INVITED PAPER

Multichannel Audio Coding Based on Analysis by Synthesis

An overview of familiar multichannel coding techniques and a new coding framework for improving the objective fidelity of decoded signals are presented in this paper.

By Ikhwana Elfitri, Banu Günel, *Member IEEE*, and Ahmet M. Kondoz, *Senior Member IEEE*

ABSTRACT | Spatial hearing enables translation of an auditory scene into a perceived 3-D image by interpreting the acoustic cues related to the sounding objects, their locations, and the physical characteristics of the space. Spatial audio production requires multichannel audio signals in order to convey this information and increase the realism of a real or virtual environment for applications such as the home entertainment, virtual reality, and remote collaboration. As demand to spatial audio continues to expand, efficient coding of multichannel audio content becomes more and more important. This paper provides an overview of some well-known multichannel audio coding techniques and presents a new coding framework for improving the objective fidelity of the decoded signals. A closed-loop encoding system based on analysis-by-synthesis (AbS) principle applied on the MPEG surround (MPS) architecture is described. Comparison results are presented, which show that significant improvements can be achieved with a closed-loop system instead of the conventional open-loop system.

KEYWORDS | Analysis by synthesis (AbS); MPEG surround (MPS); multichannel audio coding; two-to-one (TTO) structure

I. INTRODUCTION

Spatial audio describes the audio signals that convey information about a 3-D sound scene [1]. This scene describes the individual sounding objects, their positions, and the acoustics of the environment even when no visual information is available [2]. In order to convey this spatial information and improve the realism of perceived 3-D scenes, multiple audio signals need to be reproduced by multiple loudspeakers. Multichannel audio is important in many applications for improving the realism of a rendered acoustic environment. Started by the cinema industry, it is now used in home entertainment, digital audio broadcasting (DAB), computer games, music downloading and streaming services, as well as other internet applications, such as teleconferencing and remote collaboration. Moreover, the importance of 3-D audio is projected to increase even further due to the introduction of 3DTV services and the spread of 3-D displays.

There are various 3-D audio reproduction systems available today, such as binaural [1], stereo, 5.1, 7.1, and similar multichannel systems [3], Ambisonics [4] and wave field synthesis (WFS), where a very large number of channels, such as 32, 64, or more, can be used [5]. These are based on either fooling the ears by creating an auditory illusion at the ear location, or creating the exact wavefront of a sound source using secondary sources. The fidelity of the auditory environments reproduced with these techniques differs according to the listening position, the acoustics of the listening environment, and the loudspeaker setup [6]. However, the major difficulty lies with the encoding of the multichannel content in a way that preserves the spatial fidelity for efficient transmission and rendering.

Although compressing multiple channels using mono audio coders has been proposed before as a straightforward solution, further reduction of bandwidth is possible by exploiting the interchannel relations resulting in a more compact representation. Considering interchannel relations also ensures the preservation of spatial cues, which are important for localization of sound sources. Some approaches exploiting multichannel information include parametric stereo (PS) [7]–[9], binaural cue coding (BCC) [10]–[12], spatial audio scene coding (SASC) [13]–[15],

Manuscript received March 7, 2010; revised October 11, 2010; accepted December 7, 2010. Date of publication February 14, 2011; date of current version March 18, 2011. The work of I. Elfitri was supported by the Ministry of National Education, Republic of Indonesia.
The authors are with the I-Lab Multimedia Communications Research, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: I.Elfitri@surrey.ac.uk; B.Gunel@surrey.ac.uk; A.Kondoz@surrey.ac.uk).
Digital Object Identifier: 10.1109/JPROC.2010.2102310

directional audio coding (DirAC) [16]–[18], spatially squeeze surround audio coding (S^3AC) [19]–[21], MPEG surround (MPS) [22]–[25], and finally the spatial audio object coding (SAOC), whose standardization is ongoing [26], [27]. All these approaches have unique advantages and disadvantages. However, they can all be classified as open-loop systems, which means that the decoding process is not considered during encoding.

In open-loop multichannel audio coding, spatial parameters are extracted and multiple signals are typically downmixed into mono or stereo audio signals. The spatial parameters are quantized and then transmitted as side information. The major drawback of open-loop systems is the error introduced by quantizing the spatial parameters and the coding of the downmixed signals. Therefore, when using such open-loop systems, it is difficult to reach even perceptual lossless quality. In order to minimize quantization errors and improve the reconstructed audio quality, a close-loop system can be implemented.

In this paper, we provide an overview of multichannel audio coding techniques and propose an analysis-by-synthesis (AbS) framework, which takes advantage of a closed-loop system, for enhancing the quality of multichannel audio reproduction. Although we present a specific case to prove the AbS concept in spatial audio coding (SAC), we believe that this concept is applicable to other configurations to achieve the highest possible objective quality of reconstructed audio for a given bit rate since the AbS feedback mechanism will naturally attempt to minimize the error resulting from the signal processing and quantization processes. As an example, the AbS framework has been applied on the tree-structured two-to-one (TTO) encoder of the MPS standard.

The rest of this paper is organized as follows. Section II provides an overview of multichannel audio coding techniques with special emphasis on the tree structure utilized by the MPS. Section III provides a comparative analysis of some of the multichannel audio coding techniques. The principles of the AbS framework and its exemplary implementation on MPS are provided in Section IV. The experimental setup and the results are presented in Section V, followed by Section VI, which concludes the paper. A table listing all the abbreviations used in this paper can be found in Table 3.

II. OVERVIEW OF MULTICHANNEL AUDIO CODING TECHNIQUES

For digital transmission and storage, multichannel audio signals should be represented in a more compact form [28]. This can be done either by exploiting the statistical redundancy within each channel or the interchannel redundancy between the channels. Among the techniques in the first group, Dolby AC-3 [29] and MPEG advanced audio coding (AAC) [30], [31] are the most powerful ones. These apply transform coding as well as exploiting the human auditory system. Each channel of the multichannel audio content can be coded and transmitted by such an audio coder. However, this method is clearly not effective in terms of the amount of bits to be transmitted, since it increases linearly with the number of channels.

The techniques in the second group represent multiple channels in fewer channels, usually mono or stereo, which is called downmixing. Matrix surround audio codecs, such as Dolby Surround/Pro Logic [32] and Lexicon Logic 7 [33], aim to increase or decrease the number of channels by up/down conversion to meet the bandwidth requirements and the reproduction system. However, this is usually done without ensuring the preservation of spatial information essential for 3-D perception.

SAC reduces the number of channels by exploiting human spatial hearing [34]. In BCC [10]–[12], interchannel level differences (ICLDs), interchannel time differences (ICTDs), and interchannel coherence (ICC) are extracted as spatial parameters. Techniques such as PS [7]–[9], MPS [22]–[25], and SAOC [26], [27] may also utilize other parameters such as interchannel predictability [35] and utilize signal processing techniques such as decorrelators [36]. These spatial parameters are transmitted together with the downmixed signals and the residual signals. At the decoder, multichannel signals are reconstructed and reproduced, ideally creating the same 3-D perceptual effect as the original multichannel signals, which is called the perceptual transparency. The great benefit of these perceptual-based coders is that they can achieve bit rates as low as 3 kb/s for transmitting spatial parameters only, as in the case of MPS [37].

Other techniques also exist that are based on extracting the source location information from multiple channels, thereby simplifying the representation of the spatial scene. These techniques calculate direction vectors representing the directional composition of a scene. At the decoder side, virtual sources are created from the downmixed signal at positions given by the direction vectors. This approach gives the coder the capability to reproduce different number of output channels than the input channels. Examples to this technique are SASC [13]–[15] and DirAC [16]–[18]. DirAC is slightly different in that it incorporates a recording system using a microphone array [38]. Therefore, the direction vectors in this coder are calculated from the microphone signals.

