OPEN ACCESS Freely available online

T. Dipassio, M. C. Heilemann and M. F. Bocko,
"Audio Capture Using Structural Sensors on Vibrating Panel Surfaces"
J. Audio Eng. Soc., vol. 70, no. 12, pp. 1027–1037, (2022 December).
DOI: https://doi.org/10.17743/jaes.2022.0049

Audio Capture Using Structural Sensors on Vibrating Panel Surfaces

TRE DIPASSIO, *AES Student Member*, **MICHAEL C. HEILEMANN,** *AES Member* **AND**
(tredipassio@rochester.edu) (mheilema@ur.rochester.edu)

MARK F. BOCKO, *AES Member*
(mbocko@ur.rochester.edu)

University of Rochester, Rochester, NY

The microphones and loudspeakers of modern compact electronic devices such as smartphones and tablets typically require case penetrations that leave the device vulnerable to environmental damage. To address this, the authors propose a surface-based audio interface that employs force actuators for reproduction and structural vibration sensors to record the vibrations of the display panel induced by incident acoustic waves. This paper reports experimental results showing that recorded speech signals are of sufficient quality to enable high-reliability automatic speech recognition despite degradation by the panel's resonant properties. The authors report the results of experiments in which acoustic waves containing speech were directed to several panels, and the subsequent vibrations of the panels' surfaces were recorded using structural sensors. The recording quality was characterized by measuring the speech transmission index, and the recordings were transcribed to text using an automatic speech recognition system from which the resulting word error rate was determined. Experiments showed that the word error rate (10%–13%) achieved for the audio signals recorded by the method described in this paper was comparable to that for audio captured by a high-quality studio microphone (10%). The authors also demonstrated a crosstalk cancellation method that enables the system to simultaneously record and play audio signals.

0 INTRODUCTION

The market for smart speaker technology is growing rapidly. Reports by Strategy Analytics show smart speaker sales consistently increasing year after year, and sales of display-enabled smart devices are growing faster than their display-less counterparts [1]. Although adding a display to a smart speaker enables additional modes of interaction for the user, it limits the amount of space in the device available for the audio system. Because smart devices are frequently used to play music [2], the development of display-based smart speakers with enhanced sound quality is desirable.

A promising alternative to conventional audio reproduction on smart displays is to radiate sound from the screen of the display itself, employing one or more force actuators affixed to the back of the screen to induce vibrations in the display panel [3–5]. Techniques for improving sound quality from vibrating panels have been studied extensively in the literature [6, 7], and can be applied to sound radiation from monolithic organic LED display panels [8]. With proper design and tuning, panel loudspeakers can produce sound quality comparable to prosumer-grade bookshelf speakers [9].

Although audio reproduction via bending waves in a panel has been extensively reported in the literature, little work has been done to analyze the fidelity of audio recorded by measuring the vibrations of panels induced by acoustic waves from speech and other audio sources. This work explores the accuracy with which automatic speech recognition (ASR) systems can transcribe speech signals recorded by monitoring panel vibrations, both with and without simultaneous audio reproduction from actuators on the same panel. If sufficient transcription accuracy is achievable, a display panel could effectively be employed as a full-duplex audio interface.

The proposed surface audio system can provide several advantages for smart devices. A physical vulnerability in audio-sensing electronic devices is that embedded microelectromechanical systems (MEMS) microphones require enclosure penetrations, making devices susceptible to environmental factors. Employing structural sensors mounted to the back side of the display (internal to the enclosure) would eliminate the need for case penetrations. Additionally, the extended surface of a display panel enables source direction information to be inferred, which can be used for noise reduction improvement, de-reverberation, and speech enhancement [10, 11]. Smart devices utilizing a surface audio system have form factor advantages, and seamless integration of smart speaker technology into existing environments is possible.

The usefulness of surface vibrations captured by structural sensors has been demonstrated in various control and sensing applications. Meirovitch [12] and Fuller [13] summarize several theoretical approaches to feedback control of structures using force actuators and sensors. Rubenstein et al. [14] used accelerometers to update Kalman filters in a feedback loop to control bending modes of a thin metal sheet. Active surfaces made from layers of piezorubber have been used to control transmissions and reflections simultaneously in a confined acoustic channel [15, 16]. Recently, vibration data from active surfaces has been used in machine learning algorithms to extract information about systems and acoustic environments. Gamboa-Montero et al. have shown how surface vibrations within a social robotic system can be used to localize and distinguish touching gestures on the robot's body [17]. Kita and Kajikawa have used surface vibrations on a structure to replace microphone-array sound source localization techniques for sources inside a structure [18].

