HIGH QUALITY ENHANCED WAVEFORM INTERPOLATIVE CODING AT 2.8 KBPS

Oded Gottesman and Allen Gersho

Signal Compression Laboratory, Department of Electrical and Computer Engineering
University of California, Santa Barbara, California 93106, USA
E-mail: [oded, gersho]@scl.ece.ucsb.edu, Web: http://scl.ece.ucsb.edu

ABSTRACT

This paper presents a high quality *Enhanced Waveform Interpolative* (EWI) speech coder at 2.8 kbps. The system incorporates novel features such as: dual-predictive *analysis-by-synthesis* (AbS) quantization of the *slowly-evolving waveform* (SEW), efficient parametrization of the *rapidly-evolving waveform* (REW) magnitude, and AbS *vector quantization* (VQ) of the REW parameter. Subjective tests indicate that its quality exceeds that of G.723.1 at 5.3 kbps, and it is slightly better than that of G.723.1 at 6.3 kbps.

1. INTRODUCTION

In recent years, there has been increasing interest in achieving toll-quality speech coding at rates of 4 kbps and below. Currently, there is an ongoing 4 kbps standardization effort conducted by the ITU-T. The expanding variety of emerging applications for speech coding, such as third generation wireless networks and Low Earth Orbit (LEO) systems, is motivating increased research efforts. The speech quality produced by waveform coders such as *code-excited linear prediction* (CELP) coders [1] degrades rapidly at rates below 5 kbps. On the other hand, parametric coders such as the *waveform-interpolative* (WI) coder [4]-[10], the *sinusoidal-transform coder* (STC) [2], and the *multiband-excitation* (MBE) coder [3] produce good quality at low rates, but they do not achieve toll quality. This is largely due to the lack of robustness of speech parameter estimation, which is commonly done in open-loop, and to inadequate modeling of non-stationary speech segments. In this work we propose a paradigm for WI coding that incorporates *analysis-by-synthesis* (AbS) for parameter estimation, offers higher temporal and spectral resolution for the *rapidly-evolving waveform* (REW), and more efficient quantization of the *slowly-evolving waveform* (SEW).

Commonly in WI coding, the similarity between successive REW magnitudes is exploited by downsampling and interpolation and by constrained bit allocation [5]. In our past EWI coder [12][13], the REW magnitude was quantized on a waveform by waveform base, and at excessive number of bits – more than is perceptually required. Here we propose a novel parametric representation of the REW magnitude and an efficient paradigm for AbS predictive vector quantization of the REW parameter sequence. The new method achieves a substantial reduction in the REW bit rate.

In very low bit rate WI coding, the relation between the SEW and the REW magnitudes was exploited by computing the magnitude of one as the unity complement of the other [5]-[10]. Also, since the sequence of SEW magnitude evolves slowly, succeeding SEWs exhibit similarity, offering opportunities for redundancy removal. Additional forms of redundancy that may be exploited for coding efficiency are: (a) for a fixed SEW/REW decomposition filter, the mean SEW magnitude increases with the pitch period and (b) the similarity between succeeding SEWs, also increases with the pitch period. In this work we introduce a novel "dual-predictive" AbS paradigm for quantizing the SEW magnitude that optimally exploits the information about the current quantized REW, the past quantized SEW, and the pitch, in order to predict the current SEW.

This paper is organized as follows. In Section 2 we explain the REW parameterization, and the corresponding AbS VQ. The dual predictive SEW AbS VQ and its performance are discussed in Section 3. The bit allocation is given in section 4. Subjective results are reported in Section 5. Finally, we summarize our work.

2. REW QUANTIZATION

Efficient REW quantization can benefit from two observations: (1) the REW magnitude is typically an increasing function of the frequency, which suggests that an efficient parametric representation may be used; (2) one can observe similarity between succeeding REW magnitude spectra, which may suggest a potential gain by employing predictive VQ on a group of adjacent REWs. The next three sections introduce the REW parametric representation and the associated VQ technique.

