and the necessary wavelet reconstruction network. Speech coding rate for the proposed architecture is estimated to be about 2.37Kbps.

1. INTRODUCTION

To reduce the performance degradation of speech processing systems operating in noisy environment, it is a well-known practice to include an enhancement system as a preprocessing unit. There are quite a few successful noise reduction techniques used in speech coding and recognition [1-3]. The primary objective of these techniques is to suppress the perceivable background noise without affecting the signal quality. It is highly desirable if such an algorithm can reduce complexities in subsequent stages as well.

Speech enhancement is usually considered as a preprocessing stage in speech

coding and recognition applications operating in noisy environments. Even though many of speech coding systems in literature use similar kind of processing algorithms for noise removal and coding, they are not integrated together with a common processing structure for formal parameter extraction in the coding stage due to the complexities involved. Advancements in wavelet transforms and its applications in denoising and subband coding can solve this problem to a great extent. In this paper we propose an architecture, which integrates speech enhancement and coding using principles of wavelet analysis and vector quantization.

This paper is organized as follows: Section 2 discusses the procedure for enhancement of noisy speech, which includes decomposition in wavelet domain, constrained Wiener filtering algorithm for denoising speech, and extraction of parameters for subsequent coding. Section 3 discusses a variable rate multistage vector quantizer for coding different levels of wavelet analysis. Performance evaluation of the proposed method is presented in section 4.

2. ENHANCEMENT OF NOISY SPEECH

The first processing step in our approach is to decompose incoming speech in the wavelet domain. This is followed by a constrained Wiener filtering algorithm to enhance noisy speech and to generate appropriate parameters for subsequent VQ and bit assignment tasks.

Following subsections discuss the above procedures in detail.

2.1. Wavelet Decomposition of Noisy Speech

Wavelet transforms decompose signals in terms of their wavelet coefficients onto shifted and dilated versions of a prototype bandpass wavelet function [4]. The discrete, dyadic, and orthogonal wavelet transform of a causal, finite energy signal sequence, $x(k)$, is given by the ensemble of projections, $X_{n,m}$, of $x(k)$ over an orthogonal and complete set of bandpass sequences, $\psi_{n,m}(k) = \psi_{n,0}(2^n k - m)$. Here $n = 1, 2, 3, \dots$ represents scale and $m = 0, 1, 2, \dots$ is associated with time-shift. The Wavelet Transform (WT) of $x(k)$ is defined by:

$$x(k) = \sum_{n=1}^N \sum_m X_{n,m} \psi_{n,m}(k) \quad (1)$$

where

$$X_{n,m} = \sum_k x(k) \psi_{n,m}(k) \quad (2)$$

The block diagram for the proposed wavelet decomposition is shown in Figure 1. The outputs at HPF and LPF are labeled as DWT coefficients (D1, D2, and D3) and approximations (A3), respectively. This architecture has a number of inherent properties for data compression. Separation of signal into a number of channels restricts the quantizing noise in a given channel to that subband and permits different bit assignment budgets for each channel.

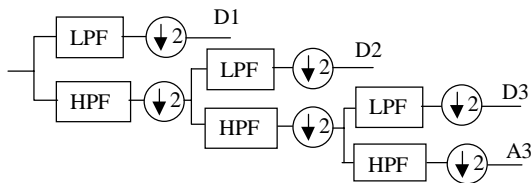


Figure 1. Block diagram for proposed wavelet architecture

2.2. ARMA Wiener filtering for Denoising:

Wiener filter models speech signal as a response of an all-pole network and it processes with respect to a posteriori probability of the signal from the noisy signal. Its performance depends on a number of factors, such as, coherency of signal time imprint, degree of spectral overlap between signal and noise components, the SNR content, and the match between time-frequency structure of the signal and its decomposition.

It is also noted that a posteriori probability based Wiener filtering always assumes the signal is stationary. This, in turn, makes its application to noisy speech fairly difficult. The proposed Wiener filtering algorithm overcomes this constraint by decomposing the speech signal into different time-frequency planes, thereby, exploiting the properties of discrete wavelet transforms. However, this necessitates inclusion of a multirate filter network that matches with frequency ranges at each level. The enhanced wavelet coefficients at each level, however, can be retrieved from the filter structure using the principle of minimum mean square estimation (MMSE).

