ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095986

ESTIMATING ACOUSTIC DIRECTION OF ARRIVAL USING A SINGLE STRUCTURAL SENSOR ON A RESONANT SURFACE

Tre DiPassio, Michael C. Heilemann, Benjamin Thompson, Mark F. Bocko*

Department of Electrical and Computer Engineering, University of Rochester, USA

ABSTRACT

The direction of arrival (DOA) of an acoustic source is a signal characteristic used by smart audio devices to enable signal enhancement algorithms. Though DOA estimations are traditionally made using a multi-microphone array, we propose that the resonant modes of a surface excited by acoustic waves contain sufficient spatial information that DOA may be estimated using a singular structural vibration sensor. In this work, sensors are affixed to an acrylic panel and used to record acoustic noise signals at various angles of incidence. From these recordings, feature vectors containing the sums of the energies in the panel's isolated modal regions are extracted and used to train deep neural networks to estimate DOA. Experimental results show that when all 13 of the acrylic panel's isolated modal bands are utilized, the DOA of incident acoustic waves for a broadband noise signal may be estimated by a single structural sensor to within $\pm 5^{\circ}$ with a reliability of 98.4%. The size of the feature set may be reduced by eliminating the resonant modes that do not have strong spatial coupling to the incident acoustic wave. Reducing the feature set to the 7 modal bands that provide the most spatial information produces a reliability of 89.7% for DOA estimates within $\pm 5^{\circ}$ using a single sensor.

Index Terms— Direction of arrival (DOA), Vibration Sensing, Audio Feature Extraction, Structural Sensors

1. INTRODUCTION

Smart audio devices employ microphone arrays to record acoustic sources [1]. As these devices typically exist in chaotic, noisy, and untreated environments, signal enhancement must be applied to the recorded signal before it is interpreted by a smart-home service to ensure that the user's request is executed correctly.

One such signal enhancement method commonly employed in smart devices is acoustic beamforming, which aims to improve the signal-to-noise ratio of the recorded audio signal by attenuating sounds that do not arrive from the direction of the desired source [2]. Focusing the beam pattern toward the desired source location requires an estimation of the signal's direction of arrival (DOA). Techniques such as inter-sensor time difference of arrival (TDOA), correlative techniques such as generalized cross-correlation with phase transform (GCC-PHAT), and the multiple signal classification algorithm [3, 4, 5, 6] are widely used for DOA estimation. These techniques each require the simultaneous measurement of the arriving wave at multiple points in space. While the estimation accuracy may be improved by increasing the spatial resolution of the microphone array, this comes higher power consumption, increased hardware expense, and additional computational cost. As such, there is a need to develop systems that accurately estimate DOA using as few sensors as possible.

We propose that a single structural sensor affixed to a panel surface may be employed to reliably estimate DOA. When an acoustic wave induces vibrations on a panel, the contribution of each of the panel's bending modes to the total vibration response is dependent on the incident angle of the wave [7, 8, 9]. This vibration response may be recorded by a structural vibration sensor, and deep neural networks (DNNs) may be trained to distinguish the subtle variations in the relative modal amplitudes to infer DOA. Since the panel's resonant modes produce many peaks and dips in the vibration response across the audible frequency bandwidth, a previous experiment utilized mel-frequency cepstral coefficients (MFCCs) to reduce the recorded signal to a spectral feature vector where the relative modal excitations could be inferred [10]. MFCC feature vectors typically use 40 overlapping mel bands to provide a detailed representation of the auditory spectrum. The feature-space for estimating DOA proposed in this work is made significantly more compact by summing the energy in bands that align with the panel's resonant modes, and rejecting bands that may not contain significant spatial information and may cause over-fitting and other training errors.

Although the panel's resonances that are leveraged in this work to estimate DOA will inevitably introduce reverberation into the recorded signal, intelligibility measurements have shown that automatic speech recognition systems are still able to transcribe speech recorded by structural sensors without a significant accuracy reduction when compared to conventional microphones [11]. Combining these methods with the emergence of panel-based audio reproduction systems [12] gives the potential to create new, multimodal interfaces for

*This work supported by NSF Award 2104758

smart devices that meet the form-factor requirements of thin, lightweight displays.

