JASA ARTICLE

Estimating direction of arrival in reverberant environments for wake-word detection using a single structural vibration sensor[a)]

Jenna Rutowski,[b)] Tre DiPassio, Benjamin R. Thompson, Mark F. Bocko, and Michael C. Heilemann
Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York 14624, USA

ABSTRACT:
The vibrational response of an elastic panel to incident acoustic waves is determined by the direction-of-arrival (DOA) of the waves relative to the spatial structure of the panel's bending modes. By monitoring the relative modal excitations of a panel immersed in a sound field, the DOA of the source may be inferred. In reverberant environments, early acoustic reflections and the late diffuse acoustic field may obscure the DOA of incoming sound waves. Panel microphones may be especially susceptible to the effects of reverberation due to their large surface areas and long-decaying impulse responses. An investigation into the effect of reverberation on the accuracy of DOA estimation with panel microphones was made by recording wake-word utterances in eight spaces with reverberation times (RT60s) ranging from 0.27 to 3.00 s. The responses were used to train neural networks to estimate the DOA. Within ±5°, DOA estimation reliability was measured at 95.00% in the least reverberant space, decreasing to 78.33% in the most reverberant space, suggesting an inverse relationship between RT60 and DOA accuracy. Experimental results suggest that a system for estimating DOA with panel microphones can generalize to new acoustic environments by cross-training the system with data from multiple spaces with different RT60s.
 https://doi.org/10.1121/10.0032367

(Received 29 April 2024; revised 18 September 2024; accepted 26 September 2024; published online 17 October 2024)
[Editor: Efren Fernandez-Grande] Pages: 2619–2629

I. INTRODUCTION

Determination of the direction of arrival (DOA) of sound waves is an essential signal processing tool in applications, such as acoustic source localization and tracking and acoustic beamforming. As an example, DOA estimation is the first step in methods designed to enhance the signal-to-noise and interference ratio for directional acoustic signals. Specifically, in smart speakers (internet-enabled loudspeakers that respond to voice commands), if the source direction for a received "wake-word" can be inferred, then interfering sounds from other directions may be suppressed via spatial filtering.

DOA estimation is required for smart speakers to be able to capture, recognize, and respond appropriately to user commands. Room reverberation can be a significant complicating factor for DOA estimation, and various strategies that attempt to mitigate the adverse effects of reverberation on DOA estimation have been developed.[1] Such methods have been employed in smart audio devices, but their implementation requires multi-microphone arrays and directional signal processing algorithms to analyze the temporal and spatial characteristics of the incoming signals.[2–6] These methods often incur high costs in power consumption, manufacturing, and computation.[5] Developing reliable DOA estimation techniques that are more efficient is important for optimal device performance.

An extensive review of DOA and acoustic source localization techniques was conducted by Bianco *et al.*[7] Most traditional systems utilize multi-microphone arrays to carry out DOA estimation and other advanced tasks, and it is assumed that the microphones in the array have relatively consistent, flat frequency responses with angle and short impulse responses.[8–14] Single-sensor panel microphones are fundamentally different from traditional microphone arrays due to their large surface areas, long ringdown times, and uneven frequency response with respect to angle of incidence.[15] These systems cover a large surface area but only record vibration data at a single point, making them more susceptible to the influences of reflections and reverberation, which could potentially exacerbate issues in DOA estimation and intelligibility for transcription compared to traditional systems. Because of this difference, surface-based, single-sensor DOA estimation systems present unique challenges compared to conventional multi-microphone setups. Using a single sensor for DOA and source localization purposes would represent a significant reduction in hardware and computational costs, so a detailed analysis of the limitations on the technology is an important step toward future commercialization.

Previous studies have demonstrated the feasibility of estimating the DOA of an incident wave using a single structural vibration sensor affixed to an elastic panel under

[a)]This paper is a continuation of work initially presented at the 186th Meeting of the Acoustical Society of America.
[b)]Email: jrutowsk@ur.rochester.edu

controlled, semi-anechoic conditions.[16] This is possible because the panel's bending modes exhibit different coupling characteristics to incoming acoustic waves depending on the angle of incidence. Deep neural networks (DNNs) can then be employed to estimate the incoming wave's DOA by analyzing the panel's response. While previous studies explored using a single sensor to capture directional information, the efficacy of these systems under more typical acoustic conditions remains uncertain. It is unclear how a system employing panel-based microphones will perform in acoustic environments with strong early reflections and long reverberation times. It has been shown that the performance of DOA estimation systems using conventional microphone arrays suffers as a function of the amount of environmental reverberation present.[17–19] It is as yet unknown whether single-sensor surface microphone-based systems are susceptible to the same issue and to what degree.

Given that one of the primary anticipated use cases for this technology is in smart speakers, it is essential that the speech recorded by these systems can be accurately transcribed. Previous research has shown that the long ringdown times and non-uniform frequency response of surface microphones do not hinder accurate transcription in non-reverberant environments.[16] However, it remains uncertain if this holds true when these systems are subjected to the long reverberation times and the poor clarity of realistic acoustic environments.

To be of practical utility, it is essential that DOA estimation systems based on panel microphones can adapt to diverse acoustic environments, particularly in residential settings where smart audio devices are commonly used.[20] The panel microphones in these systems must be able to capture speech with sufficient quality that it can be accurately transcribed. The focus of this work is to explore the performance of a single-sensor DOA estimation system in a range of acoustic environments.

We begin with an overview of the vibrations of an elastic panel excited by an incident acoustic wave and room acoustics analysis.

II. THEORETICAL DEVELOPMENT

A. Vibrational response of a baffled panel

For a damped, isotropic panel with Young's modulus E, Poisson's ratio ν, density ρ, thickness h, and mechanical loss factor b, excited by external load $p(x, y, t)$, the out-of-plane displacement w may be found using the equation from Cremer *et al.*:[21]

$$p(x,y,t) = \frac{Eh^3}{12(1-\nu^2)}\nabla^4 w(x,y,t) + b\dot{w}(x,y,t) + \rho h\ddot{w}(x,y,t). \tag{1}$$

The displacement $w(x, y, t)$ is separable as functions of space and time, given by

$$w(x,y,t) = \phi(x,y)e^{j\omega t}. \tag{2}$$

The spatial response $\phi(x, y)$ may be expressed as a superposition of the panel's bending modes $\Phi_r(x, y)$, where the rth mode has amplitude α_r, given by Fuller *et al.*[22] as

$$\phi(x,y) = \sum_{r=1}^{\infty} \alpha_r \Phi_r(x,y). \tag{3}$$

B. Panel response to an oblique pressure wave

To derive the response of a rectangular panel with dimensions (L_x, L_y) excited by incident pressure waves across various angles of acoustic incidence in the azimuthal plane, consider a plane wave with amplitude P_i incident on a panel. The resulting pressure distribution $P(x, y)$ on the surface is given by

$$P(x,y) = 2P_i e^{-jk\cos\theta_i x}, \tag{4}$$

where k is the wave number and θ_i is the angle between the projection of the propagation vector within the plane and the horizontal axis.

