An Analysis-by-Synthesis Encoding Approach for Multiple Audio Objects

Ziyu Yang, Maoshen Jia, Changchun Bao, and Wenbei Wang
Speech and Audio Signal Processing Laboratory, College of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing 100124, China
E-mails: yangziyu@emails.bjut.edu.cn, {jiamaoshen, chchbao}@bjut.edu.cn, wwb@emails.bjut.edu.cn

***Abstract*—Object-based audio techniques are becoming popular as they provide the flexibility for personalized rendering. For encoding multiple audio objects, a recent approach based on the intra-object sparsity was proposed. However, the allocation strategy of the number of preserved time-frequency (TF) instants (NPTF) utilized in this approach usually leads to an unbalanced perceptual quality for the decoded audio objects. To overcome this issue, an analysis-by-synthesis (ABS) encoding approach for multiple audio objects is proposed in this work. By using the ABS framework, the allocated NPTF for each object is adjusted through an iterative processing, such that the maximum difference of preserved frame energy among all objects is minimized. Thereafter, multiple audio objects are encoded into a downmix signal plus side information. Both objective and subjective evaluations validated that the proposed approach is robust to different types of audio objects whilst confirming all the decoded audio objects with similar perceptual quality.**

I. Introduction

Multichannel audio is becoming common since it can render a vivid audio scene. Various multichannel audio formats, e.g. the ITU-5.1 [1], have been widely applied in cinemas as well as home theaters. However, such channel-based audio formats (e.g. the ITU-5.1) have to be rendered in a fixed manner, and cannot be adjusted for different rendering requirements in the receiver end. In contrast, the object-based audio can provide the flexibility for personalized rendering by preserving the audio scene in the form of individual audio objects (e.g. piano, horn, vocal, etc.). Such object-based audio techniques have been applied in commercial cases, e.g. in Dolby ATMOS [2].

Several techniques for encoding multiple audio objects have been proposed, such as the MPEG Spatial Audio Object Coding (SAOC) [3], the Informed Source Separation (ISS) [4], [5] approaches, and the Psychoacoustic-Based Analysis-by-Synthesis (PABS) approach [6]. These approaches are based on the inter-object relationship to compress multiple audio objects. Recently, a novel encoding approach for multiple audio objects based on *intra-object sparsity* was proposed [7]. In contrast to the aforementioned approaches exploring the inter-object relationship, the approach based on the intra-object sparsity maintains the perceptual quality of all active objects when multiple audio objects are simultaneously active in a TF bin. The evaluations validated that this approach performs better perceptual quality of decoded object signals compared to the PABS approach when eight simultaneously occurring objects are jointly encoded. However, this approach leaves the problem of allocating the number of preserved TF instants (NPTF) for each active object. In [7], the average allocation strategy was adopted, i.e., all active objects share the same NPTF. This allocation strategy cannot guarantee all decoded objects with similar perceptual quality. Especially, the uneven quality can be perceived significantly if the big difference of intra-object sparseness exists among objects (this sparseness can be measured c.f [7]).

To overcome the aforementioned issue, a novel encoding approach for multiple audio objects is proposed in this work. This approach is also based on the intra-object sparsity whilst utilizing an Analysis-by-Synthesis (ABS) framework to balance the perceptual quality for all object signals. The ABS framework for spatial audio coding has been proposed in [8] to choose the parameters such that the residual signal is minimized. Unlike [8], the ABS framework proposed in this work aims to determine the NPTF allocation strategy such that the maximum frame energy difference among all objects is minimized. Thereafter, multiple audio objects are compressed into a mono downmix signal plus side information by extracting the dominant TF instants of all object signals. The downmix signal can be further encoded using a legacy mono audio codec at various bitrate according to the channel bandwidth. Meanwhile, the side information is lossless transmitted.

The remainder of the paper is organized as follows: Section II introduces the proposed encoding approach. Experimental results are presented in Section III, while conclusions are drawn in Section IV.

