2016

Encoding and communicating navigable speech soundfields

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

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

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

Publication Details

X. Zheng, C. Ritz & J. Xi, "Encoding and communicating navigable speech soundfields," Multimedia Tools and Applications, vol. 75, pp. 5183-5204, 2016.

Encoding and communicating navigable speech soundfields

Abstract
This paper describes a system for encoding and communicating navigable speech soundfields for applications such as immersive audio/visual conferencing, audio surveillance of large spaces and free viewpoint television. The system relies on recording speech soundfields using compact co-incident microphone arrays that are then processed to identify sources and their spatial location using the well-known assumption that speech signals are sparse in the time-frequency domain. A low-delay Direction of Arrival (DOA)-based frequency domain sound source separation approach is proposed that requires only 250 ms of speech signal. Joint compression is achieved through a previously proposed perceptual analysis-by-synthesis spatial audio coding scheme that encodes sources into a mixture signal that can be compressed by a standard speech codec at 32 kbps. By also transmitting side information representing the original spatial location of each source, the received mixtures can be decoded and then flexibly reproduced using loudspeakers at a chosen listening point within a synthesised speech scene. The system was implemented based on this framework for an example application encoding a three-talker navigable speech scene at a total bit rate of 48 kbps. Subjective listening tests were conducted to evaluate the quality of the reproduced speech scenes at a new listening point as compared to a true recording at that point. Results demonstrate the approach successfully encodes multiple spatial speech scenes at low bit rates whilst maintaining perceptual quality in both anechoic and reverberant environments.

Keywords
communicating, navigable, speech, encoding, soundfields

Disciplines
Engineering | Science and Technology Studies

Publication Details
X. Zheng, C. Ritz & J. Xi, "Encoding and communicating navigable speech soundfields," Multimedia Tools and Applications, vol. 75, pp. 5183-5204, 2016.

Encoding and Communicating Navigable Speech Soundfields

Xiguang Zheng[1,2], *Christian Ritz*[1], *Jiangtao Xi*[1]

[1]ICT Research Institute/School of Electrical Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia, 2522

[2]Dolby Laboratories (Beijing), No. 1, East 3rd Ring Middle Road, Beijing, China, 100020

xzhen@dolby.com, critz@uow.edu.au, jiangtao@uow.edu.au

Abstract

This paper describes a system for encoding and communicating navigable speech soundfields for applications such as immersive audio/visual conferencing, audio surveillance of large spaces and free viewpoint television. The system relies on recording speech soundfields using compact co-incident microphone arrays that are then processed to identify sources and their spatial location using the well-known assumption that speech signals are sparse in the time-frequency domain. A low-delay Direction of Arrival (DOA)-based frequency domain sound source separation approach is proposed that requires only 250 ms of speech signal. Joint compression is achieved through a previously proposed perceptual analysis-by-synthesis spatial audio coding scheme that encodes sources into a mixture signal that can be compressed by a standard speech codec at 32 kbps. By also transmitting side information representing the original spatial location of each source, the received mixtures can be decoded and then flexibly reproduced using loudspeakers at a chosen listening point within a synthesised speech scene. The total bit rate for transmission of a navigable scene of up to three speech sources is 48 kbps. Subjective results demonstrate the approach successfully encodes multiple spatial speech scenes at low bit rates whilst maintaining perceptual quality in both anechoic and reverberant environments.

Key words: Immersive Audio/Visual Conferencing, Interactive Audio/Visual Applications, Spatial Audio, Speech Soundfields

1. Introduction

Traditional audio/visual communication systems, such as a standard Voice over Internet Protocol (VoIP)-based video conference, do not replicate a face-to-face meeting experience when more than two participants are engaged in conversation. While large video displays can be used, most commonly available systems utilise only mono or stereo playback, which limits the ability to provide for spatially accurate reproduction of each participants voice. It has been demonstrated that allowing a listener to flexibly locate the speech reproduced for each participant in a teleconference provides benefits such as increased realism of the meeting as well as improving their ability to effectively multitask [1]–[5]. This is referred to in this paper as a navigable soundfield and achieving this requires more sophisticated approaches for recording, encoding and reproducing the speech scenes compared with traditional audio/visual conferencing solutions. Key to this application is to firstly consider the soundfield as consisting of multiple independent speech signals arriving from different directions, which are typically referred to as spatial audio objects [6]. Such a parametric description provides the flexibility required to achieve navigable soundfields [6] and is the state-of-the art approach for efficient encoding of 3D soundfields [7], [8] as well as allowing for flexible reproduction that is not tied to a specific loudspeaker setup [7], [9], [10]. This paper describes a complete system based on the spatial audio object approach for providing navigable soundfields. There are three main components to the system: soundfield recording and derivation of spatial speech sources; joint compression of the speech sources; and reproduction via loudspeakers. The system is designed for applications where a listener wishes to flexibly choose how and when to reproduce individual talkers or complete speech scenes, for example, within an immersive audio/visual communication application, for 'zooming in' to an area of interest in large indoor space monitored by audio recordings or reproducing a chosen listening point within free viewpoint television applications for interactive audio/visual experiences [11].

For soundfield recording to derive individual speech objects, one approach is to have each participant wear a close talking microphone. While this can lead to high quality recordings of the speech objects, the use of individual microphones is cumbersome. Further, when there are multiple participants at multiple geographic locations, this can lead to high transmission bandwidth requirements as each recorded speech signal is separately compressed and transmitted. A more realistic and natural experience can be achieved using a microphone array [12][13] to record the speech scenes at each remote site. This frees participants from using close talking microphones and allows participants to more easily join or leave an online meeting at each site. The use of microphone arrays for multichannel speech and audio processing has attracted significant attention over many years, with applications including sound source localisation, multichannel speech enhancement/noise reduction, Blind Source Separation (BSS) to identify individual sound sources and soundfield recording for subsequent playback using spatial audio reproduction approaches [13]–[16]. Sound source localisation is often simplified to estimating the Direction of Arrival (DOA), which suffices for the spatialized playback applications targeted in this paper. Time Delay of Arrival (TDOA) is a common approach for estimating the DOA from recordings from microphone arrays, such as the Uniform Linear Array [12][17]. This is based on the principle of estimating time differences of arrival for sound sources recorded by adjacent microphones based on knowledge of the microphone separations. Robust methods to achieve good performance in additive noise and reverberant environments have been proposed including the Generalised Cross Correlation with PHAse Transform (GCC-PHAT) [18] and the Steered Response Power with PHAse Transform (SRP-PHAT) [19].

In addition to designing robust signal processing algorithms that achieve high DOA estimation accuracy, two important factors for practical deployment are the size of the array and the processing delay of the chosen algorithm. One disadvantage of the TDOA-based approach is that the signals must be recorded by microphone arrays consisting of several spatially separated microphones. While increasing the number of microphones is important to improve the directional response and hence highly accurate soundfield recording [12][13], this leads to larger arrays and increases the number of channels to process. This spacing is also governed by the spatial sampling theorem, which is analogous to the time-domain Nyquist sampling theorem and requires the spacing to be less than half the wavelength of the sound of interest (for narrowband speech this equates to a minimum spacing of 5 cm to ensure spatial aliasing does not occur for an assumed maximum frequency of 3.4 kHz). Delay of the chosen algorithm can be characterised by two components: the computational complexity and the amount of signal samples required to achieve accurate DOA estimates. Both these components impact on the ability to achieve real-time performance, which is a critical issue for speech communication systems where signals must be recorded, processed and transmitted within a maximum time delay. Whilst computational complexity can be addressed by hardware choices and algorithmic optimisations, the chosen DOA method must be designed to require minimal signal samples to achieve high accuracy. To address these issues, this paper describes the use of more compact co-incident microphones for DOA estimation within the proposed immersive spatial audio communication system. Such arrays have been previously investigated for speech DOA estimation and include the acoustic vector sensor [20] and the soundfield microphone [21]. While existing techniques have been shown to achieve high accuracy, they often require a relatively high delay in terms of the length of signal required (e.g. 2 s of signal is required in [21]). Hence, this paper proposes the use of a low delay DOA estimation method applied to soundfield microphone recordings that requires approximately 250 ms of speech and hence is more suitable to immersive audio communication systems.

