14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

AN ADAPTIVE EQUALIZER FOR ANALYSIS-BY-SYNTHESIS SPEECH CODERS

Mark Jasiuk and Tenkasi Ramabadran

Speech Processing Research Lab, Motorola
1301 East Algonquin Road, Schaumburg, IL 60196, USA
Mark.Jasiuk@motorola.com, Tenkasi.Ramabadran@motorola.com

ABSTRACT

An equalizer to enhance the quality of reconstructed speech from an analysis-by-synthesis speech coder, e.g., CELP coder, is described. The equalizer makes use of the set of short-term predictor parameters normally transmitted from the speech encoder to the decoder. In addition, the equalizer computes a matching set of parameters from the reconstructed speech. The function of the equalizer is to undo the computed set of characteristics from the reconstructed speech and impose the set of desired characteristics represented by the transmitted parameters. Design steps for the equalizer and its implementation both in time and frequency domain are described. Experimental results of applying the equalizer to the output of a standard coder, viz., EVRC (Enhanced Variable Rate Coder) operating at half-rate (4000 bps), are presented. Objective evaluation using an ITU recommended voice quality tool shows that the equalizer can help enhance the quality of the reconstructed speech significantly.

1. INTRODUCTION

One of the characteristics of Analysis-by-Synthesis speech coders, e.g., Code Excited Linear Predictive (CELP) coders, which typically use Mean Squared Error (MSE) minimization criterion, is that as the bit rate is reduced, the error matching at higher frequencies becomes less efficient. This is because MSE criterion tends to emphasize signal modeling at lower frequencies where the energy is relatively higher. Any training procedure for optimizing excitation codebooks, when used, likewise tends to emphasize lower frequencies and attenuate higher frequencies in the trained codevectors, with the effect becoming more pronounced as the excitation codebook size is reduced. The perceived effect of the above characteristic is that the reconstructed speech becomes increasingly muffled as the bit rate is reduced. This muffling effect can be mitigated to some extent by the first order high-pass filter, whether fixed or adaptive, that is part of a standard adaptive spectral post-filter commonly used with CELP speech coders although the original purpose of this filter is to compensate for the spectral tilt introduced by the other terms in the spectral post-filter [1]. Another solution to this problem can be found in a 3GPP2 standard document [2] in the context of an algebraic excitation codebook. It involves the use of a shaping filter for the excitation codebook of the form $H_{SHAPE}(z) = 1 - \mu z^{-1}$, where $0 \le \mu \le 0.5$. The value of μ is selected based on the degree of periodicity of the preceding sub-frame, which when high, causes a value close to 0.5 to be selected. This imposes a high-pass characteristic on the excitation codevector being evaluated and thereby on the excitation codevector that is ultimately selected. Notice that the shaping filter does not necessarily optimize the MSE criterion. However, it is still used because it mitigates the attenuation at higher frequencies to some extent and the resulting reconstructed speech sounds more similar to the target input speech.

In this paper, we describe an adaptive equalizer that attempts to impose the overall frequency characteristics of the input speech onto the reconstructed speech thereby mitigating the above muffling effect. The idea is to design an equalizer that would bridge the gap between the short-term spectral characteristics of the input and reconstructed speech and apply it to the reconstructed speech. The European patent specification EP 1141946B1 [3] describes an adaptive equalizer to reduce the distance between the reconstructed signal and the input signal. In the patent specification, a transfer function is computed in the frequency domain, which, when applied to the reconstructed signal renders it identical to the input signal. The transfer function is simplified, quantized, and transmitted as enhancement information to the speech decoder. Clearly, this entails additional bandwidth. On the other hand, the adaptive equalizer described in this paper is designed using parameters already transmitted to the speech decoder as part of the coded bit stream and parameters derived from the reconstructed speech. Therefore, there is no additional bandwidth requirement. The design and implementation of the adaptive equalizer are described in Section 2. Results of applying the equalizer to a standard speech coder are presented in Section 3. Our conclusions are summarized in Section 4.

