Advancing Audio Surveillance in Simulated Environments: Real-World Soundscapes and Targeted Noise Detection through Enhanced Beamforming Techniques

Stephen Stroud[a], Karl Jones[a], Gerard Edwards[a], Colin Robinson[a],
Sebastian Chandler-Crnigoj[a], David Ellis[a]

[a]Applied Forensic Technology Research Group (AFTeR),
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@ljmu.ac.uk, g.edwards@ljmu.ac.uk, c.robinson1@ljmu.ac.uk,
s.l.chandlercrnigoj@ljmu.ac.uk, d.l.ellis@ljmu.ac.uk

ABSTRACT

This paper introduces an innovative beamforming approach designed for audio surveillance, executed through a virtual simulation of a real-world environment based at Liverpool John Moores University. Our research is driven by the increasing requirement for sophisticated audio analysis methods to isolate and enhance specific sounds within noisy environments for forensic analysis, for example, in criminal court cases. By leveraging a time-delay beamforming algorithm, our work offers a novel solution to discern and amplify targeted noises amidst complex soundscapes, a challenge commonly encountered in urban surveillance and forensic audio analysis. Our approach's foundation lies in utilising a carefully arranged, robust array of omnidirectional microphones, which are instrumental in capturing a wide range of real-world sound signals. The core of our methodology involves processing captured sounds using the proposed algorithm, followed by evaluating the system's effectiveness in capturing the desired localised audio sources.

This paper explores the system's resilience against microphone array degradation, showcasing its robustness in scenarios of partial system functionality. The experiments, grounded in the simulation of real-world acoustic environments, demonstrate the algorithm's adeptness at managing sound reflections and reverberation, critical factors in the realistic replication of urban soundscapes. This paper also considers the broader implications of our findings, exploring the potential for adopting this technology in various domains beyond law enforcement, including broadcast solutions, advanced audio engineering applications, and animal conservation in the wild.

In conclusion, this research showcases a creative approach to audio surveillance and opens the door to numerous applications that can benefit from enhanced audio isolation and analysis. Our findings contribute to the ongoing discourse on developing advanced surveillance technologies, offering insights that could help shape the future of audio processing and analysis.

KEYWORDS: *Audio Zooming, Surveillance, Forensic Evidence Gathering, Beamforming, Digital Signal Processing.*

1 INTRODUCTION

The concept of audio zooming, similar to a visual camera's ability to narrow down on a specific portion of a scene, aims at allowing users to concentrate on particular sounds within an audio environment. This notion dates back to the early 1950s (Cherry, 1953). However, mimicking the human brain's capacity to segregate and focus on specific audio signals amidst noise is still a formidable challenge for current technologies (Hawley, Litovsky, & Culling, 2004). Huang, Benesty, and Chen (2006) suggested the creation of a system capable of filtering out non-essential sounds while preserving the audio of interest. Such advancements could significantly benefit areas like video surveillance and media broadcasting.

The team's prior research (Stroud et al., 2023) detailed the development of a resilient beamforming audio zoom system specifically designed for video surveillance, utilising a microphone array to pinpoint and amplify sound from designated areas. This setup utilised time-delay beamforming techniques to offset issues with three of the sixteen microphones in the array failing.

Advancing from our previous work, we present a simulation model that offers the use of arrays in various configurations and the simulation of scenes with adjustable sizes and dimensions inspired by an actual crime scene in Liverpool, England. This model applies time-delay beamforming techniques accounting for audio signal reflections and incorporates an active noise-cancelling feature. Through a MATLAB-based program, our simulation aims to provide a reliable and accurate audio surveillance tool that could be integrated with video surveillance systems, representing a step forward in developing novel surveillance technologies.

