MULTIPLE-DESCRIPTION PCM SPEECH CODING BY COMPLEMENTARY ASYMMETRIC VECTOR QUANTIZERS

Stephen D. Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado 80305, USA, svoran@its.bldrdoc.gov

ABSTRACT

We describe new 2-channel multiple-description speech coders based on the ITU-T Recommendation G.711 PCM speech coder. The new coders operate in the PCM code domain in order to exploit the companding gain of PCM. They apply pairs of complementary asymmetric 2-dimensional vector quantizers to each pair of PCM codes, thus exploiting the correlation between adjacent speech samples. If both quantizer outputs are received (two channels working), they are combined to generate an approximation to the original pair of PCM codes. If only one quantizer output is received (one channel failed, one channel working), a coarser approximation is still possible. The vector quantizers use rectangular cells, and the aspect ratio of the cells controls the speech-quality trade-off between the two-channel and one-channel cases.

INTRODUCTION

It is often necessary to transmit speech signals over lossy communication channels. Important examples of lossy channels include noisy and fading radio channels (as in wireless telephony) and congested packet data networks (as in Internet telephony). Receiver-based packet loss concealment (PLC) algorithms can be used to reduce the effects of short-duration channel losses on received speech quality. After a 60-90 ms gap in the received speech stream, these algorithms mute or strongly attenuate their outputs because they cannot conceal longer losses.

Multiple-description coding (MDC) offers a different way to gain robustness to channel losses and MDC is effective for both long and short losses. The original theory of MDC is set out in [1]-[2] and examples of additional development and applications can be found in [3]-[7]. In MDC an encoder forms multiple partial descriptions of a signal and these descriptions are sent over different physical or virtual channels. If all descriptions arrive at the decoder, a high quality reconstruction is available. If any descriptions are lost, a lower-quality reconstruction is produced.

Our earlier work in this area includes the development of 2-channel multiple-description speech coders [7] that extend the international standard for Pulse Code Modulation (PCM) speech coding, ITU-T Recommendation G.711 [8]. The extension is inserted between a PCM speech encoder and decoder as shown in Figure 1.

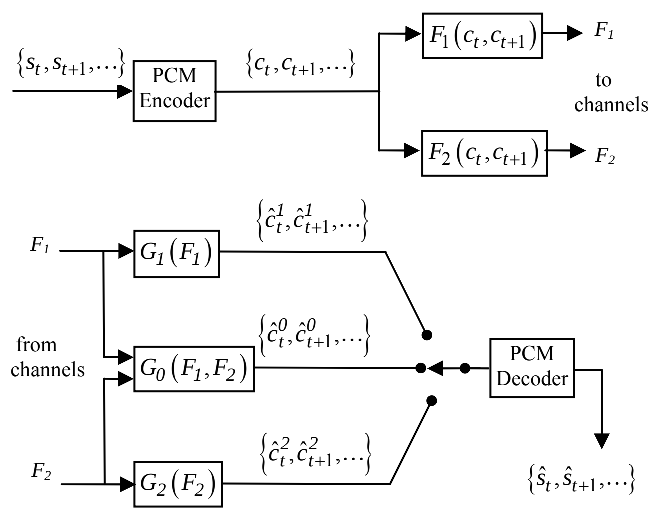

Figure 1. Block diagram for multiple-description PCM speech encoding and decoding.

This approach is motivated by three main principles. First, by incorporating the PCM encoder and decoder, we can reap the benefits of PCM companding. Second, by applying a vector quantizer (VQ) to each pair of successive PCM codes (c_t,c_{t+1}) (generated from a pair of successive speech samples (s_t,s_{t+1})), we can exploit the correlation between adjacent speech samples to reduce quantization noise and/or data rate. Finally, by developing paired VQ's, we can generate a pair of descriptions $F_1(c_t,c_{t+1})$ and $F_2(c_t,c_{t+1})$ for each pair of PCM codes (c_t,c_{t+1}). In general, each description F_1 and F_2 carries some information about both c_t and c_{t+1}. Thus if only F_1 or F_2 is received, a coarse reconstruction of the PCM code-pair is possible. By picking these VQ's appropriately, we can ensure that when both of the descriptions F_1 and F_2 are received, they can be combined to generate a more refined reconstruction of the PCM code-pair. The symbols F_1 and F_2 will refer to the two VQ's, their associated outputs, or their associated partitions of the PCM code plane, depending on context.

