Developing Audio Zoom in Virtual Environments: Real-World Soundscapes and Targeted Noise Detection

Stephen Stroud, Dr Karl O. Jones, Dr Gerard Edwards, Colin Robinson, Dr Sebastian Chandler-Crnigoj and Dr David Ellis
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, G.Edwards, C.Robinson1, D.L.Ellis, S.L.ChandlerCrnigoj }@ljmu.ac.uk

Abstract – **This study presents an innovative beamforming method tailored for audio surveillance applications, developed through virtual simulations conducted at Liverpool John Moores University. Motivated by the growing need for advanced audio analysis techniques, this method focuses on segregating and enhancing particular sounds within acoustically cluttered environments, such as those presented in forensic scenarios, including criminal court cases. Utilising a time-delay beamforming algorithm, this study introduces a novel strategy to identify and magnify specific noises within intricate acoustic settings, addressing challenges often faced in surveillance and forensic audio examinations.**
The foundation of our technique lies in the strategic deployment of a robust omnidirectional microphone array, which is crucial for capturing environmental sounds. Our methodology involves the application of a MATLAB algorithm to process these sounds, followed by comprehensive evaluations to gauge the system's effectiveness in isolating targeted audio sources.
The investigation examines the system's robustness against microphone array degradation, demonstrating its reliability even with compromised functionality. Simulations of real-world acoustic conditions reveal the algorithm's effectiveness in handling sound reflections and reverberations, crucial for urban acoustic landscapes. This study introduces a novel beamforming method for audio surveillance, with potential applications in broadcasting, advanced audio engineering, and wildlife conservation, highlighting its versatility.
In summary, this research introduces a novel approach to audio surveillance, paving the way for numerous practical applications that could leverage improved audio isolation and analytical capabilities. Our results contribute significantly to the ongoing development of sophisticated surveillance technologies, offering valuable perspectives that could influence the future landscape of audio engineering and analysis.

Keywords – **Audio Zooming, Surveillance, Forensic Evidence Gathering, Beamforming, Digital Signal Processing**

I. Introduction

The principle of audio zooming, analogous to how a visual camera zooms into specific scene segments, focuses on enabling listeners to isolate and engage with selected sounds within an auditory environment. This concept has historical roots in the early 1950s [1]. Despite its longevity, replicating the human brain's ability to filter and concentrate on particular sounds amid background noise remains a significant technological challenge [2]. Huang, Benesty, and Chen [3] proposed a system to eliminate irrelevant noises while capturing and preserving the desired audio signals. Such advancements can profoundly impact sectors such as media broadcasting and video surveillance.
Previous investigations [4] have led to the creation of a robust audio zoom system using beamforming tailored for video surveillance applications. This system features an array of microphones that specifically target and enhance audio from selected zones. This arrangement relied on time-delay beamforming techniques to manage the failure of three out of sixteen microphones in the setup.
Building on this foundational work, we introduce a simulated model that employs microphone arrays in diverse configurations and simulates environments of varying sizes and dimensions, mirroring a real-life crime scene [5] in Liverpool, England. This model leverages time-delay beamforming methods to address the challenges posed by audio reflections and includes an active noise-cancellation functionality. Executed through a MATLAB-based platform, our model strives to deliver a precise and dependable audio surveillance tool that could seamlessly integrate with existing video surveillance frameworks, marking an advancement in the evolution of innovative surveillance technologies. [6]

II. Related Work

Initial progress in developing devices for enhancing auditory scenes began when Olson and Preston [7] introduced a single ribbon cardioid microphone that attenuated sounds from behind. This device demonstrated an increased Super-Cardioid response at higher sound frequencies, showing that frequencies like 10kHz exhibited a more pronounced Super-Cardioid pattern compared to lower frequencies such as 1kHz. This innovation was influenced by Cherry's earlier work [1] on "The Cocktail Party Problem" (CPP), which was the first to address the challenge, marking a preliminary attempt to tackle this issue with technology.
The notion of true audio magnification, paralleling then-recent advancements in video zoom technology, was realised in 1980 by Ishigaki et al. [8], who developed a

Second Order gradient unidirectional microphone for JVCTM. This device utilised a pair of closely matched electret microphones, characterised by a frequency spectrum of 100Hz to 10KHz and a unidirectional polar pattern. Advancing this concept, Matsumoto and Naono [9] created a stereo-zooming microphone that used psychoacoustic effects to improve the directionality of sound capture, adding a spatial dimension akin to stereo sound. This technique was built on prior mono-zoom methods but did not yet match the capabilities of conventional video zoom technology.

