2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3503-2372-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/WASPAA58266.2023.10248068

ESTIMATING THE DIRECTION OF ARRIVAL OF A SPOKEN WAKE WORD USING A SINGLE SENSOR ON AN ELASTIC PANEL

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

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

ABSTRACT

The vibrations induced in an elastic panel from an incident acoustic pressure wave are a function of the resonant mode structure of the panel and the angle of incidence of the acoustic wave. In this paper it has been demonstrated how measurement of the panel's modal response with a single structural vibration sensor may be employed to infer the direction of arrival (DOA) of the incident sound. The method is dependent on the frequency content of the acoustic wave, as modes that provide important spatial information about the source may not be excited if the acoustic signal lies outside their resonant bandwidths. This work explores techniques for extending this single-sensor approach to DOA estimation for speech signals, which represent a realistic use case for applications such as smart audio devices. Feature sets including Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs), and linear spectrograms, were used to train convolutional and feedforward neural networks to estimate the DOA of a wake word recorded by a single structural vibration sensor affixed to a panel. The experiments were carried out in semi-anechoic conditions and are thus presented as proof of concept. Additionally, the models presented are compact enough to be deployed on embedded/edge hardware commonly used in smart audio devices. The trained models estimated the DOA of the wake word utterance to within ±5° with an average reliability of 83.1% when using MFCCs as features. This average reliability improved to 92.23%, with a maximum reported reliability of 99.9%, when using Mel and magnitude spectrograms and an additional hardware-specific feature set, suggesting that single-sensor DOA estimates for speech signals may be improved by using more spectrally complete feature sets.

Index Terms— Direction of Arrival (DOA), Vibration Sensing, Source Localization, Structural Sensors

1. INTRODUCTION

Smart audio devices utilize multi-microphone arrays to determine the direction of arrival (DOA) of a user's speech before performing directional speech enhancement [1]. Common approaches for estimating DOA, such as inter-sensor time difference of arrival, generalized cross-correlation with phase transform, and the multiple signal classification algorithm, require an array of transducers and typically become more reliable as the size of the array is increased [2, 3, 4, 5]. However, adding elements to the array increases the device's power consumption, manufacturing expense, and computational complexity. Therefore, developing techniques to reduce the number of sensors needed for reliable DOA estimation is an important design consideration.

It has recently been demonstrated that structural vibration sensors affixed to an elastic panel are able to record acoustic speech signals while preserving intelligibility sufficient for transcription by automatic speech recognition systems [6, 7]. The relative excitations of the panel's bending modes change depending on the angle of incidence of the acoustic wave [8, 9, 10], and deep neural networks (DNNs) may be trained to reliably estimate a wave's DOA by associating the magnitude response of the panel measured by a single structural vibration sensor with the incident angle of an acoustic wave. This previous experiment provided a proof of concept of the single-sensor DOA estimation method, although the conclusions were limited by the use of source signals containing only stationary white noise and isolated phonemes [11, 12].

An arrangement that more closely approximates an eventual use case is the estimation of a source's DOA from full phonetic utterances, such as device-specific wake words. The scope of this work is to provide experimental evidence that the DOA of complete speech signals can be estimated using information from a single sensor affixed to an elastic panel. Additionally, the feature sets and neural networks used in the experimental portion of this work are compact enough to be deployed on an embedded processor. As such, the application of this work is as a single-sensor DOA estimation system that is deployable on the edge hardware that is becoming ubiquitous in commercially available smart devices [13]. We begin with a brief overview of the vibrations of an elastic panel excited by an incident acoustic wave.

2. THEORETICAL DEVELOPMENT

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

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

where b is the panel's mechanical loss factor. Solutions for (1) are well known (see for example [14]). The displacement $w(x, y, t)$ is a separable function of space and time, given by

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

The spatial component $\varphi(x, y)$ can be written as a sum of the panel's bending modes, expressed 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 excited with amplitude α_r.

