BREAKING DOWN THE COCKTAIL PARTY: CAPTURING AND ISOLATING SOURCES IN A SOUNDSCAPE

Anastasios Alexandridis[⋆†], *Anthony Griffin*[⋆], *and Athanasios Mouchtaris*[∗⋆†]

[⋆]FORTH-ICS, Heraklion, Crete, Greece, GR-70013
[†]University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013

ABSTRACT

Spatial scene capture and reproduction requires extracting directional information from captured signals. Our previous work focused on directional coding of a sound scene using a single microphone array. In this paper, we investigate the benefits of using multiple microphone arrays, and extend our previous method by allowing arrays to cooperate during spatial feature extraction. We can thus render the sound scene using both direction and distance information and selectively reproduce specific "spots" of the captured sound scene.

Index Terms— Microphone array, beamforming, source separation, spatial audio, sensor network

1. INTRODUCTION

Spatial audio refers to the reproduction of a soundscape by preserving the spatial information. The soundscape is usually encoded into multiple channels and reproduction is performed using multiple loudspeakers or headphones [1–5]. The use of microphone arrays for spatial audio recording has attracted attention, due to their ability to perform operations such as Direction-of-Arrival (DOA) estimation and beamforming.

Different approaches for recording spatial audio with microphone arrays have been investigated, such as high-order differential arrays [6, 7], and DOA estimation combined with Head-Related Transfer Functions (HRTFs) for binaural spatial audio [8]. Microphone arrays have also been proposed as a recording option for Directional Audio Coding (DirAC) [5]. In [9] a planar microphone array is employed and the microphone signals are converted to B-format for DirAC processing. The work in [10] combines DirAC with a linear array. DOA and sound diffuseness are estimated for each time-frequency element, using microphone array processing: a modified version of ESPRIT estimates the DOA, while the estimation of diffuseness is based on the Magnitude Squared Coherence (MSC) between the two outer microphones.

These techniques either restrict the loudspeaker configuration according to the microphone configuration, or ignore the spatial-aliasing that occurs in microphone arrays. The latter makes the accurate estimation of spatial features (direction and/or diffuseness) very challenging across the whole spectrum of frequencies, and degrades the quality of reproduction. In our previous work [11, 12] we proposed a real-time method for spatial audio recording using a circular microphone array that mitigates some of these problems by counting the number of active sources and estimating their DOAs for each time-frame—and not individually for each frequency. Based on the estimated DOAs, we separate the source signals through spatial filtering with a superdirective beamformer. Finally, all source signals—and thus the entire soundscape—are downmixed into one monophonic signal and side-information.

Based on our method, we developed ImmACS, an Immersive Audio Communication System. The goal of ImmACS is to capture the soundscape at the recording side using a microphone array and reproduce it using multiple loudspeakers or headphones in real-time. The capturing and reproducing sides of ImmACS can be located far apart, so the encoded soundscape needs to be transferred through the Internet (Fig. 1). ImmACS also gives the listeners the ability to select the directions they want to hear and attenuate the sources that come from other directions. For these features, source isolation is important to provide accurate spatial impression or reproduce specific sources while attenuating others in the soundscape.

In this paper, we review ImmACS and investigate the use of multiple microphone arrays for recording spatial audio. Motivated by situations where a single microphone array cannot provide sufficient spatial coverage—such as when the angular separation of sources is very small or the sources have the same DOA with respect to the array—we extend ImmACS by allowing multiple arrays to cooperate in order to provide better and more robust source isolation.

2. IMMACS: IMMERSIVE AUDIO COMMUNICATION SYSTEM

In this section we summarize the basics of ImmACS. ImmACS consists of two parts: the capturing and the reproduction sides (Fig. 1). The capturing side uses a circular microphone array to estimate the DOAs of all active sound sources and separate the source signals. It encodes the soundscape

[∗]This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES-MusiNet.

