Journal of Sound and Vibration 553 (2023) 117671

Contents lists available at ScienceDirect

Journal of Sound and Vibration

journal homepage: www.elsevier.com/locate/jsvi

Direction of arrival estimation of an acoustic wave using a single structural vibration sensor

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

Department of Electrical and Computer Engineering, University of Rochester, United States

ARTICLE INFO

Keywords:
Surface audio
Direction of arrival
Vibro-acoustics
Source localization

ABSTRACT

The modal vibrational response of a flexible panel to an incident acoustic pressure wave is dependent on the direction of arrival (DOA) of the wave. This work presents a proof of concept whereby vibration measurements of flat panels made by structural sensors were used to infer the DOA of an incident acoustic wave. In the experiments reported here, rectangular panels with various material damping factors were excited in an anechoic space by acoustic waves incident at angles between −90° and +90° in the horizontal plane. Mel-frequency cepstral coefficients (MFCCs) were computed from the recorded panel responses and used to train a deep neural network (DNN) to estimate the DOA. Experimental results show that under controlled conditions, the DOA of incident acoustic waves containing broadband noise may be estimated to within ±5° with a reliability of 99.8% utilizing recordings from a single structural sensor. For acoustic waves containing less spectrally uniform human speech signals, a DNN trained using data from a single sensor correctly estimated DOA to within ±5° with a reliability of 86.5% and to within ±10° with a reliability of 96% within the experimental conditions. The scope of this work is to suggest that a panel's modal response may contain sufficient information to enable DOA estimation with as few as one sensor. As such, a proof of concept is presented in place of a final engineering solution, along with a discussion of experimental limitations and future improvements.

1. Introduction

Smart speakers utilize multiple acoustic signal processing techniques such as source localization and tracking [1–4], source separation [5], and acoustic beamforming [6,7] to allow the device to better understand and execute a users' commands. These techniques require an estimation of the user's position relative to the device. One way this position may be estimated is by measuring the direction-of-arrival (DOA) of the acoustic pressure waves produced by the user's speech. Traditionally, devices such as smart speakers employ multi-microphone arrays to estimate the DOA of speech signals [8] using methods such as inter-sensor time difference of arrival (TDOA), generalized cross-correlation with phase transform (GCC-PHAT) and the multiple signal classification (MUSIC) algorithm [9–12]. These techniques require an array of microphones to measure changes in acoustic pressure at multiple points in space. In general, the performance of these algorithms is improved by increasing the number of microphones in the array at the expense of increased computational requirements and higher manufacturing cost. Techniques for reducing the number of sensors required to perform DOA estimation are of interest to researchers and manufacturers working on low-cost smart audio devices.

* Corresponding author.
E-mail addresses: tredipassio@rochester.edu (T. DiPassio), mheilema@ur.rochester.edu (M.C. Heilemann), mark.bocko@rochester.edu (M.F. Bocko).

https://doi.org/10.1016/j.jsv.2023.117671
Received 17 August 2022; Received in revised form 16 January 2023; Accepted 11 March 2023
Available online 13 March 2023

Flat panel loudspeakers have been demonstrated to offer a promising alternative to conventional moving-coil loudspeaker systems in thin electronic devices, where weight and form-factor are key design constraints [13]. In these loudspeaker systems, sound-radiating bending waves are induced on the panel by one or more force actuators [14–16]. A number of methods have been developed to improve the quality of the radiated sound, and blind listening tests have shown that flat-panel loudspeakers are rated comparably to traditional loudspeakers [17]. The bending motion of a panel is also sensitive to changes in acoustic pressure, and the vibrations induced in a panel from acoustic pressure waves can be recorded using inexpensive structural vibration sensors affixed to the panel. Although the resonant properties of the panel inevitably introduce reverberant artifacts to the recorded speech signals, the audio quality is sufficient for automatic speech recognition (ASR) systems to transcribe the recorded speech without a significant loss in accuracy in comparison to recordings made with conventional microphones [18].

For devices performing audio capture in noisy environments, signal enhancement is commonly performed before the audio is processed by an ASR system. Little work has been done on using recorded surface vibrations in signal processing algorithms, particularly those such as DOA estimation, beamforming, and signal enhancement. However, recent work by Kita and Kajikawa demonstrates the potential of recording surface vibrations in the audio band to enable signal processing tasks by training deep neural networks (DNNs) to localize acoustic sources internal to a structure utilizing surface vibrations on the outside of the structure [19]. Additionally, sound source localization was achieved by Badawy and Dokmanić using a single microphone placed in an arbitrary scattering structure [20], though in this case the structure was a complex three-dimensional assembly.

For panels excited by an acoustic wave, the response of each bending mode is dependent on the incident angle and frequency content of the acoustic wave [21–23]. The relative amplitudes of each mode may be measured directly by structural sensors affixed to the surface of the panel. Since the vibration response is angle-dependent for a given source signal, DNNs may be employed to estimate the DOA of an incident acoustic pressure wave when trained with frequency response data from one or more vibration sensors. As a proof of concept, we demonstrate that the resonant properties of a panel excited by acoustic waves containing broadband noise, and individual speech sounds enable an estimation of the excitation signals' DOA using audio recorded by a single structural sensor. As the scope of this work is to present a preliminary study, the panels will be excited under anechoic conditions using recordings of individual phonetic utterances made by a single male speaker. We begin with a brief overview on the mechanics of acoustically excited panels.

2. Mechanical system modeling

2.1. Mechanics of a baffled panel

Consider a damped, isotropic panel with Young's Modulus E, Poisson's ratio ν, density ρ, and thickness h. The out-of-plane displacement w in response to an external load $p(x, y, t)$ at time t and point (x, y) on its surface may be found by solving the following equation (see for example Cremer et al. [24]):

$$D\nabla^4 w(x,y,t) + \rho h\ddot{w}(x,y,t) + b\dot{w}(x,y,t) = p(x,y,t), \tag{1}$$

where b is the panel's mechanical loss factor and the bending stiffness D is given by,

$$D = \frac{Eh^3}{12(1-\nu^2)}. \tag{2}$$

The displacement $w(x, y, t)$ may be separated into functions of space and time as,

$$w(x,y,t) = \varphi(x,y)e^{j\omega t}. \tag{3}$$

The spatial response $\varphi(x, y)$ may be expressed as a superposition of modes $\boldsymbol{\Phi}_r(x, y)$ with amplitudes α_r as,

$$\varphi(x,y) = \sum_{r=1}^{\infty} \alpha_r \boldsymbol{\Phi}_r(x,y). \tag{4}$$

For a panel with simply supported boundaries, the mode shapes are sinusoidal functions of space, and each mode has resonant frequency ω_r. Substituting Eq. (3) and Eq. (4) into Eq. (1) gives,

$$\alpha_r = \frac{4}{\rho h L_x L_y(\omega^2 - \omega_r^2 - \frac{j\omega_r\omega}{Q_r})} \int_0^{L_x}\int_0^{L_y} P(x,y)\boldsymbol{\Phi}_r(x,y)\mathrm{dydx}, \tag{5}$$

where $P(x, y)$ is the pressure magnitude at each point on the panel and the quality factor of each mode Q_r is given by,

$$Q_r = \frac{\omega_r \rho h}{b}. \tag{6}$$

The material properties of the panels that were constructed for this experiment are summarized in Table 1. The panels are, respectively, made of acrylic, Gatorboard, and two aluminum (Al.) sheets sandwiching a layer of viscoelastic damping material. Note that since the panels used in this experiment have boundary conditions that approximate clamped edges, the resonant frequencies were computed following the approximation given by Mitchell and Hazel [25]. The mode indices r_m and r_n represent the number of half-wavelengths in the horizontal and vertical dimensions respectively. In the next section, we derive the response of a rectangular panel excited by incident pressure waves for varying angles of incidence.

Table 1
Properties for the panels used in this experiment. The reported average value for b can be used with Eq. (6) to demonstrate that lesser-damped panels made of acrylic and Gatorboard experience higher-Q reverberant modes than a highly-damped panel made of the Al. sandwich material. Values of r for reporting f_r were chosen to yield the six lowest frequency modes of the Gatorboard panel.