Squeezing or mapping the auditory space from 360° into 60° has also been proposed for reducing the number of channels to be transmitted. Such a mapping relies on estimating virtual sources, via the inverse amplitude panning technique applied between pairs of input channels. In order to provide a stereo downmixed signal, subsequent panning is applied into estimated virtual sources. This work, known as spatially squeeze surround audio coding (S^3AC) [19]–[21], is unique such that it does not need to transmit any additional side information. The major drawback of S^3AC is that it introduces the ambiguity of sound

positions caused by representing audio channels by a set of virtual sources.

A. MPS Architecture

MPS typically works by extracting spatial parameters and downmixing multiple audio signals into either one channel (mono audio) or dual channel (stereo audio) signals [22]–[25]. As shown in Fig. 1, the downmixed signals are subsequently compressed by an existing audio coder and then transmitted accompanied by spatial parameters as side information. MPS basically extracts three spatial parameters: channel level differences (CLDs), interchannel coherences (ICCs), and channel prediction coefficients (CPCs).

Any receiver system that cannot handle multichannel audio can simply remove this side information and just render the downmixed signals. This provides the coder backward compatibility, which is important for implementation in various legacy systems. For high-quality reproduction, the residual signal can also be transmitted.

There are two pairs of elementary building blocks on MPS that are two-to-one (TTO) and two-to-three (TTT) encoder–decoder. CLDs, ICCs, and the residual signal are extracted from the TTO encoder, whereas CPCs, ICCs, and the residual signal are calculated from the TTT encoder. The whole process in the encoder and decoder is built up by combining several TTOs and TTTs in a tree structure. In this section, CLDs and ICCs as well as the residual signal will be discussed as they are implemented within the proposed framework. Also, a tree-structured TTO for converting five audio channels into a single channel is described. For details of the other parameters and the other tree structures, we refer the readers to [39].

Fig. 2 shows the schematic of the TTO encoder and decoder, denoted as E and D, respectively. The encoder converts two input channels into one downmixed output channel. CLDs and ICCs are extracted as spatial parameters in a parameter band, which can be a subband or a group of subbands. Reversely, the decoder resynthesizes two channels from one channel by utilizing the spatial parameters.

The CLD, denoted as C, is defined as the ratio between the energies of the signals in the first and second channels

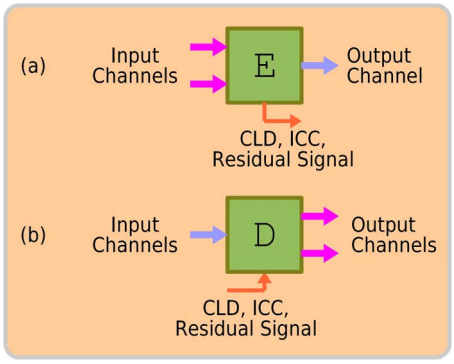

Fig. 2. *Block diagram of (a) the TTO encoder E and (b) the decoder D used in MPS. At the encoder side two channels are converted into one channel. Reversely, one channel is split into two channels at the decoder side. For downmixing more than two channels, TTO encoder and decoder is structured in a tree scheme.*

for a parameter band b

$$C^b = \frac{e^b_{x_1}}{e^b_{x_2}} \tag{1}$$

where $e^b_{x_k} = \sum_n x^b_k[n] x^{b*}_k[n]$, $k \in \{1, 2\}$. For transmission, the logarithmic values of the CLDs are preferred.

The second parameter ICC, denoted as I, is determined by

$$I^b = \mathrm{Re}\left\{ \frac{\sum_n x^b_1[n] x^{b*}_2[n]}{\sqrt{e^b_{x_1} e^b_{x_2}}} \right\}. \tag{2}$$

This parameter describes the degree of correlation between the input channels.

One set of CLD and ICC parameters is calculated for each parameter band b. However, the downmixed signal and the residual signal are calculated for each subband s.

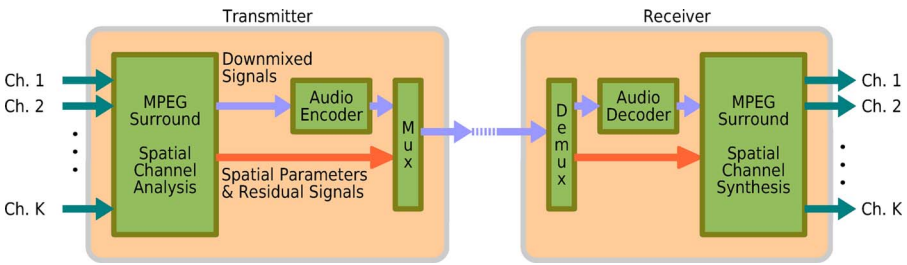

Fig. 1. *Block diagram of MPS. Multichannel audio signals are downmixed into either mono or stereo signals. The spatial parameters are extracted and the residual signals are calculated. The downmixed signals are subsequently compressed by a core audio coder.*

The downmixed signal $y[n]$ is a scaled sum of the input signals in each subband

$$y^s[n] = \frac{x_1^s[n] + x_2^s[n]}{a^s + b^s} \tag{3}$$

where the constants a and b represent the energy preservation constraint [39] calculated as

$$(a^s + b^s)^2 = \frac{e_{x_1}^s + e_{x_2}^s + 2I\sqrt{e_{x_1}^s e_{x_2}^s}}{e_{x_1}^s + e_{x_2}^s}. \tag{4}$$

The residual signal $r[n]$ in each subband is determined from the following decomposition:

$$x_1^s[n] = a^s y^s[n] + r^s[n] \tag{5a}$$
$$x_2^s[n] = b^s y^s[n] - r^s[n]. \tag{5b}$$

The downmixed signal is further coded by the mono audio coder. Spatial parameters (CLD, ICC) as well as the residual signal are quantized and then transmitted to the decoder.

At the decoder side, both audio signals are recreated by estimating a and b as follows:

$$\hat{a} = X\cos(A + B) \tag{6a}$$
$$\hat{b} = Y\cos(A - B) \tag{6b}$$

where

$$X = \sqrt{\frac{\hat{C}}{1 + \hat{C}}} \tag{7a}$$
$$Y = \sqrt{\frac{1}{1 + \hat{C}}} \tag{7b}$$
$$A = \frac{1}{2}\arccos(\hat{I}) \tag{7c}$$
$$B = \tan\left[-\left(\frac{X - Y}{X + Y}\right)\arctan(A)\right]. \tag{7d}$$

Hence, both signals can be reconstructed as

$$\hat{x}_1[n] = \hat{a}\hat{y}[n] + \hat{r}[n] \tag{8a}$$
$$\hat{x}_2[n] = \hat{b}\hat{y}[n] - \hat{r}[n]. \tag{8b}$$

The symbols $\hat{C}$, $\hat{I}$, and $\hat{r}[n]$ are the quantized values of the C, I, and $r[n]$, respectively, and $\hat{y}[n]$ is the resynthesized downmixed signal. The indexes for the subbands and parameter bands have been ignored for notation simplicity.

Fig. 3 shows the tree-structured TTO encoder for encoding K-channel audio signals. The spatial parameters are extracted in each TTO encoder $E_{p,q}$ where $p \in \{1, 2, \ldots, P\}$ and $q \in \{1, 2, \ldots, Q\}$ representing TTO indexes. For simplicity, the structure of TTO encoders and decoders is assumed to be symmetric, which means that $K = 2^Q = 2P$. Using (3), the intermediate downmixed signal $y_{p,q}[n]$ can be written as

$$y_{p,q}[n] = \frac{y_{2p-1,q+1}[n] + y_{2p,q+1}[n]}{a_{p,q} + b_{p,q}} \tag{9}$$

where $y_{2p-1,q+1}[n]$ and $y_{2p,q+1}[n]$ are the intermediate downmixed signals.

