16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

NOISE SHAPING IN AN ITU-T G.711-INTEROPERABLE EMBEDDED CODEC

Jimmy Lapierre[1], *Roch Lefebvre*[1], *Bruno Bessette*[1], *Vladimir Malenovsky*[1], *Redwan Salami*[2]

[1]Université de Sherbrooke, Sherbrooke (Québec), Canada, [2]VoiceAge Corporation, Montréal (Québec), Canada
phone: +1 (819) 821-8000, email: roch.lefebvre@usherbrooke.ca
http://www.gel.usherbrooke.ca/audio

ABSTRACT

In the transition from narrowband to wideband speech communications, there is a need in some applications for a high quality wideband coding scheme interoperable with the ITU-T G.711 narrowband coding standard. This can be accomplished using a multi-layer coding scheme with a G.711 compatible core layer. For optimal wideband quality in the upper layers, this requires using full frequency range (50-4000 Hz instead of 300-3400Hz) in the core layer. In this context, the 8-bit non-uniform PCM quantizer of the ITU–T G.711 standard can produce highly perceptible noise. The purpose of this paper is to demonstrate how efficient noise masking can be applied at the encoder in a G.711-interoperable manner, and how the same noise masking can be extended at the decoder to one or more enhancement layers to implement a perceptually optimized multilayer codec.

1. INTRODUCTION

The demand for efficient digital wideband (50-7000Hz) speech and audio encoding techniques with good subjective quality is increasing for numerous applications such as audio/video teleconferencing, multimedia, IP telephony and other various wireless applications. Historically, speech coding systems were only able to process telephony-band signals (300-3400Hz), achieving good intelligibility. Nevertheless, the bandwidth of 50-7000Hz is required to increase the intelligibility and naturalness of speech and to offer a better face-to-face experience to the end user. For audio signals such as music, this frequency range enables good quality, even though it is lower than CD quality (20Hz-20kHz). To this effect, ITU-T has approved Recommendation G.729.1 in May 2006 [1], which is an embedded multi-rate coder with a core interoperable with G.729 at 8kbps. Similarly, a new activity has been launched in March 2007 for an embedded wideband codec based on a narrowband core interoperable with G.711 (both μ-law and A-law) at 64kbps. This new G.711-based standard is known as the ITU-T G.711 wideband extension (G.711.1). Since G.711 is widely deployed in voice communication systems, extending to wideband services while keeping interoperability with the legacy end devices is a desirable feature.

The G.711 standard defines a well known 8-bit non-uniform scalar quantization law operating at an 8 kHz sampling rate [2]. It was designed specifically for narrowband voice telephony (300-3400Hz) with a preconditioned input signal (high frequency emphasis), and produces good quality within that configuration. However, to allow efficient wideband coding in the upper layers of an embedded structure, the G.711-interoperable core should be capable of processing flat narrowband (50-4000Hz) inputs. When used in that context, the G.711 quantization noise becomes audible and often annoying, especially at high frequencies (Figure 1). Thus, even if the upper frequency band (4000-7000Hz) of the embedded wideband codec is properly coded, the quality of the synthesized signal would often be poor due to the limitations of the G.711 core layer.

Although noise feedback was introduced in the early sixties [3,4] to shape the quantization noise of a scalar quantizer, it is not part of the G.711 standard (which is not an issue for standard voice telephony). This paper presents a noise shaping scheme that is interoperable with the existent ITU-T G.711 standard, while providing higher quality for full frequency range speech and audio. In the proposed approach, the quantization noise is shaped according to a psychoacoustic model very similar to the ITU-T AMR-WB standard encoder [5]. Furthermore, similar noise shaping can be achieved in an embedded scheme with one or more enhancement layers above the G.711-interoperable core. This is accomplished by adding the post-processing algorithm described in this paper. We consider a simple uniform scalar quantizer in the second layer, with a post-processor effectively lowering the shaped noise level by an additional 6dB per bit/sample.

