A MULTIPLE-DESCRIPTION PCM SPEECH CODER USING STRUCTURED DUAL VECTOR QUANTIZERS

Stephen D. Voran

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

ABSTRACT

We describe a 2-channel multiple-description speech coder based on the ITU-T Recommendation G.711 PCM speech coder. The new coder operates in the PCM code domain in order to exploit the companding gain of PCM. It applies a pair of 2-dimensional structured vector quantizers to each pair of PCM codes, thus exploiting the correlation between adjacent speech samples. If both quantizer outputs are received, they are combined to generate an approximation to the original pair of PCM codes. If only one quantizer output is received, a coarser approximation is still possible. When using 6 bits/sample/channel (for a total data rate of 96 kbps) the coder provides an equivalent PCM speech quality of 7.3 bits/sample when both channels are working and 6.4 bits/sample when one channel is working.

1. 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 typically mute or strongly attenuate their outputs because they cannot even attempt to 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]. 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.

This paper describes a new 2-channel multiple-description speech coder that extends the international standard for Pulse Code Modulation (PCM) speech coding, ITU-T Recommendation G.711 [8]. This extension is inserted between a PCM speech encoder and decoder as shown in Figure 1. The extension uses PCM code statistics, so it can be described as a source-aware channel coder. We call G.711 PCM speech coding with this extension the Structured Dual Vector Quantizer PCM speech coder or SDVQ-PCM.

SDVQ-PCM works with A-law or μ-law companding and can be implemented with very low complexity, using only look-up tables and no mathematical computations. SDVQ-PCM offers a rate-distortion trade-off. We describe five versions of SDVQ-PCM that generate b=4, 5, 6, 7, and 8 bits/sample/channel and we identify a version that has, when compared to conventional PCM, a major speech quality improvement for lossy channels, a minor speech quality reduction for lossless channels, and a modest total data rate increase.

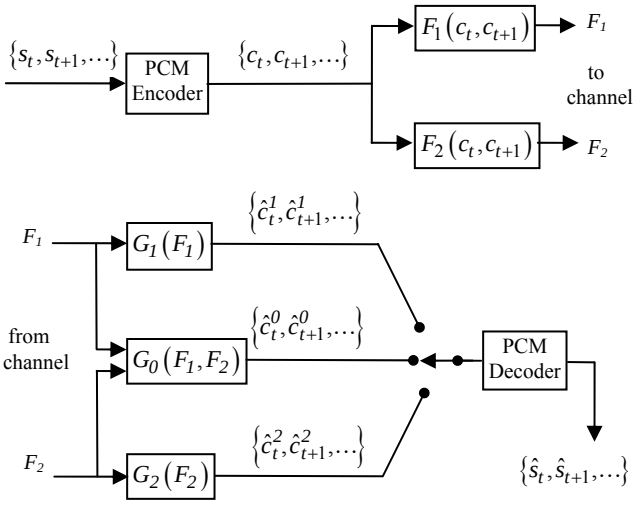

Fig. 1. Block diagram for SDVQ-PCM encoder and decoder.

2. SDVQ-PCM PRINCIPLES

Figure 1 is a block diagram of SDVQ-PCM encoding and decoding. There are three main principles behind SDVQ-PCM design. First, by incorporating the PCM encoder and decoder, we can reap the benefits of PCM companding. This very basic form of perceptual coding minimizes perceivable quantization noise by distributing small quantization noises to the small speech signal samples and larger quantization noises to the larger samples. The result is a higher perceived speech quality than that provided by uniform quantization at the same data rate.

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. This helps to combat the data rate increase that is inherent when replacing one description with two descriptions.

Finally, by developing a pair of VQs 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}). Each description F_1 and F_2 carries coarse 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 forcing an appropriate structure on this pair of VQs, we can ensure that when both of the coarse descriptions F_1 and F_2 are received, they can be combined to generate a more refined reconstruction of the PCM code-pair.