2. DESIGN AND IMPLEMENTATION

Figure 1 shows a block diagram with a typical CELP speech decoder along with the adaptive equalizer. The inputs to the speech decoder block are: i) the quantized linear predictor (LP) coefficients A_q, ii) the long-term predictor delay L and coefficients β_j's, and iii) the excitation codevector index I and gain factor γ. The quantized LP coefficients $A_q = \{a_1, a_2, \ldots, a_P\}$ describing the short-term spectral envelope are transmitted to the decoder typically once per frame. Other parameters are transmitted once per sub-frame. For speech signals sampled at 8000 Hz, the frame duration ranges typically from 10 to 30 ms, the number of sub-frames per frame is typically 2 to 4, and the order P of the linear predictor is typically 10.

The transmitted LP coefficients A_q usually correspond to the last sub-frame of a frame and LP coefficients corresponding to other sub-frames are obtained through interpolation between the A_q parameters of the current and preceding frames. For each sub-frame, the speech decoder selects from a fixed codebook an excitation codevector C_I corresponding to the index I, scales it by the gain factor γ, and filters the gain-scaled excitation vector by the long-term and short-term predictor filters corresponding to the sub-frame to generate the reconstructed speech $\hat{s}(n)$ as an estimate of the input speech $s(n)$ at the speech encoder. Here n is the sample index and ranges from 0 to N-1, where N is the sub-frame length in samples. The long-term predictor (LTP) filter is described by the system function

$$H_{LTP}(z) = \frac{1}{1-\sum_{j=-J1}^{j=J2}\beta_j z^{-L+j}}, J1 \ge 0, J2 \ge 0, J = J1 + J2 + 1$$

where, the LTP filter order J is typically between 1 and 3, and the short-term predictor (STP) filter (for the last sub-frame) is described by the system function

$$H_{STP}(z) = \frac{1}{1-\sum_{i=1}^{P} a_i z^{-i}} \; .$$

In Figure 1, the inputs to the equalizer block are the quantized LP coefficients A_q, and the reconstructed speech $\hat{s}(n)$. The output of the equalizer block is the equalized, reconstructed speech $\hat{s}_{eq}(n)$.

2.1 Equalizer Design

Figure 2 illustrates a flowchart for the equalizer design based on the reconstructed speech $\hat{s}(n)$ and the transmitted quantized LP coefficients A_q. The reconstructed speech is first windowed and an LP analysis is performed on the windowed speech to obtain the LP coefficients $A_r = \{b_1, b_2, \ldots, b_Q\}$ where, $Q \le P$, is the order of the LP model. Ideally, the window used for LP analysis of the reconstructed speech is the same as the window that was used on the input speech at the encoder in obtaining the transmitted coefficients A_q. Furthermore, the windowing is synchronous – that is, the window placement is such that corresponding speech samples are used at the encoder and decoder in obtaining A_q and A_r respectively for each frame. It is also desirable that the same LP analysis technique is used for computing A_r as that was used in computing A_q. The order Q of the LP model for the reconstructed speech is typically chosen to be equal to P. But if Q is chosen to be less than P, then the model order of A_q has to be reduced to Q using well-known techniques before being used in subsequent computations.

The objective of the equalizer is to undo the spectral characteristics corresponding to A_r and impose the spectral characteristics corresponding to A_q on the reconstructed speech. To accomplish this, we first compute from A_r the zero-state impulse response of the zero-filter $A_r(z) = 1 - \Sigma\, b_i z^{-i}$ as the sequence $\{1 - b_1 - b_2 \ldots -b_Q\}$. This sequence is then filtered by the zero-state pole filter $1/A_q(z) = 1/(1 - \Sigma\, a_i z^{-i})$ to obtain an *initial* estimate of the equalizer impulse response. This equalizer response, however, causes phase distortion. In order to obtain an equalizer response with zero phase distortion, the initial estimate of the equalizer impulse response is processed further by truncating it to a suitable length (preferably a power of 2, e.g., 512), transforming into the frequency domain by means of a Fourier transform, e.g., an FFT, computing the magnitude response, and setting the phase response to zero. The resulting magnitude-only frequency response $G_{eq}(\omega_i)$ is referred to as the *intermediate* equalizer frequency response. The corresponding time-domain sequence $g_{eq}(n)$ can be obtained by means of an inverse Fourier transform, e.g., IFFT, and is referred to as the intermediate equalizer impulse response. This impulse response is real and symmetric because of the imposed zero-phase characteristic.