Efforts to solve the CPP continued into the 21st century across multiple disciplines. Haykin & Chen [10], citing Wang and Brown [11], explored the application of Machine Learning and Computational Auditory Scene Analysis (CASA) to develop computational models that could identify and follow sound sources within an auditory environment. Schultz-Amling et al. [12] investigated Acoustical Zooming via Directional Audio Coding (DirAC), analysing sound parameters such as Direction of Arrival (DOA) and diffuseness. Initially intended for teleconferencing, their findings also suggested uses in synchronising drone audio and video for targeted recording. Van Waterschoot et al. [13] further advanced Acoustical Zooming techniques using a multi-microphone array, proposing a comprehensive framework for controlling sound levels independently without requiring explicit sound source separation algorithms, thus lowering computational requirements. Based on spatial and spectral noise reduction, their methods appeared promising for audio-visual applications employing affordable microphones. Acoustic Zoom (AZ) manipulates various acoustic cues that affect the perceived proximity of sound sources, focusing on factors like sound intensity, Direct to Reverberant Ratio (DRR), Spectral distortion, Interaural differences (both Time and Level), and changes in intensity due to source movement. Thiergart, Kowalczyk, and Habets [14] recommended spatial filtering as the most effective technique for Acoustic Zooming. Following this approach, Christensen et al. [15] proved the efficacy of a rank Wiener subspace filter with Dynamic rank limitation over conventional methods in enhancing speech within CPP simulations. Wilson's [16] research emphasised the innate human ability to improve the Signal to Noise Ratio (SNR) in CPP situations through binaural hearing, comparing natural auditory processes and sophisticated audio engineering practices.

With no existing integrated audio and video zoom systems for surveillance available, beginning experiments with a computationally efficient time-delay beamformer was considered a practical step forward.

III. Proposed Research

A. Simulation Environment Preparation

This research presents a modular MATLAB codebase to enhance environmental sound recognition through sophisticated beamforming and noise detection strategies in a digital simulation setting. The methodological framework is structured as follows:

The process initiates within the MATLAB environment by precisely specifying the dimensions of the experimental area, including the length, width, and height of the space around the 'Exemplar Houses' at Liverpool John Moores University. This crucial first step establishes an accurate spatial context necessary for the realistic simulation of acoustic environments.

A three-dimensional scene introduces a 'World Objects' configuration comprising realistic architectural forms and materials. This phase integrates actual Sabine acoustic coefficients to mirror real-life conditions when dealing with sound reflections, which is crucial for realistic environmental sound simulations.

Users are invited to determine the location of the microphone array in the scene, with options to select either a standard central position or customised coordinates.

The system design permits the selection of the microphone array's configuration (Square, Circle, Cross, Octagon or Eight-pointed star). Also, it enables users to turn individual omnidirectional microphones within the array on or off, thus providing adaptability in the sound detection setup.

B. Noise Reduction Array Integration

An advanced microphone array, specifically designed for noise suppression, is positioned atop the main array. It was developed assuming it could be utilised on a police surveillance drone. Eliminating ambient noise is critical for improving signal clarity, as noted by Harrison *et al.* [17], audio forensic evidence frequently is obscured by undesired noise; thus, employing noise reduction techniques to discern human speech more effectively was considered a logical strategy.

C. Grid Selection, Speaker Placement and Sound Wave Simulation

The experimental framework enables users to choose the grid size, offering configurations from a standard large scale (40m x 60m) to customisable dimensions. This choice impacts the 3D scene's dynamic visual representation. Users can configure the placement of virtual speakers within the experiment, with a default arrangement of eight sound sources uniformly spaced along the grid's edge and a ninth central speaker (Speaker 5) directly below the array, which is deactivated to avoid overloading the system.

Furthermore, users can modify the acoustic properties of each speaker, such as azimuth and elevation angles, beam width, and sound pressure levels, to closely mimic real-world acoustic environments. After setting up the speakers, the system generates visual simulations of the sound waves interacting with the 3D scene. These waves reflect off surfaces according to the Sabine absorption coefficients of the virtual materials encountered and diminish in intensity following the material-specific absorption rates and the inverse square law. Upon completing this setup, users select one of nine numbered grid segments as the target for the beamforming analysis.

A secondary 3D scene is created, providing a zoomed-in view of the chosen grid segment for detailed examination, as depicted in Figure 1.