II. Proposed Encoding Approach

According to the investigation in [7], it has been validated that audio object signal satisfies the intra-object sparsity (also referred to as the approximate k-sparsity) in the TF domain. That is, for an audio object, the majority of the frame energy concentrate in k TF instants where k is less than the DFT points in each frame. Therefore, the perceptual quality of the audio object can be maintained by preserving the k dominant

This work was supported in part by the National Natural Science Foundation of China under Grants 61231015 and 61201197, in part by the Specialized Research Fund for the Doctoral Program of Higher Education of the Peoples Republic of China under Grant 20121103120017, and in part by the Scientific Research Project of Beijing Educational Committee under Grant KM201310005008.

TF instants of the original object signal, which is a sparse approximation of the original one.

This work is based on the intra-object sparsity, and performed on a frame-by-frame basis. As shown in Fig. 1, multiple audio objects are transformed into TF domain using a Short Time Fourier Transform (STFT). All active objects are encoded after taking the active object signal detection. As illustrated in Fig. 2, according to the initial NPTF allocation strategy, the dominant TF instants of each active object are extracted independently. Followed by a Frame Energy Preservation Ratio (FEPR) analysis, the NPTF strategy for each active object will be updated through a ABS framework. Thereafter, the extracted TF instants of all active objects are downmixed in to a mono downmix signal plus side information. The downmix signal can be further encoded using the legacy mono audio codec while the side information is lossless transmitted. The detailed contents are described below.

A. Dominant Time-Frequency Instants Extraction

Suppose there are M audio objects are jointly encoded. The time-frequency representation of the audio object signals $s_m, m = 1, \cdots, M$, denoted by $S_m(n,l)$, can be obtained by using a STFT, where $n(1 \le n \le N)$ and $l(1 \le l \le L)$ represent frame number and frequency index, respectively, and L is the DFT points in each frame.

Subsequently, the active object detection is performed via a Voice Activity Detection (VAD) technique applied in speech signal processing [9]. For the sake of brevity, we assume that all the input M audio objects are active in the following discussion.

By sorting the TF components $S_m(n,l), l = 1, \cdots, L$ according to their module in descending order, an index $P_m(n,l)$ is found to indicate the module order of the corresponding $S_m(n,l)$ among all TF components in the frame. If we define the vectors $\boldsymbol{s}_{m,n}$ and $\boldsymbol{p}_{m,n}$ as:

$$\boldsymbol{s}_{m,n} \equiv [S_m(n,1), \cdots, S_m(n,L)], \tag{1}$$

$$\boldsymbol{p}_{m,n} \equiv [P_m(n,1), \cdots, P_m(n,L)], \tag{2}$$

the elements of these two vectors satisfy

$$|\boldsymbol{s}_{m,n}(P_m(n,u))| \ge |\boldsymbol{s}_{m,n}(P_m(n,v))|, \forall u < v, \tag{3}$$

For the given NPTF denoted by $k_{m,n}$, a TF mask of $\boldsymbol{s}_{m,n}$ can be determined via

$$\boldsymbol{i}_{m,n} \equiv [I_m(n,1), \cdots, I_m(n,L)], \tag{4}$$

where

$$I_m(n,l) = \begin{cases} 1, & \text{if } l = \arg\left(P_m(n,l) \le k_{m,n}\right), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Thereafter, a sparse approximation signal $\boldsymbol{s}'_{m,n}$ of $\boldsymbol{s}_{m,n}$ containing the dominant TF instants is attained by

$$\boldsymbol{s}'_{m,n} = \boldsymbol{i}_{m,n} \odot \boldsymbol{s}_{m,n}, \tag{6}$$

where $\odot$ denotes element-by-element multiplication.

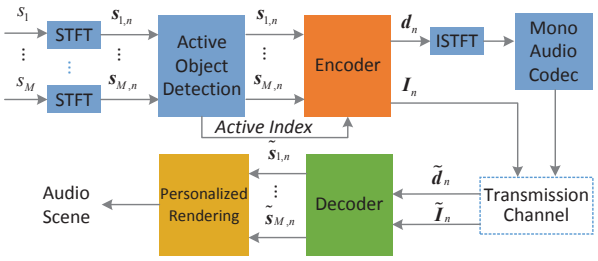

Fig. 1. Diagram for the proposed encoding approach.