The intermediate equalizer impulse response is then truncated to a suitable length by means of a symmetric time-domain window, e.g., a rectangular window. The resulting sequence $h_{eq}(n)$, which is still symmetric, is the desired *final* equalizer impulse response. The sequence $h_{eq}(n)$ is zero-padded on both sides to maintain its symmetry and Fourier transformed, e.g., by an FFT, to yield the magnitude-only final equalizer frequency response $H_{eq}(\omega_i)$. Notice that since $H_{eq}(\omega_i)$ has been obtained from the zero-padded sequence

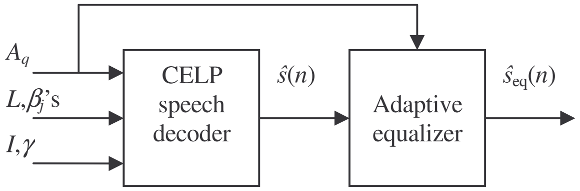

Figure 1. CELP speech decoder with Adaptive equalizer

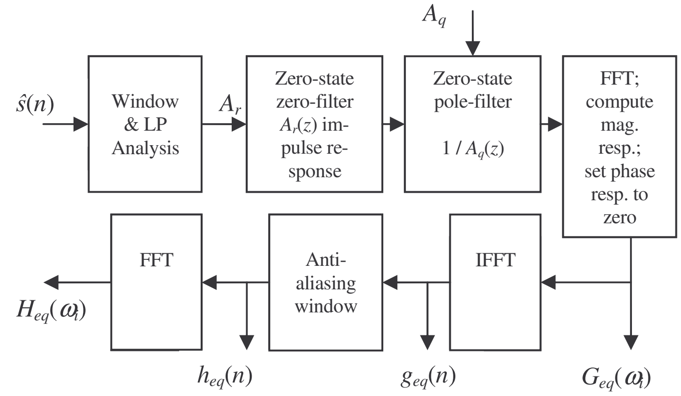
A_q
$\hat{s}(n)$
Window & LP Analysis
A_r
Zero-state zero-filter $A_r(z)$ impulse response
Zero-state pole-filter $1 / A_q(z)$
FFT; compute mag. resp.; set phase resp. to zero
FFT
Anti-aliasing window
IFFT
$H_{eq}(\omega_i)$
$h_{eq}(n)$
$g_{eq}(n)$
$G_{eq}(\omega_i)$

Figure 2. Equalizer design flowchart

$h_{eq}(n)$, it can be used to filter, i.e., equalize, a reconstructed speech frame without aliasing provided the length of the frame is not greater than one plus the number of zeros in the zero-padding. This is not the case with $G_{eq}(\omega)$ and this is why the window used to obtain $h_{eq}(n)$ from $g_{eq}(n)$ is referred to as an anti-aliasing window in Figure 2. $H_{eq}(\omega)$ is the desired equalizer frequency response although in order to reduce complexity $G_{eq}(\omega)$ can also be used for equalization if a certain amount of aliasing distortion can be tolerated.

2.2 Equalizer Implementation

As shown in Figure 3, the equalizer can be implemented in the frequency domain or the time domain. For this purpose, the overlap-add (OLA) analysis/synthesis approach [4] is used. To implement the equalizer in the frequency domain, the reconstructed speech $\hat{s}(n)$ is first windowed by means of a suitable OLA analysis window. Ideally, the OLA window will satisfy the perfect reconstruction property, that is, the overlapped sections of adjacent windows will add up to unity. An example of such a window is a non-symmetric version of the *Hanning* window with 50% overlap. Such a window is expressed by $w(n) = 0.5[1 - cos(2\pi n/2L)]$, $n = 0, 1, \dots , 2L\text{-}1$, where $2L$ is the length of the window and the distance between adjacent windows is L. If this type of OLA window is used, we choose L to be equal to the frame length used at the speech encoder. We also place the window such that it coincides with the LP analysis window used for obtaining A_r to the extent possible. The windowed reconstructed speech frame is then zero-padded and Fourier transformed into the frequency domain, e.g., by means of an FFT. The number of zeros padded should be at least equal to the length of $h_{eq}(n)$ minus one to avoid aliasing and the size of the Fourier transform should be the same as that of $H_{eq}(\omega)$. The Fourier transformed signal is equalized in the frequency domain by multiplying each frequency coefficient by the corresponding term of $H_{eq}(\omega)$ or $G_{eq}(\omega)$. The equalized frequency-domain signal is next inverse Fourier transformed, e.g., by means of an IFFT, and overlap-added with the adjacent frames to yield the equalized, reconstructed signal $\hat{s}_{eq}(n)$. In OLA synthesis, special care must be taken to appropriately handle the "*sample tails*" at both ends of the windowed speech frame caused by the equalizer impulse response. To implement the equalizer in the time domain, the windowed reconstructed speech frame is directly convolved with the equalizer impulse response $h_{eq}(n)$ or $g_{eq}(n)$ and overlap-added with adjacent frames to yield $\hat{s}_{eq}(n)$ once again ensuring that the sample tails at both ends are handled appropriately. To provide a concrete example of the different lengths involved, let the frame length L be 160 samples. The OLA analysis window is then 320 samples long, the FFT length may be chosen as 512, and the length of the anti-aliasing window, i.e., the length of $h_{eq}(n)$, may be chosen as 193, which would result in a 96-sample "sample tail" at each end.