In [7] we developed the VQ's F_1 and F_2 so that each partitions the PCM code plane into a regular grid using square cells of size $w \times w$. Further, we offset the grids of F_1 and F_2 by $w/2$ in each dimension, so that the new VQ defined by intersecting the cells of F_1 and F_2 has a regular grid and square cells of size $w/2 \times w/2$ as shown in Figure 2. This halving of cell dimensions results in a 6 dB reduction of quantization noise in the PCM code domain:

$$10\log_{10}\left(\frac{\varepsilon_0^2}{\varepsilon_k^2}\right)=10\log_{10}\left(\frac{\mathrm{E}\left(\hat{c}_t^0-c_t\right)^2}{\mathrm{E}\left(\hat{c}_t^k-c_t\right)^2}\right)=-6\,\mathrm{dB},\ k=1,2\,. \qquad (1)$$

COMPLEMENTARY ASYMMETRIC VECTOR QUANTIZERS

A more general case uses a pair of complementary VQ's each with asymmetric (i.e., rectangular) cells. The cells of the VQ F_1 have size $wr^{+0.5} \times wr^{-0.5}$ (width × height) and the cell aspect ratio is r. The cells of the VQ F_2 have size $wr^{-0.5} \times wr^{+0.5}$ and are offset from the cells of F_1 by $\frac{1}{2}wr^{-0.5}$ in dimension. In F_2 the cell aspect ratio is r^{-1}. An example of this rectangular partitioning for the case r=4 is given in Figure 3.

For either VQ, the geometric mean of the cell dimensions is w and the cell area is w^2. The cell area determines the number of cells required to partition the 255×255 (A-law) or 256×256 (μ-law) PCM code plane and thus it also determines the channel bit-rate requirement. Since the cell area is independent of r, the bit-rate is also independent of r. We will see that the aspect ratio r can thus be used to make speech quality trade-offs while keeping the total bit-rate constant.

The work in [7] is the important special case r=1 of this more general case. Equation (1) shows that the case r=1 gives a fixed 6 dB difference between the one- and two-channel cases. By increasing the aspect ratio we can increase the speech quality when both channels are working (relative to the r=1 two-channel speech quality) if we are willing to accept a decrease in the speech quality when only one channel is working (relative to the r=1 one-channel speech quality). This may be desirable for mitigating the effects of long but infrequent channel losses.

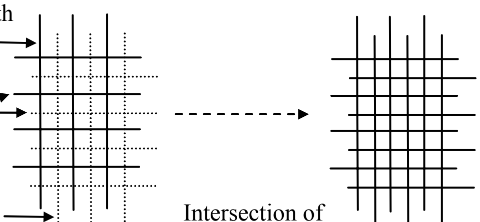

Figure 2. Example of intersecting two offset partitions to create a new partition.

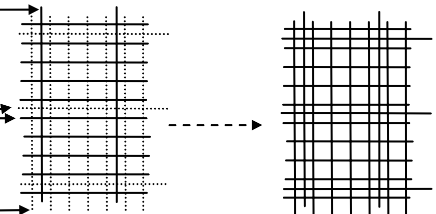

Figure 3. Example of intersecting two asymmetric offset partitions to create a new partition, aspect ratio r=4.

This behavior of the aspect ratio r is intuitive. For convenience in this discussion, associate the x-direction of the PCM code-plane with speech samples and PCM codes with even-numbered time indices, and the y-direction with those having odd-numbered time indices. If r is large, the VQ F_1 generates, and channel 1 carries, little information about the even samples (VQ cells are wide) and lots of information about the odd samples (VQ cells are short). The opposite is true for VQ F_2 and channel 2.

