A PSYCHOACOUSTIC-BASED ANALYSIS-BY-SYNTHESIS SCHEME FOR JOINTLY ENCODING MULTIPLE AUDIO OBJECTS INTO INDEPENDENT MIXTURES

Xiguang Zheng, Christian Ritz, and Jiangtao Xi

ICT Research Institute/School of Electrical Computer and Telecommunications Engineering,
University of Wollongong, Wollongong, NSW, Australia, 2522
{xz725, critz, jiangtao}@uow.edu.au

ABSTRACT

Perceptually accurate representation of audio objects obtained from multi-track audio signals is desired for applications such as interactive soundfield rendering and browsing. Presented in this work is a scalable psychoacoustic analysis-by-synthesis approach to extract the perceptually dominant time-frequency audio objects from a multi-track audio signal. The proposed compression framework exploits sparsity in the perceptual time-frequency domain where up to eight audio objects can be efficiently encoded using only two audio mixtures with side information representing the origin of the time-frequency instances in the mixture signals. The proposed approach, judged by both objective and subjective tests, results in superior audio quality compared to existing techniques when encoding more than 5 audio objects.

Index Terms— Multichannel audio compression, Audio object coding, Soundfield navigation

1. INTRODUCTION

Multichannel surround audio gives the audience an improved listening experience by providing a vivid surround soundfield compared to the traditional stereophonic techniques. Among these multichannel audio formats, the most popular format is the 5.1 surround audio format which can be compressed by the standardized MPEG-Surround (MPS) Codec [1], [2]. However, such multichannel audio formats do not provide interaction and personalization for the individual audiences, i.e. each audience can only perceive the received soundfield in a pre-rendered manner that cannot be adjusted by the individual user.

An alternative solution for preserving the audio scene is to compress the audio objects within the soundfield such that the audio objects can be separated and rendered according to different requirements of the audiences in the receiver end. Such personalized soundfield rendering or soundfield browsing is becoming increasingly demanded for the applications in the emerging 3DTV [3] and free viewpoint TV (FTV) [4]. Techniques for compressing multitrack audio objects (i.e. piano, guitar, vocal, etc) have been proposed such as Spatial Audio Object Coding (SAOC) [5], [6] and Informed Source Separation (ISS) [7], [8]. The SAOC approach is inspired by the MPEG-Surround (MPS) framework [1] such that these "dry" audio sources are treated as separate audio objects in the time-frequency domain and encoded into a stereo or mono downmix signal representing the dominant (highest energy) time-frequency object along with side information representing the inter-object relationships in the time-frequency domain. The side information is encoded using 2-3 kbps per audio object [6]. Decoding of the downmix and transcoding of the side information allows derivation of the playback loudspeaker signals using the existing MPS decoder. To achieve the same goal, ISS is aimed more specifically at deriving a stereo mixture signal such that the most dominant audio objects in a scene can be separated from the mixture with high quality. This is based on selecting the time-frequency instants amongst all objects with maximum energy (results in [7] suggest that only two sources per time-frequency are required to capture the majority of energy when there are less than 5 audio objects).

In this work, a Psychoacoustic Analysis-By-Synthesis (PABS) framework, previously applied to speech sources in [9], [10], is generalized to the scalable compression of audio objects. The PABS technique is employed ensuring that the most perceptually important components of each audio object are maintained. Similar to the compressed audio format of [2], the proposed framework extracts the perceptually dominant audio sources in the time-frequency domain which are then preserved in a two-channel mixture signal with side information indicating the origin of the audio objects preserved at each time-frequency. Alternatively, the side information can be embedded within the mixture using the methods presented in [7].

The key contribution of the proposed framework is the generalization of the PABS technique [9], [10], previously applied to speech sources in [9], [10] using one dominant source for each time-frequency, to the scalable compression of audio objects where more than one dominant audio object in the time-frequency domain can be determined. The generalization is achieved by the proposed Multi-level Psychoacoustic Analysis-By-Synthesis (M-PABS) framework. Compared to [7], the proposed approach allows for selective decoding and playback of more than 5 individual audio objects using the same number of downmixing channels whilst maintaining perceptual quality as judged by the subjective listening test. In comparison to SAOC [5], [6], the proposed M-PABS scheme decomposes the audio objects into multi-level audio mixtures ordered by their perceptual importance as judged by psychoacoustic masking curves derived for each object. Thus instead of uniformly quantizing the time-frequency domain inter-object parameters, the perceptual-based multi-level decomposition highlights the perceptual importance among the time-frequency domain overlapping audio objects and thus is more suitable for bandwidth constrained compression (i.e. the available bitrates can be allocated to more perceptually important parameters in a water filling manner). In addition, decomposing the audio objects into multiple levels is more suitable for scalable or distributed transmission and packet loss concealment. For instance, the decomposed mixtures can be transmitted through different independent channels by Multiple Description Coding [11] (MDC) as well as protected by