While the above is a description of the standard equalizer implementation, several variations are possible: 1) Adaptive spectral post-filtering is commonly used in CELP speech coders to reduce the perceived level of coding noise [1]. The standard equalizer undoes the effect of post-filtering, so it may be useful to include adaptive spectral post-filtering as part of the equalizer transfer function or as a separate block after the standard equalizer block. 2) Bandwidth widening technique may be employed in the design of the equalizer by replacing $A_q(z)$ and $A_r(z)$ by $A_q(z/\alpha)$ and $A_r(z/\alpha)$ respectively, where $\alpha < 1$. This places a reduced demand on the equalizer and leads to improved performance especially when the transmitted coefficients A_q are coarsely quantized and/or when there is a mismatch between the LP analysis window and the OLA analysis window. 3) The quantized LP coefficients are transmitted typically once per frame and this may result in long OLA analysis frames and a correspondingly long delay in the standard equalizer implementation. By using interpolated values of A_q in addition to the transmitted A_q values and more frequently computed values of A_r, the effective length of OLA analysis frames and the corresponding delay can be reduced.

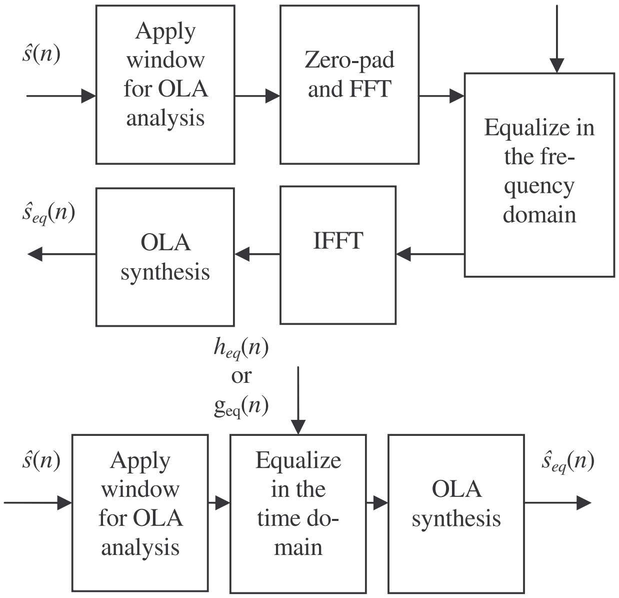
$\hat{s}(n)$
Apply window for OLA analysis
Zero-pad and FFT
Equalize in the frequency domain
$\hat{s}_{eq}(n)$
OLA synthesis
IFFT
$h_{eq}(n)$ or $g_{eq}(n)$
$\hat{s}(n)$
Apply window for OLA analysis
Equalize in the time domain
OLA synthesis
$\hat{s}_{eq}(n)$

