2013

Encoding navigable speech sources: a psychoacoustic-based analysis-by-synthesis approach

Xiguang Zheng
University of Wollongong, xz725@uowmail.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

Jiangtao Xi
University of Wollongong, jiangtao@uow.edu.au

Publication Details
Encoding navigable speech sources: a psychoacoustic-based analysis-by-synthesis approach

Abstract
This paper presents a psychoacoustic-based analysis-by-synthesis approach for compressing navigable speech sources. The approach targets multi-party teleconferencing applications, where selective reproduction of individual speech sources is desired. Based on exploiting sparsity of speech in the perceptual time-frequency domain, multiple speech signals are encoded into one mono mixture signal, which can be further compressed using a standard speech codec. Using side information indicating the active speech source for each time frequency instant enables flexible decoding and reproduction. Objective results highlight the importance of considering perception when exploiting the sparse nature of speech in the time-frequency domain. Results show that this sparsity, as measured by the preserved energy level of perceptually important time-frequency components extracted from mixtures of speech signals, is similar in both anechoic and reverberant environments. The proposed approach is applied to a series of simulated and real reverberant speech recordings, where the resulting speech mixtures are compressed using a standard speech codec operating at 32 kbps. The perceptual quality, as judged both by objective and subjective evaluations, outperforms a simple sparsity approach that does not consider perception as well as the approach that encodes each source separately. While the perceptual quality of individual speech sources is maintained, subjective tests also confirm the approach maintains the perceptual quality of the spatialized speech scene. © 2012 IEEE.

Keywords
synthesis, analysis, psychoacoustic, sources, approach, speech, encoding, navigable

Disciplines
Engineering | Science and Technology Studies

Publication Details
Encoding Navigable Speech Sources: A Psychoacoustic-based Analysis-By-Synthesis Approach

Xiguang Zheng, Christian Ritz, Senior Member, IEEE, and Jiangtao Xi, Senior Member, IEEE

Abstract — This paper presents a psychoacoustic-based analysis-by-synthesis approach for compressing navigable speech sources. The approach targets multi-party teleconferencing applications, where selective reproduction of individual speech sources is desired. Based on exploiting sparsity of speech in the perceptual time-frequency domain, multiple speech signals are encoded into one mono mixture signal, which can be further compressed using a standard speech codec. Using side information indicating the active speech source for each time frequency instant enables flexible decoding and reproduction. Objective results highlight the importance of considering perception when exploiting the sparse nature of speech in the time-frequency domain. Results show that this sparsity, as measured by the preserved energy level of perceptually important time-frequency components extracted from mixtures of speech signals, is similar in both anechoic and reverberant environments. The proposed approach is applied to a series of simulated and real reverberant speech recordings, where the resulting speech mixtures are compressed using a standard speech codec operating at 32kbps. The perceptual quality, as judged both by objective and subjective evaluations, outperforms a simple sparsity approach that does not consider perception as well as the approach that encodes each source separately. While the perceptual quality of individual speech sources is maintained, subjective tests also confirm the approach maintains the perceptual quality of the spatialized speech scene.

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

I. INTRODUCTION

The compression and reproduction of multichannel speech is becoming increasingly common for applications such as multiparty spatialized teleconferencing. Such applications provide a natural communication experience for geographical dislocated participants, whilst also allowing for improved multi-tasking [1] where the listeners are requested to participate or monitor the teleconference while performing a current task. It has been suggested in [1] and [2] that allowing the users to personalize the reproduced sound scene (i.e., place the sound sources at the desired positions) will enhance the multitasking efficiency when participating or monitoring the teleconference. To ensure high perceptual quality, separate encodings of the individual speech sources is most commonly adopted, where the required bitrates increases linearly as more participants become involved. For bandwidth constrained environments, an efficient compression technique supporting personal sound scene rendering of simultaneous speech sources without requiring this linear increase in bit rate is required.

Assuming multiple participants at each site, one approach for the efficient compression of the multiparty speech scene is to apply a spatial audio coder at each site. Spatial audio coders including MPEG-Surround (MPS) [3–5], and Spatial Squeezed Surround Audio Coding (S3AC) [6], [7], have focused on the compression and reproduction of the original multichannel audio signals by applying spatial psychoacoustic models in the time-frequency domain. However, this approach has two problems. Firstly, each site will have a compressed spatial audio scene (represented as compressed downmix signals) and hence as the number of sites increases, so does the total bit rate. Secondly, due to their reliance on spatial perception, existing spatial audio coders are designed to allow perceptually accurate reproduction of an entire audio scene. However, the quality of reproduction of part of the audio scene (e.g., the speech from one participant) cannot be guaranteed. A further problem is that in a multiparty teleconferencing application, each site must merge and reproduce multiple audio scenes (one for each remote site), which may be difficult without accurate representation of each audio object. In addition, the sound scene signals compressed by the spatial audio coders are in general less flexible for personalized sound scene rendering (i.e., by receiving the exact same compressed signals, different users cannot render the talker at any preferred spatial location in the reproduction site), which is also not desired especially for the teleconferencing with multitasking applications.