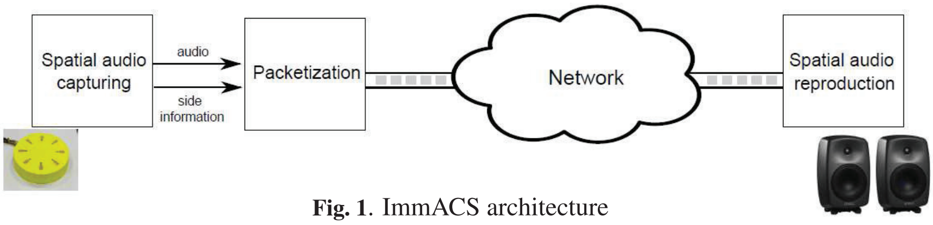

Fig. 1. ImmACS architecture

using one monophonic audio signal and side-information facilitating its use for spatial audio transmission through the Internet. The reproduction side receives the encoded acoustic environment and reproduces it using multiple loudspeakers.

2.1. Capturing side: directional coding of the soundscape

Assume that P active sources are in the far-field of a circular microphone array with M microphones. The microphone array signals are transformed to the Short-Time Fourier Transform (STFT) domain. Then the number of active sound sources and their DOAs are estimated using our previously proposed method of [13, 14], which is capable of estimating the DOAs with high accuracy in reverberant environments. The DOA estimation is applied in each time-frame k and outputs the number of active sources $\hat{P}_k$ and the estimated DOA vector—with 1° resolution—$\boldsymbol{\theta}_k = \left[\theta_1 \cdots \theta_{\hat{P}_k}\right]$.

Based on the estimated DOA vector, we employ a fixed superdirective beamformer in order to separate the source signals that come from different directions. The beamformer is designed to maximize the array gain while maintaining a minimum constraint on the white noise gain [15]. Thus, the beamformer filter coefficients are given by:

$$\mathbf{w}(\omega, \theta_s) = \frac{\left[\epsilon\mathbf{I} + \boldsymbol{\Gamma}(\omega)\right]^{-1} \mathbf{d}(\omega, \theta_s)}{\mathbf{d}^H(\omega, \theta_s)\left[\epsilon\mathbf{I} + \boldsymbol{\Gamma}(\omega)\right]^{-1} \mathbf{d}(\omega, \theta_s)} \quad (1)$$

where $\mathbf{w}(\omega, \theta_s)$ is the $M \times 1$ vector of filter coefficients for frequency ω and steering direction θ_s, $\mathbf{d}(\omega, \theta_s)$ is the steering vector of the array, $\boldsymbol{\Gamma}(\omega)$ is the $M \times M$ noise coherence matrix (assumed diffuse), $(\cdot)^H$ is the Hermitian transpose operation, $\mathbf{I}$ is the identity matrix, and ϵ controls the white noise gain constraint. As fixed beamformers are signal-independent, the filter coefficients can be estimated offline.

In the k-th time frame, we employ $\hat{P}_k$ concurrent beamformers resulting in the beamformed signals:

$$B_s(k, \omega) = \sum_{m=1}^{M} w_m(\omega, \theta_s) X_m(k, \omega), s = 1, \cdots, \hat{P}_k \quad (2)$$

where $X_m(k, \omega)$ is the STFT of the signal recorded at the m-th microphone of the array and $w_m(\omega, \theta_s)$ denotes the m-th component of $\mathbf{w}(\omega, \theta_s)$.

The beamformed signals are given as input to a post-filter. The goal of the post-filter is twofold: it produces the final separated source signals and it allows us to downmix the source signals into one monophonic signal. Based on the beamformed signals, the post-filter estimates $\hat{P}_k$ binary masks:

$$U_s(k, \omega) = \begin{cases} 1, & \text{if } s = \arg\max_p |B_p(k, \omega)|^2, p = 1, \cdots, \hat{P}_k \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

According to (3), for each frequency element the post-filter keeps only the source with the highest energy (i.e., the most dominant) and sets all the other sources at that frequency element to zero. Thus, the masks are orthogonal to each other, meaning that for each frequency element only one source is maintained while the other sources are set to zero. Each binary mask is applied to its corresponding beamformed signal to yield the final separated source signals:

$$\hat{S}_s(k, \omega) = U_s(k, \omega) B_s(k, \omega), \quad s = 1, \cdots, \hat{P}_k \quad (4)$$

Finally, the orthogonality property of the binary masks, allows us to efficiently downmix all the source signals into one full spectrum signal by summing them up. Hence, one audio signal and side-information—consisting of the DOA of the source that is dominant in each time-frequency element—are used to encode the soundscape. In [12] we demonstrated that it is possible to encode the audio signal with an MP3 encoder without any loss in spatial impression and we also proposed a coding scheme for the side-information channel.

