Enhancing Environmental Sound Recognition in Digital Simulations: A Novel Approach to Beamforming and Signal Identification

A Novel Approach to Beamforming and Signal Identification for Police Forensic Evidence Gathering.

Stephen Stroud
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
s.stroud@2022.ljmu.ac.uk

Dr Karl Jones
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
k.o.jones@ljmu.ac.uk

Dr Gerard Edwards
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
g.edwards@ljmu.ac.uk

Colin Robinson
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
c.robinson1@ljmu.ac.uk

Dr Sebastian Chandler-Crnigoj
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
s.l.chandlercrnigoj@ljmu.ac.uk

Dr David Ellis
Applied Forensic Technology Research Group (AFTeR), School of Engineering, Faculty of Engineering and Technology, Liverpool John Moores University
d.l.ellis@ljmu.ac.uk

ABSTRACT

This paper advances the field of environmental sound recognition, presenting a refined approach to beamforming and noise identification through digital simulations of realistic environments at Liverpool John Moores University. Amidst the growing demand for precise audio detection techniques in cluttered acoustic environments, our research introduces a method for identifying and highlighting specific sound signals. We use an advanced time-delay beamforming algorithm to achieve strategic audio zooming, addressing topical issues in urban surveillance and forensic sound examination for potential analysis in criminal cases.

Our methodology is rooted in deploying a carefully configured array of virtual omnidirectional microphones, crucial for collecting real-world audio signals. Our technique's core lies in applying our advanced algorithm to the captured sound data, thoroughly assessing our system's capability to identify and isolate targeted sound sources.

Our investigation further measures the robustness of our system to microphone failure, which continues to function even when microphones completely fail, highlighting its reliability, even when operating under compromised conditions. Through simulations that capture actual acoustic environments, our experiments reveal the algorithm's proficiency in coping with both sound reflections and reverberations, critical elements in authentically reproducing real-world scenarios.

Finally, this study explores the extended applicability of our research findings, considering their potential impact across various sectors, including environmental surveillance, animal conservation, broadcasting, and sound engineering. Our work offers a forward-thinking strategy for environmental sound recognition within a digital simulation framework. It paves the way for practical applications that stand to gain from improved sound separation and analysis techniques. Our contributions engage with the broader dialogue on the evolution of surveillance technology, providing valuable perspectives that could influence the future of audio research.

CCS CONCEPTS

• **Hardware** → Communication hardware, interfaces and storage; Signal processing systems; Beamforming; Communication hardware, interfaces and storage; Signal processing systems; Noise reduction; • **Computing methodologies** → Modeling and simulation; Simulation evaluation; • **General and reference** → Cross-computing tools and techniques; Experimentation.

KEYWORDS

Audio Zooming, Surveillance, Forensic Evidence Gathering

1 INTRODUCTION

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

In a previous study [4], we presented a real-world robust beamforming audio zoom system for video surveillance applications. The system used an array of omni-directional microphones to capture audio signals and then isolate and enhance sounds from a user-defined grid area. The experiment employed time-delay beamforming methodologies to compensate for the malfunction of 3 out of the 16 microphones in the array.

Building upon our previous work, this paper introduces a simulation system that enables a choice of virtual multi-shaped arrays and flexibly sized scenes and dimensions based on real-life experimental housing in Liverpool John Moores University, England, United Kingdom. The system employs time-delay beamforming methodologies that consider the multiple reflections of audio signals and introduce an effective noise-cancellation system. The MATLAB program ensures the delivery of dependable and precise audio signals in our virtual surveillance system, aiming to establish a novel audio surveillance option that could also be synchronised with video.

2 LITERATURE REVIEW

Early steps towards a machine capable of ”zooming” into an audio scene were taken by Olson and Preston [5] when their Single ribbon cardioid microphone was shown to eliminate sounds to the rear. Their microphone displayed an increasingly Super-Cardioid response, depending on the speed of the recorded frequency (10kHz signals introduced a more significant super-cardioid response from the microphone than 1kHz signals).

