DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **15**, No 6
Special Issue on Logistics, Informatics and Service Science

Sofia • 2015

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2015-0074

# Multi-Stage Encoding Scheme for Multiple Audio Objects Using Compressed Sensing

*Ziyu Yang, Maoshen Jia, Wenbei Wang, Jiaming Zhang*

*School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, 100124 China*
*Emails: yangziyu@emails.bjut.edu.cn jiamaoshen@bjut.edu.cn wwb@emails.bjut.edu.cn zhangjiaming@emails.bjut.edu.cn*

***Abstract:*** *Object-based audio techniques have become common since they provide the flexibility for personalized rendering. In this paper a multi-stage encoding scheme for multiple audio objects is proposed. The scheme is based on intra-object sparsity. In the encoding phase the dominant Time Frequency* (*TF*) *instants of all active object signals are extracted and divided into several stages to form the multi-stage observation signals for transmission. In the decoding phase the preserved TF instants are recovered via Compressed Sensing* (*CS*) *technique, and further used for reconstructing the audio objects. The evaluations validated that the proposed encoding scheme can achieve scalable transmission while maintaining perceptual quality of each audio object.*

***Keywords:*** *Audio object coding, sparsity, compressed sensing, multi-stage encoding.*

## 1. Introduction

With the development of multimedia technology, multi-channel audio is becoming more widespread. From ITU-5.1 [1] to NHK 22.2 [2], various audio formats provide more and more vivid listening experience, as the number of channels increases. However, such channel-based audio formats, e.g. ITU-5.1, require a fixed manner for rendering, and hence, are less flexible in adjusting to practical applications. Moreover, with the advent of 3DTV [3] and free viewpoint TV [4], the interactive and personalized playback for audio sources is increasingly expected.

But the channel-based audio formats cannot provide such personalized choices with respect to the audio scene.

To provide such flexibility, one solution is to preserve the audio scene in the form of multiple audio objects, e.g., piano, violin, vocal, drums, etc., such that each audio object can be rendered independently according to the requirements. The object-based audio techniques have been applied in the commercial case, e.g., Dolby, ATMOS [5].

For encoding and transmitting multiple audio objects several approaches have been proposed, such as the MPEG Spatial Audio Object Coding (SAOC) [6], the Informed Source Separation (ISS) approaches [7, 8], and the Psychoacoustic-Based Analysis-by-Synthesis (PABS) approach [9]. These approaches explore the inter-object relationship to present the multiple audio objects as a mono/stereo down-mix signal plus side information. Recently, a new encoding approach [10] based on *intra-object sparsity* was proposed. Compared to the aforementioned techniques, this approach confirms that the dominant TF instants of all active object signals are preserved, and hence it maintains the good perceptual quality of all decoded object signals. The evaluation results validated that this approach has achieved better performance than the reference approach [9] when the multiple objects are simultaneously active in a TF bin.

In this work, a multi-stage encoding scheme for multiple audio objects is proposed. This approach is based on intra-object sparsity. In the encoding phase, the dominant TF instants from all active object signals are preserved and divided into several stages according to their energy. Specifically, the 1st stage contains the most significant time-frequency instants extracted from all object signals. The secondary significant TF instants are contained within the 2nd stage, etc. Thereafter, the preserved TF instants along with their original information are multiplied by a sensing matrix to form the observation signals, which can be further encoded via the Scalar Quantized Vector Huffman Coding (SQVH) [11] for transmission. In the decoding phase, the preserved TF instants can be attained via the Compressed Sensing (CS) techniques [12, 13], i.e., solving the $l_1$-norm minimization problems with respect to the received observation signals. Finally, the object signals can be reconstructed by exploiting the TF instants and further rendered according to the requirements.

The key contributions of the proposed scheme include the following two aspects. On the one hand, the proposed scheme generalizes the framework proposed in [10] by means of introducing the multi-stage encoding framework. Compared to the single-stage technique, such a framework is more suitable for scalable transmission. Especially in the bandwidth constrained case, the multi-stage approach enables bit rate adaption according to the channel condition. On the other hand, unlike the existing techniques which encode the objects into a down-mix signal plus side information for transmission [9, 10], no side information is needed to be transmitted in this work. Moreover, compared to the existing approaches [9, 10], in which the side information must be transmitted in a lossless manner, the observation signals generated by the proposed approach can be transmitted in a loss manner, and hence are more robust for transmission.

The remainder of the paper is organized as follows: Section 2 provides the overview of intra-object sparsity. Section 3 presents the proposed encoding scheme. Evaluation results are presented in Section 4, while conclusions are drawn in Section 5.

