A SPATIAL PRIORITY BASED SCALABLE AUDIO CODING

Li Gao, Ruimin Hu, Yuhong Yang

National Engineering Research Center for Multimedia Software,
Computer School of Wuhan University, China
Research Institute of Wuhan University in Shenzhen, China

ABSTRACT

A spatial priority scheme for scalable audio coding is presented in this paper. To improve the coding quality of important sounds with high attention, especially the moving sound, spatial information is introduced to assign the priorities of frequency subbands. Spatial cues and distance features are extracted in frequency subbands to represent the sound with fast changing direction and distance. Coding priorities are assigned to different frequency subbands according to the energy and spatial information. With trivial added side information and complexity, experimental results show that the perceptual quality is improved especially for the sound with high attention, especially the moving sound in scalable audio coding.

Index Terms— Spatial priority, Scalable audio coding, Spatial information

1. INTRODUCTION

Scalable audio coding (SAC) allows encoder to compress audio signal at high bit-rate and decoder to restore signal at different bit-rates with different fidelity levels. SAC is an good optional choice to provide stable service quality for various different systems and user terminals according to heterogeneous networks, fluctuating bandwidth and the different user terminal devices. The structure of SAC includes a most important core layer and several enhancement layers in terms of finer quantization or higher sampling rate. To encode the most important information and get the gradually improved audio quality with the increased layers, numerous approaches have been proposed in SAC schemes. The scalability criteria in SAC schemes mainly focus on the scalability in signal-to-noise ratio (SNR), bandwidth and bit-plane.

The early methods of scalability in SNR encode the scalar quantization indexes of residual signal from the most significant data to the least significant data[1]. In bandwidth scalability, the frequency spectrum coefficients are coded from the most important subbands to the least important subbands. Since the subbands with higher energies are considered more important than the subbands with lower energies, the priorities of subbands are assigned according to the energies of subbands[2], which was standardized as ITU-T G.729.1 standard for the scalable coding of speech and audio signals. In bit-plane coding, the bit-planes of frequency spectrum coefficients are coded from the most significant bit (MSB) to the least significant bit (LSB), as done in MPEG-4 scalable lossless (SLS) coding standard[3]. Since the human auditory sensory system has different sensitivities in different frequency regions, auditory perception importance criterion was used in [4] as a substitute for the sequential bit-plane coding in MPEG-4 SLS, and the frequency regions were divided into different regions in which the bit-planes were assigned with different priorities to ensure the lower frequency spectrum coefficients to be coded first.

According to the priority assignment criteria in SAC schemes mentioned above, the frequency spectrums with higher energy or in lower frequency regions which are more sensitive to auditory sensation are assigned higher priorities. Energy and frequency are two of the basic and major attributes of sound source and mainly considered in current priority schemes. However, since the human auditory sensory system has different sensitivities in different frequency regions with different energies, the frequency spectrums with highest energy or in lowest frequency do not always correspond to the subbands of the most important sound. For example, in the scene with ambience or background sound, the important sound we pay attention to may have relatively lower energy or higher frequency to the ambience and background sound, which would not be the focus and would suffer with the loss of coding quality in the current SAC schemes. Besides the basic sound quality influenced by energy and frequency, we also concern other important sound features, such as spatial location of the sound. The researches on psychology have discovered that spatial information plays an important role in the auditory attention of human[5]. Spatial information can be used to identify the important sounds, moving sounds and the sounds people pay more attention to. However, energy and frequency in the original mono channel or downmixed channel can provide few information about the spatial information of the sound. It will have to, more or less, bring quality loss to encode some important sounds with high

This work is supported by NSFC(No.61231015, No.61201340, No.61102127, No.61201169), Hubei NSF(2012FFB04205).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

attention, especially the fast moving sound.

In this paper spatial information is considered for the priority assignment of frequency subbands in SAC. A spatial priority based scalable audio coding (SP-SAC) method is proposed to improve the quality of important sounds with high attention, especially the fast moving sound. Comparisons of proposed SP-SAC with energy priority based SAC (G.729.1) and frequency priority based SAC (MPEG-4 SLS) are performed to justify the performance.