The super-cardioid microphone was obviously influenced by the work of Cherry [1], who first mentioned ”*The Cocktail Party Problem*” (CPP) when the idea of a machine capable of solving the problem was first introduced.

The Cocktail Party Problem describes the difficulty of isolating and understanding a specific human voice amidst numerous competing sound sources in a noisy environment. While the human brain accomplishes this task effortlessly, it remains a significant challenge for technological systems.

It was not until 1980 that the idea of an actual audio zoom, which attempted to replicate the functions of the recently developed video zoom, was introduced. A Second Order gradient unidirectional microphone as developed for JVC^{TM} by Ishigaki *et al.* [6], with a frequency response of 100Hz - 10KHz and a unidirectional polar pattern. Their experiment was reliant upon a well-matched pair of electret microphones to work.

Matsumoto and Naono [7] developed a stereo-zooming microphone to improve the super directional aspect of previous experiments by manipulating psychoacoustics, essentially recording in stereo, to create a feeling of ’spaciousness’.

This stereo zoom improved upon the earlier mono zoom but was still limited compared to a conventional video zoom.

The research into Cherry's CPP continued in multi-disciplined fields into the 21st century, attempting to answer the question, ”Is it possible to build a machine capable of solving it *[The Cocktail Party Problem]* [8] in a satisfactory manner?” The authors citing the earlier work of Wang and Brown [9], concluded that by using Machine learning, and specifically Computational Auditory Scene Analysis (CASA), it could be possible to build a computational model of the auditory scene and then automatically extract and track a sound signal.

Schultz-Amling *et al.* [10] published an acoustic zooming paper that examined directional audio coding (DirAC). DirAC is a parametric approach to the analysis and reproduction of spatial sound. The parameters dealt with were the Direction of Arrival (DoA) and the diffuseness of the sounds. While this research was undertaken with teleconferencing in mind, their aim of the video automatically steering to the active talker could potentially have useful parallels with drone audio and video zoom alignment. Van Waterschoot *et al.* [11] continued the line of work on Acoustical zooming using a multi-microphone array. The paper contained a general theory for independent sound source level control, which could be exploited for an acoustic zoom effect. An important feature of the research was that, unlike the previous papers, it did not consist of an explicit sound source separation algorithm, which relieved it of extreme computational requirements. Instead, it used spatial and spectral noise reduction algorithms and was tested using a small number of low-cost microphones attached to an array. The Acoustic Zoom (AZ) effect relies on altering one or more acoustic cues related to the auditory perception of sound source distance. Several factors determine auditory distance perception, the main one being sound intensity. Other factors include Direct to Reverberant Ratio (DRR), which measures the relative levels of direct sound to reflected sound in an acoustic environment—spectral distortion results from an alteration of a signal's frequency content. Interaural Time Difference (ITD) is the difference in the arrival time of sound between two ears, while Interaural Level Difference (ILD) is the difference in the sound pressure level that reaches each ear.

Motion-induced intensity rate of change refers to the variation in sound intensity perceived as a result of the relative motion between the sound source and the listener. This approach was particularly suited to audio-visual capture applications. Thiergart, Kowalczyk and Habets [12] concluded that Acoustic Zooming was best achieved using spatial filters. Similarly, a full-rank Wiener subspace filter with dynamic rank limiting was used for speech enhancement in a CPP simulation by Christensen *et al.* [13] and performed better than the classical Wiener filter approach for noise and distortion reduction. Wilson [14] found that humans naturally increased the Signal to Noise Ratio (SNR) in CPP situations by taking advantage of their binaural hearing, interaural time differences (Time of arrival of signal), interaural level differences (amplitude) and interaural decorrelation (Coherence of speech). The observation that natural human solutions to sound problems are similar to gain, pan and audio engineering techniques supports Cherry's original theory. To date, to the author's knowledge, no audio and video zoom alignment has been developed for a surveillance system, so choosing to begin simulation studies with a time-delay beamformer, which would be computationally low, is a sensible initial step.