The modal shapes of a panel under simply supported boundary conditions are sinusoidal functions of space, with each mode characterized by a resonant frequency ω_r, given by Cremer *et al.*[21] as

$$\alpha_r = \frac{4}{\rho h L_x L_y\left(\omega^2 - \omega_r^2 - \frac{j\omega_r\omega}{Q_r}\right)} \times \int_0^{L_x}\int_0^{L_y} P(x,y)\Phi_r(x,y)\, dy\, dx, \tag{5}$$

where the quality factor Q_r determines the bandwidth of each mode in terms of the resonant frequency w_r of the rth mode.

Following Refs. 23–25, the relative excitations of the panel's modes are functions of the pressure amplitude P_i, frequency ω, and incident angle θ_i of the acoustic azimuthal plane wave, given by

$$\alpha_r = \frac{8P_i I_{r_m}(\theta_i,\omega) I_{r_n}}{\rho h(\omega_r^2 - \omega^2 + j\omega_r\omega/Q_r)}, \tag{6}$$

where $I_{r_m}(\theta_i, \omega)$ and $I_{r_n}(\omega)$ are coupling factors resulting from evaluating the integral in Eq. (5) using the pressure distribution given in Eq. (4). These coupling factors delineate the correlation between the pressure distribution induced on the panel by the incident wave and the spatial response of individual modes expressed as

$$I_{r_m}(\theta_i,\omega) = \frac{m\pi\left[1-(-1)^m e^{-j\sin\theta_i(\omega L_x/c)}\right]}{m^2\pi^2 - \left[\sin\theta_i\left(\frac{\omega^2 L_x^2}{c^2}\right)\right]}, \tag{7}$$

$$I_{r_n} = \frac{n\pi\left[1-(-1)^n\right]}{n^2\pi^2}, \tag{8}$$

https://doi.org/10.1121/10.0032367

where m and n represent the modal indices, which are the number of half-wavelengths in the horizontal and vertical dimensions, respectively, and c is the propagation speed of the incident wave.

C. Recorded characteristics of induced vibrations

In this experiment, an acoustic source emits a signal $s(t)$ to a panel with a structural vibration sensor affixed at point (x_0, y_0) along its surface. The impulse response from the source to the sensor, $h_{\theta_i}(t)$, depends on the incident angle θ_i. Since the panel operates within the linear region,[25] the velocity response at the sensor's position can be expressed with convolution as

$$\dot{w}(x_0, y_0, t) = s(t) * h_{\theta_i}(t). \tag{9}$$

D. Acoustic characteristics of rooms

In the context of this experiment, the RT60 is approximated using the impulse response of each room.[26] For the purposes of this work, we adopted the convention of using the point at which the Schroeder curve dips 5 dB below its peak as the starting point for our RT60 calculation.[27]

The clarity (C_{80}) is a measure of the intelligibility of a sound signal. This early-to-late energy ratio, expressed in decibels, is the ratio of the integral of sound energy within the initial 80 ms, considered the "early phase," to the subsequent sound energy after 80 ms.[28]

The early decay time (EDT) reflects the interval it takes for the acoustic pressure level to decrease by 10 dB after the onset of the decay process. It is computed analogously to the RT60 but only corresponds to the slope of the Schroeder curve over the first 10 dB of reduction, providing insight into the promptness of the sound pressure reduction during this defined interval.[27]

To calculate these metrics at a specific frequency, analysis was performed using a frequency-domain representation of the impulse response rather than a time-domain representation. The impulse response is filtered at a specified center frequency using an octave band filter, the energy decay curve is calculated, and the metric at the chosen frequency is determined.

III. METHODS

A. Dataset

A total of 600 sentences containing the wake-word, "Hey Alexa," were recorded by male and female participants (300 each).[29] The popularity of this wake-word is attributed to its spectral complexity and high number of phonemes.[20] A collection of sentences typical of smart speaker interactions was used to introduce natural variations in inflection and pronunciation over the corpus. These recordings were conducted in an acoustically treated studio using a Shure SM58 microphone (Shure Inc, Chicago, IL) with a sample rate of 48 kHz and later downsampled to 16 kHz during processing. Subsequently, the sentences were trimmed to isolate and extract only the wake-word, "Hey Alexa," for DOA estimation.

B. Experimental setup and procedure

The experimental setup used to record the panel's response to each stimulus at various angles of incidence is shown in Fig. 1. A 0.003-m-thick acrylic panel with Young's modulus $E = 3.2$ GPa, Poisson's ratio $v = 0.35$, density $\rho = 1180\,\text{kg/m}^3$, and dimensions $(L_x, L_y) = (0.36\,\text{m}, 0.26\,\text{m})$ was mounted on a rotary table to allow the acoustic wave to be recorded at angles between $\theta = -90°$ to $\theta_i = +90°$.

A KEF LS50 loudspeaker was used to reproduce the excitation signal, and was positioned 2 m away from the center of the panel in Rooms 3–8.[30] The loudspeaker was positioned 1 m away from the center of the panel in Rooms 1 and 2 due to size constraints of the room. This difference in distance from the source to the panel was compensated for by adjusting the output gain of the speaker according to the inverse square law.

The panel was outfitted with a set of PCB Piezotronics U352C66 accelerometers.[31] While data were simultaneously recorded by many sensors during the experiment, only the data from one sensor were used for training and testing the neural network at a time, allowing for a comparison of the model's performance for various sensor locations. The results presented in this work utilized a sensor that was positioned off-center on the panel at position (x_0, y_0) as seen in Fig. 1. Compared to sensors positioned at the center of the panel, off-center sensors provide a more comprehensive representation of the panel's modal behavior, as they lie in the antinodal region of many of the panel's low-frequency bending modes.

The panel's response to each wake-word was recorded at each angle of acoustic incidence from $\theta = -90°$ to $\theta_i = +90°$ in 10° increments by moving the experimental setup shown in Fig. 1 into eight rooms with varying RT60s.

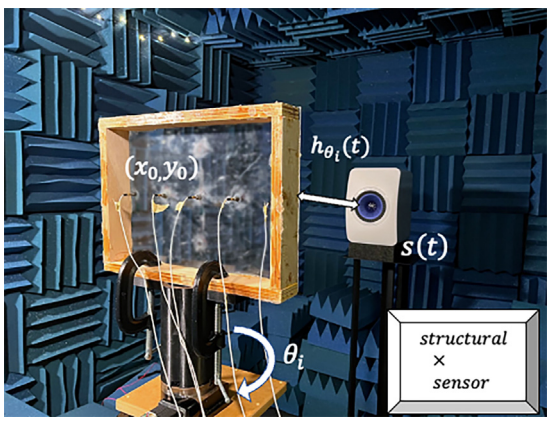

FIG. 1. (Color online) Experimental setup used to record the panel's response to incident pressure waves at varying angles of acoustic incidence. Although the panel is depicted with five sensors, only the structural sensor denoted by an x in the lower-right insert was used for DOA estimation.