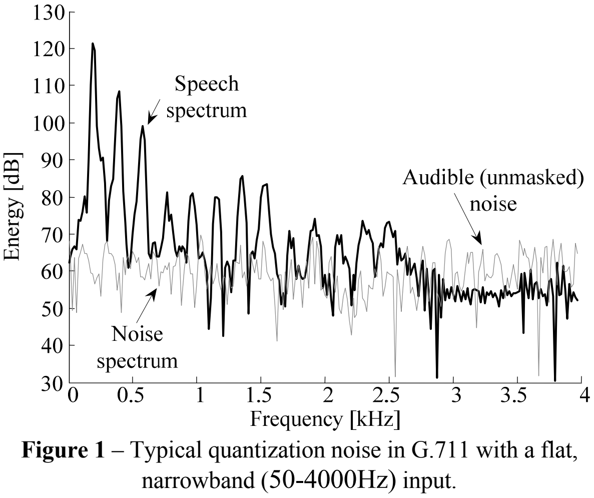

Figure 1 – Typical quantization noise in G.711 with a flat, narrowband (50-4000Hz) input.

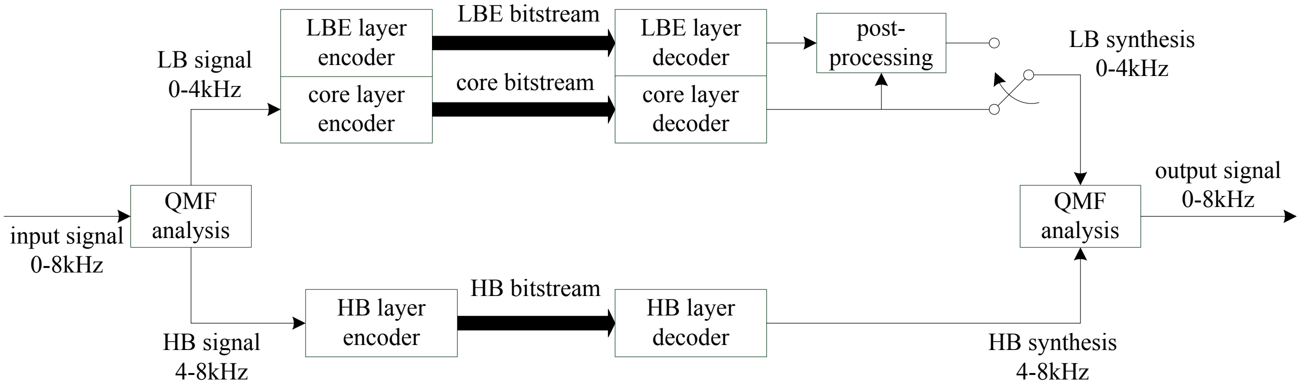

Figure 2 – Schematic diagram of the multilayer codec based on the ITU-T G.711

2. CODEC FRAMEWORK OVERVIEW

For illustration, we consider the specific structure of the G.711.1 codec. This wideband multilayer codec has two upper layers added to a G.711-interoperable core layer. The first upper layer (Low Band Enhancement, or LBE) adds 16 kbit/s to the narrowband core, improving the quality of the 50-4000Hz band. The second upper layer (High Band Extension, or HBE) adds another 16 kbit/s to encode the 4000-7000Hz band to provide a wideband signal. This structure offers four modes of operation and three bitrates, depending on which layers are used at the decoder. Table 1 shows all supported combinations. In Figure 2 we see a high level overview of the multilayer encoder/encoder. The input signal is sampled at 16kHz and split into two bands by means of a QMF filter. The lower band signal is encoded by the G.711-interoperable core layer and the LBE layer. The higher band is encoded using the HBE layer. Note that both the core layer and the LBE layer operate on signals sampled at 8kHz. The inclusion of the LBE layer in the codec allows decreasing the quantization noise level in the lower band by 12dB (6dB per bit/sample). The HBE layer encodes the HB signal, downsampled to the 0-4kHz range. Thus, the HB signal has also a sampling frequency of 8 kHz. The inclusion of the HB extension layer in the codec allows the transmission of wideband signals, providing a significantly higher quality and a more natural sound over the narrowband signal. The bitstream produced by this codec has an embedded structure, which allows the transmission facilities to select the transmitted layers according, for example, to the capabilities of terminals. The G.711.1 codec uses 10-ms frames at the encoder.

