NOVEL CODEC STRUCTURES FOR NOISE FEEDBACK CODING OF SPEECH

Juin-Hwey Chen

Broadcom Corporation, Irvine, California, USA

ABSTRACT

This paper presents several novel codec structures for Noise Feedback Coding (NFC) incorporating both long-term and short-term noise spectral shaping, as well as long-term and short-term prediction. In addition, the paper generalizes the conventional scalar-quantization-based NFC to vector-quantization-based NFC, and it lays the foundation for the associated efficient VQ codebook search and closed-loop VQ codebook design. BroadVoice®16, a PacketCable 1.5 mandatory narrowband speech codec standardized by CableLabs® for Voice over Cable in North America, is based on one of such novel NFC codec structures.

1. INTRODUCTION

Early predictive speech codecs such as DPCM and ADPCM used only single-stage short-term prediction and produced white coding noise. Recent popular predictive codecs such as Code-Excited Linear Prediction (CELP) added long-term prediction and noise spectral shaping. These codecs achieve noise spectral shaping by using a perceptual weighting filter in a closed-loop excitation codebook search. However, an alternative noise spectral shaping method known as Noise Feedback Coding (NFC) had been proposed nearly three decades before CELP came into existence.

The basic ideas of NFC date back to Cutler [1] in 1954. Based on Cutler's ideas, Kimme and Kuo proposed a noise feedback coding system for television signals [2]. Enhanced versions of NFC for Adaptive Predictive Coding (APC) of speech were later proposed by Makhoul and Berouti [3] and by Atal and Schroeder [4]. More recently, NFC has also been used to improve ADPCM, as proposed by Lee in [5].

Early NFC codecs only used a short-term predictor and a short-term noise feedback filter. The resulting short-term noise spectral shaping only affected the spectral envelope of the coding noise. The codec in [4] added a long-term predictor based on pitch periodicity but still used only short-term noise spectral shaping. Gerson and Jasiuk [6] introduced long-term noise spectral shaping to improve the speech quality of CELP codecs. They used a long-term perceptual weighting filter to make the noise spectrum follow the harmonic fine structure of the voiced speech spectrum. The codec in [5] has both long-term and short-term noise spectral shaping but only a short-term predictor.

As described above, both long-term prediction and long-term noise spectral shaping are desirable features that can make the output speech less noisy. Atal and Schroeder [4] used long-term prediction but not long-term noise spectral shaping. Lee [5] used long-term noise spectral shaping but not long-term prediction. Gerson and Jasiuk [6] used both, but in a CELP codec rather than an NFC codec.

Due to the required Vector Quantization (VQ) of excitation, CELP codecs have much higher complexity than conventional NFC codecs, which are based on scalar quantization (SQ). For applications that require low complexity and high speech quality, it is desirable to improve the conventional NFC so it incorporates both long-term prediction and long-term noise spectral shaping.

Addressing such applications, this paper introduces novel NFC codec structures [7] that employ both long-term prediction and long-term noise spectral shaping, in addition to the short-term prediction and short-term noise spectral shaping that are already used by conventional NFC. Furthermore, the paper generalizes the conventional SQ-based NFC to VQ-based NFC to improve the codec performance. The paper also introduces the fundamental approaches for the efficient VQ codebook search and closed-loop VQ codebook design for VQ-based NFC.

2. CONVENTIONAL NOISE FEEDBACK CODING

The structure of the most common conventional NFC is shown in Fig. 1. The short-term predictor $P(z)$ and the short-term noise feedback filter $F(z)$ have the transfer functions of

$$P(z) = \sum_{i=1}^{M} a_i z^{-i} \quad \text{and} \quad F(z) = \sum_{i=1}^{L} f_i z^{-i} \ ,$$

where M and L are the orders of the two filters, respectively. It can be shown that the codec reconstruction error, or coding noise, is given by

$$r(n) = s(n) - sq(n) = \sum_{i=1}^{M} a_i r(n-i) + q(n) - \sum_{i=1}^{L} f_i q(n-i) \ ,$$