2.1 REW Parameterization

Direct quantization of the REW magnitude is a variable dimension quantization problem, which may result in spending bits and computational effort on perceptually irrelevant information. A simple and practical way to obtain a reduced, and fixed, dimension representation of the REW is with a linear combination of basis functions, such as orthonormal polynomial [8]-[10]. Such a representation usually smoothens the REW magnitude, and improves the perceptual quality. Suppose the REW magnitude, $R(\omega)$, is represented by a linear combination of orthonormal functions, $\psi_i(\omega)$:

This work was supported in part by the University of California MICRO program, ACT Networks, Inc, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., DSP Group, Inc., Fujitsu Laboratories of America, Inc., General Electric Corp., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products NV, Lucent Technologies, Inc., Nokia Mobile Phones, Panasonic Speech Technology Laboratory, Qualcomm, Inc., Sun Microsystems Inc., and Texas Instruments, Inc.

$$R(\omega) = \sum_{i=0}^{I-1} \gamma_i \psi_i(\omega) \quad , \quad 0 \le \omega \le \pi \tag{1}$$

where ω is the angular frequency, and I is the representation order. The REW magnitude is typically an increasing function of the frequency, which, for perceptual considerations, is coarsely quantized with a low number of bits per waveform. Therefore, it may be advantageous to represent the REW magnitude in a simple, but perceptually relevant manner. Suppose the REW is modeled by the following parametric representation, $\hat{R}(\omega,\xi)$:

$$\hat{R}(\omega,\xi) = \sum_{i=0}^{I-1} \hat{\gamma}_i(\xi)\psi_i(\omega) \quad , \quad 0 \le \omega \le \pi \quad ; \quad 0 \le \xi \le 1 \tag{2}$$

where $\hat{\gamma}(\xi) = [\hat{\gamma}_0(\xi),...,\hat{\gamma}_{I-1}(\xi)]^T$ is a parametric vector of coefficients within the representation model subspace, and ξ is the "unvoicing" parameter which is zero for a fully voiced spectrum, and one for a fully unvoiced spectrum.

2.2 Piecewise Linear REW Representation

For practical considerations we may assume that the parametric representation is piecewise linear, and may be represented by a set of N uniformly spaced spectra, $\{\bar{R}(\omega,\bar{\xi}_n)\}_{n=0}^{N-1}$, as illustrated in Figure 1. This representation is similar to the hand-tuned REW codebook in [9][10]. The parametric surface is linearly interpolated in between by:

$$\bar{R}(\omega,\xi) = (1-\alpha)\bar{R}(\omega,\bar{\xi}_{n-1}) + \alpha\bar{R}(\omega,\bar{\xi}_n) \tag{3}$$

$$; \quad \bar{\xi}_{n-1} \le \xi \le \bar{\xi}_n \quad ; \quad \alpha = \frac{\xi - \bar{\xi}_{n-1}}{\Delta} \quad ; \quad \Delta = \bar{\xi}_n - \bar{\xi}_{n-1}$$

From the linearity of the representation:

$$\bar{\gamma}(\xi) = (1-\alpha)\bar{\gamma}_{n-1} + \alpha\bar{\gamma}_n \tag{4}$$

where $\bar{\gamma}_n$ is the coefficient vector of the n-th REW magnitude representation:

$$\bar{\gamma}_n = \bar{\gamma}(\bar{\xi}_n) \tag{5}$$

Suppose for a REW magnitude, $R(\omega)$, represented by some coefficient vector, γ, we search for the parameter value, $\xi(\gamma)$, in $\bar{\xi}_{n-1} \le \xi \le \bar{\xi}_n$, whose respective representation vector, $\bar{\gamma}(\xi)$, minimizes the MSE distortion between the two spectra:

$$D\left(R,\hat{R}(\xi)\right) = \int_0^{\pi} \left| R(\omega) - (1-\alpha)\bar{R}(\omega,\bar{\xi}_{n-1}) - \alpha\bar{R}(\omega,\bar{\xi}_n) \right|^2 d\omega \tag{6}$$

From orthonormality, the distortion is equal to:

$$D\left(R,\hat{R}(\xi)\right) = \left\|\gamma - \bar{\gamma}(\xi)\right\|^2 = \left\|\gamma - (1-\alpha)\bar{\gamma}_{n-1} - \alpha\bar{\gamma}_n\right\|^2 \tag{7}$$

The optimal interpolation factor that minimizes the MSE is:

$$\alpha_{opt} = \frac{(\bar{\gamma}_n - \bar{\gamma}_{n-1})^T(\gamma - \bar{\gamma}_{n-1})}{\left\|\bar{\gamma}_n - \bar{\gamma}_{n-1}\right\|^2} \tag{8}$$

and the respective optimal parameter value, which is a continuous variable between zero and one, is given by:

$$\xi(\gamma) = (1-\alpha_{opt})\bar{\xi}_{n-1} + \alpha_{opt}\bar{\xi}_n \tag{9}$$

This result allows a rapid search for the best unvoicing parameter value needed to transform the coefficient vector to a scalar parameter, followed by the corresponding quantization scheme, as described in the next section.