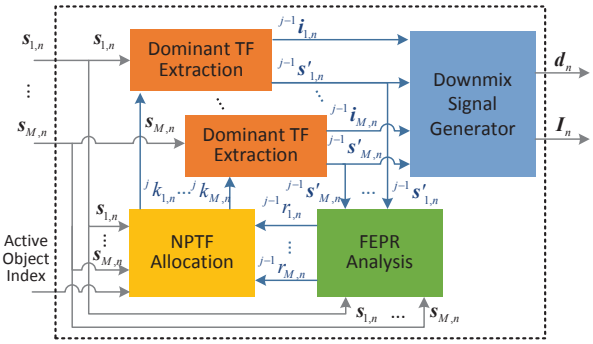

Fig. 2. Diagram for the encoder.

B. Analysis-by-Synthesis Framework

In the beginning of each frame, allocating the initial NPTF ${}^0k_{m,n}$ for each active object m can be taken in various manner provided

$$\sum_{m=1}^{M} \left({}^0k_{m,n}\right) \le L, \tag{7}$$

where L is the DFT points of each frame. This constraint guarantees a mono downmix signal can be generated whilst preserving the dominant TF instants of all active objects. In this work, the average allocation strategy is employed as the initial NPTF allocation scheme, i.e.,

$${}^0k_{m,n} = \lfloor L/M \rfloor, m = 1, \cdots, M, \tag{8}$$

where $\lfloor \cdot \rfloor$ represents the floor function.

The key problem is to find a optimal allocation strategy with respect to the NPTF $k_{m,n}, m = 1, \cdots, M$, such that all active object signals share approximately the same perceptual quality. To formulate the problem, we firstly employ the FEPR, denoted by $r_{m,n}$, as the measure for the quality of each object $\boldsymbol{s}_{m,n}, m = 1, \cdots, M$, which can be calculated through

$$r_{m,n} = \frac{\|\boldsymbol{s}'_{m,n}\|_1}{\|\boldsymbol{s}_{m,n}\|_1}, \tag{9}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm.

Subsequently, for arbitrary object pair $(m_u, m_v) \in \{1, \cdots, M\}$, the FEPR difference $\Delta r_n(m_u, m_v)$ is defined as

$$\Delta r_n(m_u, m_v) \equiv |r_{m_u,n} - r_{m_v,n}|. \tag{10}$$

Therefore, the aforementioned problem is converted into the following minimax problem:

$$\text{minimize} \quad \max_{m_u, m_v \in \{1,\cdots,M\}} \{\Delta r_n(m_u, m_v)\}. \tag{11}$$

According to the theory of minimax methods [10], the optimal NPTF allocation strategy $k_{m,n}, m = 1, \cdots, M$ calculated by solving (11) can yield the approximate evenly distributed FEPR for all objects.

The problem can be solved via the iterative processing. For the j-th iteration, we denote the object signal with the highest and lowest FEPR by ${}^{j}\boldsymbol{s}'_{m_h,n}$ and ${}^{j}\boldsymbol{s}'_{m_l,n}$, respectively. Thus, the maximum FEPR difference among all active objects is defined by

$$\begin{aligned} {}^{j}\Delta r_n^{max} &\equiv {}^{j}r_{m_h,n} - {}^{j}r_{m_l,n} \\ &= \max_{1\le m\le M} \left\{{}^{j}r_{m,n}\right\} - \min_{1\le m\le M} \left\{{}^{j}r_{m,n}\right\}. \end{aligned} \tag{12}$$

Here, the expression $({}^{j}\Delta r_n^{max} < {}^{j-1}\Delta r_n^{max})$ is employed as the stop criterion for the iteration. Specifically, if the expression is true, the NPTF of all active objects are updated for the $(j+1)$-th iteration as follow:

$${}^{j+1}k_{m_h,n} = {}^{j}k_{m_h,n} - 1, \tag{13}$$

$${}^{j+1}k_{m_l,n} = {}^{j}k_{m_l,n} + 1, \tag{14}$$

$${}^{j+1}k_{m,n} = {}^{j}k_{m,n}, \quad m \neq m_h, m_l. \tag{15}$$

Otherwise, if $({}^{j}\Delta r_n > {}^{j-1}\Delta r_n)$ is true, the NPTF allocation strategy calculated from the previous (i.e., the $(j-1)$-th) iteration should be adopted as the final strategy. Meanwhile, all the sparse approximation signals ${}^{j-1}\boldsymbol{s}'_{m,n}, m = 1, \cdots, M$ along with their mask vectors ${}^{j-1}\boldsymbol{i}_{m,n}, m = 1, \cdots, M$ from the previous iteration should be sent to the downmix module to generate the downmix signal.