2 LITERATURE REVIEW

Initial advancements towards an apparatus designed for auditory scene magnification commenced with Olson and Preston (1949) introducing a single ribbon cardioid microphone capable of diminishing rearward sounds. This invention exhibited a heightened Super-Cardioid response as the sound frequency increased, demonstrating that higher frequencies (e.g., 10kHz) elicited a stronger Super-Cardioid pattern than lower frequencies (e.g., 1kHz). The influence of Cherry's work (1953), which first addressed "*The Cocktail Party Problem*" (CPP), is evident in this development, marking the initial consideration of machinery to navigate this issue.

The concept of genuine audio magnification, mirroring the capabilities of then-recent video zoom technologies, emerged in 1980 through the work of Ishigaki *et al*. (1980), who crafted a Second Order gradient unidirectional microphone for JVC^{TM}. This innovation, characterised by a frequency range of 100Hz to 10KHz and a unidirectional polar pattern, depended on a closely matched pair of electret microphones.

Furthering the endeavour, Matsumoto and Naono (1989) designed a stereo-zooming microphone that leveraged psychoacoustic manipulation to enhance the directionality of sound capture, providing a sense of spatiality similar to stereo recording. This method expanded upon previous mono-zoom techniques but still fell short of the capabilities of traditional video zooms. Investigations into resolving the CPP persisted across various disciplines into the 21st century. Haykin & Chen (2005), referencing the work of Wang and Brown (1999), proposed the utilisation of Machine Learning and Computational Auditory Scene Analysis (CASA) to construct computational models capable of discerning and tracking sound signals within an auditory scene.

Schultz-Amling *et al*. (2010) explored Acoustical Zooming through Directional Audio Coding (DirAC), focusing on the parameters of sound's Direction of Arrival (DOA) and diffuseness. While this research originated with teleconferencing applications in mind, their work suggested potential applications in synchronising drone audio and video for focused capture. Van Waterschoot *et al*. (2013) further considered Acoustical Zooming using an array of multiple microphones, presenting a comprehensive theory for independent sound level control without necessitating explicit sound source separation algorithms, thereby reducing computational demands. Their approach, relying on spatial and spectral noise reduction techniques, showed promise for audio-visual applications using cost-effective microphones.

The principle behind Acoustic Zoom (AZ) involves manipulating acoustic cues that influence the perceived distance of sound sources, with sound intensity being a primary factor alongside the Direct to Reverberant Ratio (DRR), Spectral distortion, and Interaural differences (Time and Level), as well as the rate of intensity change due to motion. Thiergart, Kowalczyk, and Habets (2014) endorsed spatial filtering as the most effective method for Acoustic Zooming. Following this, Christensen *et al*. (2016) demonstrated the superiority of a rank Wiener subspace filter with Dynamic rank limitation over traditional approaches for speech enhancement in CPP simulations.

Wilson's (2017) findings highlighted the natural human propensity to enhance the Signal to Noise Ratio (SNR) in CPP scenarios by leveraging binaural hearing capabilities, drawing parallels between natural auditory strategies and professional audio engineering techniques. Studies focusing on Sound Source Localisation (SSL) using an onboard microphone array on a drone have been completed by Manamperi, Abhayapala, Zhang, and Samarasinghe (2022), even under extreme SNR levels.

Research into the cocktail party problem that started in 1953 continues to this day, with recent work by Zhang et al. (2022) exploring end-to-end dereverberation and beamforming related to human speech recognition. Given the absence of integrated audio and video zoom systems for surveillance, initiating experiments with a computationally efficient time-delay beamformer was deemed a pragmatic approach.

3 METHODOLOGY

3.1 Simulation Environment Preparation

This study introduces a modular MATLAB codebase designed to improve environmental sound recognition by employing advanced beamforming and noise detection techniques within a digital simulation. The methodology is organised as follows: The setup begins in the MATLAB environment by accurately defining the experimental dimensions, including the simulated space's length, width, and height surrounding the 'Exemplar Houses' at Liverpool John Moores University (See Appendix for information). This initial step ensures the creation of a precise spatial framework essential for the realistic modelling of acoustic scenarios.

