(W)EARABLE MICROPHONE ARRAY AND ULTRASONIC ECHO LOCALIZATION FOR COARSE INDOOR ENVIRONMENT MAPPING

Felix Pfreundtner, Jing Yang, Gábor Sörös

Department of Computer Science, ETH Zurich, Switzerland

ABSTRACT

We present a microphone array structure for spherical sound incidence angle tracking that can be attached to headphones or directly integrated into earphones. We show that this microphone array together with an ultrasonic sound source, e.g., a home assistant speaker in the room, allows to estimate the direction and distance of sound reflections on wall surfaces in the room. With our presented method, we achieved sound incidence angle estimation errors of around $14°$ in the horizontal plane and around $5°$ in the vertical plane, and reflection position estimation errors of around $5\,cm$ within $2\,m$. Having the reflection points in 3D allows us to sparsely capture the closest room geometry at interactive rate, which is a necessary step towards indoor audio augmented reality applications.

Index Terms— Wearable computing, hearables, echolocation, acoustic environment capture, audio augmented reality

1. INTRODUCTION

We envision future entertainment scenarios of 3D experiences with virtual sounds that interact with the real environment, also called audio augmented reality [1, 2]. In our anticipated scenario, the user is wearing headphones or earphones and is interacting with spatial audio content at home. Prior work has shown the possibility to use the room impulse response (RIR) for room identification and tracking the user's location inside a room. We have also witnessed the inference of the room shape from RIR measurements [3, 4], which led to a whole new research direction of localization and mapping based on acoustic signals. In particular, acoustic mapping and head tracking might be applicable when visual tracking is not feasible, for example in a dark cinema, home theater, or a gaming room, when the face is partially or fully occluded for an external camera, or low battery consumption is required. In contrast with most prior works, we target the simplest possible hardware setup that can give a coarse but real-time environment map and could be integrated into hearables.

State-of-the-art acoustic tracking and mapping includes sources that emit audio signals in the space to track the pose of a receiver that carries one or more microphones [5, 6, 7], and/or to map out the surrounding walls in the room [8, 9, 10, 11]. However, the complex microphone arrays or the active sources required by these techniques are difficult to be integrated into hearables. Binaural microphones with known speaker array anchors [5] or human speech anchors [6] can also be exploited for head tracking, with the main advantage of having the power-hungry active source(s) off the body. There also exist a number of methods that estimate the wall positions based on the reflected sounds from these surfaces [12, 13, 14]. Some approaches use several microphones that are distributed in the space [4] or a spherical microphone array [15, 16], some combine them with panoramic cameras [17]. Some methods [9, 11, 18, 19] even work with smartphones to derive the geometry of polygon-shaped rooms.

In this work, we aim for boundary mapping with hearables and ultrasonic signals with simple hardware. We propose an attachment for normal headphones with pairs of microphones, which could be miniaturized and easily integrated into earpods too. The microphone pairs capture ultrasound signals emitted from a loudspeaker source, which can be a home assistant smart speaker, a smartphone, or a mounted speaker in the room to generate chirps in the unhearable ultrasonic range. Based on the presented (w)earable microphone array, we propose an algorithm to estimate the spherical incidence angles of the incoming signals and derive the orientation and the distance of the hearables (the user) with respect to the sound source and the points of reflections in 3D space. In contrast to most prior works, the positions of the source and the listener are almost arbitrary, no prior assumption about the room geometry is required, and no explicit synchronization is needed. We show in qualitative and quantitative evaluation that our method is able to capture and update the coarse geometry in the user's proximity at interactive rate.
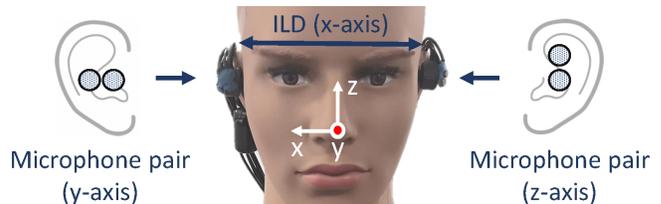
Fig. 1: The proposed ILD-Y-Z microphone array for ultrasonic measurements.

2. WEARABLE MICROPHONE ARRAY

The basic design concept is to split the 3D space around the head into three orthogonal directions (X, Y, Z) and align microphone pairs with these three directions. More specifically, a 4-mic ILD-Y-Z microphone array shown in Fig. 1 was used in our experiments. The microphone pair on the left ear corresponds to the Z-direction and the microphone pair on the right ear corresponds to the Y-direction. Instead of having a third pair, we derive the microphone signal along the X-axis by calculating the sound pressure level difference (interaural level difference, ILD [20]) of the head from the Y and Z microphones.

