Synthesis Filter/Decoder Structures in Speech Codecs

Jerry D. Gibson, Electrical & Computer Engineering, UC Santa Barbara, CA, USA
gibson@ece.ucsb.edu

Abstract
Using the Shannon backward channel result from rate distortion theory, we derive new decoder/synthesizer structures for common linear predictive speech models. We calculate the transfer functions for these new structures for common source models and discuss the effect on the reconstructed signal. The perceptual weighting filter and the average distortion both play a prominent role. We show that common CELP encoders and decoders lack these components, and note that at least at the decoder, postfiltering as currently used in CELP may partially compensate for the missing structure.

1. Introduction

[1]Codebook excited linear predictive coding (CELP) is the underlying principle used in all narrowband and some wideband standardized speech codecs today [1]. Most recent efforts toward the development of fullband codecs (20 Hz to 20 kHz) utilize a combination of the CELP approach and the transform/filter bank approaches with well-designed switching between the coding methods, as is evident in the recently standardized USAC codec [2] and as is expected in the EVS codec for LTE Mobile Systems [3]. While linear prediction had long had success for speech waveform coding and was later used for low bit rate voice codecs by modeling the vocal tract [4], codebook excited analysis-by-synthesis coding using linear prediction was first motivated by rate distortion theory principles [5-10].

In this paper, we return to rate distortion theory fundamentals to examine the structure of the decoder and the synthesizer used in the encoder for optimal speech coding subject to the squared error fidelity criterion. It is shown that the usual CELP decoder should have additional excitation filtering, currently not present in CELP codecs, that is dependent on the perceptual weighting filter and on the average distortion.

The paper is organized as follows. Section II contains a quick review of the linear prediction model and CELP codecs. Section III then uses the classical Shannon lower bound and Shannon backward channel concept to derive the form of the optimal code generator or decoder for sources satisfying the linear prediction model with perceptual weighting of the distortion. Section IV presents examples of the new structures for three simple representative source models subject to weighted and unweighted squared error fidelity criteria. Section V presents comparisons to current CELP codec structures, and Section VII discusses possible implementation approaches. Section VII contains some conclusions.

2. Linear predictive Voice Codecs

Linear predictive coding (LPC) is the dominant paradigm for narrowband speech coding in the last 40 years. In LPC, the decoder or synthesizer has the form shown in Fig. 1, wherein

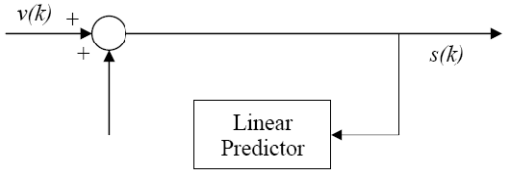

Figure 1. The Linear Prediction Model

The output speech is reconstructed according to

$$s(k) = -\sum_{i=1}^{m} a_k s(k-i) + v(k)$$

This decoder structure has been carried over to the code-excited linear predictive (CELP) analysis-by-synthesis codecs with encoders of the form shown in Fig. 2 and decoders as shown in Fig. 3. In these figures, the Synthesis filter is the linear predictor given in Fig. 1.

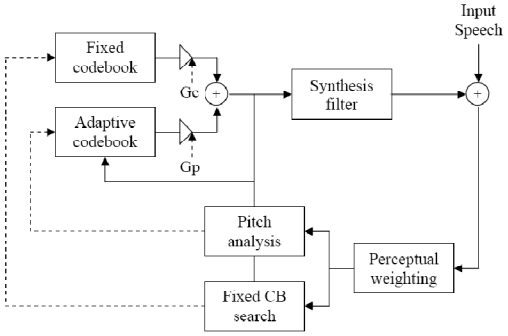

Figure 2. CELP Encoder

[1] This research has been funded by NSF Grant Nos. CCF-0728646 and CCF-0917230

Having a synthesis filter that mimics the linear prediction model appears to be intuitive and well-motivated. However, it is well known that other modifications such as postfiltering following the decoder may improve the reconstructed speech in some instances. In this paper we explore alternative decoder structures implied by rate distortion optimal lossy source coding theory.