Panel Material and Resonant Properties			
	Al. Sandwich Panel	Acrylic Panel	Gatorboard Panel
E (GPa)	68.9	3.2	1.5
υ	0.334	0.35	0.35
ρ (kg m^{-3})	2700	1180	222
b (kg s^{-1})	10270	2172	241.5
h (mm)	1.0	2.0	3.0
L_x (m)	0.26	0.26	0.26
L_y (m)	0.36	0.36	0.36
f_r (Hz) for $(r_m, r_n) = (1,1)$	115	65	154
f_r (Hz) for $(r_m, r_n) = (2,1)$	190	108	254
f_r (Hz) for $(r_m, r_n) = (1,2)$	280	159	375
f_r (Hz) for $(r_m, r_n) = (3,1)$	314	178	421
f_r (Hz) for $(r_m, r_n) = (2,2)$	350	198	468
f_r (Hz) for $(r_m, r_n) = (3,2)$	467	265	626

2.2. Response of a panel excited by an obliquely-incident pressure wave

Consider a plane wave p_i incident on a panel as shown in Fig. 1 such that the pressure $P(x, y)$ along the panel's surface is given by,

$$P(x,y) = 2P_i e^{-\mathrm{j}k\sin\phi_i\cos\theta_i x - \mathrm{j}k\sin\phi_i\sin\theta_i y}, \tag{7}$$

where P_i is the amplitude of the incident wave at frequency ω, k is the wave number, ϕ_i is the angle between the incident wave's propagation vector and the axis normal to the panel, and θ_i is the angle between the in-plane projection of the propagation vector and the horizontal axis. Following [21–23], the mode amplitudes given in Eq. (5) are a function of the incident angles of the plane wave θ_i and ϕ_i given by,

$$\alpha_r = \frac{p_r}{\rho h(\omega_r^2 - \omega^2 + \mathrm{j}\omega_r\omega/Q_r)}, \tag{8}$$

where,

$$p_r = 8P_i I_{r_m}(\theta_i,\phi_i,\omega) I_{r_n}(\theta_i,\phi_i,\omega). \tag{9}$$

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

$$I_{r_m}(\theta_i,\phi_i,\omega) = \begin{cases} \dfrac{m\pi\left[1-(-1)^m e^{-\mathrm{j}\sin\theta_i\cos\phi_i(\omega L_x/c)}\right]}{m^2\pi^2 - [\sin\theta_i\cos\phi_i(\omega^2 L_x^2/c^2)]}, & \text{if } (m^2\pi^2) \neq [\sin\theta_i\cos\phi_i(\omega^2 L_x^2/c^2)] \\ \dfrac{-\mathrm{j}}{2}\mathrm{sgn}(\sin\theta_i\cos\phi_i), & \text{if } (m^2\pi^2) = [\sin\theta_i\cos\phi_i(\omega^2 L_x^2/c^2)] \end{cases}, \tag{10a}$$

$$I_{r_n}(\theta_i,\phi_i,\omega) = \begin{cases} \dfrac{m\pi\left[1-(-1)^n e^{-\mathrm{j}\sin\theta_i\sin\phi_i(\omega L_y/c)}\right]}{n^2\pi^2 - [\sin\theta_i\sin\phi_i(\omega^2 L_y^2/c^2)]}, & \text{if } (n^2\pi^2) \neq [\sin\theta_i\sin\phi_i(\omega^2 L_y^2/c^2)] \\ \dfrac{-\mathrm{j}}{2}\mathrm{sgn}(\sin\theta_i\sin\phi_i), & \text{if } (m^2\pi^2) = [\sin\theta_i\sin\phi_i(\omega^2 L_y^2/c^2)] \end{cases}, \tag{10b}$$

where $c = \frac{\omega}{k}$ is the prorogation speed of the incident pressure wave. Eqs. (3), (4), and (8) may be combined to predict the total vibration response of a panel due to an obliquely incident plane wave by a sensor located at position (x_0, y_0) on the panel surface.

2.3. Harmonic features of induced vibrations from broadband sources

The velocity response of a panel at sensor location (x_0, y_0) due to an acoustic source incident at angles ϕ_i and θ_i may be modeled as the convolution of the source signal $s(t)$ and the effective transfer function $h_{\phi_i,\theta_i}(t)$ from the source to the sensor location as,

$$\dot{w}(x_0,y_0,t) = s(t) \circledast h_{\theta_i,\phi_i}(t). \tag{11}$$

Therefore, the signals recorded by the structural sensor contain harmonic properties that vary depending on the angle of incidence of the acoustic wave. Incident angles ϕ_i and θ_i may be estimated from the recorded signal $\dot{w}(x_0, y_0, t)$ if information about $s(t)$ is known. In this work, panels were excited by acoustic waves containing broadband noise bursts and speech at various incident angles in the horizontal (azimuthal) plane ($\phi_i = 0°$). For $s(t)$ containing white noise, $h_{\phi_i,\theta_i}(t)$ can be obtained directly from $\dot{w}(x_0, y_0, t)$. For $s(t)$ containing speech, the harmonics contained in human speech become a component of the excitation signal, which complicates

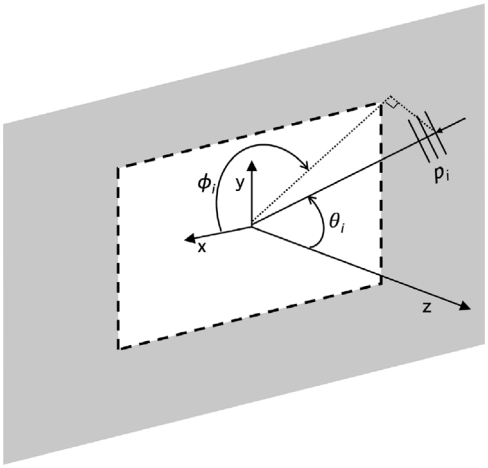


Fig. 1. A pressure wave p_i incident to a baffled panel surface at angles ϕ_i and θ_i as shown in [21].

the solution of the inverse problem. To address this, θ_i was estimated only after extracting information about the speech sound or word contained in the incident waveform.

For acoustic waves containing broadband white noise, the incident angles may be inferred by comparing the response recorded by the sensor to a bank of previously measured sensor responses where the angles of incidence were known. A computationally efficient way of making this comparison is to calculate and compare the energy contained in a number of frequency bands in the recorded response to those in the bank of previously measured responses. Consider a filter bank G_l containing l band-pass filters. The energy $E(l)$ contained in each band after applying the lth filter to the signal recorded by the sensor is given by,

$$E(l) = \int_{\omega} G_l |S(\mathrm{j}\omega) H_{\phi_i,\theta_i}(\mathrm{j}\omega)|^2 \mathrm{d}\omega, \tag{12}$$

where $S(\mathrm{j}\omega)$ and $H_{\phi_i,\theta_i}(\mathrm{j}\omega)$ are the Fourier transforms of $s(t)$ and $h_{\theta_i,\phi_i}(t)$ respectively.

A filter bank of 14 triangular filters whose center frequencies f_c are linearly spaced between 230 Hz and 1261 Hz is shown in Fig. 2(a). Consider the Gatorboard panel used in this work, whose properties are given in Table 1, excited by acoustic waves containing broadband white noise. For simplicity, we simulate three waves incident at $(\phi_i = 0°, \theta_i = -30°)$, $(\phi_i = 0°, \theta_i = 0°)$, and $(\phi_i = 0°, \theta_i = 45°)$ for panel structural sensor locations at $(0.75L_x, 0.75L_y)$. Following Eq. (12) the summed energies in the frequency bands shown in Fig. 2(a) are given in Fig. 2(b) for the panel's responses at each angle of incidence. The results of this simulation demonstrate that the relative energies contained in each frequency band depend on the incident angle of the acoustic wave. A DNN may use this vector to create decision boundaries in 14-dimensional space to estimate the DOA of a recorded response.

Note that the resonant modes of the panel contribute to the relative energies in each band. For instance, the center frequency of Band 4 is close to the resonant frequency of the (2,2) mode of the Gatorboard panel, while the center frequency of Band 6 is close to the resonant frequency of the (3,2) mode. Since these two modes have different spatial responses, the acoustic wave will couple differently to each mode depending on the angle of incidence as shown in Eq. (9). An optimal filter bank may be designed such that the center frequencies of the filters align with the resonant frequencies of panel modes that give the most spatial information about the source location. Though the optimization of the filter bank is outside the scope of this paper, the idea warrants consideration in future work. Section 3 discusses the use of a mel filter bank to extract the energy contained in bands of the panel's response to acoustic waves containing speech.

