2023 International Conference on Information Technologies (InfoTech-2023)
20-21 September 2023, Bulgaria

Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array

Stephen Stroud, Dr Karl O. Jones, Dr Gerard Edwards, Colin Robinson, Dr David Ellis and Dr Sebastian Chandler-Crnigoj
School of Engineering, Faculty of Engineering and Technology
Liverpool John Moores University
James Parsons Building, Byrom Street, Liverpool, L3 3AF, United Kingdom
S.Stroud@2022.ljmu.ac.uk{ K.O.Jones, G.Edwards, C.Robinson1, D.L.Ellis, S.L.ChandlerCrnigoj }@ljmu.ac.uk

Abstract – **This paper presents a time-delay beamforming audio zoom process that aims to overcome the issue of broken microphones in a video surveillance system. The proposed method uses a reduced number of 13 omnidirectional microphones to capture audio signals and employs time-delay beamforming techniques to enhance the audio signals from a specific grid area defined by the user. The system provides an audio zoom capability to enhance the sound coming from a particular direction, making it useful for video surveillance applications where the user needs to focus on a specific location. The script, developed in MATLAB, includes a polar plot radiation pattern to visualise the beamforming direction. A comparison between the response of the 13 omnidirectional microphone array and the full 16-microphone array after beamforming is also given. Experimental results demonstrate the effectiveness of the proposed system in overcoming the issue of broken microphones in a surveillance system.**

Keywords – **Audio Zooming, Beamforming, Forensic Evidence Gathering, Microphones**

I. Introduction

Audio zooming is a concept that allows users to focus on specific sounds within an auditory scene, akin to the way a camera can zoom in and magnify a particular area within a visual frame. The idea has been around since the 1950s. However, replicating the human brain's ability to filter out unwanted noise remains a challenging problem for technology.[1] In a typical audio environment, interfering sounds can make it challenging to extract the desired signals [2]. To address this issue, a system should be developed that can effectively remove unwanted audio and retain only the desired sounds[3]. This technology would have a wide range of applications, including video surveillance systems and broadcast media.

In a previous study [4], we presented a delay and sum beamforming audio zoom system designed for video surveillance applications. The system used an array of microphones to capture audio signals and isolate and enhance sounds from a user-defined grid area. However, microphone damage and malfunctions can lead to an incomplete or unreliable audio signal, making it necessary to develop a system that can operate effectively with a reduced number of microphones. Building upon our previous work, this paper proposes a system that employs time-delay beamforming methodologies and considers the application of the Minimum Variance Distortionless Response (MVDR) beamforming algorithm to compensate for the malfunction of 3 out of the 16 microphones in the array. With the remaining 13 operational omnidirectional microphones, the system ensures the delivery of dependable and precise audio signals in our surveillance system, aiming to establish a novel audio surveillance system that synchronizes video zooming with beamforming-based audio zooming."

II. Related Work

Olson and Preston [5] introduced the Single ribbon cardioid microphone, which eliminated rear noises and had a frequency-dependent Super-Cardioid response. This microphone was inspired by Cherry's [1] work on "The Cocktail Party Problem" (CPP). Ishigaki *et al.*[6] developed a second-order gradient unidirectional microphone with a frequency response of 100Hz - 10KHz, while Matsumoto and Naono [7] created a stereo-zooming microphone that improved upon earlier mono-zoom efforts. The use of machine learning, notably Computational Auditory Scene Analysis (CASA), was proposed as a means of solving the CPP [8]. Extending the work of Wang and Brown [9], Schultz-Amling *et al.*[10] explored Directional Audio Coding (DirAC) for teleconferencing applications, while Van Waterschoot *et al.* [11] investigated acoustical zooming using a multi-microphone array, without an explicit sound source separation algorithm. Thiergart, Kowalczyk, and Habets [12] suggested that spatial filters were the most effective means of achieving Acoustic Zooming. In a CPP simulation, Christensen et al. [13] used a full-rank Wiener subspace filter with dynamic rank limiting for speech enhancement. At the same time, Wilson [14] observed the similarities between natural human solutions to auditory challenges and audio engineering techniques. Fahim *et al.* [15] demonstrated the effectiveness of utilising sparse arrays for sound separation in complex acoustic environments, highlighting the potential of sparse arrays for beamforming while reducing the number of microphones required, which ties into this research. In the domain of beamforming, different algorithms have been explored with varying levels of complexity and performance. The delay-and-sum method, owing to its simplicity and robustness, has been

widely used in various applications including our previous work [4]. However, as computational power has increased in recent years, researchers have begun exploring more complex methods such as the Minimum Variance Distortionless Response (MVDR). The MVDR method, discussed by Yang, McKay, and Couillet [16], offers the potential for better interference suppression, although it is more computationally intensive and its performance relies heavily on the accurate estimation of the covariance matrix. No combined audio and video zoom alignment system has been developed for a surveillance system, therefore experiments using a delay and sum beamformer with relatively low computational demands is a sensible research avenue.

