A NEW BANDWIDTH EXTENSION TECHNOLOGY FOR MPEG UNIFIED SPEECH AND AUDIO CODING

Yuki Yamamoto, Toru Chinen, Masayuki Nishiguchi

Sony Corporation, Tokyo, Japan

ABSTRACT

In January 2012, MPEG finalized the new MPEG-D Unified Speech and Audio Coding (USAC) standard, which enables the coding of a variety of audio content at low bitrates. USAC provides low-bitrate coding by integrating a speech codec and an audio codec into a unified system. In USAC, Predictive Vector Coding (PVC) is added to Enhanced Spectral Band Replication (eSBR) to improve the subjective quality, especially for speech at low bitrates. For speech signals, there is generally a relatively high correlation between the spectral envelopes of low- and high-frequency bands. The PVC scheme exploits this by predicting the high-frequency envelopes from the low-frequency ones, with the coefficient matrices for the prediction being coded by means of vector quantization.

Index Terms— Unified Speech and Audio Coding (USAC), Bandwidth extension, Predictive Vector Coding (PVC)

1. INTRODUCTION

In the field of audio coding, there are two main types of codecs: speech codecs for speech signals and audio codecs for music signals. For low-bitrate coding, separate codecs have been developed and optimized for the two types of signals. However, with the advent of multifunctional devices, there is increasing demand for devices that can handle a variety of audio content, not just speech or just music. So, a codec that can deal with both speech and music signals at low bitrates is needed. To meet this demand, MPEG standardized the Unified Speech and Audio Coding (USAC) in January 2012 [1, 2].

USAC enables the low-bitrate coding of speech and music signals by adaptively selecting either a speech or an audio codec, depending on the characteristics of the input signal. USAC is based on a time domain coding method called Algebraic Code Excited Linear Prediction (ACELP) [3] and a frequency domain coding method called Advanced Audio Coding (AAC) [4]. An ACELP-based codec is used for speech signals and an AAC-based codec is used for music signals.

On the other hand, MPEG Audio uses Spectral Band Replication (SBR) as a bandwidth extension technology [5]. However, since USAC needs to handle not only music but also speech at low bitrates, the subjective quality of the SBR output at low bitrates needs to be carefully examined. Previously developed bandwidth extension technologies include a Linear Predictive Coding (LPC)-based bandwidth extension technology in Extended Adaptive Multi-Rate Wideband (AMR-WB+) [6], Spectral Band Replication (SBR) technology in MPEG Audio, and prediction-based technologies that use codebook mapping methods to predict a high-frequency (HF) envelope from a low-frequency (LF) one [7, 8]. This paper describes a new bandwidth extension technology that outperforms previous technologies. It is prediction-based, and it provides better performance than other technologies by integrating the information on predictions and prediction errors into a codebook that is generated by means of vector quantization. The results of subjective listening tests comparing our new technology with SBR technology show that our new technology provides better quality at low bitrates, especially for speech signals. As a result, this technology was adopted in MPEG USAC.

2. PREVIOUS APPROACHES

An SBR [5] decoder generates an HF signal by applying copy-up methods to a core decoded signal in the QMF domain and by envelope adjustment using the transmitted HF envelope data which is coded according to a scheme of delta and entropy coding. The delta-coding scheme calculates the differential values of the energy in HF bands along the time or frequency direction and uses entropy coding methods to code the values. At low bitrates, this scheme provides high quality for music signals, but not for speech signals, because the differential values along the time or frequency direction are generally much larger for speech than for music.

The LPC-based bandwidth extension technology in AMR-WB+ [6], which was developed for low rather than high bitrates, does not provide as high a quality as SBR technology. One way of improving the quality to the level of SBR technology is to increase the transmission ratio of the coefficients of the Linear Prediction (LP) filter, but that necessitates a higher bitrate.