For $q = Q$, the intermediate downmixed signals are calculated from the input signals as

$$y_{p,Q}[n] = \frac{x_{2p-1}[n] + x_{2p}[n]}{a_{p,Q} + b_{p,Q}}. \tag{10}$$

At the decoder side, the audio signals can be estimated using (8) as

$$\hat{x}_{k-1}[n] = \hat{a}_{p,Q}\hat{y}_{p,Q} + \hat{r}_{p,Q} \tag{11a}$$
$$\hat{x}_k[n] = \hat{b}_{p,Q}\hat{y}_{p,Q} - \hat{r}_{p,Q} \tag{11b}$$

where $p = k/2$ and $\hat{y}_{p,Q}$ is the estimated intermediate downmixed signal. For arbitrary TTO decoder $D_{p,q}$, the estimated intermediate downmixed signal can be represented as

$$\hat{y}_{p-1,q}[n] = \hat{a}_{p/2,q-1}\hat{y}_{p/2,q-1} + \hat{r}_{p/2,q-1} \tag{12a}$$
$$\hat{y}_{p,q}[n] = \hat{b}_{p/2,q-1}\hat{y}_{p/2,q-1} - \hat{r}_{p/2,q-1} \tag{12b}$$

where p should be an even number. For $p = 2$ and $q = 2$, (12) becomes

$$\hat{y}_{1,2}[n] = \hat{a}_{1,1}\hat{y}_{1,1} + \hat{r}_{1,1} \tag{13a}$$
$$\hat{y}_{2,2}[n] = \hat{b}_{1,1}\hat{y}_{1,1} - \hat{r}_{1,1} \tag{13b}$$

where $\hat{y}_{1,1}$ is the estimated downmixed signal.

Tree-structured MPS encoding has certain limitations that become significant as the number of input channels increases. Typically, the 5.1-channel configuration is used for multichannel format even though more complex configurations, such as 7.1 and others up to 27 channels,

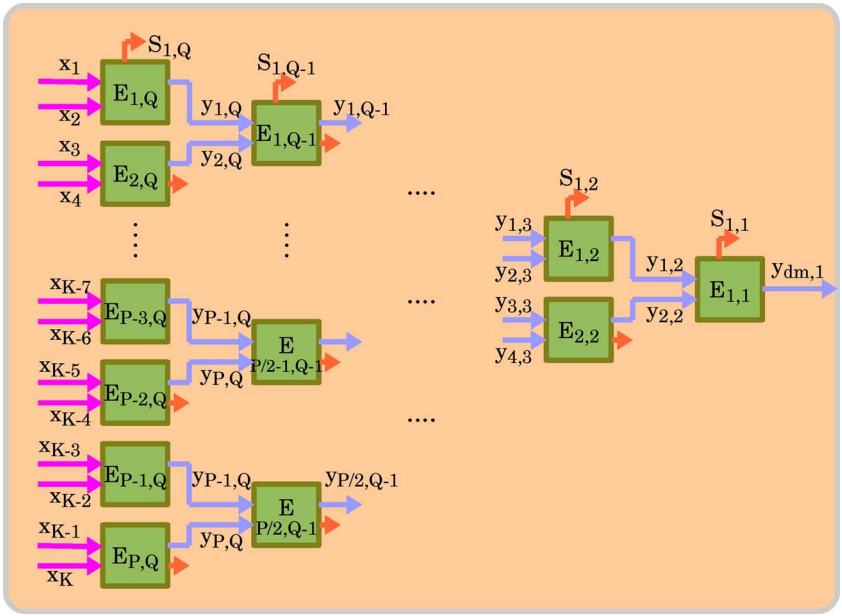

Fig. 3. *Tree-structured TTO for downmixing K-channel audio into one channel. A set of spatial parameters are extracted in each TTO encoder.*

are supported. However, using a larger number of audio channels causes higher distortion introduced by the use of the tree structure. Moreover, the pairing of input channels starts to affect the overall quality.

B. Spatial Audio Object Coding

Three-dimensional audio is essential to create the aural spatial awareness required for high-quality presence and immersiveness. Spatial fidelity becomes especially important when 3-D audio accompanies the 3-D video as discrepancies between the perceived aural and visual environments become more noticeable. There are two main challenges in achieving spatial audiovisual fidelity: perceived quality of spatial audio degrades when recording system does not match the reproduction system, and immersiveness degrades when aural cues conflict with visual cues. Whether it is a live production or a postproduction, recomposing a natural scene and rendering its spatial audio in a way agnostic of the reproduction system requires access to individual audio objects. This also makes it possible to preserve spatial synchronicity of the captured audiovisual scenes so that perceived positions of sound objects match with the visual ones.

SAOC is promising as it is designed to offer such flexibility [40]. The block diagram of the SAOC is shown in Fig. 4. Similar to MPS, the coder creates mono or stereo downmixed signals and side information according to the inter object relationships, such as the object level differences (OLDs) and interobject cross coherence (IOC) [26]. These are transmitted to the receiver where the decoder creates the estimated audio objects from the downmixed signals and the side information. The objects can then be remixed according to the user preferences on the scene composition and rendered in a format suited to the loudspeaker setup [27]. Such a system has obvious advantages in broadcasting where the content producers have no way of predicting the loudspeaker systems at the receivers. SAOC also makes it possible for users to remove vocals for karaoke, or remove background music and other unwanted sounds for improving intelligibility of speech.

For reduced computational complexity, the decoding and rendering blocks can be combined. This combined block utilizes a SAOC transcoder and a downmixed signal preprocessor. According to the user interaction and rendering system selection, new side information and downmixed signals are produced in MPS format. Therefore, the decoder of the SAOC systems includes an MPS decoder.

The block diagram in Fig. 4 assumes that the audio objects are available. At the moment, such a functionality is not offered by any system due to the difficulty of extracting audio objects from a scene and rendering a new scene in real time [41]. Extraction of audio objects for SAOC is an area of prime interest for spatial synchronization of audio and video and an approach has recently been proposed [42] based on real-time sound source separation exploiting intensity vector directions [43].

III. COMPARISON OF MULTICHANNEL AUDIO CODING TECHNIQUES

Some key parameters such as the bit rate (coding efficiency), perceptual quality, source localization, and complexity can be used in benchmarking existing multichannel audio coders. Depending on the application, one or more

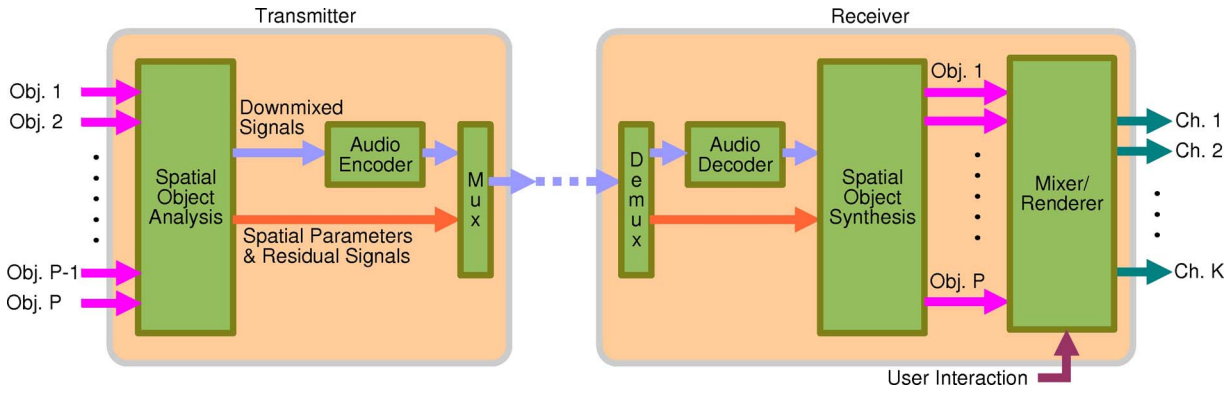

Fig. 4. ***Block diagram of the SAOC. Multiple audio objects are downmixed and transmitted. At the receiver, channel signals are calculated from the decoded audio objects based on the rendering system and the user preferences.***

key parameters may be given priority over others as there is usually a tradeoff between them. For home consumer applications, perceptual quality may be considered as the most important factor. The coding efficiency and low complexity may be the key parameters for mobile communications, while the source localization accuracy may be targeted for advanced, spatialized teleconferencing.