The other main task performed in the first stage of the system is to derive the individual speech objects. This typically requires a BSS algorithm to separate the multichannel recordings into individual speech signals. To improve performance in reverberant environments where convolutive mixing occurs, it is common to perform separation in the time-frequency domain [15]. When applied to microphone arrays, BSS approaches can utilise spatial location information to help resolve the permutation problem typically encountered in frequency domain methods by assigning time-frequencies to unique sources based on their estimated DOA. This helps to improve the separation performance and

is the approach adopted in this paper. Various time-frequency based BSS techniques using single spatial microphone recordings have been proposed [15],[22]. This typically involves first transforming time domain components to the short time frequency domain using, for example, a standard Short Time Fourier Transform (STFT) and then estimating the DOA for the individual time-frequency components. Individual sources are then identified based on labelling of time-frequency components with similar DOA estimates using clustering or other statistical analysis techniques applied to histograms formed from the DOA estimates. One limitation of this approach is the assumption that only one source is active (or has significant energy) in each time-frequency instant. While this assumption has been shown to be valid for up to 80% of time frequency components when two speech signals are mixed together, there remain time-frequency instants where more than one source is active, which becomes more common in reverberant environments and when more than two sources are active [23]. This can result in distortion of the separated speech sources. To address this, the system of this paper adopts the Collaborative BSS (CBSS) approach [24] that was previously investigated by the authors and shown to provide significantly improved separation performance in terms of estimated perceptual quality. CBSS utilises time-frequency-based DOA estimates derived from multiple co-incident microphone arrays to achieve separation and while this previous work analysed complete audio recordings (up to 10s) to maximise the DOA estimation accuracy, in this paper the low delay DOA estimation methods are used within CBSS to examine the practical performance limits.

The second stage of the system, joint compression of the speech sources, relies on spatial audio coding designed based on a spatial audio object approach. State of the art approaches include [7], [8], [25] . A unique aspect of this work is the joint compression of multiple speech sources derived from multiple spatial sound scenes to allow for efficient transmission as a single compressed mixture signal. This exploits the advantages provided by a parametric description of the soundfield, which allows for flexible encoding of multiple speech objects regardless of their originating location. In this paper, results are presented when using the previously proposed Perceptual Analysis-by-Synthesis (PABS) approach to spatial audio coding but adapted here to encoding speech objects derived using the low delay DOA and CBSS methods introduced above.

The third and final stage in the proposed system is the flexible reproduction of speech scenes via loudspeakers. In this work, a 5.1 surround sound system was chosen for the loudspeaker reproduction as it was determined suitable for spatialising the limited number of speech sources in each scene as well as being a common reproduction setup. An alternative approach could be binaural reproduction using headphones, however loudspeakers were chosen as a more natural reproduction experience targeting multiple participants at each geographical site. A key task in this stage is to also selectively create the so called “listening point”. This allows a user to effectively “zoom in” to a particular location within the original soundfield to listen to a specific talker. This is achieved by also encoding information representing the original spatial location of each speech object. Such information is then used to derive virtual microphone recordings at the chosen listening point that can then be used to reproduce a personalised soundfield using loudspeakers. A subjective evaluation using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) approach [26] was conducted to measure the performance of the proposed system for reproducing speech scenes of high quality.

Section 2 of this paper presents the system overview of the proposed soundfield while Section 3 describes the low delay DOA estimation approach for speech sound objects. Section 4 will review the CBSS approach based on low delay DOA estimation. Section 5 will review the PABS approach used for soundfield compression while Section 6 will describe the flexible reproduction of speech soundfields. Results from subjective testing are presented in Section 7 with conclusions and future work presented in Section 8.

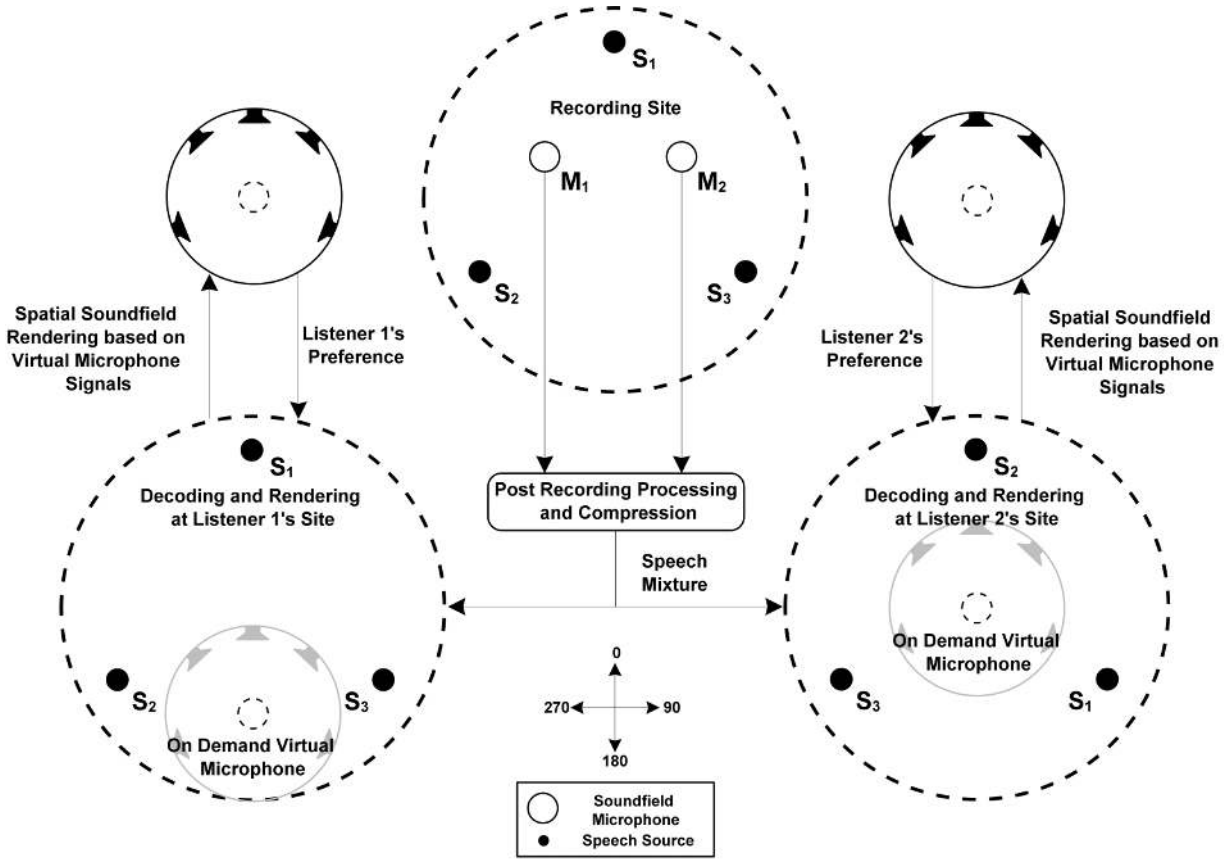

Fig. 1 An example soundfield navigation scenario. A speech soundfield consisting of three sources is recorded at one site using two co-incident microphones. Listener 1 and Listener 2 at two different remote sites each select a different 'listening' position within the original soundfield, which is reproduced by deriving a virtual microphone recording at the selected 'listening point'.

