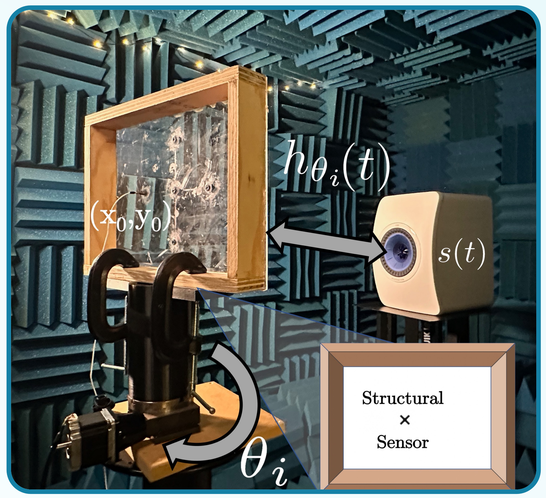
Estimating the Direction of Arrival of a Spoken Wake Word using a Single Sensor on an Elastic Panel

Tre DiPassio, Michael C. Heilemann, Benjamin R. Thompson, Mark F. Bocko
Department of Electrical and Computer Engineering

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.

Experimental Setup



Motivation

Smart acoustic surfaces allow for **seamless integration** of a smart speaker into existing environments, as any surface (such as picture frames and artwork) can be used

Mounting sensors internally to the display **eliminates the need for case penetrations,** improving the device's water resistance and durability

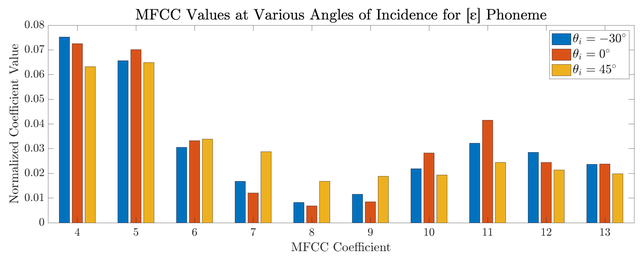
Extended surfaces allow for **signal processing advantages**, as sensors can be placed further apart than the standard 1-4 cm on modern smart devices

VS

By coupling to a modal surface, direction of arrival estimation and beamforming can occur with as few as one sensor, which can **lower manufacturing cost**

Smart devices with display panels can include **higher-fidelity audio without changing form factor** by using their screen's surface for audio reproduction

Estimating DOA using an Elastic Panel's Harmonic Properties



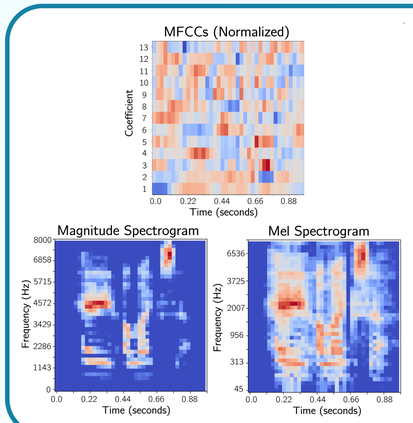
Selected **mel-frequency cepstral coefficients (MFCCs)** extracted from recordings of an elastic panel's vibrational response to acoustic waves containing the **speech sound "eh"** incident at −30˚, 0˚, and 45˚, measured using a single structural vibration sensor. This figure demonstrates that the MFCCs are dependent on the incident angle of the acoustic wave. A neural network may therefore utilize an MFCC vector to create decision boundaries and estimate the DOA of the excitation using information from a **single structural vibration sensor**.

Material	# Sensors	Average Reliability (%) within: ±5°	±10°	±20°
Acrylic	1	77.2	93.3	98.0
	3	98.6	99.8	100
	5	99.0	99.9	100

Tabulated is the average reliability of the DOA estimates made by a recurrent neural network trained with **MFCC feature vectors** extracted from recordings of a panel's vibrational response to incident **speech sounds**. The recordings were made using 1, 3, and 5 structural vibration sensors. The results demonstrate that MFCC feature vectors can be employed to reliably estimate DOA using data from a **single structural vibration sensor**.

Results are reproduced from our recent publication in the Journal of Sound and Vibration, linked here

Feature Extraction, Model Training, and Experimental Results



Two participants, one male and one female, each recorded 300 sentences containing common phrases used to interact with smart audio devices. The participants started each sentence with **"Hey, Alexa",** the wake word phrase commonly used to activate Amazon's line of smart devices. These wake word recordings were then set incident to an acrylic panel at angles of incidence ranging from −90° to +90˚ in 5˚ increments. The panel's response to each excitation was recorded using a single structural vibration sensor affixed to the panel.

From these recorded panel responses, feature vectors containing **MFCCs, Mel spectrograms, and magnitude spectrograms** were extracted. Examples of these features are visualized at left using Edge Impulse (edgeimpulse.com). Additionally, a proprietary feature set created for edge hardware developed by **Syntiant** (syntiant.com) was extracted from the panel's responses. The Syntiant's tiny machine learning (TinyML) development board features an always-on neural decision processor (NDP) for performing wake word detection and other real-time speech processing tasks.

These feature vectors were used to train two architectures that are compatible with TinyML and are compact enough to be embedded onto commercially available edge devices. The first architecture is a two-dimensional **convolution neural network (CNN)** with a regression output layer. The second model, a **feedforward neural network (FNN)** was chosen because of its compatibility with the hardware design of the Syntiant NDP.

The reliability with which each model estimated the DOA of the speech source to within various angular tolerances is tabulated at right. The trained CNN models demonstrated the ability able to estimate the DOA of both participant's voices to within ±5˚ with up to **98.3% reliability** using a single structural vibration sensor. The FNNs trained with the non-proprietary feature sets were able to estimate the DOA of both participant's voices to within ±5˚ with up to **94.3% reliability**.

The FNN trained with the proprietary feature set created for the Syntiant hardware performed very well, estimating the DOA of both participant's voices to within ±5˚ with up to **99.9% reliability**. Although this feature set 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.**

Network	Feature	Reliability (%) of DOA Estimates to within: ±5°	±10°	±20°	±5°	±10°	±20°
		Male			Female		
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