III. Proposed Research

A. Experimental Setup

A studio space measuring 457.2 cm × 342.9 cm was partitioned into a 3 × 3 grid. Nine speakers were positioned within each grid cell and connected to a nine-channel amplifier. Various pink noise and real-world sounds were played through each speaker, ensuring each speaker could produce a unique sound within its grid cell. The sounds were initiated using a digital audio workstation (DAW) at a sample rate of 48 kHz and with a 24-bit depth.

A square microphone array (70 cm × 70 cm) holding 16 omnidirectional microphones was placed in the centre of the room, 100 cm above the floor. The microphone array captured the soundscape from all directions, with the signals being recorded in a secondary DAW. The resulting sixteen audio files were subsequently transferred to MATLAB for signal processing.

In this study, omnidirectional microphones were chosen for several key reasons. Their ability to capture sound from all directions accommodates unpredictable or mobile sound sources, making them particularly suited to environments with varying sound source locations. In addition, the omnidirectional microphones' ability to capture a broader range of environmental noise facilitates phase cancellation techniques for noise reduction. Furthermore, practical considerations were taken into account: omnidirectional microphones are generally cheaper and lighter than their cardioid counterparts, making them a cost-effective and practical solution given budget and weight constraints. Lastly, their use simplifies the experimental setup, eliminating the need for precise orientation towards the sound source.

B. Data Processing and Analysis in MATLAB

The experimental setup of the room for the audio zooming experiment was represented and visualised in MATLAB. The Euclidean distance between each speaker and microphone was computed using the Pythagorean theorem, incorporating z-axis values to enable a 3D visualisation of the experimental setup.

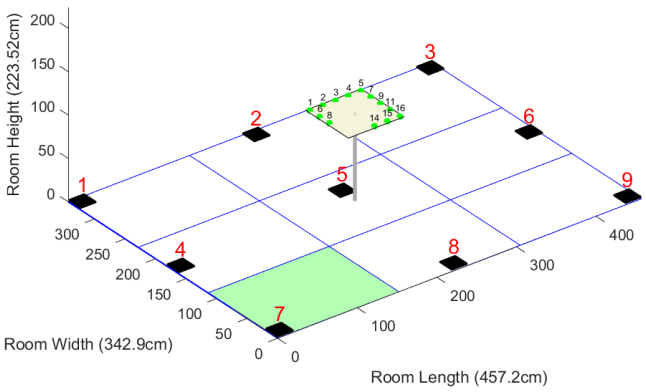

Figure 1. A plot of the user-selected grid seven is highlighted within MATLAB. The user has chosen to steer the beamformer towards grid seven, highlighted in green in the bottom left corner of the figure. Note the central microphone array in grid five with three microphones disabled.

The MATLAB script determined the distance between microphones, the resulting time delay between the microphones and speakers, and the average time delay for any speaker signal reaching each microphone. The time delay between each pair of microphones was also calculated. The script proceeded to load the audio files from the experiment for all active microphones and plot them individually. The MATLAB script prompted the user to select a grid number to isolate and emphasise the chosen grid in the room plot based on user input.

C. Beamforming with Reduced Microphone Array

To assess the effectiveness of the beamforming algorithm with a reduced microphone array, three of the sixteen microphones were disabled in MATLAB, simulating microphone failure.

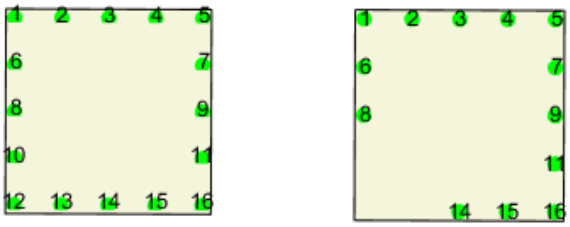

Figure 2. Comparison diagrams of the full microphone array with all sixteen working microphones (left-hand side) and the reduced array with only thirteen working microphones (right-hand side).