C. Downmix Processing

After taking the ABS framework, the final sparse approximation signals $\boldsymbol{s}'_{m,n}, m = 1, \cdots, M$ containing the dominant TF instants of all active objects are grouped together to form a matrix $\boldsymbol{D}_n \equiv [\boldsymbol{s}'_{1,n}, \cdots, \boldsymbol{s}'_{M,n}]^T$. Note that $\boldsymbol{D}_n$ is a sparse matrix containing at most L nonzero entries with the total number of $M \times L$ entries. Hence, the TF representation of the downmix signal can be generated through redistributing the nonzero entries of $\boldsymbol{D}_n$ (i.e., the extracted TF instants) from 1 to L in the frequency axis. Here, we denote the downmix signal by

$$\boldsymbol{d}_n \equiv [d(n,1), \cdots, d(n,L)], \tag{16}$$

which can be generated through a column-wise scanning of the nonzero entries of $\boldsymbol{D}_n$, cf. Section III-C in [7]. Then, the downmix signal is transformed into time-domain using a Inverse Short Time Fourier Transform (ISTFT) and further encoded via a legacy audio codec as illustrated in Fig. 1.

Meanwhile, the mask vectors are grouped together to form a matrix $\boldsymbol{I}_n \equiv [\boldsymbol{i}_{1,n}, \cdots, \boldsymbol{i}_{M,n}]^T$, which is also a sparse matrix with the same size of $\boldsymbol{D}_n$. Note that $\boldsymbol{I}_n$ indicate the origin and the position of the extracted TF instants, and can be further lossless compressed as side information signal.

D. Decoding and Rendering

In the receiver end, the extracted TF instants of all objects can be decoded from the downmix signal by analyzing the side information. Thereafter, the extracted TF instants are assigned to recover the object signals. All audio object signals are obtained by transforming back to the time domain using the ISTFT, and can be further rendered according to the requirement of practical application.

III. Evaluations

To examine the performance of the proposed encoding approach, both objective and subjective evaluations are taken. All the test data are selected from the QUASI audio database [11], which provides various types of audio object signal (e.g. piano, vocal, drums, etc.) sampled at 44.1 kHz. In this work, 6 multi-track audio files are produced where each file is served as a group of 8 simultaneously occurring audio objects. Both the instruments and playing notes vary for each track. The duration of each file is 15 seconds. A 2048-points STFT with 50% overlapping is applied in this work, where the number of the DFT points in each frame is also 2048.

A. Objective Evaluations

In the objective evaluations, both the downmix signal and the side information are encoded using lossless techniques. For each multi-track audio file, the average FEPR $\overline{r}$ is employed as the measurement to compare the quality of decoded audio objects by using the reference approach [7] (condition 'SPA') and our proposed approach (condition 'SPA+ABS'), respectively. Specifically, $\overline{r}$ can be calculated through:

$$\overline{r} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} r_{m,n}, \tag{17}$$

where $r_{m,n}$ is defined by (9), and M and N represent the number of active objects and frames, respectively.

Results are presented in Fig.3 with 95% confidence intervals. It can be observed that the proposed approach achieves the similar average FEPR compared to the reference approach while presenting a smaller variation, which can be seen by the smaller 95% confidence intervals. Furthermore, in order to observe the quality differences of decoded objects significantly, the average maximum FEPR difference is employed as the measurement, which can be calculated by:

$$\Delta\overline{r} = \frac{1}{N} \sum_{n=1}^{N} \Delta r_n^{max}, \tag{18}$$

where Δr_n^{max} is defined by (12).

Results are presented in Fig.4 with 95% confidence intervals. It can be observed that the proposed approach leads to a much smaller differences among object signals, which indicates a more balanced quality of decoded objects compared to the reference approach. In addition, this test validates that the proposed approach is robust to different types of audio objects.