The fidelity of these surface vibration signals in the audio band has not been reported previously. An apparent issue in audio recording via panel vibration monitoring is the added reverberation due to the panel's resonant properties. Human speech contains wide-bandwidth bursts, such as oral stops and plosives [19], and when a panel is excited by a signal containing such impulsive sounds, reverberation can be observed at the frequencies where the panel has modal resonances, leading to reverberant audio artifacts. This can be combated to some extent by using more heavily damped materials [20]. Gelfand and Silman have shown that consonant intelligibility is reduced under reverberant conditions [21], and the resonances in flat-panel loudspeakers were shown to degrade the quality of reproduced speech [22]. Although speech quality may be reduced by the resonant modes of a panel, the effects on the performance of speech recognition tasks remain unknown.

This work will explore whether audio captured from surface vibrations can be used with ASR systems with negligible effect on transcription accuracy, thus demonstrating that these systems are viable for use in smart audio devices. This paper begins with a brief overview of the physics of vibrating panels to provide a mathematical basis for the discussion of the design challenges faced when capturing audio with a structural vibration sensor.

1 THEORETICAL DEVELOPMENT

1.1 Design of Flat Panel Acoustic Surfaces

The out-of-plane displacement $\tilde{\varphi}$ at time t and point (x, y) on a damped, isotropic panel's of bending stiffness D and density ρ subject to external load $p(x, y, t)$ is shown by Cremer et al. [23] to be

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

where h is the thickness of the panel and R_m is the panel's mechanical resistance. The bending stiffness D is determined by the Young's Modulus E and Poisson's ratio v,

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

In addition to the panel's physical properties, the response of a vibrating panel is also determined by its shape and boundary conditions. In this work, the authors assume that the panel is rectangular, with dimensions (L_x, L_y) and that the edges are fully clamped. Under these boundary conditions, an approximation for the resonant frequency of each bending mode ω_r is given by Mitchell and Hazel [24].

Following Fuller [13], the out-of-plane displacement of a panel can be expressed as a superposition of the panel's modes,

$$\tilde{\varphi}(x, y, \omega) = \sum_{r=1}^{\infty} \frac{P_r\Phi_r(x, y)}{\rho h(\omega_r^2 - \omega^2 + j\omega_r\omega/Q_r)}, \tag{3}$$

where $\Phi_r(x, y)$ is the shape of each resonant mode's shape, P_r is the pressure on each resonant mode due to the input disturbance, ω_r is the resonant frequency of each mode, and Q_r is the quality factor of each mode given by