2. Overview of Soundfield Navigation and Speech Sparsity

This section describes the soundfield navigation framework investigated in this paper, introduces the proposed system for achieving soundfield navigation and highlights the key concept of speech sparsity used within the proposed system.

2.1 Soundfield Navigation

Fig. 1 is an illustrative example of a soundfield navigation scenario where three sources (S_1, S_2 and S_3) are recorded by two microphone arrays M_1 and M_2 at a given recording site. It is desirable that:

- the information of the interested soundfiled can be efficiently captured by a limited number of observations (i.e. the spatial recordings)
- these spatial recordings can be efficiently compressed such that the listening points (on demand virtual microphone signals) can be interactively selected by different users based on the same transmitted signal.

The proposed soundfield navigation system starts by employing multiple co-incident microphone arrays (M_1 and M_2 in the example of Fig. 1) to record the speech soundfield. The recorded signals are then processed to derive DOA estimates for each source which is then followed by a source triangulation process to pinpoint the location of each source within the recording site. The triangulation procedure requires prior knowledge of the microphone locations, which is encoded within the transmitted side-information.

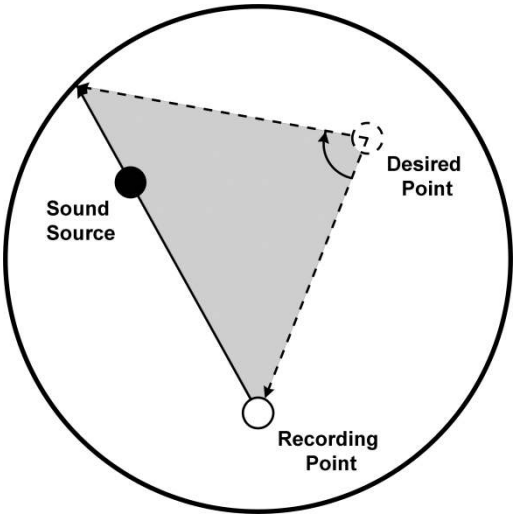

Fig. 2 Using DOA derived at a single recording point does not allow for accurate reproduction at a desired listening point.

It should be noted that while other BSS techniques may also successfully achieve source separation using DOA estimates, source location (i.e. spatial position including distance from the microphone) cannot usually be derived with a single microphone array. An example is given in Fig. 2. It can be observed that while source DOA can be estimated from a single microphone recording, the possible source direction with respect to the desired listening point can be in any direction within the grey area. Hence, two co-incident microphones are used within the scenario of Fig. 1 whilst the system proposed for achieving soundfield navigation is described in the next sub-section.

2.2 Soundfield Navigation Framework

The system of the proposed soundfield navigation framework is illustrated in Fig. 3. Recordings from multiple co-incident microphone arrays are transformed to the frequency domain and processed to estimate the DOA of each speech source using the low delay technique described in Section 3. Based on these DOA and source location estimates, the CBSS technique described further in Section 4 is performed to jointly separate the sources from the microphone recordings where each microphone will have one set of separated sources. These separated speech sources are then further processed by a source ownership estimation stage and followed by employing the Psychoacoustic-based Analysis-By-Synthesis (PABS) compression scheme [23], as discussed in Section 5. The compressed mixture signal is further encoded by the AMR-WB+ [27] codec and transmitted along with side information representing the spatial location of speech sources. By receiving the same compressed speech signal with spatial side information, each user can then select a desired reproduced soundfield by "zooming in" to a preferred location and performing selective playback of the simultaneous speech sources as shown in Figure 1. This is achieved by simulating a virtual microphone recording signal at the desired listening point from the received speech sources and their spatial location information. The personalized soundfield signal can be reproduced by standard 5.1 surround playback system while the listening point can be adjusted freely by the user. A key to implementing this framework is a reliance on the sparsity of speech soundfields in the time-frequency domain.

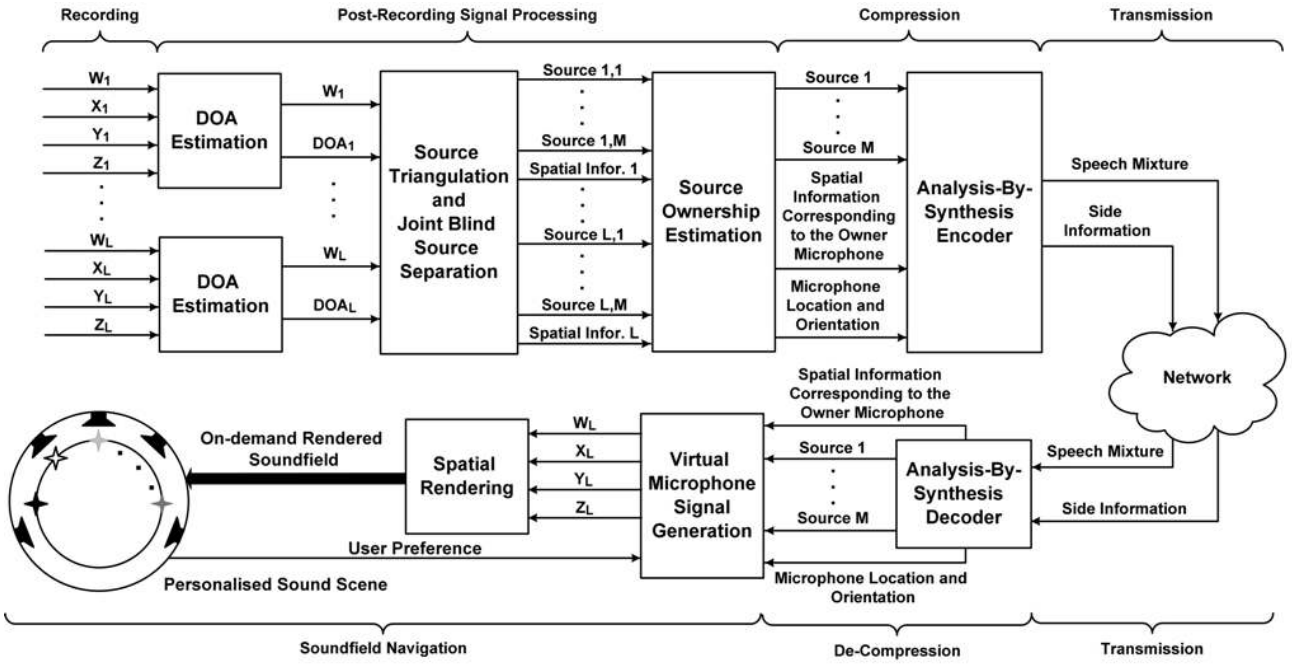

Fig. 3 Soundfield Navigation System including all stages from recording to reproduction.