Another approach for compressing the soundfield is Directional Audio Coding (DirAC) [8], [9], which focuses on the compression of the B-format soundfield microphone recordings [10] of a sound scene. However, a disadvantage of this approach is that it relies on B-format recordings of a single sound scene, which may not be available in practice and is not directly applicable to a set of mono audio object signals that may have been extracted from more than one sound scene (e.g., recordings at multiple participant sites).

Recently, Spatial Audio Object Coding (SAOC) [11], [12] was proposed to operate within the MPEG-Surround Codec such that each audio object may be interactively chosen by a listener, which is more suitable for speech teleconferencing and browsing applications as targeted in this paper. However, the MPEG surround codec was designed for multichannel audio compression and relies on spatial sound perception derived...
for multiple loudspeaker signals. Hence, this approach is not optimized for speech as the primary input signal or a set of mono signals representing real sources. More recently, Informed Source Separation (ISS) [13], [14] has been proposed as an alternative solution to compress the audio objects. Specifically in [15], ISS has been adopted for compressing speech sources by selecting two dominant time-frequency instants per time-frequency bin. It should be noted that the aim of the proposed framework in this paper is to achieve the same goal by using only one dominant time-frequency instant per time-frequency bin.

The spatial audio coders discussed so far can be classified as open loop multichannel audio coding techniques, where the spatial parameters are extracted from the input signals without considering the impact on the decoding quality. More recently, an analysis-by-synthesis framework for spatial audio coding was proposed [16], where coding parameters are chosen based on error minimization between the original and decoded multichannel audio signals. Results showed improved objective quality of the decoded signals compared to the existing MPEG Surround approach.

An Analysis-By-Synthesis (ABS) framework for the compression of simultaneous speech sources was recently proposed in [17], which is based on the exploration of the speech orthogonality in the time-frequency domain and was shown to provide good subjective results for anechoic overlapping speech sentences. Here, this approach is significantly extended by proposing a Psychoacoustic-based Analysis-By-Synthesis (PABS) coding scheme designated for the compression and navigation of multiple practical (reverberant) speech objects. A thorough investigation into speech sparsity within several practical (reverberant) environments is presented prior to describing the proposed PABS framework. In contrast to [16] where the MPEG-Surround coder is employed, the proposed approach is based on the S³AC framework [6], [7], which was previously investigated for the compression of multichannel audio signals. Compared to other spatial audio coding approaches, S³AC preserves the audio scene by generating a virtual audio object for each time-frequency bin based on the psychoacoustic assumption that for one frequency bin, one can only perceive one audio object. Although this approach succeeds to preserve the entire audio scene, when there are simultaneously occurring sources, the virtual audio object cannot be used to recover each real audio object. In this paper, this approach is redesigned based on the sparse property of speech and used within a psychoacoustic-based analysis-by-synthesis framework for the compression of simultaneously occurring speech signals, such as in realistic multiparty meeting scenarios. Significantly higher perceptual quality of the individual sources when compared to a non-ABS approach or separate encoding of each source is achieved.

The proposed framework relies on exploiting the sparse property of speech in the time-frequency domain to maintain the individual quality of speech objects in the receiver ends. This property has been successfully analyzed and employed for Blind Source Separation (BSS) of anechoic speech [18], [19]. In this work, this property is firstly analyzed for reverberant speech signals and then used to create a separable mono mixture signal in practical (reverberant) conditions. This is achieved using an Analysis-By-Synthesis (ABS) optimization iteration, where musical distortion (caused by missing perceptually important frequency components) can be significantly reduced by applying a psychoacoustic motivated approach as successfully exploiting in speech enhancement postfilters [20]. The approach is designed so that the mixture signal can be compressed for transmission using a standard speech coder (in this paper, the AMR-WB+ coder [28] is chosen). At the reproduction site, the compressed 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 [21], 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 II investigates the time-frequency sparsity of simultaneous speech signals in several reverberant conditions. Section III presents the architecture of the proposed framework. Experimental results are presented and discussed in Section IV, while conclusions are drawn in Section V.

II. EXPLORING SPEECH SPARSITY IN REVERBERANT ENVIRONMENTS

Speech signals are known to be sparse in the short-term time-frequency domain. This property has been successfully used for BSS [19]. Before employing this property to compress simultaneous real speech recordings, it is necessary to validate if this property still holds for real recording environments. While the sparse property has been well examined in idealized anechoic environments, it will be examined using reverberant recordings in this section.