The audio data from the remaining microphones were then processed by a time-delay beamforming script to isolate the sound originating from the chosen grid in the room. The beamformer was applied to the audio, and the resulting audio was normalised and stored as a single-channel PCM audio file.

$$S_{Out}(t) = \sum w_i S_{In}(t - \tau_i) \qquad (1)$$

$S_{out}(t)$ denotes the beamformed output signal at time t, w_i represents the weight applied to the first microphone

signal, S_{In} (t-τ_i) is the input signal from the microphone delayed by τ_i, and τ_i is the time delay for the microphone, calculated based on the distance between the microphone and the speaker, as well as the speed of sound. The sum in equation (1) is computed over all 16 microphones in the array.

Although our time-delay beamforming algorithm (1) doesn't require explicit detection and switching off of failed microphones, as non-operational channels naturally do not contribute to the beamforming process, we acknowledge potential complications arising from failures causing decreased sensitivity or increased self-noise. In such cases, a failure detection mechanism could be beneficial. Further investigation into this could be a valuable direction for future research

The study also investigated the MVDR beamforming algorithm, known for managing multiple interferences. A comparison between MVDR and the delay-and-sum methods helped identify the most effective under varying conditions.

D Comparison and Cross-Correlation Analysis

A plot was created to showcase the waveforms of the reduced microphone array (with failed microphones) after beamforming, against the full array after beamforming. The beamformed audio signals from the full and reduced arrays will be plotted and compared. Additionally, normalised cross-correlation was calculated between the beamformed audio signals from the reduced and full microphone arrays. The cross-correlation was then plotted to help assess the similarity between the two signals and understand the impact of microphone failures on the beamforming algorithm's performance.

In summary, the methodology of this study focuses on evaluating the performance of a time-delay beamforming algorithm in the presence of microphone failures. The results are visualised and compared to assess the robustness of the beamforming technique with a reduced microphone array.

IV. Results Discussion

The following section presents the results of the audio zooming experiment using time-delay beamforming. Once the three microphones were disabled in the setup, the script recognised any data with zero amplitude, highlighted it in red, and removed it from the beamforming algorithm. The average time delay between the microphones on the array and the target speaker was calculated at 8.5ms.

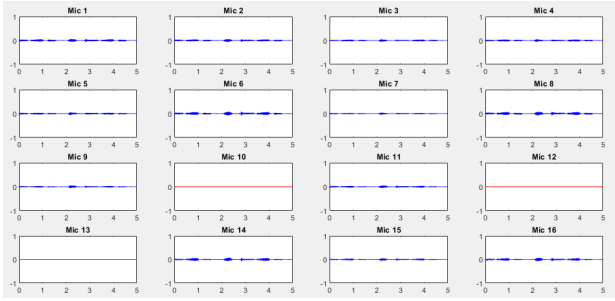

Figure 3. Plots of the sixteen microphone waveforms. The script identifies any signal with zero amplitude, indicating a broken microphone, highlights it in red, and removes it from the algorithm.

A. Polar Plot Analysis

The beamforming pattern for the selected grid was displayed using a polar pattern plot. Figure 4 shows the polar pattern plot of the combined response of the 16 omnidirectional microphones before beamforming occurred, compared to the response of the reduced array with only 13 microphones and the beamforming pattern steered towards the selected grid seven for both the 13 and the 16 mic array.

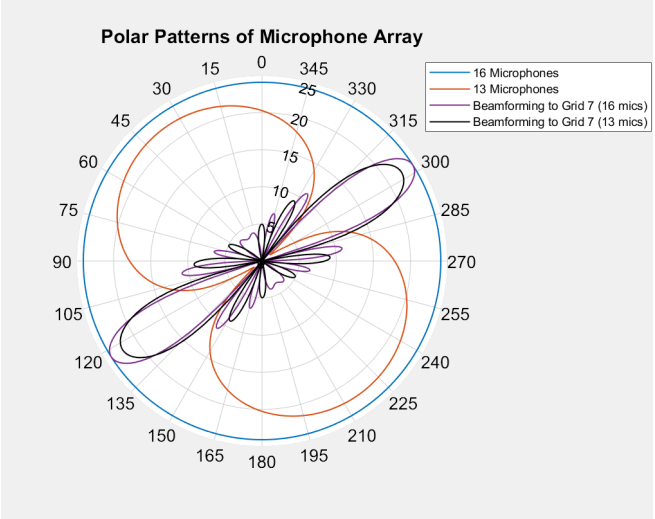

Figure 4. The polar plots of the combined polar responses of the array with 16 microphones (blue) and 13 microphones (orange) and then the beamform pattern steered to grid 7 with 16 microphones (purple) and 13 microphones (black).