2.3 Sparsity of Speech Soundfields

Speech signals are known to be sparse in the time-frequency domain. The STFT has been employed in the BSS technique described in [28] to separate the speech signals via sparse-based binary mask. The sparse property of speech can be generally described by:

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

where $S_i(n,k)$ and $S_j(n,k)$ is the time-frequency representation of simultaneously occurring speech signal s_i and s_j, respectively, n is the frame number and k is the frequency index. This speech orthogonality (1) has been verified in [28] to be satisfied for time-frequency components corresponding to 94% of the energy of 2 simultaneous speech sources. While the orthogonality of (1) reduces as the number of sources increases, results in [28] still show that 79% of the energy of simultaneous speech sources satisfies (1). Hence, due to the sparseness of the speech resulting from (1), peaks in the histogram formed from time-frequency DOA estimates [20]–[22] correspond to unique sources, i.e. one time-frequency component is contributed by no more than one source. Thus if the time-frequency DOA estimates have the same DOA, they correspond to the same source. Hence, the source locations can be estimated by finding the peaks of the DOA histogram. This is the basis of the time-frequency DOA estimation algorithms used in this work and forms the basis of the source separation approach using CBSS (Sections 4) as well as the soundfield compression approach using the PABS spatial audio coder [23] (Section 5).

3. Low Delay DOA Estimation

Fig. 4 is an illustration of the system used for low delay DOA estimation consisting of the following four stages: Time-Frequency DOA estimation; Energy Weighted DOA Histogram Estimation; Kernel Density estimation; and Local Maximum Estimation. Each of these stages will be described in more detail in this section.

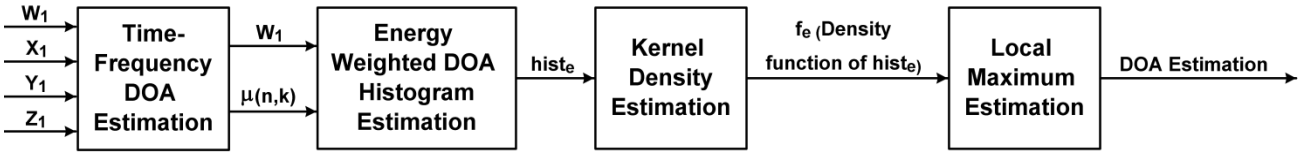

Fig. 4 Overview of Low-Delay DOA Estimation System.

3.1 Time-Frequency DOA Estimation

A soundfield microphone is adopted as the co-incident microphone array in this work, which records four channels that are converted to the Ambisonics B-format that includes one omnidirectional channel (w) and the three orthogonal directional components of the sound field (x, y and z). Converting these signals to the time frequency domain results in the corresponding frequency domain signals indicated in Fig. 4 as W_1, X_1, Y_1 and Z_1). Following [20], [22], [24] the azimuth of the DOA is estimated for each time-frequency as:

$$\mu_{m,l}(n,k) = tan^{-1}(\frac{Y_1(n,k)}{X_1(n,k)}) \tag{2}$$

While we assume 2D sources in this work the extension to 3D is straightforward. The DOA of each speech source is identified as a peak of the histogram formed from estimates of (2) [20], [22], [24]. An example DOA histogram is shown in Fig. 5 (a) for a 10 s recording of 3 simultaneously occurring speech sources recorded in a reverberant environment and sampled at 20 kHz and using a 1024 point Modified Discrete Cosine Transform (MDCT) with 50% overlapping windows. Estimated locations of the three sources are indicated by the peak labels S1, S2 and S3.

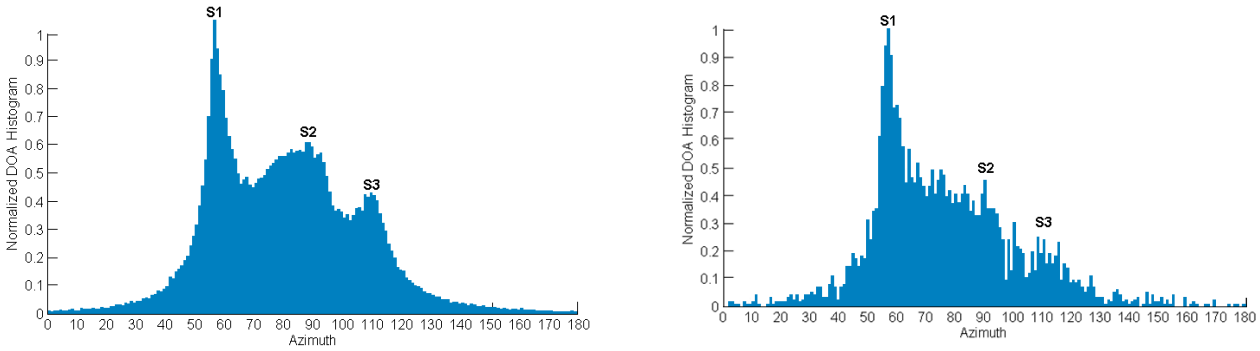

Fig. 5 The normalised DOA Histogram obtained from a recording of three simultaneous sources labelled as the peaks (a) using the entire 10 s duration of the sources (b) using 250 ms of data (approximately 9 frames for 20 KHz sampling, 1024 point MDCT and 50% overlapping).

3.2 Energy Weighted DOA Histogram Estimation

While reliable DOA estimation has been achieved by analysing the intensity vector statistics [21], this DOA estimation process requires 2 seconds of the recordings, which is not desirable for low-delay applications, where DOA estimation over a short period of time e.g. less than 400 ms is more desirable for applications such as speaker tracking and segmentation [29]. However, using less data makes it more difficult to accurate determine the peaks in the DOA histogram. This can be seen in the example of Fig. 5 (b) for the same recording as Fig. 5 (a) but now using only 250 ms of data. To address this, it is common to adopt techniques to consider only time-frequencies where the direct component

of the source signal has significant energy compared with the background noise or reverberant components. In this paper, an energy weighting is applied to the DOA histogram such that time-frequencies with higher energy make a greater contribution to the histogram count compared to low energy sources. It was found that this resulted in more pronounced histogram peaks when estimating the DOA using a limited number of frames of time-frequency components.

3.3 Kernel Density Estimation and Local Maximum Estimation

While energy weighting leads to stronger peaks in the DOA histogram, an appropriate peak picking method is still required. Here, a stochastic approach is considered whereby the DOA histogram is assumed to be an undersampled representation of the probability distributions of the DOAs corresponding to each of the underlying sources. The Kernel Density Estimation [30] (KDE) method is used to estimate these probably density functions and results in smooth density distribution curves with peaks that can be then processed to determine source locations. Fig. 6 shows an example of applying this approach to the energy weighted DOA histogram using a Gaussian kernel and for three sources and using only 9 frames of data. The final step then involves a local maximum estimation of the peaks of the KDE derived distribution function to identify a unique source DOA. These are indicated in green in Fig. 5 for the example multisource recording.