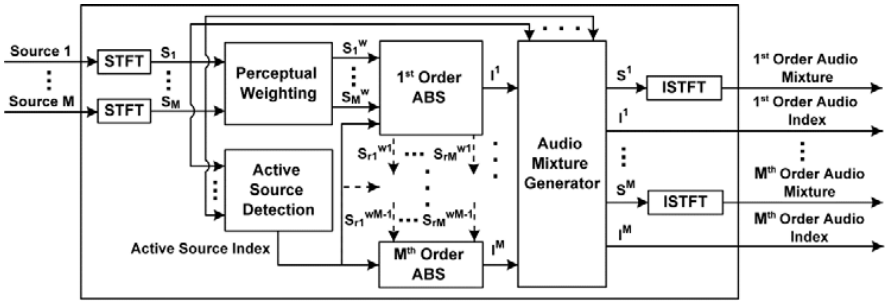

Fig.1 Overview of the proposed framework

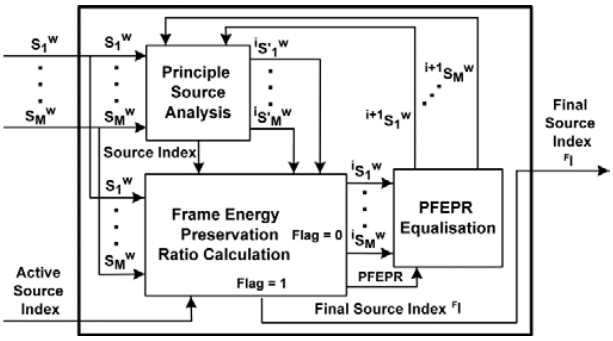

Fig.2 General ABS Scheme

packet loss concealment techniques such as Forward Error Correction (FEC) by introducing different amount of redundancy according to the perceptual importance.

The remainder of the paper is organized as follows: Section 2 describes the derivation of the generalized PABS framework for compressing audio objects. Objective and subjective experimental results are presented in Section 3, while conclusions are drawn in Section 4.

2. PROPOSED FRAMEWORK

The proposed framework as shown in Fig.1 consists of multi-level Analysis-By-Synthesis (ABS) encoders. The audio objects (Source 1 to Source M) are transformed to the time-frequency domain by Short Time Fourier Transform (STFT) and then perceptually weighted prior to the multi-level ABS encoders. For instance, at the 1^{st} order ABS encoder, the 1^{st} order perceptual dominant time-frequency audio objects are analyzed (see also Fig. 2 for detailed illustration of the general ABS encoder) which is represented by the Final Source Index (I^1) indicating the origin of these 1^{st} order perceptual dominant audio objects. The perceptual domain 1^{st} order residue signals ($S_{r1}{}^{w1}$ to $S_{rM}{}^{w1}$) are inputs to the next (i.e. the 2^{nd} order) ABS encoder to obtain the 2^{nd} order perceptually dominant time-frequency audio objects. This multi-level decomposition proceeds until the majority of the energy of the input objects have been preserved by the j^{th} ($1 \le j \le M$) order ABS encoder. In the Audio Mixture Generator, the 1^{st} to j^{th} order time-frequency domain mono audio mixtures are derived from the input audio sources using the 1^{st} to j^{th} order source indexes, respectively. It should be noted that the proposed framework in Fig. 1 generalizes the PABS framework originally designed to compress simultaneous speech signals in [9], [10] by preserving only the 1^{st} order mixture and source index. Here, up to M^{th} order mixtures are supported in order to preserve sufficient perceptually important time-frequencies for simultaneous audio objects.

