A Scalable Variable Bit Rate Audio Codec Based on Audio Attention Analysis

Bo Hang
School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang, Hubei, 441053, China

Yi Wang
School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang, Hubei, 441053, China

Changqing Kang
School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang, Hubei, 441053, China

Abstract
In surveillance, the audio signal of emergency event is required to have a high quality in order to reserve the event information as much as possible. In the surveillance heterogeneous networks, network bandwidth variation is large. We need a flexible and efficient audio compression coding scheme to adapt to bandwidth variation. For this requirement, we propose a scalable variable bit rate encoding method. Firstly, the attention value of each audio signal frame is calculated based on the audio attention model, thereby judging whether the frame of the audio signal contains attention events or not. If yes, a bandwidth extension enhancement layer and a signal-noise-ratio (SNR) scalable enhancement layer are introduced to enhance the quality of base layer encoded audio signal frame; if not, only the base layer is used to encode the audio signal. The final bit-stream contains the base layer bits for all frames and the enhancement layer bits for the audio signals which contain attention events. The test results show that the coding bit rate of proposed coding algorithm have an average increase of 5.92% for the test sequences. In the subjective audio quality tests, the audio coding quality improvement is about 1 score. The attention audio quality of the proposed scalable variable bit rate encoding scheme is improved, with the cost of increasing a few bits of side information, then saving storage space and transmission rate, and enhancing the adaptability of stream for bandwidth variation.

Keywords: Audio coding, Attention analysis, Scalable, Variable bit rate, Bandwidth

1. INTRODUCTION

In the audio surveillance, the audio signal needs to be compressed because of the limitation of network bandwidth and storage space (Yan, 2016). High compression rate encoding method can save transmission and storage resources, but also bring audio signal quality loss. On the other hand, some audio segments, which may include some interesting events, should be encoded with high quality, to reserve more information. Therefore, we can use variable bit rate encoding method. Audio segments with high degree of attention should be encoded with higher coding bit rate for high quality. And other audio segments can be encoded with a lower bit rate for high compression efficiency.

Variable bit rate encoding methods use different bit rate to encode audio signals according to audio signal features and network bandwidth (Westerink, Rajagopalan and Gonzales, 1999; Song and Kuo, 2001; Wang, Ji, Zhao, Xie and Kuang, 2013). In the case of a wide network bandwidth or complex audio signals, high bit rate is adopted to encode audio signals. Otherwise, low bit rate will be adopted.

There are two types of variable bit rate encoding methods. One type is different coding schemes for different situations. For example, in speech coding, comfortable noise filling is used for silent audio signal, and speech synthesis models coding for speech signals (Freeman, Cosier, Southcott and Boyd, 1989) . Furthermore, when the network bandwidth is narrow, the lower precision quantization codebooks are used. And higher precision quantization codebooks are used when network bandwidth is wide (Makinien, Bessette, Bruhn, Ojala, Salani and Taleb, 2005).

Another type is embedded bit rate coding, that is scalable coding method.

In scalable audio coding scheme, there are one core layer and one or more enhancement layers. The core layer coding is use to ensure the basic quality of the reconstructed signals. The enhancement layers improve the quality of reconstructed audio signal gradually by increasing encoding SNR or extending the frequency bands. The more enhancement layers received, the higher the decoded audio quality is (Petrovsky, Herasimovich and Petrovsky, 2015).
When the network bandwidth variation is large, the scalable coding scheme can drop enhancement layers in bit stream, instead of dropping audio segments, reserving only the core layer for decoding, that ensure a continuous sequence with basic decoding quality. In a better network conditions, scalable coding scheme can use the enhancement layer in bit stream to enhance the quality of the core layer. Therefore, the scalable encoding requires only one encoding to get a bit stream, which is flexible to adapt to different network bandwidth.

Compared with the former type of variable bit rate audio coding methods, the scalable audio coding methods have two advantages. Firstly, scalable coding methods have good compatibility with present audio codec (Kandadai and Creusere, 2008). The existing audio codec do not need make any change, just enhancing audio quality by enhancing layer coding. Secondly, scalable audio coding can be more flexible to adapt the fluctuation of network bandwidth. The encoder encodes the audio signal in highest bit rate without knowing the network bandwidth situation. In transmission or decoder, the network devices or terminal devices can abandon or reserve the enhancement layer, according to the situation of network or other resource constraints, to adjust bit rate and adapt to the fluctuation of network bandwidth (Shu, Yu and Rahardja, 2011), without feedback to encoder.