3 METHODOLOGY

3.1 Simulation Environment Preparation

In this study, we developed a modular MATLAB-based simulator to enhance environmental sound recognition through innovative beamforming and noise identification techniques. The methodology is structured as follows:

The simulator input quantities consist of the actual experimental dimensions, including the length, width, and height of the grounds surrounding the 'Exemplar Houses' in Liverpool John Moores University, to establish an accurate spatial framework essential for realistic acoustic scenario modelling. The dimensions for the houses and surrounding grounds were obtained through on-site measurements and blueprints obtained from the local government planning permission website [15]. An area of 80 meters (X-axis) by 50 meters (Y-axis) with a height of 50 meters (Z-axis) was selected, incorporating the experimental housing and the surrounding grounds.

A three-dimensional scene was generated using basic 3D shapes in a function, constituting a 'World Objects' structure. This structure contains crafted realistic building constructions and materials, incorporating real-life Sabine acoustic coefficients, which are values that represent the amount of sound absorbed by a material in a room. Incorporating these Sabine acoustic coefficients into the simulation helps to mimic real-world environments, a critical component for authentic environmental sound simulation.

The user is then prompted to specify the location of the microphone array, choosing between a default central position or custom coordinates within the scene. Users can then select the shape of the microphone array (Square, Circle or Cross) and turn the individual omnidirectional microphones within the array on or off.

3.2 Noise Reduction Array Integration

A specialised microphone array designed for noise reduction is added on top of the main array, created with the assumption that a police surveillance drone could carry the system, so removing the local noise would be necessary to enhance signal clarity. Harrison et al. [16] stressed that audio forensic data is frequently masked in surplus noise; therefore, applying noise reduction techniques to detect speech effectively seemed reasonable.

3.3 Grid Selection, Speaker Placement and Sound Wave Simulation

The user selects the grid size for the experimental setup, ranging from full-sized (40m x 60m) to custom dimensions. This choice updates the visual representation of the 3D scene. The coordinates for the placement of virtual speakers within the experiment are determined by equally positioning eight sound sources around the grid's perimeter, with the ninth sound source (Speaker 5) located in the scene's centre or through custom coordinates. Speaker number 5 is turned off to avoid overdriving the array. Users can also customise sound wave parameters for each speaker, such as azimuth and elevation angles and sound pressure levels, to better simulate realistic acoustic environments. Each speaker's Sound Profile data was chosen based on standardised 80 dB SPL at 1 meter and typical speaker placement angles for uniform coverage. Azimuth and elevation angles represent various audio source positions, and the 10° beam widths standardise the directional characteristics across all speakers. (See Table 1 below).

Virtual sound waves are produced, interacting with the environment, as governed by the Sabine coefficients and attenuating by the inverse square law. Users then choose a grid segment (1 to 9) for beamforming, as illustrated below in Figure 1, an example of a 3D scene created in our MATLAB simulator.

3.4 Beamforming and Noise Reduction

Multiple calculations for distances and delays between speakers, microphones, and within the microphone array are performed to model sound propagation and reception accurately.

An interactive legend is created for dynamic interaction with Figure 1, enhancing the user's ability to interpret the simulation data. Virtual or real-world audio signals are incorporated, enriching the simulation with diverse acoustic environments. A scenario involving a male voice in Speaker 1 and music files in the remaining speakers, supplemented with drone noise bleeding into the microphones, serves as a test case.

Time delay beamforming focuses the array on precise sound sources within the chosen grid segment (Grid 1 in this example) by applying the following formula.

$$s_{Out}(t) = \quad w_i s_{in}(t - r_i) \tag{1}$$

$S_{out}(t)$ is the beamformed output signal at time t, while w_{i} represents the weight applied to the first microphone signal from the array, S_{In} $(t\text{-}r_i)$ is the input signal from the microphone delayed by $r_i = l_i / v_s$ which is the time delay for the microphone signal, calculated based on the distance between the microphone and the speaker (l_i) and the speed of sound (v_s). This process is applied to all the microphones on the array.