The panels used in the experimental portion of this work are rectangular panels with clamped boundary conditions, and therefore $\Phi_r(x, y)$ contains separable sinusoidal functions along the panels

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

length L_x and width L_y. Modal indices r_m and r_n represent the number of half-wavelengths in the horizontal and vertical dimensions, respectively. The bandwidth of each mode is determined by the mode's quality factor Q_r, which can be expressed as,

$$Q_r = \frac{\omega_r \rho h}{b}, \tag{4}$$

where ω_r is the resonant frequency of the r^{th} mode.

When the panel is excited by an acoustic plane wave with pressure amplitude P_i at frequency ω incident at angle θ_i in the azimuthal plane, the relative excitation of the panel's modes can be computed following [8, 9, 10] as,

$$\alpha_r = \frac{8 P_i I_{r_m}(\theta_i, \omega) I_{r_n}(\theta_i, \omega)}{\rho h(\omega_r^2 - \omega^2 + j\omega_r \omega / Q_r)}, \tag{5}$$

where $I_{r_m}(\theta_i, \omega)$ 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.

An acoustic source playing signal $s(t)$ to a panel with a structural vibration sensor affixed at point (x_0, y_0) on its surface is shown in Fig. 1. From (3) and (5), the transfer function from source to sensor $h_{\theta_i}(t)$ is dependent on the incident angle θ_i. Since the panel displacement is small such that it operates in a linear regime [8, 11], the velocity response at the sensor's position can be expressed with convolution as,

$$\dot{w}(x_0, y_0, t) = s(t) \circledast h_{\theta_i}(t). \tag{6}$$

In the experimental portion of this work, $s(t)$ contained full utterances of the wake word, and the panel's response to these phonetic excitations were recorded at various angles of incidence. Following (5), varying θ_i causes subtle variations in the spectral properties of the recordings. In this work, we report experimental evidence that neural networks can be trained to recognize these variations and estimate the DOA of the speech source using information from one structural vibration sensor.

3. METHODOLOGY

3.1. Dataset

Two participants, one male and one female, each recorded 300 sentences containing common phrases used to interact with smart audio devices [15]. The recordings were made in an acoustically treated studio environment with a Shure SM58 microphone at a sample rate of 48 kHz, and later downsampled to a sample rate of 16 kHz. The participants started each sentence with "Hey, Alexa", the wake word phrase most commonly used to activate Amazon's line of smart devices. Pronunciation and inflection were naturally varied based on the context of the rest of the command. For this experiment, the wake word in each recording was isolated and used to train the neural networks as described in the following sections.

3.2. Experimental Setup

The experimental setup used to record panel vibrations induced at various angles of incidence is shown in Fig. 1. A 2 mm thick acrylic panel with Young's Modulus $E = 3.2$ GPa, Poisson's ratio $\nu = 0.35$, density $\rho = 1,180$ kg/m^3, and dimensions $(L_x, L_y) = (26 \text{ cm}, 36 \text{ cm})$ was mounted in a semi-anechoic chamber. The panel was placed on a rotary table to allow the incident angle of the acoustic wave to be measured between $\theta_i = -90°$ and 90° in 5° increments. A KEF LS50 loudspeaker was used to reproduce the source signal, and was placed at a distance of 50 centimeters in front of the center of the panel [11].

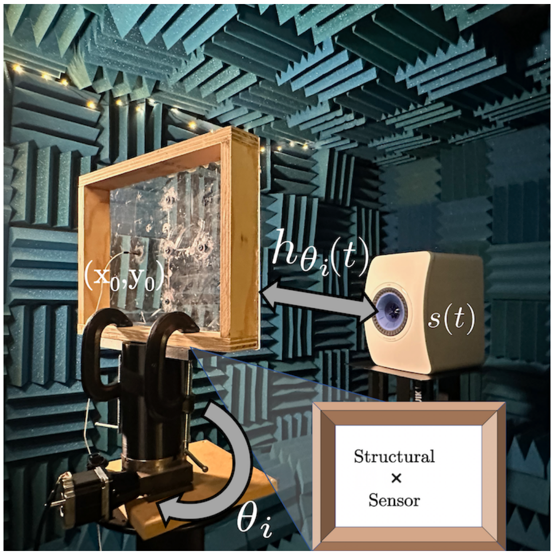

Figure 1: Experimental setup used to record panel vibrations induced by incident acoustic waves with various angles of incidence.