It is hard to provide optimal audio coding stream for all surveillance network and devices in the heterogeneous surveillance networks (Kumar, Pande and Mittal, 2010). Scalable audio coding method is an effective means to solve the problem. The method can ensure the basic audio quality of surveillance via core layer and improve the reconstruction quality via enhancement layer (Reyes and Candéas, 2010). Therefore, we use scalable variable bit rate scheme to enhance the surveillance audio signal quality in this work. In section 2, we introduce the audio coding framework for surveillance. Section 3 and section 4 describe the bandwidth scalable enhancement layer coding method and SNR scalable enhancement layer coding method respectively. Experiments are presented in section 5. And section 6 is conclusions.

2. SCALABLE AUDIO CODING FRAMEWORK

In this work, we propose a scalable audio coding for surveillance based on attention analysis (Hang and Hu, 2010; Hang, Wang and Kang, 2016). The encoder frame is shown in figure 1.

![Figure 1. Surveillance audio encoder frame](image)

In this frame, the results of attention audio will be put into the bit stream for real time alarm, event retrieval and audio signal enhancement etc. For audio signal enhancement coding, we use scalable audio enhancement layers to improve the attention audio signal coding quality, together with the application of the audio attention model.
The adaptive scalable variable bit rate frame is proposed on the basis of audio attention analysis and detection. The principle is shown in Figure 2.

**Figure 2.** The framework of adaptive variable bit rate method

Since the data, which include high attention level audio events and need to be encoded with high bit rate and high quality, are just a small part of the mass surveillance data, it will consume large amounts of network bandwidth to encode enhancement layers for all audio signals, resulting in unnecessary network bandwidth and storage space resources waste. Therefore, in the proposed frame, when there is no emergency attention event detected, enhancement layer encoder is turned off. At this time, only the core layer of the encoder will code the acquired audio signal, in order to ensure the basic quality of the background sound. When there is any attention events detected by the attention model, the encoder will automatically turn on the enhancement layer coding to improve the quality of encoded audio signal including attention event. When the attention events ending is detected, the enhancement layer turn off again, to reduce coding bit rate. Through this adaptive enhanced coding framework, with a minimum average bit rate cost, surveillance audio quality is enhanced, and storage space saved. In the frame, the enhancement layers include bandwidth scalable extension layer and SNR scalable enhancement layers. The two types of layers are introduced in following sections.

### 3. BANDWIDTH SCALABLE LAYER

Bandwidth scalable coding framework is based on bandwidth extension coding technology. The main information of the audio signal is concentrated in the low frequencies. And the human ear is more sensitive to low frequency signal. So, the low frequency bands are coded by a fine coding method. And High frequency bands do not need to be coded by such a fine coding method, because human ears are not as sensitive to the fine structure high frequency bands as to low frequency bands. The method, which encodes high frequency bands, decode high frequency signal with a slight increase of side information on the basis of low-band signal bit stream. Therefore, in this bandwidth scalable framework, the core layer is used to decode narrowband signal with basic quality. And based on the core layer bit stream, the high frequency band signal is obtained by extension layer bit stream to construct the full band signal to enhanced the decoded audio quality.

Bandwidth extension algorithm used currently by various standardization organizations, include Spectral Bandwidth Replication (SBR) of MPEG (Moving Picture Experts Group) (Dietz, Liljeryd, Kjorling, and Kunz, 2002), Time Domain BandWidth Extension (TDBWE) of Telecommunication Standardization Sector of the International Telecommunications Union (ITU-T) G.729.1 audio standard (Geiser, Jax, Vary, Taddei, Schandl, Gartner and Ragot, 2007), ultra wideband mobile audio codec standard AMR-WB + bandwidth extension of 3rd Generation Partnership Project (3GPP) (Makinen, Bessette, Bruhn, Ojala, Salami and Taleb, 2005) and the Audio and Video Coding Standard Workgroup of China (AVS) Part 10 mobile audio bandwidth extension algorithm (Zhan, Choo and Oh, 2009). In order to adapt complex audio signals in surveillance, a bandwidth extension algorithm based on multi-mode prediction is proposed.