There is an alternative technology that uses the correlation between LF and HF envelopes to generate an HF signal [7, 8]. It employs one-to-one codebook mapping to

generate the LP coefficients of an HF signal from the LP coefficients of an LF signal. Although this technology does not require that the encoder send side information to the decoder, the quality is very low due to the error in the predicted LP coefficients of the HF signal. To solve this problem, M. Werner and G. Schuller devised a technology that uses the codebook mapping of one entry of LP coefficients of the LF signal to several entries of LP coefficients of the HF signal. The encoder sends the index of the entries of the LP coefficients of the HF signal, which the decoder side needs, as side information [9].

3. NEW APPROACH

This paper describes a new bandwidth extension technology that provides high quality at low bitrates for both speech and music signals. As mentioned above, in SBR technology, the scheme of delta and entropy coding for an HF envelope works fine with music signals, but not for speech signals, due to the larger delta values, which result in a higher bitrate. We solved this problem by adding Predictive Vector Coding (PVC) to SBR technology. The PVC scheme exploits the correlation between LF and HF envelopes: It combines information on the HF envelopes predicted from LF envelopes with prediction errors into a codebook, thereby providing high quality for speech signals at low bitrates. In this technology, the PVC scheme and the delta coding scheme for HF envelopes are adaptively switched, depending on the characteristics of the input signal.

To make an encoder containing a PVC encoder (Fig. 1), we added a core decoder, an analysis QMF bank, and a PVC encoder to an SBR encoder. The core decoder decodes the data encoded by the core encoder and feeds the decoded data to the analysis QMF bank. This QMF bank sends the QMF subband samples in the LF range of the core decoder output samples to the PVC encoder. The other QMF bank sends the QMF subband samples in the HF range of the input PCM samples to the PVC encoder. The PVC encoder first obtains energies for each group of subbands in the LF and HF ranges. The LF and HF energies are used as the LF and actual HF envelopes, respectively. Next, the LF envelope is multiplied by a prediction coefficient matrix, and a prediction error vector is added to it, which yields a predicted HF envelope. The PVC scheme employs a codebook containing various prediction coefficient matrices and error vectors. The encoder selects the particular prediction coefficient matrix and error vector that provide the lowest difference between the predicted and actual HF envelopes. The index of the selected entry is transmitted as a 7-bit value in the bit stream. The following sections describe how the codebook is generated.

3.1. Prediction coefficient matrix

Unlike music signals, speech signals generally exhibit a relatively high correlation between the LF and HF envelopes.

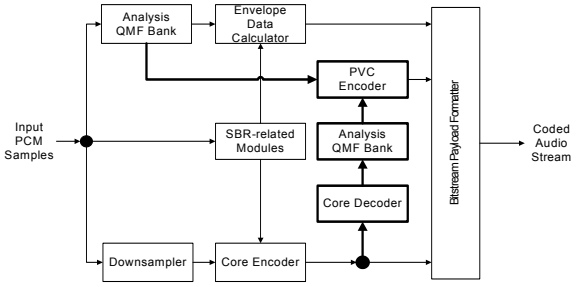

Fig. 1. Block diagram of encoder containing PVC encoder

This is because speech signals consist of a single audio object, while music signals are composed of multiple objects. In the PVC scheme, this characteristic is exploited in predicting the HF envelope from the LF envelope. The predicted HF envelope is given by the equation

$$\begin{pmatrix} P(1) \\ P(2) \\ \cdot \\ \cdot \\ \cdot \\ P(8) \end{pmatrix} = \begin{pmatrix} C(1,1), C(1,2), C(1,3) \\ C(2,1), C(2,2), C(2,3) \\ \cdot \\ \cdot \\ \cdot \\ C(8,1), C(8,2), C(8,3) \end{pmatrix} * \begin{pmatrix} L(1) \\ L(2) \\ L(3) \end{pmatrix} \qquad (1)$$