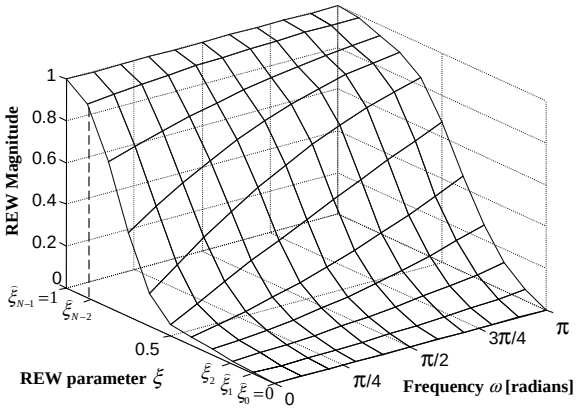

Figure 1. REW Parametric Representation $\hat{R}(\omega,\xi)$

2.3 REW Quantization

The encoder maps the REW magnitude to an unvoicing parameter, and then quantizes the parameter by AbS VQ, as illustrated in Figure 2. Initially, the magnitudes of the M REWs in the frame are mapped to coefficient vectors, $\{\gamma(m)\}_{m=1}^{M}$. Then, for each coefficient vector, a search is performed to find the optimal representation parameter, $\xi(\gamma)$, using equation (9), to form an M-dimensional parameter vector for the current frame, $\{\xi(\gamma(m))\}_{m=1}^{M}$. Finally, the parameter vector is encoded by AbS VQ. The decoded spectra, $\{\hat{R}(\omega,\hat{\xi}(m))\}_{m=1}^{M}$, are obtained from the quantized parameter vector, $\{\hat{\xi}(m)\}_{m=1}^{M}$, using equation (3). This scheme allows for higher temporal as well as spectral REW resolution, since no downsampling is performed, and the continuous parameter is vector quantized in AbS.

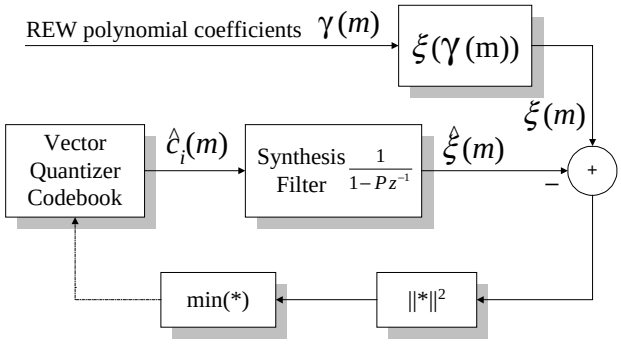

Figure 2. REW Parametric Representation AbS VQ.

3. AbS SEW QUANTIZATION

Figure 3 illustrates a Dual Predictive SEW AbS VQ scheme which uses the quantized REW as well as the past quantized SEW to predict the current SEW. Suppose $|\hat{\mathbf{r}}_M|$ denotes the spectral magnitude vector of the last quantized REW in the current frame. An "implied" SEW vector is calculated by:

$$\left|\hat{\mathbf{s}}_{M,implied}\right| = 1 - \left|\hat{\mathbf{r}}_M\right| \quad (10)$$

and from which the mean vector is removed. The mean removed vectors are denoted by apostrophe. Then, we compute a (mean-removed) estimated "implied" SEW magnitude vector, $\left|\tilde{\mathbf{s}}'_{M,implied}\right|$, using a diagonal estimation matrix $\mathbf{P}_{REW}$,

$$\left|\tilde{\mathbf{s}}'_{M,implied}\right| = \mathbf{P}_{REW}\left|\hat{\mathbf{s}}'_{M,implied}\right| \quad (11)$$