When the SAC technique was first introduced, the key parameter was the coding efficiency defined in terms of bit rate required for the transmission. For instance, MPS in combination with AAC stereo has been shown to perform better in terms of perceptual quality for transmitting 5.1 audio than MP3 Surround, Dolby Prologic II in combination with AAC stereo and AAC multichannel, while maintaining the transmitted bit rate at 160 kb/s [39]. However, for applications where more bandwidth is available, such as the HDTV, the quality of the reconstructed audio is very significant. The perceptual quality of AAC for transmitting 5.1 channel audio at 320 kb/s still outperforms the quality of MPS in combination with HE-AAC at 160 kb/s [24].

A comprehensive listening test has been conducted for comparing codecs for high-quality multichannel audio rendering intended for HDTV applications. Older codecs such as the digital theatre sound (DTS), which operated at 1.5 Mb/s, and Dolby Digital, which operated at 448 kb/s, performed better than other codecs including MPS in combination with HE-AAC. It was shown that at present there is no audio codec that can meet the requirement of delivering high-quality surround sound at bit rates lower than 448 kb/s [44], [45].

Some experiments have been conducted to evaluate source localization consistency. It has been reported that MP3 Surround maintains the consistency of the spatial sound image delivered, while the Dolby Prologic II results in changes to the positions of sound components in some cases [46]. Another test showed that S^3AC, which is mapping five channel audio into stereo signal without side information, performs better in terms of source localization than MPS, which is coding five channel audio into stereo downmixed signal with side information. In objective evaluation, MPS introduced more than 10° average azimuth error while S^3AC gave only 1.39°, even though subjective evaluation tests did not show much differences in source positions detected while using these coders [20].

With regards to complexity in terms of mega cycles per second (MCPS) and the memory requirements, HE-AAC v1/v2 and MPS decoders, in high-quality mode, score the highest when compared to AAC-LC, AC-3, and Enhanced AC-3 decoders [47].

IV. ANALYSIS-BY-SYNTHESIS FRAMEWORK

AbS technique is a general method implemented in the area of estimation and identification. Several decades ago, this concept was proposed as a framework for encoding of speech signals [48] and determining the excitation signal on an linear predictive coding (LPC)-based speech coder [49]. Since then, many other speech coders have been proposed within this framework, such as the code-excited linear prediction (CELP) and regular pulse excitation long-term prediction (RPE-LTP) [50], [51].

The principle of AbS is depicted in Fig. 5. A model that is able to synthesize a signal by a set of parameters is defined. These parameters are usually variable in order to produce the best matched synthesized signal. The difference between the observed signal and the synthesized signal is utilized in an error minimization block in order to find the optimal parameters [50].

A. Implementation of Tree-Structured TTO Within AbS Framework

The tree-structured TTO, as used in MPS, has been implemented within the AbS framework and named as AbS TS-TTO. Despite its limitations, MPS was chosen as an example application of the AbS framework, due to its

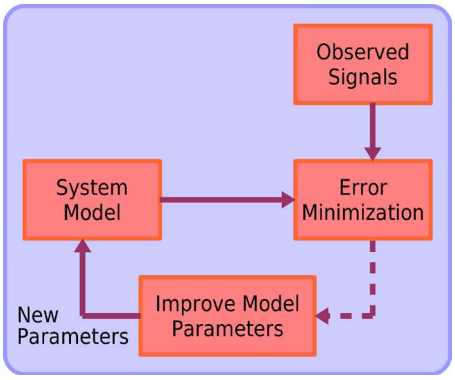

Fig. 5. *Block diagram of a typical AbS technique.*

widespread use. Moreover, AbS technique naturally reduces the error introduced by the tree structure, making it a useful testing platform.

The tree-structured TTO decoder is performed as a model for reconstructing multichannel audio signals so that (11) becomes the formula of the model. The purpose of the feedback mechanism is to minimize the errors introduced by the coding processes. The optimum signals can be determined by minimizing an error criterion, which is usually the mean squared error (mse). Fig. 6 shows an example AbS block diagram where an SAC decoder is utilized in the encoder as the system model to calculate the optimum downmixed signals.

B. Determining the Optimum Signals and Parameters

The error signals that are the differences between the original audio signals $x_k[n]$ and the synthesized signals $\hat{x}_k[n]$ can be written in vector form as

$$\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k. \tag{14}$$

By substituting (11) into (14), the errors can be represented as

$$\mathbf{e}_{k-1} = \mathbf{x}_{k-1} - \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} - \hat{\mathbf{r}}_{p,Q} \tag{15a}$$

$$\mathbf{e}_k = \mathbf{x}_k - \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} + \hat{\mathbf{r}}_{p,Q}. \tag{15b}$$

The mse for each channel can then be represented as

$$\text{mse}_{k-1} = (\mathbf{x}_{k-1} - \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} - \hat{\mathbf{r}}_{p,Q})\mathbf{e}_{k-1}^T \tag{16a}$$

$$\text{mse}_k = (\mathbf{x}_k - \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} + \hat{\mathbf{r}}_{p,Q})\mathbf{e}_k^T \tag{16b}$$

where T denotes transpose.

Assuming that the errors are not zero, the minimum mse is obtained by requiring each component of the errors to be orthogonal to the error transpose. These are described as follows.

1) First component: The input signal $\mathbf{x}_k$ has to be orthogonal to the errors transpose $\mathbf{e}^T$ (i.e., $\mathbf{x}_{k-1}\mathbf{e}_{k-1}^T = 0$ and $\mathbf{x}_k\mathbf{e}_k^T = 0$), hence

$$\mathbf{x}_{k-1}\mathbf{e}_{k-1}^T = \mathbf{x}_{k-1}\mathbf{x}_{k-1}^T - \hat{a}_{p,Q}\mathbf{x}_{k-1}\hat{\mathbf{y}}_{p,Q}^T - \mathbf{x}_{k-1}\hat{\mathbf{r}}_{p,Q}^T = 0 \tag{17a}$$

$$\mathbf{x}_k\mathbf{e}_k^T = \mathbf{x}_k\mathbf{x}_k^T - \hat{b}_{p,Q}\mathbf{x}_k\hat{\mathbf{y}}_{p,Q}^T + \mathbf{x}_k\hat{\mathbf{r}}_{p,Q}^T = 0. \tag{17b}$$

Therefore

$$\hat{\mathbf{r}}_{p,Q}^1 = \mathbf{x}_{k-1} - \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} \tag{18a}$$

$$\hat{\mathbf{r}}_{p,Q}^2 = \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} - \mathbf{x}_k \tag{18b}$$

where each residual signal is indexed assuming that they have different values.

2) Second component: The downmixed signals $\hat{\mathbf{y}}_{p,Q}$ have to be orthogonal to the errors $\mathbf{e}^T$ (i.e., $\hat{\mathbf{y}}_{p,Q}\mathbf{e}_{k-1}^T = 0$ and $\hat{\mathbf{y}}_{p,Q}\mathbf{e}_k^T = 0$). Then

$$\hat{\mathbf{y}}_{p,Q}\mathbf{e}_{k-1}^T = \hat{\mathbf{y}}_{p,Q}\mathbf{x}_{k-1}^T - \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q}\hat{\mathbf{y}}_{p,Q}^T - \hat{\mathbf{y}}_{p,Q}\hat{\mathbf{r}}_{p,Q}^T = 0 \tag{19a}$$

$$\hat{\mathbf{y}}_{p,Q}\mathbf{e}_k^T = \hat{\mathbf{y}}_{p,Q}\mathbf{x}_k^T - \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q}\hat{\mathbf{y}}_{p,Q}^T + \hat{\mathbf{y}}_{p,Q}\hat{\mathbf{r}}_{p,Q}^T = 0. \tag{19b}$$

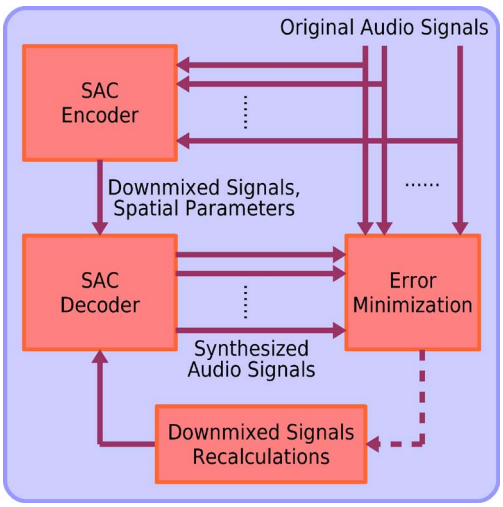

Fig. 6. *Block diagram of analysis-by-synthesis spatial audio coding (AbS SAC) when the downmixed signal is optimized.*

These equations can be simplified as

$$\mathbf{x}_{k-1} = \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} + \hat{\mathbf{r}}^1_{p,Q} \quad (20a)$$

$$\mathbf{x}_k = \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} - \hat{\mathbf{r}}^2_{p,Q} \quad (20b)$$

where they are actually the same equations as (18).