This means that when either channel has failed, the other channel working alone can generate lower quality speech (aliasing becomes prominent as we approach the case of 2:1 subsampling without the requisite preliminary lowpass filtering). When both channels are working, they together provide enough information to generate higher quality speech. Thus for large values of r the two-channel speech quality can be dramatically higher than the one-channel speech quality. As $r \to 1$, each channel carries more balanced information regarding the odd and even time samples and this means that either channel alone can generate medium quality speech. In this case however, there is less to be gained by combining the information carried by the two channels. Thus the two-channel speech quality is only slightly higher than the one-channel speech quality as $r \to 1$.

More specifically, when only one channel is operating, the PCM code domain quantization step size is $wr^{-0.5}$ for half of the PCM codes (e.g., those with odd time indices) and $wr^{+0.5}$ for the other half of the PCM codes. Thus the average quantization noise power in the PCM code domain is

$$\varepsilon_k^2 = \mathrm{E}\left(\hat{\boldsymbol{c}}_t^{\boldsymbol{k}} - \boldsymbol{c}_t\right)^2 = cw^2 \frac{1}{2}\left(\frac{1}{r} + r\right), \; k = 1,2\,, \quad (2)$$

for some constant c.

To analyze the case where both channels are operating we restrict r to integer values so that the smaller cell dimension $wr^{-0.5}$ will evenly divide the larger cell dimension $wr^{+0.5}$. Due to the offset of $\frac{1}{2}wr^{-0.5}$ between the two VQ's, each large quantizer step will be divided into $r+1$ (rather than r) smaller quantizer steps; r-1 of these have size $wr^{-0.5}$ and 2 of them have size $\frac{1}{2}wr^{-0.5}$. Assuming a uniform probability density for the data across the larger cell dimension, this gives an average quantization noise power in the PCM code domain of

$$\varepsilon_0^2 = \mathrm{E}\left(\hat{\boldsymbol{c}}_t^{\boldsymbol{0}} - \boldsymbol{c}_t\right)^2 = cw^2\left(\frac{r-1}{r}\frac{1}{r} + \frac{1}{r}\frac{1}{4r}\right) = cw^2\left(\frac{r - \frac{3}{4}}{r^2}\right). \quad (3)$$

Thus ε_1^2 and ε_2^2 increase with r, while ε_0^2 decreases for all integer values of $r \geq 2$. The noise power ratio in dB (both channels operating referenced to one channel operating) is

$$10\log_{10}\left(\frac{\varepsilon_0^2}{\varepsilon_k^2}\right) = 10\log_{10}\left(\frac{2\left(r - \frac{3}{4}\right)}{\left(r^3 + r\right)}\right), \; k=1,2\,. \quad (4)$$

This result is shown in Figure 4. Increasing r from 1 to 2 causes ε_0^2, ε_1^2, and ε_2^2 to each increase by a factor of 5/4 (1.0 dB). Thus the ratio (4) is unchanged when r changes from 1 to 2. Equation (4) reduces to (1) when r=1 and as r gets large it reduces to

$$10\log_{10}\left(\frac{\varepsilon_0^2}{\varepsilon_k^2}\right) = 3 - 20\log_{10}(r), \; k=1,2\,. \quad (5)$$

Results (2) and (3) indicate that if channel failures are infrequent, then larger values of r are desirable, but if channel failures are frequent, then smaller values of r are desirable. If it is known a priori that the probability of only one channel working is α and the probability of both channels working is $(1-\alpha)$, then we can appropriately weight results (2) and (3) to find the average quantization noise power in the PCM code domain:

$$\xi^2 = \alpha\frac{\boldsymbol{cw}^2}{2}\left(\frac{1}{r} + r\right) + (1-\alpha)\boldsymbol{cw}^2\left(\frac{r - \frac{3}{4}}{r^2}\right). \quad (6)$$

This relationship is shown in Figure 5, and one could identify a minimizing value of r for any fixed value of α. Figure 5 indicates again that as channel failures become less likely, larger values of r become optimal.