2.2. Reproduction side

On the reproduction side, the downmixed signal is transformed into the STFT domain. Based on the side-information we apply VBAP [16] at each frequency element. A low-bitrate version of ImmACS features the *beamformer cutoff frequency*. The beamformer cutoff frequency defines the frequency up to which directional information is extracted. The frequencies above the beamformer cutoff frequency are reproduced from all loudspeakers after appropriate scaling by the reciprocal of the square root of the number of loudspeakers for energy preservation. This version was shown to be more appropriate for speech applications [12].

3. INCORPORATING MULTIPLE MICROPHONE ARRAYS

ImmACS and other related methods usually assume that the microphone array is placed in the middle of the acoustical environment that is encoded. While this is suitable for applications like teleconferencing where people are located around a room, or recording a music performance where the orchestra is placed in the front area of the microphone array, there are other scenarios where a single array cannot provide sufficient spatial coverage. In such scenarios, the sound sources may be located such that their angular separation is too small for the array to isolate them, or the sources may even be located such that they have the same DOA with respect to the array, making the discrimination of the sources impossible.

For these reasons, we investigate the use of multiple microphone arrays combined with location information about the sound sources in order to isolate them and encode the soundscape. Source isolation is an important aspect, as in order to provide accurate spatial impression each source signal that will be reproduced from a specific direction must not contain interfering sources. Moreover, it enables listeners to “focus” the reproduction on a specific sound source by choosing to reproduce that source only and attenuate all the other sources present in the soundscape.

On the recording side, multiple arrays are placed to monitor the area. Assuming that the locations of the sources are known—or can be estimated for example using our work in [17] by fusing DOA estimates from the different arrays—each microphone array can calculate the DOAs of the sources with respect to that array by:

$$\theta_{n,s} = \arctan \frac{p_{\mathrm{y},s} - q_{\mathrm{y},n}}{p_{\mathrm{x},s} - q_{\mathrm{x},n}} \qquad (5)$$

where $\theta_{n,s}$ is the DOA of the s-th source with respect to the n-th microphone array, $\mathbf{p}_s = \begin{bmatrix} p_{\mathrm{x},s} & p_{\mathrm{y},s} \end{bmatrix}^T$ and $\mathbf{q}_n = \begin{bmatrix} q_{\mathrm{x},n} & q_{\mathrm{y},n} \end{bmatrix}^T$ are the locations of the s-th sound source and the n-th microphone array respectively. We also assume that the microphone arrays are connected to a central node that carries the spatial audio capturing operations, providing synchronized signals. We will try to address the following question: what is the best policy for microphone array selection so as to achieve the best source isolation for reproduction?

3.1. Beamforming and post-filtering from the closest array for each source

As the locations of the sources are known—or estimated—a natural approach would be to isolate each source using the closest array to the source, as it is expected that this array would have the highest Signal-to-Noise Ratio (SNR) for the source of interest. This approach works in the following way:

1. The microphone array closest to a source is selected, based on the source’s location.
2. The DOAs of all the active sources to that array are found via (5) so as to perform beamforming and post-filtering through (2)–(4) using the signals from that array.

From the $\hat{P}_k$ final separated source signals only those of the sources that are closest to that array are maintained, while the separated signals of the other sources are discarded, as they will be estimated from the array that is closest to them. Finally, each microphone array will contribute with the separated signals of the sources that are closest to it.

In this scheme, each microphone array estimates its own post-filter. Thus, the binary masks are no longer orthogonal which does not allow the encoding of the soundscape in one audio signal. Moreover, each array has to beamform to all sources—in order to estimate and apply the post-filter—even though only the closest ones are maintained. As a result, unnecessary beamforming operations are carried out and the computational complexity increases proportionally to the number of microphone arrays. An important problem may arise when the sources are far apart but at a small angular separation with respect to an array. As the post-filter compares energies and energy decreases with distance, the array aiming to separate its closest source will provide poor beamformed signals for the sources that are far away—and act as interferers—degrading the source isolation performance.