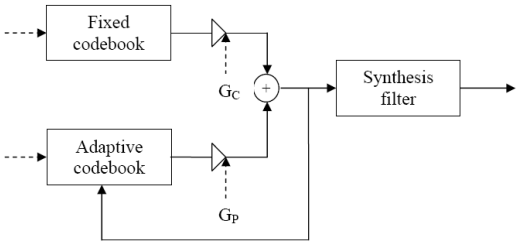

Figure 3. CELP Decoder

3. Rate Distortion Analysis for Difference Distortion Measures

To introduce the basic underlying principle from rate distortion theory, we begin by considering the problem of encoding a memoryless Gaussian source *X* subject to the mean squared error (MSE) fidelity criterion. A classic lower bound on the rate distortion function for difference distortion measures, first introduced in Shannon's original paper on rate distortion theory in 1959, is called the Shannon lower bound [11].

When this bound is satisfied with equality for an average distortion D_1, the optimally encoded output, $\hat{X}_1$, satisfies the Shannon backward channel condition expressed by [5]

$$X = \hat{X} + Z \tag{1}$$

where if *X* is zero mean, Gaussian with variance σ^2, then $\hat{X}$ is zero mean, Gaussian with variance $\mathrm{var}(\hat{X}) = \sigma^2 - D$, and Z a zero mean Gaussian random variable that is statistically independent of $\hat{X}$ with variance D. Memoryless sources do not give us much insight into the coding of actual speech signals so we turn our attention now to sources that satisfy the linear prediction model. However, after suitable transformations, the Shannon backward channel condition can still be imposed to provide the essential results.

An *m*th-order, time-discrete AR source can be expressed as

$$X_t = -\sum_{k=1}^{m} a_k X_{t-k} + Z_t \tag{2}$$

where $a_1, \ldots, a_m$ are the AR coefficients, and $\{Z_t\}$ is a sequence of independent and identically distributed (iid) random variables, and X_r and Z_s are statistically independent if *s*>*r*. The Shannon backward channel formulation has been used by Berger to analyze optimal tree encoding of Gaussian autoregressive (AR) sources subject to the MSE distortion measure. In his analysis, not repeated here, Berger shows that the power spectral density (psd) of the optimal reconstructed value is of the form [5]

$$\Phi_Y(z) = \Phi_X(z) - D \tag{3}$$

for average distortion *D*. For an AR source, the source psd is

$$\Phi_X(z) = \frac{\sigma^2}{|A(z)|^2} \tag{4}$$

which upon substitution into the above yields

$$\Phi_Y(z) = \frac{\sigma^2 - |A(z)|^2 D}{|A(z)|^2} = \frac{|B(z)|^2}{|A(z)|^2} \tag{5}$$

This perhaps surprising result shows that the reconstructed output is no longer purely AR, but it is now an autoregressive moving average (ARMA) sequence, where the MA part is dependent on the average distortion and on the linear prediction coefficients through the numerator polynomial *B(z)*.

The analysis can be extended to the optimal encoding of this AR source subject to a weighted squared error distortion measure to obtain

$$\Phi_Y(z) = \Phi_X(z) - \frac{D}{|W(z)|^2} \tag{6}$$

where $W(z)$ is the frequency weighting of the reconstruction error (the distortion). Substituting as before for the psd of the input, we obtain the expression [12]

$$\Phi_Y(z) = \frac{\sigma^2 - D\dfrac{|A(z)|^2}{|W(z)|^2}}{|A(z)|^2} = \frac{|B(z)|^2}{|A(z)|^2} \tag{7}$$

The results in Eqs. (5) and (7) are subject to the small distortion condition, namely,

$$\frac{D}{|W(z)|^2} \le \frac{\sigma^2}{|A(z)|^2} \tag{8}$$

Which also guarantees that the *B(z)* polynomial in the numerator of the needed form exists. The numerator polynomial in the expression for the psd of the reconstructed source now depends upon the perceptual weighting function as well as the average distortion and linear prediction coefficients.

These results imply that the synthesizer in the encoder and the decoder in CELP codecs should not simply be the linear prediction model with appropriate excitation. In the following sections, we investigate the impact of the change in the numerator on the reconstructed spectrum.

4. The Excitation Shaping Filter

We denote the numerator polynomial *B(z)* as the excitation shaping filter and we illustrate the form of this filter with and without perceptual weighting for three different source spectra, a 3rd order Butterworth shaping, a 3rd order AR shaping based on the coefficients in [6,13], and a 10th order AR model. The power spectral densities of these sources are shown in Fig. 4.