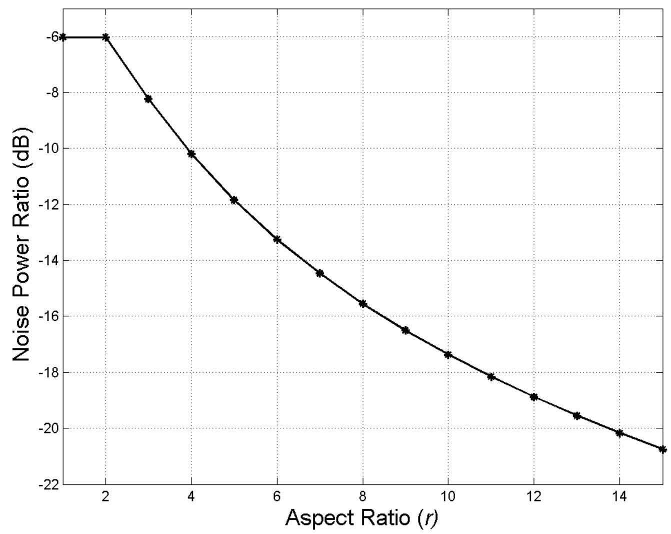

Figure 4. Ratio of two-channel noise power to one-channel noise power in PCM code domain as a function of the aspect ratio r.

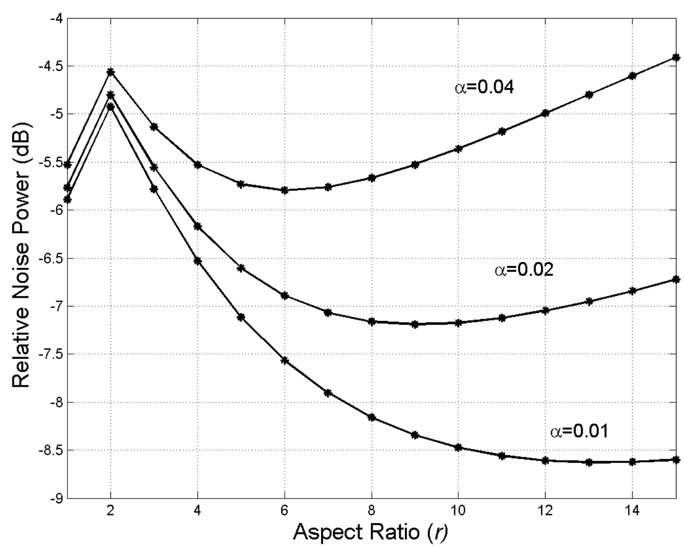

Figure 5. Relative noise power as a function of aspect ratio r for single channel failure probabilities of 1, 2, and 4%.

SPECIFIC DESIGNS

We have designed a family of multiple-description PCM speech coders that use complementary asymmetric VQ's. The starting point for VQ designs is the distribution of the data to be quantized. Figure 6 provides a contour plot representation of a smoothed histogram of μ-law PCM code-pairs. This histogram was generated from 40 different English sentences taken from the Harvard phonetically-balanced sentence lists [9]. Two female and two male talkers each provided ten sentences for a total of approximately two minutes of speech. Consistent with standard G.711 PCM operation, the speech was bandpass-filtered (300-3400 Hz) and adjusted to an active speech level of 26 dB below overload before PCM encoding.

Due to correlations between adjacent speech samples, this histogram takes the value zero over approximately half of the PCM code-plane. This indicates that about half of the code-pairs will appear very infrequently in practice. The histogram for A-law PCM is similar. From these histograms one could use conventional techniques to design VQ's that minimize mean-squared error (MSE). But a VQ design driven by MSE would effectively result in a non-optimized speech companding law. As expected, our experiments indicate that uniform quantization of PCM codes generates the highest speech quality. Thus we generally use a fixed VQ cell size across the entire region (R_1) where the histogram is non-zero. We use a single larger cell size in the region (R_0) where the histogram is zero and PCM code-pairs will rarely appear, thus exploiting the correlation between sequential PCM codes in order to reduce data rate.