3. Phonetic considerations

3.1. Modeling excitation signals containing speech

The source-filter model is a model for speech production detailed by Chiba and Kajiyama [26] and formalized by Fant [27]. The source $x_s(t)$ and filter $h_s(t)$ are modeled as independent linear systems to express the speech signal $s(t)$ as,

$$s(t) = x_s(t) \circledast h_s(t). \tag{13}$$

We consider two basic types of sources for this source-filter model: the so-called "voiced source" created by the periodic vibration of human vocal folds, and a "noise source" created by turbulent airflow through the vocal system. The combination of these sources can lead to voiced sounds that contain pitch, aspirated sounds (when the noise stems from the glottis), fricative sounds (when the noise stems from elsewhere), and transient bursts.

Formant frequencies are the resonant frequencies of the speaker's vocal tract, and appear as narrow high-frequency bands containing large amounts of energy — particularly when observing spectrograms made from recordings of voiced utterances such as

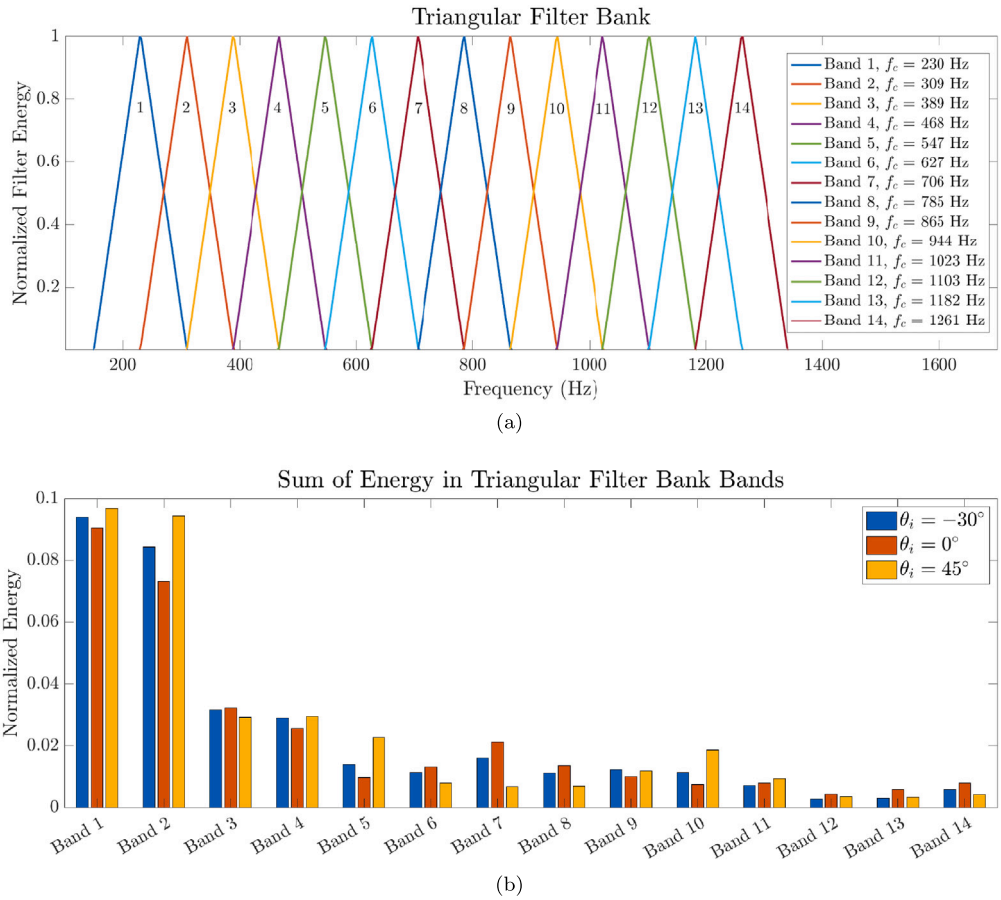


Fig. 2. (a) Triangular filter bank used to extract the energy in linearly spaced bands of the Gatorboard panel's response to incident pressure waves. The center frequencies f_c of each band are reported in the legend. (b) Total energy in the panel's response to simulated plane waves containing broadband white noise incident at $(\phi_i = 0°, \theta_i = -30°)$, $(\phi_i = 0°, \theta_i = 0°)$, and $(\phi_i = 0°, \theta_i = 45°)$ after filtering with the triangular filter bank.

vowels. As the vocal tract changes shapes to pronounce different sounds, the formant frequencies change in a predictable manner. Therefore, the filter system of the source-filter model will vary depending on the type of speech sound produced. In general, Eqs. (11) and (13) may be combined to model the velocity response of a panel at sensor location (x_0, y_0) to acoustic waves containing speech as,

$$\dot{w}(x_0, y_0, t) = x_s(t) \circledast h_s(t) \circledast h_{\theta_i,\phi_i}(t). \tag{14}$$

For panels excited by non-voiced fricative sounds, $\dot{w}(x_0, y_0, t)$ is similar to the simulated results of a panel excited by broadband white noise from Section 2.3, though it should be noted that these utterances more closely resemble pink noise than they do white noise. For voiced utterances, the harmonics of the vocal tract couple with the resonances of the panel to create a $\dot{w}(x_0, y_0, t)$ unique to each speech sound. In this work, we propose that the direction of arrival of acoustic waves containing speech may be inferred by comparing the panel's response recorded by the sensor to a bank of previously measured sensor responses to the same phoneme where the angles of incidence are known.

As the precise frequencies of formants change from speaker to speaker, this proposed method operates as a speaker-dependent model. Generalization to a speaker-independent model is possible as the relative formant frequencies for different speech sounds vary less between speakers than potential variances in the speaker's fundamental frequency [28]. However, the creation of speaker-independent models is highly resource intensive and falls outside of the scope of the present work. A brief discussion regarding these challenges is included in Sections 5.2.1 and 5.2.2.

3.2. Harmonic features of induced vibrations from speech sources

The harmonics introduced by "the filter" within the source-filter model of speech production lie in bands that are not optimally sampled by filter banks with linearly-spaced center frequencies, such as the one used in Section 2.3. Therefore, speech processing algorithms commonly employ features derived from observing the energy in the bands of a mel filter bank, whose center frequencies

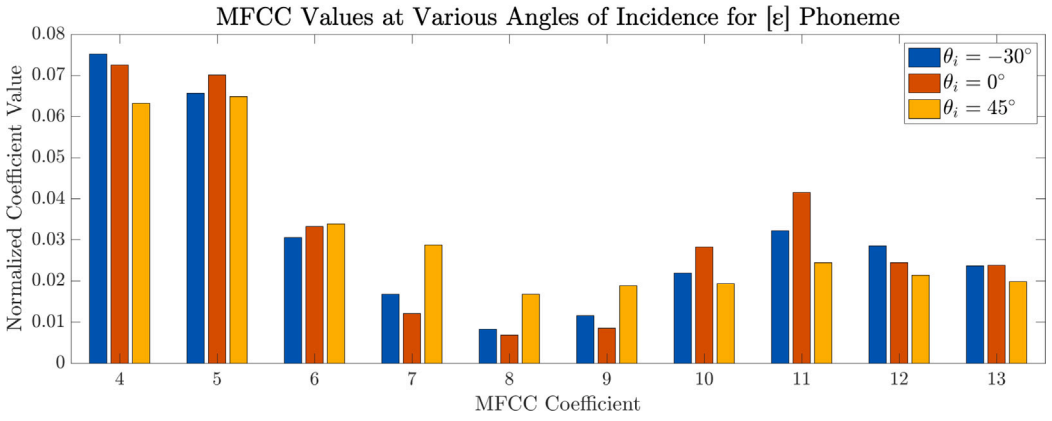


Fig. 3. Selected MFCC Coefficients extracted from a recording made of the Gatorboard panel's response to pressure waves containing the speech sound [ɛ] incident at $(\phi_i = 0°, \theta_i = -30°)$, $(\phi_i = 0°, \theta_i = 0°)$, and $(\phi_i = 0°, \theta_i = 45°)$. The first three coefficients are abbreviated as they are large enough that when all 13 coefficients are plotted on the same axes, the spatial information contained in later coefficients is not readily visible.

are spaced according to their prevalence in the human auditory system [29,30], with mel-frequency cepstral coefficients (MFCCs) commonly employed as a feature set [31].