A subtractive noise reduction technique is employed to refine the beamformed signal using the feedforward microphone input. A textual summary of experimental outputs is provided to outline the key findings of the experiment, correlating each figure with specific aspects of the simulation and analysis process. This methodology leverages modular coding practices in MATLAB to create a flexible and comprehensive framework for studying environmental sound recognition. The developed simulator offers advancements in digital acoustic simulations through detailed, flexible, user-friendly setup options, dynamic 3D modelling and incorporates sophisticated signal processing techniques.

The MATLAB implementation of the beamforming algorithm utilised the Audio Toolbox and the Phased Array System Toolbox. Specifically, the 'phased.ULA' [17] was used to create and configure the sixteen microphone array, and the 'phased.TimeDelayBeamformer' [18] applied the appropriate beamforming techniques.

Table 1: Sound Profile for each Speaker

Speaker	dB$_{SPL}$ at 1 meter	Azimuth Angle (°)	Elevation Angle (°)	Horizontal Beam Width (°)	Vertical Beam Width (°)
1	80	210	10	10	10
2	80	180	10	10	10
3	80	150	10	10	10
4	80	270	10	10	10
5	0	90	10	10	10
6	80	45	10	10	10
7	80	330	10	10	10
8	80	0	10	10	10
9	80	30	10	10	10

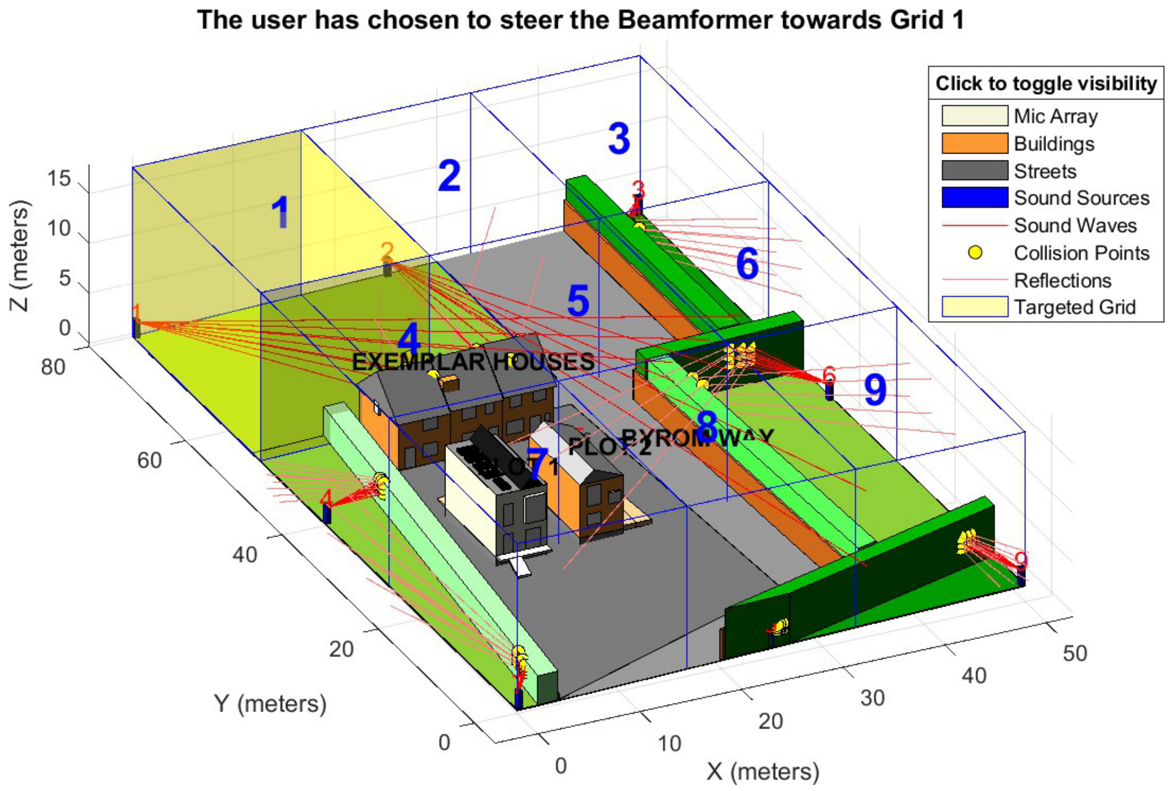

Figure 1: The plot of the 3D simulation includes a full grid, microphone array, speakers, soundwaves and reflections.