The impulse response $h_{\theta_i}(t)$ from the loudspeaker to the panel was also recorded with 5° resolution. This was done using two maximum length sequence excitations sampled at 48 kHz, each with a 2 s duration, generated and interpreted by the MATLAB impulse response measurer app.[32] Using these recorded impulse responses, the response of the panel to wake-word excitation can be simulated using convolution. Utilizing a convolution-based approach significantly decreased the resources required for conducting this experiment at a higher angular resolution while also reducing the likelihood of dynamic noise interference during data collection. Ultimately, a dataset consisting of 22 200 wake-word recordings was created with various angles of incidence and a variety of acoustical conditions.

C. Room characteristics

A calibrated miniDSP UMIK-1 USB microphone was used to measure the characteristics of each room.[33] A Mackie SRT215 loudspeaker (Mackie, Bothell, WA) was used to produce the excitation signal, which was a 5.5-s-long swept sine sampled at 48 kHz.[34] Three different loudspeaker positions and three different microphone positions were used, yielding a total of nine microphone and speaker configurations to ensure consistency of the measured data. The impulse response data were recorded by Room EQ Wizard and then exported to MATLAB to calculate RT60, C_{80}, and EDT for each room following ISO 3382 standards.[35–37]

The rooms in which experiments were conducted can be seen in Fig. 2, which provides a visual representation of their layout and dimensions. The characteristics of each room can be found in Table I. Room 1 was a (2.4 m × 3.0 m × 2.4 m) rectangular-shaped Whisper Room.[38] Room 2 was a (2.7 m × 3.4 m × 2.4 m) rectangular-shaped mixing room with acoustic treatment. Room 3 was a (8.4 m × 15.2 m × 2.7 m) rectangular-shaped classroom with acoustic treatment and no windows. Room 4 was a (7.9 m × 13.4 m × 2.7 m) rectangular-shaped classroom with windows and no acoustic treatment. Room 5 was a (7.8 m × 14.0 m × 7.3 m) live room for a recording studio with sloped ceilings. Room 6 was a (9.1 m × 13.4 m × 5.5 m) lecture hall with many windows and angled walls. Room 7 was a (15.5 m × 7.8 m × 6.9 m) atrium with many windows and a sloped ceiling. Room 8 was a large rotunda with a radius of 9.1 m and a ceiling height of 24.4 m. These rooms are organized in order of increasing RT60.

D. Spectral features

Prior studies have shown that linear spectrograms (STFTs), mel spectrograms, and mel frequency cepstral coefficients (MFCCs) are effective feature sets for speech analysis.[39] These feature sets are also effective for estimating the DOA of wake-word recordings captured by a single sensor on an elastic panel under semi-anechoic conditions.[40] Among these, STFTs emerged as the best-performing feature set, but were also the largest feature vector tested.[41] Consequently, all three of these feature sets were extracted and used to train three different neural networks, one for each feature set, to compare DOA accuracy for feature sets of different sizes.[16]

E. Network architecture

The two model architectures utilized in this work are illustrated in Figs. 3 and 4. The first is a two-dimensional convolutional neural network (CNN) with a regression output layer, shown in Fig. 3. The second is a feedforward neural network (FNN) with a regression output layer, shown in Fig. 4. While deploying these models on the embedded

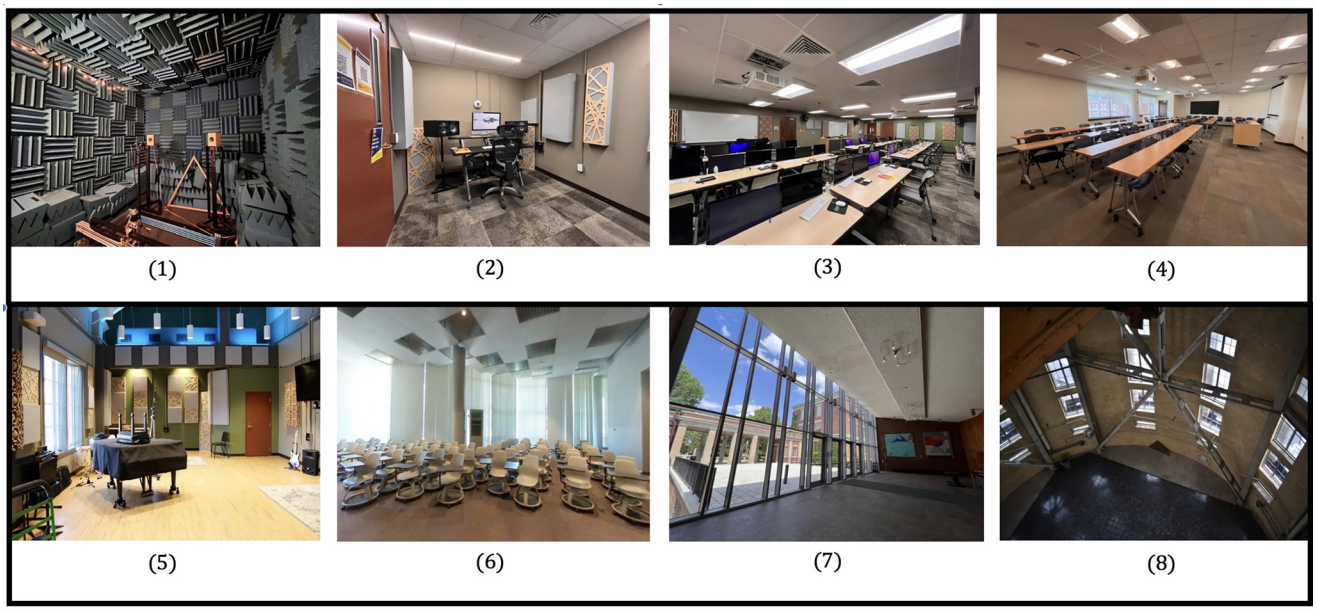

FIG. 2. (Color online) Layout of each room used in this experiment. Room numbers correspond to the described spaces in Table I and the corresponding paragraph in Sec. III C. The complete dataset of room impulse responses is available for reference and download via the University of Rochester Research Repository (URRR) (Ref. 48).

https://doi.org/10.1121/10.0032367

TABLE I. Measured acoustic properties of each room. Data measured using Room EQ Wizard and results calculated in MATLAB.

Room	RT60 (s)	Volume (m^3)	C_{80} @ 1 kHz (dB)	EDT (s)
1	0.3	17.28	31.0	0.18
2	0.5	22.03	22.6	0.24
3	0.6	344.74	15.5	0.54
4	0.9	285.82	10.0	0.64
5	1.2	797.2	3.9	0.86
6	1.6	670.67	9.0	1.40
7	2.0	834.21	−4.0	2.00
8	3.0	6347.78	−5.0	3.30

systems that are typically found in smart audio devices is out of the scope of this study, this practical implementation is an important design consideration for future work. Similar models have shown promise for estimating the DOA of speech signals when deployed on an embedded system.[42]

Two different network architectures were tested to determine whether one would significantly outperform the other and to verify general network performance. Between the two regression architectures, the CNN does slightly outperform the FNN. However, further testing would be required to verify if this is true in general once each network architecture has been optimized.