Consider again the Gatorboard panel used in this work excited by acoustic waves incident at $(\phi_i = 0°, \theta_i = -30°)$, $(\phi_i = 0°, \theta_i = 0°)$, and $(\phi_i = 0°, \theta_i = 45°)$, this time containing the vowel pronounced 'eh' (or [ɛ] in the International Phonetic Alphabet). The transfer function $h_{\phi_i,\theta_i}(t)$ can again be simulated used to compute the expected response of the panel at sensor location $(0.75L_x, 0.75L_y)$. MFCCs are extracted from these simulated responses and plotted in Fig. 3. The results again demonstrate that the extracted MFCC values depend on the incident angle of the acoustic wave. A DNN may use the MFCC vector to create decision boundaries to estimate the DOA of a recorded response.

3.3. Practical implementation

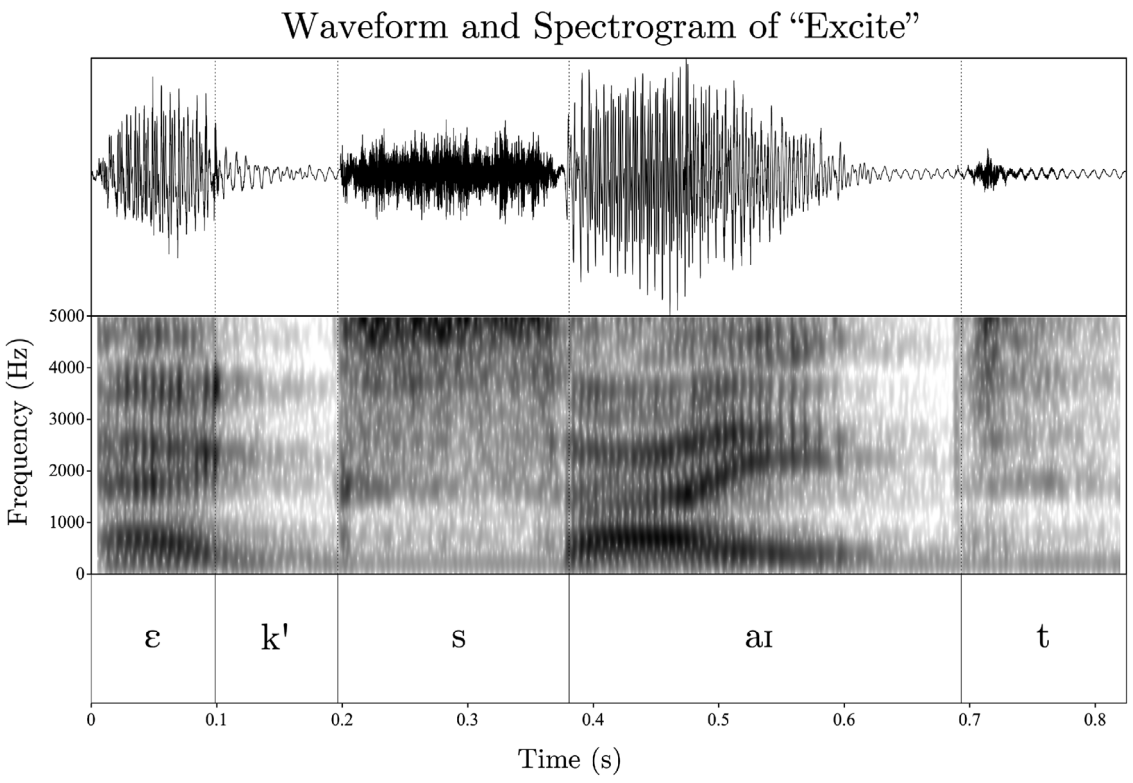
In practice, generating a set of panel responses to every speech sound in a given language at all possible angles of incidence would require significant resources and was not necessary for the scope of this work. Instead, we train DNNs to estimate the DOA for speech sounds contained within the word "excite" (or [ɛkˈ saɪt] in the International Phonetic Alphabet). This word contains a wide variety of speech sounds, including a vowel [ɛ], diphthong [aɪ], fricative [s], and plosive [t]. [kˈ] is a velar ejective stop and is therefore omitted for DOA estimation. The time-domain and spectrogram representation of a recording of the word "excite" separated into phonemes using PRAAT [32] is shown in Fig. 4(a).

While the reliability of the DOA estimates made by the DNNs for individual phonemes is reported in Section 5.2, a smart audio system rarely handles phonemes in isolation, and instead processes full words or phrases. For activation, many modern smart devices use wake-words. If we consider "excite" to be a wake-word, a phoneme segmentation method (such as the method presented by Arik et al. [33]) can be used to break a recording made by the structural sensors into frames containing the different phonemes. Separate DNNs trained to estimate DOA for speech sounds [ɛ], [s], [aɪ], and [t] can then be used jointly or independently. Once DOA is established using the wake-word, time-invariance can be assumed for the following phrase spoken to the audio system as the source and panel are assumed to be stationary [8]. However, continuous DOA estimation on the phonemes in the post-activation phrase for source tracking is a promising application for future work. An overall proposed DOA estimation system is shown in Fig. 4(b).

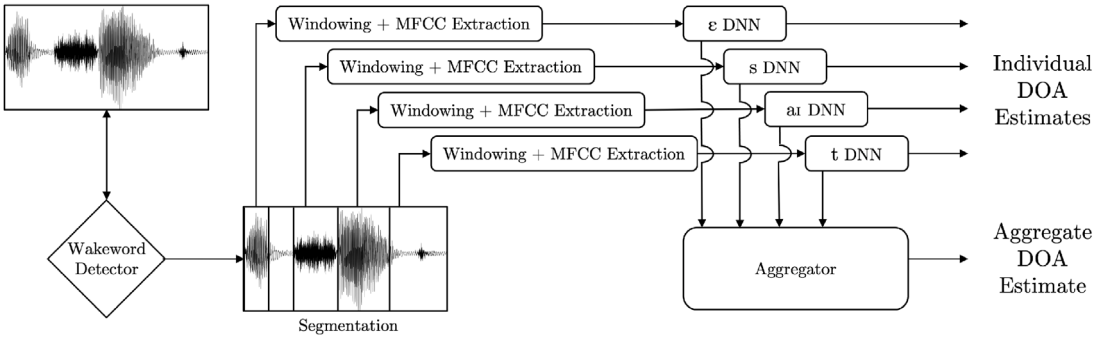
4. Experiment layout and procedure

4.1. Data acquisition

The experimental setup is shown in Fig. 5. Five PCB Piezotronics U352C66 accelerometers were attached to a panel mounted to a rotary table so that the panel could be rotated between $\theta_i = -90°$ and 90° in 5° increments relative to a KEF LS50 loudspeaker placed on-axis half a meter away (implying $\phi_i = 0°$ for all measurements since all incident waves lie in the azimuthal plane). The impulse response from the loudspeaker to the panel under test was recorded at each angle of incidence and sensor location using the maximum length sequence (MLS) method at 71 dBSPL to simulate a human speaking at this distance [34]. Using the recorded impulse responses, the response of the panel excited by speech incident at each angle was simulated by convolving the speech signal with each impulse response.



(a)



(b)

Fig. 4. (a) Waveform and spectrogram of the word “excite” with the individual phonemes labeled. Formants are visible in the spectrogram of the vowel and diphthong, as are the broadband characteristics of the fricative and plosive. (b) A proposed algorithm for estimating the DOA of incident speech sounds using one or an aggregate of DNNs trained to estimate DOA from individual phoneme sounds.