or in terms of z-transform representation,

$$R(z) = \frac{1-F(z)}{1-P(z)} Q(z) \ .$$

1-4244-0469-X/06/$20.00 ©2006 IEEE

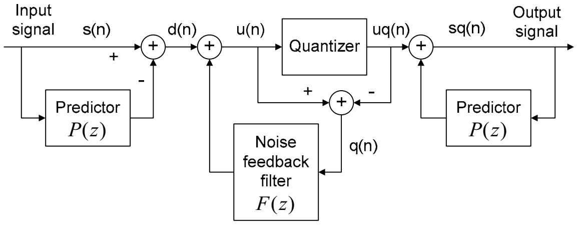

Fig. 1 Conventional Noise Feedback Coding

If the quantizer bit rate is high, the quantization error $q(n) = u(n) - uq(n)$ is roughly white. Then, the spectrum of the coding noise $r(n)$ will have the same shape as the frequency response of the filter $N(z) = [1 - F(z)] / [1 - P(z)]$. Atal and Schroeder [4] used $F(z) = P(z/\alpha)$, where $0 < \alpha < 1$.

Makhoul and Berouti [3] proposed the alternative NFC codec structure in Fig. 2, where $N(z)$ has a frequency response corresponding to the desired noise spectral shape.

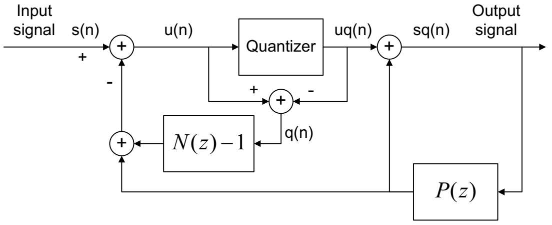

Fig. 2 Alternative form of Conventional Noise Feedback Coding

With proper $N(z)$, the structure in Fig. 2 is equivalent to the structure in Fig. 1. However, this structure in Fig. 2 gives more flexibility in the choice of $N(z)$, because $N(z)$ is not required to be a ratio of two polynomials as is the case for the structure in Fig. 1.

3. TWO-STAGE NOISE FEEDBACK CODING

To achieve Two-Stage Noise Feedback Coding (TSNFC) using two stages of noise spectral shaping (long-term and short-term) as well as two stages of prediction (long-term and short-term), one possible approach is to combine the long-term and short-term filters into a single composite filter and then re-use the structure of Fig. 1.

Let $P_s(z), P_l(z), F_s(z)$, and $F_l(z)$ represent the transfer functions of the short-term predictor, long-term predictor, short-term noise feedback filter, and long-term noise feedback filter, respectively. It can be shown [7] that replacing the predictor $P(z)$ in Fig. 1 by a composite predictor $P'(z) = P_s(z) + P_l(z) - P_s(z)P_l(z)$ will cause the single-stage analysis filter in the leftmost one-third of Fig. 1 to have a transfer function of $1 - P'(z) = [1 - P_s(z)][1 - P_l(z)]$. Similarly, it will cause the single-stage synthesis filter in the rightmost one-third of Fig. 1 to have a transfer function of

$$\frac{1}{1 - P'(z)} = \frac{1}{[1 - P_s(z)][1 - P_l(z)]}.$$

Therefore, both long-term prediction and short-term prediction are achieved. Furthermore, replacing the noise feedback filter $F(z)$ in Fig. 1 by a composite noise feedback filter $F'(z) = F_s(z) + F_l(z) - F_s(z)F_l(z)$ will cause the noise spectral shape to follow the frequency response of

$$\frac{1 - F'(z)}{1 - P'(z)} = \left(\frac{[1 - F_s(z)]}{[1 - P_s(z)]}\right)\left(\frac{[1 - F_l(z)]}{[1 - P_l(z)]}\right).$$

Hence, both long-term noise spectral shaping and short-term noise spectral shaping are also achieved.

Similarly, using Fig. 2 but with $P(z)$ replaced by $P'(z) = P_s(z) + P_l(z) - P_s(z)P_l(z)$, one can choose a suitable noise feedback filter $N(z) - 1$ such that it includes the effects of both short-term and long-term noise spectral shaping.

The composite filter approach described above achieves its goal, but the long-term and short-term filters are mixed together. A second approach allows us to separate short-term filter sections from long-term filter sections, but it requires a modification of the structures in Figures 1 and 2.

It is not obvious how these structures should be modified to achieve two-stage prediction and two-stage noise spectral shaping at the same time. For example, assuming the filters in Fig. 1 are all short-term filters, then, a seemingly reasonable way is to cascade a long-term analysis filter after the short-term analysis filter, cascade a long-term synthesis filter before the short-term synthesis filter, and cascade a long-term noise feedback filter to the short-term noise feedback filter in Fig. 1. However, this will not give a codec that achieves the desired result.