where $P(1), P(2), ..., P(8)$ are the predicted subband energies in the HF range, $\boldsymbol{P}$; $L(1), L(2), L(3)$ are the subband energies in the LF range, $\boldsymbol{L}$; and $C(1,1), C(1,2), C(1,3), C(2,1), C(2,2), C(2,3), ..., C(8,3)$ are the prediction coefficients, $\boldsymbol{C}$. Here, we assume that the number of HF subbands is 8. To predict the HF envelope from the LF envelope (Fig. 2), linear regression is used to produce the prediction coefficient matrices, $\boldsymbol{C}$, so as to minimize the cost function

$$Res = \sum_{ib=1}^{8} \{\hat{E}(ib) - P(ib)\}^2 \qquad (2)$$

where $\hat{E}(1), \hat{E}(2), ..., \hat{E}(8)$ are the actual subband energies in the HF range, $\hat{\boldsymbol{E}}$; and *ib* is the index of the subband.

3.2. Prediction error vector

As mentioned above, although prediction-based technologies enable a great reduction in the bitrate, they do not provide high quality because of the large prediction error. We reduce this error by adding the prediction error vector, $\boldsymbol{R}$, that minimizes the prediction error to the predicted HF envelope, $\boldsymbol{P}$, to yield the predicted and error-compensated HF envelope, $\boldsymbol{E}$ (Fig. 3):

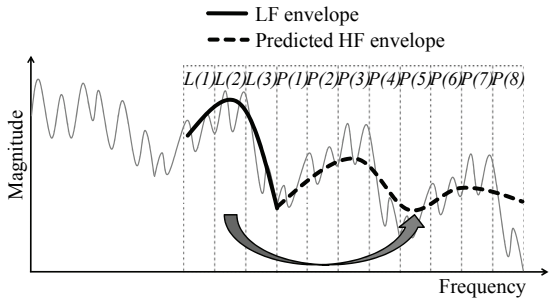

Fig. 2. Prediction of HF envelope

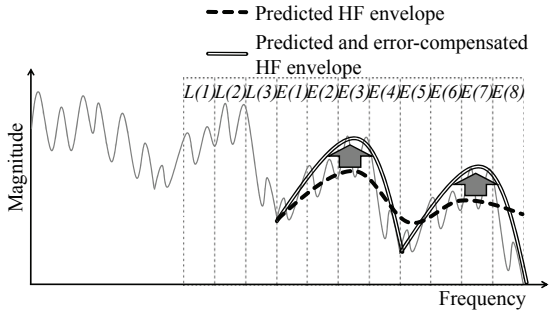

Fig. 3. Minimization of prediction error

$$\boldsymbol{E} = \boldsymbol{P} + \boldsymbol{R} \tag{3}$$

A codebook for the prediction error vector, $\boldsymbol{R}$, is generated by vector quantization so as to minimize the cost function

$$Dis = \sum_{ib=1}^{8} \{\hat{R}(ib) - R(ib)\}^2 \tag{4}$$

where $R(1), R(2), ..., R(8)$ are the elements of $\boldsymbol{R}$; and $\hat{R}(1), \hat{R}(2), ..., \hat{R}(8)$ are the elements of $\hat{\boldsymbol{R}}$, the actual prediction error vector, which is defined by the following equation:

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{E}} - \boldsymbol{P} \tag{5}$$

The actual prediction error vector, $\hat{\boldsymbol{R}}$, is the difference between the actual HF envelope, $\hat{\boldsymbol{E}}$, and the predicted HF envelope, $\boldsymbol{P}$.

3.3. Combining entries for prediction coefficient matrix and prediction error vector

An analysis of prediction coefficient matrices and prediction error vectors shows that several prediction error vectors can correlate with a given prediction coefficient matrix. The error vector providing the highest correlation is selected from among them and is integrated into the codebook, and its entry is mapped to the entry of the selected prediction coefficient matrix. In the PVC scheme, there are a total of 128 entries in the codebook. This number of entries requires 7 bits. These bits are transmitted to the decoder as side information every time when the correlation between the LF and HF envelopes changes. As for the size of the codebook in the PVC scheme, it is 1096 bytes, which is much smaller than for other prediction-based technologies [9].