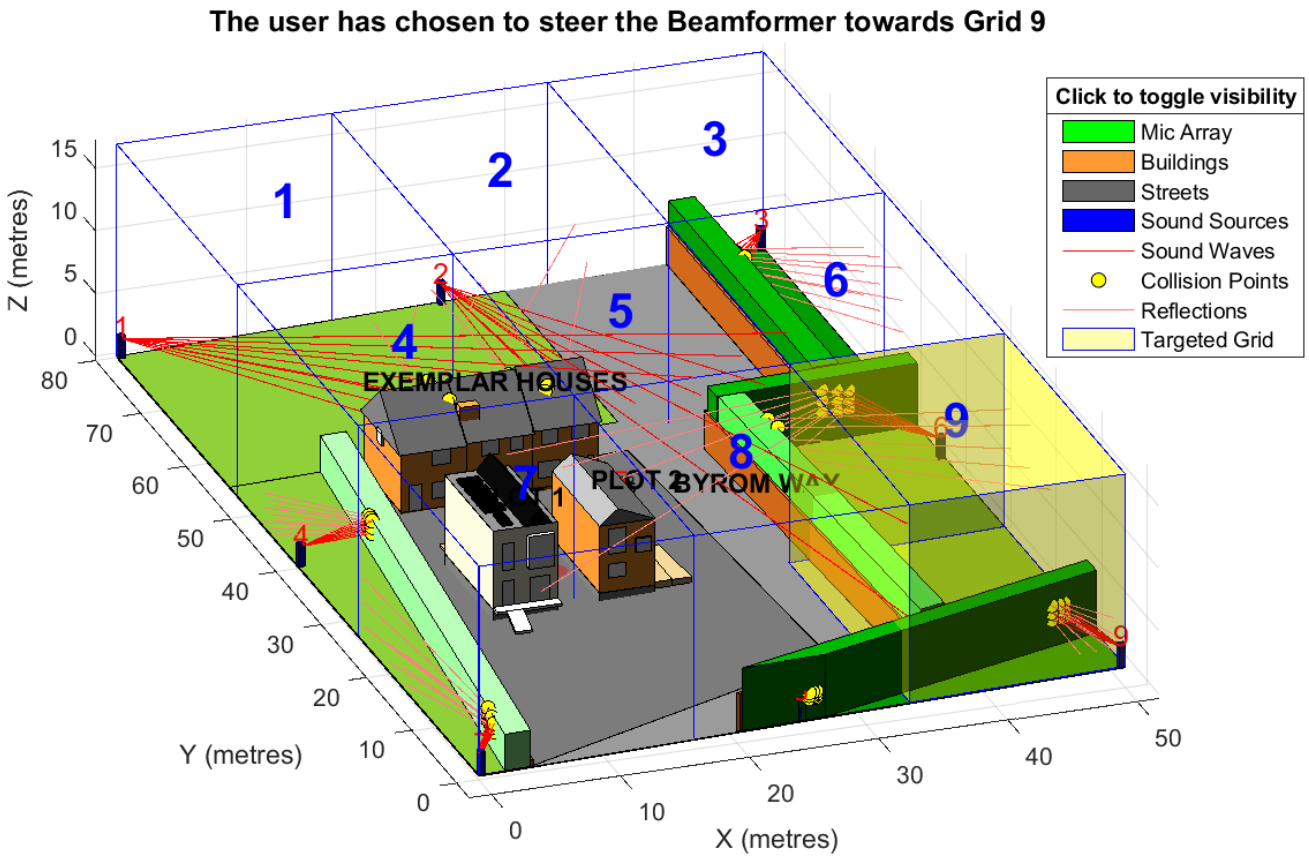

Figure 1. A plot of the user-selected grid nine is highlighted within MATLAB. The user has chosen to steer the beamformer towards grid nine, highlighted in yellow in the bottom right corner of the figure.

D. Beamforming and Noise Reduction

Distance and delay calculations among speakers, microphones, and elements within the microphone array are precisely executed to depict sound transmission and reception accurately. An interactive legend is incorporated next to Figure 2 to enhance the user's ability to interact with and understand the simulation data. The simulation allows for integrating either virtual or actual audio signals, accommodating a range of acoustic environments for detailed examination. In a particular test scenario, a male voice is emitted from Speaker 1, and music tracks play from other speakers while drone noise permeates the microphone setup.

The time delay beamforming method focuses the microphone array on specific audio sources within a designated grid section (such as grid nine in the described scenario), applying a particular formula to achieve precise targeting.