The key to a correct structure is to recognize that the quantizers in Figures 1 and 2 can be replaced by a single-stage NFC system. If the quantizer in Fig. 1 is replaced by a long-term NFC structure based on Fig. 1, the resulting "nested", or "layered", two-stage NFC structure is shown in Fig. 3. The shaded area Q_1 used to be a simple quantizer in Fig. 1, but now it is a single-stage NFC structure.

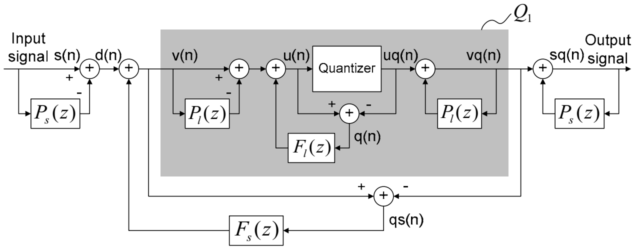

Fig. 3 Nested Two-Stage Noise Feedback Coding – Form 1

From the equations in Section 2, we have

$$R(z) = S(z) - SQ(z) = \frac{1 - F_s(z)}{1 - P_s(z)} QS(z) = \frac{[1 - F_s(z)]}{[1 - P_s(z)]} \frac{[1 - F_l(z)]}{[1 - P_l(z)]} Q(z).$$

This proves that the nested structure in Fig. 3 indeed performs both short-term and long-term noise spectral shaping in addition to short-term and long-term prediction.

The long-term section and short-term section in Fig. 3 are completely decoupled. This allows us to use different structures for long-term NFC and short-term NFC. For example, we can replace the shaded area in Fig. 3 by the NFC structure in Fig.2 to arrive at the equivalent nested TSNFC structure in Fig. 4. As another example, we can even replace the outer-layer short-term NFC structure in Fig. 4 by the NFC structure in Fig. 2 to get the equivalent structure in Fig. 5. For convenience, we will refer to the structures in Figures 3, 4, and 5 as TSNFC Forms 1, 2, and 3, respectively. We can even have a Form 4 (not shown) by replacing the shaded area of Form 3 by that of Form 1.

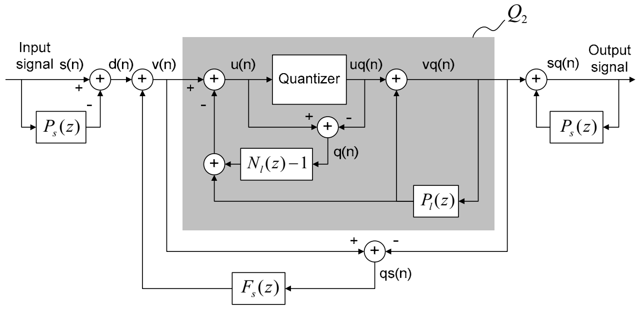

Fig. 4 Nested Two-Stage Noise Feedback Coding – Form 2

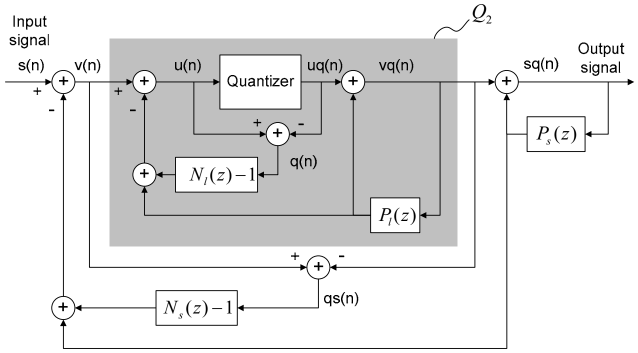

Fig. 5 Nested Two-Stage Noise Feedback Coding – Form 3

There is an advantage for such flexibility to mix and match different single-stage NFC structures in different parts of the nested TSNFC structure. For example, although Form 2 mixes two different types of single-stage NFC structures, it actually has the lowest complexity among the four forms discussed above. This is because the long-term NFC in Form 2 has only two filters rather than three as in Form 1, and for excitation VQ codebook search (to be discussed later), only the filter $F_s(z)$ in the outer-layer needs to be considered. In contrast, the VQ codebook search for Form 3 needs to include both $P_s(z)$ and $[N_s(z)-1]$.