Different types of high-frequency portion of the audio signals have different characteristics. Some noise-like, and some have significant harmonic components. Thus, different high frequency encoding mode should be chosen for different types of signals. We can use multi-mode including inner-frame, inter-frame signals or a white noise to predict high-frequency signals of the current frame, and extract high frequency sub-band gain factors for each mode, and then compare signal to noise ratio (SNR) values of all prediction modes to select the most accurate coding mode. In this paper, we use the copy or folded low frequency signals to predict high frequency signals as the intra-frame prediction, which can predict the harmonic components effectively. High-frequency signals before the current frame are used to predict the high-frequency signals of current frame as the inter-frame prediction, which is suitable for short term stationary signals. And white noise is used to predict noise-like signal. Using this approach we can obtain more optimum prediction method from multi prediction modes, so as to improve the quality of encoded audio high frequency signal.

Four excitation modes are encoded by 2 bits (00,01,10,11). The definitions of the bits are shown in Table 1. Gain factors of eight sub-bands compose two four-dimensional vectors, and coded by 7-bit vector quantization respectively. So each frame signal is coded by a total of 16 bits, with encoding bit rate of 0.8 kbps. Coding diagram is shown in Figure 3.
### Table 1: Four excitation modes and corresponding coding bits

<table>
<thead>
<tr>
<th>Predicting mode</th>
<th>Mode coding (2 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low frequency copied</td>
<td>00</td>
</tr>
<tr>
<td>Low frequency folded</td>
<td>01</td>
</tr>
<tr>
<td>High frequency signal of previous frame</td>
<td>10</td>
</tr>
<tr>
<td>White noise</td>
<td>11</td>
</tr>
</tbody>
</table>

### 4. SNR SCALABLE CODING

The above bandwidth scalable coding technology can improve decoded audio quality by using bandwidth extension, with only a few bits increasing for high frequency signal parameters on the basis of the core layer encoder. But since the bandwidth extension technology can only coding high frequency signal, this technology can not reduce the coding noise of core layer codec. Therefore, except bandwidth scalable coding, some signal-noise-ratio scalable coding methods were proposed to reduce coding noise progressively and improve coding signal quality [17, 18].

This paper proposes a scalable coding method with two-layer structure, wherein the core layer mode code rate 12kbps, 16kbps rate enhancement layer mode to the first layer, the second layer pattern 20kbps rate enhancement layer, the quality of the signal may be raised step by step. The method is based on AVS-S (AVS for Surveillance) design and implementation of the core layer. Coding method shown in FIG. 5

![Figure 3. High frequency band encoding method](image1)

#### 5. Experimental results

##### 5.1. Adaptive bit rate test

The bandwidth scalable extension layer and SNR scalable enhancement layer form the enhancement layers for surveillance oriented scalable coding. The bit rate of each frame of the enhancement layer is 8 kbps. Audio signals are encoded according to the attention audio detection results. If the present frame is an attention frame, the enhancement layer will be added. The frame is encoded by only the core layer if it is not an attention frame. In our test, the audio sequences with different SNR are encoded by the core layer with and without adaptive enhancement layer separately. The encoding bit rates of these sequences are recorded and compared with each
other. The test results are shown in table 2.

**Table 2.** The bit rate comparing between encoder with/without adaptive enhancement layer

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Core layer</th>
<th>Adaptive variable bit rate</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mix1f1b0db.wav</td>
<td>16</td>
<td>16.558</td>
<td>3.49%</td>
</tr>
<tr>
<td>Mix1f1b5db.wav</td>
<td>16</td>
<td>16.627</td>
<td>3.92%</td>
</tr>
<tr>
<td>Mix1f1b15db.wav</td>
<td>16</td>
<td>17.017</td>
<td>6.35%</td>
</tr>
<tr>
<td>Mix2f1b0db.wav</td>
<td>16</td>
<td>17.333</td>
<td>8.33%</td>
</tr>
<tr>
<td>Mix2f1b5db.wav</td>
<td>16</td>
<td>16.643</td>
<td>4.02%</td>
</tr>
<tr>
<td>Mix2f1b15db.wav</td>
<td>16</td>
<td>17.51</td>
<td>9.44%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>16</strong></td>
<td><strong>16.948</strong></td>
<td><strong>5.92%</strong></td>
</tr>
</tbody>
</table>

From the table 2, we can see that the average bit rate increase about 5.92%. It is because that the ratio of the audio signals with attention events in all audio signals acquired is low. Therefore, the quality of the attention audio in the surveillance audio can be improved with increasing a small mount of bit rate, which can save a lot of storage space and transmission rate.