## 2. Overview of intra-object sparsity

Intra-object sparsity for an audio object signal was investigated in [10]. Generally, this intra-object sparsity can be explained as the energy of an object signal when concentrated in a small number of TF instants. More specifically, this sparsity is also referred to as the *approximate k-sparsity*, i.e., $k$ TF instants occupying the majority of the energy of the object signal.

To measure and examine the sparsity assumption for audio objects, a measure named Number of Preserved TF instants (NPTF) was proposed in the recent paper [10]. Explicitly, in a frame, $\boldsymbol{x}$ denotes the TF representation of an audio object signal $\boldsymbol{s}$ under a basis function.

Suppose that $\boldsymbol{x}$ is an $L$-dimensional vector. The *sparse approximation signal of* $\boldsymbol{x}$, denoted by $\boldsymbol{\theta}$, can be attained by preserving the portion of the TF instants of $\boldsymbol{x}$ while setting the other TF instants to zero.

Thus, for a given $\boldsymbol{\theta}$, the Frame Energy Preservation Ratio (FEPR) $r$ can be calculated through

$$r = \frac{\| \boldsymbol{\theta} \|_1}{\| \boldsymbol{x} \|_1}, \tag{1}$$

where $\|\cdot\|_p$ denotes the $l_p$-norm. Therefore, the NPTF, denoted by $k$, is defined as a function of FEPR:

$$k(r) = \inf \left\{ \|\boldsymbol{\theta}_i\|_0 \left| \frac{\| \boldsymbol{\theta}_i \|_1}{\| \boldsymbol{x} \|_1} \geq r, i = 1, 2, \cdots \right. \right\}, \tag{2}$$

where $\inf\{\cdot\}$ represents the infimum. This measure describes the least achievable preserved TF instants for an arbitrary FEPR.

The statistical results, presented in [10], have shown that the function $k(r)$ is a convex function with respect to $r$, which means that the audio object signals satisfy the approximate $k$-sparsity. Furthermore, another statistical analysis in [10] validated that preserving the portion of the TF instants can maintain the perceptual quality of the audio object signal.

## 3. Proposed multi-stage encoding scheme

The proposed encoding approach is based on the intra-object sparsity, discussed in Section 2, and performed on a frame-by-frame basis. As shown in Fig. 1, multiple audio objects are converted into TF domain. After taking the active object detection, the dominant TF instants of all active object signals are extracted independently and then jointly represented as the 1st stage observation signal, which is illustrated in Fig. 2.

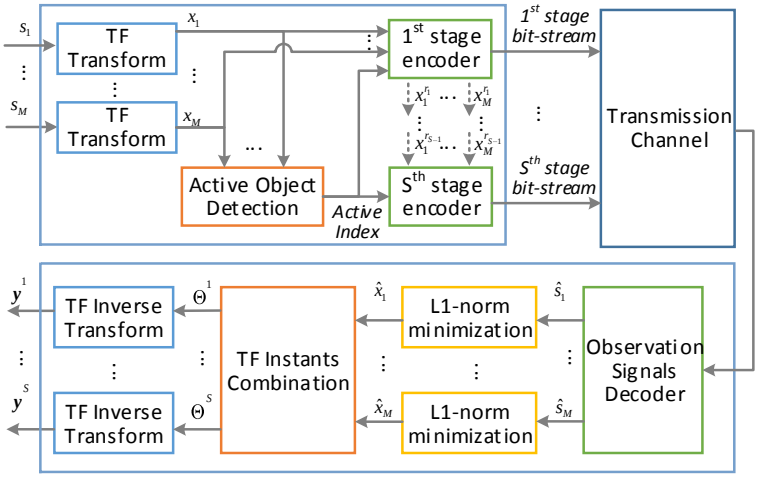


Fig. 1. Diagram for the proposed multi-stage encoding scheme

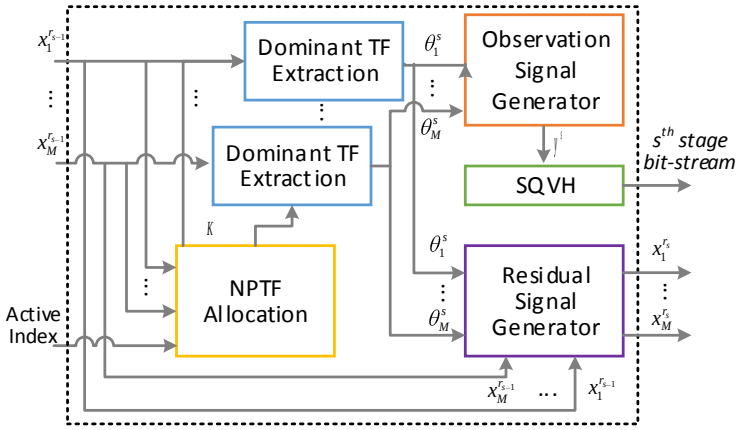


Fig. 2. Diagram for the *s*-th stage encoder

The same procedures are taken again for the residual signal generated from the last stage to form the 2nd observation signal, etc. The multi-stage observation signals can be further encoded using SQVH to form the multi-stage bit streams for transmission throughout the network.