Due to its lowest complexity, TSNFC Form 2 is used in BroadVoice®32, a 32 kb/s speech codec for 16 kHz wideband speech. On the other hand, due to the need to have more flexible control of the noise spectral envelope, TSNFC Form 3 is used in BroadVoice®16, a PacketCable 1.5 mandatory narrowband speech codec standardized by CableLabs® for Voice over Cable in North America [8].

To keep the complexity low, the quantizer in Figures 3 through 5 can be a scalar quantizer. We found that essentially transparent speech quality can be achieved with such SQ-based TSNFC codecs if such a scalar quantizer operates at a bit rate of 2 bits/sample or higher. However, to achieve good speech quality at a lower encoding rate for the prediction residual, vector quantization will need to be used.

4. VECTOR QUANTIZATION OF RESIDUAL

A “chicken-and-egg” problem will arise when attempting to vector-quantize the prediction residual signal *u*(*n*) in Figures 3 through 5. Normally the input vector to a vector quantizer is formed before the VQ operation can begin. However, due to the feedback filter structures in Figures 3 through 5, the second through the last samples of the VQ input vector cannot be determined until the VQ output vector is determined, but a normal vector quantizer cannot determine its output vector until its input vector is determined.

To solve this problem, we note that the quantizer in Figures 3 through 5 minimizes the energy of the quantization error *q*(*n*), and that the VQ codebook only has a limited number of codevectors. Hence, we can try all VQ codevectors one at a time, use the feedback filter structure to calculate the resulting quantization error vector *q*(*n*) for each of the codevectors, and select the VQ codevector that minimizes the energy of *q*(*n*) as the winning codevector.

The complexity of such a VQ codebook search can be reduced greatly. Take TSNFC Form 2 in Fig. 4 as an example. Figure 6 shows the portion of the Form 2 structure that affects the VQ codebook search. For the purpose of VQ codebook search, the structure in Fig. 6 is a linear system with the scaled VQ codevector *uq*(*n*) as the input signal and the quantization error vector *q*(*n*) as the output signal. According to the linear system theory, we can decompose the output vector *q*(*n*) into two components: the Zero-Input Response (ZIR) vector *qzi*(*n*) and the Zero-State Response (ZSR) vector *qzs*(*n*).

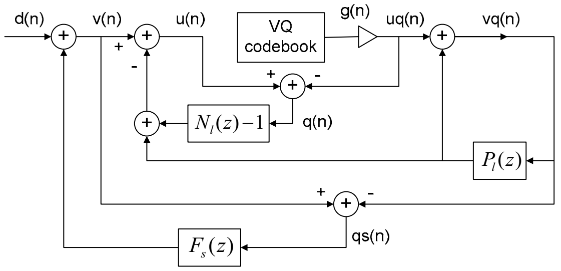

Fig. 6 Portion of Form 2 structure used in VQ codebook search

During the calculation of the ZIR vector *qzi*(*n*), the system input vector *uq*(*n*) is set to zero. In this case, Fig. 6 reduces to Fig. 7. Since this structure in Fig. 7 is unaffected by the VQ codevector, the ZIR vector *qzi*(*n*) only needs to be calculated once for each input speech vector.

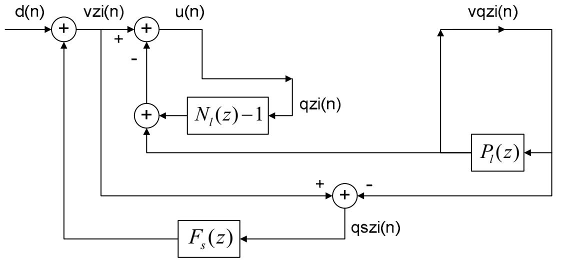

Fig. 7 Portion of Form 2 structure used in ZIR calculation

During the calculation of the ZSR vector *qzs*(*n*), the system internal states (i.e. filter memory and *d*(*n*)) are set to zero. If the vector dimension is chosen to be smaller than the minimum bulk delay of the two long-term filters $P_l(z)$ and $[N_l(z)-1]$ in Fig. 6, then these two filters will produce zero output signals for ZSR. In this case, Fig. 6 reduces to Fig. 8, which further reduces to Fig. 9.