4.1.1. Linearity of panel vibrations

It is shown by Fahy and Gardonio [35] that non-linear vibrational behavior of flat panels can be produced only by significant transverse deflection. Plane waves incident on a panel’s surface cause displacements on the order of tens of microns, for which the curvature is a fraction of the panel’s minimum dimensions and is well within the panel’s linear vibrational region. To demonstrate this linearity experimentally, broadband noise bursts with a duration of 100 ms were reproduced through the loudspeaker at each rotational angle to excite a vibrational response on the panel’s surface, yielding roughly 10,000 recorded bursts per angle. In Section 5.1, the reliability of the DOA estimates for panels excited by broadband noise is reported for both recorded and convolved bursts. The similarity of the results suggests that, for large amounts of speech data, convolution with the panel’s impulse response may be used to efficiently simulate the panel’s response to incident speech without loss of accuracy. This convolution approach

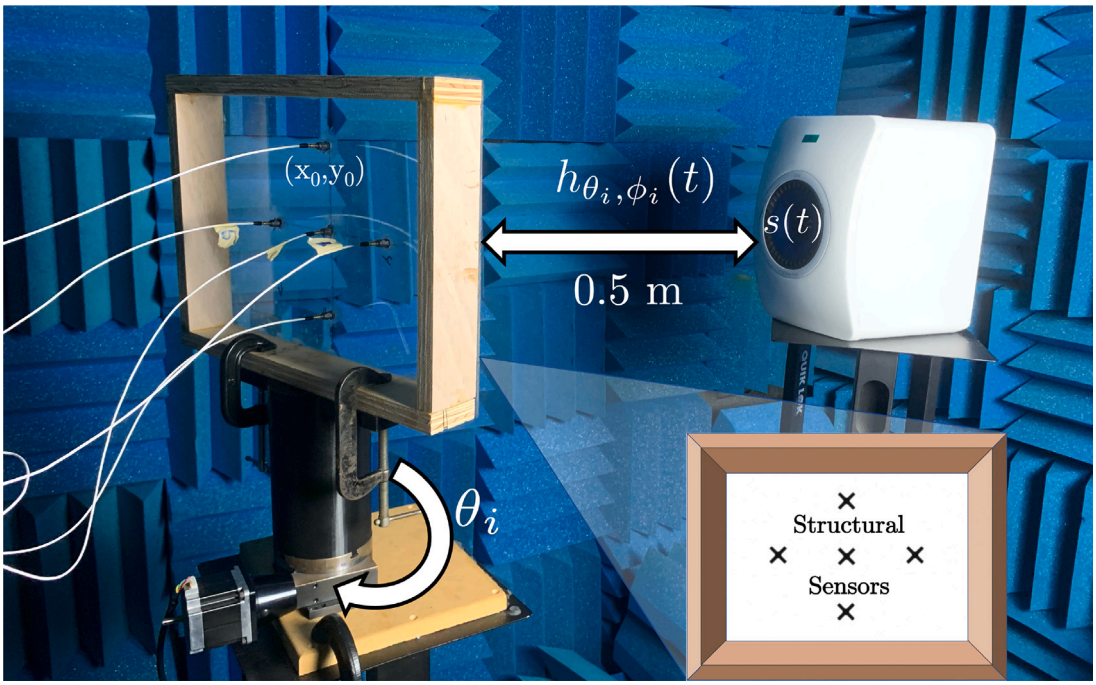


Fig. 5. Panels with a 0.26 m by 0.36 m active surface area are mounted to a turntable in a semi-anechoic setting. Five sensors were affixed to the panel, one centrally and one 20% along each of the panel's dimensions in L_x and L_y.

greatly reduced the laboratory resources needed to carry out this experiment and eliminates the potential for non-stationary noise pollution during unsupervised data collection.

4.2. Training and testing the deep neural networks

In this work, five distinct DNNs were trained: one for estimating DOA from incident broadband noise bursts and four for estimating the DOA from incident bursts containing each individual usable phoneme in the word "excite" (omitting the velar ejective stop). The training data for each DNN is noise or speech bursts convolved with the panel's impulse responses for each rotational angle. Each DNN is trained with 185,000 total bursts (1000 per angle over five sensors) that were split into training and validation sets with a ratio of 80:20. An additional 148,000 bursts (800 per angle over five sensors) were used to test the performance of each DNN. The broadband noise bursts are snippets of independently generated white noise. For the phoneme data, a single male speaker recorded each individual usable phoneme in isolation 1800 times such that the training, validation, and testing sets could be sufficiently populated. The speaker slightly varied volume and pronunciation while performing each phoneme for a degree of robustness on a speaker-dependent level.

MFCC vectors were extracted from each burst in the training and validation sets. During training and testing, each structural sensor was in either an 'on' or 'off' state, whereby the DNN model either utilized the data vectors from that sensor or ignored them entirely. We define N as the number of sensors that were 'on' while training a particular model. For each class of DNN, 31 total models were trained corresponding to the number of unique sensor combinations out of the 5 affixed sensors, ranging from models utilizing a single sensor ($N = 1$) to a model that utilizes data from all sensors ($N = 5$). Models were trained with the loss function L_{RMSE} that minimizes the root-mean-square error (RMSE) between the known incident angle θ_i and the incident angle θ_i^e estimated by the model, where L_{RMSE} is given by,

$$L_{\mathrm{RMSE}} = \sqrt{\frac{\sum_{i=1}^{n}\left(\theta_i - \theta_i^e\right)^2}{n}}. \tag{15}$$

The models were evaluated by their ability to make correct estimates of the incident angle within a defined angular tolerance. An estimate was deemed correct if it was within $\pm\Delta\theta$ of the known incident angle. Following [5,36], the reliability with which the model correctly estimates the DOA at each angle θ_i is given as the ratio of the number of correct predictions within $\pm\Delta\theta_i$ to the total number of bursts tested whose incident angle is known to be θ_i. The average reliability with which the DNNs estimated the incident angle of the bursts in the test sets are reported in Section 5 for a $\Delta\theta$ of 5°, 10°, and 20°, consistent with the resolutions reported in the literature [37].

Table 2
Tabulated are the validation RMSEs and the reliability of the DOA estimates made by the DNNs trained with broadband noise using both the convolution and *recording* approaches when acting on their respective testing sets, with the results from the recording approach italicized. The tabulated results are those given by the models with the best validation RMSE for each value of N.

Material	N	Validation RMSE (°)		Reliability of DOA Estimates to within:					
				±5°	±10°	±20°	±5°	±10°	±20°
Acrylic	1	0.909	*1.08*	0.998	0.999	0.999	*0.998*	*0.999*	*0.999*
	3	0.601	*0.483*	1.00	1.00	1.00	*1.00*	*1.00*	*1.00*
	5	0.377	*0.261*	1.00	1.00	1.00	*1.00*	*1.00*	*1.00*
Al. Sandwich	1	2.84	*3.52*	0.933	0.991	0.999	*0.897*	*0.983*	*0.998*
	3	1.35	*1.24*	0.999	1.00	1.00	*1.00*	*1.00*	*1.00*
	5	1.13	*1.05*	1.00	1.00	1.00	*1.00*	*1.00*	*1.00*
Gatorboard	1	2.41	*2.82*	0.989	0.999	0.999	*0.993*	*0.997*	*0.998*
	3	1.03	*0.802*	1.00	1.00	1.00	*1.00*	*1.00*	*1.00*
	5	0.547	*0.479*	1.00	1.00	1.00	*1.00*	*1.00*	*1.00*

5. Results and discussions

5.1. DOA estimation with broadband noise bursts

In the first experiment, a DNN was trained to estimate the direction of arrival of broadband noise bursts. Two distinct training, validation, and testing sets were created as described in Section 4.2: one where the noise bursts were convolved with the measured transfer function from the speaker to the vibration sensor at each angle of incidence, and one where the noise bursts were directly recorded at each angle by the structural sensor. For both the convolved and recorded data sets, a DNN was trained using each possible combination of the five affixed sensors, and the model that yielded the smallest validation RMSE for $N = 1$, 3, and 5 was applied to the testing set. For these trained models, the validation RMSE during training and the average reliability of their DOA estimates when they acted on the testing set is shown in Table 2.

The results suggest preliminary evidence of the potential for a DNN to reliably estimate the incident angle of broadband noise bursts with as few as one structural sensor. A DNN trained with data from a single sensor on the acrylic panel estimated the DOA of the bursts in the testing set to within ±5° up to 99.8% of the time. A DNN trained with data from a single sensor on the Gatorboard panel estimated the DOA of the bursts in the testing set to within ±5° up to 99.3% of the time. On the more highly damped Al. sandwich panel, a DNN trained with data from a single sensor estimated the DOA of the bursts in the testing set to within ±5° up to 93.3% of the time. It is also worth noting that for all panel materials, utilizing information from additional sensors increases the reliability of the DOA estimates within ±5° to 100%. The distributions of estimates made by the DNNs trained with data from a single sensor on the Al. sandwich and acrylic panels for bursts at selected angles of incidence in the testing set are shown in Fig. 6(a) and (b). For the Al. sandwich panel, the majority of estimates (79%) fall within the correct bin and only 1.57% of estimates fall outside of the bins directly adjacent to the correct bin. When the acrylic panel is used, the histogram shows a near-perfect distribution of estimates with 99.2% of estimates falling in the correct bin.

