UNIVERSAL SUCCESSIVE REFINEMENT OF CELP SPEECH CODERS

Hui Dong, Jerry D. Gibson

Electrical Engineering Department
Southern Methodist University
Dallas, TX 75275
(huidong, gibson)@seas.smu.edu

ABSTRACT

Many speech coding standards are based upon code-excited linear prediction (CELP), and it is desirable to develop layered coding methods that are compatible with this installed base of coders. We propose a layered speech coding structure that is universally compatible with all CELP-based coders. This structure encodes the reconstruction error signal from layer 1 using a low-delay, adaptive tree coder based upon the mean squared error (MSE) criterion. We note that rate distortion optimal successive refinement is achievable using two different distortion criteria and we derive expressions for the rate distortion function under autoregressive Gaussian assumptions on the source and the two different distortion measures. We demonstrate the universality of the approach by developing two-layer coders for a 3.65 kbps CELP coder, G.723.1, and G.729. We show that our layering method is favorably competitive with the MPEG-4 layering method at 8.7 kbps for both clean and noisy speech. Using tree coding and the MSE criterion in layer 2 improves speech naturalness when coding noisy speech.

1. INTRODUCTION

Many speech coding standards are based upon code-excited linear prediction (CELP), and these coders are deployed in a variety of wireless and wireline applications. As examples we cite the TIA 8-kbps VSELP for use with TDMA in the North American digital cellular mobile radio system, the variable rate TIA QCELP coder for use with CDMA in the North American digital cellular mobile radio system, G.728 LD-CELP for 16 kbps toll quality wireline speech coding, G.729 8 kbps CS-ACELP for use in wireline telephony, and the dual-rate speech coder G.723.1 CS-ACELP for visual telephony and videoconferencing in H.323 and H.324 [1].

In a networked multimedia environment, it is desirable to have a very flexible speech coding system, and layered or scalable compression schemes are important for such applications. In early 1980s, Jayant [2] discussed an embedded ADPCM system which consists of an ADPCM core layer and an APCM enhancement layer, operating from 16 kbps to 48 kbps. Recently MPEG-4 adopted a number of functionalities, including additional scalability of the transmitted bitstream and scalability of the complexity of the decoder. The core coder of MPEG-4 CELP coding is based on a CELP algorithm and encodes the input speech signal at a predetermined bitrate range. Bitrate scalability is achieved by encoding the speech signal using a combination of the core coder and the bitrate scalable tool [3].

In order to increase network flexibility and improve overall system performance, it is desirable to develop layered coding methods that are compatible with this installed base of coders. Ramprashad [4] proposed a two stage hybrid embedded speech/audio coding structure in which an existing fixed core codec is combined with another audio codec. This system improves the quality of music and speech with background noise. Here we propose a layered speech coding structure, depicted in Fig. 1, that is universally compatible with all CELP-based coders and does not require the modification of the deployed base. This structure, called CELPTree, encodes the reconstruction error signal from layer 1 using a low-delay, adaptive tree coder based upon the mean squared error (MSE) criterion. This approach is unique in comparison to most existing layered speech coding schemes in that we incorporate two different fidelity criteria in our layered coding structure. We note that Rimoldi [5] has shown that rate distortion optimal successive refinement is achievable based upon two different distortion criteria, and we derive the rate distortion function for layered coding under the assumption of a Gaussian autoregressive (AR) source and the use of a weighted squared error criterion in layer 1 and the MSE distortion criterion in layer 2.

We demonstrate the universality of the approach by developing two layer coders for a 3.65 kbps CELP coder, G.723.1, and G.729. We show that our layering method is favorably competitive with the MPEG-4 layering method at 8.7 kbps for both clean and noisy speech. Using tree coding and the MSE criterion in layer 2 yields substantial increases in output SNR and improves speech naturalness when coding noisy speech.