2. SPATIAL PRIORITY BASED SCALABLE AUDIO CODING

Besides energy and frequency, there are still other important information and attributes about the sound, such as some attributes about the spatial information of the sound which play important roles to the auditory cognition of human. Energy, frequency and spatial location information should all be taken into account in the priority criteria of SAC. Although the importance of sound is not decided only by energy or moving speed, but the sound with high energy, or fast changing energy and location will attract more auditory attention of human except when someone selectively pays more attention to the motionless object among moving objects (as in the cocktail party effect). A spatial priority based SAC is proposed here to use the spatial information of the sound to pay more attention to the sound with fast changing location and energy in the priority assignment of frequency subbands.

2.1. Direction criterion

The information of spatial location is contained in the sound. The change of sound location will result in the change of spatial information. Spatial cues are used to represent the direction of sound in each subband. One of the spatial cues, Interaural Level Difference (ILD) [6] between two ears is the dominative spatial acoustic cues for people to discriminate sound direction in horizontal plane [7]. Inter Channel Level Difference (ICLD) between two channels is usually used instead of ILD to represent the directional information of the sound. The difference of ICLD is used here to represent the azimuth change of sound. It is inspired by our previous work in [8] which presented a spatial audio cues based audio attention model.

ICLD is a spatial cue related to the spatial direction of sound source. The value of ICLD is nearly zero for the sound source with no obvious direction information (such as ambience noise) or the sound source locates at the vertical plane. The ICLD difference in instantaneous time is also nearly zero for the motionless sound source. Although ICLD is not enough to indicate the exact direction of the sound source with the existence of ambience noise, the nonzero value of ICLD is still enough to tell the existence of directional sound source and the value of ICLD difference can show the directional change of sound source.

As shown in Fig. 2, ICLD difference of each subband at four discontinuous time frames of the audio signal in Fig. 1 is calculated. T1 refers to some moment before a woman's voice appears and T2-T4 correspond to different moments when the woman is at different locations.

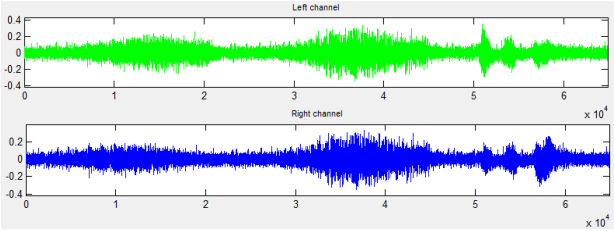

Fig. 1. Street sound with rushing motorcycle sound and running woman's voice.

The ICLD difference at T1 (ICLD difference between the time frame T1 and the time frame just before T1) is zero for there is no obvious directional sound source in the ambience noise. ICLD difference is nonzero at T2-T4 when speaker moves and location changes, which means that the moving sound results in the change of ICLD and also that of ICLD difference. It can be shown in Fig. 2 that apparent differences of ICLD appear mainly at the subband 3,5,6,12,13 and 14, corresponding to the main sound frequencies of the speaker and the most important frequencies we need to focus on.

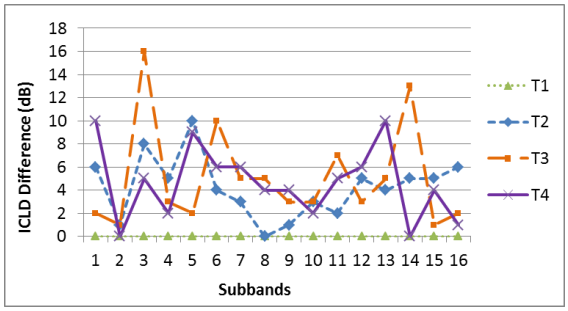

Fig. 2. ICLD difference of each subband at four discontinuous frames of the audio signal in Fig. 1.