3. SDVQ-PCM DEVELOPMENT

To design an SDVQ-PCM speech coder we must find an appropriate pair of 2-dimensional VQs $G_1\left(F_1\left(c_t,c_{t+1}\right)\right)$ and $G_2\left(F_2\left(c_t,c_{t+1}\right)\right)$ along with a combining function $G_0\left(F_1\left(c_t,c_{t+1}\right),F_2\left(c_t,c_{t+1}\right)\right)$. We view any PCM codeword c_t as an integer

$$c_t \in C=\{1,2,...,c_{max}\},\ c_{max}=256 \text{ (A-law) or } 255 \text{ (}\mu\text{-law)}. \quad (1)$$

When SDVQ-PCM uses b bits/sample/channel, F_1 (or F_2) is a function that defines a partition of the PCM code-plane $C\times C$ into 2^{2b} different cells. We can also view the values F_1 and F_2 as integers that point to these cells:

$$F_1\left(c_t,c_{t+1}\right),\ F_2\left(c_t,c_{t+1}\right)\in\left\{1,2,3,...,2^{2b}\right\}. \quad (2)$$

This allows F_1 or F_2 to describe any PCM code-pair (c_t, c_{t+1}) using $2b$ bits (equivalently b bits/code). G_1 and G_2 then associate a representation point (i.e., a PCM code-pair) with each cell of the appropriate partition

$$G_k\left(F_k\left(c_t,c_{t+1}\right)\right)=(\hat{c}_t^k,\hat{c}_{t+1}^k)\in C\times C,\ k=1,2. \quad (3)$$

Together F_1 and G_1 define a VQ operating on the PCM code-plane. F_1 defines a partition of that plane and G_1 assigns a single representation point to each cell in the partition. In the same way, F_2 and G_2 define a second VQ.

The combining function $G_0\left(F_1\left(c_t,c_{t+1}\right),F_2\left(c_t,c_{t+1}\right)\right)$ associates a representation point (i.e., a single PCM code-pair) $(\hat{c}_t^0,\hat{c}_{t+1}^0)$ with each possible pairing of cell numbers (F_1, F_2):

$$G_0\left(F_1\left(c_t,c_{t+1}\right),F_2\left(c_t,c_{t+1}\right)\right)=(\hat{c}_t^0,\hat{c}_{t+1}^0)\in C\times C. \quad (4)$$

Together F_1, F_2, and G_0 define a third VQ operating on the PCM code-plane. For example, with t fixed, $F_1(c_t, c_{t+1}) = n$ indicates that (c_t, c_{t+1}) is in the n^{th} cell of the partition defined by F_1. $F_2(c_t, c_{t+1}) = m$ indicates (c_t, c_{t+1}) is in the m^{th} cell of the partition defined by F_2. Thus (c_t, c_{t+1}) must be in the intersection of these two cells, and G_0 assigns a representation point to this new cell formed by the intersection of those two cells.

3.1. Design Considerations

The starting point for VQ designs is the distribution of the data to be quantized. Figure 2 contains 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 PCM operation, speech was bandpass-filtered (300-3400 Hz) and adjusted to an active speech level of 26 dB below overload before PCM encoding.

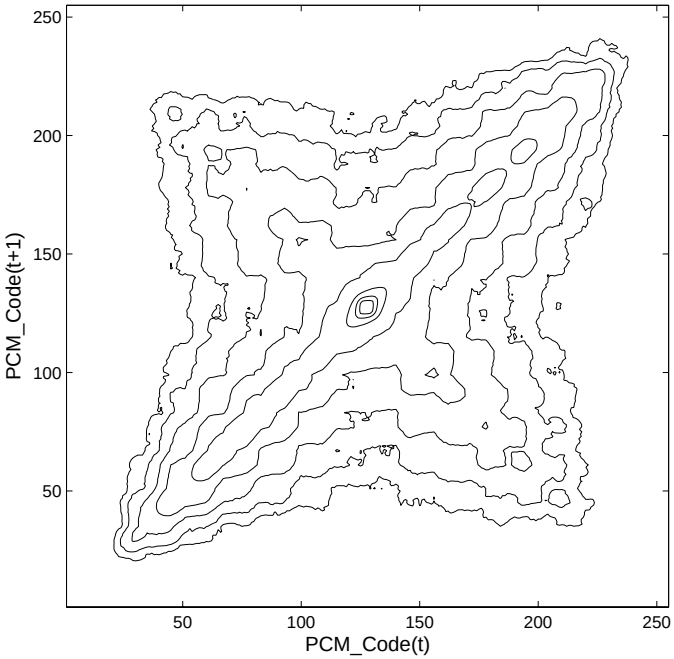

Fig. 2. Contours of smoothed histogram of μ-law PCM code-pairs.

