A LOW-DELAY WIDEBAND SPEECH CODER AT 24 KBPS

Anil Ubale and Allen Gersho

Dept. of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA

ABSTRACT

A novel low-delay wideband speech coder, called Multi-band CELP (MB-CELP) achieves a delay of about 10 ms, by exploiting time-domain correlations with a two-stage linear prediction scheme. A low-order forward-adaptive LP stage models coarse shape, and a high-order backward-adaptive LP stage models fine structure of the input spectrum. A conditional pitch prediction method improves the performance of the coder for speech without degrading music performance. A multi-band bank of off-line filtered codebooks generates the excitation signal. A 24 kbps version of the coder has nine multi-band codebooks with nonuniform bandwidth. Subjective comparison tests show that this coder outperforms the G.722 coder at the bit-rate of 56 kbps.

1. INTRODUCTION

With recent advances in speech coding bit rates as low as 8 kbit/s are achievable while preserving audio quality, as demonstrated by the ITU-T Recommendation G.729. However, the input is limited to the narrow-band telephone bandwidth of 300-3400 Hz. Further improvement in audio quality (at the price of an increased bit-rate) is obtainable by widening the bandwidth to roughly 50 to 7000 Hz using a sampling rate of 16 kHz. The extended upper band limit gives crisper more intelligible speech, while the extended lower band limit gives more natural sounding speech. In 1988, the CCITT (now known as the ITU-T) established an international coding standard for high-quality 7 kHz audio, known as G.722 [1].

The ITU-T is pursuing a new standard for wideband coding of speech and music and recently specified terms of reference which include a maximum algorithmic delay requirement of 40 ms and a maximum frame size of 20 ms. However, the algorithmic delay objective is 20 ms or less, with a maximum frame size of 10 ms.

The performance requirements for the standard, are to achieve at 24 kbps a quality equivalent to that of G.722 at 56 kbps, and at 16 kbps to obtain a quality equivalent to that of G.722 at 48 kbps for clean speech and music inputs.

In this paper, we describe a 24 kbps Mutli-band CELP coder which meets the quality requirements and also achieves the delay objective. We are currently working on 16 kbps bit-rate version of this coder, and if this work is successful, the MB-CELP coder might be a worthy candidate for submission to the ITU-T.

2. BACKGROUND

Contemporary approaches to wideband speech coding are typically either CELP [2] or perceptual transform coders [3]. Perceptual transform coders exploit the frequency masking phenomenon [4] in the human hearing process. These coders achieve compression by the use of high-resolution frequency-domain representation of the signal, adaptive bit allocation to transform coefficients based on psychoacoustic model and entropy coding. In order to obtain a high-resolution frequency-domain decomposition of the signal, these coders require large number of samples per frame, and hence high algorithmic delay (usually $\geq$ 32 ms).

The CELP coding paradigm remains the most effective speech coding method for narrowband speech e.g. G.729. Compared to CELP coders, the transform coders generally perform poorly for speech at bit rates lower than 2 bits/sample. An important factor in the success of CELP coders for speech is the use of pitch prediction. On the other hand, conventional CELP coders perform very poorly for music. This is due to the inadequacy of pitch prediction in modeling the fine structure in the music spectrum.

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACC, ACT Networks, Advanced Computer Communications, Cisco Systems, DSP Group, DSP Software Engineering, Fujitsu, General Electric Company, Hughes Electronics, Intel, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments.

3. INTRODUCTION TO MULTI-BAND CELP

As mentioned above, two methods of wideband speech coding, namely transform and CELP coding are well suited for music and speech respectively. However, none of these methods provide acceptable performance for both music and speech.

In our pursuit of a universal wideband coder (for both speech and music), we decided to explore the CELP coding technique. The main reason for this choice is to meet the low-delay objective.

In this paper, we describe our novel encoding scheme, MB-CELP, and show that it achieves high-quality wideband speech and music compression at low bit-rate while maintaining low delay.

To achieve a delay of about 10 ms, the MB-CELP wideband speech coding algorithm as shown in Figure 1, exploits time-domain correlations using linear prediction, incorporates quantization noise shaping, and uses off-line filtered stochastic multi-band excitation codebooks. The multi-band codebook sizes can be dynamically tailored in accordance with the perceptual importance of the frequency bands.