A three-dimensional scene is generated, creating a 'World Objects' structure that includes realistic building structures and materials. This step incorporates actual Sabine acoustic coefficients to emulate real-world environments, a vital element for authentic environmental sound simulation. Participants are prompted to select the position of the microphone array within the scene, with the option to choose a default central location or custom coordinates.

The design allows for the selection of the microphone array's shape (Square, Circle, or Cross) and provides the capability to activate or deactivate individual omnidirectional microphones within the array, offering flexibility in configuring the sound detection system.

3.2 Noise Reduction Array Integration

A specialised microphone array intended for noise reduction is integrated on top of the primary array, designed under the presumption that a police surveillance drone could deploy the system. Consequently, eliminating ambient noise is deemed necessary for enhancing signal clarity. As Harrison *et al.* (2023) documented, audio forensic evidence is often buried in unwanted noise; therefore, applying noise reduction to identify human speech better seemed a sensible approach.

3.3 Grid Selection, Speaker Placement and Sound Wave Simulation

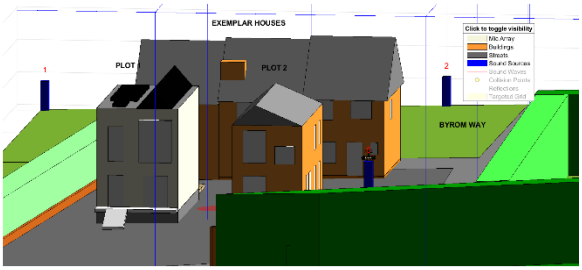

Figure 1. The 3D simulation showing Speaker 5 is located below the array.

Figure 2. Exemplar Houses, Liverpool John Moores University.

The experimental setup allows the user to select the grid size, with options ranging from a full-sized configuration (40m x 60m) to custom dimensions. This selection dynamically alters the visual representation of the 3D scene. The placement of virtual speakers within the experiment is configurable, offering a standard setup with eight sound sources evenly distributed around the grid's perimeter and a ninth sound source (Speaker 5) located in the scene's centre or alternative placement via custom coordinates. To prevent overdriving the array, the central speaker (Speaker 5), located directly underneath the array, is disabled.

Additionally, users have the capability to adjust sound wave characteristics for each speaker, including azimuth and elevation angles, beam width, and sound pressure levels, facilitating a more accurate recreation of realistic acoustic scenarios. Further details are available in Table 1.

Table 1: Sound Profile for each Speaker

Speaker	SPL at 1 meter (dB $_{SPL}$)	Azimuth Angle (Degrees °)	Elevation Angle (Degrees °)	Horizontal Beam Width (Degrees °)	Vertical Beam Width (Degrees °)
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

Following the configuration of speakers, visual representations of the sound waves are produced to interact within the 3D scene. The simulated sound waves engage with the scene, reflecting off surfaces, with the intensity of these reflections being determined by the Sabine absorption coefficients of the virtual materials they encounter. Additionally, the waves attenuate based on the specific absorption levels of these materials and the inverse square law.

Once this setup is complete, users select a particular grid segment, numbered 1 to 9, to serve as the focus of the beamforming process. A secondary 3D scene is generated, zooming into the selected grid size for detailed observation, as illustrated in Figure 3.