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 VQs 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. We have considered VQs for PCM code m-tuples with m>2. But correlation falls off rapidly and complexity increases exponentially with m, and we expect little benefit from extending SDVQ-PCM into higher dimensions.

An efficient SDVQ-PCM system also requires that the VQ partitions F_1 and F_2 relate to each other so that $G_0\left(F_1\left(c_t,c_{t+1}\right),F_2\left(c_t,c_{t+1}\right)\right)$ carries as much information about (c_t, c_{t+1}) as possible for a given rate constraint. $G_0\left(F_1,F_2\right)$ is another VQ with cells defined by the intersections of the cells of F_1 and F_2. The rate-distortion theory for MDC [1-2] states that for the case of an independent, identically-distributed (iid) Gaussian source and a squared-error distortion measure, when the individual "side coders" $G_1\left(F_1\right)$ and $G_2\left(F_2\right)$ are near their respective rate-distortion limits and each has a distortion of ε^2, then the "central

coder" $G_0(F_1,F_2)$ will have a distortion of at least $\varepsilon^2/2$ (i.e. distortion is reduced by 3 dB at most)

We force structure on F_1 and F_2 so that each partitions the PCM code-plane into a regular grid using 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 will have a regular grid and cells of size $w/2 \times w/2$ as shown in Figure 3. This halving of cell size results in a 6 dB reduction of quantization noise for each sample or a data rate reduction of about 1 bit/sample/channel,

$$10\log_{10}\left(\frac{\mathrm{E}\left(\hat{c}_t^0 - c_t\right)^2}{\mathrm{E}\left(\hat{c}_t^k - c_t\right)^2}\right) = -6\,\mathrm{dB},\ k=1,2\,. \qquad (5)$$

We cannot directly compare this 6 dB result for coding of PCM codes with the 3 dB bound for the coding of an iid Gaussian source with rate-distortion limited "side coders." Yet these two results suggest that SDVQ-PCM may be a relatively efficient approach.

3.2. Resulting Designs

We have designed SDVQ-PCM speech coders using b=4, 5, 6, 7, and 8 bits/sample/channel. In the case b=4 bits/sample/channel, the VQs use a 13×13 cell size in R_1 and a 26×26 cell size in R_0. 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 require a few cells of other sizes as well.) F_1 defines 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 (128,128) for μ-law.) F_2 mimics F_1 except for a shift of 7 PCM codes in both dimensions in the region R_1 and a shift of 13 PCM codes in both dimensions in the region R_0.

The third VQ has cells that are the intersections of the cells defined by F_1 and F_2. There are 950 such cells indicating that the reception of both F_1 and F_2 will give us approximately $\log_2(950)$ = $9.9 \approx 2(b+1)$ bits/sample or $(b+1)$ bits/sample/channel of information. Here we see again the approximate 1 bit or 6 dB improvement in distortion when both channels can be used.

To complete these three VQs we need three sets of representation points that will be stored in the functions G_0, G_1, and G_2. For each cell in R_1, we define the representation point to be the centroid (using the PCM code-pair histogram) of that cell. For each cell in R_0, we define the representation point to be the geometric center of that cell.

The case b=5 is similar. We use a 6×6 cell size (and a shift of 3) in R_1 and a 12×12 cell size (and a shift of 6) in R_0 for a total of about 2^{2b}=1024 cells.

The case b=6 is a bit more involved because cell sizes that are small odd numbers do not generate uniform intersecting cells, and hence do not give the expected 1 bit of gain when the two channels are combined. Thus we elected to use two different cell sizes in R_1. In regions of R_1 where the histogram is greater than approximately 10^{-3}, we use a cell size of 2×2; in the remainder of R_1, we use a cell size of 4×4. Throughout R_0 we use cells of size 6×6. The result is about 2^{2b}=4096 cells. We shift F_2 relative to F_1 in both dimensions by 1, 2, or 3 PCM codes in each of the appropriate regions. Intersecting the cells defined in F_1 and F_2 results in 15,650 cells, indicating that the reception of both F_1 and F_2 will give us approximately $\log_2(15{,}650) = 13.9 \approx 2(b+1)$ bits of information.