Distinct instances of both architectures were trained with each of the STFT, mel spectrogram, and MFCC feature sets previously described. Training utilized 4560 wake-word recordings from each room partitioned into training and validation sets at an 80:20 ratio. The remaining 1140 recordings per room were reserved for testing each model. Model training using the wake-word data synthesized from the measured room impulse responses was also performed, with 8880 wake-word recordings from each room split into training and validation sets with a ratio of 80:20. The remaining 2220 responses were used to test each model. Training was conducted with a batch size of 512 for 100 epochs and a dropout rate of 0.2.

These models leveraged the extracted features to minimize the MSE between the predicted incident angle and the ground truth angle, θ_i, within $\pm\Delta\theta_i$ of angular tolerance for each room or combination of rooms. The models in this study were trained individually with a single voice, making them inherently speaker-dependent as a proof of concept.

F. Evaluation

Two evaluations were conducted during this experiment. The first was to determine how DOA accuracy was affected by the room reverberation. The second was to determine whether the recorded signal, now processed through the transfer functions of both the room and panel, maintained enough intelligibility that the panel microphone was still useful for speech recognition.

Following Ref. 43, the reliability of the DOA estimation can be expressed as

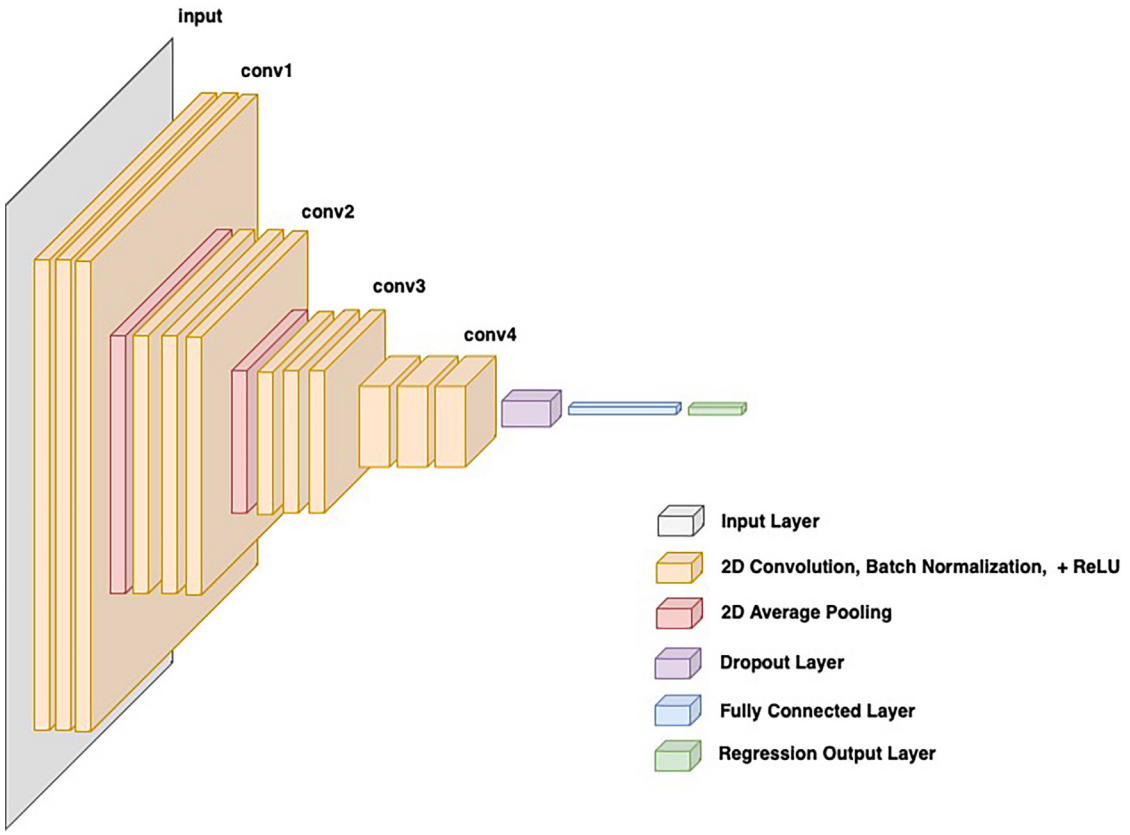

FIG. 3. (Color online) CNN architecture implemented in MATLAB.

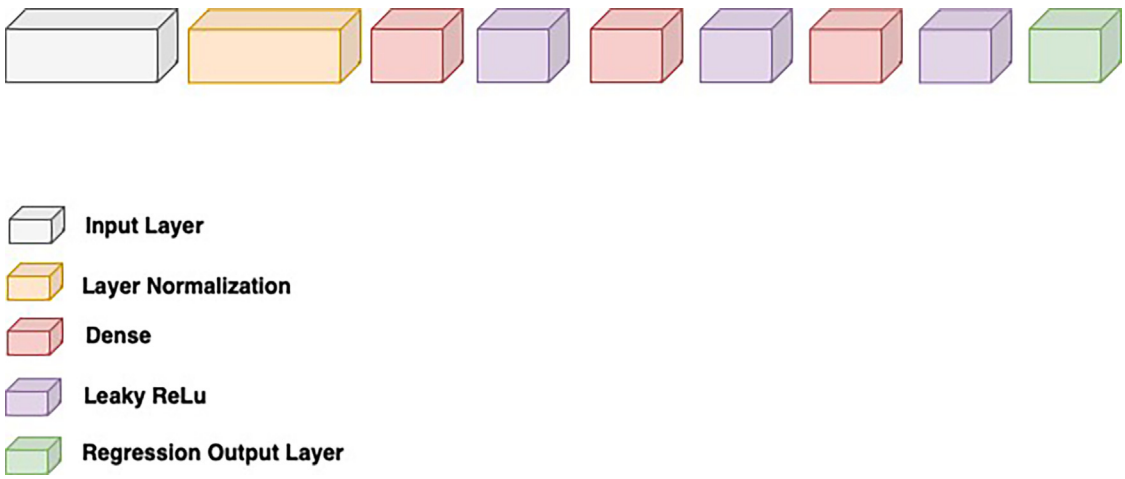

FIG. 4. (Color online) FNN architecture implemented in MATLAB.

$$\frac{\text{Number of Correct Predictions}}{\text{Total Number of Utterances Tested}}. \tag{10}$$

Results for angular tolerances are presented within $\pm 5°$, $\pm 10°$, and $\pm 20°$, consistent with previous work.[16,42]

IV. RESULTS

The DOA results, utilizing both recorded and synthesized reverberant audio data, are depicted in Fig. 5. Discrepancies between the synthesized reverberant dataset and the recorded dataset can be attributed to the variable background noise during the recording of the datasets. Since the difference in DOA accuracy for all reverberant environments between the two datasets is less than 1.9% for the male voice and 4.7% for the female voice, we report only the results from the synthesized dataset, which contains data from additional rooms, for brevity.[44] Similarly, since the CNN and FNN architectures also displayed an accuracy difference of less than 2.0% in Room 1 and 3.0% in Room 8 as shown in Fig. 6, all future outcomes are presented using the CNN architecture.