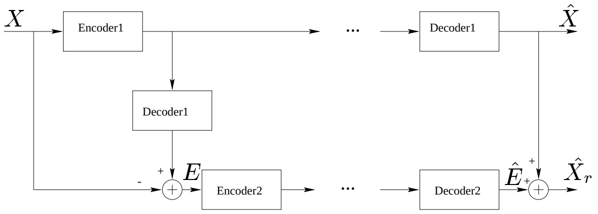

Fig. 1. The two layered codec system

The paper is outlined as follows. Section 2 provides the theoretical foundation for the CELPTree refinement approach. Section 3 briefly introduces the coding schemes: CELP coding and tree coding. Section 4 gives simulation results and performance comparison tests. Finally, Section 5 presents the summary of this work.

This work was supported by NSF Grant No. NCR 9796255

2. SUCCESSIVE REFINEMENT OF THE CELPTREE STRUCTURE

Assume that the source X is encoded as $\hat{X}$ at rate R_1 bits per symbol with distortion D_1. Then the error signal information is added to the first layer at rate $R_e = R_2 - R_1$ bits per symbol so that the two-stage resulting reconstruction $\hat{X}_r$ now has distortion D_2 and rate R_2. Based on the information-theoretic results on successive refinement that have been reported in the literature [6], [5], we show that this structure is successively refinable in the rate-distortion optimal sense, and provide an interpretation of the different distortion measures.

By the Shannon backward channel theorem [7] for layer 1 and 2 shown as Figs. 2 and 3, where Δ is the coding error from layer 2 and E is statistically independent of $\hat{X}$, we have using Figs 2 and 3

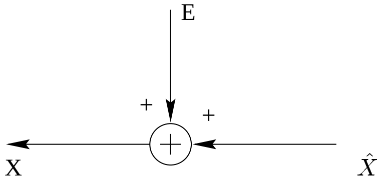

Fig. 2. The optimum backward channel for layer 1

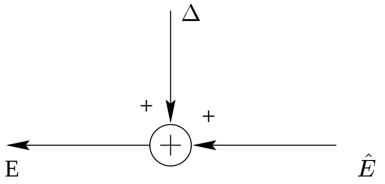

Fig. 3. The optimum backward channel for layer 2

Therefore the refined output is

$$\hat{X}_r = \hat{X} + \hat{E} \tag{1}$$

$$= X - \Delta \tag{2}$$

That is $X - \hat{X}_r = \Delta$ which means that the distortion of the refined output is due to the distortion of layer 2 only. Thus, we have the following theorem for this structure [6], [5].

Theorem: Successive refinement with distortion D_1 and D_2 ($D_1 \geq D_2$) can be achieved if there exists a conditional distribution $p(\hat{x}, \hat{x_r} \mid x)$ with

$$Ed_1(X, \hat{X}) \leq D_1, \tag{3}$$

and

$$Ed_2(E, \hat{E}) \leq D_2, \tag{4}$$

such that

$$R_1 \geq I(X; \hat{X}), \tag{5}$$

$$R_e \geq I(E; \hat{E}), \tag{6}$$

Suppose the input signals to the two layers are Gaussian autoregressive (AR) sources [8], characterized by a zero mean value with a variance σ_x^2 and a zero mean value with a variance σ_e^2 respectively.

The rate-distortion function in the first layer which is based on the frequency weighted distortion measure [8] is

$$R_1(D_1) = \frac{1}{2}\log\frac{\sigma_x^2}{D_1} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log W(\lambda)d\lambda \tag{7}$$

where $W(\lambda)$ is the frequency weighting parameter [8].