Our specific designs use b=4, 5, and 6 bits/sample/channel. These rates correspond to total data rates of 32, 40, and 48 kbps when one channel is working and 64, 80, and 96 kbps when two channels are working.

First we describe the case b=4 bits/sample/channel with r=1. In this case both VQ's use a 13×13 cell size in R_1 (w=13) and a 26×26 cell size in R_0 (w=26). We elected to use a cell dimension ratio (between cells of R_0 and cells of R_1) of 2 and then calculated the necessary cell sizes to give approximately 2^{2b} = 256 cells. We then slightly adjusted the definition of R_0 (from "histogram = 0" to "histogram < ε") to get precisely 256 cells. Boundary conditions will generally require a few cells of other sizes as well. Throughout this section, this technique is used to make bit rates exact.

The VQ F_1 uses a partition with a cell centered on the origin of the PCM code-plane. (We define the origin of the PCM code-plane to be (129,129) for A-law and

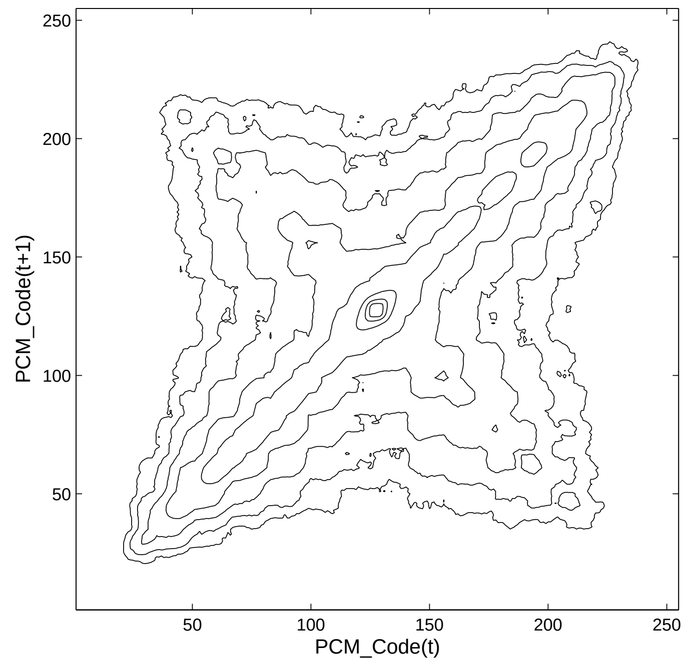

Figure 6. Contours of smoothed histogram of μ-law PCM code-pairs.

(128,128) for μ-law.) The VQ F_2 is the same except for shifts of 7 PCM codes in each dimension in the region R_1 and shifts of 13 PCM codes in each dimension in the region R_0.

As r gets larger, the VQ F_1 uses cells that are wider $\left(wr^{+0.5}\right)$ and shorter $\left(wr^{-0.5}\right)$ as demonstrated in Figure 3. The VQ F_2 uses cells that are narrower $\left(wr^{-0.5}\right)$ and taller $\left(wr^{+0.5}\right)$. In addition, offsets of $\frac{1}{2}wr^{-0.5}$ are maintained between the two VQ partitions in each direction so that intersecting the partitions generates the greatest possible number of cells. In practice, VQ cell sizes and offsets are constrained to integer values so these cannot always be exactly $wr^{+0.5}$, $wr^{-0.5}$, and $\frac{1}{2}wr^{-0.5}$, but rather they take approximately these values. In the following, we report r as the actual F_1 cell aspect ratio (i.e., the ratio of the integer cell width to the integer cell height).