As shown in the Figure 1, we use a two-stage time-domain linear prediction. One stage operates in forward-adaptive mode and exploits the time-domain correlations due to overall coarse spectral shape of the input spectrum. This filter is also useful for formant prediction for the speech input. The other stage is intended to model the fine structure in music spectrum. It operates in a backward-adaptive mode and relies on the slowly changing fine structure of the music spectrum. The backward-adaptive time-domain prediction in the second-stage achieves low bit-rate. Further, the backward- adaptive prediction gain is high for low frame size. This makes the two-stage forward-backward prediction an effective scheme for low-delay wideband music coding.

However, backward-adaptive linear prediction filter is not sufficient to encode the pitch structure of speech, even with a high order of 100. In order, to improve the performance for speech, we also include a pitch prediction filter. The pitch predictor is implemented as an adaptive codebook, and is disabled when it does not provide enough coding gain to justify its use.

To efficiently represent the excitation signal we use mutli-band excitation codebooks.

In the following sections, we will describe a specific MB-CELP coder configuration which operates at the rate of 24 kbps. The frame size is 10 ms, and the subframe size is 5 ms. The look-ahead for the LP analysis is 3.75 ms, yielding an algorithmic delay of 13.75 ms. The forward-adaptive LP filter order is 16, and backward-adaptive LP filter order is 100. The perceptual weighting filter has order 30 and the form $A(z/\gamma_1)/A(z/\gamma_2)$.

4. FORWARD-ADAPTIVE LP-ANALYSIS AND CODING

Forward-adaptive linear prediction (LP) analysis is performed once per input frame using the autocorrelation method with a 12.5 ms Hamming window. The autocorrelation values of the windowed speech are computed and a bandwidth expansion of 10 Hz is introduced by windowing the autocorrelations. The LP coefficients are converted to line spectral frequencies (LSFs) and quantized using 21 bits for each frame. The LSFs are quantized using switched-predictive multi-stage vector quantization [5]. The switch between two predictors is quantized using one bit. We use second-order autoregressive interframe predictors. The prediction residual is coded using multi-stage vector quantization (MSVQ) with 5 stages of 4 bits each. In order to provide good performance for both speech and music, the two predictors and corresponding prediction error codebooks are designed using both speech and music training data.

The distortion measure employed for predictive multi-stage vector quantization is a weighted mean-square-error (WMSE). The weights are proportional to the distance between the neighboring LSFs.

The multi-stage vector quantization scheme uses a multiple-survivor method for an effective trade-off between complexity and performance. Six residual survivors are retained from each stage and are tested by the next stage. The final quantization decision is made at the last stage, and a backward search is conducted to determine the entries in all stages. The multi-stage vector quantizer is designed by a joint optimization procedure [6].

The LP coefficients are computed once per frame with the LP analysis window situated at the center of the second subframe. These LP coefficients are directly applied to the second subframe of each frame. For the first subframe, the quantized LP coefficients are obtained by a linear interpolation of the corresponding parameters in the adjacent subframes.

5. BACKWARD-ADAPTIVE LP ANALYSIS

Backward-adaptive LP analysis can be done by processing the past excitation signal which is applied to the forward-adaptive LP synthesis filter. However, since the forward-adaptive LP synthesis filter changes ev-

ery subframe, better performance is achieved if the backward-adaptive LP analysis is computed from the residual that is obtained by forward-adaptive LP filtering the previously decoded speech [7], as shown in Figure 1.

6. PITCH ANALYSIS AND CODING

Pitch prediction is implemented using the adaptive codebook method where pitch delays greater than the subframe length are searched in the range of (81-336) samples. Fractional pitch delays with nonuniform spacing are quantized using 9 bits per subframe. The highest resolution for pitch delay is equal to 1/4 of a sample. For the selected pitch delay, the pitch gain is scalar quantized using 4 bits.