The rate-distortion function for the second layer based on MSE [7] is

$$R_e(D_2) = \begin{cases} \frac{1}{2}\log\frac{\sigma_e{}^2}{D_2} & \sigma_e{}^2 > D_2 \\ 0 & \sigma_e{}^2 \leq D_2 \end{cases} \tag{8}$$

however $D_1 = \sigma_e{}^2$ for this system, therefore

$$R_e(D_2) = \begin{cases} \frac{1}{2}\log\frac{D_1}{D_2} & D_1 > D_2 \\ 0 & D_1 \leq D_2 \end{cases} \tag{9}$$

The rate for the refined output including both layers can be shown to be

$$\begin{aligned} R_2(D) &= R_1(D_1) + R_e(D_2) \\ &= \frac{1}{2}\log\frac{\sigma_x^2}{D_1} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log W(\lambda)d\lambda + \frac{1}{2}\log\frac{D_1}{D_2} \end{aligned}$$

Thus

$$R_2(D_2) = \frac{1}{2}\log\frac{\sigma_x^2}{D_2} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log W(\lambda)d\lambda \tag{10}$$

As a result, the refinement is achievable ($D_1 > D_2$) by sending layer 2 information, shown in Eq. (9). Also the refined output is subject to a frequency weighted distortion measure if the first layer is coded by a frequency weighted distortion and the second layer is coded by the MSE distortion criterion.

3. CODING SCHEMES

Based upon its widespread deployment in applications, we use CELP as the core speech coder to deliver an acceptable quality at a limited bit rate. The second layer coder is designed to enhance the core output by coding the error signal (waveform matching) for the core. Tree encoding with a MSE distortion measure is a promising approach for layer 2 since it is known to be capable of performing arbitrarily close to the rate distortion bound for any memoryless source and single-letter fidelity criterion [9], [10].

The auditory masking effect in the human hearing system is included in the error criterion of all CELP coders. MSE distortion measure is used for the second layer coding to match the error signal waveform, and from Sec. 2, the distortion measure for the refined output is still frequency weighted.

3.1. CELP coder for the core

Though the proposed scheme is applicable to any CELP coder, three Algebraic-Code-Excited Linear Predictive (ACELP) coders have been investigated for use as a core: a CELP coder at 3.65 kbps (labeled CELP 1), G.723.1 at 5.3 kbps and G.729 at 8 kbps.

All three CELP coders use a 10th order LPC filter, but use different structures for the codebook, pitch filter and weighting filter. They also have different frame sizes, windowing schemes, and quantization methods for the parameters. Table 1 gives a bit allocation comparison of the three CELP coders. More details on each coder are given in [3], [11], [12].

Table 1. Bit allocation for CELP coders (bits/second)

Parameter	CELP 1	G.723.1	G.729
LPC	550	800	1800
Pitch delay	800	600	1400
gains	600	1600	1400
fixed-codebook	1700	2267	3400
Total	3650	5267	8000

3.2. Tree coder for the refinement layer

Tree coding has been studied widely in [13]-[18]. In the tree coder for this work, the excitation sequences are generated from variables placed on the branches of a tree, where the tree can be populated with quantizer output levels or random variates from a particular distribution. Only the path map symbols, specifying the branches of the tree that contain the desired output sequence, are transmitted, and hence, a fractional number of bits/symbol is possible. Unlike the block-oriented CELP coders with delay, tree coders may send one path map symbol at a time, and may be classified as waveform coders.

In this experiment, backward adaptive tree coders are used that incorporate a robust backward coefficient adaptation structure [18]. Bits are allocated to the index of the path map only, and 12 kbps and 16 kbps tree coders in [19] are adopted for layer 2.

3.3. Bit rate control

Bit-rate scalability is provided by adding the refinement layer. The refinement layer can be added in two schemes: the voiced part only of error signal as refinement and the full error signal as refinement.

The human auditory system is much more sensitive to voiced speech sounds than to unvoiced and transitional sounds of speech [20]. So, a voiced/unvoiced detector is used for layer 2 , and only the information on voiced parts are transmitted to refine the core. This results in comparative bitrates and quality as in MPEG4 since typically only 30% of the speech is voiced.

Since CELP coders do not work well for noisy speech due to their specialization to speech signals, Layer 2 tree coding can compensate for this drawback and get natural sounding noisy speech by sending the full error signal as refinement under very noisy conditions.

