ENCODING NAVIGABLE SPEECH SOURCES: AN ANALYSIS BY SYNTHESIS APPROACH

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

This paper presents an analysis-by-synthesis coding architecture for compressing navigable speech sources. The proposed coding scheme encodes multiple overlapped speech sources recorded, for example, during a multi-participant meeting or teleconference, into a mono or stereo mixture signal that can be compressed with an existing speech coder. The individual speech sources can be separated from the received compressed mixture, which allows the listener to determine the active sources and their spatial locations at the reproduction site. The approach was applied to the compression of a series of speech soundfields created from multiple clean speech sentences and real meeting recordings, where each soundfield contained four participants with up to three simultaneous speech sources. At a total bit rate of 48 kbps, the perceptual quality of each decoded speech source, as judged by subjective listening tests, was found to be significantly better than either a non-a-by-s approach or separate encoding of each source at the same overall total bit rate. Subjective listening tests also confirm that the quality of the spatialised speech scene is maintained as well.

Index Terms— Multichannel Speech Coding, Soundfield Navigation, Spatial Teleconferencing

1. INTRODUCTION

The compression and reproduction of multichannel speech is becoming increasingly common for applications such as spatialised teleconferencing as well as personalised soundfield browsing. While spatialised teleconferencing provides a natural communication experience for geographical dislocated participants, soundfield browsing is also desired for entertainment or training applications where user defined preferences are enabled to provide a personalised entertaining and learning environment. These applications require efficient compression techniques that support navigation of simultaneous speech sources at the reproduction site. However, existing techniques generally assume separate recording and compression of each speech source, where the required bitrate increases as more participants become involved.

An alternative compression approach is to employ a spatial audio coder, where each source is treated as a separate channel. Compared to separate encoding of each source, this approach can achieve significantly lower bit rates through joint compression of the multichannel signals [1]. Existing approaches, such as MPEG-Surround [2] and S^3AC [3], have focused on the compression and reproduction of the original multichannel audio signals by applying spatial psychoacoustic models in the time-frequency domain. However, the psychoacoustic models used in spatial audio coders were originally designed for compressing multichannel loudspeaker signals representing a surround audio soundfield, rather than for signals representing unique audio sources Recently, Spatial Audio Object Coding (SAOC) [4] was proposed for compression multiple spatial audio sources rather than loudspeaker channels, which is more suitable for speech teleconferencing and browsing applications as targeted in this paper.

More recently, an analysis-by-synthesis (A-by-S) framework for spatial audio coding was proposed [1], where coding parameters are chosen based on error minimisation between the original and decoded multichannel audio signals. Results showed improved objective quality of the decoded signals compared to the existing MPEG Surround approach. In this paper, an A-by-S coding scheme designated for the compression and navigation of multiple speech objects is proposed. In contrast to [1] where the MPEG-Surround coder is employed, the proposed approach is based on the S^3AC framework in [3], which was previously investigated for the compression of multichannel audio signals into a stereo downmix that can be efficiently compressed with an existing audio coder. Here, this approach is used within an A-by-S coding framework for the compression of simultaneously occurring speech signals, such as in realistic meeting scenarios. Significantly higher perceptual quality of the individual sources when compared to a non-a-by-s approach or separate encoding of each source is achieved. As opposed to [1], the proposed approach does not require any side information to be transmitted along with the stereo mixture signal.

The proposed approach relies on exploiting the sparse property of speech in the time-frequency domain to maintain the individual quality of speech objects. This property has been successfully used for Blind Source Separation (BSS) of speech [5] and here it is used to create a separable mono or stereo mixture signal. At the reproduction end, the mixture signal can be decoded such that the listener is able to interactively choose the speech objects of interest and their spatial locations. It should be noted that merging speech signals into one stream has been proposed in [6], but here the aim is to encode multiple sources into one stream such that the individual speech sources can be decoded with high perceptual quality.

The remainder of the paper is organized as follows: Section 2 provides the theoretical basis for time-frequency sparsity of overlapping speech signals. Section 3 presents the architecture of the proposed approach. Experimental results are presented and analysed in Section 4, while conclusions are drawn in Section 5.