2.1. Deriving Principle Audio Time-Frequency Objects

The proposed M-PABS framework is illustrated in Fig. 1 and is performed on a frame-by-frame basis. For frame n ($1 \le n \le N$), input time-frequency representation $S_m(n,k)$ ($1 \le m \le M$, k is the frequency index) is firstly transferred to the perceptual domain as $S_m{}^w(n,k)$ by applying a perceptual weighting function derived separately for each object:

$$S_m{}^w(n,k) = A_m(n,k) \cdot S_m(n,k), m \in [1,M] \quad (1)$$

where $A_m(n,k)$ is the perceptual weighting function of the m^{th} source determined as the inverse of the perceptual masking threshold energy in the MP3 audio codec [12]. In the general ABS encoder shown in Fig. 2, the set of perceptual domain signals $S_m{}^w(n,k)$ are analyzed by the Principle Source Analysis block to obtain the initial perceptual principle time-frequency sources of the input time-frequency audio objects ${}^0S_p{}^w(n,k)$. Note that only the 1^{st} order initial perceptual principle source ${}^0S_p{}^w(n,k)$ is generated from $S_m{}^w(n,k)$. The j^{th} ($2 \le j \le M$) principle source is obtained from the j-1^{th} ($2 \le j \le M$) order residue signals (as shown in Fig. 1 and discussed further in Section 2.3). For the updated perceptual principle source (i.e. ${}^iS_p{}^w(n,k)$, $i > 0$), it is obtained from the previous iteration. The initial perceptual principle time-frequency sources are given by:

$${}^0S_p{}^w(n,k) = \max_m (S_m{}^w(n,k)), m \in [1,M] \quad (2)$$

If m_p denotes the audio object of the corresponding principle time-frequency source, a initial source index ${}^0I$ is employed to indicate the origin of the initial principle time-frequency instants given by:

$${}^0I(n,k) = m_p \quad (3)$$

Thus the initial principle time-frequency components of audio object m in the perceptual domain (i.e. ${}^iS'_m{}^w(n,k)$ with $i = 0$) can be recovered from the initial principle time-frequency mixture ${}^0S_p{}^w(n,k)$ using the initial source index ${}^0I(n,k)$ as:

$${}^0S'_m{}^w(n,k) = \begin{cases} S_m{}^w(n,k), \text{ if } {}^0I(n,k) = m \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

2.2. Analysis-by-Synthesis Scheme

After the initial principle time-frequency is obtained, the Analysis-By-Synthesis (ABS) loop starts to refine the principle time-

frequencies by updating the source index. The aim of the ABS iteration is to preserve the perceptually weighted energy for each audio object in the final principle time-frequency mixture on a frame-by-frame basis. Suppose ${}^{i}S'_{m}{}^{w}(n,k)$ is the selected time-frequency instant for the i^{th} ABS loop (for the initial loop of the 1^{st} order ABS, $i = 0$ where ${}^{0}S'_{m}{}^{w}(n,k)$ is obtained by (1) to (4); for the initial loop of the j^{th} ($2 \le j \le M$) ABS, ${}^{0}S'_{m}{}^{w}(n,k)$ is obtained by (2) - (4) where the input is the j-1^{th} ($2 \le j \le M$) order perceptual domain residue signal), it will be sent to the Frame Energy Preservation Ratio Calculation block where the Perceptual Frame Energy Preservation Ratio [10] (*PFEPR*) is analyzed. The *PFEPR* is the ratio between the selected ${}^{i}S'_{m}{}^{w}(n,k)$ and $S_m{}^{w}(n,k)$, which is given by:

$$ {}^{i}PFEPR_{m}{}^{n} = \sum_{k=1}^{K} \left\| {}^{i}S'_{m}{}^{w}(n,k) \right\| \Big/ \sum_{k=1}^{K} \left\| {}^{i}S'_{m}{}^{w}(n,k) \right\| \quad (5) $$

Hence, the approximated equal energy preservation for each audio object in the final principle source mixture can be achieved if the maximum difference among the *PFEPR*s of the active audio sources in the operating frame is minimized, i.e. for the n^{th} frame, the iteration terminates at $i = F$ if

$$ F = \arg\left[\min_{i} \left(\max_{m} \left({}^{i}PFEPR_{m}^{n} \right) - \min_{m} \left({}^{i}PFEPR_{m}^{n} \right) \right) \right] \quad (6) $$