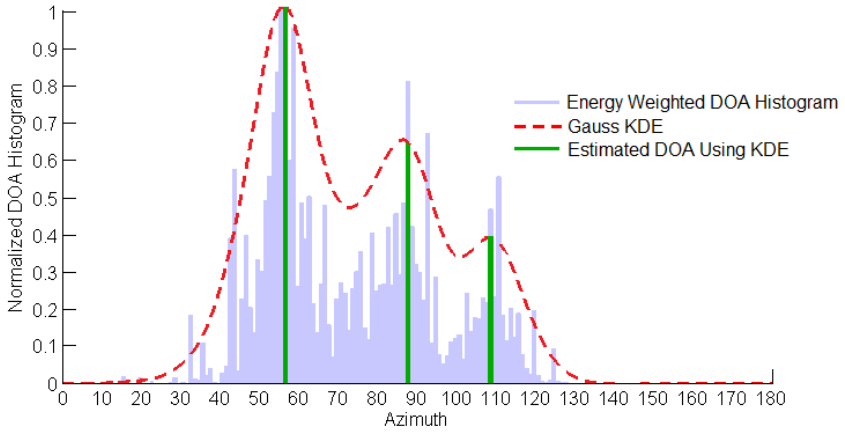

Fig. 6 KDE applied to the energy weighted DOA histogram obtained for three sources and 9 frames of data. The Gaussian-based KDE derived probability density function is shown as the dashed red curve while estimated peaks corresponding to a source DOA are indicated by the green bars.

4. Low delay Collaborative Blind Source Separation (CBSS)

This section provides and overview of the Collaborative Blind Source Separation (CBSS) approach used for deriving unique speech sources from the co-incident microphone array recordings. Since the approach uses multiple co-incident microphones to improve the separation and localisation, this results in mutiple versions of each source (one for each microphone) and hence a source ownership stage is described for selecting one version for subsequent compression and transmission.

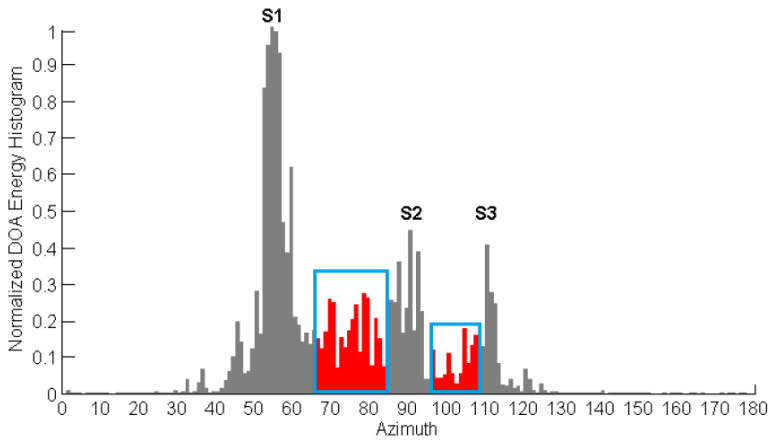

Fig. 7 Example cause of musical distortion in separated speech signals. The section of the DOA histogram illustrated in red corresponds to time-frequency components with multiple active sources.

4.1Overview of CBSS

When the w-disjoint orthogonality of simultaneously occurring speech signals (1) is met, DOA estimates performed in the time-frequency domain using (2) will correspond to the location of a true speech source. In practice, simultaneously occurring speech signals are not strictly w-disjoint orthogonal for all time-frequencies [24] and the separated speech signals using the sparse-based approaches applied to the mixture suffer spectral distortion. This is a result of the non-sparse components combining in the mixture and hence DOA estimates using the recorded channels X_1 and Y_1 in (2) do not correspond to true source DOAs. An illustrative example is provided for the DOA histogram of Fig. 7, where the sections highlighted in red occur in between true source DOAs. Since the time-frequency of the DOA estimates in the highlighted sections must be assigned to one source or discarded, this can lead to a distortion of the spectrum of one or more sources when the w-disjoint orthogonality assumption is violated. Further, if three frontal sources of equal energy are considered, one directly in line with the array and two at equal angles but opposite sides of the array, the non-sparse components contributed by the left and right sources may lead to the same DOA estimate as the middle source. This causes crosstalk distortion, where the separated sources contain spectral content from more than one source at the corresponding time-frequency. A similar problem can exist in the Linearly Constrained Minimum Variance (LCMV) [31] beamformer, where the distortionless constraint can be difficult to maintain when there are multiple overlapping time-frequency sources.

To address these problems, the Collaborative Blind Source Separation (CBSS) technique is adopted here. This was originally proposed in [24] and aims to decompose the mixture of non-sparse components into their corresponding sources using a pair of coincident microphone arrays with known location. This assumes that no more than two speech sources contribute to one time-frequency instant in the mixture. Based on the possible contributor source pairs for one coincident microphone array, their corresponding DOA for the second coincident microphone array is estimated. The non-sparse components can then be correctly decomposed by comparing these estimates with the DOA obtained from the second coincident microphone array recordings. A detailed description of the methods used to resolve musical and cross-talk distortion is provided in [24], where results from objective PESQ and subjective MUSHRA tests verified the significant improvement in the quality of separated speech signals compared with existing BSS approaches. While [24] utilized 10 s of data for source DOA estimation, in this paper the low delay DOA estimation technique of Section 3 is adopted within the proposed soundfield navigation framework.

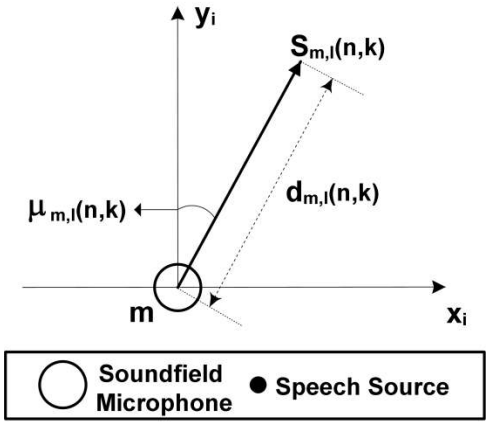

Fig. 8 Spatial parameters used to represent the original location of the speech source within the recorded soundfield.

4.2 Source Ownership Estimation

Applying CBSS process to each microphone will result in multiple estimates of the each source. Denoting $S_{m,l}(n,k)$ as the m^{th} source separated from the l^{th} microphone recording, for time-frequency instant (n,k), spatial parameters are required to represent the spatial information of this time-frequency source. These spatial parameters (as shown in Fig. 8) are obtained from the source triangulation stage and can be represented as the combination of the azimuth $\mu_{m,l}(n,k)$ and the distance $d_{m,l}(n,k)$ corresponding to microphone l, which is given by:

$$P_{m,l}(n,k) = [\mu_{m,l}(n,k), d_{m,l}(n,k)] \tag{3}$$

Thus, if the spatial location and orientation of the l^{th} microphone is known, the spatial location of source m with respect to microphone l can be derived from $d_{m,l}(n,k)$. Since the separated sources from different microphone recordings are duplicated, it is redundant to transmit all of these sources to represent the original soundfield. Here, the source ownership estimation aims to preserve the best version of the separated source among all available versions. Suppose $Q(\cdot)$ represents the quality of the separated source. Microphone l^o owns one particular source S_m if

$$l^o = arg\ max_l \left(Q\left(S_{m,l}\right)\right) \tag{4}$$

Here, the selection criterion used for $Q(\cdot)$ is based on the minimum source to microphone distance and hence denoting $d_{m,l}$ is the distance between the m^{th} source and the l^{th} microphone leads to

$$l^o = \arg min_l\left(d_{m,l}\right) \tag{5}$$

Note that if the source is located at the same distance to the microphones, the owner is assigned to either of the microphone. Thus, for each source, only one version is sent to the compression stage with side information indicating the spatial parameter corresponding to the owner microphone. For source S_m, the spatial parameters P_m corresponding to the owner microphone is:

$$P_m = \left[l_m^o, P_{m,l_m^o}\right] = \left[l_m^o, \mu_{m,l_m^o}, d_{m,l_m^o}\right] \tag{6}$$

S_m and P_m will be sent to the PABS compression stage along with the microphone location and orientation.