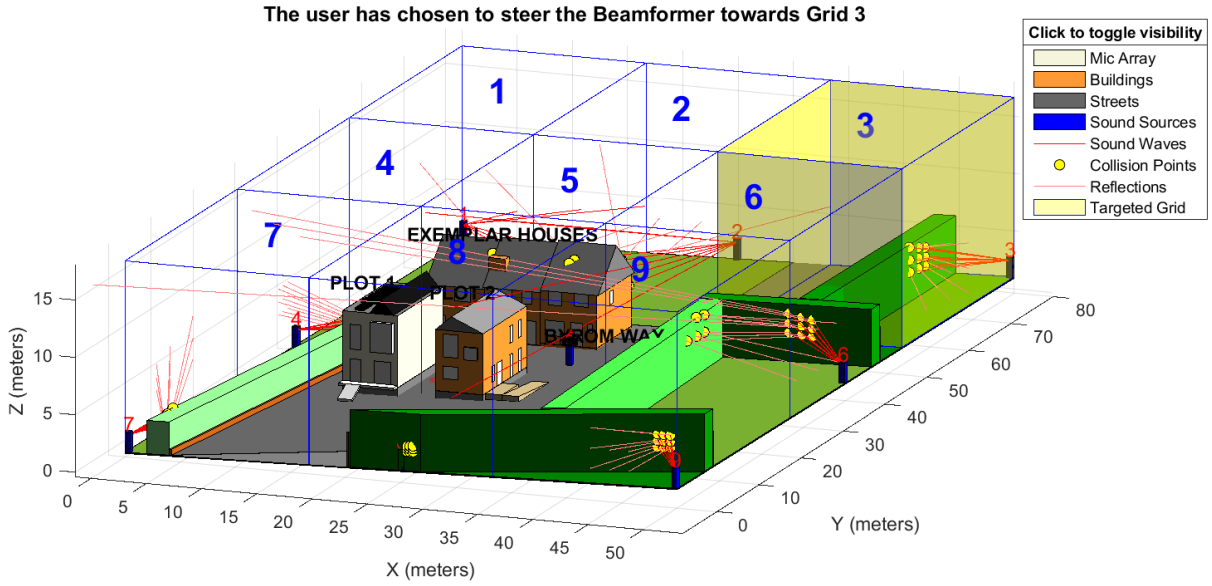

Figure 3. The 3D simulation plot displays the entire grid, with the microphone array, speakers and soundwaves.

3.4 Beamforming and Noise Reduction

Calculations of distances and delays between speakers, microphones, and those within the microphone array are meticulously conducted to model sound propagation and reception accurately. To

facilitate dynamic interaction with the simulation data, an interactive legend is added alongside the secondary 3D scene, enhancing the ability to navigate and interpret the presented information.

The simulation integrates virtual or real-world audio signals, providing various acoustic environments for analysis. A specific scenario is tested, featuring a male voice from Speaker 1 and music files from the remaining speakers, with additional drone noise infiltrating the microphone setup.

The technique of time delay beamforming is employed to concentrate the microphone array on precise sound sources located within a selected grid segment (for instance, Grid 3 in the given scenario), utilising a specific formula to achieve this focus.

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

As an explanation for the time delay beamforming algorithm, $S_{out}(t)$ is the beamformed output signal at time t, while w_i represents the weight applied to the first microphone signal from the array, $S_{In}(t-\tau_i)$ is the input signal from the microphone delayed by τ_i, which is the time delay for the microphone signal, calculated based on the distance between the microphone and the speaker and the speed of sound. This process is applied to all the microphones on the array.

A subtractive noise reduction method is applied to improve the clarity of the beamformed signal, taking advantage of the input from the feedforward microphone.

$$S_{Filter}(t) = S_{Out}(t) - \alpha N_{Drone}(t) \quad (2)$$

$S_{Filter}(t)$ is the cleaned signal after noise reduction, and α is the scaling factor that adjusts the contribution of the noise signal to the subtraction process. The experiment's results are summarised in a textual format, highlighting the principal outcomes and linking each figure to distinct elements of the simulation and analytical procedures.

This strategy adopts modular coding techniques in Matlab, establishing a versatile and extensive system for exploring environmental sound recognition. The approach presented promotes progress in digital acoustic simulations by offering elaborate setup choices, dynamic 3D modelling, and sophisticated signal processing methodologies.

4 RESULTS

The following section presents the results of the audio zooming experiment using time-delay beamforming.

4.1 Sound Capture by the Virtual Array

The setup of 16 virtual omnidirectional microphones demonstrated precision in recording audio mixtures that mimic real-world situations, encompassing elements of speech, music, and drone noise. The algorithm employed considered several essential parameters for each sound source in the experiment. These included the initial Sound Pressure Level (SPL) at 1 meter, azimuth angle, elevation angle, and the horizontal and vertical beam widths. This detailed approach facilitated an intricate capture of the complex sound environment, considering the effects of distance, reflections, material absorption properties, and reverberation.