Models trained using more spectrally complete feature sets gave more accurate predictions than more compact feature sets with lower spectral resolution as seen in Fig. 7. Additionally, the differences between the male and female speakers increase as the spectral resolution decreases, and as the reverberation time increases. The models trained using linear spectrogram feature sets outperformed those trained with mel spectrogram feature sets, which outperformed models trained with MFCC feature sets. Given that this work centers on investigating the impact of reverberation on the accuracy of DOA estimation, we will focus on reporting the results obtained using the highest-performing STFT feature set.

The effect of angular tolerance on DOA accuracy is shown in Fig. 8. With $\pm 5°$ tolerance, the model achieved up to 95.00% reliability in the least reverberant environment, decreasing to 78.33% in the most reverberant environment.

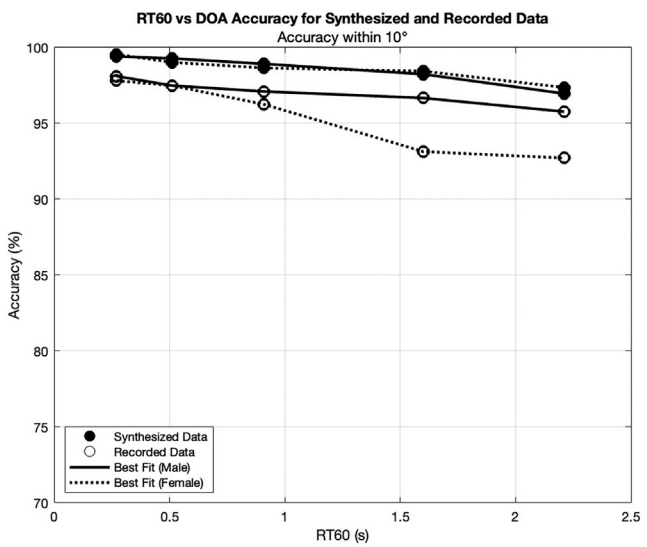

FIG. 5. DOA accuracy within $\pm 10°$ using STFT extraction on synthesized and recorded data for male and female voice.

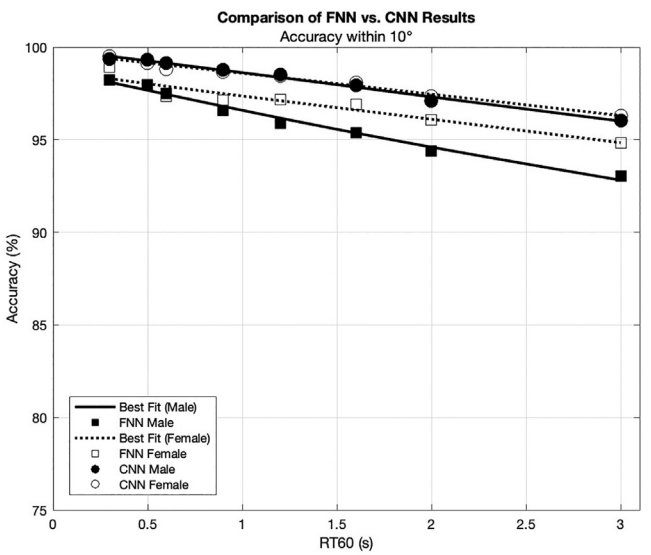

FIG. 6. FNN vs CNN DOA accuracy results within $\pm 10°$ using STFT extraction.

https://doi.org/10.1121/10.0032367

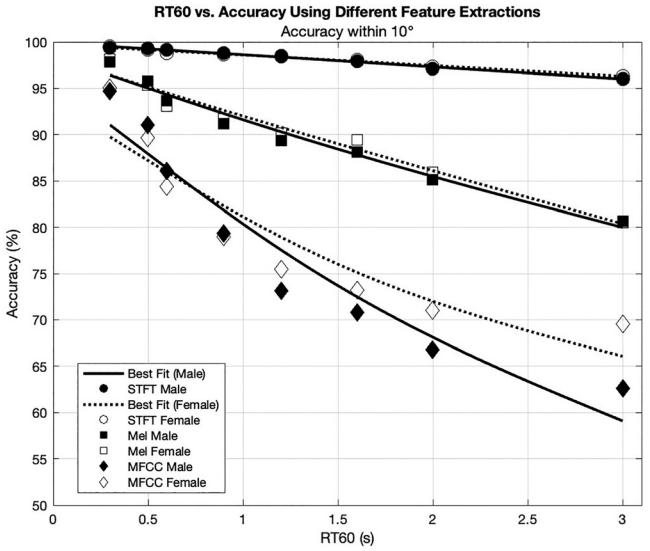

FIG. 7. The reliability of the CNN in estimating DOA within ±10° angular tolerance using multiple feature sets across eight rooms with varying RT60s.

With ±10° tolerance, the model achieved up to 99.80% reliability in the least reverberant environment, decreasing to 96.28% in the most reverberant environment. Results shown in Table II indicate an inverse relationship between DOA accuracy and RT60.

These findings suggest that an increase in RT60 does not markedly impede performance if the angular tolerance is at least 10 degrees, underscoring the robust suitability of this single-sensor DOA method for practical applications in reverberant environments. All of the trained models were able to estimate the DOA of both the male and female voices in all eight rooms within ±10° with greater than 96% reliability. The estimation accuracy is upwards of 98% in all cases within 10 degrees for rooms with RT60 < 1.5 s, suggesting high accuracies are achievable for in-home settings, which represent the most common use cases.

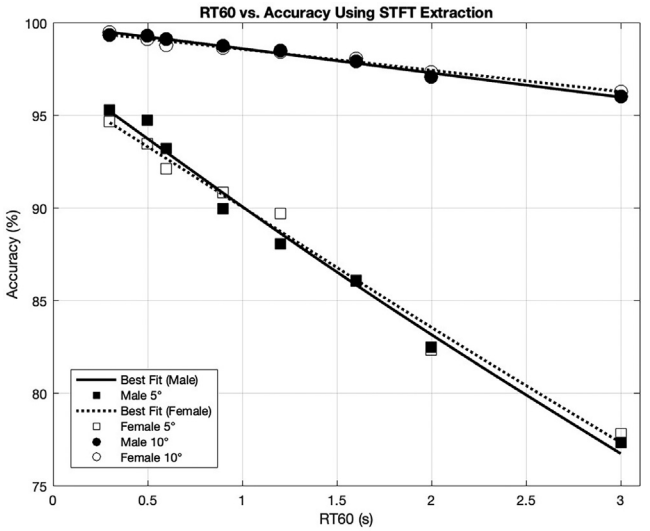

FIG. 8. The reliability of the CNN in estimating DOA within ±5° and ±10° angular tolerances using STFT feature sets across eight rooms with varying RT60s.