2. EXPLORING SPEECH SPARSITY

Speech signals are known to be sparse in the short-term time-frequency domain. In [5], this is defined by the W-Disjoint Orthogonality (W-DO) [5] as: Two signals s_i and s_j are W-DO if the corresponding time-frequency components of the Short Time Fourier Transform (STFT) can be described by:

$$S_i(n,k) \cdot S_j(n,k) = 0, \ \forall n,k \qquad i \neq j \tag{1}$$

where $S_i(n,k)$ and $S_j(n,k)$ are the time-frequency representations of signal s_i and s_j, respectively, n is the frame number and k is the frequency index. It has been shown in [5] that this property is met 90% for a large database of clean speech sentences and is exploited in Section 3.1. In practice, this assumes that speech sources are recorded by close talking microphones or accurately derived from microphone array recordings (e.g. using blind source separation techniques). To account for times where (1) is not met, this paper proposes the analysis-by-synthesis approach of Section 3.3 to maximise the quality of the decoded individual speech signals.

3. ANALYSIS BY SYNTHESIS APPROACH FOR ENCODING NAVIGABLE SPEECH SOURCES

In Fig.1, the A-by-S coding scheme encodes multi-channel speech signals into a mono or stereo mixture that can be further compressed by speech codecs such as AMR-WB+ [7]. At the reproduction site, the individual speech sources can be decoded and separated from the mixture, which produces a navigable speech soundfield, where a listener can interactively choose to active a speech source (or sources) and move them to desired positions in the reproduced audio scene. The detailed framework of the proposed A-by-S encoder is illustrated in Fig.2 and is described further below.

3.1. Speech Source Mixture Generation

Input mono signals from speaker 1 to speaker M (as shown in Fig.1 and Fig.2), transformed into the time-frequency domain, are denoted by $S_m(n,k)$ where $1 \le m \le M$ and n and k are frame number and frequency index, respectively. In the encoder block of Fig.2, under the assumption of the sparsity of speech, a speech source mixture in the time-frequency domain can be generated by:

$$S_d(n,k) = \max_m (S_m(n,k)), m \in [1, M] \tag{2}$$

where $S_d(n,k)$ is the time-frequency component with maximum energy out of all M speech sources, corresponding to the dominant speaker at this time-frequency instant. If m_d denotes the speech source of the corresponding dominant speaker, the azimuth corresponding to $S_d(n,k)$ is given by:

$$\theta_d(n,k) = \theta_{m_d}(n,k) \tag{3}$$

$S_d(n,k)$ could be transferred back to time domain using an Inverse Short Time Fourier Transform (ISTFT) to create a mono source mixture. The azimuth of (3) is preserved in the time-frequency domain transmitted with $S_d(n,k)$ as side information for decoding.

Alternatively, the information in $S_d(n,k)$ and $\theta_d(n,k)$ could be preserved in a stereo downmix by applying the spatial squeezing approach [3] where no side information is required. The advantages of this approach is that the stereo signal can then be further compressed with a standard stereo compatible audio or speech coder and stored or transmitted in this format. In this approach, a new (squeezed) azimuth is generated by mapping the original azimuth covering 0° to 360° to a new azimuth in the rage from 0° to 60°, which is represented by the two downmix channels. The squeezed azimuth $\theta_S(n,k)$ for the m^{th} source (when $m_d = m$) is:

$$\theta_S(n,k) = \begin{cases} -30, & if\ M = 1 \\ -30 + \dfrac{m-1}{M-1} \cdot 60, & otherwise \end{cases} \tag{4}$$

where the mapping is chosen such that all sources are maximally separated by azimuth in the stereo soundfield.