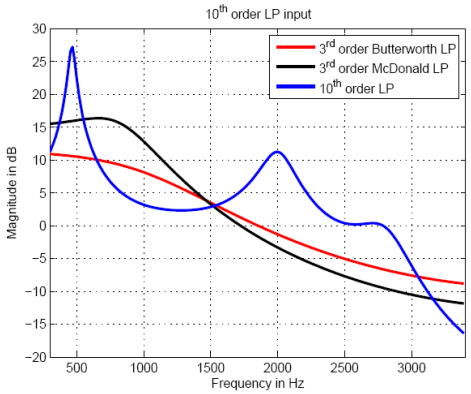

Figure 4. Power Spectral Densities of Example Sources

For the unweighted case, the excitation shaping filter has the form shown in Fig. 5. Expanding both sides of the numerators in Eq. (5), we obtain n+1 nonlinear equations in n+1 unknowns. Using a nonlinear optimization technique, these equations can be solved for the coefficients of *B(z)* [12].

For the case with weighting, we use the weighting function

$$W(z) = A(z/\gamma_1)/A(z/\gamma_2) \qquad (9)$$

where γ_1 and γ_2 are 0.94 and 0.6, respectively, since these parameters are used for some modes of the AMR-NB codec. With weighting, the structure of *B(z)* becomes more complicated and has the form shown in Fig. 6. In this case, upon expanding the numerator of the expression in Eq. (7), where *R(z)* is the numerator of *B(z)* and *Q(z)* is the denominator of *B(z)*, we obtain 2n+1 nonlinear equations in 2n+1 unknowns. The solutions of this nonlinear optimization are nonunique and depend on the initial conditions and the optimization scheme [12].

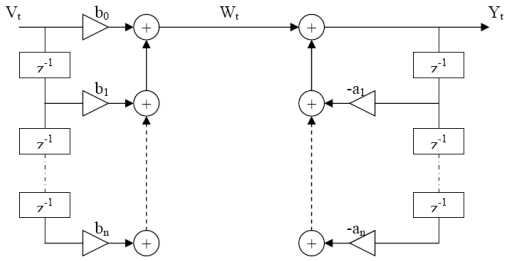

Figure 5. Excitation Shaping Filter *B(z)* for No Perceptual Weighting

The details of calculating the *B(z)* expressions are given elsewhere, but we exhibit the magnitude response of the resulting filters for the unweighted and weighted cases and different average distortions for each of the three sources in Figs. 7, 8, and 9.

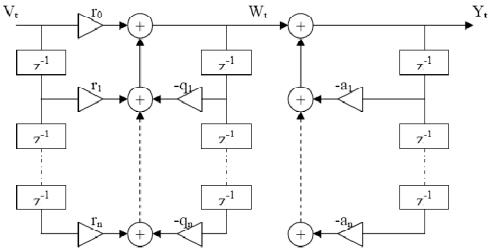

Figure 6. Excitation Shaping Filter *B(z)* with Perceptual Weighting

From Figs. 7 and 8 for the 3rd order sources, we observe the following:

- *B(z)* has a low-pass filtering effect, the intensity of which increases with an increase in distortion.
- For the same distortion, the frequency response of *B(z)* with no weighting has a more severe low-pass filtering effect relative to the frequency response of *B(z)* with weighting.
- As distortion *D* is increased, we reach a point where the small distortion condition is not valid for MSE distortion but remains valid for weighted MSE distortion. This is true in the case of *D*=0.15 for the Butterworth coefficients and *D*=0.1 for the McDonald coefficients.

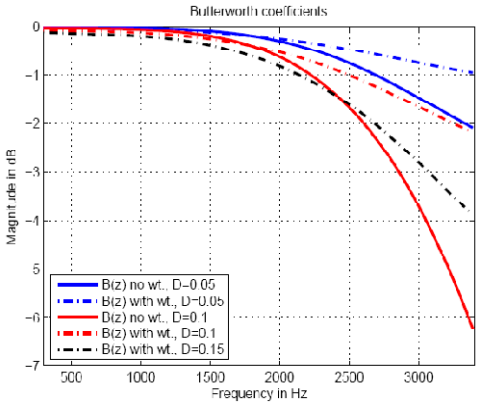

Figure 7. *B(z)* for 3rd Order Butterworth Source

The low-pass filtering effect of *B(z)* has been attributed in [6] to the rate distortion theory trading off the high frequency signal component (where the quantization effects are centered) against a reduction in noise. When weighting is used, the weighted distortion at high frequencies is reduced due to a redistribution of the noise across the spectrum. This may explain why the low-pass effect is less severe in case of weighted MSE, relative to when no weighting is used.