In Figure 2, we depict the block diagram for an iterative Wiener filter structure. Simplified algorithms for Wiener filtering have been proposed in literature for sequential estimation of all-pole parameters and gain from the estimated signal. To the best of our knowledge, the representation of noisy speech as

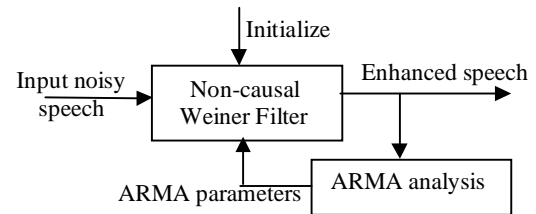


Figure 2. Block diagram of ARMA Wiener filtering for the speech enhancement

an Autoregressive (AR) process has not been successful. In fact, it has been reported that addition of white noise to speech makes it an Autoregressive Moving Average (ARMA) class signal. Also, a model containing zeros and poles rather than an all-pole model can better represent several speech sounds (e.g., nasals). For this reason, we think ARMA modeling is a better alternative for representing noisy speech.

3. VARIABLE RATE VECTOR QUANTIZER

In our proposed architecture we employ a multistage VQ approach in which Line Spectral Pair (LSP) parameters are quantized at the first stage and the quantization errors are encoded in subsequent stages to exploit the coarse and fine structures in a sequential manner. Multistage VQ systems are normally assumed to have identical cardinality and hence the same bit rate. However, in the proposed architecture, we have relaxed this assumption to use a scheme that matches with the concept of variable rates used for discrete wavelet transforms and subsequent Wiener filtering in the speech enhancement stage. A structured variable rate vector quantization algorithm has been designed to take care of wavelet coefficients at different locations of time and frequency. This has effectively reduced the bit rate for coding and at the same time, has preserved the perceptual aspects of speech.

In Figure 3 we present the structure of our vector quantizer at each level of the wavelet transform process. The input to the quantizer is a set of AR parameters obtained from the constrained Wiener filtering. These are mapped onto LSP frequencies and the codewords are generated in the usual nearest-neighbor sense. After obtaining the nearest-neighbor codeword, the distortion is computed and the search proceeds to subsequent stages for minimization of quantization errors. In our case, we have used scalar quantizers for encoding source parameters, including pitch, gain, voicing decision and the residual information. As in all analysis-by-synthesis techniques, a synthesis loop is also incorporated at the encoder side to complete the structure.

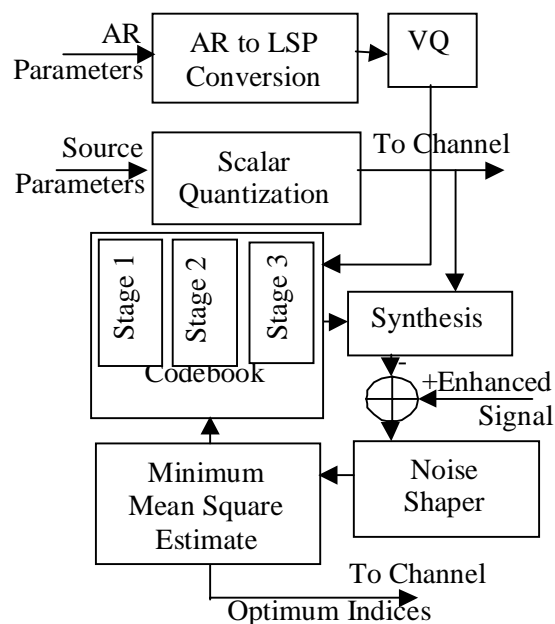


Figure 3. Block diagram for basic structure of vector quantizer at each level of wavelet transform

4. THE SYSTEM ARCHITECTURE

The noisy speech is decomposed into three levels of DWT coefficients and one approximation signal. Both constrained Wiener filtering and VQ are applied at each level individually. The optimum indices are transmitted to receiver side together with source parameters consisting of pitch, gain, voicing decision and residual information for each band. These parameters are extracted from the second level of approximation signal. On the receiver side, the DWT coefficients and approximation are reconstructed by a table look-up to the reproduction codebooks. A three level inverse DWT analysis is used for reconstructing the enhanced output speech.