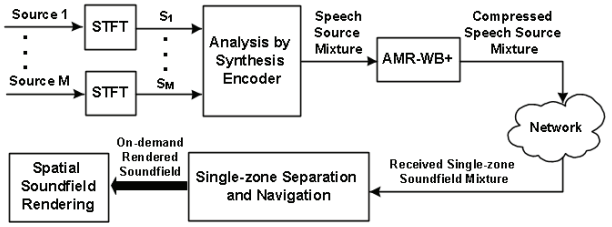

Fig.1. Proposed Analysis by Synthesis Scheme

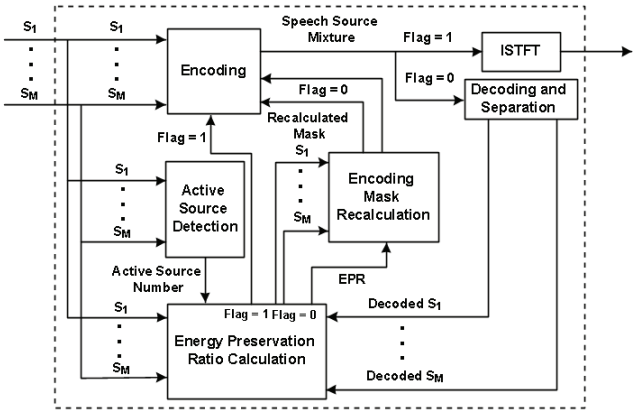

Fig.2. Analysis by Synthesis Encoder

The left and right channel of the proposed stereo mixture in the time-frequency domain, $L_S(n,k)$ and $R_S(n,k)$, respectively, are given by:

$$L_S(n,k) = \frac{S_d(n,k) \cdot (\tan\varepsilon + \tan\theta_S(n,k))}{\sqrt{2\tan^2\varepsilon + 2\tan^2\theta_S(n,k)}} \tag{5}$$

$$R_S(n,k) = \frac{S_d(n,k) \cdot (\tan\varepsilon - \tan\theta_S(n,k))}{\sqrt{2\tan^2\varepsilon + 2\tan^2\theta_S(n,k)}} \tag{6}$$

where ε is the half angle between the left and right channels. The left and right time-frequency stereo signals could be transferred back to the time domain using ISTFT similar to the mono approach. The mono or stereo downmix is then further compressed by AMR-WB+ [7] speech codec that supports both mono and stereo speech signals.

3.2. Source Separation from One Mixture

To separate the speech sources, the mono mixture is transferred back to the time-frequency domain signal $S'_d(n,k)$ by using the STFT (if no compression of the mixture is performed, $S'_d(n,k) = S_d(n,k)$. It should be noted that for the separation process of the frame by frame A-by-S encoder (decoding and separation block), this is conducted directly in the time-frequency domain. Based on the W-DO assumption in (1), the separation mask $M(n,k)$ is generated based on analysing the localization cue $\theta'_d(n,k)$.

For a stereo downmix, the same information is recovered by solving (5) and (6) for two unknowns, namely, $S'_d(n,k)$ and $\theta'_S(n,k)$, and reversing (4) for $\theta'_d(n,k)$. Suppose the azimuth for the m^{th} speech source is θ_m, the time-frequency mask for extracting the m^{th} source is given by:

$$M_m(n,k) = \begin{cases} 1, & \theta_d(n,k) = \theta_m \\ 0, & otherwise \end{cases} \tag{7}$$

The m^{th} speech source is extracted in time-frequency domain from the mixture signals by:

$$S'_m(n,k) = M_m(n,k) \cdot S_d(n,k) \ \forall n,k \quad (8)$$

After an inverse STFT, simultaneous speech sources from the mixture are recovered in the time domain.

3.3. Analysis-by-Synthesis Scheme

In Fig.2, the A-by-S encoder aims to preserve the quality of each individual speech source. The encoded speech source mixture using the method described in Section 3.1 will be separated back into individual sources using the approach discussed in Section 3.2 before transmission. The Energy Preservation Ratio Calculation module will evaluate the percentage of energy preserved between each separated source and the corresponding original source for the current frame. For speech source m in frame n, this ratio is:

$$P_m^{\ n} = \sum_{k=1}^{K} \left\| S'_m(n,k) \right\| \Big/ \sum_{k=1}^{K} \left\| S_m(n,k) \right\| \quad (9)$$

where $S'_m(n,k)$ and $S_m(n,k)$ is the separated and original speech source m, respectively. In this module, this ratio is evaluated to check if the energy for all active speech sources in the current frame is approximately equal in the mixture. The active source detection module will detect the active speech sources in the current frame, and is achieved using a Voice Activity Detector [8].