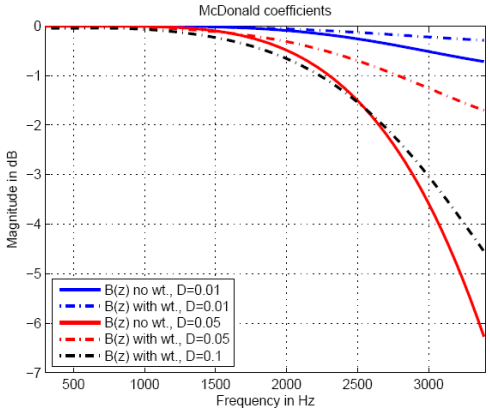

Figure 8. *B(z)* for 3rd Order MacDonald Source

For the AR(10) source in Fig. 9, the observations are similar to those that we had observed earlier for AR(3), but the weighting filter parameters play a more prominent role.

From Fig. 9, we note that:

- The low pass filtering effect in the frequency response of *B(z)* increases in severity as the distortion increases for both MSE and weighted MSE.
- In comparing the frequency response of *B(z)* without and with weighting, we see that weighting reduces the severity of the low pass filtering, relative to the case when no weighting is used. This is true for the cases where both the unweighted and weighted MSE satisfy the small distortion condition (D=0.005).
- For D=0.01 and D=0.05, the case without weighting does not satisfy the small distortion condition, and hence *B(z)* cannot be determined. However, for these specified values of *D*, when weighting is used, the small distortion condition is satisfied, allowing us to determine *B(z)*.
- We observe some shaping in the low frequencies visible for *B(z)* with weighting when D=0.05. To investigate this effect,we adjusted the weighting filter parameters, and the *B(z)* for a value of γ_2 =0.2 is shown in Fig. 10.

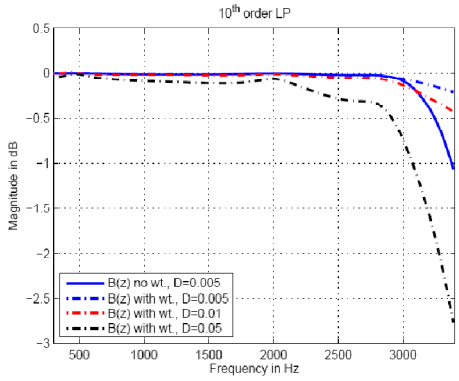

Figure 9. *B(z)* for 10th Order AR Source

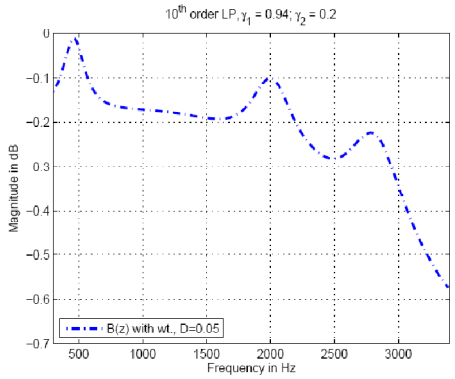

Figure 10. *B(z)* for 10th Order AR Source with Different Weighting

We see that the shaping is more prominent and the formants on the source spectrum start to appear. This value for the parameter γ_2 may appear due to adaptation of the perceptual weighting filter, but some codecs put upper and lower bounds on the value of this parameter. For

example, G.729 bounds γ_2 to be between 0.4 and 0.7 [14]. Following the explanation in [6] for the high-frequency effect, the shaping of the *B(z)* observed in Fig. 10 that emphasizes formant frequencies and de-emphasizes formant valleys and high frequency components can be said to be trading-off signal fidelity at formant valleys and high frequencies against a reduction in perceived noise.

5. Decoder Structures in CELP Codecs

For comparison purposes, we now examine the reconstructed output of the common standardized CELP codecs. The weighted reconstruction error in the analysis by synthesis calculation has the form

$$E(z) = W(z)[S(z) - \hat{S}(z)] \tag{10}$$

which upon rewriting yields,

$$S(z) = \hat{S}(z) + \frac{1}{W(z)} E(z) \tag{11}$$

By comparing to the Shannon backward channel condition, we see that this has the same form, and if the sequences are Gaussian and if we argue that $\hat{S}(z)$ is selected to satisfy the orthogonality condition from optimal estimation, the error will be independent of the reconstructed output. However, the expression for the transfer function of the decoder is

$$H(z) = \frac{1}{A(z)} \tag{12}$$

Therefore, the common CELP decoders (and encoders) lack the excitation shaping filter implied by rate distortion optimal encoding.