3) Third component: The residual signal $\hat{\mathbf{r}}_{p,Q}$ has to be orthogonal to the error $\mathbf{e}^T$ (i.e., $\hat{\mathbf{r}}_{p,Q}\mathbf{e}^T_{k-1} = 0$ and $\hat{\mathbf{r}}_{p,Q}\mathbf{e}^T_k = 0$), hence

$$\hat{\mathbf{r}}_{p,Q}\mathbf{e}^T_{k-1} = \hat{\mathbf{r}}_{p,Q}\mathbf{x}^T_{k-1} - \hat{a}_{p,Q}\hat{\mathbf{r}}_{p,Q}\hat{\mathbf{y}}^T_{p,Q} - \hat{\mathbf{r}}_{p,Q}\hat{\mathbf{r}}^T_{p,Q} = 0 \quad (21a)$$

$$\hat{\mathbf{r}}_{p,Q}\mathbf{e}^T_k = \hat{\mathbf{r}}_{p,Q}\mathbf{x}^T_k - \hat{b}_{p,Q}\hat{\mathbf{r}}_{p,Q}\hat{\mathbf{y}}^T_{p,Q} + \hat{\mathbf{r}}_{p,Q}\hat{\mathbf{r}}^T_{p,Q} = 0. \quad (21b)$$

Therefore

$$\hat{\mathbf{r}}^1_{p,Q} = \mathbf{x}_{k-1} - \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} \quad (22a)$$

$$\hat{\mathbf{r}}^2_{p,Q} = \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} - \mathbf{x}_k \quad (22b)$$

where the results are also exactly the same as obtained from the first and second components.

C. RSR TS-TTO: Using the Optimum Residual Signal

The residual signal recalculation (RSR) algorithm can be developed based on (18) and its block diagram is given in Fig. 7. If it is assumed that the first and second residual signals are different, two residual signals can be transmitted or a variety of approximations can be performed in order to obtain one residual signal, for example, by choosing the signal with the maximum energy. If residual signals are not modified (i.e., a single resynthesized residual signal is given), the spatial parameter optimization can be implemented. The optimum CLD and/or ICC is chosen in such a way that $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are minimum, as follows:

$$\mathbf{Q}_1 = \hat{\mathbf{r}}^1_{p,Q} - \mathbf{x}_{k-1} + \hat{a}_{p,Q}\hat{\mathbf{y}}_{p,Q} \quad (23a)$$

$$\mathbf{Q}_2 = \hat{\mathbf{r}}^2_{p,Q} - \hat{b}_{p,Q}\hat{\mathbf{y}}_{p,Q} + \mathbf{x}_k. \quad (23b)$$

D. DSR TS-TTO: Using the Optimum Downmixed Signal

From (18), the optimum downmixed signals are obtained by assuming that both residual signals are exactly the same. Therefore

$$\hat{\mathbf{y}}_{p,Q} = \frac{\mathbf{x}_{k-1} + \mathbf{x}_k}{\hat{a}_{p,Q} + \hat{b}_{p,Q}}. \quad (24)$$

By substituting (24) in (15) and simplifying the expressions, the errors can be represented as

$$\mathbf{e}_k = -\mathbf{e}_{k-1} = \frac{\hat{a}_{p,Q}(\mathbf{x}_k + \hat{\mathbf{r}}_{p,Q}) - \hat{b}_{p,Q}(\mathbf{x}_{k-1} - \hat{\mathbf{r}}_{p,Q})}{\hat{a}_{p,Q} + \hat{b}_{p,Q}} \quad (25)$$

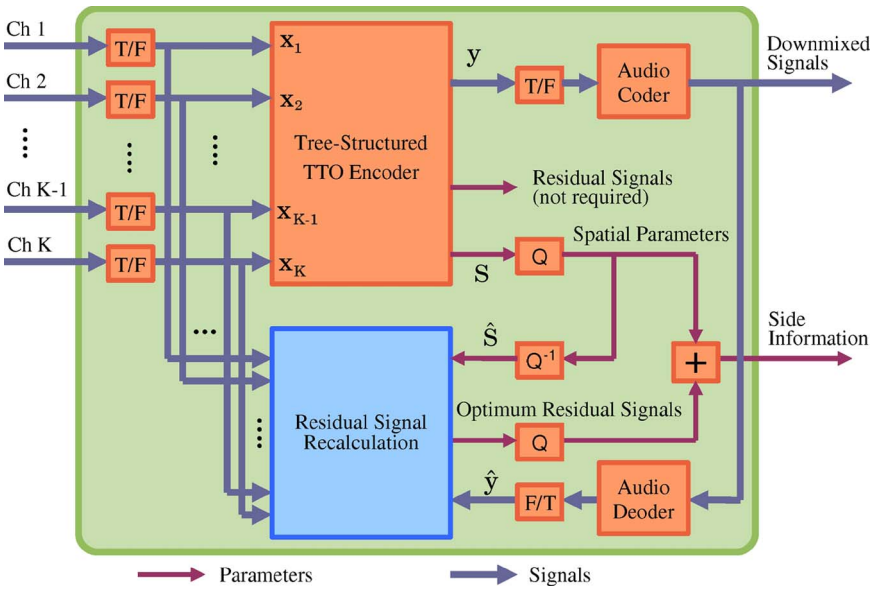

Fig. 7. *AbS tree-structured TTO is performed by feeding the resynthesized downmixed signal, the quantized spatial parameters, and the resynthesized residual signal into RSR block.*

where each channel has the same error but in different phase.

In a subband, the samples of the optimum downmixed signals can be written as

$$y_{p,Q}[n] = \frac{x_{k-1}[n] + x_k[n]}{\hat{a}_{p,Q} + \hat{b}_{p,Q}} \tag{26}$$

and generalized for any p and q as

$$y_{p,q}[n] = \frac{x_{2p-1}[n] + x_{2p}[n]}{\hat{a}_{p,q} + \hat{b}_{p,q}} \tag{27}$$

where the tree structure is assumed to be symmetric.

Using the quantized spatial parameters, downmixed signal recalculation (DSR) is performed by using (27). The new downmixed signals are then transmitted replacing the downmixed signals from the original TS-TTO encoder. For more than two channels (i.e., $K > 2$), an algorithm in DSR block performing (27) is associated for each TTO encoder in the original TS-TTO. The new intermediate downmixed signals are fed back into the original TS-TTO replacing the original intermediate downmixed signal as shown in Fig. 8, which is used by the next TTO encoder in the tree structure.

V. RESULTS

Since the proposed AbS system naturally attempts to reduce the error by trying to make the output waveform similar to the input waveform, an objective performance metric, rather than a subjective one, has been found more suitable, particularly for comparing it to the conventional open-loop system. Due to its simplicity signal-to-noise ratio (SNR) measurement has been chosen.

In terms of multichannel audio coding, we define SNR as

$$\mathrm{SNR}_k = 10 \log_{10} \left[\frac{\sum_n (x_k[n])^2}{\sum_n (x_k[n] - \hat{x}_k[n])^2} \right] \tag{28}$$

where $x_k[n]$ and $\hat{x}_k[n]$ are the original and reconstructed audio signals, respectively for the kth channel and for a frame length L, where $n = 0, \cdots, L-1$, and $k \in \{1, 2\}$ and $k \in \{1, 2, 3, 4, 5\}$ for the two-channel and five-channel encoders, respectively.