2. RESONANCES OF PANEL VIBRATIONS

2.1. Vibration of a Baffled Panel

For a damped isotropic panel with Young's Modulus E, Poisson's ratio ν, density ρ, and thickness h, the out-of-plane displacement w when the panel is excited by external load $p(x, y, t)$ may be expressed as,

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

where b is the panel's mechanical loss factor. Solutions for (1) are found extensively in the literature, such as by Cremer et al. [13]. The displacement w is a separable function of space and time that may be written as,

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

The resonant modes of the surface can be used to fully describe the spatial component $\varphi(x, y)$ as,

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

where $\Phi_r(x, y)$ is the spatial function of the r^{th} mode and α_r is the mode's amplitude. For rectangular panels with clamped boundary conditions, $\Phi_r(x, y)$ contains separable sinusoidal functions along the panels length L_x and width L_y, and modal indices r_m and r_n may be used to represent the number of half-wavelengths in the horizontal and vertical dimensions, respectively. The resonant frequency ω_r of each mode under these conditions has been approximated by Mitchell and Hazel [14] as,

$$\omega_r = \pi^2 \sqrt{\frac{D}{\rho h}} \left[\left(\frac{m_r + \Delta m_r}{L_x} \right)^2 + \left(\frac{n_r + \Delta n_r}{L_y} \right)^2 \right], \tag{4}$$

where Δm_r and Δn_r are edge effect factors used to compensate for clamped boundary conditions. The bandwidth of each mode is subject to quality factor Q_r, given by,

$$Q_r = \frac{\omega_r \rho h}{b}. \tag{5}$$

2.2. Incident Pressure Wave Excitation

For a plane wave p_i incident on a baffled panel at angle θ_i between the in-plane projection of the propagation vector and the horizontal axis as shown in Fig. 1, the pressure $P(x, y)$ on the panel's surface is given by,

$$P(x, y) = 2P_i e^{-jk\cos\theta_i x - jk\sin\theta_i y}, \tag{6}$$

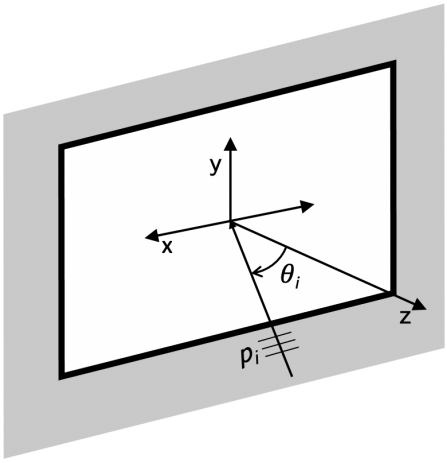

Fig. 1. A pressure wave p_i incident with angle θ_i in the horizontal plane on a baffled panel surface, from [7].

where P_i is the amplitude of the wave at frequency ω and k is the wave number. The panel mode amplitudes excited by (6) are given following [7, 8, 9] as,

$$\alpha_r = \frac{p_r(\theta_i, \omega)}{\rho h(\omega_r^2 - \omega^2 + j\omega_r\omega/Q_r)}, \tag{7}$$

with $p_r(\theta_i, \omega)$ given by,

$$p_r(\theta_i, \omega) = 8P_i I_{r_m}(\theta_i, \omega) I_{r_n}(\theta_i, \omega). \tag{8}$$

$I_{r_m}(\theta_i, \omega)$ and $I_{r_n}(\theta_i, \omega)$ are coupling factors between the pressure distribution on the panel due to the incident wave and the spatial response of each mode and are given by,

$$I_{r_m}(\theta_i, \omega) = \frac{m\pi\left[1 - (-1)^m e^{-j\sin\theta_i(\omega L_x/c)}\right]}{m^2\pi^2 - [\sin\theta_i(\omega^2 L_x^2/c^2)]}, \tag{9a}$$

$$I_{r_n}(\theta_i, \omega) = \frac{n\pi\left[1 - (-1)^n e^{-j\sin\theta_i(\omega L_y/c)}\right]}{n^2\pi^2 - [\sin\theta_i(\omega^2 L_y^2/c^2)]}, \tag{9b}$$

where $c = \frac{\omega}{k}$ is the incident wave's prorogation speed. Substituting (7) into (3) allows the frequency response of a panel excited by an incident plane wave to be fully described when the wave's incident angle is known.

2.3. Signal Recorded by Structural Sensor

Consider an acoustic source radiating signal $s(t)$ toward a panel that has a sensor affixed to its surface at position (x_0, y_0). If the transfer function from source to sensor is $h_{\theta_i}(t)$, the recorded velocity response becomes,

$$\dot{w}(x_0, y_0, t) = s(t) \circledast h_{\theta_i}(t), \tag{10}$$

where $h_{\theta_i}(t)$ varies by incident angle as described in Section 2.2. For $s(t)$ containing broadband white noise, $h_{\theta_i}(t)$ can be inferred directly from the recorded signal.