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

The time delay beamforming algorithm is elucidated as follows: $S_{out}(t)$ represents the beamformed output signal at time t, w_i denotes the weight applied to the initial microphone signal in the array, and $S_{in}(t - \tau_i)$ is the input signal from a microphone, delayed by τ_i. The delay τ_i is determined by the distance between the microphone and the speaker and the sound's velocity. This procedure is replicated across all microphones in the array.

To enhance the clarity of the beamformed signal, a subtractive noise reduction technique is employed, leveraging the input from the feedforward microphone. The experiment's findings are concisely summarised in text, underscoring key outcomes and associating each figure with specific aspects of the simulation and analytical methods. This approach utilises modular coding practices in MATLAB, creating a flexible and comprehensive framework for investigating environmental sound recognition. The methodology advances digital acoustic simulations by providing complex setup options, dynamic 3D modelling, and advanced signal processing techniques.

IV. Results Discussion

A. Sound Capture by the Virtual Array

The configuration involving 16 virtual omnidirectional microphones achieved high precision in recording audio compositions that resemble actual environments, featuring components such as speech, music, and drone noise. The applied algorithm considered several crucial parameters for each audio source used in the experiment. These parameters included the initial Sound Pressure Level (SPL) at 1 meter, azimuth angle, elevation angle, and horizontal and vertical beam widths. Such a comprehensive method allowed for a sophisticated recording of the complex auditory scene, accounting for factors like distance, reflections, material absorption qualities, and reverberation. Adherence to the inverse square law was maintained to ensure a realistic portrayal of how sound intensity

diminishes with distance. The result was the generation of a 5-second audio clip for each microphone, encapsulated in a MATLAB array, which precisely depicted the intricate interactions among various sound elements. This accuracy in capturing the acoustic environment underscores the efficacy of the virtual microphone array setup in simulating real-world audio dynamics within a digital framework. The effectiveness of this simulation setup is demonstrated in Figure 2, which displays the audio waveforms captured by each microphone.

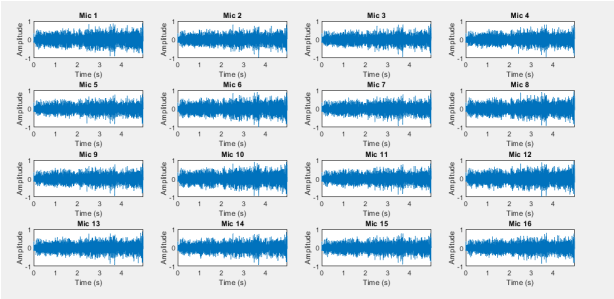

Figure 2. Plots of the sixteen microphone waveforms. The script identifies any signal with zero amplitude, indicating a broken microphone, highlights it in red, and removes it from the algorithm.

B. Beamforming Results

The results from the beamforming procedure affirm the effectiveness of the time-delay beamforming approach. Utilising the signals gathered by the 16 virtual omnidirectional microphones, the algorithm adeptly merged and directed these inputs towards a specific grid direction (notably, Grid 9 in this case). This precise control of audio signals illustrates the algorithm's capacity to selectively concentrate on distinct sound sources, amplifying sounds within the targeted area and reducing those outside the desired polar pattern. Consequently, the algorithm significantly enhanced the clarity and volume of the focal sounds and substantially reduced the levels of background noise and other extraneous audio elements. The optimised beamformed audio, spanning a 5-second duration, was recorded and analysed across both time and frequency domains. This analysis was complemented by an auditory playback feature within MATLAB, offering users immediate and interactive access to the results. The effective deployment of beamforming technology underscores its utility as a sophisticated tool for refining audio signals, paving the way for more accurate and precise acoustic recordings in digital simulation settings.

C. Noise Reduction on the Beamformed Audio

After isolating target audio through beamforming, the code applied a subtractive noise reduction strategy to the beamformed audio, markedly improving sound clarity. This method utilises a virtual feedforward microphone, part of the noise reduction array situated above the primary microphone array, specifically designed to capture ambient sounds, such as the hum of drone blades and environmental wind noise. The algorithm identifies and removes specific noise patterns from the beamformed audio, reducing background noise and improving sound quality. This technique notably enhances the audibility of human speech in grid nine. This stage underscores the effectiveness of subtractive noise reduction methods in refining audio recordings for surveillance purposes, particularly in settings where it is critical to distinguish foreground speech from pervasive background noise. The outcomes of this technique are depicted in Figure 3.