The Active Source Detection block will detect the active audio objects in the operating frame using the Voice Activity Detector in [13]. For "dry" audio sources, an inactive source means it is muted in the current frame. If the maximum *PFEPR* difference in the current iteration is smaller than the previous iteration, more time-frequency components from a lower *PFEPR* object will be included in the next iteration (i.e. ${}^{i+1}S_{p}{}^{w}(n,k)$). In the *PFEPR* Equalization block, the active source with the lowest *PFEPR* in the current frame is amplified by a factor a. The selection of this factor is based on the trade-off between the number of iterations and the accuracy. $a = 1.01$ is used in the evaluation section to achieve the presented results in Fig. 3 and Fig. 4. Assuming the m_l^{th} active source has the lowest *PFEPR*, the amplified source ${}^{i}S_{m_l}{}^{w}$ and other active sources for the next ABS iteration is given by:

$$ {}^{i+1}S_{m_l}{}^{w}(n,k) = {}^{i}S_{m_l}{}^{w}(n,k) \cdot a \quad (7) $$

$$ {}^{i+1}S_{m}{}^{w}(n,k) = {}^{i}S_{m}{}^{w}(n,k) \quad 1 \le m \le M,\ m \ne m_l \quad (8) $$

${}^{i+1}S_{m}{}^{w}(n,k)$ ($1 \le m \le M$) will be sent back to the Principle Source Analysis block to obtain ${}^{i+1}S'_{m}{}^{w}(n,k)$ for the $i+1^{th}$ iteration. Principle Source Generation proceeds (Flag = 1 as shown in Fig. 2) if the maximum *PFEPR* difference among the active objects in the operating frame is minimized (i.e. (6) is satisfied). The final source index (${}^{F}I$), representing the origin of selected audio objects in the F^{th} iteration of the current order ABS encoder that minimize (6), is the output of the current order ABS encoder which will be sent to the Audio Mixture Generator block to generate the j^{th} ($1 \le j \le M$) order audio mixture.

2.3. Multi-level Audio Object Extraction

Once the 1^{st} order time-frequency source index is obtained, as shown in Fig. 1 and Fig. 2, the residue signal is fed into the 2^{nd} order ABS block to obtain the 2^{nd} order time-frequency audio objects by applying the same ABS method. This process proceeds to the j^{th} ($1 \le j \le M$) order until the majority of the time-frequencies of the input audio objects have been preserved in the 1^{st} to j^{th} order audio mixtures. It should be noted that scalable compression can be achieved by using M-PABS since the importance of the time-frequency audio objects is indicated by different orders of the audio mixture extraction, i.e. lower order audio mixtures are perceptually prioritized over higher order audio mixtures, which enables bit rate adaptation based channel bandwidth constraints. The j^{th} ($1 \le j \le M$-1) order perceptual domain residue signal (S_{rm}^{wj}) for the m^{th} source is obtained by:

$$ S_{rm}^{wj}(n,k) = \begin{cases} S_{m}^{w}(n,k) - {}^{F}S'^{wj}_{m}(n,k), & j=1 \\ S_{rm}^{wj-1}(n,k) - {}^{F}S'^{wj}_{m}(n,k), & 2 \le j \le M \end{cases} \quad (9) $$

where ${}^{F}S'^{wj}_{m}$ is the preserved time-frequencies of source m in the j^{th} order ABS encoding, obtained using the final encoding mask ${}^{F}I^{j}$ (i.e. using (4) replace ${}^{0}I^{j}$ with ${}^{F}I^{j}$). ${}^{F}I^{j}$ is simplified by I^{j} in Fig. 1.

After the multi-level ABS time-frequency source extraction stage, the final source indices generated by each level of the ABS encoder are sent to the audio mixture generator where the time-frequency domain audio mixtures are formed. Since the perceptually important source origin of the j^{th} order time-frequency audio objects has been specified in ${}^{F}I^{j}$, the j^{th} order time-frequency audio mixture S^{j} can be extracted from the input audio objects in the time-frequency domain by:

$$ S^{j}(n,k) = S_{m}(n,k), if\ {}^{F}I^{j}(n,k) = m \quad (10) $$

Thus the audio mixtures and indices indicating the origin of the audio objects are obtained.

