A ROBUST VARIABLE-RATE SPEECH CODER

A. Shen, B. Tang, A. Alwan, and G. Pottie

Department of Electrical Engineering, UCLA
405 Hilgard Ave.
Los Angeles, CA 90024

ABSTRACT

The goal of this study is to develop a robust and high-quality speech coder for wireless communication. The proposed coder is a perceptually-based variable-rate subband coder. The perceptual metric ensures that encoding is optimized to the human listener and is based on calculating the signal-to-mask ratio in short-time frames of the input signal. An adaptive bit allocation scheme is employed and the subband energies are then quantized using a Max-Lloyd quantizer. The coder is fully scalable–increasing the bit rates, improves the quality of encoded speech. Subjective listening tests, using quiet and noisy input signals, indicate that the proposed coder produces high-quality speech when operating at 12 kbps or higher. In error-free conditions, our coder has comparable performance to that of QCELP or GSM coders. For speech in background noise, however, our coder, at 12 kbps, outperforms QCELP significantly, and for music, it outperforms both QCELP and GSM.

1. INTRODUCTION

Speech codec design is typically driven by bandwidth-efficiency considerations; CELP-based coders, for example, are popular because of their low bit rates. The performance of these coders, however, is poor for female speech and for non-speech signals, such as music; in addition, the performance deteriorates significantly in the presence of background noise. As a result, new standards for personal communication services are likely to use high-quality, medium bit-rate speech codecs, such as that proposed in this paper. With the increasing use of wireless communication devices and consumer need for high-quality services, robustness of the speech coder for varying channel conditions and speaker differences becomes increasingly important.

2. SYSTEM OVERVIEW

The proposed encoder consists of four components: analysis/synthesis filterbanks, a perceptual model, a bit allocation block, and a quantization block. The coder processes input frames of 160 samples (20 ms at 8 kHz). The analysis/synthesis filterbank is an 8-channel IIR QMF bank. The perceptual model estimates the SNR required to mask quantization noise for transparent coding and the bit allocation scheme translates the SNR prescribed by the model into a bit assignment used for quantizing the subband samples. Finally, the subband energies are quantized.

Filterbank Design: Although FIR QMF banks are commonly implemented in subband coders because of their linear phase properties, IIR QMF banks are more computationally efficient. For our perceptual coder, an 8-channel tree-structured IIR QMF bank was implemented; the filterbank is a modified version of the 4-channel design given in [9]. The amount of phase distortion depends mainly on the order of the filter and transition bandwidth required. Informal listening tests of reconstructed tones and several TIMIT sentences, with no quantization, indicate that the 7th and 9th order filters provide reconstructed signals which are free from audible phase distortions. Because of the transparent quality and low complexity, 7th order elliptic filters were chosen for the 8-channel QMF banks. These filters provide over 60 dB attenuation in the stop bands; a minimum of 40 dB stop-band attenuation is typically used in speech coding applications [3]. The alias-free property of the QMF bank becomes invalid when quantization is performed between the analysis and synthesis filterbanks. Moreover, the amount of aliasing distortion is directly related to the number of bits used to quantize the subband signal. Careful proportional bit allocation, however, can minimize the audibility of this effect.

Masking Estimation: The metric which has been shown to be perceptually relevant to the quality of coded speech is the signal-to-mask ratio (SMR) [1] [1]. This metric determines the amount of noise which is masked in each subband due to in-band and out-of-band masking. If the 'critical' SMR is achieved in all frequency bands, then the reconstructed signal should sound identical to the original. An estimate of quantization-noise masking by the signal spectrum is performed for every input frame. For masking analysis, the input samples are windowed by a 160 point Hanning window and an FFT is performed on this sequence. The spectral estimate is performed efficiently via a subband FFT; this approach is described in [12]. The FFT components are then used to approximate the spectral energy density in each frame. The magnitude and frequency axes

Supported in part by NSF grant IRI-9309418 and ARPA/CSTO Contract J-FBI-93-112. Albert Shen is now at Intel. Inc., OR. The coder is described in detail in his thesis [11].

[1] SMR is the signal-to-noise ratio (SNR) between the speech signal and the noise masking threshold.