As a comparison, the energy of each subband at the same time frames as in Fig. 2 is shown in Fig. 3. It is observed that although there are mainly ambience sounds with noises at T1, the energies of different subbands are not at the identical level, even abrupt change between subbands exists. At T2-T4, although the woman's voice is at different locations, obvious energy differences between subbands do not exist. It can be concluded that energy is not effective enough to reflect the importance of subbands.

Based on the fact and analysis above, few information about the location of the woman's voice can be observed from

the energy distributions in subbands. However with the priority assignment criteria described as followed, spatial cues (such as ICLD) in subbands can be a guide to find out the most important frequency spectrum subbands and accordingly to improve the coding quality for the moving sound with high attention.

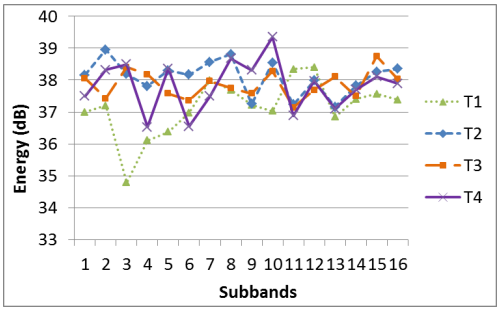

Fig. 3. Energy of each subband at four discontinuous frames of the audio signal in Fig. 1.

Firstly a current frame signal is obtained from the audio stream. Duration of each frame is t_f. Then the ICLD of subband i is calculated as

$$s_i = 10\, lg \frac{I_{Li}}{I_{Ri}}, \tag{1}$$

where $i \in [1, N]$ and I_{Li}, I_{Ri} denote the energy of left channel and right channel of the subband i, respectively. Suppose the current frame is frame k, a vector of ICLD is obtained as $S_i = \{s_1, s_2, \ldots, s_{N-1}, s_N\}$. N denotes the number of subbands.

If the azimuth of the sound source changes quickly, the short-term variation of ICLD would be large apparently. So ICLD difference between the frame k and the frame with time interval ΔT before is calculated to represent the azimuth change of the fast moving sound. ICLD difference between frame k and frame $k - \Delta T/t_f$ of each subband is computed as

$$D_{S_i} = \{d_{S1}, d_{S2}, \ldots, d_{S(N-1)}, d_{SN}\}, \tag{2}$$

where $d_{Si} = |s_i(k) - s_i(k - \Delta T/t_f)|$.

Since the value of each subband in D_{S_i} represents the change of spatial direction, the subband order sorted by these values could be used to assign the coding priority of frequency spectrum subbands. Priorities of subbands are expressed as

$$P_{S_i} = \{p_{S1}, p_{S2}, \ldots, p_{S(N-1)}, p_{SN}\}, \tag{3}$$

where p_{Si} denotes the sorted order of each subband.

2.2. Distance criterion

The distance feature used here is the same feature as used in [9] to estimate sound source distance, namely the frequency-dependent coherence between the left and right channel signals. The magnitude-squared coherence is calculated by

$$r_i = \frac{|G_{lr}(f,t)|^2}{G_{rr}(f,t)G_{ll}(f,t)} \tag{4}$$

$$G_{lr}(f,t) = < X_l^*(f,t)X_r(f,t) > \tag{5}$$

$$G_{ll}(f,t) = < |X_l(f,t)|^2 > \tag{6}$$

$$G_{rr}(f,t) = < |X_r(f,t)|^2 > \tag{7}$$

Where $X_l(f,t)$ and $X_r(f,t)$ is the short-time spectra of the left and right channel of frequency f at time t, respectively. $G_{lr}(f,t)$ is the estimated cross-spectrum between left and right channel. $G_{ll}(f,t)$ and $G_{rr}(f,t)$ are the estimated spectra of the left and right channel signals, respectively.