For the case b=4 bits/sample/channel, we consider r=1.0, 3.6, 6.8, and 10.5. The associated cell sizes in R_1 (width×height) for the VQ F_1 are 13×13, 25×7, 34×5, and 42×4. The associated cell sizes in R_1 for F_2 are 13×13, 7×25, 5×34, and 4×42. Thus when either channel is working alone, there is a mixture of larger and smaller quantization noises associated with alternating time samples. When both channels are working, the effective quantization cell sizes are 13×13, 7×7, 5×5, and 4×4 respectively. As in the case r=1, when moving from R_1 to R_0 we simply scale all dimensions by a factor of 2.

For the case b=5 bits/sample/channel, we use r=1.0, 4.0, and 9.0. The associated cell sizes in R_1 for F_1 are 6×6, 12×3, and 18×2 respectively. For the case b=6 bits/sample/channel, we use r=1.0, 2.0, and 9.0. The associated cell sizes in R_1 for F_1 are 3×3, 4×2, and 9×1. All dimensions are scaled by a factor of 2 in R_0.

For all of the VQ's the representation point or reconstruction point associated with each cell in R_1 is the centroid (using the PCM code-pair histogram) of that cell. For any cell in R_0, we define the representation point to be the geometric center of that cell.

RESULTING SPEECH QUALITY

Equation (6) and Figure 5 indicate that average quantization noise power can be minimized when using complementary asymmetric VQ's to perform MDC. In some applications this may be sufficient for optimal performance. In the present application this only guarantees that the average noise power in the PCM code domain is minimized. The relationship between this average noise power and perceived speech quality is a somewhat indirect one. In addition, it is unclear whether averaging accurately reflects how listeners perceive the quality of speech signals that move between lower and higher quality levels. Thus, minimization of (6) will not necessarily lead to the highest perceived speech quality in this application.

Ultimately, the perceived speech quality will depend not only on the probability of a channel failure α, but also on its temporal statistics (e.g., the lengths of the failures and to what extent they occur randomly or in bursts). This makes generalized testing and optimization beyond the scope of this paper. Instead we have conducted a listening experiment to investigate the perceived speech quality of the ten multiple-description PCM speech coder designs described above. For each of the ten designs we consider three cases: channel 1 working, channel 2 working, and both channels working. This gives a total of 30 cases. As expected, the results for the two cases "channel 1 working" and "channel 2 working" are very close and hence they have been combined here. The 20 resulting cases are described in the first 3 columns of Table 1.

We conducted the listening experiment in an environment consistent with the specifications given in [10]. The experiment used sentences from phonetically-balanced

Table 1. Mean equivalent speech quality results for 20 multiple-description PCM speech coding conditions.

b (bits/sample/ channel)	r (aspect ratio)	Number of Channels Working	Equivalent PCM Speech Quality (bits/sample)
4	1.0	1	4.63
4	3.6	1	4.52
4	6.8	1	4.38
4	10.5	1	4.32
5	1.0	1	5.75
5	4.0	1	5.04
5	9.0	1	4.77
6	1.0	1	6.54
6	2.0	1	6.29
6	9.0	1	5.68
4	1.0	2	5.29
4	3.6	2	5.43
4	6.8	2	6.07
4	10.5	2	6.14
5	1.0	2	6.21
5	4.0	2	6.79
5	9.0	2	6.86
6	1.0	2	7.00
6	2.0	2	7.07
6	9.0	2	7.93

sentence lists [9], and eight different talkers (four female and four male). Listeners were given the task of matching the overall perceived speech quality of the various conditions under test with one of six different reference conditions. For each case, each listener played different recordings until that listener determined a match to one of the reference conditions. The six reference conditions were created using a modified version of conventional, single-description G.711 PCM that uses 4.0. 4.5, 5, 6, 7, or 8 bits/sample. Thus the experiment generated results in terms of equivalent PCM speech quality using units of bits/sample. There was no overlap between the material used in this listening experiment and the material used to design the vector quantizers.