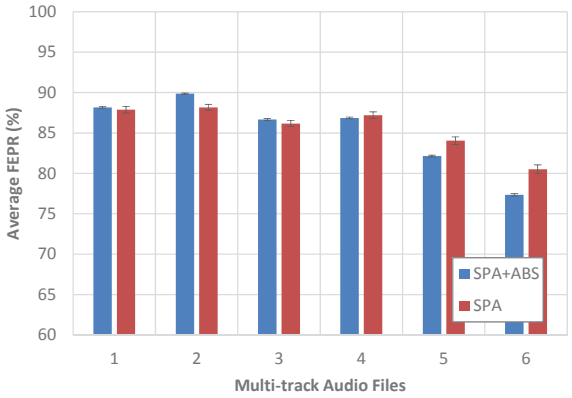

Fig. 3. Average FEPR results for the 6 groups of multiple audio objects.

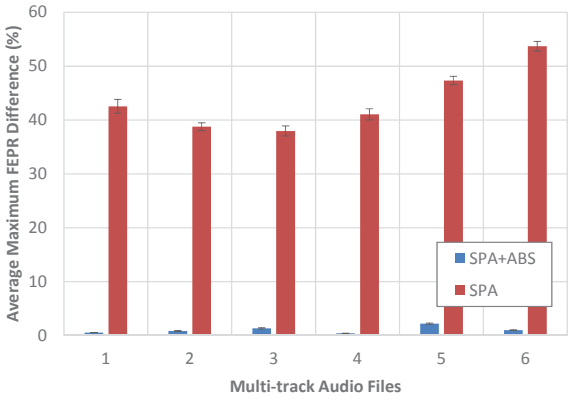

Fig. 4. Average maximum FEPR difference results for the 6 groups of multiple audio objects.

B. Subjective Evaluations

The MUSHRA [12] listening test is further employed to measure the perceptual quality of decoded audio objects from the objective evaluation. In addition to the conditions 'SPA' and 'SPA+ABS', the original object signals served as the Hidden Reference (condition 'Ref') and the 3.5 kHz low pass filtered anchor signals are included in this test. Noted that each decoded audio object is independently evaluated using headphones for playback. Ten listeners participated in the test.

Results are presented in Fig. with 95% confidence intervals. It can be observed that both the reference approach and the proposed approach achieve around 80 marks indicating 'Good' perceptual quality compared to the 'Hidden Reference'. Moreover, the proposed approach has the smaller 95% confidence intervals indicating the more balanced perceptual quality compared to the reference approach.

IV. Conclusions

This work presented a novel encoding approach for multiple audio objects based on the intra-object sparsity. To achieve a balanced perceptual quality for all object signals, an analysis-by-synthesis framework is employed for obtaining the optimal

Fig. 5. MUSHRA test results for the 6 groups of multiple audio objects.

allocation strategy of the NPTF. This optimal allocation strategy can be attained by the iterative method. Both objective and subjective evaluations showed that the proposed approach performed good perceptual quality whilst achieving the more balanced perceptual quality compared to the existing approach. Further research could include adding the perceptual model to the encoding stage such that more perceptually important TF bins can be preserved.

References

[1] BS.775 Int. Telecommunication Union, "Multichannel stereophonic sound system with and without accompanying picture," 2006.

[2] Dolby Laboratories, "Dolby ATMOS cinema specifications," 2014. [Online]. Available: http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-specifications.pdf

[3] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG Spatial Audio Object Coding - the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc*, vol. 60, no. 9, pp. 655–673, 2012.

[4] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed audio source separation: A comparative study," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO '12)*, Aug. 2012, pp. 2397–2401.

[5] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1699–1712, Aug. 2013.

[6] X. Zheng, C. Ritz, and J. Xi, "A psychoacoustic-based analysis-by-synthesis scheme for jointly encoding multiple audio objects into independent mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '13)*, May 2013, pp. 281–285.

[7] M. Jia, Z. Yang, C. Bao, X. Zheng, and C. Ritz, "Encoding multiple audio objects using intra-object sparsity," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1082–1095, June 2015.

[8] I. Elfitri, B. Günel, and A. Kondoz, "Multichannel audio coding based on analysis by synthesis," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 657–670, Apr. 2011.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, Jan. 1999.

[10] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer Science & Business Media, Mar. 2007.

[11] "QUASI database - a musical audio signal database for source separation." [Online]. Available: http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

[12] BS.1534 Int. Telecommunication Union, "Method for the subjective assessment of intermediate quality levels of coding systems," 1997.