The polar pattern plots in this study were generated using MATLAB's 'polarpattern' function and custom algorithms to simulate the response of the microphone array rather than relying on direct measurements. The polar plot indicates that the reduced array maintains enough combined response to support a steered beamforming pattern. Although the pattern appears slightly different compared to the entire array, the overall response is preserved, suggesting the system can still effectively focus on the desired grid in case of a microphone failure.

B. Beamforming Results and Analysis

The time-delay beamforming was applied to the audio data to isolate the sound from a specific grid in the room; in

this example, speaker seven in grid seven of the room was used. The beamforming algorithm delays and sums the microphone signals to emphasise the sound coming from a specific direction while reducing noise and interference from other directions.

Figure 5 shows the normalised beamformed audio in the time domain for the selected grid, with the results for the full 16-microphone array and the beamformed output for the reduced 13-microphone array.

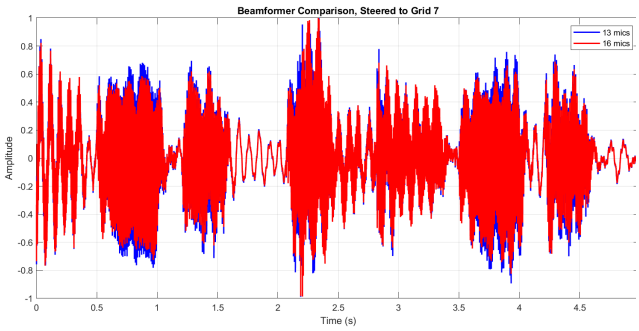

Figure 5. Comparison of the beamformed signals with all sixteen microphones (red waveform) and the reduced array with only thirteen microphones (blue waveform) steered to grid seven.

The script also plotted a side-by-side comparison of the time-domain waveforms after normalisation, revealing that even with a reduced number of microphones near the desired target area, the array can still achieve similar results after beamforming and normalisation of the signal.

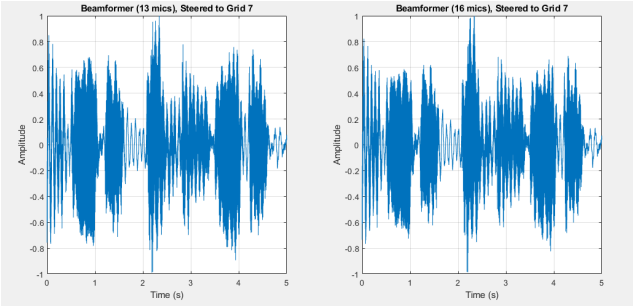

Figure 6. Side-by-side comparison of the beamformed signals' waveforms. The reduced array with only thirteen mics (left-hand side) against the entire array with all sixteen mics (right-hand side) steered to grid seven.

The results demonstrate that the adapted beamforming algorithm, which dynamically adjusts to the 13 functional sensors while excluding the three malfunctioning channels, performs adequately compared to the entire array, with only minor variations in the beamformed signals. These plots suggest that the beamforming system is robust and can maintain performance even when some microphones are lost or malfunctioning by implementing a technique that compensates for the non-operational sensors.

V. Conclusion

In conclusion, this study has demonstrated the resilience of the time-delay beamforming algorithm against microphone failures, maintaining satisfactory performance even when the microphone array is reduced. This is a pilot study with a particular configuration of microphones, with three adjacent microphones 'broken' out of sixteen investigated. This finding is particularly promising for drone surveillance systems, as it suggests that these systems can function effectively under less-than-ideal conditions, such as microphone malfunctions.

The results also highlighted the algorithm's effectiveness in accurately capturing sound from a chosen grid, as evidenced by the average time delay of 8.5ms between the microphones and the speaker and the polar plots illustrating the beamformed audio's directional radiation pattern. This study, therefore, paves the way for further research and development in audio zooming, with potential applications in forensic evidence collection and broadcasting.

Future research should focus on enhancing the beamforming algorithm to minimise noise and interference, further reducing the number of microphones, investigating the influence of microphone placement, and evaluating alternative beamforming algorithms for their robustness under various conditions. Developing a system that aligns audio zoom with video zoom, enabling users to select a grid on a GUI control panel for simultaneous audio and video zooming, would be a novel contribution to the field.