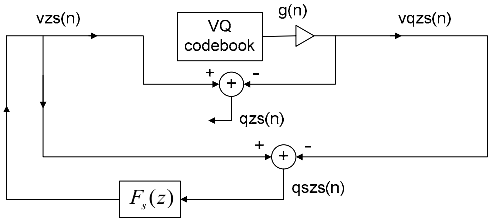

Fig. 8 Portion of Form 2 structure used in ZSR calculation

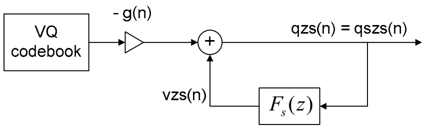

Fig. 9 A structure equivalent to the structure in Fig. 8

Since a ZSR vector needs to be calculated for each of the VQ codevectors, using the much simpler filter structure in Fig. 9 rather than Fig. 6 greatly reduces the complexity of the VQ codebook search, in a way similar to how ZIR and ZSR decomposition greatly reduces the excitation VQ codebook search complexity of CELP codecs.

The complexity can be further reduced by using a sign-shape-structured VQ codebook, where each codevector has its mirror-image vector with regard to the origin as another codevector. As only half the codevectors are independent, the complexity of ZSR calculation is cut in half.

The VQ codebook can be closed-loop optimized [7]. Let *K* be the vector dimension, y_j be the *j*-th codevector, and **H**(*n*) be the *K* × *K* lower triangular Toeplitz matrix with *h*(*k*), *k* = 0, 1, …, *K*-1 as the first column, where *h*(*k*) is the impulse response of the feedback filter structure in Fig. 9. Let N_j be the set of time indices *n* when y_j is the winning codevector in actual encoding of a training file. Then, it can be shown [7] that the centroid condition for updating y_j is

$$\sum_{n\in N_j} g^2(n)\mathbf{H}^T(n)\mathbf{H}(n)y_j = \sum_{n\in N_j} g(n)\mathbf{H}^T(n)qzi(n)$$

During each iteration of the codebook design, solving the above linear equation for y_j will optimize y_j for that iteration. Doing so for all codevectors in the codebook and for many iterations will optimize the entire VQ codebook.

With VQ applied to TSNFC, the resulting codec can compete favorably with CELP codecs in terms of output speech quality, codec complexity, and coding delay. BroadVoice16 and BroadVoice32 are good examples of VQ-based TSNFC codecs that achieve high quality, low complexity, and low delay.

5. CONCLUSION

This paper introduces a new class of speech codecs called Two-Stage Noise Feedback Coding (TSNFC) that utilizes two stages of prediction and two stages of spectral noise shaping. Several novel TSNFC codec structures have been described. It has been shown that vector quantization can be applied to NFC and TSNFC, and the codebook search complexity can be reduced significantly by ZIR and ZSR decomposition. The closed-loop optimization method for the TSNFC residual VQ codebook has been introduced. VQ-based TSNFC codecs, such as BroadVoice16 and BroadVoice32, compete favorably with conventional CELP codecs in terms of speech quality, complexity, and delay.

6. REFERENCES

[1] C. C. Cutler, US Patent No. 2,927,962, "Transmission Systems Employing Quantization," filed April 1954, issued March 1960.

[2] E. G. Kimme and F. F. Kuo, "Synthesis of Optimal Filters for a Feedback Quantization System," *IEEE Trans. Circuit Theory*, pp. 405-413, September 1963.

[3] J. D. Makhoul and M. Berouti, "Adaptive Noise Spectral Shaping and Entropy Coding in Predictive Coding of Speech," *IEEE Trans. Acoust., Speech, Sig. Proc.*, pp.63-73, February 1979.

[4] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. Acoust., Speech, Sig. Proc.*, pp. 247-254, June 1979.

[5] C. C. Lee, "An enhanced ADPCM Coder for Voice Over Packet Networks," *Int'l J. Speech Tech.*, pp. 343-357, May 1999.

[6] I. A. Gerson and M. A. Jasiuk, "Techniques for Improving the Performance of CELP-Type Speech Coders," *IEEE J. Selected Areas in Communications*, pp. 858-865, June 1992.

[7] J.-H. Chen, US Patent Application No. 20000722077, "Method and Apparatus for One-Stage and Two-Stage Noise Feedback Coding of Speech and Audio Signals," filed November 2000.

[8] J.-H. Chen, BroadVoice®16 Speech Codec Specification, Version 1.2, October 2003. (For further information, contact PacketCable, Cable Television Laboratories, Inc.)