WIDEBAND CODING OF SPEECH USING NEURAL NETWORK GAIN ADAPTATION

Cheung-Fat Chan and *Man-Tak Chu*

Department of Electronic Engineering
City University of Hong Kong
83, Tat Chee Avenue, Kowloon, HONG KONG
Email : eecfchan@cityu.edu.hk

ABSTRACT

In this paper, a high-quality wideband speech coder is proposed. The coding structure resembles a LD-CELP coder, however, several novel improvements are made. The gain adapter for the stochastic codebook is driven by a neural network and it updates the excitation gain in a sample-by-sample fashion. The purpose of incorporating a neural network is to exploit both the intra- and inter-frame correlation of speech signal in a non-linear manner. A psychoacoustic model instead of a simple perceptual weighting filter is used to shape the quantization noise. Simulation result shows that the proposed coder can achieve transparent coding of wideband speech at 16 kbps.

1. INTRODUCTION

Low delay code-excited linear predictive (LD-CELP) coding has been demonstrated capable of coding narrowband (3.4 kHz bandwidth) speech at 16 kbps[1]. Recently, LD-CELP coder has been extended to code wideband (7 kHz bandwidth) speech signals at 24 kbps[2]. In conventional LD-CELP coding, the excitation gain is backward-adapted in a block-by-block fashion. Obviously, block adaptation can not cope with time-varying signals during rapid transitions since the excitation gain is fixed for the entire analysis block. Moreover, in order to increase the dynamic range, the gain is usually adapted in logarithmic domain by using an ad hoc procedure[1]. In this paper, a novel LD-CELP coder for transparent coding of wideband speech and audio signals at 16 kbps is proposed. Several improvements on conventional LD-CELP coding are made in order to achieve this goal. First, a noise shaping filter based on psychoacoustic model is developed to increase the coding efficiency. Second, a sample-by-sample gain adapter which is driven by a neural network is developed to code rapid time-varying signals with a wider dynamic range. Finally, improvements are also obtained by using a trained stochastic codebook.

2. THE PROPOSED WIDEBAND SPEECH CODER

The block diagram of the proposed coder is illustrated in Fig. 1. This coder composes of four main components; a stochastic excitation codebook, a neural network driven gain adapter, a backward adaptive prediction filter and a noise shaping filter based on psychoacoustic model. The structure of the proposed coder is the same as conventional LD-CELP coder excepted that, the excitation gain is adapted in a sample-by-sample fashion by using a neural network, and a noise shaping filter based on psychoacoustic model is applied during the analysis. In this coder, short-term correlation in the input signals is removed by a 20^{th}-order adaptive predictor. The coefficients of the predictor are derived from linear predictive analysis on the synthetic signals. A 16ms recursive window is used for deriving the autocorrelation for analysis[3]. The optimum codeword from excitation codebook is determined such that the weighted mean square error between the input and synthetic signals is minimized.

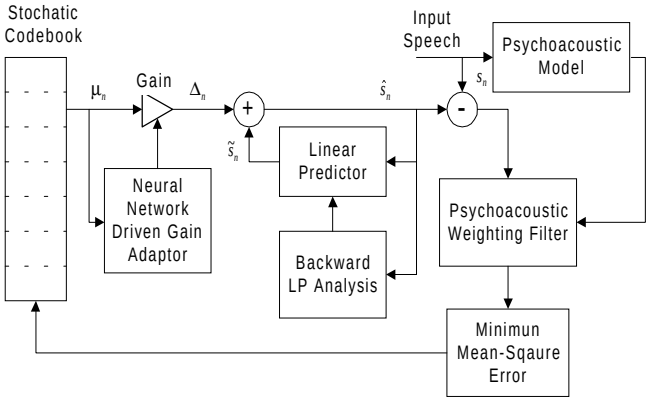

Fig. 1 Block Diagram of the Proposed Coder

2.1 Noise Shaping Filter Using Psychoacoustic Model

The perceptual weighting filter utilized in the conventional CELP coder is normally derived from linear predictive analysis of the input signals[1]. In practice, the performance of this weighting filter is far from satisfactory if it is compared to frequency-domain noise shaping using psychoacoustic model of human auditory system. This is particularly true if the coder is used for coding audio signals instead of speech signals. For frequency-domain audio coders such as MPEG's MDCT coder, psychoacoustic noise shaping can be applied in a straight-forward manner. In this paper, a psychoacoustic noise shaping filter suitable for time-domain coders such as LD-

CELP is proposed. Fig. 2 illustrates the procedures for obtaining the shaping filter's coefficients. In this method, the windowed input speech is first transformed to frequency-domain via a 256-point FFT. Then, the power spectrum of the noise masking curve is derived from a psychoacoustic model. The power spectrum is then converted to autocorrelation via IFFT and a 20th-order all-pole model is fitted to the noise masking curve and the model coefficients are derived from the autocorrelation via Levinson algorithm. Note that in order not to increase the delay, the model coefficients are derived from the past input signal with an exponential window. In this work, the psychoacoustic model used is a standard model based on noise masking threshold[4].