4. EVALUATION

In order to compare the subjective quality of our new technology with that of the SBR technology, subjective listening tests were conducted using MUSHRA methodology [10] at 8 and 12 kbps mono. The number of experienced listeners (25-33) who participated in a given test varied depending on the test items. The listeners assessed the subjective quality of 15 standard test items selected by the MPEG Audio sub group and 3 additional speech items. In the tests, the border frequencies between the LF and HF envelopes were 4.5 kHz for 8 kbps mono and 4.0-6.2 kHz for 12 kbps mono; and the upper frequencies of the HF envelopes were 9.6 kHz for 8 kbps mono and 12.0 kHz for 12 kbps mono.

4.1. Listening test results

Figures 4 and 5 show the differential scores of our new technology (SBR/PVC) - SBR technology in MUSHRA scale for each test item and for all the items together. Vertical bars around each score indicate the 95 % confidence intervals using a Student's t-distribution. For 8 kbps mono, the performance of our new technology is statistically superior to that of SBR technology for 8 test items and for all the items together. For 12 kbps mono, the performance is statistically superior for 5 test items and for all the items together. Thus, our new technology provides better quality, especially for speech test items, for both 8 and 12 kbps mono.

4.2. Bitrate results

Figures 6 and 7 compare the average bitrates of side information for our new technology (SBR/PVC) and SBR technology for each category of test items and for all the items together. For 8 kbps mono, the average bitrates are almost the same for the two technologies. Furthermore, the side information was analyzed from a viewpoint of how often HF envelopes are transmitted. The analysis results show that for our new technology, HF envelopes are transmitted 37.0 times per second and for SBR technology, HF envelopes are transmitted 28.9 times per second. This means that HF envelopes are transmitted more often for our new technology than for SBR technology. Therefore, this is considered to contribute to improve the quality of HF signal.

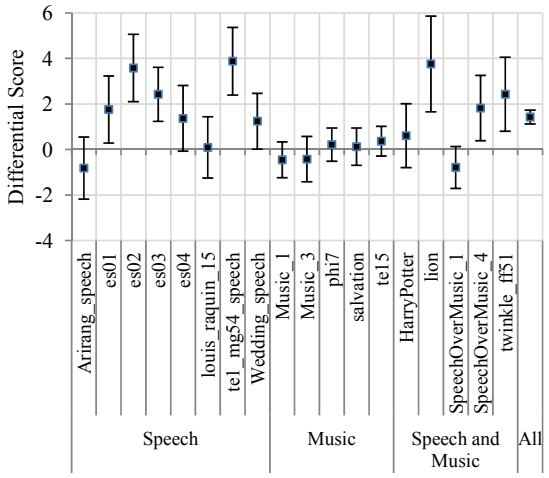

Fig. 4. Differential scores of our new technology (SBR/PVC) - SBR technology for 95% confidence intervals at 8 kbps mono

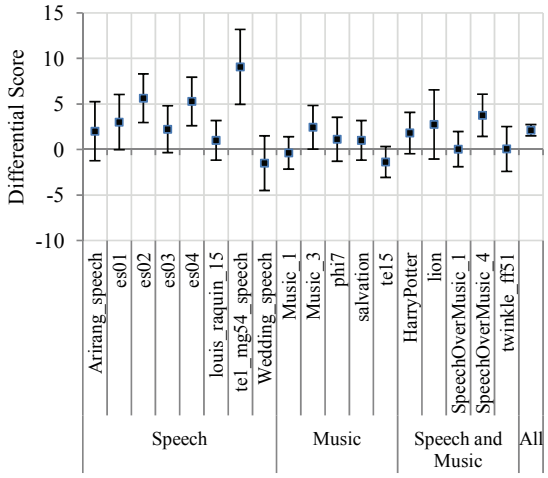

Fig. 5. Differential scores of our new technology (SBR/PVC) - SBR technology for 95% confidence intervals at 12 kbps mono