3.2. Beamforming and cooperative post-filtering

An alternative approach is to allow the microphone arrays to cooperate in order to design a single post-filter that separates all source signals. In this scheme, each microphone array remains responsible for the sources that are closest to it, but it does not individually estimate its own post-filter. This approach works in the following way:

1. Based on the sources’ locations, the closest microphone array for each source is selected and the DOA for that source with respect to that array is calculated using (5).
2. In contrast to the method in Section 3.1, each array beamforms only to the sources that are closest to it using (2).
3. The beamformed signals $B_s(k,\omega),\ s = 1, \cdots, \hat{P}_k$ that now come from different arrays are used to estimate a single post-filter using (3).
4. The final separated signals are estimated via (4).

This scheme is more computationally efficient than the one of Section 3.1, as for $\hat{P}_k$ number of sources only $\hat{P}_k$ beamforming operations are needed. Moreover, as a single post-filter is used, the orthogonality property holds, which allows ImmACS to encode the entire soundscape into one monophonic audio signal and side-information. Note that, as the locations of the sources are known, the side-information can contain the locations—and not DOAs only—of the sources. Our previously proposed encoding scheme for the side-information channel in [12] can also support the encoding of location information. Finally, this approach is expected to perform better isolation, as the beamformed signals that take part in the post-filtering stage are all beamformed from the closest array (i.e., with the highest SNR) in contrast to the method of Section 3.1.

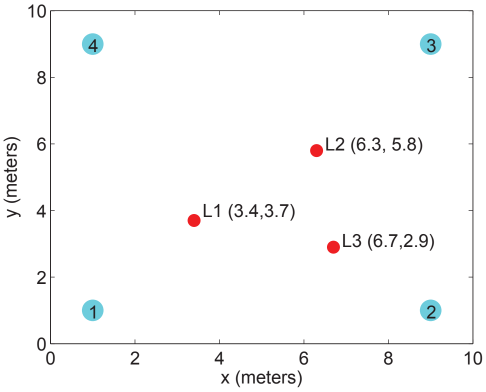

Fig. 2. Microphone array placement (blue circles, numbered 1–4) and locations of active sound sources (red circles) used for the listening test.

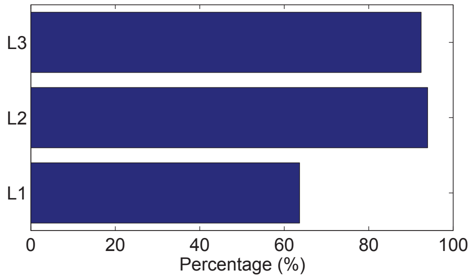

Fig. 3. Preference test results that indicate the percentage of listeners that preferred the method of Section 3.2 over the method of Section 3.1 for the three test locations.

4. RESULTS

In order to evaluate the source isolation performance of the two methods described in Sections 3.1 and 3.2 we performed a listening test. The test scenario is described in Fig. 2 and consists of three simultaneously active sources at locations L_1, L_2, and L_3. In a room of dimensions $10 \times 10 \times 3$ meters there are $N = 4$ circular microphone arrays at locations $(1, 1), (9, 1), (9, 9), (1, 9)$ meters. Each microphone array has a radius of 2 cm and consists of $M = 4$ omnidirectional microphones. The DOAs of the sources at the three locations with respect to the 4 microphone arrays are shown in Table 1. Note that the sources are located close together in terms of angular separation with respect to all arrays (Table 1) making the source isolation problem quite challenging.

We used the image-source method [18] to produce simulated signals of omnidirectional sources in a room with reverberation time $T_{60} = 0.4$ seconds. The signals were processed using frames of 2048 samples with 50% overlap, windowed with a von Hann window. The FFT size was 4096. The approaches of Sections 3.1 and 3.2 were used in order to isolate the three source signals. The experiment was repeated 6 times with different speakers at locations L_1, L_2, and L_3 (Fig. 2), resulting in 18 isolated source signals for each method.

Table 1. DOAs for the source locations used in the listening test with respect to each microphone array.

	L_1	L_2	L_3
Mic. array 1	48°	42°	18°
Mic. array 2	154°	119°	140°
Mic. array 3	223°	229°	249°
Mic. array 4	294°	328°	313°

We employed a preference test, where listeners used headphones to listen to the reverberant source signal of the target source and the output of the two methods (Section 3.1 and 3.2) and they were asked to indicate which method of the two they preferred in terms of speech quality, intelligibility, and source isolation (always comparing to the original reverberant source). The samples were randomized and the subjects did not know to which method they belonged. Eleven volunteers participated in the listening test (authors not included).