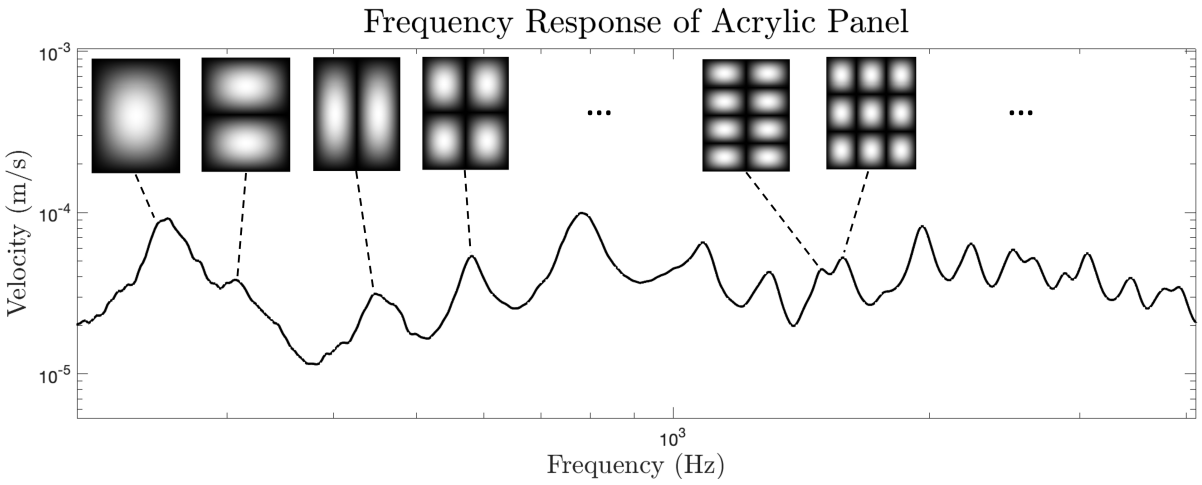

Fig. 2. Velocity response of the panel showing the isolated modal regions that would be excited by incident acoustic waves.

3. METHODOLOGY

3.1. Experimental Setup

A 2 mm thick acrylic panel with $E = 3.2$ GPa, $\nu = 0.35$, $\rho = 1,180$ kg/m^3, and $(L_x, L_y) = (18 \text{ cm}, 23 \text{ cm})$ was constructed. The panel's spatially-averaged velocity response with corner excitation was measured using a Polytec PSV-500 scanning laser vibrometer and plotted in Fig. 2 to show the modes of the panel that would be excited by incoming acoustic waves.

The panel was mounted in a semi-anechoic space to a rotary table capable of rotating between $\theta_i = -90°$ and $90°$ in $5°$ increments relative to a KEF LS50 loudspeaker placed on-axis at a distance of one-half meter [10]. At each angle of incidence, the panel was excited by 1,800 broadband noise bursts from the loudspeaker, each with a duration of 100 ms, and the panel's response was recorded by a single PCB Piezotronics U352C66 accelerometer arbitrarily positioned off-center in each dimension.

3.2. Feature Extraction

The vibrometer scan shown in Fig. 2 was used to determine the center frequencies and bandwidths of the isolated modal bands in the panel's vibration response. Note that some bands may contain degenerate modes, such as the band containing the $(2, 4)$ and $(3, 3)$ modes. At sufficiently high frequencies, so many modes are excited simultaneously that individual modes can no longer be observed in the panel's response [15, 16]. For this panel, this effect occurs at approximately 4 kHz, so the response above this frequency may be ignored. The 13 isolated modal bands below this threshold were used to make a band-pass filter bank G_l with center frequencies f_c and bandwidths Δf shown in Table 1. The modes contained in isolated modal bands with significant degeneracy are labeled "unclear" in the table. The energy contained in the l^{th} band, $E(l)$, can be computed by,

$$E(l) = \int_\omega G_l |S(j\omega) H_{\theta_i}(j\omega)|^2 d\omega, \tag{11}$$

where $S(j\omega)$ and $H_{\theta_i}(j\omega)$ are the Fourier transforms of $s(t)$ and $h_{\theta_i}(t)$ respectively. The proposed feature vector is an array containing $E(l)$ values for the recorded panel vibrations. The filter bank may be abbreviated to contain only the bands with modes whose excitation varies strongly with θ_i, as these modes are hypothesized to be the most useful for determining DOA. Algorithm 1 shows how (7) is used to rank modes by variance in θ_i.