As our test data, we have used 8 utterances from various male and female speakers. These utterances are sampled at a rate of 8KHz, which are then segmented into non-overlapping frames of 800 samples. The computation of data rate for the proposed architecture is closely related to the principles of

DWT. At each level of wavelet analysis, the number of data output is reduced by a factor of 2. With this in mind, three level wavelet analysis is done on each frame to yield 400 samples at the first level output (D1 in Figure 1), 200 samples in the second level (D2) and 100 samples in the third level (D3 and A3). It is worth noting that the samples at D1 are subdivided into 2 frames of size 200 samples, whereas samples at other levels are treated as single frames. Consequently, the total number of frames is five, i.e., two in D1 and one each in D2, D3 and A3.

In our four-stage VQ codebook, we have used 128, 64, 64, 64 code words, respectively. Source parameters including residual information are computed from the second level approximation by subdividing into four frames of equal size and then the values are extracted. The pitch, gain and residual information are scalar quantized using eight bits and voicing decision is quantized into four bits. Hence, the total number of bits to represent a block of 800 samples are 237, i.e., $5 \times 25 + 4(8+8+8+4)$, which amounts to an overall bit rate of 2.37 KBPS.

5. PERFORMANCE EVALUATION OF THE SYSTEM

We have used white Gaussian noise (AWGN) of different levels to mix with clean speech to generate noisy speech. This noisy speech is then enhanced, quantized and encoded using the proposed method. The quality of output speech is compared with the input noisy speech in terms of frame-wise segmental SNR (SegSNR). Since the Wiener filter algorithm is a minimum mean square estimation technique, the segmental mean square error between the coded speech and clean speech can be considered as a good measure for performance of this enhancement algorithm. The SegSNR is computed for both white noise and bandlimited colored noise of levels -5 dB, 0 dB, 5 dB, 10 dB and 20 dB. The experiments are carried out for various size codebooks in the proposed multistage VQ. The results are given in Table 1. As expected, the segSNR improves from 2 to 4dB when the number of stages is increased. It is also noted that segSNR is improved

significantly between the input noisy speech and the clean speech. The average segmental SNR of the enhanced speech decreases with the noise level.

No of Stages	SegSNR					Clean speech
	-5dB	0dB	5dB	10dB	20dB	
1	-6.47	-1.26	4.26	6.72	8.01	10.02
2	-4.48	-1.67	4.85	7.23	8.57	10.56
3	-3.38	-0.12	5.58	7.69	8.46	11.02
4	-2.15	1.12	6.65	8.12	9.24	11.47

Table 1. Segmental SNR of the output speech subject to different level of white noise

7. CONCLUSIONS

Architecture for vector quantization of noisy speech is proposed based on principles of discrete wavelet transforms. An iterative Wiener filter like estimation technique based on ARMA modeling is employed on the decomposed speech for enhancement. The performance of the proposed model is tested and found satisfactory for speech signal mixed with different levels of white noise. Since the enhancement and coding use the same set of assumptions, the design of our overall system is less complex and a number of tradeoff scenarios can be developed easily among quality, intelligibility and bit rate.

4. REFERENCES

1. J. S. Lim, Ed., *Speech Enhancement*: Prentice-Hall, New Jersey, 1983
2. J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp 795-805, Apr. 1991
3. Y. Ephraim, "Statistical model based speech enhancement systems," *Proc., IEEE*, vol. 80, pp 1526-1555, Oct. 1992
4. I. Daubechies, "Orthonormal bases for compactly supported wavelets," *Commun. Pure Appl. Math.* vol. 41, pp. 909-996, Nov. 1988.