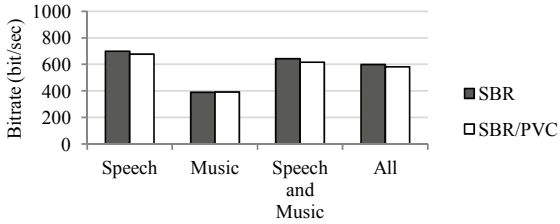

Fig. 6. Bitrate results at 8kbps mono

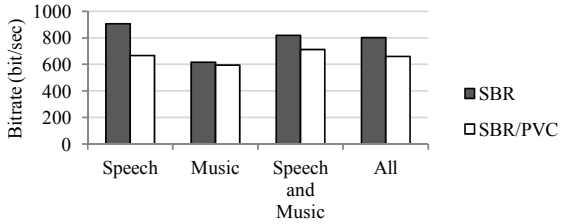

Fig. 7. Bitrate results at 12kbps mono

For 12 kbps mono, according to the analysis of the side information, the number of HF envelopes transmitted per second are 37.5 for our new technology and 37.8 for SBR technology, which are almost the same. However, it is shown in Figure 7 that the average bitrate for our new technology is reduced by 26.5% for speech test items and 12.9% for speech and music test items compared to SBR technology. These saved bits for our new technology are consumed by the core codec. Therefore, this is considered to contribute to increase the quality of LF signal that the core codec encodes.

5. CONCLUSION

This paper concerns a new bandwidth extension technology that provides high quality at low bitrates for both speech and music signals. The results of subjective listening tests show that the performance of the technology is statistically superior to that of SBR technology for 8 test items and for all the items together for 8kbps mono and is statistically superior for 5 test items and for all the items together for 12 kbps mono. Thus, this new technology offers a great improvement in subjective quality over SBR technology. It has been adopted in MPEG USAC to improve subjective quality at low bitrates. Future work will involve further testing of the techniques for generating prediction coefficient matrices and prediction error vectors and improvement of the cost functions in the regression and vector quantization techniques.

6. REFERENCES

[1] M. Neuendorf et al., "MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of all Content Types", in 132nd Convention of the Audio Engineering Society, Budapest, Hungary, 2012.

[2] ISO/IEC 23003-3:2012, "MPEG audio technologies, Part 3: Unified speech and audio coding", 2012.

[3] C. Laflamme, J. -P. Adoul, R. Salami, S. Morissette and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic celp", in IEEE International Conference on Acoustics,

Speech, and Signal Processing, vol. 1, pp. 13-16, Toronto, Canada, 1991.
[4] ISO/IEC 14496-3:2009, "Coding of audio-visual objects, Part 3: Audio", 2009.
[5] M. Dietz, L. Liljeryd, K. Kjörling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding", in 112th Convention of the Audio Engineering Society, Munich, Germany, 2002.
[6] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami and A. Taleb, "AMR-WB+: A new audio coding standard for 3rd generation mobile audio services", in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1109-1112, Philadelphia, PA, USA, 2005.
[7] J. Epps and W. Holmes, "A new technique for wideband enhancement of coded narrowband speech", Proc. IEEE Workshop on Speech Coding, pp. 174-176, 1999.
[8] J. A. Fuemmeler, R. C. Hardie and W. R. Gardner, "Techniques for the Regeneration of Wideband Speech from Narrowband Speech", EURASIP Journal on Applied Signal Processing, 2001.
[9] M. Werner and G. Schuller, "An SBR tool for very low delay applications with flexible crossover frequency", in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 353-356, Dallas, Texas, USA, 2010.
[10] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality levels of coding systems", International Telecommunication Union, Geneva, Switzerland, 2003.