Then we get the frequency-dependent coherence of each subband as a vector $R_i = \{r_1, r_2, \ldots, r_{N-1}, r_N\}$. Next the coherence difference between frame k and frame $k - \Delta T/t_f$ of each subband is computed as

$$D_{R_i} = \{d_{R1}, d_{R2}, \ldots, d_{R(N-1)}, d_{RN}\}, \tag{8}$$

where $d_{Ri} = |r_i(k) - r_i(k - \Delta T/t_f)|$. Since the value of each subband in D_{R_i} represents the change of sound distance, the subband order sorted by these values could be used to assign the coding priority of frequency spectrum subbands. Priorities of subbands are expressed as

$$P_{R_i} = \{p_{R1}, p_{R2}, \ldots, p_{R(N-1)}, p_{RN}\}, \tag{9}$$

where p_{Ri} denotes the sorted order of each subband.

2.3. Energy criterion

From energy I_{Li} and I_{Ri} of the left and right channel of subband i, respectively, the energy of subband i is calculated as $I_i = I_{Li} + I_{Ri}$. Then we get the energy of each subband as a vector $I_i = \{I_1, I_2, \ldots, I_{N-1}, I_N\}$. Next the energy difference between frame k and frame $k - \Delta T/t_f$ of each subband is computed as

$$D_{I_i} = \{d_{I1}, d_{I2}, \ldots, d_{I(N-1)}, d_{IN}\}, \tag{10}$$

where $d_{Ii} = |I_i(k) - I_i(k - \Delta T/t_f)|$.

Since the value of each subband in D_{I_i} represents the change of energy, the subband order sorted by these values could be used to assign the coding priority of frequency spectrum subbands.Priorities of subbands are expressed as

$$P_{I_i} = \{p_{I1}, p_{I2}, \ldots, p_{I(N-1)}, p_{IN}\}, \tag{11}$$

where p_{Ii} denotes the sorted order of each subband.

With the criteria described above, the normalized priority of each subband is obtained by

$$np_i = \lambda_1 p_{Si} + \lambda_2 p_{Ri} + \lambda_3 p_{Ii}, \tag{12}$$

where λ_1, λ_2, and λ_3 are weighting coefficients. It can be set that $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$. The priorities of subbands can be transferred in the bit stream as the side information and can be coded for further compression. For the SAC with core codec, the priorities of subbands can be calculated from the decoded core signals without any additional side information.

3. EXPERIMENTAL RESULTS

To verify the performance of the above SP-SAC method, a series of subjective listening tests have been carried out based on comparison mean opinion score (CMOS). Subjective evaluations were conducted based on extended G.729.1 and MPEG-4 SLS codec to compare the proposed spatially priority scheme with those used in G.729.1 and MPEG-4 SLS. The subjective tests were performed by 7 experienced listeners with ages between 22 and 32 who actively working in the domain of audio compression. The test items are listed in Table1.

Table 1. Test items in the listening test

items	Content description	Duration(s)
item 1	es03 female voice from MPEG standard stereo test sequences	7.6
item 2	si02 castanets panned by VBAP with the azimuth from 0° to 180° and mixed with sc01 trumpet	7.7
item 3	audio clip from movie Ice age 3 from time 19:51.000 to 19:56.000 with moving scream and hooves	5.0
item 4	recording street sound including crying woman's voice and rushing motorcycle sound with background sound as Fig. 1	6.3

The subjective result CMOS of proposed SP-SAC compared with energy based priority scheme in G.729.1 at bit-rates of 36, 40, 48, and 64kbps (double mono) is presented in Fig. 4. Compared with the perceptual quality of G.729.1 in terms of CMOS scores, SP-SAC displays obvious improvements for most test items. The improvements are relatively small for one test item es03.wav as this item contains a relatively stationary female voice and its spatial attributes are more stable.

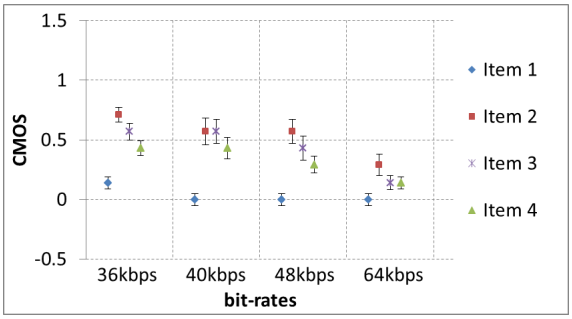

Fig. 4. CMOS of proposed SP-SAC compared with G.729.1 at different bit-rates with 95% confidence interval.