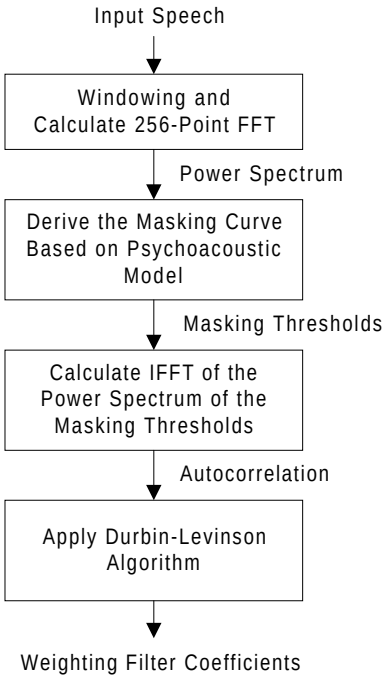

Fig. 2. Estimation of Noise Shaping Filter's Coefficients

2.2 Neural Network Driven Gain Adaptation

In conventional LD-CELP coders, the excitation gain is adapted in a block-by-block fashion by applying linear prediction on the logarithmic of previous gain values. The reason to apply linear prediction in logarithmic domain is to extend the dynamic range of the excitation gain so as to make the prediction more effective. In this paper, a neural network driven gain adapter is proposed. Fig. 3 shows the structure of the neural network gain adapter. This adapter is designed to update a gain vector of dimension 6 in each time sample. The network has two layers, an input layer which contains 11 neurons and an output layer which contains 6 neurons.

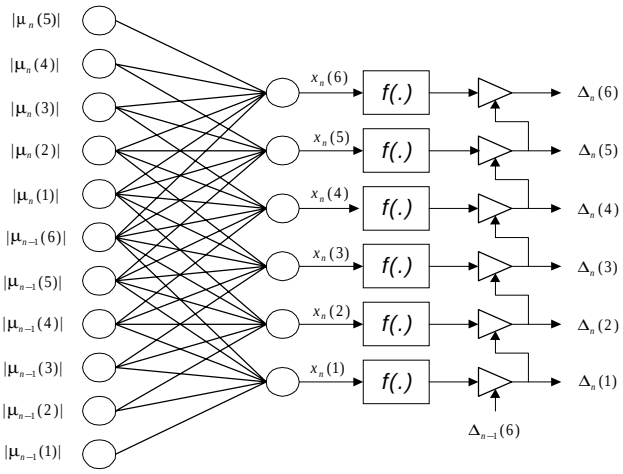

Fig. 3 Neural Network Driven Gain Adaptor

This neural network is a single layer perceptron. The input layer of the network consists of the magnitudes of 6 previous stochastic excitation elements. Each output neuron produces an adaptation factor which is then used to determine the current gain value by multiplying it with the previous gain value. The neurons between the two layers are partially interconnected via the weighting coefficients $\{w_{i,j}\}$. Only 6 consecutive magnitudes are actually connected to a single output neuron. The output from neuron *i* in the n^{th} time frame is calculated as (see Fig. 3),

$$x_n(i) = \sum_j w_{i,j} y_i^{(n)}(j) \qquad (1)$$

where $y_i^{(n)}(j)$ is the magnitude of the previous excitation signal which may contain both elements from the current and previous excitation vectors. The output of the neural network is obtained via an activation function $f(x)$. Generally, $f(x)$ is a sigmoid function, e.g. $f(x) = 1/(1+e^{-x})$, which provides the non-linearity to the neural network. In this work, the activation function is chosen as

$$f(x) = \begin{cases} 1.25 & x \geq 1.0 \\ 1.0 + 0.25x & 1.0 > x \geq 0.0 \\ 1.0 + 0.05x & 0.0 > x > -1.0 \\ 0.95 & x \leq -1.0 \end{cases} \qquad (2)$$

The clipping of $f(x)$ for $|x| \geq 1.0$ is to limit the signal magnitude to an allowable range. The reason for choosing this activation function is that, practically, the energy of real-world signal tends to increase more rapid than when the signal is decreasing, e.g., during a speech onset. The current gain value is determined as the product of the output of the neural network and the previous gain value,

$$\Delta_n(i) = f(x_n(i))\Delta_n(i-1) \qquad (3)$$

The advantages of using a neural network to control gain adaptation are that, non-linearity is inherently built-in and also, by operating the neural network directly on the stochastic excitation signal, both intra- and inter-frame

correlation can be exploited in a sample-by-sample manner.