2.4. Recovering Audio Objects from the Audio Mixtures

The audio objects can be recovered from the audio mixtures with the aid of audio indices. For the m^{th} audio object, $S'_{m}(n,k)$ is recovered by:

$$ S'_{m}(n,k) = \sum_{j=1}^{L} S'^{j}_{m}(n,k), \ 1 \le L \le M \quad (11) $$

where L is the number of audio object mixtures, $S'^{j}_{m}(n,k)$ is the audio object m extracted from the j^{th} order audio mixture $S^{j}(n,k)$ using the corresponding audio index:

$$ S'^{j}_{m}(n,k) = \begin{cases} S^{j}(n,k), \ if\ I^{j}(n,k) = m \\ 0, \ otherwise \end{cases} \quad (12) $$

It should be noted that compression of the audio objects can be achieved by using fewer mono audio mixtures with audio object indices than the number of audio objects (i.e. $L < M$). In the ISS method [7], only two of the highest time-frequency components among five audio objects are considered ($L = 2$ without using M-PABS). Thus as suggested by [7], at least two time-frequencies among up to five overlapping audio objects should be preserved in a stereo downmixed mixture. In this work, results in Section III indicate that up to eight audio objects can be reliably compressed using the same downmixing format (i.e. $L = 2$) by applying the proposed M-PABS framework. The bitrate when using two order PABS is limited to approximately 134 kbps (i.e. 64 kbps for perceptual lossless legacy encoding of the mono audio mixture and 2-3 kbps for the side information of each mixture [10]).

3. EVALUATIONS

Both objective and subjective testing results are presented in this section. A test database of multi-track audio signals [14] sampled

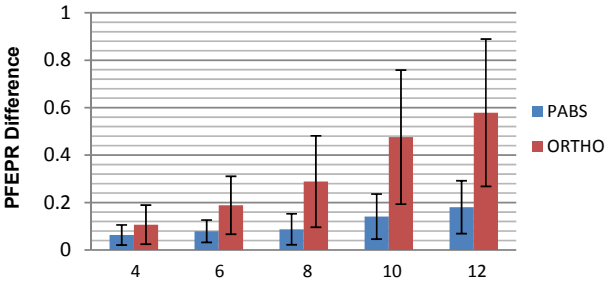

Fig. 3 Maximum PFEPR Differences

at 44.1 kHz and containing simultaneous audio objects including one or more guitars, violins, drums, pianos, horns and vocal tracks was created.

3.1. Objective Evaluation

Eight multi-track audio signals were created for each case of 4, 6, 8, 10 and 12 simultaneous objects. These were processed by the orthogonal approach in [7] and the proposed M-PABS approach using a short time Fourier Transform with a window size of 2048 samples and 50% overlapping to obtain the 1^{st} and 2^{nd} audio mixtures. The objective evaluation compares the *PFEPR*s generated from the recovered individual audio objects (S'_m) and the original audio objects (S_m) using the proposed M-PABS approach (condition 'PABS') and the orthogonal approach (condition 'ORTHO') in [7] (two highest energy time-frequencies among all active sources are selected in [7]). Results are illustrated in Fig. 3 where the maximum *PFEPR* difference among all active sources is presented (error bars represent 95% confident interval). It can be observed in Fig. 3 that while the maximum *PFEPR* differences are similar for the 4 audio objects case, a significant increase of the maximum *PFEPR* difference is observed when the number of audio objects increases for the 'ORTHO' condition (up to 0.6 on average for 12 audio objects). The *PFEPR*s differences for 'PABS' are less (below 20%) than those for 'ORTHO' in all conditions, especially when the number of the audio objects increases (the difference is as high as 40% on average).

3.2. Subjective Evaluation

A MUSHRA [15] listening test was also employed. Six audio objects covering all instruments and vocal tracks are recovered from 4, 8 and 12 simultaneous audio object mixtures created in the objective evaluation. Besides condition 'PABS' and 'ORTHO', a Hidden Reference and a 3.5 kHz low pass filtered anchor are included for the MUSHRA Test. 15 listeners participated in this test and the results are shown from Fig. 4 (a) to Fig. 4 (c) with 95% confidence intervals.

It can be observed that the proposed framework shared similar perceptual quality with the orthogonal approach in [7] for 4 overlapping object conditions, as shown in Fig. 4 (a). This is expected since for smaller number of overlapping objects, the probability for more than two active audio objects in one time-frequency region is less than the cases for more overlapping audio objects. However, for more complicated cases (i.e. 8 simultaneous audio objects), the proposed PABS approach significantly outperformed the orthogonal approach, which is consistent with the objective evaluation results. This can be confirmed in Fig. 4 (b) where the MUSHRA scores for the proposed PABS condition are around 80 indicating 'Good' subjective quality compared to the Hidden Reference. In comparison, the ORTHO condition significantly degrades when