4 RESULTS

4.1 Sound Capture by the Virtual Array

The 16 virtual omnidirectional microphones showcased the ability of the simulator to capture audio mixtures that closely emulate a real-world scenario, incorporating speech, music, and drone noise elements. The underlying algorithm considered several critical parameters for each sound source within the experiment, including the initial Sound Pressure Level (SPL) at 1 meter, azimuth angle, elevation angle, and horizontal and vertical beam width. This nuanced approach allowed for a sophisticated capture of the composite sound environment, factoring in the complexities of distance, reflections, absorption properties of various materials, and reverberation. The implementation enforces the inverse square law to ensure a realistic attenuation of sound intensity with distance. This culminated in recording a 5-second audio file into a MATLAB array for each microphone, faithfully representing the intricate interplay of various audio components. This precision in capturing the acoustic environment underscores the effectiveness of the virtual microphone array setup in simulating real-world audio dynamics within the digital simulation framework. An example of the captured audio waveforms for each microphone is given in Figure 2.

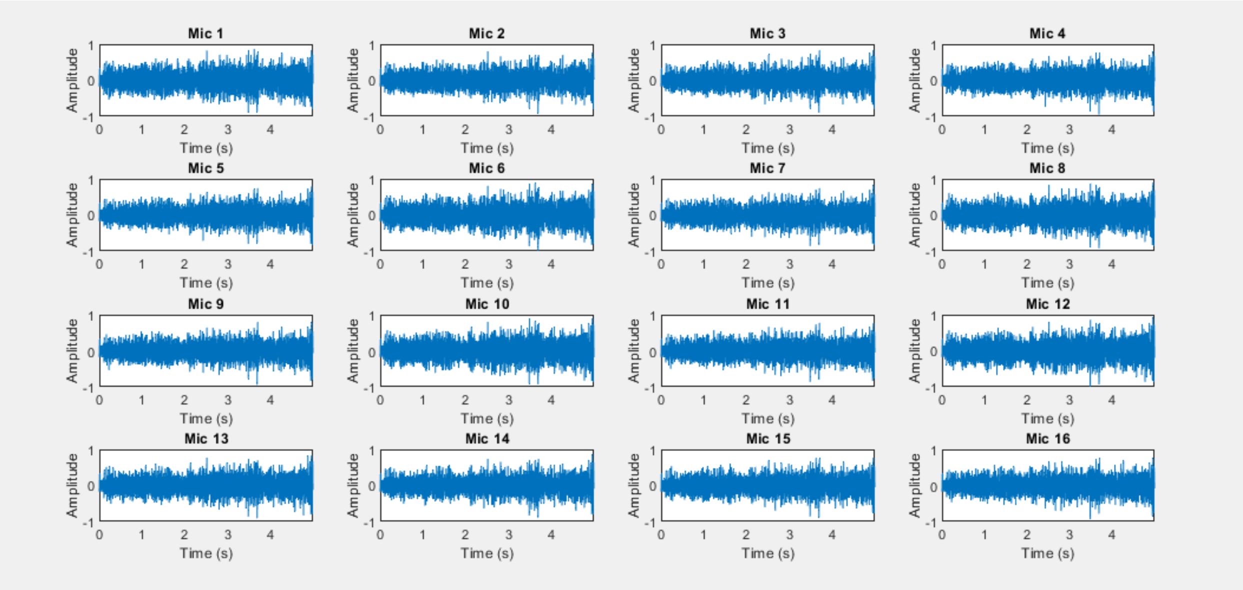

Figure 2: Real-world sound mixtures were captured by the 16 virtual microphones from within the 80 meters (X-axis) by 50 meters (Y-axis) by 50 meters (Z-axis) simulation space.