0-7803-2431-5/95 $4.00 © 1995 IEEE

of the magnitude spectrum are transformed into dimensions which are more closely related to the characteristics of the human auditory system. First, the magnitude values of the FFT spectrum are transformed to dB SPL using calibration curves; the calibration curves were obtained by measurements. Second, frequency is transformed into the critical band rate, or Bark scale, by integrating the power of the spectral components over the 17 bark bands which comprise the 4 kHz signal bandwidth. The perceptual spectra generated can now be used to estimate the masking curve for this frame of input speech. To estimate the masking thresholds, the ISO/MPEG psychoacoustic model [10] is convolved with the perceptual spectrum. First, a local noise masking curve is calculated from the energy in each of the 17 Bark bands. Next, an overall noise-masking curve is estimated by power summing the masked power curves due to individual Bark bands [14]. Groups of Bark bands are then integrated to approximate the eight 500Hz channels of the QMF bank. The maximum Bark signal level and minimum Bark noise masking level in each channel are chosen as the levels for the SMR (in dB) for effective quantization noise masking in that channel [2].

Bit Allocation: Bit allocation is a two-tiered process. First, a perceptual measure is used to estimate the channel SNR required for the signal to mask the quantization noise. Second, the estimate is used by the bit-allocation block to assign bits to quantize the samples in each subband. Three bit-allocation schemes were implemented and evaluated: uniform bit allocation, reverse water-filling bit allocation, and proportional bit allocation. Uniform bit allocation, in which all subband frames are quantized with the same number of bits, does not permit allocating more bits to frames where more SNR is required for noise masking. Reverse water-filling, such as that used in the MPEG standard, assigns bits to the channel which requires the most SNR first. The major drawback of this scheme is that channels with low SNR requirements may be given few or no bits for quantization; this would result in disproportionate amounts of noise in the coded speech. To overcome these limitations, we developed a bit allocation scheme which provided the best perceptual quality when compared to the uniform and reverse water-filling schemes.

The proportional bit allocation scheme allocates bits to each channel in proportion to the prescribed SNR for masking. If there are enough bits to meet the SNR requirement for the frame, the bits are assigned directly. Otherwise, the allocation is done according to the following equation:

$$Bits[n] = Bits_{rqd}[n] * \left(\frac{BPF}{Bits_{rqd-all}}\right) \qquad (1)$$

where $Bits[n]$ is the number of bits assigned to channel n, $Bits_{rqd}[n]$ is the number of bits calculated by the perceptual block for noise-masking in channel n, BPF is the total bits per frame (determined from the coding rate), and $Bits_{rqd-all}$ is the total number of bits needed to meet the SNR requirements for all subbands for that particular frame. For example, for coding at 16 kbps with an input frame of 160 samples, if the calculated $Bits_{rqd}[n]$ values were [80, 80, 80, 40, 80, 60, 60, 40], then, from Eq. (1) the proportional bit allocation would scale the bit allocation vector to [49, 49, 49, 24, 49, 36, 36, 24]. Any remaining bits are allocated evenly across the channels starting with channel 0. In informal listening tests, an allocation of 4 bits per subband sample provided transparent speech coding. Hence, there was a maximum limit of 4 bits per sample (20 samples per band) for quantizing subband samples. All remaining bits are allocated evenly to the other channels starting from the low-frequency channel.

Quantization: Two quantizers were implemented and evaluated: a uniform quantizer (PCM) and a Max-Lloyd quantizer. The Max quantizer was superior, both perceptually and in the MMSE sense, to the uniform quantizer. This result is expected since Max quantizers are less sensitive to the source PDF than PCM quantizers. To design the Max quantizer which takes into account the statistical distribution of the input signal, analysis was performed on the subband speech signals to be quantized. A set of subband signals were generated from TIMIT sentences and processed through the 8-channel QMF bank. When normalized using the subband signal variances, the average subband sample distribution resembled a Gaussian (Fig. 1). Hence, a Max quantizer using a Gaussian table was designed and implemented in the perceptual subband coder. The quantizer was made embedded by considering the effect of truncating the LSBs of the index. This effectively results in the union of two adjacent levels. The optimal reconstruction value is given by the centroid over the union of these two levels. Reconstruction values can be derived in this manner until all bits have been truncated. Embedded Max quantizers differ from non-embedded ones in that only the reconstruction rules, in the event of bit truncation, are changed. It should be noted that the embedded Max quantizer used is different than that implemented in [3]. That approach results in a suboptimum quantizer at full rates. In our approach, the quantizer is optimum at full rates, and yields optimum reconstruction values if bits are lost or packets are dropped in transmission.

3. PERFORMANCE EVALUATION

Traditionally, the performance of encoding systems has been evaluated using the SNR criteria. For encoded speech and audio signals, however, the SNR is a poor indicator of distortion in the coded signals. Signals with high SNR may contain significant levels of audible distortion, whereas signals with moderate to low SNR may contain noise levels that are not perceptible [1]. Hence, subjective tests were used to evaluate our coder (8-channel, 7th order IIR QMF bank, a perceptual model, proportional bit allocation, and embedded Max quantizer.)