5. Soundfield compression via PABS

This section provides an overview of the psychoacoustic-based analysis-by-synthesis approach for spatial audio coding that is employed within the soundfield navigation system.

5.1 Overview of PABS

A psychoacoustic-based analysis-by-synthesis approach is employed to compress the navigable speech sources [23]. 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. The mono mixture signal is formed from the time-frequency components estimated from the CBSS stage of Section 4. Side information is used to store information about the corresponding source label and original spatial location, which enables flexible decoding and reproduction. For time-frequency instants where there is more than one active source estimated from the CBSS stage, an iterative process based on a perceptual distortion measure as described in [24] is used to maximise an overall objective estimation of the perceptual quality of each source in each frame. The mono mixture signal is further compressed using a standard speech codec, which in this case as in [24] is the AMR-WB+ codec operating at 32 kbps. Full details can be found in [23] where results from subjective testing showed that the approach can both main perceptual quality of individual speech sources as well as the perceptual quality of the spatialised speech scene. While [24] original sources were available or recorded with close-talking (lapel) microphones, in this paper sources are obtained using the low delay DOA and CBSS approaches applied to the microphone array recordings (Sections 3 and 4). The next sub-section further describes how the corresponding speech sources are decoded using the received mixture signal and side information.

5.2 PABS for Soundfield Navigation

Side information in the time-frequency domain indicates the origin of the preserved time-frequency sources. Assume $S_d{}'(n,k)$ represents the decoded time-frequency from the received mono PABS mixture and $P_d{}'(n,k)$ is the received spatial parameter. For the m^{th} source, a separation mask $M_m(n,k)$ can be obtained as:

$$M_m(n,k) = \begin{cases} 1, P_d{}'(n,k) = P_m(n,k) \\ 0, otherwise \end{cases} \tag{7}$$

where $P_m(n,k)$ indicates the desired source. Thus, the reconstructed speech source $S_m{}'(n,k)$ with the spatial parameter $P_m{}'(n,k)$ can be extracted in the time-frequency domain by:

$$S_m^{\prime(n,k)} = M_m(n,k) \cdot S_d^{\prime(n,k)} \forall n,k \tag{8}$$

The extracted time-frequency sources and corresponding spatial parameter will be used to achieve free listening point navigation by generating the virtual microphone signal at the desired listening point. Details of virtual microphone signal generation is presented in the next Section.

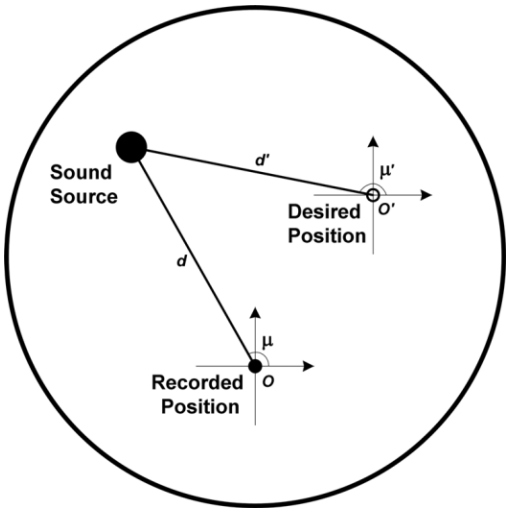

Fig. 9 Selective Listening Point.

6. Flexible Reproduction of Speech Soundfields

As discussed in Section 5, the speech sources are separated from the mixture along with their spatial parameters. Thus, for one speech source, the available information in the receiver end is the separated speech source with the azimuth, μ, distance from the owner microphone to the source, d, and the source ownership information, l. As shown in Fig. 9, if the source is owned by the microphone located at position O, the aim is to generate the virtual microphone signal at O' based on the available information at recorded position O. In order to simulate a high quality virtual microphone signal, the following two requirements need to be ensured:

- The power (volume) of the simulated virtual microphone signal should be generated based on the distance difference, i.e. (the difference between d and d')
- The spatial location of the sources in respect to the new location, i.e. μ'

The spatial location of the source for the virtual microphone can be obtained by geometrical calculation. Suppose the owner microphone is located at $O(0,0)$ (note that this information is transmitted from the recording site and assumed to be known) and the desired position is located at $O'(x, y)$, the position of the source with respect to the owner microphone can be calculated from μ and d as $(d \cdot cos\ \mu\ , d \cdot sin\ \mu)$. The position of the source with respect to the virtual microphone is ($d \cdot cos\mu - x, d \cdot sin\mu - y$). Thus, the source azimuth in respect to the virtual microphone μ' is given by:

$$\mu' = \begin{cases} \arctan\frac{d \cdot sin\ \mu - y}{d \cdot cos\mu - x}, if\ d \cdot sin\ \mu - y > 0 \\ \arctan\frac{d \cdot sin\ \mu - y}{d \cdot cos\mu - x} + 180, if\ d \cdot sin\ \mu - y < 0 \\ 0, if\ d \cdot sin\ \mu - y = 0, and\ d \cdot cos\mu - x > 0 \\ 180, if\ d \cdot sin\ \mu - y = 0, and\ d \cdot cos\mu - x < 0 \end{cases} \tag{9}$$

d' is given by:

$$d' = \sqrt{(\ d \cdot sin\ \mu\ - y\ )^2\ \ +\ (\ d \cdot cos\ \mu\ - x\ )^2} \tag{10}$$

Hence, by using the inverse-square law of sound propagation [32], the virtual microphone signal S' is given by:

$$S' = S \cdot d^2 / d'^2 \tag{11}$$

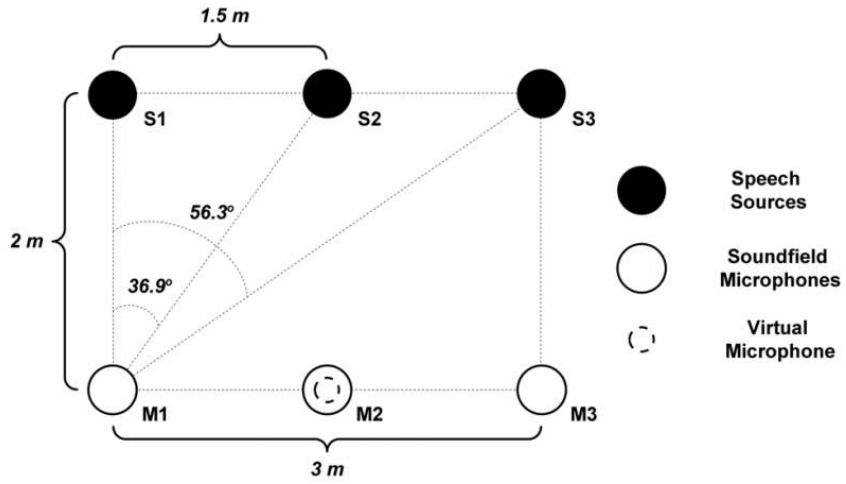

Fig. 10 Recording Configuration for the Subjective Evaluation.