Additionally, a "self-predicted" SEW vector is computed by multiplying the delayed quantized SEW vector, $\left|\hat{\mathbf{s}}'_0\right|$, by a diagonal prediction matrix $\mathbf{P}_{SEW}$. The predicted (mean-removed) SEW vector, $\left|\tilde{\mathbf{s}}'_M\right|$, is given by:

$$\left|\tilde{\mathbf{s}}'_M\right| = \mathbf{P}_{REW}\left|\hat{\mathbf{s}}'_{M,implied}\right| + \mathbf{P}_{SEW}\left|\hat{\mathbf{s}}'_0\right| \quad (12)$$

The quantized vector, $\hat{\mathbf{c}}_M$, is determined in AbS by:

$$\hat{\mathbf{c}}_M = \underset{\mathbf{c}_i}{\operatorname{argmin}}\left\{ \left(|\mathbf{s}'_M| - |\tilde{\mathbf{s}}'_M| - \mathbf{c}_i\right)^T \mathbf{W}_M \left(|\mathbf{s}'_M| - |\tilde{\mathbf{s}}'_M| - \mathbf{c}_i\right) \right\} \quad (13)$$

where $\mathbf{W}_M$ is the diagonal spectral weighting matrix [11]-[13]. The (mean-removed) quantized SEW magnitude, $\left|\hat{\mathbf{s}}'_M\right|$, is the sum of the predicted SEW vector, $\left|\tilde{\mathbf{s}}'_M\right|$, and the codevector $\hat{\mathbf{c}}_M$:

$$\left|\hat{\mathbf{s}}'_M\right| = \left|\tilde{\mathbf{s}}'_M\right| + \hat{\mathbf{c}}_M \quad (14)$$

In order to exploit the information about the pitch, and the voicing level, we have partitioned the possible pitch range into six subintervals, and the REW parameter range into three, and generated eighteen codebooks, one for each pair of pitch range and unvoicing range. Each codebook has associated two mean vectors, and two diagonal prediction matrices. To improve the coder robustness and the synthesis smoothness, the cluster used for the training of each codebook overlaps with those of the codebooks for neighboring ranges. Since each quantized target vector may have a different value of the removed mean, the quantized mean is added temporarily to the filter memory after the state update, and the next quantized vector's mean is subtracted from it before filtering is performed.

The output weighted SNR, and the mean-removed weighted SNR, of the scheme are illustrated in Figure 4. Evidently, a very high SNR is achieved with a relatively small number of bits. The weighted SNR of each codebook, for the 9-bit case, is illustrated in Figure 5. The differences in SNR between three REW parameter ranges is dominated by the different means. The respective mean-removed weighted SNR of each codebook is illustrated in Figure 6. Within each voicing range, the differences in SNR between each pitch range, are mainly due to the number of bits per vector sample, which decreases as the number of harmonics increases, and to the prediction gain.

Example for the two predictors for three REW parameter ranges is illustrated in Figure 7. For voiced segment the SEW predictor is dominant, whereas the REW predictor is less important since its input variations in this range are very small. As the voicing decreases, the SEW predictor decreases, and the REW predictor becomes more dominant at the lower part of the spectrum. Both predictors decrease as the voicing decreases from the intermediate range to the unvoiced range.

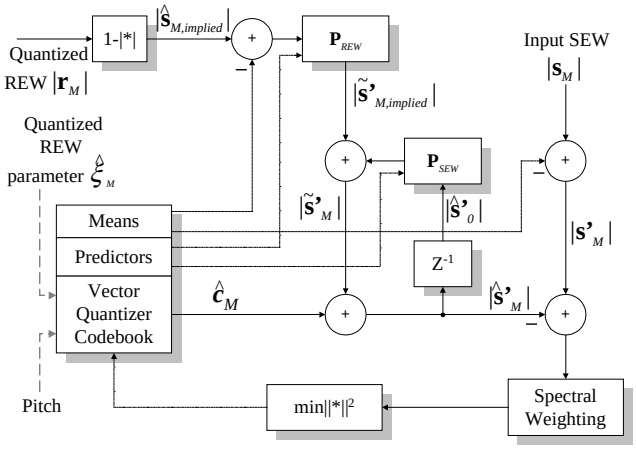

Figure 3. Block diagram of the Dual Predictive AbS SEW vector quantization.