2.2.1 Training the Neural Network

It is necessary to train the neural network to achieve good performance. The weighting coefficients $\{w_{i,j}\}$ should be determined according to a meaningful optimization criteria. The criteria used here is the mean squared error between the input signal and the reproduction signal. For an input signal vector $\mathbf{s_n}$, the corresponding reproduction signal vector $\hat{\mathbf{s}}_\mathbf{n}$ is obtained by the synthesis equation as,

$$\hat{s}_n(j) = \tilde{s}_n(j) + \mu_n(j)\Delta_n(j) \qquad 1 \le j \le N \tag{4}$$

where $\tilde{s}_n(j)$, $\mu_n(j)$ and $\Delta_n(j)$ are the predicted speech, the selected excitation signal and the corresponding gain, respectively. N is the dimension of the codeword which is equal to 6. The mean squared error is then calculated as,

$$E = \sum_n \sum_{k=1}^{N} [s_n(k) - \hat{s}_n(k)]^2 \tag{5}$$

Then, from (1), (2), (3), (4) and (5) to achieve

$$E = \sum_n \sum_{k=1}^{N} [s_n(k) - \tilde{s}_n(k) - \mu_n(k)\Delta_n(k-1) f(x_n(k))]^2 \tag{6}$$

By differentiating the error term in (6) with respect to the weighting coefficients to yield,

$$\delta_{i,j} \equiv \frac{\partial E}{\partial w_{i,j}} = -2\sum_n [s_n(i) - \hat{s}_n(i)]\mu_n(i)\Delta_n(i-1)\frac{\partial f(x_n(i))}{\partial x_n(i)} \cdot \frac{\partial x_n(i)}{\partial w_{i,j}}$$

where both the basic and activation functions are differentiable and their derivatives are;

$$\frac{\partial x_n(i)}{\partial w_{i,j}} = y_i^{(n)}(j)$$

and $$\frac{\partial f(x)}{\partial x} = \begin{cases} 0.0 & x \ge 1.0 \\ 0.25 & 1.0 > x \ge 0.0 \\ 0.05 & 0.0 > x > -1.0 \\ 0.0 & x \le -1.0 \end{cases}$$

By using the gradient descent algorithm with the computed gradient $\delta_{i,j}$, the weighting coefficients can be iteratively optimized as

$$w_{i,j}^{(t+1)} = w_{i,j}^{(t)} + \eta\delta_{i,j}$$

where η is a factor which controls the learning rate. In practice, the learning factor is set to 0.5 initially and set to 0.1 after a few iterations. It was experimentally found that the initial weighting coefficients can be set to $w_{i,j} = 1$ for $i = j$ and $w_{i,j} = 0$ for $i \ne j$.

2.3 Excitation Codebook Training

In order to further improve the performance of the coder, a training procedure is developed to optimize the excitation codebook. Suppose that ζ_i is the coding distortion for mapping input signals to the quantization cluster C_i for codeword *i* in the codebook, then

$$\zeta_i = \sum_{n \in C_i} \sum_{k=1}^{N} [s_n(k) - \hat{s}_n(k)]^2 \tag{7}$$

The total distortion ζ for quantizing a set of training signals is calculated as,

$$\zeta = \sum_i \zeta_i = \sum_i \sum_{n \in C_i} \sum_{k=1}^{N} [s_n(k) - \tilde{s}_n(k) - \mu_i(k)\Delta_n(k)]^2 \tag{8}$$

The overall distortion is minimized with respect to the codeword element $\mu_i(j)$. By setting $\frac{\partial \zeta}{\partial \mu_i(j)} = 0$ to yield,

$$\bar{\mu}_i(j) = \frac{\sum_{n \in C_i} [s_n(j) - \tilde{s}_n(j)]\Delta_n(j)}{\sum_{n \in C_i} [\Delta_n(j)]^2} \tag{9}$$

With this re-estimation equation, the training algorithm can be implemented as follows:

1. Given an initial codebook $\Psi = \{\mu_i^{(0)}(j)\}$,
2. encode the training signals by the coder. In each quantization cluster $C_i^{(m)}$, sets of signal samples $s_n(j)$, predicted samples $\tilde{s}_n(j)$, and gains $\Delta_n(j)$, that mapped to $C_i^{(m)}$ are obtained.
3. By using these sets of signals and gains, the codeword for each quantization cluster can be re-adjusted as $\mu_i^{(m+1)}(j) = \lambda\mu_i^{(m)}(j) + (1-\lambda)\bar{\mu}_i^{(m)}(j)$, where λ is a factor for controlling the training rate.
4. If the improvement in coding distortion is small, then stop, otherwise go to step 2.