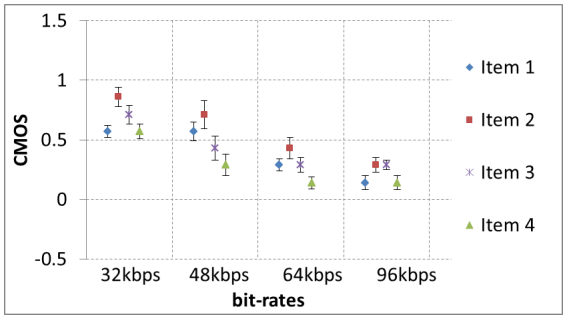

Fig. 5. CMOS of proposed SP-SAC compared with MPEG-4 SLS at different bit-rates with 95% confidence interval.

The performance of SP-SAC is also compared with that of MPEG-4 SLS noncore at bit-rates of 32,48,64, and 96kbps(in total for stereo channels) as presented in Fig. 5. As can be seen from the evaluation results of CMOS, the proposed SP-SAC improves the perceptual sound quality at the low to middle bit-rate range compared with the frequency region based priority scheme. Even at higher bit-rate of 96 kbits/sec, the proposed scheme still displays slight improvements.

4. RELATION TO PRIOR WORK AND CONCLUSION

The traditional energy priority based scalable audio coding methods (such as G.729.1) assign higher priorities to the subbands with higher energies. The frequency priority based scalable audio coding methods (such as MPEG-4 SLS) ensure the lower frequency spectrums which are more sensitive to auditory sensation to be coded first. Besides energy and frequency, there are still other important features about sound, such as spatial information about the location of sound source which can be used to identify the important sounds, moving sounds and the sounds people pay more attention to. Different from the traditional SAC, the proposed spatial priority based SAC in this paper takes into account not only energy and frequency but also spatial information of the sound in the priority assignment of frequency subbands to improve the coding quality of important sounds with high attention, especially the moving sound. Experiments verify the performance of proposed method with trivial added side information and complexity, comparing with the energy and frequency priority based SAC.

5. REFERENCES

[1] Goodman, D., "Embedded DPCM for variable bit-rate transmission," IEEE Transactions on Communications, 28(7), pp. 1040-1046, 1980.

[2] Ragot, S., et al, "ITU-T G. 729.1: An 8-32 kbit/s scalable coder interoperable with G. 729 for wideband telephony and Voice over IP," In Acoustics, Speech and Signal Processing, IEEE International Conference on, pp. IV529-IV532, 2007.

[3] Rongshan Yu, Ralf Geiger, Susanto Rahardja, Juergen Herre, Xiao Lin, Haibin Huang, "MPEG-4 scalable to lossless audio coding," In: 117th Audio Engineering Society Convention, AES, San Francisco, 2004.

[4] Li, T., S. Rahardja, S.N. Koh, "Frequency region-based prioritized bit-plane coding for scalable audio," IEEE Transactions on Audio Speech and Language Processing, 16(1), pp. 94-105, 2008.

[5] Eramudugolla, R., McAnally, K. I., Martin, R. L., Irvine, D. R. F., Mattingley, J. B., "The role of spatial location in auditory search," Hearing Research, 238, pp. 139-146, 2008.

[6] Christof Faller, "Parametric Coding of Spatial Audio," Ph.D.thesis,COLE POLYTECHNIQUE FDRALE DE LAUSANNE, 2004.

[7] Moore BCJ, "An Introduction to the Psychology of Hearing," Fifth. Amsterdam: Elsevier Academic Press, 2004.

[8] Hang, B. and R. Hu., "Spatial audio cues based surveillance audio attention model," In Acoustics, Speech and Signal Processing, IEEE International Conference on, pp. 289-292, Dallas, 2010.

[9] Vesa, S., "Binaural sound source distance learning in rooms," IEEE Transactions on Audio Speech and Language Processing, 17(8), pp. 1498-1507, 2009.