Figure 3. Equalizing in **a)** frequency domain and **b)** time domain

3. EXPERIMENTAL RESULTS

In order to evaluate the performance of the equalizer, the EVRC (Enhanced Variable Rate Codec) standard [5] used in CDMA networks was selected. This codec is a variable rate codec that operates at three different rates: 8550 bps (full-rate), 4000 bps (half-rate) and 800 bps (eighth-rate). For the purpose of evaluating the equalizer, the EVRC was configured to operate always at half-rate, i.e., at 4000 bps. The EVRC was selected for the study because it is an analysis-by-synthesis coder and when operating at 4000 bps, it has many of the shortcomings outlined in Section 1. The EVRC is a type of CELP codec known as Relaxed CELP where the target signal the coder is attempting to match is not the original input speech but a time-warped version of the input speech. For the objective evaluation of the equalizer performance, we use this time-warped version of the input speech as the reference signal. The objective evaluation itself was done using an ITU recommended voice quality tool known as PESQ [6]. Given a reference signal and a degraded signal, e.g., the unequalized or equalized EVRC output speech signal, this tool provides a predicted MOS (Mean Opinion Score) for the degraded signal on a 5-point scale. For the evaluation, a speech database consisting of 32 sentence pairs (8 speakers, 4 male + 4 female, 4 sentence pairs each) was used. The speech database is sampled at 8000 Hz, quantized at 16 bits/sample, and has no pre-processing filter.

To study the effect of window length on the equalizer performance, we used a Hanning window with 50% overlap (as specified in Section 2.2) for both LP and OLA analysis. The LP coefficients A_q and A_r were calculated respectively from high-pass filtered input speech (from which EVRC normally computes the LP coefficients) and EVRC output speech. The LP coefficients were computed using an auto-correlation technique slightly different from the one that EVRC uses and the coefficients were unquantized. The predictor order P was 10. It may be noted that the EVRC also uses a predictor order P of 10 and a frame length L of 160. The equalizer performance results for three different window lengths ($2L$) and several values of the bandwidth widening parameter α are shown in Table 1. For comparison, the unequalized EVRC output has a predicted MOS of 3.428. It is seen that the equalizer performance improves as the window length decreases and the best performance for all three lengths occurs when $\alpha = 0.96$. To study the effect of predictor order on equalizer performance, we varied the predictor order P keeping the window length ($2L$) at 160 and α at 1.00. Other conditions were similar to the experiment above. The results of this experiment are shown in Table 2. Once again, for comparison, the predicted MOS of the unequalized EVRC output is 3.428. It is seen that the equalizer performance improves with increasing predictor orders.

To study the equalizer performance under more realistic conditions, the LP analysis window was made identical to the one used in the EVRC, viz., a 160-sample Hamming window. The OLA analysis window was a 320-sample Hanning window with 50% overlap chosen such that the LP analysis window is centred within the OLA analysis window. Furthermore, the same auto-correlation technique used by the EVRC was used to compute the LP coefficients. The predictor order P was chosen as 10. The results of this study for different values of α are shown in Table 3 for unquantized LP coefficients as well as quantized LP coefficients quantized both by EVRC half-rate quantizer (22 bits/frame, 1100 bps), and full-rate quantizer (28 bits/frame, 1400 bps). The best equalizer performance occurs when $\alpha = 0.96$, 0.80, and 0.88 respectively for the unquantized, half-rate quantized, and full-rate quantized LP coefficients. Compared to the predicted MOS of 3.428 for the unequalized EVRC output, the equalized outputs for the three cases are better by 0.121, 0.054, and 0.081 respectively. Clearly, the lower the quantization error, the better is the equalizer performance.

Table 1. Predicted MOS for different window lengths

α	Window length ($2L$)		
	320	**256**	**160**
1.00	3.533	3.596	3.620
0.98	3.567	3.636	3.655
0.96	3.571	3.638	3.656
0.94	3.569	3.633	3.649
0.92	3.564	3.624	3.640
0.90	3.559	3.615	3.629

Table 2. Predicted MOS for different predictor orders

Predictor order P	Predicted MOS
10	3.620
12	3.658
14	3.682
16	3.706
18	3.725
20	3.745

Table 3. Predicted MOS for different quantizers

α	Quantizer		
	None	**Half-Rate**	**Full-Rate**
1.00	3.519	3.219	3.366
0.96	3.549	3.403	3.484
0.92	3.543	3.454	3.505
0.88	3.534	3.477	3.509
0.84	-	3.481	3.507
0.80	-	3.482	3.504
0.76	-	3.482	3.499
0.72	-	3.480	3.494

To qualitatively illustrate the function of the equalizer, Figure 4 shows the spectral plots of a voiced segment for the original input speech (solid line), the reconstructed speech (dotted line), and the equalized speech (dashed line) corresponding to the situation in Table 3, col. 2 with $\alpha = 1.00$. Notice the attenuation of the spectrum at higher frequencies for the reconstructed speech leading to the “muffling” effect. As can be seen, the equalizer clearly counteracts this effect.