Table 1 – Layers and bitrates of the embedded codec

mode				total bitrate
R1	core layer G.711-interop.			64 kbps
R2a	core layer G.711-interop.	LB enh. layer		80 kbps
R2b	core layer G.711-interop.	HB ext. layer		80 kbps
R3	core layer G.711-interop.	LB enh. layer	HB ext. layer	96 kbps

3. CORE LAYER NOISE SHAPING

As shown in Figure 1, the standard G.711 quantizer produces a quantization noise with flat spectrum on any signal. This is far from optimal when taking into account a psychoacoustic criterion. For the signal in Figure 1, the noise in the 2.5-4 kHz frequency range is easily perceived and annoying. To benefit from the masking effects of the human auditory system, noise shaping can be applied around the G.711 quantizer. We aim at keeping the complexity low, as well as maintaining interoperability with the G.711 standard decoder. Hence, the proposed noise shaping scheme introduces a quantization error feedback loop as shown in Figure 4. In this figure, all signals are indicated with *z*-transform notation.

The proposed noise shaping loop is implemented around the standard G.711 quantizer, i.e. using the difference between $S(z)$ and $Y_8(z)$. This is different from the usual form normally found in the literature, where the difference signal is calculated directly between the input and the output of the quantizer, i.e. $X(z)$ and $Y_8(z)$ in Figure 4. However, as already proposed by Atal [6], this is simply an efficient implementation of the recursive form of the noise shaping filter (as will also be shown by the equation below). Filter $F(z)$ is referred to as the perceptual filter and its form will be described in section 5. The emphasis on low complexity suggests that we

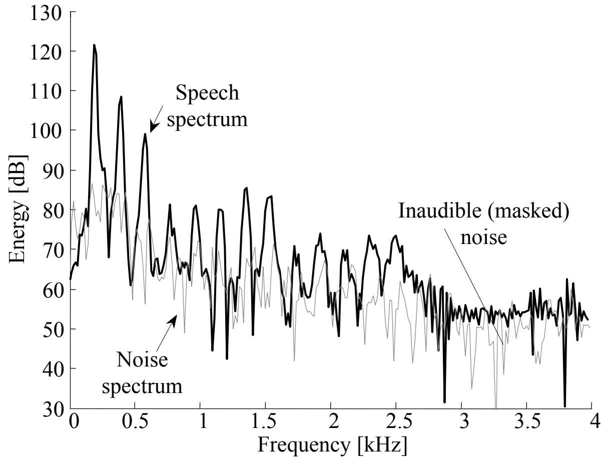

Figure 3 – The effect of error-feedback noise shaping on the

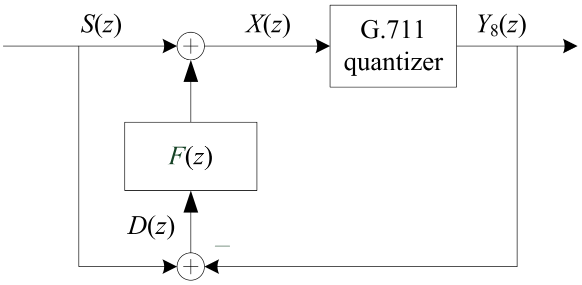

Figure 4 – Noise shaping in a G.711-interoperable encoder

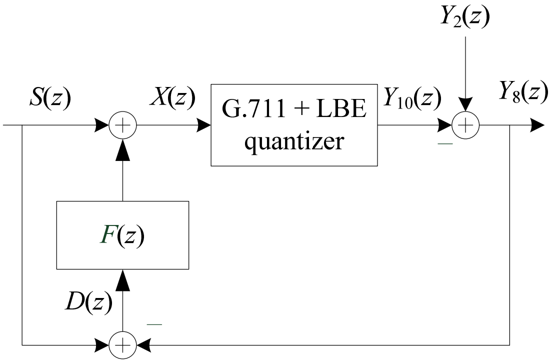

Figure 5 – Noise shaping in a G.711-interoperable encoder with a LBE layer

derive $F(z)$ based on the noise shaping filter used in AMR-WB [5], which achieves both formant and spectral tilt shaping with reduced complexity compared to other approaches. In the remaining of this section, we focus on the general input/output relationship in Figure 4.