Methodology: Four subjects participated in these tests. Training sets were used to familiarize the subjects with the types and degrees of coding distortion in the listening tests. A total of 24 sentences (2 male and 2 female talkers), from the TIMIT database, were coded at six different bit rates (ranging from 4 kbps to 32 kbps). Signals were presented monaurally to simulate audition with a telephone handset or a portable radio unit. The sentences were presented at levels ranging between 80-90 dB SPL and the additive road noise measured at 85 dB SPL. All signals were 4 sec in length. Two sequences were generated for the listening experiments. The first sequence consisted of the sentences coded without background noise and the second, consisted of the same sentences with road noise. The road noise, provided by QUALCOMM Inc., was recorded in an automobile traveling at highway speeds with the windows rolled up. This was added to the speech, and the resulting signal was coded at the six rates.

Listening Test Results: A five-point MOS scale was used; the scale reflects the 'noise' qualities: very annoying (1), annoying (2), slightly annoying (3), perceptible but not annoying (4), and imperceptible (5) [8]. Average results of the listening experiments for coded speech in quiet and in background noise at the six coding rates are shown in Fig. 2. The coder operating at 32 kbps received the highest average MOS for both conditions. As the coder bit rate decreased from 3 to 1 bit per sample (24-8 kbps), the scores decreased as well. The highly-distorted speech at 4 kbps scored poorly and consistently with or without road noise. In the presence of road noise, the clear distinction between 32 and 24 kbps diminishes (MOS difference drops an average of 0.31 points). At the same time, the MOS score for the 8-24 kbps coders increased in the presence of road noise. It appears that, at the SPL levels we used, distortions in these coders were masked by road noise and became more difficult to detect. These results indicate that at 12 kbps our coder can provide good speech quality (MOS above 3).

We also conducted subjective tests to compare the performance of our coder to that of the QCELP coder (the speech service option for wideband spread spectrum digital cellular system -EIA/TIA/IS-96-) at 8kbps and a GSM 6.10 coder (a standardized lossy speech compression algorithm employed by most European wireless telephones [6, 13]) at 13 kbps. In error-free conditions, our coder at 12 kbps had similar performance to both GSM and QCELP. In the presence of road noise, however, the average MOS score dropped by 70% for QCELP, and by 10% for GSM while the MOS score for our coder was not affected. We also conducted a subjective test using two 4 sec segments of baroque music (*Water Music* by Handel and *Autumn Concerto* by Vivaldi.) The average MOS scores in that case were 1, 2.1, 2.7, and 4.0 for QCELP, GSM, our coder at 12 kbps, and our coder at 16 kbps, respectively.

4. SUMMARY AND DISCUSSION

The purpose of this research is to develop a robust speech coder for wireless communications. A perceptual metric was used to minimize audible distortions. Our implementation of the perceptual metric is different than that suggested by the MPEG Audio Standard [2] in three ways. First, spectral analysis in the proposed coder is performed on subband signals rather than on the signal directly; thereby reducing computational complexity. Second, we implemented a proportional bit allocation scheme and an embedded Max quantizer while MPEG uses reverse-water-filling bit allocation and either a uniform quantizer (Layers I and II) or a simple non-uniform quantizer (Layer III). Third, we used a low-delay IIR QMF bank as opposed to MPEG's FIR polyphase filterbanks. The cost function used in the design of the IIR filters is the audibility of phase distortion

When integrated into a subband coding system, the perceptual model adaptively estimates the audibility of quantization noise. More resources can then be allocated to the frequency bands where a greater SNR is needed to mask quantization noise. At 12 kbps, the proposed coder achieves high quality speech. The coder is fully scalable and its performance improves with increasing bit rates. Based on studies reported in the literature on the performance of CELP-based coders [4] [5] and the results of our subjective listening tests, our coder offers several advantages. First, while the performance of the CELP-based coders degrades significantly in background noise, the performance of our coder does not. Second, our coder provides fully scalable, variable rate/variable quality coding for both speech and non-speech signals such as music. While CELP-based coders may be adequate for telephonic applications, future applications such as multimedia personal communication systems will demand high-quality speech, under varying channel conditions, which our coder could provide.

Future work will include a lower delay, wide-band implementation of the coder. In addition, we plan to develop an embedded VQ scheme which is based on a variable-length scalar quantizer. The performance of the VQ scheme will be compared with that of the Max quantizer.

To improve the performance of perceptually-based subband coders further, more complete and quantitative models of speech perception are needed.

5. REFERENCES

[1] Brandenburg, K. and T. Sporer. "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria." Audio Engineering Society Test and Measurement Conference, 1992.

[2] Brandenburg, K. and G. Stoll. "The ISO/MPEG Audio Codec: A Generic Standard for Coding of High Quality Digital Audio." Audio Engineering Society Preprint 3336, 1992.