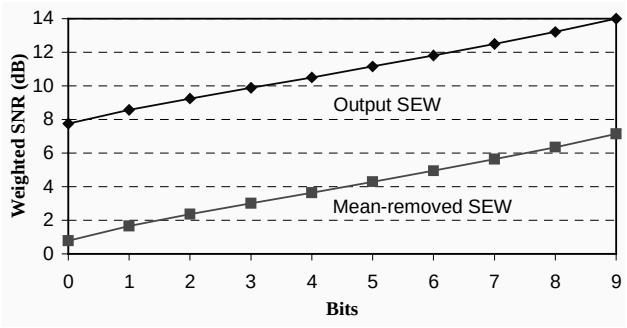

Figure 4. Weighted SNR for Dual Predictive AbS SEW VQ

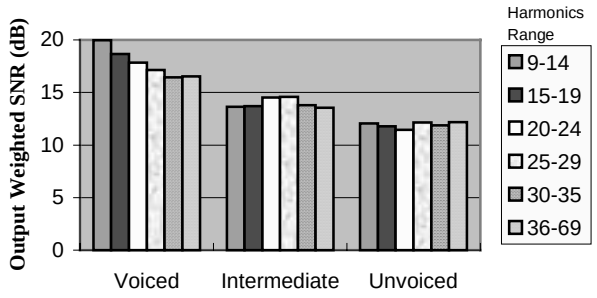

Figure 5. Output Weighted SNR for the 18 codebooks, 9-bit AbS SEW VQ

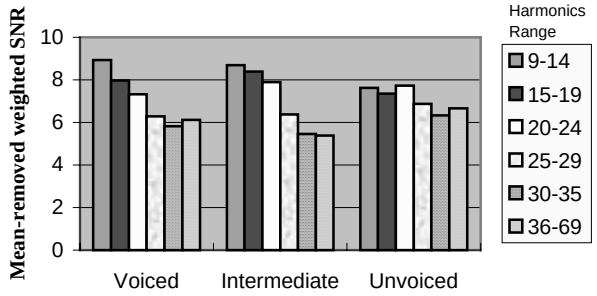

Figure 6. Mean-removed SEW's Weighted SNR for the 18 codebooks, 9-bit AbS SEW VQ

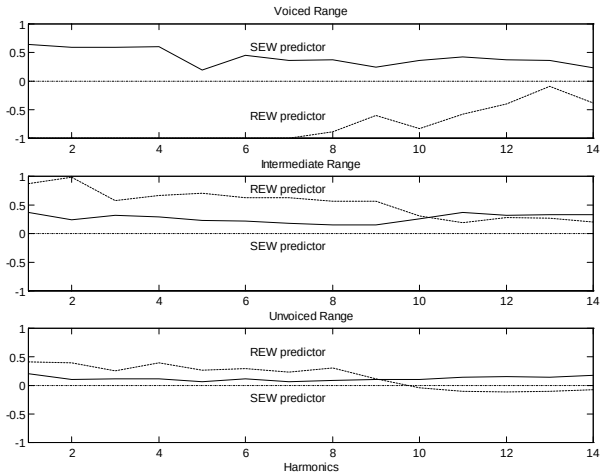

Figure 7. Predictors for three REW parameter ranges.

4. BIT ALLOCATION

The bit allocation for the 2.8 kbps EWI coder is given in Table 1. The frame length is 20 ms, and ten waveforms are extracted per frame. The *line spectral frequencies* (LSFs) are coded using predictive MSVQ, having two stages of 10 bit each, a 2-bit increase compared to the past version of our coder [12][13]. The *10*-th dimensional log-gain vector is quantized using 9 bit AbS VQ [12][13]. The pitch is coded twice per frame. A fixed SEW phase was trained for each one of the eighteen pitch-voicing ranges [11].

Parameter	Bits / Frame	Bits / second
LPC	20	1000
Pitch	2x6 = 12	600
Gain	9	450
SEW magnitude	8	400
REW magnitude	7	350
Total	56	2800

Table 1. Bit allocation for 2.8 kbps EWI coder

5. SUBJECTIVE RESULTS

We have conducted a subjective A/B test to compare our 2.8 kbps EWI coder to the G.723.1. The test data included 24 *modified intermediate reference system* (M-IRS) [14] filtered speech sentences, 12 of which are of female speakers, and 12 of male speakers. Twelve listeners participated in the test. The test results, listed in Table 2 and Table 3, indicate that the subjective quality of the 2.8 kbps EWI exceeds that of G.723.1 at 5.3 kbps, and it is slightly better than that of G.723.1 at 6.3 kbps. The EWI preference is higher for male than for female speakers.