In the decoding phase, the observation signals are initially decoded. Then, the preserved TF instants can be obtained via CS techniques and further used for recovering the audio objects. The detailed contents are described below.

### 3.1. Multi-stage dominant TF instants extraction

There are $M$ input audio objects for encoding. In the frame $n$, the $m$-th, $m \in \{1,\cdots,M\}$, audio object signal $\boldsymbol{s}_{m,n}$ has TF representation $\boldsymbol{x}_{m,n}$ via a proper basis matrix $\boldsymbol{\Psi}$:

(3) $$\boldsymbol{s}_{m,n} = \boldsymbol{\Psi}\boldsymbol{x}_{m,n},$$

where $\Psi$ is a fixed $L \times L$ orthonormal basis matrix. For the sake of brevity, the dependency of all quantities on $n$ is omitted in the following discussion.

Then a Voice Activity Detection (VAD) technique [14] is applied to detect the active audio objects. For brevity, all the $M$ audio objects are assumed to be active in the following discussion.

Therefore, $\boldsymbol{x}_m$ can be considered as a column vector that consists of $L$ TF instants, denoted by

$$\boldsymbol{x}_m \equiv [X_m(1), \cdots, X_m(L)]^{\mathrm{T}}. \tag{4}$$

For each element $X_m(l), l = 1, \cdots, L,$ a positive integer $P_m(l)$ can be determined to indicate the order according to the magnitude. For example, $P_m(l_0) = 1$ indicates that $|X_m(l_0)|$ is the largest among all $|X_m(l)|, l = 1, \cdots, L$.

Allocating the NPTF for each active object signal can be accomplished in various manners according to the applications. In this work all object signals share the same NPTF, denoted by $K$. Given $K$, a TF mask corresponding to $\boldsymbol{x}_m$ can be calculated by

$$\boldsymbol{i}_m \equiv [I_m(1), \cdots, I_m(L)]^{\mathrm{T}}, \tag{5}$$

where

$$I_m(l) = \begin{cases} 1 & \text{if } l = \arg\ P_m(l) \le K \\ 0 & \text{otherwise.} \end{cases}, \tag{6}$$

Thus, a sparse approximation signal $\boldsymbol{\theta}_m^1$ of $\boldsymbol{x}_m$ is attained by

$$\boldsymbol{\theta}_m^1 = \boldsymbol{i}_m \otimes \boldsymbol{x}_m, \tag{7}$$

where $\otimes$ denotes element-by-element multiplication. Note that $\boldsymbol{\theta}_m^1$ is a strict $K$-sparse signal, containing the $K$ dominant TF instants of $\boldsymbol{x}_m$.

Thereafter, the residual signal of $\boldsymbol{x}_m$, denoted by $\boldsymbol{x}_m^{r_1}$, can be calculated by

$$\boldsymbol{x}_m^{r_1} = \boldsymbol{x}_m - \boldsymbol{\theta}_m^1. \tag{8}$$

By applying the above-mentioned procedures again for $\boldsymbol{x}_m$, $\boldsymbol{\theta}^2$ can be obtained. Generally, for the $s$-th stage, we have

$$\boldsymbol{x}_m^{r_s} = \boldsymbol{x}_m^{r_{s-1}} - \boldsymbol{\theta}_m^{r-1}, \quad s = 2, 3, \ldots \tag{9}$$

In each stage $s$, the sparse approximation signals $\boldsymbol{\theta}_m^s$ of all active objects $m = 1, \cdots, M$ are used for generating of the observation signal.

### 3.2. Observation signals generation and quantization

In the $s$-th $(s \in \{1, \cdots, S\})$ stage, the $M$ vectors $\boldsymbol{\theta}_m^s,\ m = 1, \cdots, M,$ are grouped together to form a matrix $\boldsymbol{\Theta}^s$:

$$\boldsymbol{\Theta}^s \equiv [\boldsymbol{\theta}_1^s, \cdots, \boldsymbol{\theta}_M^s]. \tag{10}$$

Thus, $\boldsymbol{\Theta}^s$ is a sparse matrix which contains $MK$ nonzero entries, where each nonzero entry of $\boldsymbol{\Theta}^s$ corresponds to a specific preserved TF instant. It should be noted that the origin information of each preserved TF instant is indicated by the position in $\boldsymbol{\Theta}^s$. Explicitly, for a nonzero entry, the column index indicates the

object that is extracted from, while the row index is representing the frequency index of the original signal.