The output signal of the core layer quantizer is given by:

$$Y_8(z) = X(z) + Q_8(z)\,, \tag{1}$$

where $Q_8(z)$ is the (8-bit) PCM quantization noise with flat spectrum. The input to the quantizer is expressed as:

$$X(z) = S(z) + \{S(z) - Y_8(z)\}F(z)\,. \tag{2}$$

By substituting $X(z)$ into Equation (1) we get:

$$Y_8(z) = S(z) + S(z)F(z) - Y_8(z)F(z) + Q_8(z)\,. \tag{3}$$

By rearranging the terms we obtain:

$$\{1+F(z)\}Y_8(z) = \{1+F(z)\}S(z) + Q_8(z)\,, \tag{4}$$

which finally yields:

$$Y_8(z) = S(z) + \frac{Q_8(z)}{1+F(z)}. \tag{5}$$

We see that by adopting the noise shaping scheme of Figure 4, the output signal of the quantizer, $Y_8(z)$, is equal to the input signal, $S(z)$, with the quantization noise shaped by the filter $(1+F(z))^{-1}$. As shown in Figure 3, the effect is that the quantization noise is now higher in the lower frequencies, where the speech spectrum is effectively masking it. On the other hand, the quantization noise in the higher frequency range is lowered to a practically imperceptible level. Consequently, the perceived noise level is much lower even though the total noise power is slightly higher than without the noise shaping.

Filter $F(z)$ in Equation (5) can be selected in any way such that the noise is properly shaped. Section 5 will describe how we can select $F(z)$ such that the same noise shaping as in AMR-WB is accomplished. But before, the next section shows how this noise shaping is extended to the upper layers, while maintaining proper noise shaping in the core layer.

4. LOWER BAND ENHANCEMENT LAYER

In the G.711-interoperable embedded codec described in Section 2, the encoder can transmit two extra bits per sample to enhance the quality of the lower band synthesis. These refinement bits are generated by the lower band enhancement layer (LBE). The additional bits are taken from the "mantissa", extracted during core-layer quantization. When applying this 2-bit per sample LBE, the noise floor is decreased, in principle, by 12dB in the whole bandwidth. In addition, almost no calculations are necessary in the LBE quantizer, so the complexity of the encoder is not significantly increased. However, the noise feedback loop at the encoder must only take into account the quantization error of the core layer as it does not know whether the LBE will be used in the decoder or not. As a consequence, to allow noise shaping in the second layer and to maintain proper noise shaping in the core layer, some additional filtering must be applied in the decoder on the LBE decoded samples to achieve a proper shaping of the synthesized signal in case the LBE layer is used.

To derive a proper form for the post-processing that has to be applied to the LBE decoded samples, let us assume the scenario shown in Figure 5. The difference with Figure 4 is that, in Figure 5, the quantizer is a 10-bit quantizer, but the noise feedback loop is still calculated using the 8-bit, PCM core quantizer. Hence, Equation (2) holds both for Figure 4 and Figure 5. The incorporation of the LBE quantizer to the original 8-bit quantizer may be viewed as a 10-bit G.711 quantizer, which produces a signal

$$Y_{10}(z) = X(z) + Q_{10}(z)\,, \tag{6}$$

where $Q_{10}(z)$ is a quantization noise, different from$Q_8(z)$. The 2-bit LBE layer produces a signal $Y_2(z)$, which is in effect a quantized error signal related to $Y_{10}(z)$ in the following way:

$$Y_{10}(z) = Y_8(z) + Y_2(z)\,. \tag{7}$$

Using $Y_{10}(z)$ directly from Equation (7) when decoding both the core layer ($Y_8(z)$) and the NBE layer ($Y_2(z)$) would result in an improper noise shape. Indeed, by substituting $X(z)$ from Equation (6) in Equation (2) we get:

$$Y_{10}(z) - Q_{10}(z) = S(z) + \left\{ S(z) - Y_8(z) \right\} F(z) . \qquad (8)$$

Then, using Equation (7) to substitute $Y_8(z)$ in Equation (8) yields :

$$Y_{10}(z) = S(z) + \frac{1}{1+F(z)} Q_{10}(z) + Y_2(z) \frac{F(z)}{1+F(z)} . \qquad (9)$$

This is the synthesis signal we get from the core and LBE layers when no post-processing is applied to $Y_2(z)$ at the decoder. What we would really want to obtain, in order to get the proper noise shaping in the second layer, is only the first two terms at the right of Equation (9). Hence, by subtracting the last term of Equation (9) from the left hand side (i.e. from $Y_{10}(z)$) we get the desired result: a signal, generated from decoding both the core and NBE layers, which has a properly shaped quantization noise. Thus,

$$Y_D(z) = Y_{10}(z) - Y_2(z) \frac{F(z)}{1+F(z)} , \qquad (10)$$

where $Y_D(z)$ is the desired signal in the decoder. Finally, substituting $Y_{10}(z)$ from Equation (7) into Equation (10), we obtain :