Final generation and compression of the time-domain mixes proceeds if the energy preservation ratio is approximately equal for every active source within the current frame. For example, if there are three sources, the aim is to ensure that the ratios of separated to original energies for each source are approximately equal. If the largest difference among the ratios is above a threshold, more time-frequency components from the sources with a lower energy preservation ratio will be included in the mixture and the energy preservation ratio is recalculated. Informal testing found a threshold of 5% difference in ratios to provide satisfactory decoded quality.

In the Encoding Mask Recalculation block, the active source separated from the current mixture with the lowest energy preservation ratio in the current frame is amplified by a factor; here, a factor of 1.05 (i.e. 5% increase in magnitude/iteration) gave a satisfactory trade-off between the number of iterations versus quality. Assuming the m_l^{th} active source is with the lowest energy preservation ratio, the amplified source $Sm_l^i(n,k)$ for the i^{th} A-by-S iteration and other active sources for recalculating the encoding mask during the i^{th} A-by-S iteration is given by:

$$S_{m_l}^{\ i}(n,k) = S_{m_l}^{\ i-1}(n,k) \cdot a \quad (10)$$

$$S_m^{\ i}(n,k) = S_m^{\ i-1}(n,k) \quad 1 \le m \le M,\ m \ne m_l \quad (11)$$

The encoding mask will be recalculated using (2), (3), and (7) by replacing the original signals in (2) with $S_{m_l}^i(n,k)$ and $S_m^i(n,k)$. The recalculated mask $M_m^i(n,k)$ for the m^{th} source in the i^{th} iteration will be returned to the Encoding block in Fig.2, where the updated speech source mixture $S_d^i(n,k)$ and its localization information $\theta_d^i(n,k)$ in the i^{th} iteration is given by:

$$S_d^{\ i}(n,k) = \sum_{m=1}^{M} (S_m(n,k) \cdot M_m^{\ i}(n,k)) \quad (12)$$

$$\theta_d^{\ i}(n,k) = \theta_m, \ if \ \ M_m^{\ i}(n,k) = 1 \quad (13)$$

The i^{th} iteration will be terminated once the energy is equally preserved in the mixture for each active source in the operating frame. In that case, the $S_d^i(n,k)$ will be transferred back to the time domain and transferred with $\theta_d^i(n,k)$ or creating a stereo mixture by the method described in Section 3.1. Otherwise, the next A-by-S loop will be operated until the above mentioned criterion is met.

4. SUBJECTIVE EVALUATION RESULTS

The proposed approach is verified through subjective listening tests using both anechoic and real meeting recordings. 20 Australian listeners participated in the test. The anechoic recordings are used to evaluate the performance of the proposed scheme when there are overlapping speakers in the time domain. The real meeting recordings are adopted to verify the proposed codec operating in real conference scenarios.

4.1. Evaluation for the Simultaneous Speakers

Eight sentences from The Australian National Database of Spoken Language [9] are chosen for the evaluation. Each test sample contains three overlapping speech sources in the time domain formed from randomly selections of these sentences. The encoded sentences are decoded and separated from the mixture and tested using the Mean Opinion Score (MOS) methodology [10]. The results are shown in Fig.3 with 95% confidence intervals.

Conditions 'Ori' and 'Ori_AMR12' are the original and degraded original signals compressed by AMR-WB+ at 12 kbps for each speech source. Conditions 'AByS_M' and 'AByS_S' are the separated signals decoded from mixtures containing three overlapping speech sources generated by the proposed A-by-S method using mono and stereo approaches, respectively. Conditions 'M' and 'S' are conditions 'AByS_M' and 'AByS_S' without using the A-by-S method described in Section 3.3. It should be noted that the mixtures in conditions 1.3 to 1.6 are further compressed using the AMR-WB+ codec at 36 and 48 kbps for the mono and stereo mode, respectively. It has been investigated in [12] that for the mono mode, the required compression rate for the localisation cues is 10 kbps, resulting in approximately 48 kbps for both modes in total. In addition, preliminary results indicate that two (10.1%) and three (3.9%) overlapping speakers are the most common patterns when analysing 210 minutes of the AMI-Corpus real meeting recordings [11] (86.0% for non-overlapping speaker cases). It is observed that the proposed A-by-S method achieves a MOS of approximately 3.85 for both mono and stereo modes compared with a MOS of approximately 3.23 for the non A-by-S coder, indicating the quality of the overlapping speech sources is maintained through applying the proposed analysis-by-synthesis approach.