Compliance with the inverse square law was ensured to achieve a realistic simulation of sound intensity attenuation over distance. The outcome was the creation of a 5-second audio file for each microphone, stored in a MATLAB array, which accurately reflects the complex interplay of different audio elements. This level of precision in capturing the acoustic environment highlights the capability of the virtual microphone array setup to replicate real-world audio dynamics within a digital simulation context.

The audio waveforms captured by each microphone are illustrated in Figure 4, showcasing the efficacy of this simulation setup.

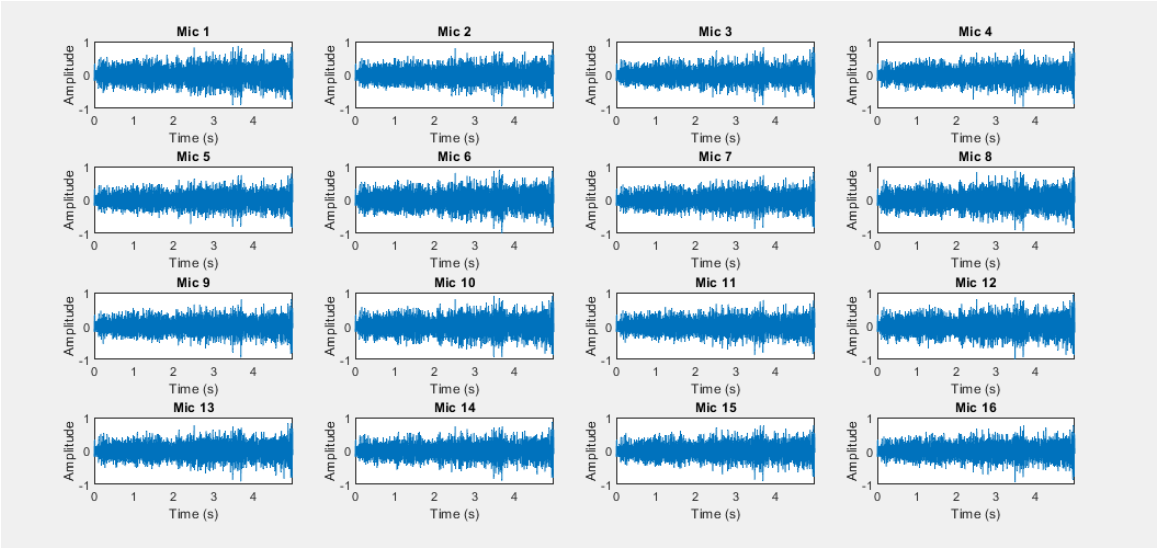

Figure 4. The array's 16 virtual microphones successfully captured real-world sound mixtures.

4.2 Beamforming Results

The outcomes from the beamforming process validate the efficacy of the time-delay beamforming technique. By leveraging the signals collected by the 16 virtual omnidirectional microphones, the algorithm successfully combined and steered these inputs towards a selected grid direction (specifically, Grid 3 for this instance). This targeted manoeuvring of audio signals showcases the algorithm's ability to selectively focus on specific sound sources, enhancing the sounds within the aimed area while diminishing those outside the intended polar pattern.

As a result, the algorithm significantly improved the clarity and loudness of the targeted sounds and markedly lowered the presence of background noise and unrelated audio components. The refined beamformed audio, encapsulating a span of 5 seconds, underwent recording and analysis in both time and frequency domains. This analytical review was further enriched by an auditory playback functionality within Matlab, providing users with immediate and accessible outcomes.

Through the successful application of beamforming technology, a valuable technique for audio signal refinement has been demonstrated, setting the stage for more precise and clearer acoustic captures within digital simulation environments.