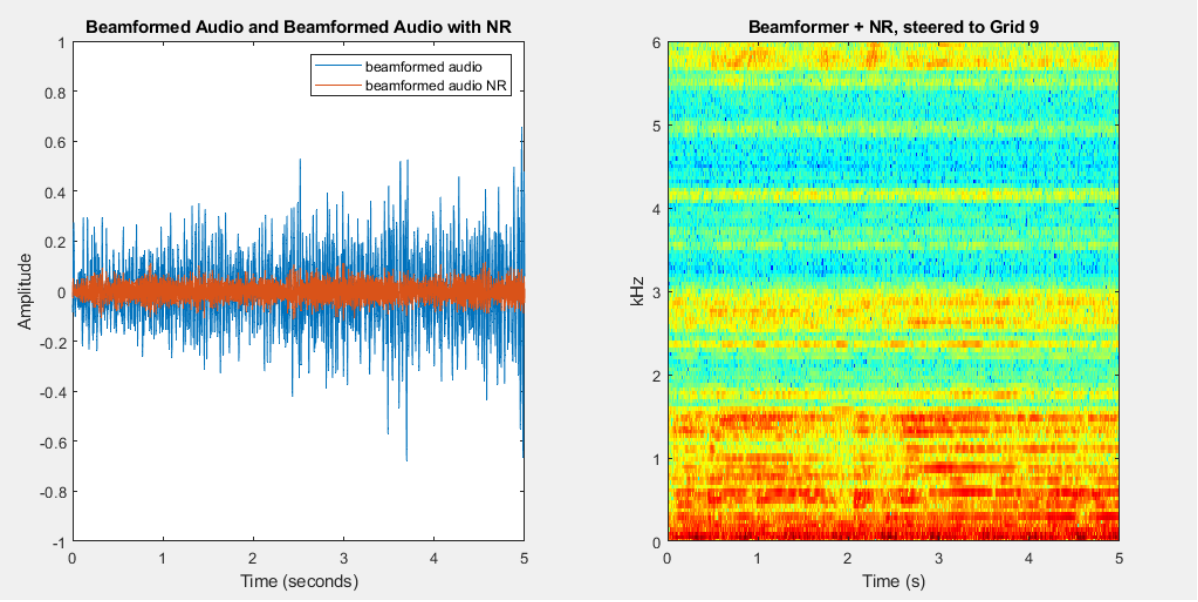

Figure 3. Waveform and Spectrogram of the beamformed signal, steered to grid nine. The original beamformed signal is in blue with the Noise reduced signal in orange.

Eliminating extraneous noise increases the clarity and intelligibility of the human voice, predominantly within the mid-range frequencies, which would aid Police with forensic audio investigations.

V. Conclusion

This study showcases the ability of a new system to accurately steer captured audio from an advanced array toward a designated grid while effectively reducing unwanted noise within the audio signal. Such enhancements significantly improve the ability to distinguish human speech from background noise in the recordings. Drawing on previous research findings, this system's durability demonstrates its considerable potential for real-world application. The system's skill in isolating and amplifying speech in a specific grid against a noisy environment underscores its state-of-the-art audio processing capabilities.

Future efforts will focus on refining the beamforming algorithm further. The selection of the time delay beamforming algorithm was driven by its straightforwardness, robustness, and minimal computational requirements, making it ideal for use in environments with limited processing power or where rapid deployment is crucial. Although the Minimum Variance Distortionless Response (MVDR) beamformer theoretically provides better interference reduction, its effectiveness depends heavily on accurate covariance matrix estimation. Time delay beamforming could deliver equal or better performance in environments with variable noise levels. The goal is to enhance noise and interference reduction further to boost sound clarity and separation. This research is directed towards achieving audio quality and distinction that renders the technology applicable for forensic surveillance, broadcasting, or wildlife conservation, pushing forward the boundaries of innovative solutions to complex audio challenges and advancing sound recognition and separation technologies.

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] BBC. "Ava White: Boy guilty of murdering girl, 12, in Snapchat row." https://www.bbc.co.uk/news/uk-england-merseyside-61554010 (accessed 13/05/2024, 2024).

[6] K. O. Jones, C. Robinson, H. Burrell, S. McColl, H. Bennett, and K. Morrison, "Audio & Video Forensics – A new direction for electronic engineering lecturing," *2023,* curriculum design; audio forensics; video forensics vol. 22, no. 3, p. 8, 2023-12-31 2023. [Online]. Available: https://sujes.selcuk.edu.tr/sujes/article/view/634.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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

[15] 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.

[16] 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.

[17] 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.