For codebook initialization, it was found that codewords distributed uniformly in the codeword space is sufficient. This codebook training has significantly improve the SNR performance of the coder by at least 5 dB. Fig. 4 shows the SNR increases as the number of iteration during the training.

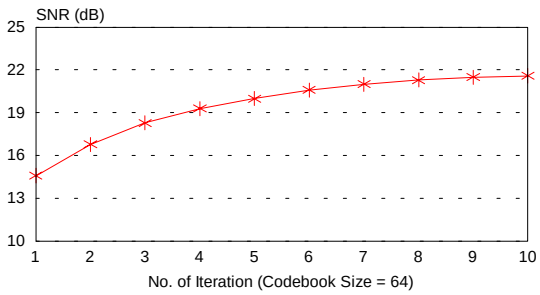

Fig. 4 SNR Improvements During Codebook Training

3. SIMULATION

The proposed coder is designed to operate at 1 bit/sample at 16 kHz sampling rate. Since the excitation gain and the short-term predictor coefficients are backward adapted, the only information sent to the decoder is the index of the optimum excitation codeword. The dimension of the codeword is 6. The excitation codebook contains 64 codewords. The codebook is searched by using an analysis-by-synthesis procedure as similar to other CELP coders. In addition, ML-search algorithm with M=4 and L=4 is applied to increase the coding gain[5]. This causes a delay of 24 samples or 1.5 ms at 16 kHz sampling rate. The order of the short-term predictor is 20 and the predictor coefficients are updated in every 48 samples. The order of the noise shaping filter is also 20 and the filter coefficients are updated in every 144 samples. Several speech sentences recorded from AM broadcast are used to test the performance of the proposed coder. The speech sentences were contributed by male and female speakers. All signals are digitized at 16 kHz with 16 bits resolution. In order to compare the coding performance, a conventional 24 kbps LD-CELP coder[2] and an ITU G.722 56 kbps subband ADPCM coder[6] were also implemented. The objective criteria for evaluating the proposed coder, the LD-CELP coder and the SB-ADPCM coder are based on average signal-to-noise ratio (SNR) and noise-to-mask ratio (NMR). NMR is defined as the ratio of the noise density to the masking threshold density for a critical band of a psychoacoustics model[7]. We also compute the percentage of the number of the masked frames to the total number of the speech frames (PMF) in the test signal. Note that a masked frame is declared if a frame containing the noise energies which are below the masking threshold of all critical bands in the spectrum. Table I listed the results achieved for the 3 coders under tested.

	Proposed Coder 16 kbps	LD-CELP 24 kbps	SB-ADPCM 56 kbps
SNR(dB)	20.63	19.8	21.2
NMR(dB) (PMF)	-6.29 (19.8 %)	-0.2 (0 %)	-1.03 (0 %)

Table I Objective Performances

Simulation results in these tests indicate that the proposed coder operating at 16 kbps outperforms conventional 24 kbps LD-CELP coder both in SNR and NMR. The quality of the proposed coder is judged to be perceptually better that the 56 kbps SB-ADPCM coder eventhough the SNR performance is lower than that of SB-ADPCM. Informal listening tests confirm that the proposed coder can achieve near transparent coding of wideband speech at 16 kbps. Currently, the proposed coder complexity is very high and future research will be carried out to reduce the complexity and increase the robustness of the coder against channel errors.

4. CONCLUSION

A novel wideband speech coder is presented in this paper. The proposed coder incorporates a noise shaping filter based on psychoacoustic model and utilizes a neural network driven sample-by-sample adaptation scheme for the excitation gain. The stochastic codebook is generated by a training procedure. Simulation results show that the proposed coder can achieve transparent quality at 16 kbps.

5. REFERENCES

[1] Juin-Hwey Chen, "A Robust Low-Delay CELP Speech Coder at 16 kb/s", Advances in Speech Coding, pp.25-35, 1990.

[2] J. Paulus, C. Antweiler and C. Gerlach, “High Quality Coding of Wideband Speech at 24 kbit/s,” EUROSPEECH-93, pp.1107-1110, 1993.

[3] Thomas P. Barnwell, "Recursive Windowing for Generating Autocorrelation Coefficients for LPC Analysis", IEEE Trans. on ASSP, VOL. ASSP-29, NO. 5, pp. 1062-1066, October 1981.

[4] M. R. Schroeder, B. S. Atal, J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear", Journal of Acoustical Society of America, Vol 66, No. 6, Dec. 1979.

[5] N.S. Jayant, Peter Noll, "Digital Coding of Waveforms : Principles and Applications to Speech and Video", Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[6] CCITT Recommendation, G.722.

[7] K. Brandenbury and T Sporer, “NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria,” AES 11th Int. Conference, pp.169-179.