$$Y_D(z) = Y_8(z) + Y_2(z) \frac{1}{1+F(z)} , \qquad (11)$$

Thus, the decoded signal from the LBE layer must first be filtered by $(1+F(z))^{-1}$ and then added to the decoded signal of the core layer, $Y_8(z)$. This ensures that the shape of the quantization noise when using both the core and the LBE layers will be coherent with the shape of the quantization noise when using only the core layer. Of course, the quantization noise when using 2 layers will be lower than the quantization noise when using only the core layer. This reasoning can be extended to as many enhancement layers as desired, providing gradual noise reduction for each additional layer. Note that, instead of transmitting $F(z)$ explicitly, it is estimated at the decoder from the decoded core layer signal, which is a good approximation of $S(z)$ since 8-bit PCM is used. A schematic description of the core and LBE layer decoder is shown in Figure 6.

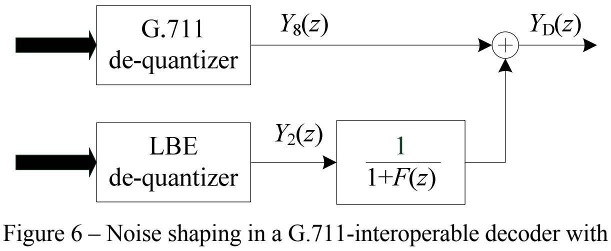

Figure 6 – Noise shaping in a G.711-interoperable decoder with a LBE layer

5. PSYCHOACOUSTIC MODEL

Filter $F(z)$ in Figures 4 and 5 must be estimated in such a way that the quantization noise has a perceptually relevant shape. The psychoacoustic model used in this paper for deriving filter $F(z)$ is based on the noise weighting filter of the AMR-WB standard speech codec [5]. The weighting filter in AMR-WB achieves both formant weighting and spectral tilt while maintaining low complexity. The noise-feedback filter $F(z)$ in Equations (5) and (11) is calculated on a frame-by-frame basis, such that $1+F(z) = A(z/\gamma)$. Here, $A(z)$ is the LPC filter calculated on the *pre-emphasized* input signal, as in AMR-WB, except that the pre-emphasis filter is adaptive. Filter $F(z)$ is updated at every frame (10 ms in the G.711.1 codec). Note that $F(z)$ does not need to be transmitted. At the encoder (for the core layer), it is calculated on the input narrowband signal. At the decoder (for the LBE layer), it is calculated on the synthesized narrowband signal from the core layer. The mismatch introduced by this approximation at the decoder is minimal, since the core layer is a high rate encoder, and the same bandwidth is used in the core and LBE layers.

The adaptive pre-emphasis operates as follows. Since one of the goals of this filtering is to reduce noise between low frequency harmonics, the level of pre-emphasis is made dependent on the level of low frequency harmonics in the input signal. This is estimated using a zero-crossing count. Significant pre-emphasis is applied when dominant harmonics are present. Contrarily, signals with limited harmonic structure, that may resemble pink noise, will have little pre-emphasis applied to them.

The LPC filter $A(z)$ is then calculated on the pre-emphasized signal, using an analysis window covering the current and previous frames. An asymmetric analysis window is selected, whose shape is designed to obtain the proper balance between simultaneous versus pre- and post-masking with the resulting filter $F(z)$.

6. MANAGING LOW LEVEL SIGNALS

The G.711 quantizer has a relatively limited dynamic range. Therefore, when the input signal amplitude decreases significantly, it gradually becomes incapable of masking the quantization noise, no matter how “perfectly” the noise is shaped. In these cases, when the noise becomes audible, the best alternative is to render that noise the least annoying possible. For the case of very low-level inputs, we propose three refinements to the basic noise shaping approach described in sections 3 to 5.

The first improvement is to make the noise shaping filter $F(z)$ converge towards a preset shape when the input signal becomes significantly low. This predetermined filter is designed so that the quantizer noise is less annoying than white noise. This feature is also very useful to avoid significant mismatches between the filter calculated at the encoder and the version calculated at the decoder from the core layer synthesis. Furthermore, for signals of even lower power that are unable to mask any quantization noise, the contribution of the feedback loop can also be progressively reduced, since in general any amount of noise shaping increases the total power of the quantization noise [7].