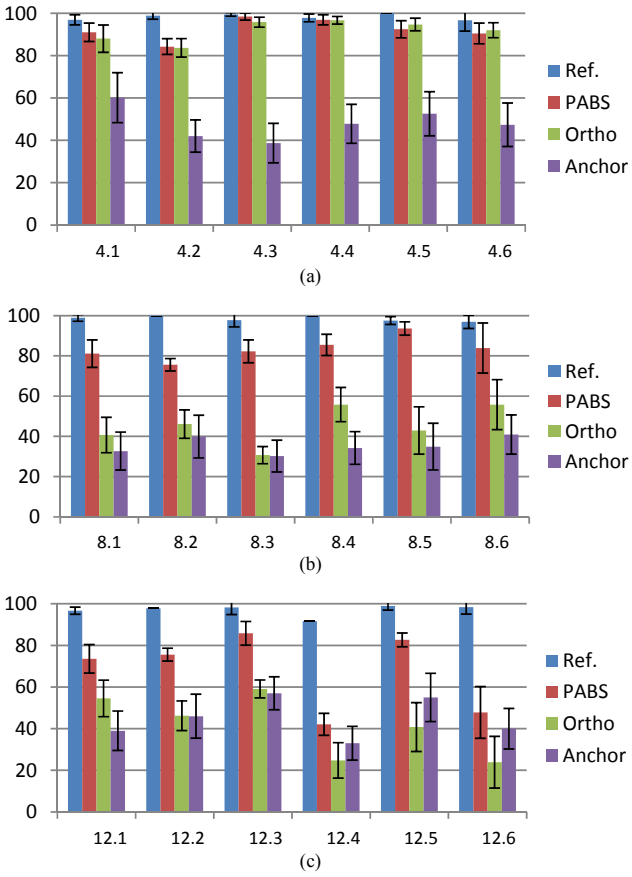

Fig. 4 MUSHRA Test Results.(a) 4 Simultaneous Audio Objects, (b) 8 Simultaneous Audio Objects, (c) 12 Simultaneous Audio Objects

there are 8 overlapping audio objects (40 in MUSHRA results of Fig. 4 (b)). For more complicated cases (i.e. the 12 object audio samples), Fig. 4 (c) shows the limitation of the proposed approach where the quality of the decoded sources degrades for some cases, although the score is still better than condition 'Ortho'.

4. CONCLUSION

A new multi-level psychoacoustic-based analysis-by-synthesis (M-PABS) framework for compressing multi-track audio objects has been proposed. The proposed compression framework generalizes from the previous works in speech by using a serial multi-stage perceptual analysis-by-synthesis technique in order to extract the perceptual important time-frequency audio objects. The perceptual quality of the recovered audio objects as judged by objective and subjective tests confirms that this approach provides significantly improved audio quality compared to the approach of [7] when encoding multi-track audio containing more than five (and up to eight) audio objects.

Acknowledgements

This work has been supported by the Australian Research Council (ARC) through the grant DP1094053.

REFERENCES

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

[2] J. Hilpert and S. Disch, "The MPEG Surround Audio Coding Standard [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 148–52, Jan. 2009.

[3] A. Smolic, "An Overview of 3D Video and Free Viewpoint Video," in *Computer Analysis of Images and Patterns*, vol. 5702, X. Jiang and N. Petkov, Eds. Springer Berlin / Heidelberg, 2009, pp. 1–8.

[4] M. Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, Jul. 2006.

[5] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, and L. Terentiev, "Spatial Audio Object Coding (SAOC) - The Upcoming MPEG Standard on Parametric Object Based Audio Coding," in *Audio Engineering Society Convention 124*, 2008.

[6] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, 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.

[7] M. Parvaix and L. Girin, "Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721–1733, Aug. 2011.

[8] M. Parvaix, L. Girin, and J.-M. Brossier, "A Watermarking-Based Method for Informed Source Separation of Audio Signals With a Single Sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1464–1475, Aug. 2010.

[9] X. Zheng, C. Ritz, and J. Xi, "Encoding navigable speech sources: an analysis by synthesis approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 405 -408, Mar. 2012.

[10] X. Zheng, C. Ritz, and J. Xi, "Encoding Navigable Speech Sources: A Psychoacoustic-Based Analysis-by-Synthesis Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 29 –38, Jan. 2013.

[11] V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[12] ISO, "Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s — Part 3: Audio," Mar, 1999.

[13] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[14] "The Sample Factory R&B Sample," *avaliable at http://www.thesamplefactory.com/baby-makers-sample-cd-p-74.html*.

[15] ITU, "BS. 1534: Methods for the subjective assessment of intermediate quality levels of coding systems." 1997.