Apart from the ILD-Y-Z geometry, we have also explored an orthogonal X-Y-Z geometry, in which an explicit microphone pair is used to derive the information along the X-axis rather than using the ILD information. With the X-Y-Z geometry, one ear is equipped with three orthogonal microphone pairs, in which the two closely aligned orthogonal pairs can share one common microphone. Compared to the X-Y-Z geometry, the ILD-Y-Z geometry requires fewer

microphones, which is more suitable for lightweight hearable applications. Therefore, we employ the ILD-Y-Z geometry and aim to solve for the ambiguities algorithmically.

With the ILD-Y-Z geometry, the two microphones in each pair should be separated by a distance that is smaller than half the shortest wave length of interest. In our case of using an ultrasonic signal in the frequency range $20\,kHz - 24\,kHz$, the inter-microphone distance was set smaller than $0.7\,cm$. Such a flexible setup might be used in hearables by exploiting the existing hybrid active noise cancellation hardware that usually includes two omnidirectional microphones at each ear. We used Primo EM258 electret microphones and covered all but one vent on each so to reach an effective intermic distance smaller than what the physical size of the individual components would allow. In our experiments, an Animal Sound Lab UMS-1 ultrasound loudspeaker played the ultrasonic signals. A Focusrite Scarlett 18i8 interface was chained with a Steinberg UR44 interface for the audio recording and playback.

3. ESTIMATING THE INCIDENCE ANGLE

Our loudspeaker repeatedly plays chirps in the ultrasonic domain and inverse filters on each stream continuously estimate the room impulse response (RIR). We detect dominant peaks in the RIRs captured by the individual microphones, which correspond to the direct sound and echos from the walls. We aim to estimate the spherical incidence angles (φ, θ) in the vertical and horizontal plane with respect to the head.

In the vertical plane, the incidence angle is estimated from the Z microphone pair. Given a vertical incidence angle φ, the sound signal arrives at these two microphones with a time delay $\tau_z = d\cos\varphi$ where d is the given distance between the two microphones. The incidence angle of our interest is where $\varphi = argmax\,(cc_{z_1,z_2}(\tau_z))$, the cross-correlation between the two z signals is maximized. The same approach is applied in the Y direction.

In the horizontal plane, there exists a mapping from the interaural level difference $ILD(\alpha_{ILD})$ to the corresponding horizontal incidence angle α_{ILD} as described in [21]. The ILD also depends on the vertical incidence angle, so taking the value φ estimated above, there exists a mapping from $ILD(\alpha_{ILD}|\varphi)$ to the corresponding α_{ILD} based on spherically measured head-related transfer functions (HRTFs) [22]. In our method, the probability distribution of the horizontal incidence angle $p(\theta)$ is approximated by a window function $p(\theta) = w(\alpha_{ILD} - \theta)$, where we use a Kaiser window as defined in [23]. This window function guarantees that $p(\theta)$ peaks at $w(0)$ (i.e. when $\theta = \alpha_{ILD}$). The window width and steepness can be tuned to adjust the angle range for the probability calculation. The microphones on the Y-axis can resolve the ambiguity in signals coming from the front and the back of the head. The mapped α_{ILD} is the desired maximum likelihood estimate for the horizontal incidence angle θ.

So far, we have explained the derivation of the spherical incidence angles (φ, θ) using the ILD-Y-Z microphone array. Because the ILD is used, we need to assume that the signal can be stably observed over the time needed for the wave to travel from one side to the other side of the head which may be violated when the head or the source is moving. We could drop this assumption by adding a dedicated X microphone pair. Note that although the derivation is based on the microphone array with orthogonal microphone pairs, it is also possible to generalize the algorithm to non-orthogonal microphone pairs by applying affine transformation matrices to map the geometry to its corresponding representation in the spherical coordinate system.

4. ESTIMATING THE POSITION OF THE REFLECTION POINT

Estimating the reflection points in the space is useful for mapping the wall surfaces near the user, and their changes may indicate the relative movements of a user in the space. Our estimation algorithm consists of two parts: the frames with sparse echos are detected within the multichannel RIR signal, and then the reflection positions are derived geometrically.