Here, our target is to encode and transmit these preserved TF instants along with their original information, i.e., the matrix $\boldsymbol{\Theta}^s$. Unlike the existing 'downmix plus side information' framework (e.g., [9, 10]), this work employs the CS techniques [12, 13] to represent jointly the preserved TF instants and their original information as an *observation signal*.

Specifically, the observation matrix $\boldsymbol{Y}^s$ corresponding to $\boldsymbol{\Theta}^s$ is attained through:

$$\boldsymbol{Y}^s = \hat{\boldsymbol{\Phi}}\boldsymbol{\Theta}^s, \tag{11}$$

where $\hat{\boldsymbol{\Phi}}$ represents the sensing matrix with size $D \times L$. The problem in choosing the type of the matrix $\hat{\boldsymbol{\Phi}}$ and determining the number of sensing measurements $D$ will be discussed in the next subsection.

Thereafter, the observation matrix $\boldsymbol{Y}^s$ is transformed into a vector to generate the observation signal (vector) $\boldsymbol{y}^s$ via a row-wise scanning. Thus, $\boldsymbol{y}^s$ is the column vector containing $R$ observation coefficients (where $R \equiv MD$), denoted by

$$\boldsymbol{y}^s = [Y^s(1), \cdots, Y^s(R)]^{\mathrm{T}}. \tag{12}$$

The observation signal $\boldsymbol{y}^s$ can be further encoded by the SQVH [11] described as follows.

The vector $\boldsymbol{y}^s$ is decomposed into two parts, i.e., the sign $\boldsymbol{y}^s_{\mathrm{sign}}$ and the magnitude $\boldsymbol{y}^s_{\mathrm{mag}}$, defined by:

$$\boldsymbol{y}^s_{\mathrm{sign}} = [\mathrm{sign}(Y^s(1)), \cdots, \mathrm{sign}(Y^s(R))]^{\mathrm{T}}, \tag{13}$$

$$\boldsymbol{y}^s_{\mathrm{mag}} = [|Y^s(1)|, \cdots, |Y^s(R)|]^{\mathrm{T}}. \tag{14}$$

These two parts are processed separately.

To quantize the magnitude, the vector $\boldsymbol{y}^s_{\mathrm{mag}}$ is divided into $W$ subvectors, where each subvector contains $B$ coefficients, i.e.,

$$\boldsymbol{y}^s_{\mathrm{mag}} = [\boldsymbol{y}^s_{(1)}, \cdots, \boldsymbol{y}^s_{(W)}]^{\mathrm{T}}. \tag{15}$$

For the subvector $\boldsymbol{y}^s_{(w)}, w = 1, \cdots, W, \quad w \in \{1, \cdots, W\}$, the mean-root-square is calculated through

$$Q_{\mathrm{rms}}(w) = \frac{\left\| \boldsymbol{y}^s_{(w)} \right\|_2}{\sqrt{B}}. \tag{16}$$

Therefore, for each element $|Y^s(r)|, r \in \{1, \cdots, R\}$, a quantization index can be obtained by

$$Q^s_{\mathrm{ind}}(r) = \min\left\{ \left\lfloor \frac{|Y^s(r)|}{Q_{\mathrm{rms}}(w) Q_{\mathrm{step}}} + Q_{\mathrm{offset}} \right\rfloor, Q_{\max} \right\}, \tag{17}$$

where: $w = \lceil r / B \rceil$; $Q_{\mathrm{step}}$ and $Q_{\mathrm{offset}}$ represent the quantization step size and the offset, respectively; $Q_{\max}$ represents the upper bound of the quantization index.

We group the $R$ quantization indices together to form a vector $\boldsymbol{q}^s_{\mathrm{ind}}$:

(18) $$\boldsymbol{q}_{\text{ind}}^{s} = [Q_{\text{ind}}^{s}(1), \cdots, Q_{\text{ind}}^{s}(R)]^{\text{T}}.$$

After that, $\boldsymbol{q}_{\text{ind}}^{s}$ is quantized via the vector Huffman coding.