SNR is then averaged for all frames of the channel k, which gives the segmental SNR (segSNR_k).

A. The Experiment Setup

A number of low-correlated and high-correlated audio signals sampled at 44 100 Hz have been provided for the experiments. For five-channel input, low-correlated signals consisted of individual audio objects that were female and male speech, cello, trumpet and percussion music,

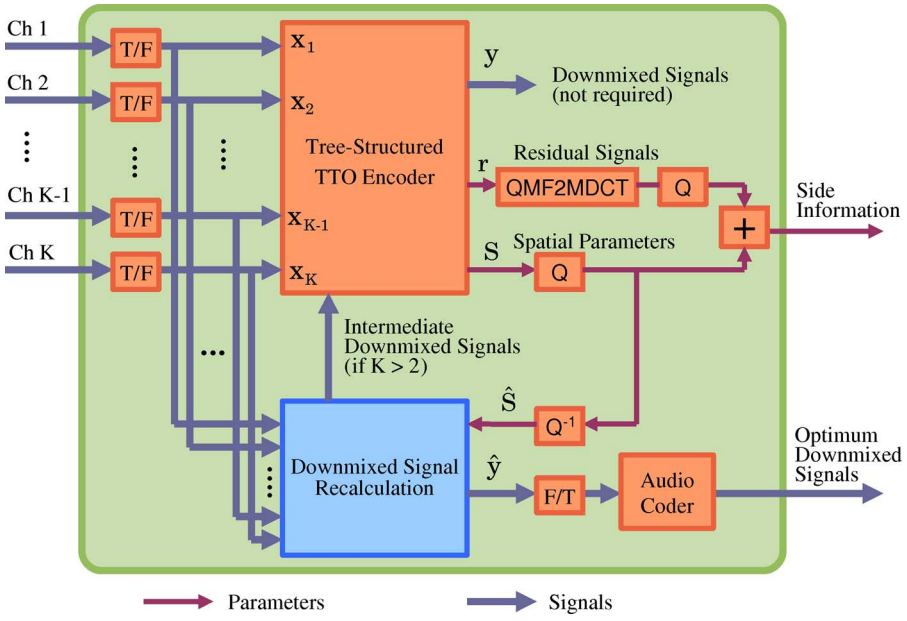

Fig. 8. *AbS tree-structured TTO is performed by utilizing downmixed signal recalculation (DSR TS-TTO). The quantized spatial parameter are used as feedback into the DSR block.*

while high-correlated signals consisted of the left (L), right (R), center (C), left surround (Ls), and right surround (Rs) channels of 5.1 recordings containing panned mixtures of the individual audio objects. For two-channel input, low-correlated signals consisted of the female speech and cello while high-correlated signals used only L and R channels.

Each of these signals were fed into a hybrid filter bank decomposing the signal into 71 subbands as in [52]. Segmentation and overlap-add windowing were performed in the subband domain. In this experiment, each audio segment consisted of 32 subband samples, which is equivalent to the effective length of 1024 time domain samples. In calculating CLDs and ICCs, 20 parameter bands were used.

For experiments using quantized spatial parameters, CLDs and ICCs were quantized as in MPS using 5 and 3 b by nonuniform quantization, respectively, as follows:

$$\begin{aligned}\mathrm{CLD} = [&150, 45, 40, 35, 30, 25, 22, 19, 16, 13, 10, 8, 6, 4,\\ &2, 0, -2, -4, -6, -8, -10, -13, -16, -19,\\ &-22, -25, -30, -35, -40, -45, -150]\\ \mathrm{ICC} = [&1, 0.937, 0.84118, 0.60092, 0.36764, 0, -0.589,\\ &-0.99].\end{aligned}$$

Furthermore, for coding the downmixed signal, AAC was used at the bit rates of 64, 80, 96, 112, 128, 144, and 160 kb/s. The residual signals were coded with the bit rates starting from 16 to 160 kb/s. To grade the performance of the closed-loop AbS TS-TTO audio coders, an open-loop audio coder (TS-TTO) has also been implemented.

In order to determine how much the closed-loop AbS system reduces the error introduced by the quantization and coding processes, a system transmitting unquantized spatial parameters (denoted as TS-TTO-ContSP) and another system transmitting unquantized downmixed signal, residual signal, and spatial parameters (denoted as unquantized TS-TTO) were included in the analysis.

The upper bound of the segSNR that can be achieved by the closed-loop AbS system can be found by transmitting both unquantized residual signals from each TTO in the subband domain (denoted as unquantized RSR TS-TTO). For this system, the error between the input and output signals in each frame occasionally reaches the zero value. For these cases, the segSNR was limited to be 80 dB.

B. Comparison of SegSNR Results for Open-Loop and Closed-Loop Systems

The experiment in this section is aimed at demonstrating that the DSR algorithm is able to improve the segSNR achieved by the TS-TTO. In this experiment, full waveform reconstruction was performed by transmitting all residual signals from each TTO encoder at 160 kb/s.

Table 1 shows the results for the two-channel coder for low-correlated and high-correlated inputs. The average segSNR measured on DSR TS-TTO is 32.30 dB, which is 4.41 dB higher than that of the conventional open-loop TS-TTO. It clearly indicates that the DSR algorithm applied on TS-TTO can minimize the error introduced by the quantization process of the spatial parameters.

Table 2 shows the results for the five-channel coder. The overall average segSNR for the DSR TS-TTO is 30.49 dB, which is 6.13 dB higher than that of the conventional open-loop TS-TTO.

In order to evaluate the effect of the AbS system when using TTO in a tree structure, the segSNR measured for the TS-TTO-ContSP can be observed. The TS-TTO cannot be used because the error contributed by the quantization of the spatial parameter and the error contributed by the tree structure of the TTO cannot be distinguished. As can be seen in Table 1, when two-channel audio coder is used, the average segSNR measured for the TS-TTO-ConstSP for low-correlated signals is 34.22 dB, which is not much different than 34.92 dB measured for high-correlated signals. However, this is not the case when five-channel audio coder is used. As shown in Table 2, the segSNR measured for five-channel TS-TTO-ContSP for low-correlated signals is 30.27 dB, which is 4.04 dB lower than that for high-correlated signals. This indicates that the tree structure of TTO used in TS-TTO also introduces error when low-correlated signals are used as input although this is not the case for high-correlated signals. The average segSNRs of DSR TS-TTO and TS-TTO-ContSP for low-correlated signals in Table 2 show that the DSR algorithm can minimize the error introduced by the spatial parameter quantizer, although it is not able to minimize the error introduced by the use of TTO in a tree structure.

Table 1 Average SegSNR (dB) of Two-Channel Audio Coders

Type of Input	Channel	TS-TTO	DSR TS-TTO	TS-TTO-ContSP
Low-correlated signals	1	24.52	30.92	31.08
	2	33.99	37.23	37.35
Average (for low-correlated signals)		29.26	34.08	34.22
High-correlated signals	L	28.83	31.55	35.66
	R	24.21	29.50	34.18
Average (for high-correlated signals)		26.52	30.52	34.92
Overall average		27.89	32.30	34.57

Table 2 Average SegSNR (dB) of Five-Channel Audio Coders

Type of Input	Channel	TS-TTO	DSR TS-TTO	TS-TTO-ContSP
Low-correlated signals	1	21.86	26.47	26.92
	2	21.72	26.21	26.56
	3	27.19	31.72	32.19
	4	28.61	32.05	32.32
	5	22.53	32.83	33.39
Average (for low-correlated signals)		24.38	29.85	30.27
High-correlated signals	L	26.38	32.06	35.31
	R	25.49	31.27	34.62
	C	20.64	28.63	33.12
	Ls	22.25	31.98	34.20
	Rs	26.92	31.70	34.32
Average (for high-correlated signals)		24.34	31.13	34.31
Overall average		24.36	30.49	32.29

Table 2 also presents that the SNR improvement achieved by the DSR TS-TTO when high-correlated signals are used is 6.79 dB, which is 2.79 dB higher than that achieved by the two-channel DSR TS-TTO for high-correlated signals as in Table 1. These results indicate that the DSR algorithm is able to maintain the segSNR, while the TS-TTO's segSNR decreases when more channels are transmitted.