We split the impulse responses recorded at all microphones into 75% overlapping frames and determine the frames $f \in \mathbb{F}$ that contain strong sparse echos by calculating the received energy in the frame $w_{f,\text{energy}}$ and normalizing it across all frames to get the metric in a range of $[0, 1]$. We set the frame duration corresponding to a wave travel distance of $1\,m$. Note that other metrics like sparsity, temporal stability, etc. can also be applied to select the frames of echos, but finding the best metric is out of the scope of this paper.

After detecting the frames with the most likely sparse and strong reflections, we derive the Cartesian position of every reflection point $R(x, y, z)$, with each point linked to one echo frame. We first derived equations to calculate the position of the reflection point $R(x, y, z)$ on a 2D ellipse (following [24]), and this algorithm does not require to know the position of the sound source. Fig. 2 illustrates the ellipse geometry, which corresponds to the special case when the loudspeaker, the reflection point, and the microphone are all on the same height (z). The positions of the sound source S and the microphone M are the two foci of the ellipse. The arrival time of an echo in a frame implies the distance the sound travels from the sound source to the reflection point and then to the microphone. The distance $d_d = a$ is travelled by the direct sound and the distance $d_r = b + c$ is travelled by the first reflection. We also generalized these equations from the special case of a 2D ellipse to the normal case of a 3D ellipsoid.
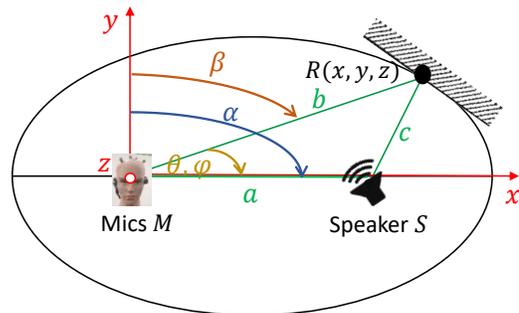
Fig. 2: Estimating the position of a reflection point $R(x, y, z)$ by using the loudspeaker and the microphone as the foci of an ellipse, with knowledge of the directions MS and MR.

Based on the method in Section 3, we can calculate the most likely incidence angle in the horizontal plane given a frame of sound signal. We can derive the incidence angle α of the direct sound along a and the incidence angle β of the reflected echo along b. From the directions α and β, we can derive the angle $\theta_R = \alpha - \beta$. In the case of a 3D ellipsoid when the vertical angle φ_R is also considered, together with the travelled distances, the location of the reflection point $R(x, y, z)$ can be calculated using the following equations: $r = b \cdot SoS$, $x = r\cos\theta_R\sin\varphi_R$, $y = r\sin\theta_R\sin\varphi_R$, and $z = r\cos\varphi_R$, where SoS denotes the speed of sound. The distance between the reflection point and the microphone can be expressed as

$b = \frac{l(1-e^2)}{1+e\cos\theta_R\cos\varphi_R}$, where $l = \frac{(b+c)}{2}$ and $e = \frac{a}{2l}$. The distances a and $b+c$ are known through the corresponding times of arrivals. Note that if the source is on the body ($a \to 0$), the method is similar to sonar sensing.

Other reflection points can be derived the same way in other frames of the same RIR, and the RIR estimation is repeated several times a second. These points thus map the prominent surfaces in the space that cause these reflections.

5. EVALUATION

5.1. Experiment setup

We evaluated the echo orientation and position estimation approaches in a room of around $50\,m^2$ surrounded by white boards as shown in Fig. 3. The hearables were attached on the head of a mannequin whose movements were controlled by a wheeled robot. To acquire the ground truth in the experiments, we used a Vicon motion capture system [25]. The Vicon system captured the pose of the mannequin and the pose of the loudspeaker during the experiments in real time, by tracking the retro-reflective markers on top of the mannequin head and the loudspeaker. The loudspeaker repeatedly played unhearable sine sweep signals between $20\,kHz - 24\,kHz$ and used sampling rate $192\,kHz$.