4.2 Beamforming Results

The beamforming results demonstrate the effectiveness of the time-delay beamforming algorithm. Utilising the signals captured from the 16 virtual omnidirectional microphones, the algorithm adeptly merged and directed these inputs towards a specifically chosen grid direction (Grid 1, in this example). This precision steering of audio signals demonstrates the algorithm's capability to selectively focus on desired sound sources, amplifying sounds within the target area while attenuating those outside the directed polar pattern. The beamforming implementation employing the Phased Array System Toolbox demonstrated effective spatial filtering of the audio signals. The algorithm enhanced the clarity and intensity of the sounds of interest and significantly reduced background noise and irrelevant audio elements. The processed beamformed audio, encapsulating a 5-second duration example, was recorded and analysed both in the time and frequency domains. The visual analysis was complemented by the auditory playback feature in MATLAB, which was provided by the MATLAB audio toolbox and offered end users quick results. Through this innovative application of beamforming, we have successfully demonstrated a refined method for audio signal manipulation, paving the way for more focused and clear acoustic captures in digital simulations.

4.3 Noise Reduction on the Beamformed Audio

Upon achieving targeted audio capture through beamforming, the simulation experiment applied a subtractive noise reduction algorithm, significantly refining the clarity of the resultant sound. This algorithm utilises the virtual feedforward microphone from the noise reduction array. The virtual feedforward microphone sits on top of the original microphone array and is specifically designed to capture ambient noise elements, such as the drone blades' whirring and wind disturbances. By recording these specific noise profiles, the algorithm effectively isolates and subtracts them from the beamformed audio. This process effectively reduces background noise, thereby enhancing the overall sound quality. The impact of this noise reduction is most notable in the improved legibility of human speech captured within grid 1. The Subsequent subtraction of the extraneous noise elements markedly amplifies the clarity and intelligibility of the human voice, which lives exclusively in the mid-range of the frequency spectrum. This stage underscored the effectiveness of subtractive noise reduction techniques in improving audio recordings for surveillance purposes. This particularly applies to scenarios where foreground speech must be disentangled from pervasive background noise. This process is exemplified in Figure 3, where the noise can be seen to be decreased using our noise reduction technique.

5 CONCLUSION

This study has successfully demonstrated via simulation the efficacy of a new system capable of accurately steering captured audio from a robust microphone array towards a predetermined grid. The new system also effectively reduced unwanted noise from the audio signal. This enhancement significantly elevates the probability of comprehending human speech amidst background noise within captured audio signals. Building upon the foundations of previous research, the robustness of this system makes it suitable for field application. The ability to isolate and clarify the spoken words occurring in a particular grid within a noisy environment is testimony to the system's advanced capabilities in audio processing.

Future work will focus on refining the beamforming algorithm further. The time delay beamforming algorithm was selected for its simplicity, robustness, and computational efficiency. Its straightforward implementation is advantageous in settings with constrained computational resources or where a rapid solution is needed. While