Note that for reverberant conditions, the virtual microphone signal generated based on (11) simulates the direction source while reverberant effects may be generated based on the simulated direct source according to specific room configuration of the recording site, which is out of the scope for this thesis. Using S' and d', the spatial recording at position O' can be simulated. Spatialised reproduction using loudspeakers is achieved using frequency domain amplitude panning [33], [34].

7. Subjective Evaluation

In this section, the proposed soundfield navigation system is evaluated. The aim of this evaluation is to compare the sound scene simulated based on the purposed navigation framework with real recordings of the same scene. As illustrated in Fig. 10, the recording setup consists of three soundfield microphones to record three speech sources. The recordings of microphone M2 is used as the ground truth where simulated virtual microphone signals are generated from recordings of M1 and M3 in the same location. The Australian National Database of Spoken Language [35] is chosen for the evaluation. A total of 24 Sentences (sampled at 20 kHz) containing 24 different Australian native speakers of different ages and genders were selected as the testing database. Three recording conditions are considered in the evaluation: an anechoic chamber; and small and large conference rooms. The anechoic condition used a Core Sound TetraMic [36] to record two to three overlapping speech sources. The two reverberant conditions using the image method [37] were implemented through RoomSim [38] to simulate the reverberant recordings of the small (RT60 = 200 ms) and large (RT60 = 500 ms) conference room. A total of 9 sessions of overlapping speech sources (3 sessions each include 2 overlapped sources, other 6 sessions each include 3 overlapped sources) are employed in the MUSHRA test. A total of 15 people participated the listening test and conditions for each test file are listed in Table 1.

The results are shown in Fig. 11 with 95% confidence intervals. Note that the PABS scheme requires 32 kbps to compress the speech mixture while up to 8 kbps to compress the spatial location parameter. Here, the distance also needs to be compressed at a total bit rate of (up to) 8 kbps depending on required accuracy leading to a total bit rate of 48 kbps for transmitting the mixture signal and side information. As shown in Fig. 11, compared to the reference (the ground truth), the proposed framework achieved excellent perceptual quality (MUSHRA scores all above 80) for generating the virtual microphone signals. Conditions for only testing the CBSS technique (i.e. assuming perfect condition for other parts such as DOA estimation and compression) to evaluating the whole system (condition DOA_CBSS_PABS) under the low-delay condition achieved similar MUSHRA scores. Note that the listeners can always pick up the hidden references and the anchors.

Table 1. MUSHRA Test Conditions for Fig. 11

Name	Descriptions
REF	The spatial reproduction based on real recording in M2
CBSS	The spatial reproduction based on simulated virtual microphone signal generated from real recording in M1 and M3 using CBSS method
CBSS_PABS	The speech sources from condition CBSS further compressed using the PABS scheme at 48 kbps then rendered similar to condition CBSS
DOA_CBSS	Condition CBSS using the low-delay DOA estimates to separated the speech sources and rendered based on these low-delay version sources
DOA_CBSS_PABS	The speech sources from condition DOA-CBSS further compressed using the PABS scheme at 48 kbps then rendered similar to condition CBSS
Anchor	The 3.5 kHz low-pass filtered unlocalised anchor

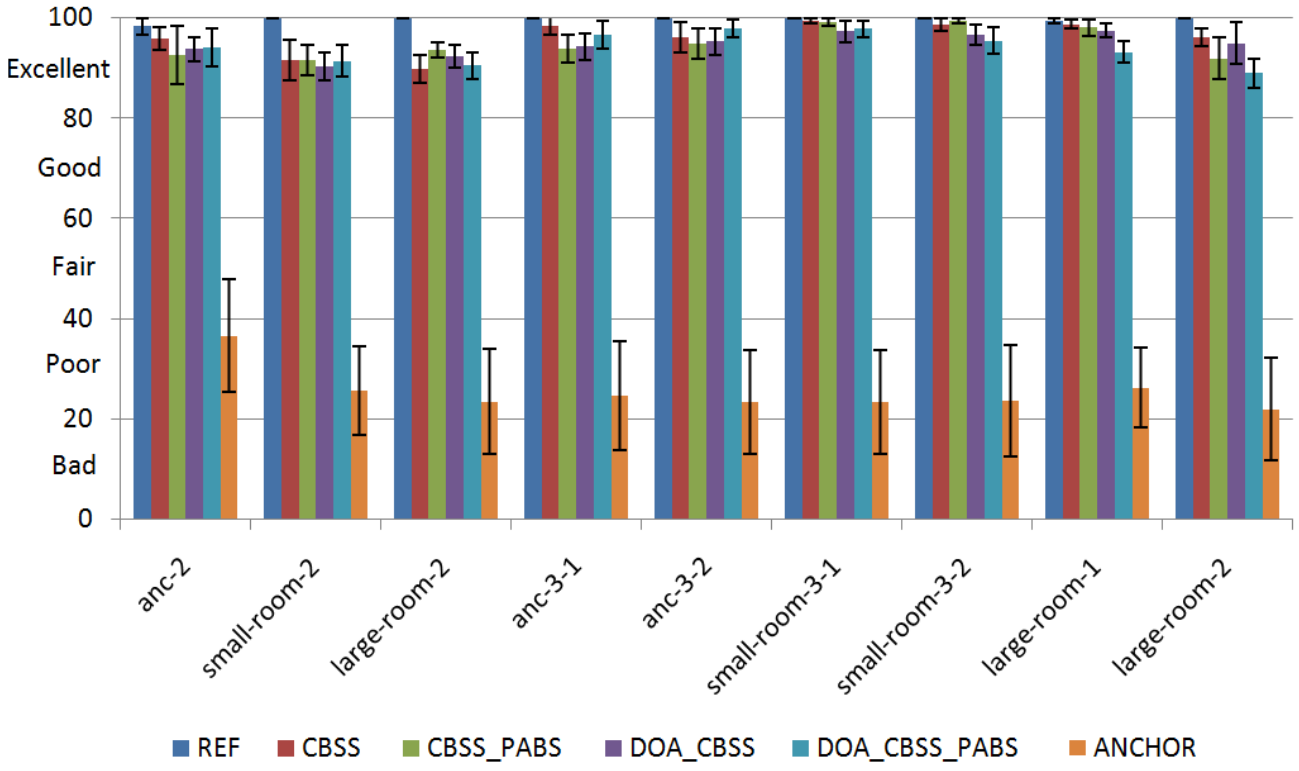

Fig. 11. MUSHRA test Results.

In order to understand the source of the minor degradation in Fig. 11, the possible types of distortions have been classified into two categories, namely, the distortion of speech quality and the inaccurate spatial location of the source. The participants of the listening test were also asked to point out the source of degradation while providing the MUSHRA score. For each condition, the options are:

- No distortion (None)
- Distortion of speech quality (Speech)
- Distortion of spatialisation (Spatialisation)
- Distortion of both speech quality and spatialisation (Both)

Table 2. Degradation Analysis for the listening test files.