The $W$ root-mean-square values are expanded to form a $R$-dimensional vector $\boldsymbol{q}_{\text{rms}}^{s}$ through

(19) $$\boldsymbol{q}_{\text{rms}}^{s} = [\underbrace{Q_{\text{rms}}(1), \cdots, Q_{\text{rms}}(1)}_{B \text{ values}}, \cdots, \underbrace{Q_{\text{rms}}(W), \cdots, Q_{\text{rms}}(W)}_{B \text{ values}}]^{\text{T}},$$

and is further quantized.

### 3.3. Recovering audio objects using compressed sensing

In the decoding stage, the received vectors $\boldsymbol{y}_{\text{sign}}^{s}$, $\boldsymbol{q}_{\text{ind}}^{s}$, $\boldsymbol{q}_{\text{rms}}^{s}$ are utilized to initially recover the observation signal $\hat{\boldsymbol{y}}^{s}$:

(20) $$\hat{\boldsymbol{y}}^{s} = Q_{\text{step}} \cdot \boldsymbol{y}_{\text{sign}}^{s} \otimes \boldsymbol{q}_{\text{ind}}^{s} \otimes \boldsymbol{q}_{\text{rms}}^{s}.$$

Subsequently, the observation matrix $\hat{\boldsymbol{Y}}^{s}$ can be attained by transforming the observation vector to a $D \times M$ matrix. For convenience, we denote $\hat{\boldsymbol{Y}}^{s}$ as a group of column vectors:

(21) $$\hat{\boldsymbol{Y}}^{s} = [\hat{\boldsymbol{y}}_{(1)}^{s}, \cdots, \hat{\boldsymbol{y}}_{(M)}^{s}].$$

According to the CS principles, recovering the vector $\boldsymbol{\theta}_{m}^{s}$ for the given $\hat{\boldsymbol{y}}_{(m)}^{s}$ depends on the *coherence* between the matrix $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. Note that $\boldsymbol{\Psi}$ is the basis matrix used in (3). $\boldsymbol{\Phi}$ is an orthonormal sensing matrix of size $L \times L$. $\hat{\boldsymbol{\Phi}}$ used in (11) is obtained by selecting $D$ rows uniformly and randomly from $\boldsymbol{\Phi}$. Thus, the coherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ is defined as

(22) $$\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) = \sqrt{L} \cdot \max_{1 \le k, j \le L} \left| \boldsymbol{\phi}_{k}^{\text{T}} \boldsymbol{\psi}_{j} \right|,$$

where $\boldsymbol{\phi}_{k}^{\text{T}}$ and $\boldsymbol{\psi}_{j}$ are the $k$-th row and $j$-th column of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, respectively. It should be noted that the co-domain of the coherence $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ is ranged from 1 up to $\sqrt{n}$, where the small value of $\mu$ usually leads to a small number of measurements, i.e., the number of sensing points $D$.

On the basis of the theorem proposed in [13], if $R$ satisfies

(23) $$R \ge C \cdot \mu^{2}(\boldsymbol{\Phi}, \boldsymbol{\Psi}) \cdot K \cdot \log L,$$

for some positive constant $C$, then $\boldsymbol{\theta}_{m}^{s}$, $m = 1, \cdots, M$, can be exactly reconstructed by solving the following convex $\ell_1$ minimization problem:

(24) $$\begin{aligned} &\text{minimize} && \| \boldsymbol{\theta}_{m}^{s} \|_{1}, \\ &\text{subject to} && \hat{\boldsymbol{\Phi}} \boldsymbol{\Psi} \boldsymbol{\theta}_{m}^{s} = \hat{\boldsymbol{y}}_{(m)}^{s}. \end{aligned}$$

Through the aforementioned procedures, the sparse approximation signals for all stages are attained. Thus, the TF representation of all object signals $\hat{\boldsymbol{x}}_{m}$, $m = 1, \cdots, M$, is recovered by

(25) $$\hat{\boldsymbol{x}}_{m} = \sum_{s=1}^{S} \boldsymbol{\theta}_{m}^{s}.$$

Lastly, taking the inverse transform (3) can yield the object signals $\hat{\boldsymbol{s}}_m$, $m = 1, \cdots, M$.

## 4. Evaluations

Both objective and subjective evaluations are taken to examine the performance of the proposed encoding approach. The test audio data are selected from the QUASI audio database [15], containing various types of audio objects (e.g., piano, vocal, violin, etc.) sampled at 44.1 kHz. There are 6 multi-track audio files produced, where each file consisting of 8 tracks served as a group of 8 simultaneously occurring audio object signals. Both the instruments and musical notes vary for each track.