(a) Scene1: planar screen   (b) Scene2: curved wall
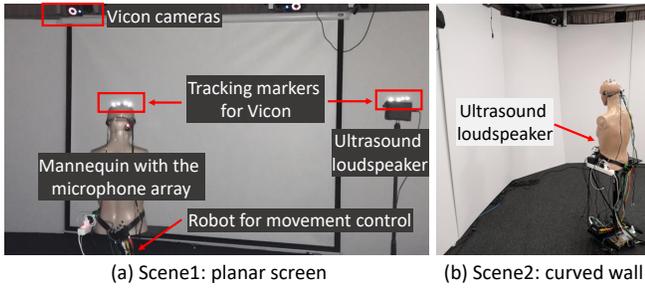
Fig. 3: The experiment setups with a Vicon-tracked mannequin wearing the microphone array, a robotic moving platform, an ultrasonic speaker, and reflective walls.

5.2. Incidence Angle Estimation

We evaluated the incidence angle estimation approach with the setup shown in Fig. 3a. A large projection screen was in front of the mannequin, and the ultrasound loudspeaker was fixed on a tripod placed in $1\,m$ to the right of the mannequin. We evaluated our algorithm via estimating the incidence angle of the direct sound from the loudspeaker to the microphones in the horizontal plane and in the vertical plane, respectively. We applied the RIR measurement technique by Farina [26] with exponential sine-sweep signals. This technique has been proven to provide higher robustness against environment noise compared to other methods (e.g., the Maximum Length Sequence technique) [27]. Hence, this technique has been commonly applied in related ultrasonic sensing work and we used it in our study as well. With this RIR measurement technique, one ultrasonic measurement period took $200\,ms$, so we updated the RIR measurements five times in a second. The first frame with the signal energy above a pre-defined threshold was selected as the frame of the direct sound.

5.2.1. Incidence Angle Estimation in the Horizontal Plane

In this experiment, we kept the loudspeaker still and controlled the robot to rotate the mannequin around (covering $360°$) for $40\,s$ in order to obtain various horizontal incidence angles. To be closer to normal human head movements, we included short pauses when rotating the mannequin.

Fig. 4 top shows that our estimations tracked the ground truth during the whole rotation trajectory. The average error of our method with respect to the ground truth was around $14°$ with a standard deviation of $16.9°$. The major source of the error might be the mapping from the ILD to the incidence angle, in particular when the mannequin rotated to positions where the microphones at the other side could not receive the sound signal directly. In addition, the pauses during the rotation process caused some jittering of the mannequin on the platform, which slightly influenced our RIR extraction method since the jittering led to marginal variations in the microphone positioning of our experimental setup. The impact of such a marginal microphone misalignment on the incidence angle estimation accuracy depends on the spacing between the microphones and the wavelength of interest. We anticipate that when the microphone positions are firmly fixed (e.g., on commercial products) with a properly designed distance between microphones, the error caused by the variation in the microphone positioning can be eliminated. Overall, even with the above limitations, we argue that our method is applicable for incidence angle tracking.
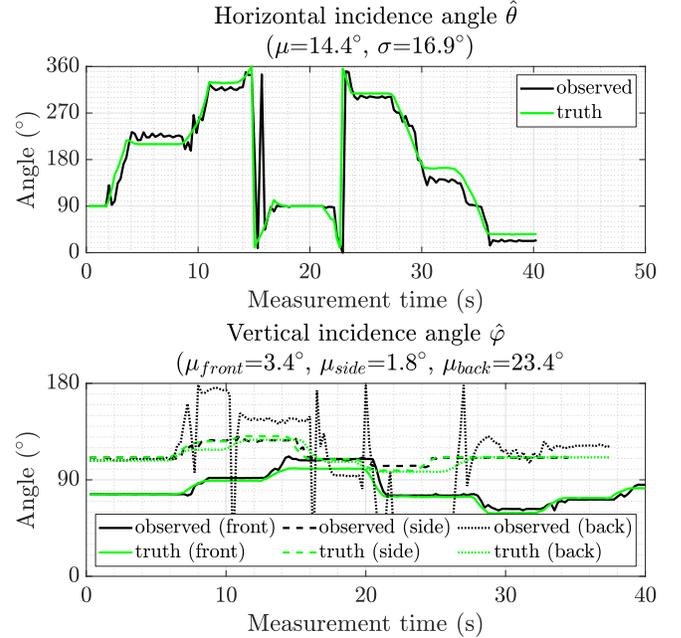
Fig. 4: Horizontal (top) and vertical (bottom, 3 cases) direct sound incidence angle estimates.

5.2.2. Incidence Angle Estimation in the Vertical Plane

In this experiment, we moved the tripod with the source up and down for $40\,s$ but kept the mannequin still because it could not move in the vertical direction. Remember that the vertical incidence angle is estimated from the Z microphone pair on the left ear, and this influences the range of detectable angles.