Fig. 3 shows the percentage of listeners that preferred the beamforming with cooperative post-filtering approach of Section 3.2 for each location. It is clear that this approach outperforms the method of Section 3.1. The cooperative post-filtering approach results in better source isolation and maintains better speech quality and intelligibility, while keeping all the attractive properties for downmixing into a single audio signal and being computationally efficient (the same number of beamforming operations as in the standard ImmACS with one array of Section 2 is required).

The binary masks during the post-filtering operation can create musical distortions in the isolated source signals. For spatial audio reproduction, the source signals are played back together albeit from different directions which eliminates the musical distortion. However, when the goal is to “focus” on the source signal from a specific location—attenuating the sources from the other locations—it is particularly important to maintain low distortion in the isolated source signal. To evaluate speech distortion we calculated the Log-Likelihood Ratio (LLR) [19]. Similar to [20], we computed the LLR by comparing the signal of the target source as received at the closest microphone and the methods’ output. Note that, as the reference signal contains reverberant parts, high values of LLR do not necessarily indicate high distortion. However, in this way, we can have a fixed reference signal and compare the LLR values for the two methods [20].

The LLR values, averaged over the different speakers, at target locations L_1, L_2, and L_3 are shown in Table 2. For each speaker, the LLR was computed using 23 ms frames with 75% overlap and a Hamming window. The mean LLR value of each speaker was then computed by taking the mean over the 95% of the frames with the smallest LLR values, as

Table 2. Log-Likelihood Ratio averaged over different speakers for locations L_1, L_2, and L_3 of Fig. 2.

	Method of Sec. 3.1	Method of Sec. 3.2
L_1	0.4080	0.3921
L_2	0.6177	0.4226
L_3	0.5838	0.3724

suggested in [19]. In good agreement with the listening test results, Table 2 shows that the beamforming with cooperative post-filtering method (Section 3.2) maintains lower distortion in the separated signals. It is of note that for the isolated signals at location L_1 both methods have similar distortion values, which can explain the discrepancy in listeners' preference between location L_1 and locations L_2 and L_3 (Fig. 3).

5. CONCLUSIONS

In this paper, we investigated the use of multiple microphone arrays to perform sound source isolation in the context of spatial audio recording and reproduction. We proposed two methods for incorporating multiple microphone arrays to ImmACS—our previously proposed system for real-time spatial audio capturing and reproduction—and discussed the advantages and disadvantages of each method. Listening test results and objective measures for speech distortion show that the beamforming with cooperative post-filtering offers better source isolation and speech quality. The results are encouraging for the use of multiple microphone arrays for spatial audio recording, and warrant further investigation of the performance of these methods in the presence of DOA and location estimation errors and for other types of signals, such as musical instruments.

REFERENCES

[1] J. Breebaart et al., "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status," in *119th Audio Engineering Society Convention*, 2005.

[2] F. Baumgarte and C. Faller, "Binaural cue coding-Part I: Psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509 – 519, Nov. 2003.

[3] C. Faller and F. Baumgarte, "Binaural cue coding-Part II: Schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[4] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, , no. 1, pp. 1305–1322, 2005.

[5] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[6] H. Hacihabiboglu and Z. Cvetkovic, "Panoramic recording and reproduction of multichannel audio using a circular microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, Oct. 2009, pp. 117–120.

[7] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002, vol. 2, pp. II–1781–II–1784.

[8] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 2010.

[9] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Hands-Free Speech Communication and Microphone Arrays,*, May 2008, pp. 37–40.

[10] O. Thiergart, M. Kallinger, G. D. Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Microelectronic Systems*, Albert Heuberger, Gnter Elst, and Randolf Hanke, Eds., pp. 321–329. Springer Berlin Heidelberg, 2011.

[11] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and Reproducing Spatial Audio Based on a Circular Microphone Array," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[12] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Directional coding of audio using a circular microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 296–300.

[13] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. on Audio, Speech, and Lang. Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[14] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit," in *European Signal Processing Conference*, Aug 2012, pp. 2303–2307.

[15] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[16] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[17] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2013, pp. 1–4.

[18] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, Aug 2010.

[19] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms.," *International Conference on Spoken Language Processing (ICSLP)*, vol. 7, pp. 2819–2822, 1998.

[20] M. Taseska and E. Habets, "Spotforming using distributed microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, Oct 2013.