Test	2.8 kbps WI	5.3 kbps G.723.1	No Preference
Female	40.28%	33.33%	26.39%
Male	48.61%	24.31%	27.08%
Total	44.44%	28.82%	26.74%

Table 2. Results of subjective A/B test for comparison between the 2.8 kbps EWI coder to 5.3 kbps G.723.1. With 95% certainty the result lies within +/-5.53%.

Test	2.8 kbps WI	6.3 kbps G.723.1	No Preference
Female	38.19%	36.81%	25.00%
Male	43.06%	31.94%	25.00%
Total	40.63%	34.38%	25.00%

Table 3. Results of subjective A/B test for comparison between the 2.8 kbps EWI coder to 6.3 kbps G.723.1. With 95% certainty the result lies within +/-5.59%.

6. SUMMARY

We have found several new techniques that enhance the performance of the WI coder, and allow for better coding efficiency. The most significant of these, reported here, dual-predictive AbS quantization of the SEW, efficient parametrization of the REW magnitude, and AbS VQ of the REW parameter. Subjective test results indicate that the performance of the 2.8 kbps EWI coder slightly exceeds that of G.723.1 at 6.3 kbps and therefore EWI achieves very close to toll quality, at least under clean speech conditions.

7. REFERENCES

[1] B. S. Atal, and M. R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rate," *Proc. Int. Conf. Comm, Amsterdam*, pp. 1610-1613, 1984.

[2] R. J. McAulay, and T. F. Quatieri, "Sinusoidal Coding," *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 4, pp. 121-173, 1995.

[3] D. Griffin, and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. ASSP*, Vol. 36, No. 8, pp. 1223-1235, August 1988.

[4] Y. Shoham, "High Quality Speech Coding at 2.4 to 4.0 kbps Based on Time-Frequency-Interpolation," *IEEE ICASSP'93*, Vol. II, pp. 167-170, 1993.

[5] W. B. Kleijn, and J. Haagen, "A Speech Coder Based on Decomposition of Characteristic Waveforms," *IEEE ICASSP'95*, pp. 508-511, 1995.

[6] W. B. Kleijn, and J. Haagen, "Waveform Interpolation for Coding and Synthesis," *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 5, pp. 175-207, 1995.

[7] I. S. Burnett, and D. H. Pham, "Multi-Prototype Waveform Coding using Frame-by-Frame Analysis-by-Synthesis," *IEEE ICASSP'97*, pp. 1567-1570, 1997.

[8] W. B. Kleijn, Y. Shoham, D. Sen, and R. Haagen, "A Low-Complexity Waveform Interpolation Coder," *IEEE ICASSP'96*, pp. 212-215, 1996.

[9] Y. Shoham, "Very Low Complexity Interpolative Speech Coding at 1.2 to 2.4 kbps," *IEEE ICASSP'97*, pp. 1599-1602, 1997.

[10] Y. Shoham, "Low-Complexity Speech Coding at 1.2 to 2.4 kbps Based on Waveform Interpolation," *International Journal of Speech Technology, Kluwer Academic Publishers*, pp. 329-341, May 1999.

[11] O. Gottesman, "Dispersion Phase Vector Quantization for Enhancement of Waveform Interpolative Coder," *IEEE ICASSP'99*, vol. 1, pp. 269-272, 1999.

[12] O. Gottesman and A. Gersho, "Enhanced Waveform Interpolative Coding at 4 kbps," *IEEE Speech Coding Workshop*, pp. 90-92, 1999, Finland.

[13] O. Gottesman and A. Gersho, "Enhanced Analysis-by-Synthesis Waveform Interpolative Coding at 4 kbps," EUROSPEECH'99, pp. 1443-1446, 1999, Hungary.

[14] ITU-T, "Recommendation P.830, Subjective Performance Assessment of Telephone Band and Wideband Digital Codecs," Annex D, ITU, Geneva, February 1996.