From the results presented above, it is seen that the equalizer has the potential to provide enhanced speech quality for low bit rate coders. However, to take full advantage of the equalizer, i) the transmitted short-term predictor parameters should be of high quality, i.e., should have low quantization error (see the results in Table 3, columns 2-4), and ii) the LP analysis window should match the OLA analysis window as much as possible (compare the results in Table 1, col. 2 with the results in Table 3, col. 2). This means that the use of the equalizer has to be taken into consideration at the design stage of the speech coder itself so that appropriate choice for the speech coder frame length, LP analysis window (type and length), and the quantizer for the short-term

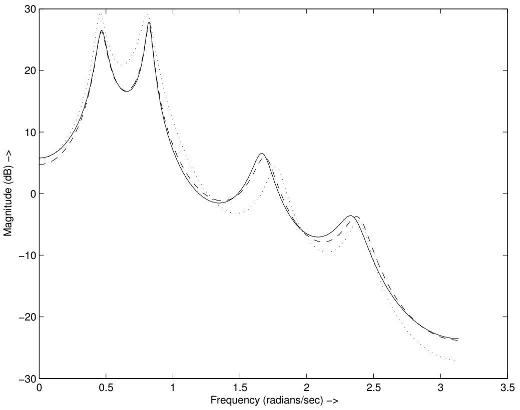

Figure 4. Spectral plots of a voiced segment for (**a**) input speech (solid line), (**b**) reconstructed speech (dotted line), and (**c**) equalized speech (dashed line)

predictor parameters can be made. It is also seen that the further away the transmitted short-term predictor parameters are from the ideal (unquantized) values and the LP analysis window from the ideal OLA analysis window, the lower the value of α (bandwidth widening parameter) at which the best equalizer performance is achieved.

4. CONCLUSIONS

In the context of low bit rate analysis-by-synthesis speech coders, an adaptive equalizer to enhance the quality of the reconstructed speech was described. The equalizer makes use of the short-term predictor parameters normally transmitted from the speech encoder to the decoder to impose the overall frequency characteristics of the input speech onto the reconstructed speech. Thus, the use of the equalizer does not entail additional bit rate. Design steps for the equalizer were presented and its implementation details were discussed. Using an ITU recommended voice quality tool (PESQ) and the EVRC operating at 4000 bps, the performance of the equalizer was evaluated under different conditions. The results show that the equalizer has the potential to enhance the quality of reconstructed speech in low bit rate coders. However, to take full advantage of the equalizer, the speech coder design has to be modified appropriately with proper choice of the speech coder frame length, LP analysis window, and quantizer for the short-term predictor parameters. Future work will include studying and evaluating the equalizer with other low bit rate coders, subjective evaluation tests, use of OLA analysis windows with smaller amount of overlap, i.e., less than 50%, and the design of a new low bit rate speech coder that incorporates the equalizer.

ACKNOWLEDGEMENT

The authors would like acknowledge with thanks the assistance provided by Bob Novorita of Motorola in conducting the objective evaluation.

REFERENCES

[1] J.-H. Chen and A. Gersho, “Real-Time Vector APC Speech Coding at 4800 BPS with Adaptive Postfiltering,” *Proc. ICASSP*, pp. 2185-2188, April 1987.

[2] 3GPP2 C.S0052-A, “Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Options 62 and 63 for Spread Spectrum Systems,” *Version 1.0*, April 2005.

[3] EP 1 141 946 B1, “Coded Enhancement Feature for Improved Performance in Coding Communication Signals,” *Inventors – R.Hagen and B. Kleijn*, April 2004.

[4] R.E. Crochiere, “A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis,” *IEEE Trans. ASSP*, Vol. 28, No. 1, pp. 99-102, February 1980.

[5] 3GPP2 C.S0014-A, “Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems,” *Version 1.0*, April 2004.

[6] ITU Recommendation P.862, “Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech,” February 2001.