The DNNs used in this work are LSTM-based recurrent neural networks modeled after networks that have shown promise in classifying colors of broadband noise from spectral features, modified in this case to estimate DOA using these energy-sum feature vectors with 13 or fewer values of $E(l)$ [17]. A set of 37,000 broadband noise bursts across all considered θ_i were used to excite the panel, and the recorded vibration responses were split into training and validation sets with a ratio of 80:20. An additional 29,600 responses were recorded as a testing set. The DNNs were trained with

Algorithm 1 Pseudo-code for selecting modes with the greatest excitation variance with respect to θ_i

```
for all isolated modes r do
    for all angles θ_i in dataset do
        Compute α_r at θ_i using (7)
    end for
    Normalize resulting α_r values
    Compute variance of normalized α_r values
end for
Rank modes r by greatest variance when varying θ_i
```

f_c (Hz)	Δf (Hz)	Mode(s)	Rank
256.5	60.0	(1,1)	5
454.9	115	(1,2)	T6
582.7	88.6	(2,1)	2
784.2	146	(2,2)	T6
1168	312	(2,3)	1
1287	210	(3,2)	T6
1564	340	(2,4),(3,3)	3
1962	224	Unclear	N/A
2233	317	(4,3)	4
2554	666	Unclear	N/A
3028	707	Unclear	N/A
3430	506	Unclear	N/A
3832	900	Unclear	N/A

Table 1. Resonant properties of the panel's modal bands that are ranked by variance in excitation with respect to θ_i.

	Reliability of DOA Estimates to within:					
# Bands	$\pm5^\circ$	$\pm10^\circ$	$\pm20^\circ$	$\pm5^\circ$	$\pm10^\circ$	$\pm20^\circ$
13	0.984	1	1			
12	0.978	1	1	*0.981*	*1*	*1*
11	0.962	0.998	1	*0.977*	*0.999*	*1*
10	0.966	0.998	1	*0.966*	*0.999*	*1*
9	0.945	0.993	1	*0.917*	*0.991*	*0.995*
8	0.921	0.993	0.998	*0.880*	*0.984*	*0.994*
7	0.897	0.983	0.998	*0.836*	*0.958*	*0.974*

Table 2. Reliability of the DOA estimates made by DNNs trained with subsets of the resonance-informed filter bank. Bands are removed in the leftmost columns by least excitation variance when varying θ_i, and removed by most variance in the *italicized rightmost columns*.

a loss function that minimizes the root-mean-square error between the known incident angle and the estimate returned by the model using regression. When acting on the testing set, the reliability of each DNN was determined by its ability to estimate DOA within an angular tolerance of $\pm\Delta\theta$ as the ratio of the number of correct predictions within $\pm\Delta\theta_i$ to the total number of bursts in the set whose incident angle was θ_i [18, 19]. Angular tolerances $\Delta\theta$ of 5°, 10°, and 20° were used [20].

4. RESULTS AND DISCUSSION

The DNN trained with each of the 13 isolated modal bands estimated DOA to within $\pm5^\circ$ with a reliability of 98.4% as shown in Table 2. Previously, when MFCC features were utilized, a reliability of 99.8% with an angular tolerance of $\pm5^\circ$ was reported [10]. Therefore, only a 1.4% reduction in reliability is observed when a feature vector constructed from 13 resonance-informed filters is utilized in lieu of a more spectrally-complete feature set utilizing 40 mel filters. In both cases, the bursts in the testing set were always estimated correctly within $\pm10^\circ$, an angular tolerance that may be sufficient for many signal enhancement algorithms. It is worth noting that the reliability values are representative of an experiment that occurred in a well-controlled and semi-anechoic environment, as the scope of this paper is to suggest the possibility of making DOA estimates from recorded vibration signals. A more robust DNN and training procedure accounting for realistic and noisy environments is left to future work.