The panel was equipped with a single PCB Piezotronics U352C66 accelerometer arbitrarily positioned off-center in each dimension [16]. The sensor was used to measure the panel vibrations in response to the wake word recordings at each incident angle. In total 11,100 wake word recordings were recorded with the panel, 300 sentences at each of 37 incident angles.

3.3. Features and Network Architectures

The models employed in this work were trained with features that contain spectral information, since the amplitudes of each resonant mode are angularly dependent as shown in (5). Since edge AI hardware has recently seen a rapid increase in deployment with smart audio devices [13], the experimental results are reported using feature sets and neural network architectures compact enough to be embedded on edge hardware; in particular, devices that are supported by the Edge Impulse AI platform [17].

3.3.1. Spectral Features

In a previous work, Mel-frequency cepstral coefficients (MFCCs) were demonstrated to be an effective feature set for estimating DOA of recordings of phonemes in isolation made by sensors mounted on elastic plates [11]. The speech signals used in this experiment contained full phonetic phrases. Therefore, in addition to the use of an MFCC feature set, Mel and magnitude spectrograms were also used as features to train the neural networks in an effort to accommodate the wider spectral and temporal variations associated with speech signals. Examples of these feature vectors are shown in Fig. 2 (a).

Additionally, a proprietary feature set created for edge hardware developed by Syntiant was utilized for model training [19]. Syntiant's tiny machine learning (TinyML) development board is a