[3] Cox, R.; Gay, S.; Shoham, Y.; Quackenbush, S.; Seshadri, N.; Jayant, N. "New Directions in Subband Coding," *JSAC*, vol. 6, Feb., 1988.

[4] Furui, S. *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker, Inc. New York, 1989.

[5] Gersho, A. "Advances in Speech and Audio Compression." *IEEE Proceedings*, June, 1994.

[6] ETSI/GSM, GSM 06.10, "GSM Full Rate Transcoding", July 1989.

[7] Jayant, N.S. and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[8] Jayant, N. S. "Signal Compression: Technology Targets and Research Directions." *JSAC*, Vol. 10, No. 5, 1992.

[9] Jiang, Z.; Alwan, A.; and A.N. Willson Jr. "High-Performance IIR QMF Banks for Speech Subband Coding." *Proc. IEEE ISCAS*, June 1994.

[10] MPEG (ISO/IEC JTC1/SC2 WG11). Audio Codec Specifications (11172-3). November, 1991.

[11] Shen, A. "Perceptually-based subband coding of speech signals," Unpublished Master's thesis, Dept. of Electrical Engineering, UCLA, June, 1994.

[12] Tang, B.; Shen, A.; Pottie, G.; and Alwan, A. "Spectral Analysis of Subband Filtered Signals," ICASSP '95 (this proceedings.)

[13] P. Vary et al., "Speech codec for the European Mobile System", Proc. ICAASP, p.227, April 1988.

[14] Veldhuis, R.N. J. "Bit Rates in Audio Source Coding." *JSAC*, vol. 10, no. 1, 1992.

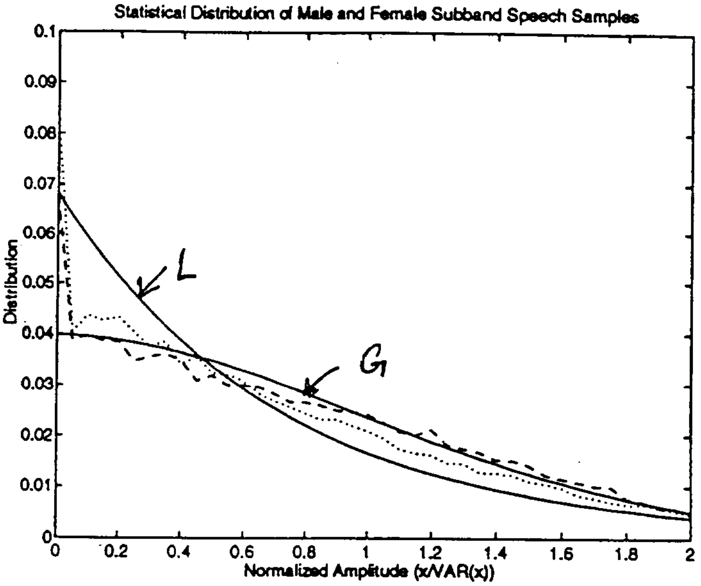

Figure 1: Distribution of the normalized amplitudes of the subband samples of male speech (dotted); female speech (dashed); Laplacian (L) distribution; and a Gaussian (G) distribution.

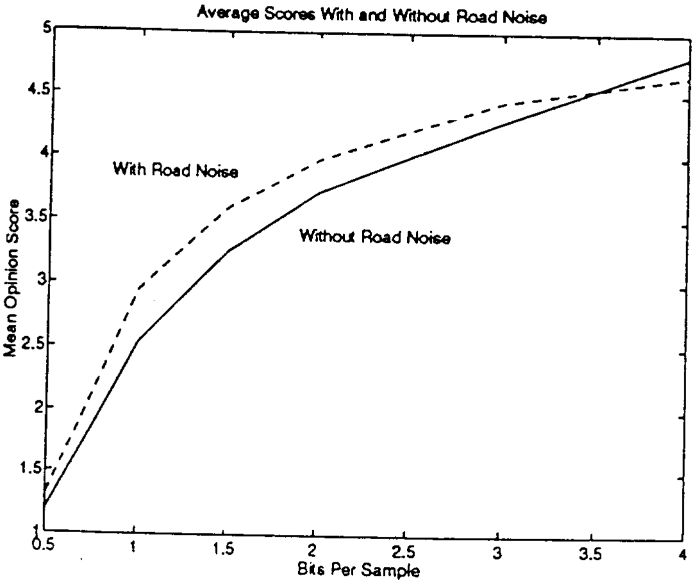

Figure 2: Average MOS for the coder at rates 4-32 kbps (.5-4 bits per sample). Dashed/solid lines are for speech coded with/without road noise.