Analyzing the results across different angular tolerances indicates that the models' DOA estimates are dispersed about the ground truth incident angles. This is evident in the illustrated confusion matrix in Fig. 9, which is the representation of data pertaining to the male voice under an angular tolerance of ±5° for the room and resolution footnoted in Table II. This trend persists across varying angular tolerances, acoustic environments, and voices.

V. GENERALIZATION

Each of the above models was trained and tested on data collected in the same room. A robust system must be able to generalize to rooms with different reverberation times. The following section contains an analysis exploring how models trained using data from rooms with varying RT60s can be generalized to different spaces.[45]

An initial assessment was made using models trained on data from seven of the rooms, and tested on the remaining room not included in the training set. For example, the model used to test Room 1 was trained using data from Rooms 2–7. The results are shown in Table III. In this case, data resizing was employed to ensure that the model was trained with the same volume of data as the models trained and tested in the previous section. Results indicate DOA estimation utilizing panel microphones can effectively adapt to diverse acoustic environments by cross-training the system with training data from multiple spaces with different reverberation times. DNNs trained by extracting features from the vibration response of the panel were able to estimate the DOA of speech signals in a novel environment within ±10° with a reliability of up to 92.15% in the least reverberant environment, decreasing to 86.79% in the most reverberant environment.

The same process was then repeated without data resizing, meaning that the model was trained with seven times the data compared to all previous cases. Results shown in

TABLE II. Reliability of the DOA estimates made by the trained CNNs for rooms of increasing RT60 with angular tolerances of ±5°, ±10°, and ±20°. Distinct models were trained for each speaker using STFTs as the extracted feature vectors.

		DOA accuracy within:					
Room	RT60 (s)	5°	10°	20°	5°	10°	20°
1	0.27	95.00[a]	99.80	100	94.98	99.55	99.80
2	0.51	94.9	99.50	99.95	92.74	99.02	99.59
3	0.57	94.59	99.08	99.95	92.58	98.88	99.45
4	0.91	88.69	98.92	99.95	90.84	98.65	98.89
5	1.05	87.88	98.24	99.91	88.96	98.44	98.61
6	1.6	87.61	98.22	99.95	87.98	98.12	98.25
7	2.21	80.80	96.96	99.59	79.90	97.36	97.79
8	3.00	78.33	96.06	99.5	78.76	96.28	97.48
			Male			Female	

[a]See Refs. 49 and 50.

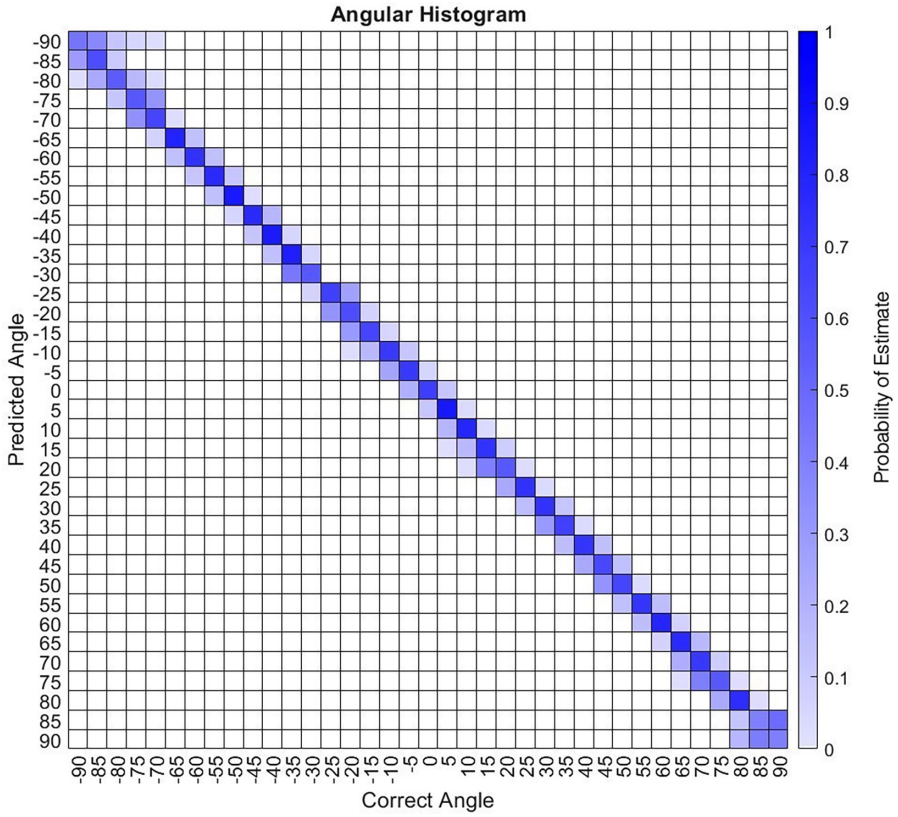

FIG. 9. (Color online) Confusion matrix within ±5° for Room 1 DOA accuracy.

Table IV demonstrate that preexisting datasets can effectively support the use of this technology for smart audio devices in environments marked by varying RT60s. Results improve when the model is provided additional training data. DNNs trained with additional data were able to estimate the DOA of speech signals in a novel environment within ±10° with a reliability of up to 99.53% in the least reverberant environment, decreasing to 98.24% in the most reverberant environment.

It is important to recognize some limitations of this experiment. Each model was trained and tested using a single participant's voice at a time. The current scope excludes the development of a speaker-independent model. Because the reported results from the trained models are relatively consistent between both the male and female voices with varying inflections, it is expected this proposed single-sensor DOA method can adapt to both diverse environments, as shown, but also to diverse speech characteristics. Any discrepancies, such as the slight underperformance of the recorded data for the female voice compared to the male voice, were beyond the scope of this study and warrant further investigation in future research. In addition, the sensor

TABLE III. Generalizing RT60 training data: Resized. Assessment of CNN's DOA estimation reliability in a generalized setting, trained on data from seven rooms and tested on an independent room. Data resizing ensured consistent training data volume as in previous cases.