CELP decoders often have a postfilter with a numerator polynomial that may inadvertently compensate for this oversight, although the dependence on the average distortion is not explicit and of the same form. The rate distortion motivated structures will have the excitation shaping at the encoder too, within the analysis by synthesis loop, which is always lacking in current CELP codecs.

6. Implementations

The process to determine the rate distortion theory motivated excitation filter is quite involved, and therefore one has to determine how this approach might be incorporated into a practical voice codec. One approach would be to use the structure in Fig. 6, with the weighting parameter updated as in CELP, but with the remaining coefficients adapted according to algorithms as in Gibson [15]. Several other approximate structures are possible.

7. Conclusions

The Shannon backward channel result from rate distortion theory is shown to require zeros in the decoder/synthesizer structures for common all pole, linear predictive speech models. This is in contrast to the usual decoders/synthesizers used in popular CELP codecs. We calculate the transfer functions for these new structures for common source models and discuss the effect on the reconstructed signal. The parameters in the perceptual weighting filter and the average distortion both change the shaping of the excitation. Although the common CELP encoders and decoders lack these components, it is noted that, at least at the decoder, postfiltering as currently used in CELP may partially compensate for the missing structure.

References

[1] J. D. Gibson, "Speech Coding Methods, Standards, and Applications," *IEEE Circuits and Systems Magazine*, vol. 5, no. 4, pp. 30 – 49, 2005.

[2] M. Neuendorf, et al, "A Novel Scheme for Low Bitrate Unified Speech and Audio Coding – MPEG RM0, Convention Paper 7713, AES 126th Convention, Munich, Germany, May 7-10, 2009.

[3] K. Jarvinen, I. Bouazizi, L. Laaksonen, P. Ojala, A. Ramo, "Media coding for the next generation mobile system LTE," *Computer Communications*, vol. 33, pp. 1916-1927, 2010.

[4] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, R. Baker, *Digital Compression for Multimedia: Principles & Standards*, Morgan Kaufmann Publishers, Inc., 1998.

[5] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[6] J. B. Anderson, J.B. Bodie, "Tree encoding of speech," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 379-387, July 1975.

[7] S. G. Wilson, S. Husain, "Adaptive Tree Encoding of Speech at 8000 Bits/s With a Frequency-Weighted Error Criterion," *IEEE Trans. Comm.*, vol. COM-27, pp. 165-170, Jan. 1979.

[8 M. R. Schroeder, B. S. Atal, "Rate Distortion Theory and Predictive Coding," ," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 201-204.

[9] L. C. Stewart, R. M. Gray, Y. Linde, "The design of trellis waveform coders," *IEEE Trans. Commun.*, vol. COM-30, pp. 702-710, Apr. 1982.

[10] M. R. Schroeder, B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1985, pp. 25.1.1-25.1.4.

[11] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Conv. Rec*., vol. 7, 1959, pp. 142-163.
[12] N. A. Shetty, "Tandeming in Multi-Hop Voice Communication," Ph. D. Dissertation, ECE Dept, UCSB, Dec.2007.
[13] R. A. McDonald, "Signal-to-noise and idle channel performance of differential pulse code modulation systems-Particular applications to voice signals," *Bell Syst. Tech. J.*, vol. 45, pp. 1123-1151, Sept. 1966.
[14] R. Salami, et al, "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 116-130, March 1998.
[15] J. D. Gibson, "Adaptive Prediction in Speech Differential Encoding Systems," Proceedings of the IEEE, vol. 68, pp. 488-525, April 1980.

Jerry D. Gibson is Professor and Chair of Electrical and Computer Engineering at the University of California, Santa Barbara. He has been an Associate Editor of the *IEEE Transactions on Communications* and the *IEEE Transactions on Information Theory*. He was President of the IEEE Information Theory Society in 1996, and he has served on the Board of Governors of the IT Society and the Communications Society. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 1992-1994, and he is currently a member of the Multimedia Communications Technical Committee of the Communications Society. He was an IEEE Communications Society Distinguished Lecturer for 2007-2008, a member of the IEEE Awards Committee (2008-2010), and a member of the IEEE Medal of Honor Committee (2009-2010). He is an IEEE Fellow, and he has received The Fredrick Emmons Terman Award (1990), the 1993 IEEE Signal Processing Society Senior Paper Award, the 2009 IEEE Technical Committee on Wireless Communications Recognition Award, and the 2010 Best Paper Award from the *IEEE Transactions on Multimedia*.