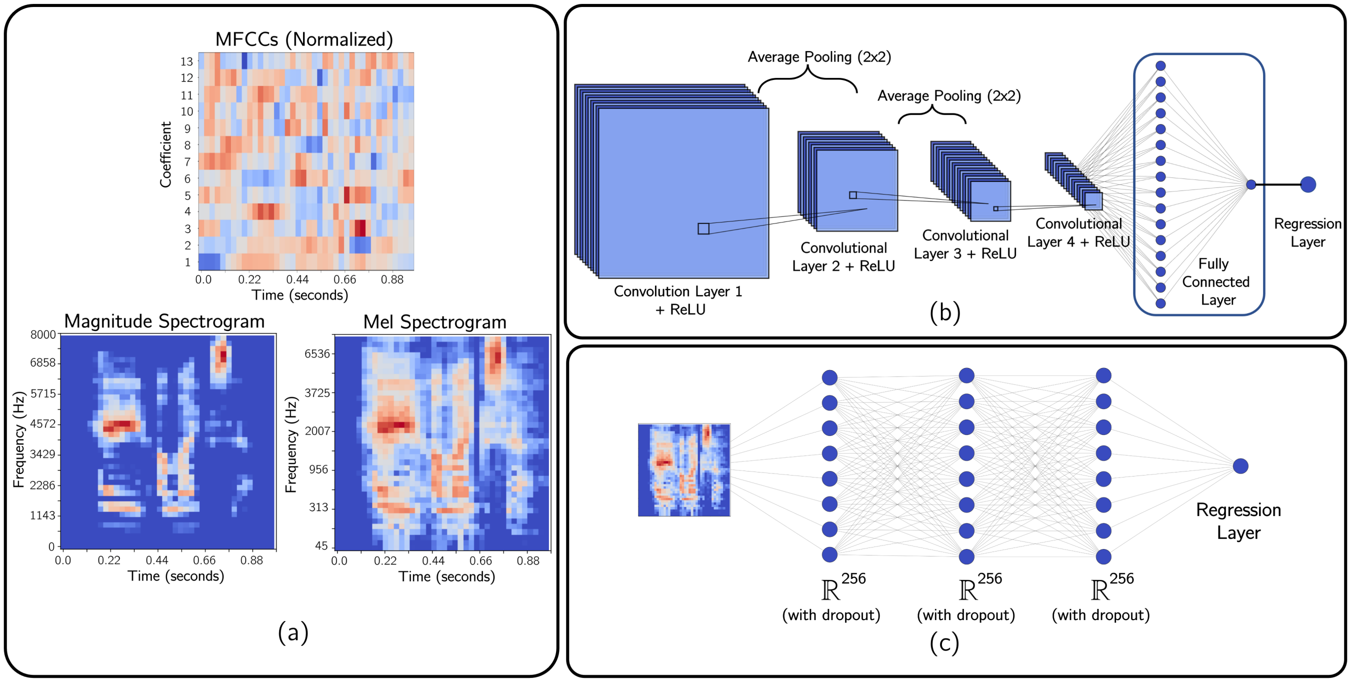

Figure 2: (a) Examples of the feature vectors extracted from the wake word recordings made by the panel using Edge Impulse [17]. The same features were also extracted locally for offline training of the CNN. (b) The architecture of the CNN used in this experiment. (c) The architecture of the FNN that is compatible with the Syntiant hardware [18].

commercially available edge device that features an always-on neural decision processor (NDP) for performing wake word detection and other real-time speech processing tasks [18].

3.3.2. Model Architecture

The models trained in this work employ two architectures that are compatible with TinyML, and are compact enough to be embedded on commercially available edge devices. The first of two architectures, visualized in Fig. 2 (b), is a two-dimensional convolution neural network (CNN) with a regression output layer [20]. Note that this CNN is not proposed as an optimal model, but is employed in this work as a proof of concept. The second model, a feedforward neural network (FNN) was chosen because it is built into the hardware on the Syntiant NDP [18]. Its architecture is shown in Fig. 2 (c).

Distinct instances of both architectures were trained with each of the feature sets shown in Fig. 2 (a). Additionally, the FNN was trained with the proprietary feature set created for the Syntiant hardware accessable on Edge Impulse. Model training was performed using the wake words spoken by each participant individually, with 8880 wake word recordings split into training and validation sets with a ratio of 80:20. The remaining 2220 recordings were used to test each model. The models were trained to minimize the mean square error between the predicted angle and the ground truth. Note that because the models were each trained with only one voice, they serve as speaker-dependent proofs of concept. Generalization to a speaker-independent model is out of the scope of this work, although the results here suggest that these methods will generalize to a wide range of voices with different spectral content.

3.4. Evaluation Metric

Each model is evaluated on its ability to correctly predict the true incident angle θ_i within a defined angular tolerance $\pm\Delta\theta_i$. Following [21, 22], the reliability with which the model estimates the DOA of the speech source is expressed as the number of correct predictions within $\pm\Delta\theta_i$, divided by the total number of utterances tested. Experimental results are reported for angular tolerances of $5°$, $10°$, and $20°$, consistent with the resolutions used in previous experiments [23].

4. RESULTS AND DISCUSSION

The reliability with which each model is able to estimate the DOA of the speech signal is shown in Table 1 various for angular tolerances. The CNN was able to estimate the DOA of both participant's voices to within $\pm 5°$ with up to 98.3% reliability using a single structural vibration sensor. The models trained with MFCC features under-performed the models trained with the more spectrally complete Mel and magnitude spectrogram feature sets. Additionally, the CNNs trained with magnitude spectrograms as features out-performed those using Mel spectrograms. This may be due to the linear spacing of the frequency bins in the magnitude spectrogram. At sufficiently high frequencies, a large number of the panel's bending modes are excited simultaneously [24, 25]. In this frequency region of high modal overlap, individual modes are no longer discernible, which mitigates the ability of the structural sensor to relate the modal excitations given by (5) to a specific angle of incidence. Therefore, the logarithmic nature of the Mel spectrogram may result in less efficient utilization of spectral information