The delay-and-sum beamforming algorithm was selected due to its simplicity, robustness, and computational efficiency. Its straightforward implementation is advantageous in settings with constrained computational resources or where a rapid solution is needed. This method also offers robustness to errors in parameters such as direction of arrival, a crucial asset in real-world applications with challenging parameter estimation. While MVDR can offer superior interference suppression in theory, its practical performance depends heavily on the accuracy of covariance matrix estimation. In situations with limited snapshots or non-stationary noise fields, delay-and-sum beamforming may yield equivalent or superior results. This algorithm also requires less computational resources than MVDR, making it advantageous for real-time applications or systems with limited computational power.

The inclusion of a microphone failure detection mechanism enhances system robustness and reliability by identifying issues before they impact the output signal, ensuring continuous operation despite faults. This is critical in real-world applications where factors like wear and tear or environmental conditions may cause failures over time. Moreover, partial failures that don't result in a complete signal loss could introduce noise or distortion; a detection system can mitigate these effects. The knowledge of functional and non-functional microphones can potentially optimize the beamforming process, as some algorithms allow for coefficient adjustments based on the number and

position of working microphones. This aligns with the research goal of assessing beamforming performance amidst microphone failures, making it essential to include a detection and handling process in the algorithm. Furthermore, this mechanism can prevent cascading failures, enhancing overall system stability.

Ultimately, this research project aims to incorporate a comparable microphone and camera array into a police drone for forensic evidence collection, capitalising on the alignment of audio and video zoom as a novel aspect of the research. The methodology could also be extended to broadcasting applications, broadening this promising audio-zooming technique's potential impact and applicability in various industries.

REFERENCES

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America,* vol. 25, no. 5, pp. 975-979, 1953, doi: 10.1121/1.1907229.

[2] M. Hawley, R. Litovsky, and J. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America,* vol. 115, pp. 833-43, 03/01 2004, doi: 10.1121/1.1639908.

[3] Y. Huang, J. Benesty, and J. Chen, "Speech Acquisition And Enhancement In A Reverberant, Cocktail-Party-Like Environment," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings,* pp. 25-28, 2006.

[4] S. Stroud, K. Jones, O., G. Edwards, C. Robinson, D. Ellis, and S. Chandler-Crnigoj, "Towards a Time Delay Beamforming Audio Zoom for a Video Surveillance System," 2023.unpublished

[5] H. F. Olson and J. Preston, "Single-Element Unidirectional Microphone," *Journal of the Society of Motion Picture Engineers,* vol. 52, no. 3, pp. 293-302, 1949, doi: 10.5594/J12528.

[6] Y. Ishigaki, M. Yamamoto, K. Totsuka, and N. Miyaji, "Zoom Microphone," *The Audio Engineering Society Convention Preprint,* vol. 1713 (A-7), 1980.

[7] M. Matsumoto and H. Naono, "Stereo Zoom Microphone For Consumer Video Cameras," *IEEE Transactions on Consumer Electronics,* vol. 35, no. 4, pp. 759-766, 1989.

[8] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation,* vol. 17, no. 9, pp. 1875-1902, 2005, doi: 10.1162/0899766054322964.

[9] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks,* vol. 10, no. 3, pp. 684-697, 1999, doi: 10.1109/72.761727.

[10] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical Zooming Based on a Parametric Sound Field Representation," *Audio Engineering Society Convention Paper 8120,* pp. 1-9, 2010.

[11] T. Van Waterschoot, W. Joos Tirry, and M. Moonen, "Acoustic Zooming by Multimicrophone Sound Scene Manipulation," *Audio Engineering Society,* vol. 61, 7/8, 2013.

[12] O. Thiergart, K. Kowalczyk, and E. A. P. Habets, "An acoustical zoom based on informed spatial filtering," 2014 2014: IEEE, doi: 10.1109/iwaenc.2014.6953348. [Online]. Available: https://dx.doi.org/10.1109/iwaenc.2014.6953348

[13] K. B. Christensen, M. G. Christensen, J. B. Boldt, and F. Gran, "Experimental Study Of Generalized Subspace Filters For The Cocktail Party Situation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2016.

[14] P. F. Wilson, "Multiple Sources in a Reverberant Environment: The "Cocktail Party Effect"," *Proc. of the 2017 International Symposium on Electromagnetic Compatibility - EMC EUROPE 2017,* 2017.

[15] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 1-3 March 2017 2017, pp. 151-155, doi: 10.1109/HSCMA.2017.7895580.

[16] L. Yang, M. R. McKay, and R. Couillet, "High-Dimensional MVDR Beamforming: Optimized Solutions Based on Spiked Random Matrix Models," *IEEE Transactions on Signal Processing,* vol. 66, no. 7, pp. 1933-1947, 2018, doi: 10.1109/TSP.2018.2799183.