Although, pitch prediction is very useful for speech, it is not suitable for music. The fine structure in music is complex, and can not be modeled using a single pitch synthesis filter. Therefore, we use pitch prediction only when it is useful. The criterion for enabling or disabling pitch prediction is the closed-loop adaptive codebook gain. If the adaptive-codebook contribution to the target vector (after removing the zero input response from weighted speech) provides a coding gain of greater than a threshold of 2 dB, the pitch prediction is used.

Figure 2 shows pitch prediction flag (0 : No pitch prediction used, 1 : pitch prediction used) for typical speech and music samples.

The pitch prediction flag is transmitted to the decoder with 1 bit per subframe.

7. MULTI-BAND EXCITATION

The fixed (non-adaptive) codebook excitation is generated from nine filtered codebooks with nonuniform bandwidths. It must be noted that since the filtered codevectors are of finite duration (equal to the subframe length) there is spectral leakage between the adjacent codebooks. However, this is not a critical issue, since the error is minimized over the fullband speech.

Previously [8], we used multi-band excitation codebooks which were bandlimited to uniform bandwidths, and we exploited the perceptual properties of the human ear by using adaptive codebook size allocation for the multi-band codebooks. The adaptive codebook size allocation, was derived from the quantized forward-adaptive LP spectrum, and hence no side information needs to be transmitted. Alternatively, we can choose nonuniform bandwidths for the multi-band excitation codebooks. We experimented with different number of codebooks with different bandwidths, and arrived with nine multi-band codebooks. To reflect the perceptual properties of the human ear, the bandwidths of these codebooks increase from low to high frequencies as listed below : 50-500 Hz, 500-1000 Hz, 1000-1500 Hz, 1500-2000 Hz, 2000-2600 Hz, 2600-3400 Hz, 3400-4400 Hz, 4400-5600 Hz, 5600-7000 Hz.

In the case of wideband music samples specified for the ITU-T qualification test plan, we observed that the signal strength in each of the nine bands varies relatively little with time, and is approximately equal for all the bands. Therefore, a fixed codebook size allocation is used. We use a 512-size codebook for each band. A fixed codebook size allocation also avoids the complexity of psychoacoustic modeling.

When pitch prediction is used, we steal 13 bits from the multi-band excitation codebook indices. The codebook sizes for the nine bands are then, 9, 9, 9, 9, 9, 7, 6, 5, 5 bits.

The excitation search is similar to ordinary multi-stage vector quantization and offers low complexity and high robustness. Furthermore, since the codevectors of the two codebooks, are nearly orthogonal (being restricted to different frequency bands), the sequential search of the codebooks provides almost the same performance as that of an optimal joint search of the codebooks.

The nine fixed codebook gains are computed, and signs of these gains are transmitted using 9 bits. The gain magnitudes are quantized using predictive split-VQ with 18 bits. The splits are of 2, 2, 2, and 3 dimensions. Fourth-order AR predictor is used for first 3 splits and second-order AR predictor is used for the last split. The predictor predicts the energy of the fixed excitation contribution based on the sequence of previously selected fixed excitation vectors. The quantized gain is expressed as a product of the predicted gain based on previous fixed excitation codebook energies and a correction factor. The correction factors for nine bands are vector quantized using 5, 4, 4, and 5 bits respectively for the splits.

The bit allocation is summarized in Table 1.

8. RESULTS

We conducted subjective tests of the MB-CELP wideband speech coder at 24 kbps and the G.722 coder at 56 kbps. The test was performed using forced-choice A/B pairwise comparisons with 12 listeners evaluating 16 sentence-pairs per listener, and 8 music samples per listener. The listeners were unfamiliar with these coders. The sentence pairs, and music samples were chosen according to the subjective test plan formalized by the ITU-T for the current wideband standardization

Pitch Prediction Flag	0	1
Parameters	Bits/Frame	Bits/Frame
LSP	21	21
Pitch Flag	2	2
Pitch Delay	0	18
Pitch Gain	0	8
MB Codebook Indices	162	136
MB Codebook Gains	54	54
Total	239	239

Table 1: Bit allocation

program. Our coder at 24 kbps was preferred over the G.722 coder at 56 kbps 64.58% to 35.42% for speech, and 61.46% to 38.54% for music.