A. Speech Sparsity in Anechoic Environments

Speech sparsity in the short-term time-frequency domain means that a small percentage of the time-frequency coefficients of speech capture a large percentage of the overall energy. This is defined by the W-Disjoint Orthogonality [18] as:

Two signals \( s_j \) and \( s_2 \) are W-Disjoint Orthogonal (W-DO) if the corresponding time-frequency component of the Short Time Fourier Transform (STFT) of \( s_j \) and \( s_2 \) are disjoint for a given window function. Mathematically, it is described by

\[
S_i(n,k) \cdot S_j(n,k) = 0, \ \forall n,k \quad i \neq j
\]  

where \( S(n,k) \) and \( S(n,k) \) is the time-frequency representation of signal \( s_i \) and \( s_j \), respectively, \( n \) is the frame number and \( k \) is the frequency index.

It has been shown in [19] for anechoic recordings that the W-Disjoint Orthogonality can be approximately satisfied. By applying a binary time-frequency separation mask, the resulting separated speech sources from mixtures containing 2 to 5 simultaneous speeches could achieve from ‘perfect’ to ‘minor artifacts or interferences’ in the perceptual test scores and re-
cover from 94% to 79% of the energy, respectively, compared to the original clean speeches.

B. Exploring Speech Sparsity in Reverberant Environments

In this section, the speech sparse property will be examined in several simulated reverberant environments. The aim of this investigation is to analyze this property for practical applications where reverberation occurs. Since the reverberation can be in general considered as the time-delayed replication of the direct (original) signal, the orthogonality for simultaneous sources is not guaranteed to be as good as in anechoic (ideal) environments, which has been illustrated in Fig. 1. The reverberation of the soundfield increases the spread of energy in the time domain for each speech source, and it can be observed in Fig. 1 that the speech sparsity is weakened.

Hence, it is important to analyze the degradation of the orthogonal property caused by reverberation. It should be noted that the reverberation is considered as the major potential disturbance of the orthogonality rather than background noise in that it is the reflection of the direct sound and thus more correlated to the direct sound.

The reverberation of audio signal can be characterized by the Sabine’s Reverberation equation ($RT_{60}$) of a room [22] as

$$RT_{60} = 0.163 \cdot \frac{V}{\alpha A}$$  \hspace{1cm} (2)

where $RT_{60}$ time is the time required for reflections of a direct sound to decay by 60dB such that it could not be perceived by the human listening system, $V$ is the volume of the room, $A$ is the total surface of the room, $\alpha$ is the average absorption coefficient of the surfaces.

For the analysis in this paper, 36 Sentences (20000Hz) selected from the Australian National Database of Spoken Language [23] containing 36 different Australian native speakers of different ages and genders are selected as the testing database. Each sentence is overlapped with other $M$-1 (2 ≤ $M$ ≤ 8) sentences in the time domain resulting in $M$ overlapping speech conditions. For the $M=2$ overlapping case, each sentence was overlapped with each of the remaining 35 sentences resulting 36x35 = 1260 combinations. For other overlapping cases, each sentence is randomly overlapped 35 times with $M$-1 sentences resulting 1260 combinations to equalize the total number of combination for different overlapping conditions. Before overlapping, different artificial reverberation conditions will be added for each sentence. Two parameters, namely, reverberation time ($T$) and direct to reverberant ratio ($R$) [24], [25] are introduced to objectively control the level of reverberant. Reverberation time represents the volume of the environment. Larger reverberation time implies a larger room. Direct to reverberant ratio indicates the absorbance level. Lower direct to reverberant ratio results lower absorbance level.

The orthogonal analysis is taken by only preserving one dominant time-frequency instant among $M$ (2 ≤ $M$ ≤ 8) simultaneous sources. After performing the Short Time Fourier Transform for each source, the dominant time-frequency instant $S_d(n,k)$ can be obtained by:

$$S_d(n,k) = \max_m(S_m(n,k)), m \in [1,M]$$  \hspace{1cm} (3)

where $n$ (1 ≤ $n$ ≤ $N$) and $k$ (1 ≤ $k$ ≤ $K$) are frame number and frequency index, respectively. Thus, for the $m^{th}$ source $S_m(n,k)$, after the orthogonal analysis described by (3), it is consisted only with the dominant time-frequency instants $S'_m(n,k)$:

$$S'_m(n,k) = \begin{cases} S_m(n,k), \text{if } m = \arg\max_m(S_m(n,k)) \\ 0, \text{otherwise} \end{cases}$$  \hspace{1cm} (4)

Instead of analyzing the Preserved-Signal Ratio (PSR) suggested in [19], which represents the energy preserved after the orthogonal analysis against the original signal, here, it is more accurate to consider the perceptual importance of each time-frequency instant. For instance, after the orthogonal analysis, the dominant time-frequency instants may not be the most important ones in the perceptual domain. In this section, the Perceptual Frame Energy Preservation Ratio (PFEPGR) and Perceptual Global Energy Preservation Ratio (PGEPR) are employed. The PFEPR aims to analyze the amount of perceptual energy kept for each frame and PGEPR is designed to analyze the amount of perceptual energy kept for each speech source. The PFEPR for the $n^{th}$ frame of source $m$ is given by:
where \( A_m(n,k) \) is the perceptual time-frequency weighting function of the \( m^{th} \) source and is equivalent to the inverse of the perceptual masking threshold energy [27].