The reduction in reliability from the acrylic and Gatorboard panels to the Al. sandwich panel demonstrates the trade-off damping presents: while the intelligibility of recorded vibration signals is moderately improved by utilizing more highly-damped panels [18], the modal and reverberant properties of the panel yield spatial information pertinent to DOA estimation. The reduction in validation and testing reliability between the acrylic and Gatorboard panel may also suggest that there is an upper limit to the improvement in DOA estimation reliability that may come with a reduction in damping. This may be due to a greater amount of audio smearing from very high-Q modes. However, it may only suggest that the center frequencies of the mel filter bank more closely relate to the acrylic panel's modes than those of the Gatorboard panel. As mentioned in Section 2.3, the use of an optimal filter bank informed by the resonances of the panel is left to future work.

From Table 2, the largest disparity in validation RMSE during training between the convolved and recorded datasets was 0.68°. When the DNNs acted on their respective training sets, the largest disparity in the reliability of their DOA estimates to within ±5° was 3.57%. Therefore, the discussion of linearly in Section 4.1.1 is supported experimentally by comparing the results from the DNNs trained and tested using both approaches, and the convolution approach to training and testing DNNs with phonetic bursts can be utilized without diminishing accuracy.

5.2. DOA estimation with individual phonemes

In the second experiment, four distinct DNNs were trained using recordings of the usable phonemes in the word "excite". The corpus of recordings of the individual phoneme sounds was convolved with the measured transfer function from the speaker to the vibration sensor at each angle of incidence. As with the previous experiment, a DNN was trained using each possible combination of the five affixed sensors, and the model with the smallest validation RMSE for $N = 1$, 3, and 5 was applied to the testing set. The average reliability of the DOA estimates made by the trained DNN models when they acted on the phonetic testing sets is shown in Table 3.

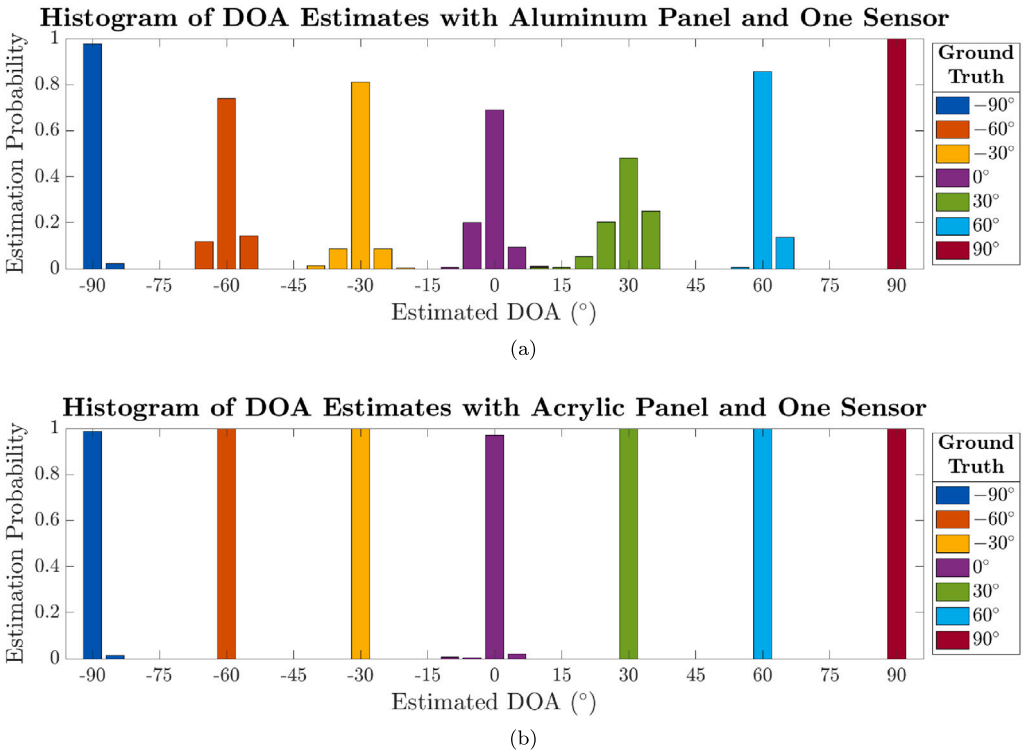


Fig. 6. (a) A histogram with a 5° bin size shows the distribution of estimates given by a DNN trained to estimate the DOA of a broadband noise burst utilizing one sensor affixed to the Al. sandwich panel acting on bursts from select angles in the testing set. The distributions are centered around the correct angle and errors spill mostly into directly adjacent bins. (b) The DNN trained to estimate the DOA of a broadband noise burst utilizing one sensor affixed to the acrylic panel shows greater reliability than that of the Al. sandwich panel, with the majority of estimates appearing in the correct bin.

Table 3
Tabulated is the reliability of the DOA estimates made by the DNNs trained with each usable phoneme in the word "excite" for $\Delta\theta$ of 5°, 10°, and 20°. The tabulated results are those given by the models with the best validation RMSE for each value of N.

Material	N	Reliability of DOA Estimates to within:											
		±5°	±10°	±20°	±5°	±10°	±20°	±5°	±10°	±20°	±5°	±10°	±20°
	1	0.762	0.921	0.972	0.752	0.928	0.978	0.754	0.926	0.982	0.821	0.955	0.987
Acrylic	3	0.976	0.995	0.998	0.98	0.999	1.00	0.988	0.999	1.00	0.999	1.00	1.00
	5	0.988	0.998	1.00	0.983	0.997	0.999	0.989	1.00	1.00	0.999	1.00	1.00
	1	0.658	0.877	0.971	0.664	0.874	0.968	0.513	0.790	0.960	0.697	0.926	0.992
Al. Sandwich	3	0.985	1.00	1.00	0.986	1.00	1.00	0.99	1.00	1.00	0.960	0.999	1.00
	5	0.982	1.00	1.00	0.992	1.00	1.00	0.998	1.00	1.00	0.999	1.00	1.00
	1	0.865	0.960	0.984	0.745	0.909	0.958	0.653	0.868	0.953	0.747	0.95	0.984
Gatorboard	3	0.993	0.996	0.998	0.978	0.998	1.00	0.959	0.999	1.00	0.995	1.00	1.00
	5	0.998	0.999	1.00	0.983	1.00	1.00	0.982	1.00	1.00	0.995	1.00	1.00
Phoneme			[aɪ]			[ɛ]			[t]			[s]	

Using a single sensor on the acrylic panel, the DNN correctly estimated the DOA of the bursts in the testing sets to within ±5° greater than 75% of the time for each phoneme. When extending $\Delta\theta$ to 10°, the DNN correctly estimated the DOA of the bursts in the testing sets greater than 92% of the time. DNNs trained using a single sensor on the Gatorboard panel yielded comparable reliability when estimating DOA over the phonetic testing sets. Once again, the acrylic and Gatorboard panels generally outperformed the highly-damped Al. sandwich panel, though the Al. sandwich panel was still able to correctly estimate the DOA of the bursts in the testing sets to within ±10° greater than 79% of the time across the tested phonemes, including 92.6% of the time for the [s] phoneme which most closely resembles broadband noise. Though the experiment is limited in nature by the use of speech from a single speaker in an anechoic environment, the results suggest that the DOA of speech sources may be estimated by a single sensor on a panel. Once again, it worth noting that utilizing information from additional sensors increases the reliability of the DOA estimates within ±5° to greater than 95% across all the tested panel materials and phonemes.