4.3 Noise Reduction on the Beamformed Audio

Following the successful isolation of targeted audio via beamforming, the code proceeded to apply a subtractive noise reduction technique to the beamformed audio, significantly enhancing the sound's clarity. This technique employs a virtual feedforward microphone from the noise reduction array, positioned atop the primary microphone array, specifically engineered to pick up ambient noise, including the drone blades' buzzing and environmental wind disturbances. By capturing these distinct noise profiles, the algorithm adeptly isolates and eliminates them from the beamformed audio, effectively diminishing background noise and thus improving the sound quality.

The effectiveness of this noise reduction is particularly evident in the enhanced audibility of human speech within grid 1. The removal of superfluous noise elements significantly boosts the clarity and comprehensibility of the human voice, primarily located in the mid-range frequency spectrum. This phase highlights the utility of subtractive noise reduction techniques in refining audio recordings for surveillance applications, especially in environments where distinguishing foreground speech from prevalent background noise is crucial.

The results of this process are illustrated in Figure 5.

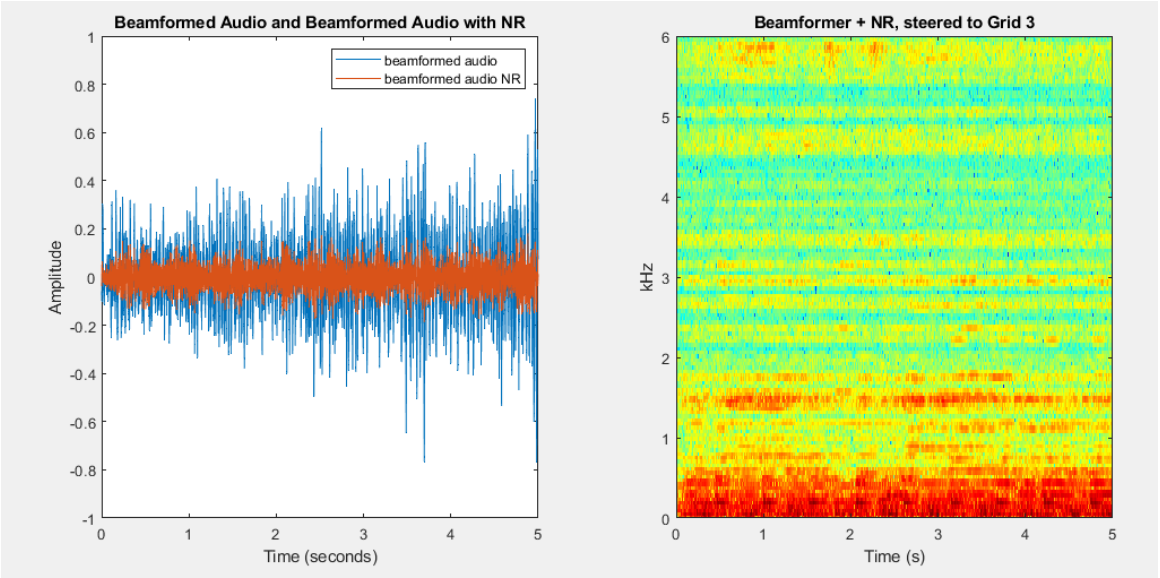

Figure 5. Time and frequency plots illustrate the application of the noise reduction algorithm to the beamformed audio.

5 CONCLUSION

The research demonstrates the capability of a novel system to precisely direct captured audio from a sophisticated array towards a specific grid whilst simultaneously diminishing extraneous noise within the audio signal. This improvement markedly increases the likelihood of discerning human speech against background noise within the audio recordings. Leveraging the insights of prior studies, the robustness of this system indicates its significant potential for practical field deployment. The system's proficiency in isolating and enhancing speech in a particular grid amidst a noisy backdrop highlights its cutting-edge audio processing advancements.