4.2. Evaluation for the Real Recordings

Five real meeting recordings captured by close talking microphones from the AMI-Corpus [11] are employed in this evaluation. Each speech meeting zone contains four participants originally located at ±45°and ±135°. Overlapping sections from the recordings are purposely selected. A MOS test similar to Section 4.1 is firstly employed testing the separated single speech quality for real meeting recordings. Six speech sources selected from the five meeting recordings are used for the MOS test. Conditions 'AByS_M_R' and AByS_S_R are single speech sources separated from the compressed mixture using mono and stereo approach, respectively. Conditions 'Ori_R' and 'Ori_AMR9_R' are Original signal and a degraded version compressed by AMR-WB+ at 9 kbps. Results are presented in Fig. 4.

It can be observed in Fig.4 that when processing real recordings where two overlapping speeches are more common, the proposed system is achieving approximately 0.3 lower MOS score than the

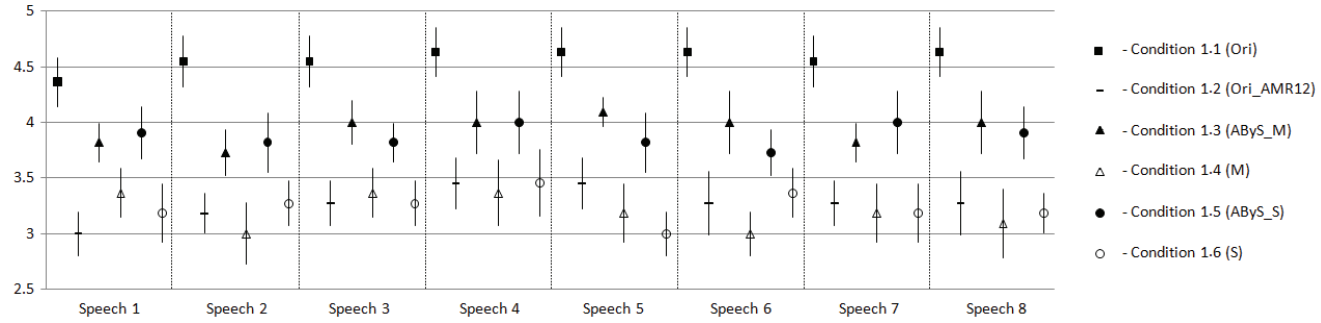

Fig.3. MOS Test Results for Anechoic Recordings

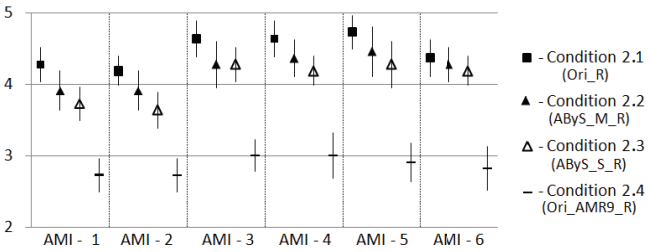

Fig.4. MOS Test Results for Real Recordings

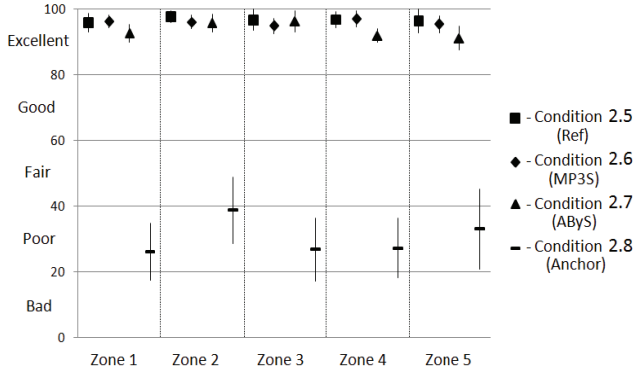

Fig.5. MUSHRA Test Results for Real Recordings

original signals, which is much higher when compressing the speech sources within a meeting zone separately using 9kbps of the AMR-WB+ (condition 2.4).