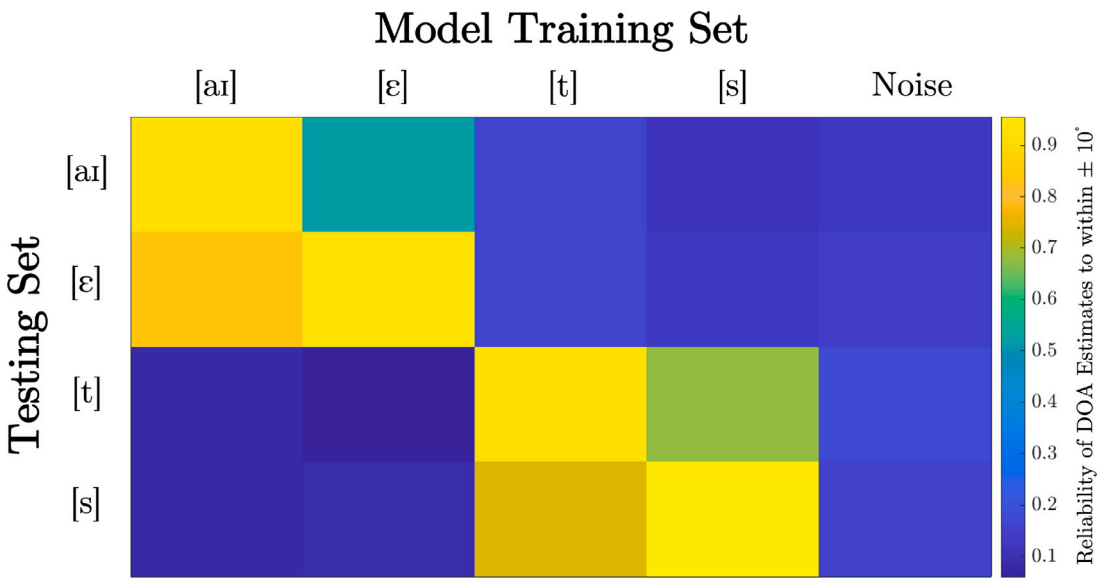


Fig. 7. The ability of a DNN trained with data from a particular sound to reliably estimate the DOA of the other speech sounds in the testing sets to within ±10° is shown. DNNs trained with phonemes of similar quality (such as the pair of [aɪ] and [ɛ] or the pair of [t] and [s]) are able to more reliably estimate DOA than DNNs trained with phonemes of dissimilar quality. The DNN trained with broadband noise bursts does a poor job generalizing to all phonemes, despite intuition that it may perform well when applied to speech sounds [s] and [t]. This is likely because "the source" in the source-filter model of human speech more closely resembles pink noise than white noise.

5.2.1. Generalization across speech sounds

The DOA estimation algorithm proposed in Fig. 4(b) would be more efficient if a single DNN could be used to estimate DOA from all four phonemes in the wake-word. In this section, we explore how well a DNN trained with data from a particular sound can estimate the DOA of a different speech sound. Each of the DNNs trained utilizing a single sensor on the acrylic panel is applied to the testing sets containing the other tested speech sounds. Fig. 7 shows the resulting matrix of estimation reliability within ±10° for this experiment. In general, DNNs trained with phonemes of similar quality (such as the pair of [aɪ] and [ɛ] or the pair of [t] and [s]) are able to more reliably estimate DOA than DNNs trained with phonemes of dissimilar quality. It is worth mentioning that the DNN trained with broadband noise bursts does a poor job generalizing to all phonemes despite intuition that it may perform well when applied to speech sounds [s] and [t]. This is likely because "the source" in the source-filter model of human speech more closely resembles pink noise than white noise. In future work, a human "source" signal may be used to see if a DNN trained with this type of noise generalizes better to speech sounds.

While the results do not demonstrate that a single DNN trained in this work can be used to estimate the DOA of any of the tested speech sounds without loss in reliability, it may be possible to reduce the number of DNNs needed for the aggregate DOA estimator shown in Fig. 4(b) by grouping speech sounds of similar quality. This will be explored in future work, along with training a model to efficiently estimate DOA from the waveform of a wake-word in its entirety.

5.2.2. Generalization between speakers

As mentioned in Section 3.1, the results shown in this work are considered speaker-dependent as the DNNs were trained using the speech sounds of a single male speaker. To see how well these DNNs can be used to estimate DOA from the speech sounds of a different speaker, a female speaker recorded a second testing set for each phoneme. The difference in the reliability of the DOA estimates made by the trained DNNs when they acted on the testing sets made by the primary male speaker and by the secondary female speaker is shown in Table 4. The models did not appear to generalize well to the female speaker, with a reduction in the probability of correct DOA estimate of up to nearly 70%. This is likely due to resonance disparities of "the filter" in the source-filter model of human speech between the two speakers. This suggests that the current proposed model exists as a speaker-dependent proof of concept, and the experimental limitation of utilizing speech from a single speaker prevents any conclusion about training a model to handle human speech in general. Generalization to a speaker-independent model is out of the scope of this work, but cost- and computationally-efficient methods for training speaker-independent models will be explored in future work.

6. Conclusions

The experiments presented in this work demonstrate that a deep neural network can utilize data from a single vibration sensor affixed to a panel's surface to estimate the direction of arrival of an incident acoustic wave to within ±5° with a high degree of

Table 4
Tabulated are the differences in the reliability of the DOA estimates made when the DNNs trained using the primary male speaker's training set acted on the testing set created by the primary male speaker and the testing set created by the secondary female speaker.

Material	N	Difference in the Reliability of DOA Estimates to within:											
		±5°	±10°	±20°	±5°	±10°	±20°	±5°	±10°	±20°	±5°	±10°	±20°
Acrylic	1	−0.648	−0.718	−0.666	−0.494	−0.507	−0.396	−0.508	−0.528	−0.428	−0.377	−0.316	−0.229
	3	−0.640	−0.513	−0.328	−0.583	−0.415	−0.233	−0.571	−0.395	−0.249	−0.173	−0.081	−0.036
	5	−0.535	−0.325	−0.127	−0.625	−0.566	−0.496	−0.576	−0.455	−0.319	−0.156	−0.095	−0.037
Phoneme			[aɪ]			[ɛ]			[t]			[s]	

reliability under constrained experimental conditions. When the incident waves contained broadband noise, the DNN trained with data from a single sensor was able to correctly estimate the DOA to within ±5° of the source location with a reliability of 99.8%. When the incident waves contained human speech, the DNN trained utilizing data from a single sensor was able to correctly estimate DOA to within ±5° of the source location with a reliability of approximately 75%, and to within ±10° of the source location with an reliability of 92% across all tested speech sounds. Although the more highly-damped Al. sandwich panel gave a reduction in reliability compared to the acrylic and Gatorboard panels, the DNN trained utilizing data from a single sensor on the Al. sandwich panel was still able to correctly estimate the DOA of the bursts in the testing set to within ±10° with a reliability of 79% for each phoneme, and up to 92.6% reliability for the [s] phoneme. It is worth noting that utilizing information from additional sensors increases the reliability within ±5° to greater than 95% across all the tested panel materials and phonemes within the experimental conditions.

The panel's damping introduces a trade-off between the intelligibility of a recorded audio signal and the reliability with which a DNN can estimate the DOA of an incident acoustic wave. For highly-damped panels such as the Al. sandwich panel used in this work, high-frequency resonant modes are quickly damped out, making it difficult for a DNN to utilize recordings of these vibrations for DOA estimation. The reduction in validation RMSE and estimation reliability between the DNNs trained with broadband noise bursts incident to the acrylic and Gatorboard panels also suggests that using very lightly-damped materials may also hinder a DNN's ability to estimate DOA due to degradation of the captured surface vibrations via audio smearing and severe reverberation [18]. While experimental results suggest that high-Q resonant modes are useful for DOA estimation, the long decay times associated with these high-Q resonances may smear speech signals that would otherwise occur during shorter time windows. As recorded signals used for this preliminary experiment were all of the same length, it will be necessary to investigate these transient effects in future work. Additionally, this discrepancy may also occur if the center frequencies of the mel filter bank are better aligned with the specific resonant frequencies of the acrylic panel modes than the resonant frequencies of the Gatorboard panel modes.

An important next step is to explore the use of an optimal filter bank informed by the resonances of the panel. Table 1 shows that the resonant frequencies of the panels tested in this work are not closely related to those of the other panels nor to the mel filter bank. The spatial information contained in the panel's modes may be more efficiently extracted using the presented energy-summing technique utilizing a filter bank whose center frequencies match the specific resonant frequencies of each panel. This would also enable the rejection of the contributions of those frequency bands that do not contain spatial information, reducing the size of the feature vectors that are used by the DNNs to increase computational efficiency. Obtaining an optimized feature vector via energy-summing distinguishes this approach from that of Badawy and Dokmanić [20] where speech signals were localized using a single microphone within a structure using non-negative matrix factorization, which requires the storage of a large dictionary of responses. Additional future work aims to improve the computational efficiency of the DOA estimation algorithm shown in Fig. 4(b), particularly in the aggregation step. The presented results demonstrate that a single DNN trained in this work cannot be used to estimate the DOA of all of the tested speech sounds without loss in reliability, though they suggest that it may be possible to reduce the number of DNNs by grouping speech sounds of similar quality. And while the DNN trained with broadband noise bursts does a poor job estimating the DOA of the tested phonemes, using a DNN trained with true human "source" signals for this task will be explored in future work. Additionally, a DNN distinct from those employed in this work may be trained to efficiently estimate DOA from the waveform of a wake-word in its entirety.