For a closer look into the SNR improvement, the segSNR for several audio frames for one of the channels of the low-correlated signals has been plotted in Fig. 9. It shows the segSNRs of the conventional TS-TTO and the DSR TS-TTO systems for the same bit rate. For comparison, the SegSNR of the unquantized TS-TTO was shown for which all of the signals and parameters are unquantized. The unquantized RSR TS-TTO was also plotted as the upper bound of the segSNR, which can be achieved by the AbS system. This figure shows that significant SNR improvement can be achieved with the DSR algorithm, although there is still a high margin before reaching the upper bound of the AbS system. This is because the DSR TS-TTO is performing an AbS loop where only the spatial quantizer is included. The SNR improvement would be higher if the downmixed and residual signal coding processes were also included within the loop.

C. Performance Comparison for Various Bit Rates

In this section, performance of the five-channel DSR TS-TTO audio coder is evaluated for various bit rates. In this experiment, lower bit rates were achieved by limiting the bandwidth of the transmitted residual signal. The spatial parameter resolution was kept constant on 20 parameter bands and the time resolution was also fixed at the effective frame length of 1024 samples. At the decoder side, the decorrelator was not performed as it provides a synthetic residual signal particularly when all residual signals are not transmitted from the encoder side.

The results are given in Fig. 10. The segSNR in decibels is plotted against the bit rate in kilobit per second. The

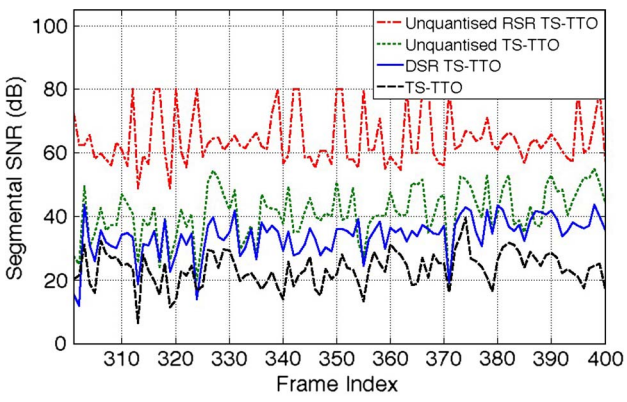

Fig. 9. *The SegSNR per frame achieved by DSR TS-TTO is compared to those measured on TS-TTO, unquantized TS-TTO and unquantized RSR TS-TTO for one of the five input signals. The SNR is plotted for frame 301 to 400.*

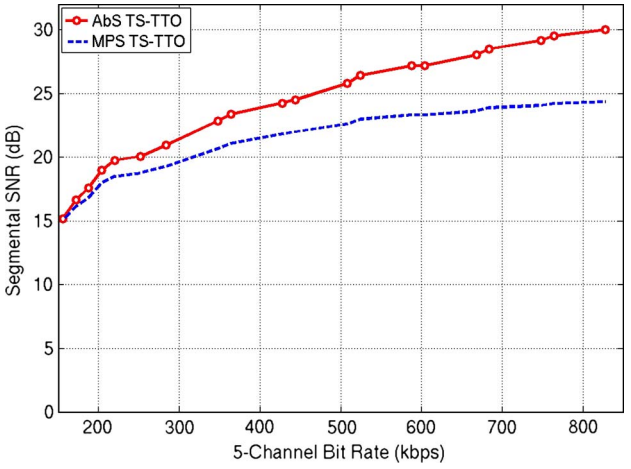

Fig. 10. *Average SegSNRs of five-channel audio coders are plotted against the bit rates. The improvement given by the closed-loop system is optimum when the residual signal is transmitted in full bandwidth.*

Table 3 List of Abbreviations

3D	three dimensions
3DTV	three dimensional television
AAC	advanced audio coding
AAC-LC	advanced audio coding low complexity
AbS	analysis by synthesis
BCC	binaural cue coding
CELP	code-excited linear prediction
CLD	channel level difference
CPC	channel prediction coefficient
DAB	digital audio broadcasting
DirAC	directional audio coding
DSR	downmixed signal recalculation
DTS	digital theatre sound
HDTV	high definition television
HE-AAC	high efficiency advanced audio coding
ICC	interchannel coherence
ICLD	interchannel level difference
ICTD	interchannel time difference
IOC	inter-object cross coherence
MCPS	mega cycles per second
MPEG	moving picture expert group
MPS	mpeg surround
mse	mean squared error
OLD	object level difference
PS	parametric stereo
RPE-LTP	regular pulse excitation long term prediction
RSR	residual signal recalculation
S3AC	spatially squeezed surround audio coding
SASC	spatial audio scene coding
SAC	spatial audio coding
SAOC	spatial audio object coding
segSNR	segmental signal-to-noise ratio
TS-TTO	tree-structured two-to-one
TTO	two-to-one
TTT	three-to-two
WFS	wave field synthesis

result of this experiment shows that the proposed framework outperforms the open-loop system for all bit rates. The highest SNR improvement is achieved when all the residual signals are transmitted. However, it decreases as the bit rate allocated for the residual signal becomes lower. From the test results one can conclude that the proposed tree-structured TTO applied within the AbS framework is suitable for all bit-rate applications.

VI. CONCLUSION

This paper has provided an overview of SAC techniques and presented an example framework for improving the objective quality of multichannel audio coding. A closed-loop system based on the AbS technique has been introduced where the reconstruction error between the input and the recovered audio channels has been minimized, which incorporates the entire signal processing and parameter quantization errors. Although we have shown this system to work on a specific case we believe that it is easily adaptable to other audio coding configurations. In our example, it has been shown that it is applicable to MPS and SAOC. Comparisons were made with the conventional open-loop MPS system using objective metrics since the AbS framework uses an objective error reduction mechanism rather than a perceptually motivated one. The results show that the proposed AbS framework significantly improves the objective quality of the reconstructed audio signals, which creates a new area for exploration in combination with other multichannel audio codecs in the future. ■

APPENDIX

See Table 3.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments in preparation of this paper.

REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. San Diego, CA: Academic, 1994.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[3] F. Rumsey, *Spatial Audio*, 2nd ed. Oxford, U.K.: Focal Press, 2001.

[4] M. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, Jan./Feb. 1973.

[5] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[6] D. R. Begault, "Challenges to the successful implementation of 3-D sound," *J. Audio Eng. Soc.*, vol. 39, no. 11, pp. 864–870, Nov. 1991.

[7] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," presented at the 114th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, Mar. 2003.

[8] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," presented at the 116th Conv. Audio Eng. Soc., Berlin, Germany, May 2004.

[9] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 1305–1322, 2005.

[10] F. Baumgarte and C. Faller, "Binaural cue coding—Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

[11] C. Faller and F. Baumgarte, "Binaural cue coding—Part II: Schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[12] C. Faller, "Parametric coding of spatial audio," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, Oct. 2004.

[13] M. M. Goodwin and J.-M. Jot, "A frequency domain framework for spatial audio coding based on universal spatial cues," presented at the 120th Conv. Audio Eng. Soc., Paris, France, May 2006.

[14] M. M. Goodwin and J.-M. Jot, "Analysis and synthesis for universal spatial audio coding," presented at the 121st Conv. Audio Eng. Soc., San Francisco, CA, Oct. 2006.

[15] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Honolulu, HI, Apr. 2007, vol. 1, pp. 9–12.

[16] J. Merimaa and V. Pulkki, "Spatial impulse response rendering," in *Proc. 7th Int. Conf. Digit. Audio Effects*, Naples, Italy, Oct. 2004, pp. 139–144.

[17] V. Pulkki and C. Faller, "Directional audio coding: Filter bank and STFT-based design," presented at the 120th Conv. Audio Eng. Soc., Paris, France, May 2006.