Column 4 of Table 1 and Figure 7 show the results of this experiment in mean equivalent bits of G.711 PCM speech quality. These results come from a total of 14 trials using 11 different listeners. The half-widths of the 95% confidence intervals about the mean values shown in Figure 7 range from 0.00 to 0.26 with a mean and median value of 0.15 bits/sample. In Figure 7, the three lines with upward trends show that perceived speech quality does

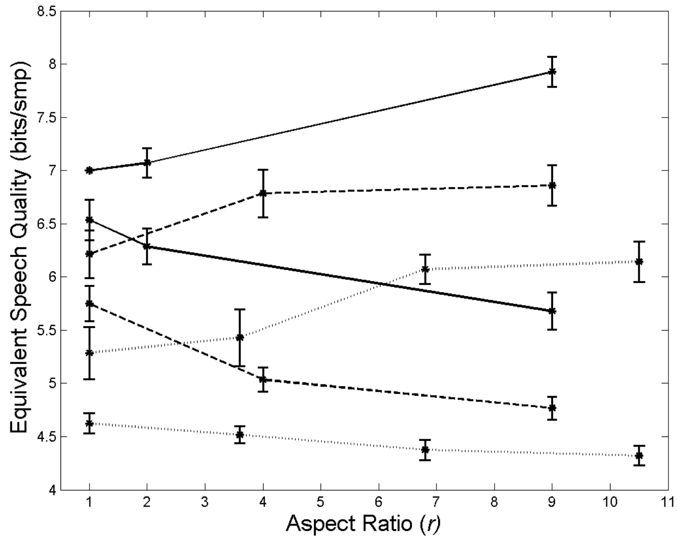

Figure 7. Equivalent speech quality results for 20 multiple-description PCM speech coding conditions. Dotted, dashed, and solid lines represent the cases *b*=4, 5, and 6 bits/sample/channel respectively. Means and 95% confidence intervals are shown.

increase with *r* when two channels are working. The three lines with downward trends show speech quality decreasing with *r* when only one channel is working.

For the case *r*=1 with one channel working, we see a 0.5 to 0.8 bit/sample increase in equivalent speech quality over conventional PCM due to the use of VQ's that exploit adjacent sample correlation. There is an additional 0.5 to 0.7 bit/sample increase in equivalent speech quality when two channels are combined. For the case *b*=4, the total bit rate is 64 kbps which matches that of conventional G.711 PCM. With *r*=1, this MDC technique can deliver 4.6 or 5.3 bits/sample of equivalent speech quality when one or two channels are working respectively. When *r* is increased to 10.5, these equivalent speech qualities drop to 4.3 and increase to 6.1 bits/sample respectively. For the case *b*=6, the total bit rate is 96 kbps which is 150% of the rate for conventional G.711 PCM. With *r*=1, the technique described here can deliver 6.5 or 7.0 bits/sample of equivalent speech quality when one or two channels are working respectively. When *r* is increased to 9.0, these equivalent speech qualities drop to 5.7 and increase to 7.9 bits/sample respectively.

Several objective estimators of perceived speech quality have been developed and verified over the years. While most have been verified for the quantization noise associated with PCM, no results were previously available for the case where odd and even samples could have different levels of quantization noise (i.e., the case $1<r$ when only one channel is working). Because of this, a listening experiment was the only known reliable way to generate reliable perceived speech-quality values. Once these values were obtained, we compared them with results from three objective estimators of perceived speech quality: Segmental SNR [11], a Measuring Normalizing Blocks algorithm [12], and the Perceptual Evaluation of Speech Quality algorithm [13]. We applied these estimators to all of the recordings used in the listening experiment including the PCM reference recordings. This allowed us to translate the objective estimates to equivalent G.711 PCM speech quality measured in bits/sample. With the listening experiment results and the objective estimates both on this common scale, we then compared them. Across the 20 conditions described above, each of the three estimators has a coefficient of correlation to the listening experiment results that is greater than 0.99. In spite of this high correlation, all three of the objective estimators did tend to slightly underestimate perceived speech quality in numerous cases, especially at the lower end of the speech quality scale. These errors never exceed about 0.4 bits/sample. We conclude that any of these three objective estimators can provide useful, but not error free, estimates of perceived speech quality under these conditions.