		Reliability (%) of DOA Estimates to within:					
Network	**Feature**	±5°	±10°	±20°	±5°	±10°	±20°
CNN	MFCC	89.1	99.3	100	82.5	98.3	99.7
	Mel-Spect	92.7	99.4	100	92.9	99.5	100
	Mag-Spect	97.1	99.8	100	98.3	99.9	100
FNN	MFCC	82.2	97.3	99.7	78.9	96.7	99.5
	Mel-Spect	94.3	99.9	100	87.9	99.6	100
	Mag-Spect	76.7	96.7	99.8	82.7	99.2	100
	Syntiant	99.7	100	100	99.9	100	100
Voice		Male			Female		

Table 1: Reliability of the DOA estimates made by the trained CNNs and the FNNs with angular tolerances of 5°, 10°, and 20°. Distinct models were trained for each feature set and speaker.

in the low-frequency region where low modal overlap occurs, and individual modes dominate the panel's spatial response. The use of panel-specific spectral features that optimize the bandwidths where individual modes are discernible is left to future work.

The FNNs trained with non-proprietary feature sets were able to estimate the DOA of both participant's voices to within ±5° with up to 94.3% reliability. As was the case for the CNNs, the FNNs trained with MFCC features under-performed those trained with the other feature sets. However, the FNNs trained with Mel spectrograms generally outperformed the models trained with magnitude spectrograms. This may be related to the limitations imposed on training time by Edge Impulse, as the magnitude spectrograms were the largest features used in this experiment. Edge Impulse recently introduced the ability to deploy pre-trained models within their framework, so re-training the FNN architecture with these feature sets in an offline setting will enable direct comparison of the results from the two networks when large feature vectors are employed, and will be explored in future work.

The FNN trained with the proprietary feature set created for the Syntiant hardware performed very well when acting on the test set, as it estimated the DOA of both participant's voices to within ±5° with up to 99.9% reliability. Although this model is currently device-specific, the reported reliability of models trained with this hardware-informed feature set is an important result that may lead to the development of an optimized, full-stack system.

It is important to note that all of the trained models were able to estimate the DOA of both participant's voices to within ±10° with greater than 96% reliability. Comparing the results across the various angular tolerances suggests that the DOA estimates returned by the models are distributed around the true incident angles. This distribution is apparent in Fig. 3, which shows the aggregate confusion matrix for the CNNs trained with the female voice with an angular tolerance of ±5°.

We wish to acknowledge some important limitations in the experimental setup. First, the wake word recordings used to train the models were made by a panel mounted in a relatively quiet semi-anechoic chamber. However, the presence of environmental noise may adversely affect the reliability of the trained models. Testing the reliability of the models in noisy environments is an important future step. It is likely that significant additional training data or the implementation of de-noising methods will be necessary for reliable model performance in more realistic environments.

In addition, each model was trained and tested on only one participant's voice at a time. As such, results are reported from models that are implicitly speaker-dependent. Generalizing to a speaker-independent model will require much more training data. However, it is encouraging that the reported results from the trained models are generally consistent across both voices. This suggests that the proposed single-sensor DOA method may be adaptable to various speech characteristics, as the voices used were inclusive of a wide range of vocal timbres.

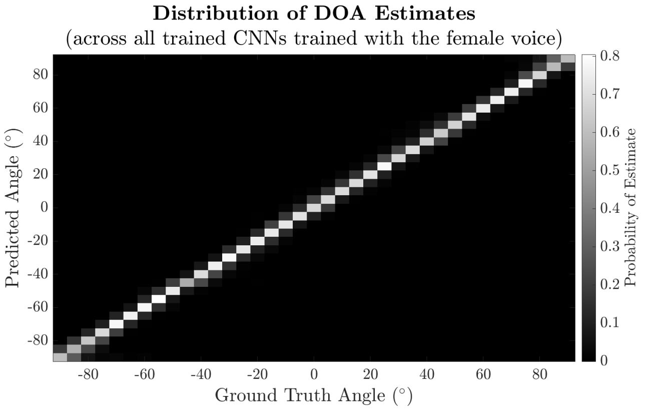

Figure 3: Confusion matrix showing the distribution of the DOA estimates returned by the CNN models trained with the female voice. The bin size is chosen to visualize an angular tolerance of ±5°.