The \( PGEPR \) for source \( m \) is given by:

\[
PGEPR_m (S_m(n)) = \frac{\sum_{k=1}^{K} [A_m(n,k) \cdot \|S_m(n,k)\|]}{\sum_{n=1}^{N} \sum_{k=1}^{K} [A_m(n,k) \cdot \|S_m(n,k)\|]} \tag{6}
\]

Fig.2. Perceptual Frame Energy Preservation Ratio (\( PFEPR \)) is analyzed from various combinations of simultaneous source numbers, reverberation times and intensities with “Anechoic” results as reference. For example, condition “0.5s - 20dB” in catalogue 3 means the \( PFEPR \) is obtained by analysing three simultaneous sources which are pre-processed using reverberant parameter \( T = 0.5s \) and \( R = 20dB \). Each bar represents the mean value analyzed from the abovementioned 1260 random overlapping combinations. The error bars are generated from the standard derivation of corresponding samples. More precisely, the width of each error are twice of the corresponding standard derivation value.

Fig.3. Perceptual Global Energy Preservation Ratio (\( PGEPR \)) is analyzed from various combinations of simultaneous source numbers, reverberation times and intensities with “Anechoic” results as reference. Legends and methodologies are similar to Fig. 2.

Statistic results are shown in Fig. 2 and Fig. 3 for Perceptual Frame and Global \( EPRs \), respectively. It can be observed that the orthogonal property degrades as the number of simultaneous sources increases. The variance (here this is reflected by the standard derivation) of \( PFEPR \) is larger than \( PGEPR \), while the average of \( PFEPR \) is lower than \( PGEPR \). This indicates that the quality of overlapping sources can be varied between frames when performing the orthogonal analysis, even though the \( PGEPR \) is in a satisfactory value with much less variance. For instance, even for the two overlapping anechoic case, the standard derivation for \( PFEPR \) is much higher (approximately 25%) than \( PGEPR \) (approximately 5%), while the average of \( PGEPR \) is roughly 85%. The results also show both reverberation time and direct to reverberant ratio can weaken
the speech sparse property by up to approximately 12%. It should be noted that for real environments, generally the larger environment will result in a lower reverberant ratio and vice versa. For instance, a typical small room \( T = 0.5s, R = 10\text{dB} \) or a concert hall \( T = 2s, R = 20\text{dB} \) did not drastically weaken the orthogonal property when compared to the anechoic reference. Hence, here it can be concluded that the sparse property does not significantly weaken in the reverberant environment. However, due to significant variance of the PFEP, the quality of recovered sources could vary when blindly applying the

![Image](Fig.4. Proposed Perceptual Analysis by Synthesis Framework)

sparse property of speech for encoding without considering the psychoacoustic and frame-by-frame factors. Therefore, it is important to intelligently choose the perceptually important time-frequency instant as well as considering the frame variances, which is the motivation of the proposed psychoacoustic based analysis-by-synthesis compression scheme.

III. PSYCHOACOUSTIC-BASED ANALYSIS-BY-SYNTHESIS COMPRESSION FRAMEWORK

As illustrated in Fig.4, the perceptual psychoacoustic-based analysis-by-synthesis (PABS) compression framework encodes multi-channel speech signals from multiple sites into a mono mixture stream (with side information) that can be further compressed using standard speech codec such as AMR-WB+ [28]. 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 activate a speech source (or sources) and move them to desired positions in the reproduced audio scene. Here it is assumed that the speech sources are obtained from close talking microphones or accurately derived from microphone array recordings using BSS techniques. While the speech sources may not achieve perfect separation, the objective of this work is to investigate the perceptually optimized coding performance assuming perfect separation. The detailed framework of the proposed framework is illustrated in Fig.5 and is described further below.

A. Perceptual Orthogonal Analysis

Input mono signals from speaker 1 to speaker M (as shown in Fig.4), transformed into the time-frequency domain using Short Time Fourier Transform (STFT) [27], are denoted by \( S_m(n,k) \) where \( 1 \leq m \leq M \) and \( n \) and \( k \) are frame number and frequency index, respectively. The time-frequency domain signals \( S_m(n,k) \) are transformed into the perceptual time-frequency domain \( S_m^w(n,k) \) by using the psychoacoustic model as described in [27]:

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

where \( A_m(n,k) \) is the perceptual time-frequency weighting function of the \( m^{th} \) source equals to the inverse of the perceptual masking threshold energy.