In this paper, the inverse Discrete Cosine Transform (DCT) matrix with a size 2048×2048 (i.e., $L$=2048) is used as the basis $\boldsymbol{\Psi}$. An orthonormal matrix with Gaussian entries with a size of 2048×2048 serves as the matrix $\boldsymbol{\Phi}$. It can be proved that the coherence $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) \approx \sqrt{2 \log L}$ in this case. The NPTF $K$ is set to 64 per stage for each active object signal. The size of the sensing matrix $\hat{\boldsymbol{\Phi}}$ is 256×2048, i.e., $D$=256.

For comparison, the encoding approach proposed in [10] serves as the reference approach. This approach employs a 2048-points Short Time Fourier Transform (STFT) with 50% overlapping, where the number of the DFT points in each frame is also 2048.

### 4.1. Objective evaluations

In this subsection, the lossless transmission case is considered, i.e., the observation signals are encoded via lossless techniques. As for the reference approach [10], both the mono down-mix signal and the side information are lossless encoded. To test the performance of the proposed multi-stage framework, the 1-stage, 2-stage, and 4-stage encoding schemes are evaluated respectively. These three schemes are respectively denoted by "MSPA-1", "MSPA-2", and "MSPA-4". Therefore, the 1-stage scheme preserves 64 TF instants per frame for each object signal. The 2-stage scheme preserves 128 TF instants per frame. The 4-stage scheme preserves 256 TF instants per frame, which is the same as the number preserved in the reference approach [10] denoted by "SPA". The FEPR defined by (1) is used as a measure in the objective evaluations.

The results presented in Fig. 3 are with 95% confidence intervals. It can be observed that the higher stage leads to higher FEPR, which validates that the proposed approach can achieve a scalable transmission. Compared to the reference approach, the suggested approach achieves slightly higher FEPR when preserving the same number of TF instants per frame for each object signal.

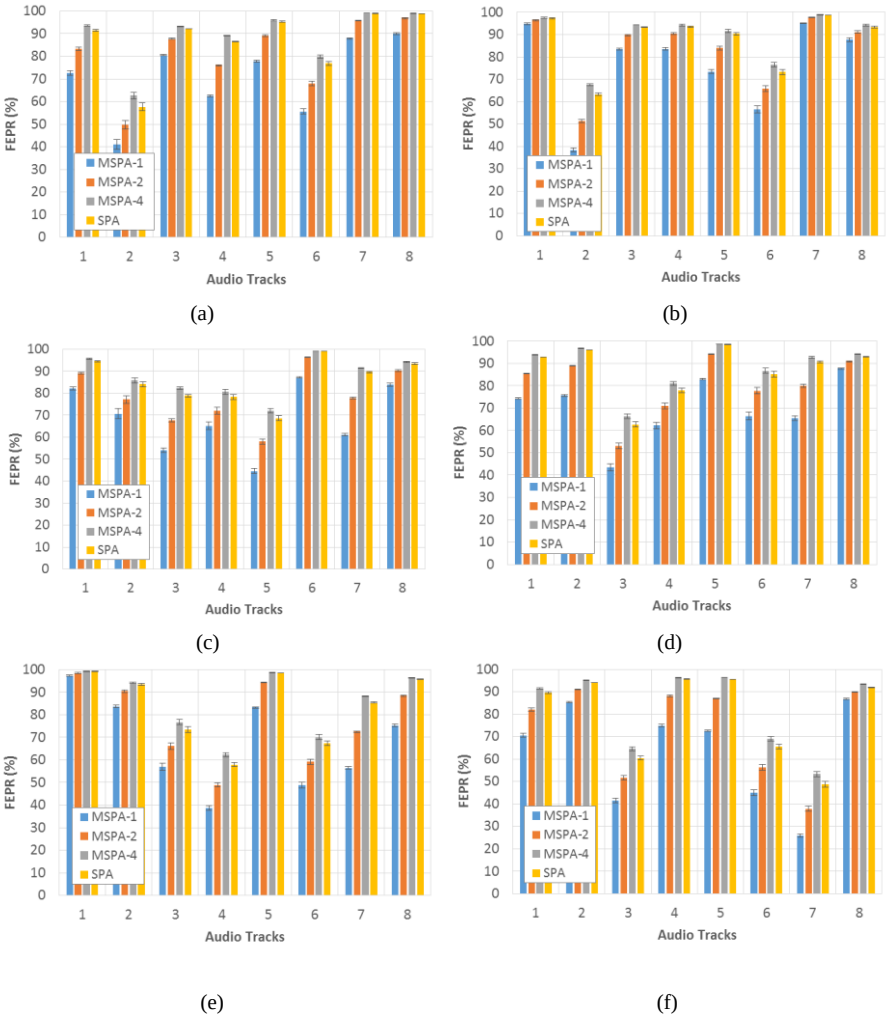


Fig. 3. FEPR results for the proposed multi-stage encoding approach and the reference approach, (a)-(f) represent the results for six multi-track audio files