5. CONCLUSIONS

The reported results provide experimental evidence that a single sensor affixed to an elastic panel may be utilized to perform reliable DOA estimation from recorded speech signals. In addition, the models and feature sets utilized in this work are all compact enough to be implemented within the constraints imposed by commercially available embedded/edge hardware. In particular, the performance of the FNN trained with the proprietary, hardware-specific feature set suggests the possibility of designing a highly-reliable, full-stack DOA estimation system utilizing the described methods. The trained models are presented here as a proof of concept, as they were determined for only two speakers and were tested without the presence of significant environmental noise. However, this does represent an important step toward demonstrating that the presented methods enable the DOA of a speech signal to be reliably estimated using a single sensor under these conditions. The ubiquitous time-delay and phase-based approaches to DOA estimation require transducer arrays with multiple sensing elements. Reducing the number of sensors needed to perform the tasks required by modern smart devices may lower their power consumption, manufacturing cost, and computational requirements, while offering the ability to integrate the sensor into built environments without sacrificing form-factor.

6. ACKNOWLEDGMENT

This work supported by NSF Award 2104758. Additionally, the authors would like to acknowledge Paula Sedlacek, Yutong (Cooper) Wen, and Joe Bumpus for their assistance in creating the speech dataset.

7. REFERENCES

[1] 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 Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[2] M. Brandstein and 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 vol.1.

[3] B. Kwon, Y. Park, and Y.-s. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *ICCAS 2010*, 2010, pp. 2070–2073.

[4] 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.

[5] 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.

[6] T. DiPassio, M. C. Heilemann, and M. F. Bocko, "Audio capture using structural sensors on vibrating panel surfaces," *Journal of the Audio Engineering Society*, vol. 70, no. 12, pp. 1027–1037, December 2022.

[7] T. DiPassio, M. C. Heilemann, B. Thompson, and M. F. Bocko, "Audio capture using piezoelectric sensors on vibrating panel surfaces," *154th Convention of the Audio Engineering Society*, In Press 2023.

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

[9] 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.*, vol. 90, no. 5, pp. 2820–2830, 1991.

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

[11] T. DiPassio, M. C. Heilemann, and M. F. Bocko, "Direction of arrival estimation of an acoustic wave using a single structural vibration sensor," *Journal of Sound and Vibration*, vol. 553, p. 117671, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022460X23001207

[12] 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," *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, In Press 2023.

[13] "Forecast shipments of artificial intelligence (ai) edge processors worldwide in 2019 and 2023 (in billion units)," June 2019. [Online]. Available: https://www.statista.com/statistics/1027163/edge-ai-processor-market-worldwide/

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

[15] B. Kinsella, "Smart speaker use case frequency in the united states as of january 2020," May 2020. [Online]. Available: https://www.statista.com/statistics/994696/united-states-smart-speaker-use-case-frequency/

[16] PCB Piezotronics. [Online]. Available: http://www.pcb.com

[17] Edge Impulse. [Online]. Available: http://https://www.edgeimpulse.com

[18] NDP101 Neural Decision Processor. [Online]. Available: https://www.syntiant.com/ndp101

[19] Syntiant. [Online]. Available: http://https://www.syntiant.com

[20] "Train a convolutional neural network for regression," *MathWorks*. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/train-a-convolutional-neural-network-for-regression.html

[21] N. Liu, H. Chen, K. Songgong, and Y. 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.

[22] 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.

[23] S. Adavanne, A. Politis, and 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.

[24] D. A. Anderson, M. C. Heilemann, and M. F. Bocko, "Measures of vibrational localization on point-driven flat-panel loudspeakers," in *Proceedings of Meetings on Acoustics 171ASA*, vol. 26, no. 1. Acoustical Society of America, 2016, p. 065003.

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