In the Orthogonal Analysis block of Fig.5, under the assumption of the sparsity of speech, the perceptual dominant source \( S_m^w(n,k) \) with maximum energy corresponding to the perceptual dominant speaker at this time-frequency bin can be obtained by:

\[
S_d^w(n,k) = \max_m (S_m^w(n,k)), m \in [1,M]
\]

If \( m_d \) denotes the speech source of the corresponding perceptual dominant speaker, an encoding mask is needed here to indicate the origin of the perceptually dominant time-frequency instants and is given by:

\[
M_d(n,k) = m_d
\]

It should be noted that the spatial information, which can be obtained through processing microphone array recordings of the sources, can be an alternative to the encoding mask where the origin of the dominant source is represented by the spatial locations of the source. For the rest of the paper, the more complicated case (speech sources with their spatial locations) will be discussed. Suppose the spatial location of the \( m^{th} \) source is \( \theta_m(n,k) \), the azimuth corresponding to \( S_d^w(n,k) \) is given by:
\[ \theta_d(n,k) = \theta_{m_d}(n,k) \]  

(10)

\( \theta_{m_d}(n,k) \) will be sent to the Frame Energy Preservation Ratio Calculation block as the current encoding mask candidate for evaluation. After the orthogonal analysis, the sources consisting of the perceptual dominant time-frequency instants only are denoted by \( S'_{m}^{w}(n,k) \), which is given by:

\[
S'_{m}^{w}(n,k) = \begin{cases} 
    S_m^{w}(n,k), & \text{if } m = \arg \max_{m} \left( \text{max}(S_m^{w}(n,k)) \right) \\
    0, & \text{otherwise} 
\end{cases} 
\]  

(11)

\( S_m^{w}(n,k) \) will be sent to the Frame Energy Preservation Ratio Calculation block as well to analyze with \( S_m^{w}(n,k) \) as described in Fig. 5. Note that \( S_m^{w}(n,k) \) is the \( i^{th} \) \( S_m^{w}(n,k) \) with \( i \) (the analysis-by-synthesis iteration number) equal to 0.

B. Psychoacoustic Analysis-By-Synthesis Framework

Due to the approximated sparse property of speech, in reality, there are time-frequency components overlapped between speech sources especially for reverberant recordings. As discussed in Section I, the variance for the PFEPR is significant which could result in unbalanced perceptual quality of simultaneous sources when applying the orthogonal analysis. The proposed ABS framework aims to provide a solution to maintain the quality of individual speech sources when time-frequency overlapping happens. As shown in Fig. 5, the Frame Energy Preservation Ratio Calculation block takes \( S_m^{w}(n,k) \) from the Orthogonal Analysis block. The PFEPR as described in (5) will be evaluated for the percentage of energy preserved between \( S_m^{w}(n,k) \) and \( S_m^{w}(n,k) \) for the current frame. Recall (5), for speech source \( m \) in frame \( n \), this ratio is:

\[
\text{PFEPR}_m^n \left( S_m^{w}(n) \right) = \frac{\sum_{k=1}^{K} \| S_m^{w}(n,k) \|}{\sum_{k=1}^{K} \| S_m^{w}(n,k) \|} 
\]  

(12)

It should be noted that this measurement is related to the Log Spectral Distortion (LSD) measurements [29], which is widely used in speech coders for objectively estimating perceptual speech quality. A higher PFEPR of a speech source indicates more frequency instants are included in the mixture, resulting in a lower LSD. In this module, the PFEPR 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 (VAD) [30] (i.e. if the current frame is active, the Voice Activity Detector will return 1).

Final speech mixture generation and compression of the time-domain mixes proceeds (Flag = 1 as shown in Fig. 5) if the PFEPR 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 PFEPR 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 PFEPR will be included in \( S_m^{w}(n,k) \) and the energy preservation ratio is recalculated (Flag = 0). Informal testing found a threshold of 5% difference in ratios to provide satisfactory decoded quality.