9. CONCLUSION

The new MB-CELP coder employs a multi-band bank of excitation codebooks, two-stage linear prediction, and conditional pitch prediction. It achieves high-quality low bit-rate universal coding of wideband speech and music while maintaining low-delay. Listening test results show that 24 kbps version of this coder performs better than the G.722 coder at the much higher bit-rate of 56 kbps.

10. REFERENCES

[1] X. Maitre,"7 kHz audio coding within 64 kbit/s", *IEEE Journal on Selected Areas in Comm.*, vol. 6, No. 2, pp. 283-298, Feb. 1988.

[2] E. Harborg, J. E. Knudsen, A. Fuldseth and F. T. Johansen, "A Real-time Wideband CELP Coder for a Videophone Application", *Proceedings of ICASSP*, 1994, pp. II-121 - II-124.

[3] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE J. Selected Areas Commun. 6:314-323, 1988.

[4] B. Moore, "An Introduction to the Psychology of Hearing", Academic Press, 1992.

[5] M. Yong, G. Davidson and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction", *Proceedings ICASSP*, April 1988, pp. 402-405.

[6] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud and V. Cuperman, "Efficient Search and Design Procedure for Robust Multi-Stage VQ of LPC Parameters for 4 kb/s Speech Coding", *IEEE Trans. Speech and Audio Processing*, vol. SAP-1, pp. 373-385, Oct. 1993.

[7] M. Serizawa, A. Murashima and K. Ozawa, "A 16 kbit/s Wideband CELP Coder With a High-order Backward Predictor and its Fast Coefficient Calculation" *Proc. of IEEE Workshop on Speech Coding*, Sept. 1997, pp. 107-108.

[8] A. Ubale and A. Gersho, "Multi-band CELP Coding of Speech and Music", *Proc. of IEEE Workshop on Speech Coding*, Sept. 1997, pp. 101-102.

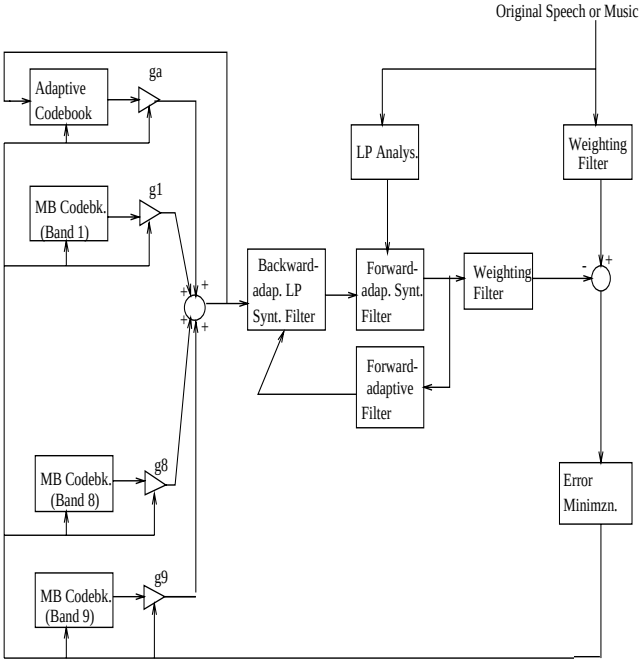

Figure 1: Multi-band CELP encoder structure.

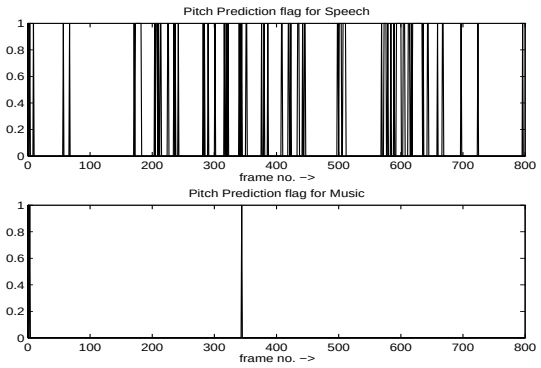

Figure 2: Use of Pitch Prediction.