4. PERFORMANCE ANALYSIS

To analyze the performance of the CELPTree system, we compare the scalability, the segmental SNRs, spectrograms and the informal listening quality of coded speech files from MPEG4 CELP coders and our CELPTree coders. The tests used 8 speech files: 4 females, 4 males, and car noise and babble noise were added to the speech at different SNRs.

4.1. Bitrate scalability

Compared to the MPEG-4 CELP enhancement layers [3], where decoders reproduce the speech signal of different quality depending on the number of received fixed codebooks, the enhancement layer in CELPTree of Fig.1 depends on the core only through the CELP core decoder output. The advantage is that the second layer can be linked to any CELP coder with no modification to either the core or refinement systems. In addition, the second layer can use an entirely different coding paradigm from the core layer. Thus, the CELPTree structure offers flexible coding schemes for the different communications environments.

When the optimized standards G.723.1 and G.729 and the lowest rate CELP coders were used as core coders, the average bitrates of the 8 sentences are given in Table 2. Note that Rf12V means the core is refined by voiced(V) parts from tree coding at 12 kbps and Rf16F means the core is refined by the full(F) error signal from tree coding at 16 kbps. Also, on the average, 22% of these 8 speech files are detected to be voiced. Therefore while the MPEG4 CELP coder covers bitrates from 3850 to 18200 bits/s, the CELPTree coder covers bitrates from 3650 to 24000 bits/s which offers needed flexibility to get natural sounding noisy speech.

Table 2. Bit rates of CELPTree coders (bits/second) with different refinement

	Core only	Core Rf12V	Core Rf12F	Core Rf16V	Core Rf16F
CELP1	3650	6290	15650	7170	19650
G723.1	5267	7907	17267	8787	21267
G729	8000	10640	20000	11520	24000

4.2. Objective measures of coded speech

The segmental SNRs for voiced parts of coded speech from CELPTree coders and MPEG4 CELP coders are compared in Table 3 at bitrates of 8787 bps and 8700 bps respectively. It shows that the CELPTree coder performs much better than the MPEG4 CELP coder in terms of SNR. Background noise decreases the performance, but the CELPTree coder is more robust than MPEG4 CELP coder.

Also, when coding noisy speech, the CELPTree coder gets better natural reconstructed noisy speech than the MPEG4 CELP coder. Figure 4 shows that the spectrogram of coded speech from CELPTree matches the original speech better than that of coded speech from the MPEG4 CELP in most areas.

4.3. Subjective evaluation of coded speech

Informal listening tests show that the CELPTree coder has a performance practically equal to that of MPEG4 CELP coders at the same bitrate for clean speech. However, for noisy speech and especially for background car noise speech, the CELPTree coder has better performance, while the speech from MPEG4 CELP coders has some audible artifacts. The primary reason for this advantage is that tree coding is waveform-like coding, while CELP coding is more model-based.

Table 3. Segmental SNRs for voiced speech (dB)

speaker/background	CELPTree	MPEG4CELP
Male1/clean	19.86	13.29
Male1/babble noise	19.33	12.97
Male1/car noise	18.50	12.12
Male2/clean	17.95	11.90
Male2/babble noise	18.03	11.68
Male2/car noise	17.08	11.06
Female1/clean	21.36	15.41
Female1/babble noise	21.00	14.81
Female1/car noise	20.33	14.20
Female2/clean	20.62	14.95
Female2/babble noise	19.92	14.52
Female2/car noise	19.28	13.76