Another possible refinement consists in tuning the quantizer for very low level signals (such as faint background noise). We refer to this as “dead-zone quantization”, since at very low levels, the input signal can fall in and out of the Voronoi region centered around the origin. Thus, it is desirable to prevent low level noise from producing a higher level of output noise, which can actually become twice as high as the level of the actual input signal. This happens when the input samples are just high enough to be rounded (quantized) to the first value larger or smaller than zero in the quantization table. For example, the A-law lowest quantization steps are 0 and ±16. Normally, input samples will be quantized to +16 if they are between 8 and 23 inclusively. Consequently, a signal with, for example, a sample distribution in the range of ±10 will produce an output covering the range of ±16. The quantizer can thus be tuned to force these very low samples to zero. This can reduce the SNR for very low level inputs, but will also reduce (or completely eliminate) annoying artefacts.

The last proposed refinement for very low level signals is to use a “noise gate” at the decoder, to progressively reduce the level of the output signal whenever the energy of the decoded signal decreases below a certain threshold. The threshold must be set very low, so that only low level (ideally, almost inaudible) noise is affected. The noise gate can be very useful to reduce the type of noise experienced when only a small proportion of samples in a segment are not set to zero by the quantizer. The effect of a properly configured noise gate is an output signal with cleaner sound between active passages. As a result, the listener focuses his attention more naturally on the speech or audio rather than on the background noise.

7. PERFORMANCE

The noise shaping techniques described in this paper have been tested extensively, since they are part of the G,.711.1 multilayer wideband codec, recently standardized by the ITU-T. Several subjective test reports are available from the ITU-T. In particular, in [8], listening test results from the selection phase demonstrate the improvements from the proposed noise shaping when applied to the G.711 encoder. The known input level dependency of G.711 essentially disappears when applying the noise feedback loop described in Figure 4 and Section 3. Of course, this is at the expense of delay (10 ms frame), since G.711 is a sample-by-sample encoder – although delay could be reduced by using backward LPC analysis. However, the increase in perceived quality for clean speech is dramatic, especially for flat 50-4000 Hz signals as was the case in these experiments. When the input is at a level -36 dB below saturation, the increase in Mean Opinion Score (MOS) is almost 2 points (from 2.5 to 4.4). At nominal level (-26 dB), the increase is still in the order of 1.2 MOS points (from 3.3 to 4.5). The noise feedback loop at the PCM encoder essentially puts the 64 kbit/s narrowband synthesis from the core layer in the saturation zone, i.e. at a quality level statistically equivalent to the original. In this context, the improvements provided by the LBE layer are not as dramatic. But, they make it possible to sustain very high quality in both the narrowband and wideband case, especially for music inputs.

8. CONCLUSION

In this paper, we have proposed techniques for noise shaping in a PCM-based multi-layer speech and audio codec. At the encoder, an adaptation of a “standard” noise feedback loop allows implementing a very efficient noise weighting filter as in AMR-WB speech coding. Further, a post-processor at the decoder allows maintaining the noise shape when using both the core and the upper layer(s). The proposed solution for the enhancement (upper) layers preserves interoperability with G.711 at the core layer, while progressively lowering the shaped noise floor when using the upper layers. This noise shaping framework has been integrated in the G.711.1 multilayer wideband speech and audio codec, recently standardized by the ITU-T. The proposed techniques could easily be extended to higher sampling frequencies, to be used for example in perceptually-optimized word length reduction of high fidelity audio.

REFERENCES

[1] “G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729,” *ITU-T Recommendation G.729.1*, May 2006.

[2] “Pulse code modulation (PCM) of voice frequencies,” *ITU-T Recommendation G.711*, Geneva, November 1988.

[3] C. C. Cutler, “Transmission systems employing quantization,” U.S. Patent No. 2,927,962; March 1960.

[4] H. A. Spang III and P. M. Schultheiss, “Reduction of Quantizing Noise by Use of Feedback,” *IRE Transactions on Communications Systems*, December 1962.

[5] “Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB),” *ITU-T Recommendation G.722.2*, Geneva, January 2002.

[6] B. Atal, M. Schroeder, “Predictive coding of speech signals and subjective error criteria,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, June 1979.

[7] R. A. Wannamaker, “Psychoacoustically Optimal Noise Shaping,” *Journal of the Audio Engineering Society*, Vol. 40, No. 7/8, pp. 611-620, July/August 1992.

[8] ITU-T Tdoc AH-07-28, “VoiceAge results of the qualification tests for the G.711 Wideband extension”, June 2007, Lannion, France.