Training rooms: All except listed testing room							
		DOA accuracy within: (%)					
Test room	RT60 (s)	5°	10°	20°	5°	10°	20°
1	0.27	85.98	92.15	92.90	85.55	91.98	92.68
2	0.51	88.23	92.22	93.00	88.63	92.71	93.45
3	0.57	90.62	92.77	93.14	90.12	92.47	93.06
4	0.91	88.11	92.18	92.95	88.91	92.36	93.65
5	1.05	87.02	92.20	92.92	87.42	92.92	93.71
6	1.6	84.62	91.09	91.44	84.16	90.09	91.11
7	2.21	80.12	89.88	90.56	79.08	88.79	89.28
8	3.00	78.72	86.79	88.77	79.65	88.12	89.10
			Male			Female	

TABLE IV. Generalizing RT60 training data: Not resized. Assessment of CNN's DOA estimation reliability in a generalized setting, trained on data from seven rooms and tested on an independent room. Data not resized such that the model received additional data compared to previous models.

Training rooms: All except listed testing room							
		DOA accuracy within: (%)					
Test room	RT60 (s)	5°	10°	20°	5°	10°	20°
1	0.27	97.30	99.53	99.95	96.42	98.98	99.90
2	0.51	97.50	99.62	100.00	98.89	99.55	100.00
3	0.57	96.57	99.82	100.00	95.97	99.91	100.00
4	0.91	94.68	99.55	100.00	93.82	99.86	100.00
5	1.05	94.23	98.96	99.95	95.11	98.99	99.98
6	1.60	94.05	99.32	99.95	93.21	99.16	99.90
7	2.21	88.64	98.33	99.86	88.53	98.19	99.79
8	3.00	88.38	98.24	99.82	88.44	98.10	99.68
			Male			Female	

https://doi.org/10.1121/10.0032367

location was chosen with the arbitrary constraint of being off-center, so as to lie in the antinodal region of a plurality of low-frequency bending modes. There is almost certainly a sensor location on the panel that provides stronger coupling to the particular set of modes relevant for determining the incident angle of an acoustic wave.

VI. INTELLIGIBILITY

To assess the system's intelligibility in reverberant environments, the impulse responses from each space were convolved with the complete dataset of "Hey Alexa" phrases as outlined in Sec. III A to create a synthesized dataset. Subsequently, these recordings underwent transcription using IBM Watson's speech-to-text automated speech recognition service, and a word error rate (WER) score was calculated by comparing the transcribed text against the transcripts of the actual spoken phrases.[46] The WER metric quantifies the Levenshtein distance between the transcription and the known text, incorporating errors, such as word insertions, deletions, and substitutions. It is represented as a percentage as

$$\mathrm{WER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Number of Words in Reference}} \times 100\%. \tag{11}$$

WER analysis was conducted on recordings captured by the reference microphone to establish the baseline error introduced by the automated speech recognition system under ideal conditions. The WER values obtained for the panel microphones were then compared to the incremental increase in WER observed in the reference case.

The results shown in Table V indicate that rooms with higher reverberation adversely affect the WER. Though the WER of the reference microphone outperforms the WER panel microphone in each space tested, the difference is never more than 3.8%. This implies that room reverberation is more important for predicting transcription accuracy, and that structural sensors can replace conventional microphones in speech-to-text applications with only a small reduction in performance.

TABLE V. WER results. Shown are results for dry reference sentences, sentences recorded in each room using a reference microphone, and sentences recorded in each room using the panel microphone.

		WER (%)	
Room	RT60 (s)	Reference microphone	Panel microphone
Dry reference	NA	1.06	NA
1	0.3	2.2	3.14
2	0.5	3.1	5.5
3	0.6	6.2	7.9
4	0.9	7.8	9.3
5	1.2	8.7	10.4
6	1.6	9.1	12.6
7	2.0	12.3	16.1
8	3.0	14.7	16.9

VII. CONCLUSION

The results in this work demonstrate that compact feature vectors informed by the resonant properties of a panel surface are sufficient for reliable DOA estimation using a single structural sensor in reverberant environments. DNNs trained by feature vectors extracted from the vibration response of the panel excited by an incident acoustic speech signal were able to estimate the DOA within $\pm 5^\circ$ with a reliability of up to 95.00% in a space with a reverberation time of 0.3 s, decreasing to 78.33% in a space with a reverberation time of 3.0 s. The accuracy improved with increased angular tolerance. Notably, the trained models were able to estimate the DOA of both male and female voices within $\pm 10^\circ$ in all eight rooms tested with greater than 96% reliability.

The utilization of panel microphones for DOA estimation showcases a robust capability to adapt across diverse acoustic environments. This adaptability implies that existing speech data corpora can be effectively integrated with this type of system, eliminating the need for new dataset recordings. Extensive collections of recorded speech from various speakers and environments which encompass a wide range of speech patterns and background noises, which have already been compiled by large companies for developing smart speakers, can continue to be utilized. By employing these diverse datasets, high levels of accuracy and reliability in DOA estimation can be achieved even in novel acoustic conditions, as results continue to improve when the model is provided an increased amount of training data. Generalized systems trained with additional data were able to estimate the DOA of speech signals in a new environment that was not part of the training set to within $\pm 10^\circ$ with a reliability of at least 98.24%.

Acknowledging the impact of reverberation on the panel's performance is vital, yet it does not markedly hinder its practical application. Results consistently achieved DOA estimation accuracy exceeding 98% within 10 degrees in environments with RT60s typical of living rooms, and the WER of transcribed speech signals was only 3.8% greater than those recorded by conventional reference microphones. This implies that panel microphones are a viable alternative to traditional microphone arrays for estimating acoustic DOA in smart audio devices. Leveraging the vibration response of a panel proves to be an accurate and more efficient method for DOA estimation compared to traditional approaches that rely on multiple sensors.[47]

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Grant No. 2104758.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The datasets that support the findings of this study are openly available in the University of Rochester Research Repository at https://doi.org/10.60593/ur.d.26417548.v1 and https://doi.org/10.60593/ur.d.26801089.v1. Additional data that support the findings of this study are available from the corresponding author upon reasonable request.

[1]J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, "DOA estimation of audio sources in reverberant environments," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016).

[2]C. M. Zannini, A. Cirillo, R. Parisi, and A. Uncini, "Improved TDOA disambiguation techniques for sound source localization in reverberant environments," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (2010).

[3]H. R. Abutalebi and H. Momenzadeh, "Performance improvement of TDOA-based speaker localization in joint noisy and reverberant conditions," EURASIP J. Adv. Signal Process. **2011**, 621390.

[4]R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Lee, and H.-M. Park, "Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments," IEEE Access **8**, 7371–7373 (2020).

[5]C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in low-noise, reverberative environments?," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008).

[6]P. Stoica and K. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," IEEE Trans. Acoust. Speech Signal Process. **38**(7), 1132–1143 (1990).

[7]M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628 (2019).

[8]F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," IEEE Trans. Audio. Speech Lang. Process. **20**(5), 1449–1460 (2012).

[9]I. L. Freire and J. A. Apolinário, "DOA of gunshot signals in a spatial microphone array: Performance of the interpolated generalized cross-correlation method," in *2011 Argentine School of Micro-Nanoelectronics, Technology and Applications* (2011), pp. 1–6.