In the PFEPR Equalization block, the active source with the lowest PFEPR in the current frame is amplified by an energy boosting factor. Assuming the \( m_i \)th active source is with the lowest PFEPR, the amplified source \( i+1S_{m_i}^{w} \) and other active sources for the \( i+1 \)th ABS iteration is given by:

\[
i+1S_{m_i}^{w}(n,k) = iS_{m_i}^{w}(n,k) \cdot a 
\]  

(13)

\[
i+1S_{m}^{w}(n,k) = iS_{m}^{w}(n,k) \quad 1 \leq m \leq M, m \neq m_i 
\]  

(14)

where \( a \) is an energy boosting factor. In here, \( a = 1.01 \) (i.e. in each iteration, the active source with the lowest PFEPR will be magnified by 1% to generate the updated source index as described in Fig. 5.) The selection of \( a \) determines the accuracy of the analysis-by-synthesis operation. In order to have a maxi-
minimum 5% of the PFEPR difference, preliminary experiments indicated \( a = 1.01 \) will suit for this accuracy requirement. It should be noted that while the analysis-by-synthesis loop leads to increased complexity, factor \( a \) of (13) could also be varied to minimize computational complexity for a given implementation platform. \({\hat{S}}_{m}^{w}(n,k) (1 \leq m \leq M)\) will be sent back to the Orthogonal Analysis block to obtain \({\hat{s}}_{m}^{w}(n,k)\) for the \( i+1 \)th iteration. It should be noted that the PFEPRs for each iteration are generated based on the non-amplified version of \( S_{m}^{w}(n,k) \) (i.e. if the first source \((S_{1}^{w}(n,k))\) is the selected source for the current time-frequency of the \( i \)th iteration, then \( S_{1}^{w}(n,k) \) in (12) equals \( S_{i}^{w}(n,k) \)). The \( i \)th iteration will be terminated once the perceptual energy is equally preserved for each active source in the operating

<table>
<thead>
<tr>
<th>Table I</th>
<th>Statistical Results for Simultaneous Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Active Sources in each Frame ((m))</td>
<td>Percentage among all Testing Meeting Recordings ((\mu))</td>
</tr>
<tr>
<td>1</td>
<td>81.3%</td>
</tr>
<tr>
<td>2</td>
<td>15.2%</td>
</tr>
<tr>
<td>3</td>
<td>3.5%</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Fig. 6. Multiparty Navigable Spatial Teleconferencing Framework

frame. In that case, the candidate encoding mask will be sent to the Speech Mixture Generation block to produce the mono mixture described by the next Section. Otherwise, the next ABS loop will be operated until the above mentioned criterion is met.

C. Speech Mixture Generation

When the criterion for the PFEPRs is met, the candidate encoding mask will be sent to the Speech Mixture Generation block. Suppose the ABS loop terminates at the \( i \)th iteration, for the more complicated case where the spatial location of the sources are available, \( \theta_{i}(n,k) \) is transmitted to the Speech Mixture Generation block instead of \( \theta_{m}(n,k) \). The encoding mask for the \( m \)th source, \( M_{m}(n,k) \) can be obtained from \( \theta_{m}(n,k) \) by:

\[
M_{m}(n,k) = \begin{cases} 1, & \theta_{m}(n,k) = \theta_{m} \\ 0, & \text{otherwise} \end{cases}
\] (15)

Hence, \( S_{d}(n,k) \) can be obtained by:

\[
iS_{d}(n,k) = \sum_{m=1}^{M} iM_{m}(n,k) \cdot S_{m}(n,k) \forall n,k
\] (16)

\( S_{d}(n,k) \) will be transferred back to time domain using an Inverse Short Time Fourier Transform (ISTFT) [27] to create a mono source mixture. The azimuth \( \theta_{m}(n,k) \) is preserved in the time-frequency domain and transmitted as side information for decoding. The mono downmix is then further compressed by the AMR-WB+ [28] speech codec.

D. Speech Separation and Navigation

The technique to separate the speech sources from the mixture is partly described in (15) and (16). To separate the \( m \)th speech source from \( S_{d}(n,k) \) and \( \theta_{m}(n,k) \), the separation mask \( M_{m}(n,k) \) can be generated by applying (15) as:

\[
M_{m}(n,k) = \begin{cases} 1, & \theta_{m}(n,k) = \theta_{m} \\ 0, & \text{otherwise} \end{cases}
\] (17)

Hence, the decoded \( m \)th speech source \( S_{m \text{-dec}}(n,k) \) is

\[
S_{m \text{-dec}}(n,k) = M_{m}(n,k) \cdot \hat{S}_{d}(n,k) \forall n,k
\] (18)

Since the sources even when overlapped in the time domain can be separated from the mono mixture, the joint compression of multiple sources can be achieved. The advantage of the joint compression rather than separated compression of each source is that the required transmission bandwidth will not be increased as the number of the participants increases. The interactive navigation as described in Fig. 4 can be achieved by sending the same stream to different users, where sources can be selected and rendered at desired spatial locations. Another feature of the proposed framework is that there is no limitation in terms of reproduction techniques since the “dry” sources are available at the reproduction site.