Name	Source of Distortion			
	None	Speech	Spatialisation	Both
REF	99%	1%	0%	0%
CBSS	78%	22%	1%	0%
CBSS_PABS	63%	34%	3%	0%
DOA_CBSS	64%	33%	2%	0%
DOA_CBSS_PABS	68%	31%	1%	0%
Anchor	0%	0%	0%	100%

The average percentages of listeners choosing each distortion type over all nine files evaluated are shown in Table 2, which shows that more than 60% of listeners indicated no distortion for all evaluated conditions. Comparing the two types of distortion, the speech distortion was the most commonly chosen option by listeners, with only up to 3% of listeners on average indicating distortion in the spatialisation of the speech soundfield. The highest percentage is for the CBSS condition and corresponds to the MUSHRA results as in this condition there is no compression and errors due to DOA estimation inaccuracies are minimized by processing the entire speech recording. Higher percentages were found for the conditions incorporating low delay DOA (DOA_CBSS) and DOA_CBSS_PABS), which all had similar results. Comparing these with the results for CBSS indicate the compression likely contributes most to the overall perceived speech distortion since results for CBSS_PABS (without low delay DOA compression) are similar to results for the two low delay DOA conditions (DOA_CBSS and DOA_CBSS_PABS).

8. Conclusions

This paper described a framework for encoding and communicating navigable speech soundfields for immersive audio/visual applications. Presented are details of each stage of a system incorporating this framework that includes a low delay approach to estimating the DOA of individual speech sources, incorporation of the DOA information within a BSS approach that utilises multiple co-incident microphones for reducing distortion of separated sources and a perceptual-based compression approach for low bit rate encoding of mixtures representing the soundfields. The subjective results indicate the proposed framework successfully achieves low-delay free listening point navigation by employing two soundfield microphone recordings while only requiring up to 48 kbps for compressing the navigable speech soundfield. The presented framework ensures satisfactory perceptual quality of the speech sources as well as their correct spatialisation. The proposed approach has application to creating personalized sound scenes for spatialised multi-site teleconferencing, remote surveillance of large spaces and free-viewpoint TV. Suggestions for future research include further reducing the data required for accurate DOA estimation and a more detailed analysis of the performance as a function of number of navigable speech sources and total transmission bit rate.

9. References

[1] J. J. Baldis, “Effects of Spatial Audio on Memory, Comprehension, and Preference During Desktop Conferences,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2001, pp. 166–173.

[2] M. J. Evans, A. I. Tew, and J. A. S. Angus, “Perceived Performance of Loudspeaker-Spatialized Speech for

Teleconferencing," *J. Audio Eng. Soc.*, vol. 48, no. 9, pp. 771–785, Sep. 2000.

[3] D. B. Ward and G. W. Elko, "Robust and adaptive spatialized audio for desktop conferencing," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1099–1099, Feb. 1999.

[4] S. N. Wrigley, S. Tucker, G. J. Brown, and S. Whittaker, "Audio spatialisation strategies for multitasking during teleconferences," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 2935–2938.

[5] A. Mieczakowski, J. Goodman-Deane, J. Patmore, and J. Clarkson, "Conversations, Conferencing and Collaboration: the effectiveness of distributed meetings," Engineering Design Centre, University of Cambridge, 2013.

[6] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric Spatial Sound Processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31–42, Mar. 2015.

[7] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Top. Signal Process.*, vol. PP, no. 99, pp. 1–1, 2015.

[8] M. Jia, Z. Yang, C. Bao, X. Zheng, and C. Ritz, "Encoding Multiple Audio Objects Using Intra-Object Sparsity," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 1082–1095, Jun. 2015.

[9] Dolby Laboratories, "Dolby ATMOS Cinema Technical Guidelines," *Available HttpwwwdolbycomuploadedFilesAssetsUSDocProfessionalDolby-Atmos-Cine.-Tech.-Guidel.*, 2012.

[10] "EBU Technology & Innovation - Audio Definition Model Ver. 1.0." [Online]. Available: https://tech.ebu.ch/publications/tech3364. [Accessed: 02-Jun-2015].

[11] R. Oldfield, B. Shirley, and J. Spille, "Object-based audio for interactive football broadcast," *Multimed. Tools Appl.*, vol. 74, no. 8, pp. 2717–2741, May 2013.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2009.

[13] G. W. Elko and J. Meyer, "Microphone Arrays," in *Springer Handbook of Speech Processing*, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 1021–1041.

[14] Y. (Arden) Huang, J. Benesty, and J. Chen, "Time Delay Estimation and Source Localization," in *Springer Handbook of Speech Processing*, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 1043–1063.

[15] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive Blind Source Separation Methods," in *Springer Handbook of Speech Processing*, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 1065–1094.

[16] R. Rabenstein and S. Spors, "Sound Field Reproduction," in *Springer Handbook of Speech Processing*, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 1095–1114.

[17] J. Eargle, *The microphone book*. Focal Press, 2004.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoust. Speech Signal Process. IEEE Trans. On*, vol. 24, no. 4, pp. 320 – 327, Aug. 1976.

[19] H. Do, H. F. Silverman, and Y. Yu, "A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 1, pp. I–121 –I–124.

[20] M. Shujau, C. H. Ritz, and I. S. Burnett, "Separation of speech sources using an Acoustic Vector Sensor," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, 2011, pp. 1 –6.

[21] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Intensity vector direction exploitation for exhaustive blind source separation of convolutive mixtures," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 41 –44.

[22] B. Gunel, H. Hachabiboglu, and A. M. Kondoz, "Acoustic Source Separation of Convolutive Mixtures Based on

Intensity Vector Statistics," *Audio Speech Lang. Process. IEEE Trans. On*, vol. 16, no. 4, pp. 748 –756, May 2008.

[23] X. Zheng, C. Ritz, and J. Xi, "A psychoacoustic-based analysis-by-synthesis scheme for jointly encoding multiple audio objects into independent mixtures," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 281–285.

[24] X. Zheng, C. Ritz, and J. Xi, "Collaborative Blind Source Separation Using Location Informed Spatial Microphones," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 83–86, Jan. 2013.

[25] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-Based Informed Source Separation: Nonnegative Tensor Factorization Approach," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 8, pp. 1699–1712, Aug. 2013.

[26] B. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*. Springer, 2003.

[27] "3GPP specification: 26.290." [Online]. Available: http://www.3gpp.org/DynaReport/26290.htm. [Accessed: 02-Jun-2015].

[28] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830– 1847, Jul. 2004.

[29] V. Rozgic, C. Busso, P. G. Georgiou, and S. Narayanan, "Multimodal Meeting Monitoring: Improvements on Speaker Tracking and Segmentation through a Modified Mixture Particle Filter," in *IEEE 9th Workshop on Multimedia Signal Processing, 2007. MMSP 2007*, 2007, pp. 60–65.

[30] G. H. Givens and J. A. Hoeting, *Computational Statistics*, 1st ed. Wiley-Interscience, 2005.

[31] S. Gannot and I. Cohen, "Adaptive Beamforming and Postfiltering," in *Springer Handbook of Speech Processing*, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 945–978.

[32] D. M. Howard and J. Angus, *Acoustics and Psychoacoustics*. Taylor & Francis, 2009.

[33] V. Pulkki and M. Karjalainen, "Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 739–752, Sep. 2001.

[34] V. Pulkki, "Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 753–767, Sep. 2001.

[35] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The Australian National Database of Spoken Language," in *, 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, 1994, vol. i, pp. I/97–I100 vol.1.

[36] "Core Sound — TetraMic." [Online]. Available: http://www.core-sound.com/TetraMic/1.php. [Accessed: 02-Jun-2015].

[37] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[38] D. Campbell, K. Palomäki, and G. Brown, "A MATLAB simulation of 'shoebox' room acoustics for use in research and teaching.," *Comput. Inf. Syst. J. ISSN 1352-9404*, vol. 9, no. 3, 2005.