A MUSHRA [13] test is also adopted to test the spatialised navigation and reproduction using a standard 5.1 speaker array. Since the stereo approach performs similar to the mono approach in previous tests, only the mono approach is used for the MUSHRA test. Results are illustrated in Fig.5 using 95% confident intervals. Conditions 'Ref' and 'Anchor' are the original speech sources rendered at frontal soundfield from ±45°and ±135° to ±60°and ±20° and an unlocalised 3.5 kHz low-pass filtered anchor signal is created. Condition 'MP3S' is Condition 'Ref' compressed by the MP3-Surround Codec at 128 kbps. Condition 'AByS' are the original signals located at ±45°and ±135° compressed using the proposed mono approach, separated and rendered at ±60°and ±20° for each speech source.

It is demonstrated in Fig. 5 that the proposed approach is able to render the location of each source as demanded by the user and achieves approximately 90 on the MUSHRA scale when compared with the conditions 'Ref' and 'MP3S', which is rendered or compressed using the original recordings.

5. CONCLUSION

An analysis by synthesis coding architecture for the compression of navigable speech sources is presented. The proposed system has been evaluated through subjective tests using both anechoic and real meeting recordings. The results show excellent quality when compressing a spatialised speech zones using only 48 kbps. Future work will extend this work to support multi-zone speech compression and navigation using this method.

ACKNOWLEDGMENT

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

REFERENCES

[1] Elfitri, I.; Gunel, B.; Kondoz, A.M.; , "Multichannel Audio Coding Based on Analysis by Synthesis," *Proceedings of the IEEE*, vol.99, no.4, pp.657-670, April 2011.

[2] Quackenbush S., Herre, J., "MPEG Surround," *Multimedia, IEEE*, vol.12, no.4, pp. 18- 23, Oct.-Dec. 2005.

[3] B. Cheng, C. Ritz, I. Burnett, "Principles and Analysis of the Squeezing Approach to Low Bit Rate Spatial Audio," *Proceedings of the ICASSP*, Vol. 1, pp 13-16, Apr. 2007.

[4] Breebaart, Jeroen; Engdegård, Jonas, "Spatial Audio Object Coding (SAOC) - The Upcoming MPEG Standard on Parametric Object Based Audio Coding," *124 AES Convention*, 2008.

[5] Yilmaz, O.; Rickard, S.; "Blind separation of speech mixtures via time-frequency masking", *IEEE Transactions on Signal Processing*, Volume 52, Issue 7, Pages 1830-1847, July 2004.

[6] Del Galdo, G.; Kuech, F.; Kallinger, M.; Schultz-Amling, R.; "Efficient merging of multiple audio streams for spatial sound reproduction in Directional Audio Coding", *ICASSP 2009,* pp.265-268, 2009.

[7] 3GPP Specification series, TS 26.290., "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions", 2009.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters.*, Vol. 6 ,Issue 1: 1–3, 1999.

[9] Millar, J.B., Vonwiller, J.P., Harrington, J.M., and Dermody, P.J., "The Australian National Database of Spoken Language", *Proceedings of the ICASSP*, vol.1, pages 97 - 100, 1994.

[10] ITU-R Recommendation P.800, "Methods for subjective determination of transmission quality", 1996.

[11] Jean Carletta, Simone Ashby, "The AMI Meeting Corpus: A Pre-announcement", *Machine Learning for Multimodal Interaction II*, pages 28-39, 2006.

[12] Cheng, B.; Ritz, C.H.; Burnett, I.S.; "Psychoacoustic-based quantization of spatial audio cues", *Electronics Letters*, Vol. 44, Issue 18, pp 1098-1099, Aug. 2008.

[13] ITU-R Recommendation BS.1534, "Methods for the subjective assessment of intermediate quality levels of coding systems", 1997.