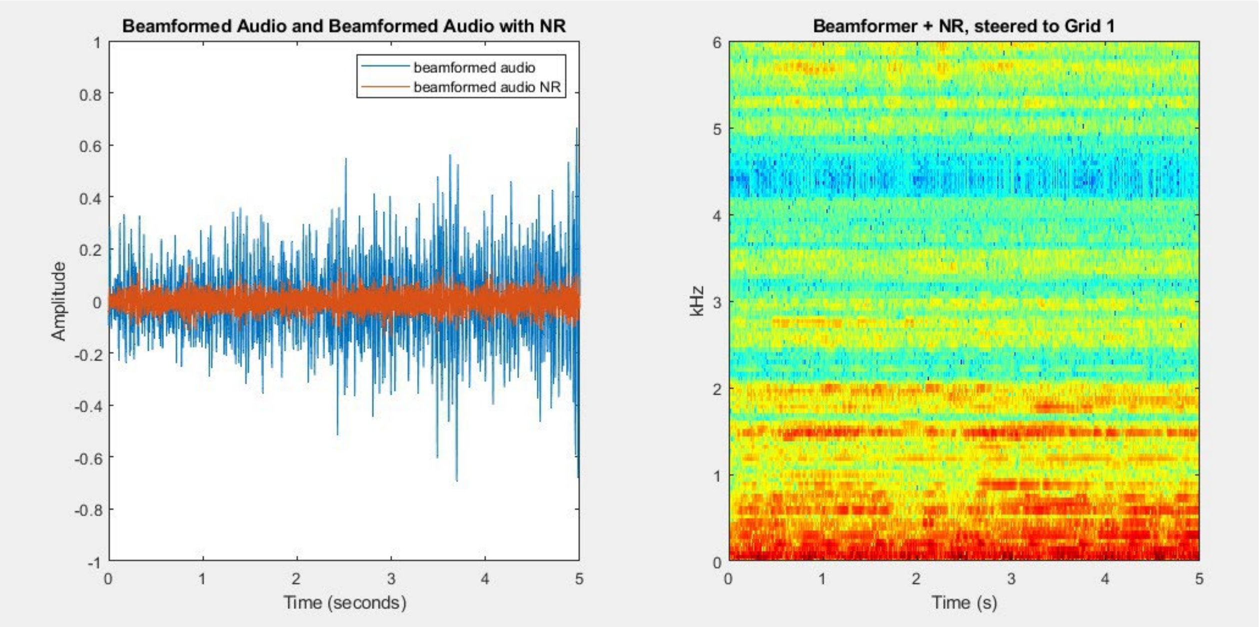

Figure 3: Time and Frequency domain plots of the noise reduction algorithm applied to the beamformed audio steered to Grid 1 within the 80 meters (X-axis) by 50 meters (Y-axis) by 50 meters (Z-axis) simulation space.

minimum variance Distortionless response (MVDR) can offer superior interference suppression in theory, its practical performance depends heavily on the accuracy of covariance matrix estimation. Time delay beamforming may yield equivalent or superior results in situations with dynamic noise fields. The aim is to diminish noise and interference to even lower levels, enhancing sound clarity and separation. Ultimately, the research aims to achieve a level of audio quality and distinction that renders the technology invaluable for forensic surveillance and broadcasting applications. This application goal is the driver for the pursuit of novel solutions to complex audio challenges, pushing the boundaries of what is achievable in sound recognition and separation.

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. O. Jones, G. Edwards, C. Robinson, D. Ellis, and S. Chandler-Crnigoj, "Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array," presented at the 37th International Conference on Information Technologies (InfoTech-2023), Bulgaria, 20-21 Sept. 2023, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10266894.

[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] Gov.UK. "Planning Applications." https://lar.liverpool.gov.uk/planning/index.html (accessed 16.05.2024, 2024).

[16] O. Harrison, K. O. Jones, J. Reed-Jones, C. Robinson, and K. Morrisson, "The Effect of Noise Reduction Upon Voiceprint Integrity.," presented at the International Conference on Intelligent Systems and New Applications (ICISNA'23), Liverpool, England, 2023.

[17] Mathworks. "Phased.ULA." https://uk.mathworks.com/help/phased/ref/phased.ula-system-object.html?searchHighlight$=$phased%20ula&s_tid$=$srchtitle_support_results_1_phased%2520ula (accessed 22.05.2024, 2024).

[18] MathWorks. "phased.TimeDelayBeamformer." https://uk.mathworks.com/help/phased/ref/phased.timedelaybeamformer-system-object.html?searchHighlight$=$phased%20beamformer&s_tid$=$srchtitle_support_results_3_phased%20beamformer (accessed 22.05.2024, 2024).