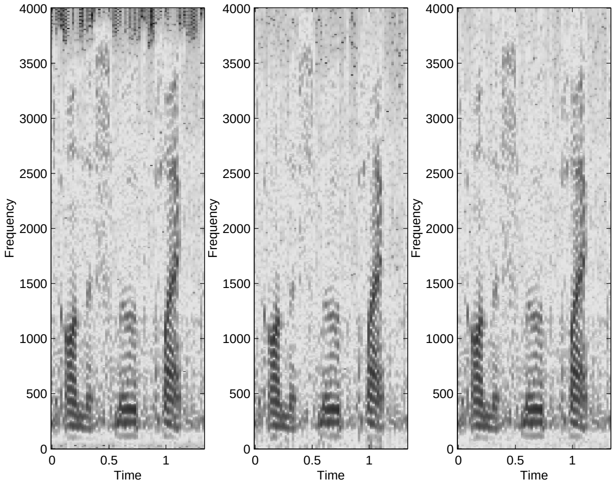

Fig. 4. Partial spectrogram of Male 1 speaker in car noise background: original(left), coded from MPEG4 CELP, coded from CELPTree(right)

5. SUMMARY

The two layer CELPTree combination provides a successively refinable speech coding structure, and a way to enhance any existing CELP-based speech coding system. The structure allows for the compensation of distortions inherent to the core coding paradigm, while still taking advantage of the high compression ratios of CELP coding. Since tree coders are amenable to interpretation as waveform coders, the proposed layer 2 coder improves the refinement for the noisy input speech under different environments.

6. REFERENCES

[1] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: principles & standards*, Morgan Kaufmann, San Francisco,CA, 1999.

[2] N. S. Jayant, "Variable rate ADPCM based on explicit noise coding," *The Bell System Technical Journal*, vol. 62, no. 3, pp. 657–677, March 1983.

[3] ISO/IEC JTC1 SC29/WG11, ISO/IEC FDIS 14496-3, *Information Technology-Coding of Audiovisual Objects-Part 3: Audio*, May 1998.

[4] Sean A. Ramprashad, "A two-stage hybrid embedded speech/audio coding structure," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*. May 1998, vol. 1, pp. 337–340, Seattle, Washington.

[5] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. on Information Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

[6] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. on Information Theory*, vol. 37, no. 2, pp. 269–275, March 1991.

[7] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[8] L. D. Davisson, "Rate-distortion theory and application," *Proc. IEEE*, vol. 60, pp. 800–808, July 1972.

[9] F. Jelinek and J. B. Anderson, "Instrumentable tree encoding of information sources," *IEEE Trans. on Information Theory*, vol. 17, pp. 118–119, 1971.

[10] C. R. Davis and M. E. Hellman, "On tree coding with a fidelity criterion," *IEEE Trans. on Information Theory*, vol. 21, pp. 373–378, 1975.

[11] ITU-T,Recommendation G.729, *Coding of Speech at 8 kbps Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)*, March 1996.

[12] ITU-T,Recommendation G.723.1, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbps*, March 1996.

[13] F. Jelinek, "Tree encoding of memoryless time-discrete source with a fidelity criterion," *IEEE Trans. on Information Theory*, vol. 15, no. 5, pp. 584–590, Sept. 1969.

[14] R. J. Dick, T. Berger, and F. Jelinek, "Tree coding for gaussian sources," *IEEE Trans. on Information Theory*, vol. 20, pp. 332–336, 1974.

[15] J. B. Anderson and J. B. Bodie, "Tree encoding of speech," *IEEE Trans. on Information Theory*, vol. 21, no. 4, pp. 379–387, July 1975.

[16] M. Foodeei and P. Kabal, "Low-delay celp and tree coders: comparison and performance improvements," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 25–28, 1991.

[17] J. D. Gibson and W. W. Chang, "Fractional-rate multitree speech coding," *IEEE Trans. on Communications*, vol. 39, pp. 963–974, June 1991.

[18] H. C. Woo and J. D. Gibson, "Low delay tree coding of speech at 8 kbps," *IEEE Trans. on speech and audio processing*, vol. 2, no. 3, pp. 361–370, July 1994.

[19] H. Dong, "Adaptive code generator for tree coding of speech," M.S. thesis, Texas A&M University, College Station, Texas, August 1998.

[20] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*, Elsevier, Boston, 1995.