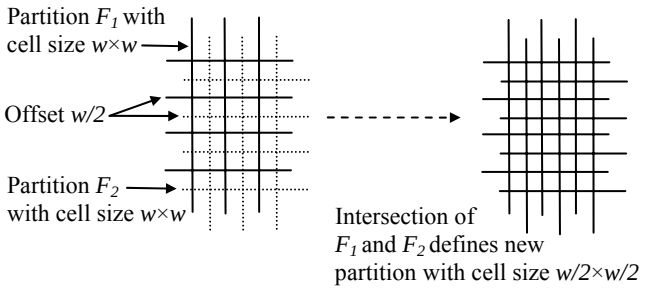

Fig. 3. Intersecting two offset partitions to create a new partition.

For b=7 we simply partition the entire plane into 2^{14}, 2×2 cells. Intersecting F_1 and F_2 results in 2^{16}, 1×1 cells, thus allowing exact recovery of (c_t, c_{t+1}). The case b=8 is the trivial case of 100% redundancy. There is no need to consider pairs of PCM codes; one simply sends each 8-bit PCM code twice, once on each channel.

4. SDVQ-PCM RESULTS

We have evaluated the speech quality of the SDVQ-PCM speech coder using 128 English sentences taken from the Harvard phonetically-balanced sentence lists [9]. The sentences come from 4 female and 4 male talkers for a total of approximately 8 minutes of speech. The speech was filtered and level-adjusted as described previously. There was no overlap between this speech and the speech used to design the VQs.

The dominant impairment introduced by SDVQ-PCM is PCM-like coding noise. Both simple and sophisticated objective quality estimators can closely track perceived speech quality for this simple impairment. We have applied three objective estimators to the 128 SDVQ-PCM processed sentences: Segmental SNR (SNRseg) [10], a Measuring Normalizing Block (MNB) algorithm [11], and the Perceptual Evaluation of Speech Quality (PESQ) algorithm [12]. The results for μ-law SDVQ-PCM measured with SNRseg are shown in Figure 4. The MNB and PESQ estimators give results that agree with the SNRseg results to within ±0.1 bits.

The two dash-dot lines in Figure 4 show SNRseg values when only one of the two channels is working. The solid line shows the improved performance when both channels are working. The figure includes conventional μ-law PCM results for reference. These were obtained from G.711 PCM with codes uniformly requantized in order to operate at 4, 4.5, 5, … 8 bits/sample.

By equating SDVQ-PCM and PCM SNRseg values we can find equivalent conventional PCM bit rates for SDVQ-PCM. For example, the SDVQ-PCM case with b=6 and 2 channels working has a conventional PCM bit rate of about 7.3 bits/sample as shown by the dotted line in Figure 4. Additional equivalences are summarized in Table 1. We have conducted blind paired-comparison listening tests to confirm these equivalences. Example SDVQ-PCM speech files are at www.its.bldrdoc.gov/audio/pubs_talks/sdvqpcm_examples.php.

For b=4 to 6, Figure 4 and Table 1 show a 0.4 bit increase in equivalent speech quality (single-channel SDVQ-PCM relative to conventional PCM) due to the use of VQs that exploit adjacent sample correlation. For b=4 to 7, there is an additional ≈1.0 bit increase in equivalent speech quality when two channels are combined due to the structured relationship between the VQs. The results for A-law SDVQ-PCM compared with A-law conventional PCM agree with Table 1 to within ±0.1 bits.