E. Multiparty Spatialized Teleconferencing Framework

As depicted in Fig. 6, a multiparty spatialized teleconferencing framework can be built based on the proposed PABS scheme. For site 2 to \( M \), the speech sources (i.e. obtained by close-talk microphone recordings for individual sources) can be firstly compressed by the Single-site PABS encoders. The single-site mixture is transmitted to the server where the sources from each site 2 to \( M \) will be separated by the proposed technique. The sources will be again encoded into one stream and transmitted to the listeners. By receiving the same stream, the listeners at each site can selectively reproduce individual speech sources to, e.g. create a personalized spatial sound scene of a teleconference or selectively reproduce one or more speakers.
IV. OBJECTIVE AND SUBJECTIVE EVALUATION FOR PRACTICAL APPLICATIONS

Both objective and subjective evaluation results are presented in this section. The evaluation conditions focus on the application of the proposed framework in practical scenarios. Statistical results regarding the number of simultaneous speakers for real meeting recordings are firstly presented to validate the most common number of overlapping speech sources. This is followed by objective results analyzing the sparsity of reverberant speech in the perceptual time-frequency domain. Subjective listening tests are then presented to analyze the perceptual quality of the individually reproduced speech sources and reproduced sound scenes containing multiple speech sources.

Fig.7. Perceptual Frame Energy Preservation Ratio between Simple Orthogonal Approach and Proposed Perceptual Frame Energy Preservation Ratio for all the reverberant Conditions in Section II. Condition Ortho-M (2≤M≤4) is the separated speech source generated by applying the simple orthogonal approach (i.e. using (3) and (4)). Condition PABS-M (2≤M≤4) is the separated speech source generated by applying the proposed PABS framework.

Fig.8. The Perceptual Frame Energy Preservation Ratio Difference between Simple Orthogonal Approach and Proposed Perceptual Frame Energy Preservation Ratio for all the reverberant Conditions in Section II.

A. Statistical Analysis of Simultaneous Speakers in Real Meeting Recordings

Before the evaluation of the proposed framework, it is necessary to understand how often simultaneous speakers occur in real meetings. Real meeting recordings are collected from the AMI Corpus [31], which is a meeting dataset consisting of four participants. A total of 210 minutes of the meeting recording in this dataset are analyzed using the voice activity detection technique described in Section III-B to find the number of simultaneous sources \( m \) (1≤m≤4) in each frame. For instance, for the \( n^{th} \) frame, if there are three active frames returned by the VAD, \( m_n \) is three. The results are shown in Table I in terms of the percentage (p) for each number of simultaneously active speakers. From Table I, 1.15% of the frames contain two simultaneously active speakers while only 3.5% of frames contain three simultaneously active speakers.
B. Objective Evaluation

The same evaluation database as used in Section I-B is employed in this section to perform the objective evaluation. The aim of this analysis is to examine the Perceptual Frame Energy Preservation Ratio (generated by (5)) for the simple orthogonal approach (i.e. only using (3) and (4) to obtain the dominant sources and separate the speech sources) and the proposed perceptual analysis-by-synthesis framework (as described in Section III). The same evaluation method as used in Section II-B is employed here. 2 ≤ M ≤ 4 simultaneous speaker cases are evaluated in this section based on the practical meeting statistics presented in Section IV-A. The PFEP results are presented in Fig. 7. Condition Ortho-M (2≤M≤4) is the separated speech source generated by applying the simple orthogonal approach (i.e. using (3) and (4)). Condition PABS-M (2≤M≤4)

![Diagram](image)

is the separated speech source generated by applying the proposed PABS framework. It can be observed from Fig. 7 that by applying the proposed framework, the averaged PFEPs are higher and the variances for PFEPs are lower for all conditions compared to the Orthogonal Approaches. It should be noted that the variances of the PFEP for the PABS conditions are caused by the variances of the orthogonality between each frame.

In addition, to evaluate the proposed framework, the maximum PFEP difference (PFEP\text{diff}) is used to examine the maximum energy difference in each frame between the simultaneous sources, which is given by (for the \(n\)th frame, 2≤M≤4):

\[
PFEP\text{diff}_n = \max_{m}(PFEP\text{m}_n) - \min_{m}(PFEP\text{m}_n), m \in [1, M]
\] (19)

The maximum PFEP differences for the same conditions in Fig. 7 are presented in Fig. 8. It can be observed in Fig. 8 that the PFEP\text{diff} for the proposed framework is below 5% as aimed in Section III, where for the simple orthogonal approach, the average of this difference is as large as 60%.