When this microphone pair faced towards the sound source or sidewards (when the nose of the mannequin faced towards the sound source), the average estimation error was as low as $3.4°$ and $1.8°$, and the corresponding standard deviations were $3.5°$ and $2.0°$. In these situations, the estimations aligned well with the ground truth as shown in Fig. 4 bottom.

However, when the microphone pair was separated from the sound source by the head, although the general tendency still matched the ground truth, we observed large fluctuations of the vertical angle estimation, resulting in an average error of $23.4°$ and a standard deviation of $24.7°$. This can be explained by the shielding effect of the head, which led to strong reflections overlapping with the direct sound from the loudspeaker. To handle this situation better, an additional Z microphone pair on the other ear would be required.

We conclude that to be able to reliably track any incidence angle, our hardware setup should be extended by another Z-pair of microphones on the other ear, but as one of them could be shared with the Y-pair, 3 microphones in a right triangle constellation on each side should be sufficient.

5.3. Reflection Position Estimation and Wall Mapping

We first tested our ellipse-based method in Scene 1 (Fig. 3a) with a planar projector canvas. We mounted the loudspeaker and the mannequin at the same height, both facing the canvas at a distance of $2\,m$. The loudspeaker was placed $2\,m$ to the right of the mannequin.

Among the first 30 frames of the received impulse response, we extracted frame 9 as the frame of direct sound (the first frame exceeding the energy threshold). Fig. 5a shows all reflection position estimates from frame 9 up to the last frame containing significant sound energy (frame 23). Each individual dot corresponds to a reflection detected in one frame of the RIR. The stronger reflection a frame had, the better we could use this frame to estimate the reflection position, and the corresponding dot is proportionally brighter in Fig. 5a (reflecting the confidence of the frame with the metric $w_{f,\mathrm{energy}}$).
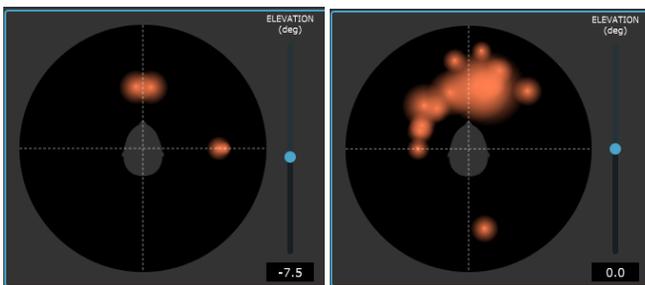
(a) Scene1: planar wall   (b) Scene2: curved wall

Fig. 5: Qualitative wall mapping around the head from estimated reflection points (cf. photos of the scenes in Fig. 3).

Quantitatively, the position estimation results of the reflection points are satisfying. From the most confident frames (the brightest dots in Fig. 5a), we observed a maximum error of $5\,cm$ out of $2\,m$ in the direction towards the wall. In the direction parallel to the wall, the reflection positions slightly varied from around $1\,m$ (at right) to $-0.5\,m$ (at left), which corresponded with the observed changes in the horizontal incidence angles along the planar wall. In Fig. 5a, we also visualized the position of the loudspeaker at the right of the

mannequin. It was detected at around $2\,m$ from the head, which corresponded to its real position.

The approach is not limited to planar walls, but also works with curved walls and varying loudspeaker positions. To test this, we positioned the mannequin in front of a curved wall as shown in Fig. 3b and mounted the loudspeaker directly at the belly of the torso $1\,m$ below the head. Fig. 5b shows that the estimated reflection points align well with the shape of the curved wall. The true distance to the wall was between $1.6\,m$ and $2.4\,m$. We measured this distance with an average error of about $10\,cm$ and a maximal error of $15\,cm$.

The method could be extended to 3D, and while it could also estimate the floor and ceiling reflections, we would need to disambiguate them from the wall reflections. It is therefore beneficial if the room has a carpet and a scattering ceiling so that we do not detect strong reflections from there and can focus on strong and distinct wall reflections.

We also note here a limitation of the RIR estimation algorithm. In these experiments, the mannequin with the microphone array and the environment were all static during individual RIR measurements. However, a quickly moving microphone array or a dynamic environment would violate the linear time-invariant (LTI) assumption, which can lead to distortion of the extracted RIR and in turn influence the reflection estimation algorithm. The extent of such a distortion depends on the extent of the LTI violation. Several measures might help to reduce such distortions, for example using shorter impulses so that less movement will occur between the impulse playbacks, or one might smoothly interpolate the sound signals that contain phase-shifted edges because of the distortion. General RIR estimation in dynamic environments requires advanced signal processing techniques that are outside the scope of this paper.