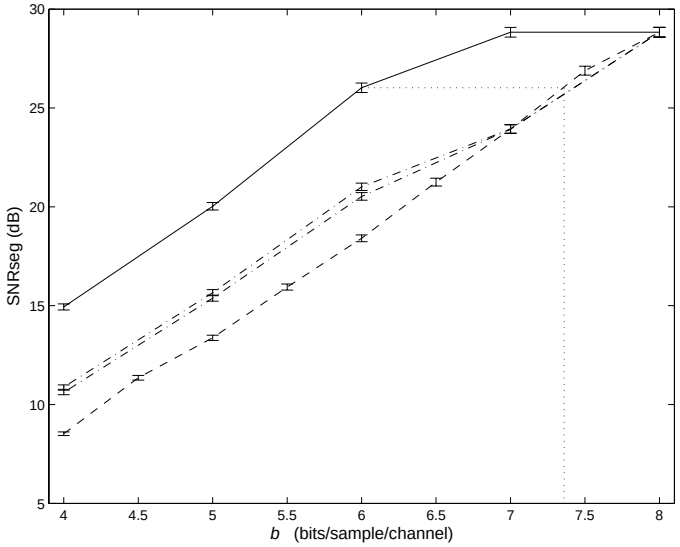

Fig. 4. Means and 95% confidence intervals for SNRseg values on μ-law SDVQ-PCM (solid line for 2 channels working, dash-dot lines for 1 channel working) and conventional PCM (dashed line).

5. 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 SDVQ-PCM speech coder offers both options.

If minimizing the data rate is the priority, one might choose b=4. The total data rate is the same as conventional PCM. This choice will provide an equivalent PCM speech quality of 5.3 bits/sample when both channels are working and 4.4 bits/sample when only one channel is working. If higher speech quality is needed, then one might choose b=6. The total data rate is 50% greater than that of conventional PCM. This choice will provide an equivalent PCM speech quality of 7.3 bits/sample when both channels are working (a quality that is very close to conventional 8 bits/sample PCM) and 6.4 bits/sample when only one channel is working (a vast improvement over a complete outage in the speech signal).

SDVQ-PCM can be implemented by simply 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, 2b-bit codes F1 and F2. If only one code (F1 or F2) arrives at the receiver, a 2b-bit look-up will generate a coarse approximation to the original pair of PCM codes. If both codes arrive at the receiver, then a 4b bit look-up is required and a finer approximation will result.

Finally, we note that partitions using square cells ($w \times w$) described here are a special case of the more general case of rectangular cells $\left(wr^{+0.5} \times wr^{-0.5}\right)$ with aspect ratio r. Consider partitions F_1 and F_2 that use rectangular cells with aspect ratios r_0 and r_0^{-1} respectively.

Data Rate b (bits/sample/channel)	Equivalent Conventional PCM Speech Quality (bits/sample)		Total Data Rate Increase over Conventional PCM
	1 Channel Working	2 Channels Working	
4	4.4	5.3	0%
5	5.4	6.3	25%
6	6.4	7.3	50%
7	7.0	8.0	75%
8	8.0	8.0	100%

Table 1. SDVQ-PCM Rate and Quality Summary.

Compared to the square case (r_0=1), increasing the aspect ratio r_0 reduces the single-channel speech quality and increases the two-channel speech quality. Thus the aspect ratio allows us to trade off single- and two-channel speech quality against each other and can be used to tune an SDVQ-PCM speech coder appropriately for known channel conditions.

6. REFERENCES

[1] A.A. El Gammal, and T.M. 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.A. 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.K. Goyal, "Multiple description perceptual audio coding with correlating transforms," *IEEE Trans. Speech and Audio Proc.*, 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.D. Gibson, and V. Cuperman, "A multiple description speech coder based on AMR-WB for mobile ad hoc networks," in *Proc. IEEE ICASSP '04,* Montreal, Canada, May 2004.

[7] K. Matty and L. Kondi, "Balanced multiple description video coding using optimal partitioning of the DCT coefficients," in *Proc. IEEE ICASSP '04,* Montreal, Canada, May 2004.

[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, no. 3, pp. 225-246, Sep. 1969.

[10] N. Kitawaki, "Quality assessment of coded speech," in *Advances in Speech Signal Processing*, S. Furui and M. Sondi, Eds., Marcel Dekker, New York, 1992.

[11] 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.

[12] J.G. Beerends, A.P Hekstra, A.W. Rix, and M.P. 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, no. 10, pp. 765-778, Oct. 2002.