C. Subjective Mean Opinion Score Test

In this test, a meeting scenario is considered. Both artificially created reverberant speeches from The Australian National Database of Spoken Language [23] and real meeting recordings from AMI-Corpus [31] are chosen for the evaluation. The test simulates a meeting scene with four participants. According to Table I, three speakers may be overlapped during the meeting. Therefore, eight sentences from the Australian National Database of Spoken Language are chosen to randomly overlap with another three sentences within the database where three out of four speech signals are overlapped in the time domain. Before overlapping, each test sample is simulated to imitate general room and concert hall recordings (parameters for the simulation are given in Section II) resulting in two sessions for room and concert hall conditions. In addition, the third session is consisted by practical meeting recordings from the AMI-Corpus [31], where four speakers are in a meeting. Eight simultaneous talker sections in the AMI Corpus are de-
luminally selected. Mean Opinion Score (MOS) methodology [32] is employed for the test with four conditions, namely, Original, PABS, Ortho and AW+8. Condition Original is the un-encoded original recording. Condition PABS is the target speech source encoded with other randomly selected sentences using the proposed PABS framework and separated back from the mixture. Condition Ortho is the target speech source encoded with other randomly selected sentences using simple orthogonal approach and separated back from the mixture. The mixtures in Condition PABS and Ortho are further compressed using the AMR-WB+ codec [28] at 32 kbps. Note that the side information can be compressed at approximately 2kbps using [33]. Condition AW+8 is the separate compression of original recordings using AMR-WB+ codec [28] at 8 kbps and thus occupying the same (8×4 = 32 kbps) total bitrate compared to Condition PABS and Ortho. It should be noted that the use of Voice Activity Detection (VAD)/Discontinuous Transmission (DTX) modes could further reduce the average bit rate required and hence the bit rates reported here represent the upper limit of total bit rate required. A total of 20 Listeners participated in the test. Results are shown in Fig.9 with 95% confidence intervals.

It can be observed from Fig. 9 that for all sections, the proposed PABS framework achieved a higher MOS score (on average around 4.0) than both Condition Ortho and AW+8 with clear statistical significant differences. It should be noted that some of the original speeches are not rated as high as 4.5, which is caused by the introduction of reverberation. Nevertheless, the MOS for the proposed PABS framework achieved a similar score compared to the original speeches with no statistical significant difference.

D. MUSHRA Test using Real Meeting Recordings

A MUSHRA [34] test is also adopted to measure the quality of spatialized versions of the encoded mixtures of AMI-Corpus Meeting Recordings [31] using a standard 5.1 speaker array for playback. Conditions ‘Original’ is the original speech sources rendered at frontal soundfield from their original location (±45°and ±135°) to the simulated desired locations (±60°and ±20°, i.e. the possible desired locations of the listener), which is the best achievable representation of the conversation and is served as the Hidden Reference of this MUSHRA test. Condition Anchor is an unlocalized 3.5 kHz low-pass filtered anchor signal. Condition ‘MP3S’ is Condition ‘Ref’ compressed by the MP3-Surround Codec at 128 kbps. Condition PABS is the original signal located at ±45°and ±135° compressed using the proposed framework, separated and rendered at ±60°and ±20° for each speech source. A total of 15 Listeners participated in the test. Results are shown in Fig.10 with 95% confidence intervals.

It is demonstrated in Fig. 10 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 Original and MP3S, which is rendered or compressed using the original recordings.

V. CONCLUSION

A psychoacoustic-based analysis by synthesis framework for the compression of navigable speech sources is presented. The proposed system has been evaluated through both objective and subjective tests using both simulated and practical reverberant recordings. Results show that the sparsity of speech in the perceptual time-frequency domain is maintained in both anechoic and reverberant environments and can be successfully used for creating mixtures of speech signals obtained from realistic meeting recordings. The results show excellent quality when compressing several simultaneous speech sources at a total bit rate of only 32 kbps. The results also indicate the proposed framework outperforms the separate encoding of the sources at the same overall bit rate. The approach has application to personalized rendering of speech sources in applications such as spatialized teleconferencing. Further research could include an investigation into combining the mixture generation and compression stages for further reducing the bit rate, e.g. using adaptive speech coding techniques for very low bit rate speech coding [35].

REFERENCES


Jiangtao Xi (M’95–SM’06) received the B.E. degree from the Beijing Institute of Technology, Beijing, China, in 1982, the M.E. degree from Tsinghua University, Beijing, in 1985, and the Ph.D. degree from the University of Wollongong, Wollongong, NSW, Australia, in 1996, all in electrical engineering.

From 1995 to 1996, he was a Postdoctoral Fellow with the Communications Research Laboratory, McMaster University, Hamilton, ON, Canada. From 1996 to 1998, he was with Bell Laboratories, Lucent Technologies Inc., Murray Hill, NJ, as a Member of Technical Staff. From 2000 to 2002, he was with TCL IT Group Company, China, as the Chief Technical Officer. In 2003, he rejoined the University of Wollongong as a Senior Lecturer, where he is currently a Professor at the School of Electrical, Computer, and Telecommunications Engineering. His research interests include signal processing and its applications in various areas, including audio and speech signal processing, optical electronics, and telecommunications.