The abbreviated versions of the resonance-informed filter bank can be used without significant reduction in reliability, particularly when removing bands that contain modes with the smallest excitation variance with respect to θ_i. A DNN trained using as few as 7 modal bands was able to estimate DOA to within $\pm5^\circ$ with a reliability of 89.7%. When removing bands that contain modes with the largest excitation variance, the reliability fell off more quickly, which supports the notion that modes whose amplitudes vary significantly with the incident angle are more effective for DOA estimation. Since several of the reported isolated modal bands contained significant degenerate modes, only 8 of the 13 bands could be ranked directly with Algorithm 1. In future work, Algorithm 1 may rank the variance of $E(l)$ directly using empirical data. Additionally, the sensor couples better to certain modes depending on its location relative to the mode's nodal lines. An optimal sensor location based on the panel's resonances should be determined to ensure that the sensor has strong coupling to all the modes within the isolated bands.

Estimating DOA from harmonic, band-limited speech signals is a required feature in smart audio devices. MFCC vectors in prior work showed promise for estimating DOA from speech-based signals recorded by structural sensors [10]. While determining DOA from the fricative sounds of human speech that more closely resemble noise bursts is a direct extension of this work, measuring the effectiveness of the resonance-informed energy-sum feature vectors proposed in this work for estimating DOA from harmonic speech sounds will be a point of emphasis in future work.

5. CONCLUSIONS

The results in this work suggest that compact feature vectors informed by the resonant properties of a panel surface are sufficient for reliable DOA estimation using a single structural sensor. DNNs trained utilizing the energy contained in isolated modal bands of a panel's vibration response were able to estimate the DOA of broadband noise signals within $\pm5^\circ$ with a reliability of up to 98.4%. The method presented is a more efficient approach to DOA estimation utilizing surface vibrations than the previous work, and is an important step in the design of panel-based smart audio devices.

6. REFERENCES

[1] Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L. Seltzer, Heiga Zen, and Mehrez Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[2] B.D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[3] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 1, pp. 375–378 vol.1.

[4] Byoungho Kwon, Youngjin Park, and Youn-sik Park, "Analysis of the GCC-PHAT technique for multiple sources," in *ICCAS 2010*, 2010, pp. 2070–2073.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[6] M. Kaveh and A. Barabell, "The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 331–341, 1986.

[7] S. Elliot C. Fuller and P. Nelson, *Active Control of Vibration*, Academic Press, 1996.

[8] Bor-Tsuen Wang, Chris R. Fuller, and Emilios K. Dimitriadis, "Active control of noise transmission through rectangular plates using multiple piezoelectric or point force actuators," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2820–2830, 1991.

[9] Louis A. Roussos, "Noise transmission loss of a rectangular plate in an infinite baffle," *NASA Technical Paper*, , no. 2398, 1985.

[10] Tre DiPassio, Michael C. Heilemann, and Mark F. Bocko, "Direction of arrival estimation of an acoustic wave using a single structural vibration sensor," *Journal of Sound and Vibration*, vol. TBD, no. TBD, pp. TBD, In Review 2022.

[11] Tre DiPassio, Michael C. Heilemann, and Mark F. Bocko, "Audio capture using structural sensors on vibrating panel surfaces," *Journal of the Audio Engineering Society*, vol. TBD, no. TBD, pp. TBD, In Press 2022.

[12] Michael C. Heilemann, David A. Anderson, Stephen Roessner, and Mark F. Bocko, "The evolution and design of flat-panel loudspeakers for audio reproduction," *J. Audio Eng. Soc*, vol. 69, no. 1/2, pp. 27–39, Jan 2021.

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

[14] A.K. Mitchell and C.R. Hazell, "A simple frequency formula for clamped rectangular plates," *Journal of Sound and Vibration*, vol. 118, no. 2, pp. 271 – 281, 1987.

[15] G. Rabbiolo, R. J. Bernhard, and F. A. Milner, "Definition of a high-frequency threshold for plates and acoustical spaces," *Journal of Sound and Vibration*, vol. 277, no. 4-5, pp. 647–667, Nov. 2004.

[16] David A Anderson, Michael C Heilemann, and Mark F Bocko, "Measures of vibrational localization on point-driven flat-panel loudspeakers," in *Proceedings of Meetings on Acoustics 171ASA*. Acoustical Society of America, 2016, vol. 26, p. 065003.

[17] "Classify sound using deep learning," *MathWorks*, https://www.mathworks.com/help/audio/gs/classify-sound-using-deep-learning.html.

[18] Nian Liu, Huawei Chen, Kunkun Songgong, and Yanwen Li, "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays," *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 1069–1084, 2021.

[19] Q. Li, X. Zhang, and H. Li, "Online direction of arrival estimation based on deep learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2616–2620.

[20] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.