5.2. **Subjective audio coding quality tests**

Experiment purpose: In this experiment, we decode the bit stream with core layer only and core layer plus enhancement layer separately, and comparing the decoded audio sequences with each other via subject listening test, to verify whether the scalable audio coding algorithm can enhance the attention audio quality. In addition, the audio sequences encoded by the proposed method and G.729.1 method are tested by subject listening test to verify the coding quality of the proposed audio codec.

Test environment: All experiments were carried out in listening laboratory. The experimental design and test procedures comply with the contents of the ITU-T P.800 proposal.

Experimental Materials: The test sequences are AVS workgroup specified 17 MPEG test sequences (es01, es02, es03, sc01, sc02, sc03, si01, si02, si03, se01, se02, sm01, sm02, sm03, na04, or07mv, or08mv).

Test standard: the tests are accordance with the international standards - ITU-T P800 Methods for Subjective Determination of Transmission Quality.

Experimental method: In audio quality subjective listening test, Comparison Mean Opinion Score (CMOS) is used. Each testing material consists of Ref/A/B, in which Ref is original uncoded signal, A/B are all decoded signals. In test, if A is the decoded signal results of proposed BWE method, then B is the decoded one for reference AVS P10; vice versa. Specific testing method is as below:

(1) The position of Ref signal in every test is fixed, A and B is allocated randomly, and they are unknown to listeners.

(2) Listeners should be trained to be familiar with the whole testing process, and try to learn about some representative distortion presented frequently in testing materials.

(3) In high-quality marking system, 7 levels are used, shown as table2, and the grades delivered by subjects should be integer.

**Table 3. Levels Comparison Standard**

<table>
<thead>
<tr>
<th>Comparison of the Stimuli</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A is much better than B</td>
<td>+3</td>
</tr>
<tr>
<td>A is better than B</td>
<td>+2</td>
</tr>
<tr>
<td>A is slightly better than B</td>
<td>+1</td>
</tr>
<tr>
<td>A is the same as B</td>
<td>0</td>
</tr>
<tr>
<td>B is slightly better than A</td>
<td>-1</td>
</tr>
<tr>
<td>B is better than A</td>
<td>-2</td>
</tr>
<tr>
<td>B is much better than A</td>
<td>-3</td>
</tr>
</tbody>
</table>

The scoring rule is that the tested one which is much closed to the original one deserves high score, and the specific score is decided by the approaching degree.

(4) The testing result consists of average score and 95% confidence interval, and all testing results are needed to calculate statistical variance.

\[
\sigma = \sqrt{\frac{N \sum_{k=1}^{N} x_k^2 - (\sum_{k=1}^{N} x_k)^2}{N(N-1)}}
\]

95% confidence interval: \( \mu \pm 1.96\left(\frac{\sigma}{\sqrt{N}}\right) \)

Test results and analysis: The subject listening tests are shown in figure 6 and figure 7.
As it can be seen from Figure 5, at 20 kbps bit rate, encoding quality of the proposed algorithm is similar with ITU G.729.1 audio coding algorithm.

From the subjective listening test in figure 6, the enhancement layer can improve the audio coding quality significantly.

6. CONCLUSIONS

In the proposed scalable coding framework based on attention analysis, the enhancement layer encoding will be turned on or off adaptively based on the attention analysis results. This method only encode the audio signal including attention events, to reduce the network bandwidth and storage space consumption of surveillance audio signals by a more flexible bit rate coding way. Test results show that by the use of adaptive variable rate mechanism, the average coding bit rate increased by only 5.92 percent. The proposed method effectively reducing the network and storage resources consumption of the scalable enhancement layer coding.

Acknowledgements

We give thanks to the reviewers and editors for their valuable comments, questions, and suggestions. This work was supported in part by National Nature Science Foundation of China (61201247) and Nature Science Foundation of Hubei Province (2011CDB322).

REFERENCES