## 4.2. Perceptual similarity evaluations

The multi-track audio files used in the last evaluation are further used to examine the perceptual quality. The Perceptual Similarity Measure (PSM) score generated by the evaluation tool PEMO-Q [16] is adopted to compare the perceptual similarity between the decoded object and the original one. It should be noted that the PSM score is ranged from −1 up to 1, where the larger value indicates the better perceptual quality of the decoded object. The average PSM score for each multi-track audio file is computed independently.

Results are shown in Fig. 4. It can be observed that the perceptual similarity becomes better as the number of stages increases, which confirms that the perceptual quality of each audio object can be maintained.

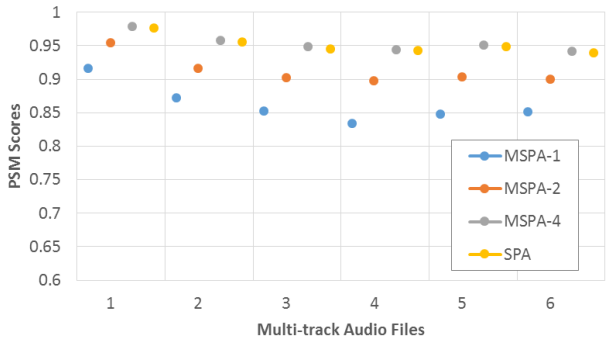


Fig. 4. PSM scores for the proposed multi-stage encoding approach and the reference approach in the lossless transmission case

Furthermore, the transmission case with loss is considered at a certain bitrate. Specifically, the proposed 2-stage encoding scheme was chosen for evaluation. Both the 1st and the 2nd stage observation signals were further encoded via SQVH technique. The quantization constants are shown in Table 1.

Table 1. Values for the quantization constants

| Constant | $Q_{\text{step}}$ | $Q_{\text{offset}}$ | $Q_{\text{step}}$ | $B$ |
|---|---|---|---|---|
| Value | $2^{-1.5}$ | 0.3 | 13 | 20 |

The bitrate is about 224 kbps for encoding the two observation signals. To occupy approximately the same total bitrate for the reference approach, the down-mix signal is encoded using the MPEG-2 AAC [17] at 128 kbps, while encoding the side information via the Run Length Coding (RLC) [18] and the Golomb-Rice coding [19] at about 90 kbps.

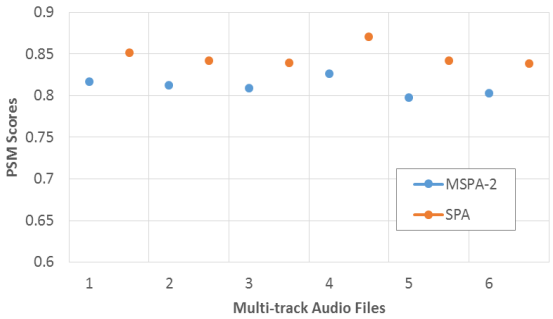


Fig. 5. PSM scores for the proposed multi-stage encoding approach and the reference approach in the transmission case with loss

The PSM scores are presented in Fig. 5. The proposed encoding approach achieves similar, but slightly lower scores compared to the reference approach at the same bitrate.

One of the reasons for this is that the reference approach preserves twice the number of TF instants per frame compared to the 2-stage condition. On the other hand, quantizing the observation signals introduces Gaussian noise distributed

throughout the whole frequency band, which degrades the perceptual quality of the decoded objects. Overcoming this issue will be considered as a future work. Nevertheless, the key advantage of the proposed scheme is that it does not require any produced signal to be encoded via lossless techniques, and hence is more robust in practical network transmission.

## 5. Conclusions

In this paper a multi-stage encoding scheme for multiple audio objects was proposed. The scheme is based on intra-object sparsity. Unlike the existing "down-mix plus side information" framework, the approach proposed encodes multiple audio objects into multi-stage observation signals for transmission. The observation signal contains the dominant TF instants of all active object signals, which is recovered via CS techniques during the decoding phase. The evaluations validated that the multi-stage scheme can achieve scalable transmission. Further work could include selecting a more suitable basis function to improve the perceptual quality of the decoded object signals.