[10]T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," Signal Process. **85**(1), 177–204 (2005).

[11]M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," IEEE Trans. Speech Audio Process. **5**, 45–50 (1997).

[12]H. Chen and W. Ser, "Sound source DOA estimation and localization in noisy reverberant environments using least-squares support vector machines," J. Signal Process. Syst. **63**, 287–300 (2011).

[13]J. Wuth, R. Mahu, I. Cohen, R. M. Stern, and N. B. Yoma, "A unified beamforming and source separation model for static and dynamic human-robot interaction," JASA Express Lett. **4**(3), 035203 (2024).

[14]J. Escolano, N. Xiang, J. M. Perez-Lorenzo, M. Cobos, and J. J. Lopez, "A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array," J. Acoust. Soc. Am. **135**(2), 742–753 (2014).

[15]T. DiPassio, M. C. Heilemann, and M. F. Bocko, "Audio capture using structural sensors on vibrating panel surfaces," J. Audio Eng. Soc. **70**(12), 1027–1037 (2022).

[16]T. DiPassio, M. C. Heilemann, B. Thompson, and M. F. Bocko, "Estimating acoustic direction of arrival using a single structural sensor on a resonant surface," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023).

[17]N. Kumar and A. Singh, "Effect of reverberation on different DOA estimation techniques using microphone array," Int. J. Technol. **3**(7), 166–170 (2017).

[18]M. Cobos, J. J. Lopez, and S. Spors, "Analysis of room reverberation effects in source localization using small microphone arrays," in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)* (2010), pp. 1–4.

[19]D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," J. Acoust. Soc. Am. **128**(4), 1800–1811 (2010).

[20]R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," IEEE Signal Process. Mag. **36**(6), 111 (2019).

[21]L. Cremer, M. Heckl, and B. Petersson, *Structure-Borne Sound Structural Vibrations and Sound Radiation at Audio Frequencies* (Springer, Berlin, Heidelberg, 2005).

[22]C. Fuller, S. Elliott, and P. Nelson, *Active Control of Vibration* (Elsevier Science, Amsterdam, 1996), available at https://books.google.com/books?id=HGP4iGWhAdEC.

[23]B. Wang, C. R. Fuller, and E. K. Dimitriadis, "Active control of noise transmission through rectangular plates using multiple piezoelectric or point force actuators," J. Acoust. Soc. Am. **90**(5), 2820–2830 (1991).

[24]T. Sors and S. Elliot, "Volume velocity estimation with accelerometer arrays for active structural acoustic control," J. Sound Vib. **258**(5), 867–883 (2002).

[25]F. Fahy, *Sound and Structural Vibration: Radiation, Transmission and Response* (Elsevier, Amsterdam, 2007).

[26]M. R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am. **37**(3), 409–412 (1965).

[27]V. L. Jordan, "Acoustical criteria for auditoriums and their relation to model techniques," J. Acoust. Soc. Am. **47**, 408–412 (1970).

[28]M. Larrosa Navarro, D. de la Prida Caballero, and A. Pedrero, "Influence of musical stimulus on the perception of clarity in rooms and its relation to C_{80}," Appl. Acoust. **208**, 109370 (2023).

[29]The complete dataset of wake-word phrases is available for reference and download via the University of Rochester Research Repository (URRR).

[30]https://us.kef.com/collections/ls50 (Last viewed 12/1/2023).

[31]https://www.pcb.com (Last viewed 12/1/2023).

[32]https://www.mathworks.com/help/audio/ref/impulseresponsemeasurer-app.html (Last viewed 7/17/2024).

[33]https://www.minidsp.com (Last viewed 12/1/2023).

[34]https://mackie.com (Last viewed 12/5/2023).

[35]J. Mulcahy, "Room EQ Wizard," https://www.roomeqwizard.com (Last viewed 12/7/2023).

[36]C. L. Christensen, G. Koutsouris, and J. H. Rindel, "The ISO 3382 parameters: Can we simulate them? Can we measure them," in *Proceedings of the International Symposium on Room Acoustics*, Toronto, Ontario, Canada (2013), Vol. 910.

[37]ISO, "Measurement of room acoustic parameters, Part 1" (International Organization for Standardization, Geneva, Switzerland, 2009).

[38]https://whisperroom.com (Last viewed 4/14/2024).

[39]K. M. Shiva Prasad, G. N. Kodanda Ramaiah, and M. B. Manjunatha, "Speech features extraction techniques for robust emotional speech analysis/recognition" (2017), available at https://ischolar.sscldl.in/index.php/indjst/article/view/139145 (Last viewed 3/33/2024).

[40]T. DiPassio, M. C. Heilemann, and M. F. Bocko, "Direction of arrival estimation of an acoustic wave using a single structural vibration sensor," J. Sound Vib. **553**, 117671 (2023).

[41]J. Oruh and S. Viriri, "Spectral analysis for automatic speech recognition and enhancement," in *Machine Learning for Networking*, edited by É. Renault, S. Boumerdassi, and P. Mühlethaler (Springer, Cham, Switzerland, 2021), pp. 245–254.

[42]T. DiPassio, M. Heilemann, B. Thompson, and M. Bocko, "Estimating the direction of arrival of a spoken wake word using a single sensor on an elastic panel" (2023), pp. 1–5.

[43]L. Qinglong, X. Zhang, and H. Li, "Online direction of arrival estimation based on deep learning" (2018).

[44]P. Guidorzi, L. Barbaresi, D. D'Orazio, and M. Garai, "Impulse responses measured with MLS or swept-sine signals applied to architectural acoustics: An in-depth analysis of the two methods and some case studies of measurements inside theaters" (2015), Vol. 78.

[45]G. Cohen, G. Sapiro, and R. Giryes, "DNN or K-NN: That is the generalize vs. memorize question" (2019).

[46]IBM, "IBM Watson speech to text," https://www.ibm.com/products/speech-to-text (Last viewed 2/7/2024).

[47]L. Weng, X. Song, Z. Liu, X. Liu, H. Zhou, H. Qiu, and M. Wang, "DOA estimation of indoor sound sources based on spherical harmonic domain beam-space MUSIC," Symmetry **15**(1), 187 (2023).

[48]J. Rutowski, "University of Rochester room impulse response dataset," University of Rochester Research Repository (URRR) (2024), available at https://rochester.figshare.com/articles/dataset/University_of_Rochester_room_impulse_response_dataset/26801089.
[49]T. DiPassio, M. Heilemann, J. Rutowski, P. Sedlacek, B. Thompson, and Y. Wen, "Smart speaker commend dataset," (2024), available at https://rochester.figshare.com/articles/dataset/Smart_Speaker_Command_Dataset/26417548.
[50]T. Dipassio, M. Heilemann, B. Thompson, and M. Bocko, "Estimating acoustic direction of arrival using a single structural sensor on a resonant surface," in *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York, 2023).