Future endeavours will aim to refine the beamforming algorithm further. The choice of the time delay beamforming algorithm was motivated by its simplicity, durability, and low computational demand, making it especially suitable for environments with limited computational capabilities or where swift implementation is necessary. While the Minimum Variance Distortionless Response (MVDR) beamformer theoretically offers enhanced interference suppression, its real-world efficacy largely depends on the precise covariance matrix estimation. Time delay beamforming might achieve comparable or even superior outcomes in scenarios characterised by fluctuating noise environments. The objective is to reduce noise and interference further, thereby improving sound clarity and separation.

While practical field conditions will undoubtedly provide additional challenges in contrast to the simulated environment, our method provides a robust and realistic model for testing. This research aims to attain an audio quality and distinction that makes the technology useful for forensic surveillance, broadcasting uses, or even animal conservation, embodying the quest for innovative solutions to intricate audio challenges and extending the frontiers of sound recognition and separation capabilities.

REFERENCES

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America, 25*(5), 975-979. doi:10.1121/1.1907229

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

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

Hawley, M., Litovsky, R., & Culling, J. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America, 115*, 833-843. doi:10.1121/1.1639908

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation, 17*(9), 1875-1902. doi:10.1162/0899766054322964

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

Ishigaki, Y., Yamamoto, M., Totsuka, K., & Miyaji, N. (1980). Zoom Microphone. *The Audio Engineering Society Convention Preprint, 1713 (A-7)*.

LJMU. (2016). Ljmu Exemplar Houses. Retrieved from https://www.ljmu.ac.uk/-/media/files/ljmu/about-us/faculties-and-schools/fet/ljmu-exemplar-houses--what-we-do.pdf

Manamperi, W., Abhayapala, T. D., Zhang, J., & Samarasinghe, P. N. (2022). Drone Audition: Sound Source Localisation Using Onboard Microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 508-519. doi:10.1109/TASLP.2022.3140550

Matsumoto, M., & Naono, H. (1989). Stereo Zoom Microphone For Consumer Video Cameras. *IEEE Transactions on Consumer Electronics, 35*(4), 759-766.

Olson, H. F., & Preston, J. (1949). Single-Element Unidirectional Microphone. *Journal of the Society of Motion Picture Engineers, 52*(3), 293-302. doi:10.5594/J12528

Schultz-Amling, R., Kuech, F., Thiergart, O., & Kallinger, M. (2010). Acoustical Zooming Based on a Parametric Sound Field Representation. *Audio Engineering Society Convention Paper 8120*, 1-9.

Stroud, S., Jones, K. O., Edwards, G., Robinson, C., Ellis, D., & Chandler-Crnigoj, S. (2023). *Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array*. Paper presented at the 37th International Conference on Information Technologies (InfoTech-2023), Bulgaria. https://ieeexplore.ieee.org/document/10266894

Thiergart, O., Kowalczyk, K., & Habets, E. A. P. (2014, 2014). *An acoustical zoom based on informed spatial filtering*.

Van Waterschoot, T., Joos Tirry, W., & Moonen, M. (2013). Acoustic Zooming by Multimicrophone Sound Scene Manipulation. *Audio Engineering Society, 61*.

Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks, 10*(3), 684-697. doi:10.1109/72.761727

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

Zhang, W., Chang, X., Boeddeker, C., Nakatani, T., Watanabe, S., & Qian, Y. (2022). End-to-End Dereverberation, Beamforming, and Speech Recognition in A Cocktail Party. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1-16. doi:10.1109/TASLP.2022.3209942

Appendix

Exemplar Test Houses, Byrom Street Campus, Liverpool John Moores University.

In 2016, Liverpool John Moores University launched three full-sized test houses at its Byrom Street campus, intended for R&D purposes. These houses represent the evolution of UK house design, with one modelled on a 1920s-style home, another on a 1970s-style house, and a third showcasing modern features. Equipped and furnished to mimic actual living conditions of the time, they serve as practical testing grounds for forensic tests, products in development, and for facilitating data gathering in authentic settings for research. (LJMU, 2016).