***Acknowledgments*:** The authors would like to thank Dr. Rainer Huber for his help in the evaluation software. This work has been supported by the National Natural Science Foundation of China (No 61231015, 61201197), Specialized Research Fund for the Doctoral Program of Higher Education of the Peoples Republic of China (No 20121103120017), and the Scientific Research Project of Beijing Educational Committee (No KM201310005008).

## References

1. BS.775 Int. Telecommunication Union. Multichannel Stereophonic Sound System with and Without Accompanying Picture. 2006.
2. Hamasaki, K. A 22.2 Multichannel Sound System for Ultrahigh Definition TV (UHDTV).– SMPTE Motion Imaging Journal, Vol. **117**, 2008, No 3, pp. 40-49.
3. Smolic, A. An Overview of 3rd Video and Free Viewpoint Video. – In: Proc. of 13th International Conference on Computer Analysis of Images and Patterns, Springer, Münster, Germany, 2009, pp.1-8.
4. Tanimoto, M. Overview of Free Viewpoint Television. – Signal Processing: Image Communication, Vol. **21**, 2006, No 6, pp. 454-461.
5. Dolby Laboratories. Dolby ATMOS Cinema Specifications. 2014. **http://www.dolby. com/ us/en/technologies/ dolbyatm-os/dolby-atmos-specifications.pdf**
6. Herre, J., H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. H¨olzer, M. L. Valero, B. Resch, H. Mundt, H. O. Oh. MPEG Spatial Audio Object Coding – The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes. – Journal of the Audio Engineering Society, Vol. **60**, 2012, No 9, pp. 655-673.
7. Liutkus, A., S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, G. Richard. Informed Audio Source Separation: A Comparative Study. – In: Proc. of 20th European Signal Processing Conference, EURASIP'12, Bucharest, Romania, 2012, pp. 2397-2401.
8. Ozerov, A., A. Liutkus, R. Badeau, G. Richard. Coding-Based Informed Source Separation: Nonnegative Tensor Factorization Approach. – IEEE Transactions on Audio, Speech and Language Processing, Vol. **21**, 2013, No 8, pp. 1699-1712.

9. Z h e n g, X., C. R i t z, J. T. X i. A Psychoacoustic-Based Analysis-Bysynthesis Scheme for Jointly Encoding Multiple Audio Objects into Independent Mixtures. – In: Proc. of 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vancouver, Canada, 2013, pp. 281-285.
10. J i a, M., Z. Y a n g, C. B a o, X. Z h e n g, C. R i t z. Encoding Multiple Audio Objects Using Intra-Object Sparsity. – IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. **23**, 2015, No 6, pp. 1082-1095.
11. J i a, M., C. B a o, X. L i u. An Embedded Speech and Audio Coding Method Based on Bit-Plane Coding and SQVH. – In: Proc. of IEEE International Symposium on Signal Processing and Information Technology, IEEE, Ajman, UAE, 2009, pp. 43-48.
12. C a n d e s, E. J., M. B. W a k i n. An Introduction to Compressive Sampling. – IEEE Signal Processing Magazine, Vol. **25**, 2008, No 2, pp. 21-30.
13. C a n d e s, E. J., J. K. R o m b e r g, T. T a o. Stable Signal Recovery from Incomplete and Inaccurate Measurements. – Communications on Pure and Applied Mathematics, Vol. **59**, 2006, No 8, pp. 1207-1223.
14. S o h n, J., N. S. K i m, W. S u n g. A Statistical Model-Based Voice Activity Detection. – IEEE Signal Processing Letters, Vol. **6**, 1999, No 1, pp. 1-3.
15. QUASI Database – a Musical Audio Signal Database for Source Separation. **http://www.tsi. telecomparistech.fr/aao/en/2012/03/12/quasi/**
16. H u b e r, R., B. K o l l m e i e r. PEMO-Q: A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. – IEEE Transactions on Audio, Speech and Language Processing, Vol. **14**, 2006, No 6, pp. 1902-1911.
17. B o s i, M., K. B r a n d e n b u r g, S. Q u a c k e n b u s h, L. F i e l d e r, K. A k a g i r i, H. F u c h s, M. D i e t z. ISO/IEC MPEG-2 Advanced Audio Coding. – Journal of the Audio Engineering Society, Vol. **45**, 1997, No 10, pp. 789-814.
18. G o l o m b, S. Run-Length Encodings (Corresp.). – IEEE Transactions on Information Theory, Vol. **12**, 1966, No 3, pp. 399-401.
19. R i c e, R., J. P l a u n t. Adaptive Variable – Length Coding for Efficient Compression of Spacecraft Television Data. – IEEE Transactions on Communication Technology, Vol. **19**, 1971, No 6, pp. 889-897.