[18] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Proc. Audio Eng. Soc. 28th Int. Conf.*, Pitea, Sweden, Jul. 2006, pp. 251–258.

[19] B. Cheng, C. Ritz, and I. Burnett, "Squeezing the auditory space: A new approach to multichannel audio coding," in *Advances in Multimedia Information Processing 2006*. Berlin, Germany: Springer-Verlag, 2006, pp. 572–581.

[20] B. Cheng, C. Ritz, and I. Burnett, "Principles and analysis of the squeezing approach to low bit rate spatial audio coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Honolulu, HI, Apr. 2007, vol. 1, pp. 13–16.

[21] B. Cheng, C. Ritz, and I. Burnett, "A spatial squeezing approach to Ambisonic audio compression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, Apr. 2008, pp. 369–372.

[22] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," presented at the 127th Conv. Audio Eng. Soc., San Francisco, CA, Oct. 2004.

[23] S. Quackenbush and J. Herre, "MPEG surround," *IEEE Multimedia*, vol. 12, no. 4, pp. 18–23, Oct./Dec. 2005.

[24] J. Roden, J. Breebaart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Holzer, "A study of the MPEG surround quality versus bit-rate curve," presented at the 123rd Conv. Audio Eng. Soc., New York, Oct. 2007.

[25] J. Hilpert and S. Disch, "The MPEG surround audio coding standard [standards in a nutshell]," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 148–152, Jan. 2009.

[26] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," presented at the 124th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, May 2008.

[27] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in *Proc. IEEE Int. Conf. Multimedia Expo*, San Francisco, CA, Oct. 2007, pp. 1894–1897.

[28] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. New York: Springer-Verlag, 2002.

[29] C. C. Todd, G. A. Davidson, M. F. Davis, L. D. Fielder, B. D. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," presented at the 96th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, 1994.

[30] *MPEG-2 Advanced Audio Coding*, ISO/IEC JTC1/SC29/WG11 N1650, IS 13818-7.

[31] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding," presented at the 101st Conv. Audio Eng. Soc., Los Angeles, CA, Nov. 1996.

[32] R. Dressier, "Dolby pro logic surround decoder principles of operation," Dolby Laboratories Inc., S93/8624/9827, 1993.

[33] D. Griesinger, "Progress in 5-2-5 matrix systems," presented at the 103rd Conv. Audio Eng. Soc., New York, Sep. 1997.

[34] J. Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 2001.

[35] G. Hotho, L. F. Villemoes, and J. Breebaart, "A backward-compatible multichannel audio codec," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 83–93, Jan. 2008.

[36] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjorling, E. Schuijers, J. Hilpert, and F. Myburg, "The reference model architecture for MPEG spatial audio coding," presented at the 118th Conv. Audio Eng. Soc., Barcelona, Spain, May 2005.

[37] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.

[38] R. S-Amling, F. Kuech, M Kallinger, G. D. Galdo, J. Ahonen, and V. Pulkki, "Planar microphone array processing for the analysis and reproduction of spatial audio using directional audio coding," presented at the 124th Conv. Audio Eng. Soc., Amsterdam, The Netherlands, May 2008.

[39] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. Van De Par, "Background, concepts, and architecture for the recent MPEG Surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 331–351, 2007.

[40] *Call for Proposals on Spatial Audio Object Coding*, ISO/IEC JTC1/SC29/WG11 (MPEG), Marrakech, Jan. 2007, Doc. N8853.

[41] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proc. IEEE*, vol. 86, no. 5, pp. 941–951, May 1998.

[42] B. Günel, E. Ekmekçioğlu, and A. M. Kondoz, "Spatial synchronization of audiovisual objects by 3D audio object coding," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, St. Malo, France, Oct. 2010, pp. 460–465.

[43] B. Günel, H. Hacıhabiboğlu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 4, pp. 748–756, May 2008.

[44] A. Mason, D. Marston, F. Kozamernik, and G. Stoll, "EBU test of multichannel audio codecs," presented at the 122nd Conv. Audio Eng. Soc., Vienna, Austria, May 2007.

[45] D. Marston, F. Kozamernik, G. Stoll, and G. Spikofski, "Further EBU test of multichannel audio codecs," presented at the 126th Conv. Audio Eng. Soc., Munich, Germany, May 2009.

[46] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "Mp3 surround: Efficient and compatible coding of multi-channel audio," presented at the 116th Conv. Audio Eng. Soc., Berlin, Germany, May 2004.

[47] M. Nema and A. Malot, "Comparison of multichannel audio decoders for use in handheld devices," presented at the 128th Conv. Audio Eng. Soc., London, U.K., May 2010.

[48] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, no. 12, pp. 1725–1736, Dec. 1961.

[49] B. S. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Paris, France, May 1982, pp. 614–617.

[50] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, 2nd ed. New York: Wiley, 2004.

[51] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 79–120.

[52] *Information Technology—Coding of Audio-Visual Objects, Part 3: Audio*, ISO/IEC 14496-3:2009(E), International Standards Organization, Geneva, Switzerland, 2009.

ABOUT THE AUTHORS

Ikhwana Elfitri received the B.Sc. and M.Sc. degrees in electrical engineering from Bandung Institute of Technology (ITB), Bandung, Indonesia, in 1998 and 2002, respectively. Currently, he is working towards the Ph.D. degree in multichannel audio coding at the I-Lab Multimedia Communication Research, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, U.K.

In 1999, he joined the Department of Electrical Engineering, Andalas University, Padang, Indonesia, as a Lecturer teaching several courses such as communication systems, speech processing, electromagnetics, and antenna and propagation. He became the Head of Telecommunication Laboratory in 2002. Between 2004 and 2008, he was the Head of Electrical Engineering Department at Andalas University. His research interests include communication systems, low bit-rate speech coding, multichannel audio systems, and 3-D audio coding.

Banu Günel (Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from Orta Dogu Teknik Universitesi, Ankara, Turkey, in 2000, the M.Sc. degree in communication systems and signal processing from the University of Bristol, Bristol, U.K., in 2001, and the Ph.D. degree in audio and acoustical signal processing from the Queen's University of Belfast, Belfast, U.K., in 2004.

She became a Research Associate in 2004 and a Senior Research Associate in 2008 at the I-Lab Multimedia Communications Research, University of Surrey, Guildford, U.K. Currently, she is an Assistant Professor at the Informatics Institute, Orta Dogu Teknik Universitesi, Ankara, Turkey and a Visiting Research Fellow at the University of Surrey. Her research interests include microphone array signal processing, 3-D audio and psychoacoustics, multimedia services, and human–computer interaction.

Dr. Günel is a member of the European Acoustics Association, Audio Engineering Society, and London Technology Network.

Ahmet M. Kondoz (Senior Member, IEEE) received the B.Sc. (honors) degree in engineering from the University of Greenwich, Greenwich, U.K., in 1983, and the Ph.D. degree in telecommunication from the University of Surrey, Guildford, U.K., in 1986.

He became a Lecturer in 1988, a Reader in 1995, and then a Professor in Multimedia Communication Systems in 1996, at the University of Surrey. Currently, he is the founding Head of the I-Lab Multimedia Communication Research Group, whose aim is to design new fixed and mobile media communication applications by researching advanced technical solutions as well as incorporating the users throughout the research involving the building blocks of the whole delivery chain. He is also the Managing Director of MulSys Limited, a University of Surrey spin-out company marketing the world's first secure voice product over the GSM/3G voice channel. He has been involved with many national and European Union projects in the networked media area. Most recently, he coordinated the EU FP6 NoE VISNET II, which developed advanced video processing and coding solutions for video surveillance and virtual collaboration applications. Currently, he is the coordinator of FP7 STREP DIOMEDES, which deals with the distribution of multiview entertainment using content aware delivery systems. His research interests are in the areas of digital signal, image/video, speech/audio processing and coding, wireless multimedia communications, error resilient media transmission, immersive/virtual/augmented environments, and the related human factors issues including human–computer interaction/interface, and measurement/modeling of quality of experience (QoE). He has published more than 400 journal and conference papers, three books, and nine patents.