6. CONCLUSIONS AND FUTURE WORK

We proposed a simple wearable microphone array that could be integrated into hearables, and by using this microphone array, we presented real-time algorithms to estimate the positions of ultrasonic reflections in the room even without knowing the sound source position. These reflections enable us to coarsely capture the reflector geometry around the user as well as to derive the orientation and position of a user with respect to the reflector with satisfying accuracy at interactive rate.

We are currently working on extending this setup to track movements of the head with respect to the sound source and with respect to the reflective surfaces. Further work will focus on more advanced metrics to automatically select and dimension the frames containing the best sparse reflections, in order to improve the wall position estimation and track dynamic changes in the impulse response. We anticipate potential to apply the microphone array and the proposed algorithms in audio augmented reality applications. Furthermore, the hardware and the method could also be adapted for localizing smart objects in a room.

7. REFERENCES

[1] J. Yang, Y. Frank, and G. Sörös, "Hearing is believing: Synthesizing spatial audio from everyday objects to users," in *10th Augmented Human International Conference (AH)*, 2019.

[2] Z. Yang, Y. L. Wei, S. Shen, and R. R. Choudhury, "Ear-AR: Indoor acoustic augmented reality on earphones," in *26th International Conference on Mobile Computing and Networking (MobiCom)*, 2020.

[3] M. Krekovć, I. Dokmanć, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[4] M. Crocco, A. Trucco, and A. D. Bue, "Uncalibrated 3D room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, no. 18, 2017.

[5] M. Tikander, A. Härmä, and M. Karjalainen, "Binaural positioning system for wearable augmented reality audio," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[6] H. Gamper, S. Tervo, and T. Lokki, "Head orientation tracking using binaural headset microphones," in *Audio Engineering Society Convention 131*, 2011.

[7] R. Kapoor, S. Ramasamy, A. Gardi, and R. Sabatini, "A bio-inspired acoustic sensor system for UAS navigation and tracking," in *2017 IEEE/AIAA 36th digital avionics systems conference (DASC)*, 2017.

[8] S. Tervo and T. Tossavainen, "3D room geometry estimation from measured impulse responses," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[9] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, "Smartphone-based acoustic indoor space mapping," *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, 2018.

[10] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatMapper: Acoustic sensing based indoor floor plan construction using smartphones," in *15th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2017.

[11] O. Shih and A. Rowe, "Can a phone hear the shape of a room?," in *18th ACM International Conference on Information Processing in Sensor Networks (IPSN)*, 2019.

[12] I. Jager, R. Heusdens, and N. D. Gaubitch, "Room geometry estimation from acoustic echoes using graph-based echo labeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[13] X. Chen, H. Sun, and H. Zhang, "A new method of simultaneous localization and mapping for mobile robots using acoustic landmarks," *MDPI Applied Sciences*, vol. 9, no. 7, 2019.

[14] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, 2018.

[15] M. Lovedee-Turner and D. Murphy, "Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, 2019.

[16] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, 2011.

[17] H. Kim, L. Remaggi, P. J. Jackson, F. M. Fazi, and A. Hilton, "3D room geometry reconstruction using audio-visual sensors," in *2017 International Conference on 3D Vision (3DV)*, 2017.

[18] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[19] T. Wang, F. Peng, and B. Chen, "First order echo based room shape recovery using a single mobile device," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[20] J. Blauert, "Spatial hearing: The psychophysics of human sound localization," MIT Press, 1997.

[21] W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of high-frequency tones," *the Journal of the Acoustical Society of America*, vol. 29, no. 9, 1957.

[22] W. G. Gardner, *3-D Audio Using Loudspeakers*, Ph.D. thesis, Massachusetts Institute of Technology, 1997.

[23] J. Kaiser and R. Schafer, "On the use of the I0-sinh window for spectrum analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, 1980.

[24] J. D. Lawrence, *A Catalog of Special Plane Curves*, Courier Corporation, 1972.

[25] "Vicon motion capture system," in `www.vicon.com`.

[26] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, 2000.

[27] M. Holters, T. Corbach, and U. Zölzer, "Impulse response measurement techniques and their applicability in the real world," in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx)*, 2009.