The results shown in this work are presented acknowledging several experimental limitations. Firstly, each of the recordings used to train the models were made under anechoic conditions. Under more realistic and noisy conditions, the reliability with which a model can estimate DOA might be significantly hindered. To address this, it may be necessary to collect significant training data for panels installed in a wide variety of acoustic environments, or to perform a pre-processing step to suppress known sources of environmental noise. The presented experiments are also limited by only utilizing speech from a single speaker. The results tabulated in Table 4 demonstrate that models trained in this way have difficulty estimating DOA from human speech in general. This suggests that the current proposed model exists as a speaker-dependent proof of concept. Training a speaker-independent model would likely require significant training data from a wide variety of speakers. While broad generalization across speakers is out of the scope of this work, further experimentation in this area is needed to alleviate the limitation of speaker-dependence in a more robust solution.

Withstanding these experimental limitations, this work serves as an important preliminary step in enabling signal processing techniques such as source localization and tracking, source separation, and beamforming to be performed by devices where form-factor and durability requirements prohibit the use of conventional microphones and microphone array systems. Using one inexpensive structural sensor (or a small array of these sensors) for these applications also may reduce the cost of audio capture systems in smart devices where form-factor and DOA estimation are functional priorities.

CRediT authorship contribution statement

Tre DiPassio: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Visualization. **Michael C. Heilemann:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Mark F. Bocko:** Resources, Writing – review & editing, Supervision.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by NSF Award 2104758. The authors would also like to acknowledge undergraduate students Luke Nash, Evan Lo, Benjamin Kevelson, and Julia Weinstock for their assistance in data collection.

References

[1] T. Nishiura, S. Nakamura, Talker localization based on the combination of DOA estimation and statistical sound source identification with microphone array, in: IEEE Workshop on Statistical Signal Processing, 2003, pp. 597–600.
[2] L.C.F. Nogueira, M.R. Petraglia, Robust localization of multiple sound sources based on BSS algorithms, in: 2015 IEEE 24th International Symposium on Industrial Electronics, ISIE, 2015, pp. 579–583.
[3] S.S. Mane, S.G. Mali, S.P. Mahajan, Localization of steady sound source and direction detection of moving sound source using CNN, in: 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT, 2019, pp. 1–6.
[4] H. Sundar, W. Wang, M. Sun, C. Wang, Raw waveform-based end-to-end deep convolutional network for spatial localization of multiple acoustic sources, in: ICASSP 2020, 2020.
[5] N. Liu, H. Chen, K. Songgong, Y. Li, Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays, J. Acoust. Soc. Am. 149 (2) (2021) 1069–1084.
[6] B. Van Veen, K. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. 5 (2) (1988) 4–24, http://dx.doi.org/10.1109/53.665.
[7] S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Trans. Signal Process. 49 (8) (2001) 1614–1626, http://dx.doi.org/10.1109/78.934132.
[8] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M.L. Seltzer, H. Zen, M. Souden, Speech processing for digital home assistants: Combining signal processing with deep-learning techniques, IEEE Signal Process. Mag. 36 (6) (2019) 111–124, http://dx.doi.org/10.1109/MSP.2019.2918706.
[9] M. Brandstein, H. Silverman, A robust method for speech signal time-delay estimation in reverberant rooms, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 1997, pp. 375–378.
[10] B. Kwon, Y. Park, Y.-s. Park, Analysis of the GCC-PHAT technique for multiple sources, in: ICCAS 2010, 2010, pp. 2070–2073.
[11] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. 24 (4) (1976) 320–327.
[12] M. Kaveh, A. Barabell, The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise, IEEE Trans. Acoust. Speech Signal Process. 34 (2) (1986) 331–341.
[13] M.C. Heilemann, D.A. Anderson, S. Roessner, M.F. Bocko, The evolution and design of flat-panel loudspeakers for audio reproduction, J. Audio Eng. Soc. 69 (1/2) (2021) 27–39, http://dx.doi.org/10.17743/jaes.2020.0057.
[14] Y. Choi, C. Oh, K. Park, S. Lee, Organic Light Emitting Display Device Including a Sound Generating Apparatus, US Patent 9818805B2, Nov. 14, 2017.
[15] T.-H. Kim, G.-C. Park, Display Device, US Patent 20190163234A1, May. 30, 2019.
[16] S. Lee, K. Park, Y. Choi, K. Kim, M. Bae, Actuator fixing device and panel vibration type sound-generating display device including the same, US Patent 20190215607A1, Jul. 11, 2019.
[17] S. Roessner, M. Heilemann, M.F. Bocko, Evaluating listener preference of flat-panel loudspeakers, in: Audio Engineering Society Convention 147, 2019, URL http://www.aes.org/e-lib/browse.cfm?elib=20603.
[18] T. DiPassio, M.C. Heilemann, M.F. Bocko, Audio capture using structural sensors on vibrating panel surfaces, J. Audio Eng. Soc. 70 (12) (2022) 1027–1037.
[19] S. Kita, Y. Kajikawa, Fundamental study on sound source localization inside a structure using a deep neural network and computer-aided engineering, J. Sound Vib. 513 (2021) 116400, http://dx.doi.org/10.1016/j.jsv.2021.116400.
[20] D. El Badawy, I. Dokmanić, Direction of arrival with one microphone, a few legos, and non-negative matrix factorization, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12) (2018) 2436–2446.
[21] C. Fuller, S. Elliot, P. Nelson, Active Control of Vibration, Academic Press, 1996.
[22] B. Wang, C.R. Fuller, E.K. Dimitriadis, Active control of noise transmission through rectangular plates using multiple piezoelectric or point force actuators, J. Acoust. Soc. Am. 90 (5) (1991) 2820–2830.
[23] L.A. Roussos, Noise Transmission Loss of a Rectangular Plate in an Infinite Baffle, NASA Technical Paper 2398, 1985.
[24] L. Cremer, M. Heckl, B. Petersson, Structure-Borne Sound: Structural Vibrations and Sound Radiation at Audio Frequencies, Springer Berlin Heidelberg, 2005.
[25] A. Mitchell, C. Hazell, A simple frequency formula for clamped rectangular plates, J. Sound Vib. 118 (2) (1987) 271–281, http://dx.doi.org/10.1016/0022-460X(87)90525-6.
[26] T. Chiba, M. Kajiyama, The Vowel: Its Nature and Structure, Vol. 652, Phonetic society of Japan Tokyo, 1958.
[27] G. Fant, Acoustic Theory of Speech Production, No. 2, Walter de Gruyter, 1970.
[28] G.E. Peterson, H.L. Barney, Control methods used in a study of the vowels, J. Acoust. Soc. Am. 24 (2) (1952) 175–184.
[29] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, J. Acoust. Soc. Am. 8 (3) (1937) 185–190.
[30] W. Koening, A new frequency scala for acoustic measurements, Bell Lab. Rec. (1949) 299–301.
[31] L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, 2010, arXiv preprint arXiv:1003.4083.
[32] P. Boersma, Praat, a system for doing phonetics by computer, Glot. Int. 5 (9) (2001) 341–345.
[33] S.Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, et al., Deep voice: Real-time neural text-to-speech, in: International Conference on Machine Learning, PMLR, 2017, pp. 195–204.
[34] N.R. French, J.C. Steinberg, Factors governing the intelligibility of speech sounds, J. Acoust. Soc. Am. 19 (1) (1947) 90–119.

[35] F.J. Fahy, P. Gardonio, Sound and Structural Vibration: Radiation, Transmission and Response, Elsevier, 2007.
[36] Q. Li, X. Zhang, 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.
[37] S. Adavanne, A. Politis, T. 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, http://dx.doi.org/10.23919/EUSIPCO.2018.8553182.