DISCUSSION

In some communication systems, channel losses are inevitable. Receiver-based PLC algorithms typically attempt to conceal losses shorter than 60-90 ms. They do not require any increase in the data rate, and they do not reduce speech quality when there are no losses. If channels present longer losses significantly often, then PLC will not suffice and MDC may be an appropriate solution. One could invoke lower-rate coders to accomplish MDC with no increase in data rate. If the complexity of lower-rate coding must be avoided, then the mitigation of longer losses will require some sacrifice in the form of either increased data rate or decreased speech quality. The multiple-description PCM speech coding techniques described here offer both options.

These techniques can provide major speech quality improvements for channels with long-duration losses and minor speech quality reductions for lossless channels. They may also require increases in total data rate. They can be implemented by inserting a set of look-up tables between a conventional PCM encoder and decoder; no mathematical computations are required. One must look up each pair of PCM codes (16-bit look-up) resulting in 2 descriptions. If only one description arrives at the receiver, a single 2*b*-bit look-up will generate a coarse approximation to the original pair of PCM codes. If both descriptions arrive at the receiver, then a pair of 2*b*-bit look-ups is required and a finer approximation will result.

Adjusting the VQ cell aspect ratio allows one to trade-off one-channel speech quality against two-channel speech quality. When channel losses are less frequent, two-channel speech quality becomes more important and this drives designs towards higher aspect ratios. When channel losses are more frequent, one-channel speech quality becomes more important and this drives designs towards lower aspect ratios.

Example recordings of speech that has been encoded and decoded using techniques described here are available at its.bldrdoc.gov/audio/pubs_talks/sdvqpcm_examples.php.

REFERENCES

[1] A. El Gammal and T. Cover, "Achievable Rates for Multiple Descriptions," *IEEE Trans. Information Theory,* vol. IT-28, pp. 851-857, Nov. 1982.

[2] L. Ozarow, "On a Source Coding Problem with Two Channels and Three Receivers," *Bell System Technical J.*, vol. 59, pp. 1909-1921, Dec. 1980.

[3] V. Vaishampayan, "Design of Multiple Description Scalar Quantizers," *IEEE Trans. Information Theory*, vol. 39, pp. 821-834, May 1993.

[4] R. Arean, J. Kovačević, and V. Goyal, "Multiple Description Perceptual Audio Coding with Correlating Transforms," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 140-145, Mar. 2000.

[5] S. Voran, "The Channel-Optimized Multiple-Description Scalar Quantizer," in *Proc. 10th IEEE Digital Signal Processing Workshop*, Pine Mountain, Georgia, USA, Oct. 2002.

[6] H. Dong, A. Gersho, J. Gibson, and V. Cuperman, "A Multiple Description Speech Coder based on AMR-WB for Mobile ad hoc Networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* Montreal, May 2004.

[7] S. Voran, "A Multiple-Description PCM Speech Coder using Structured Dual Vector Quantizers," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* Philadelphia, Mar. 2005.

[8] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," Geneva, 1988.

[9] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 225-246, Sep. 1969.

[10] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," Geneva, 1996.

[11] N. Kitawaki, "Quality Assessment of Coded Speech," in *Advances in Speech Signal Processing*. S. Furui and M. Sondi, Eds., New York: Marcel Dekker, 1992.

[12] S. Voran, "Objective Estimation of Perceived Speech Quality, Part I: Development of the Measuring Normalizing Block Technique," *IEEE Trans. Speech and Audio Proc.*, vol. 7, pp. 371-382, Jul. 1999.

[13] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual Evaluation of Speech Quality (PESQ) – The New ITU Standard for End-to-End Speech Quality Assessment, Part II – Psychoacoustic Model," *J. Audio Eng. Soc.*, vol. 50, pp. 765-778, Oct. 2002.