$$Q_r = \frac{\omega_r\rho h}{R_m}. \tag{4}$$

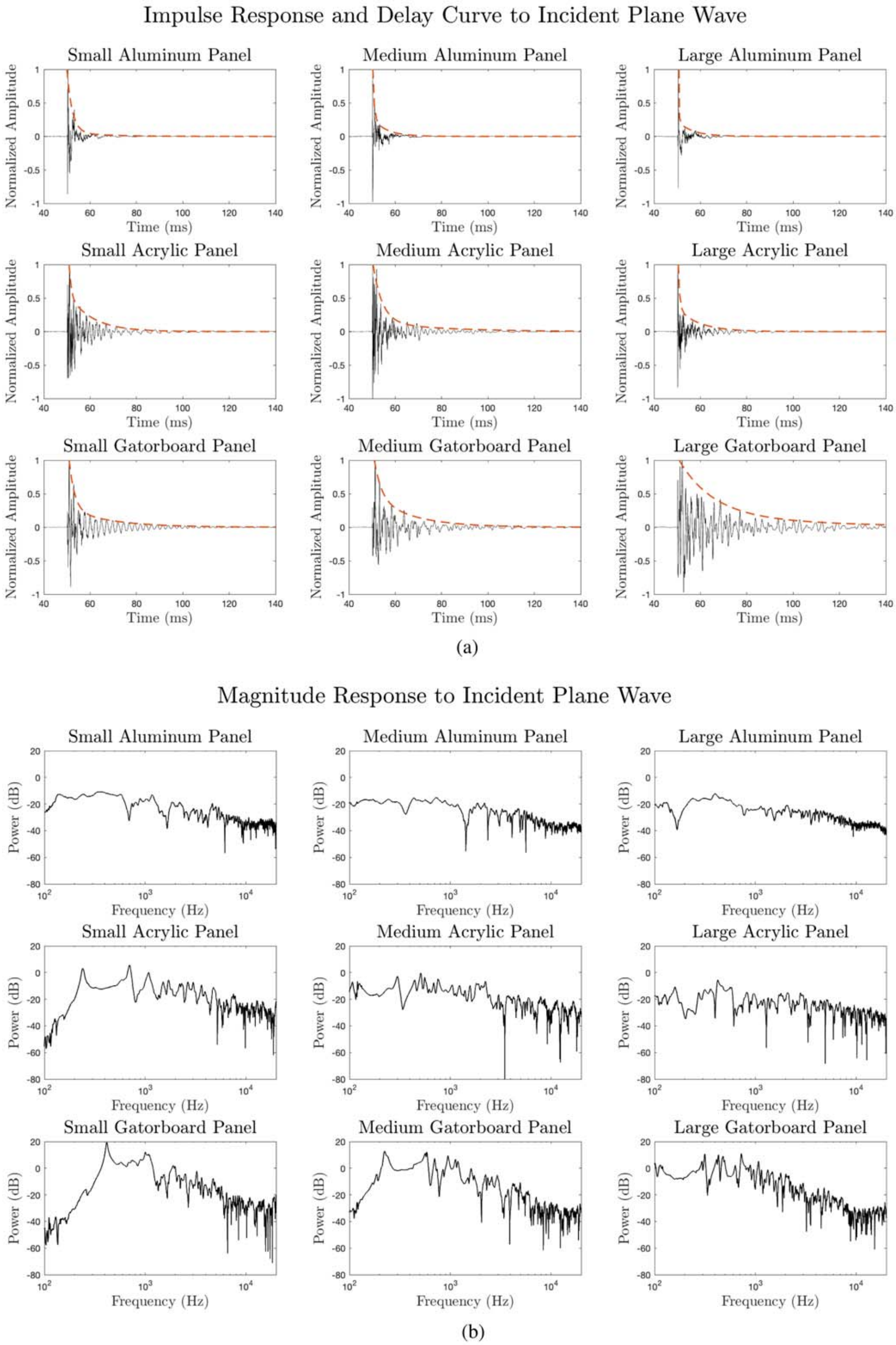
The frequency response at a particular sensor location may be derived by evaluating Eq. (3) at location (x_i, y_i) on the surface of the panel. The quality factor for isotropic plates is shown by Fahy and Gardonio to be inversely proportional to the material's damping coefficient [25].

The effective damping varies for each of the panel's bending modes; however, an average value for the damping of the panel can be expressed as

$$R_m = \frac{2\rho h \ln(2)}{t_{1/2}}, \tag{5}$$

where $t_{1/2}$ is the decay time for the impulse response of the panel to reach one-half amplitude. Impulse responses fit with exponential decay envelopes are shown in Fig. 1(a). $t_{1/2}$ can be extracted for each response in which the decay envelope reaches its half-amplitude and used in Eq. (5) to calculate average R_m.

In this work, three sizes of panels are tested: small panels with $L_x = 0.18$ m and $L_y = 0.23$ m, medium panels with $L_x = 0.26$ m and $L_y = 0.36$ m, and large panels with $L_x = 0.41$ m and $L_y = 0.51$ m. Three different panel materials are also tested: a 1-mm–thick aluminum material with an inner layer of viscoelastic adhesive to increase its damping, 2-mm–thick acrylic material, and 3-mm polystyrene-based foam board material called Gatorboard. This range of materials and sizes was chosen to show the effect of increasing damping with panel sizes comparable to existing smart devices of various classes. Properties of the materials used in this study are summarized in Table 1.



(a)

(b)

Fig. 1. The impulse responses for each panel used in this experiment are fitted with decay curves and plotted in (a). Because of the exponential nature of the decay, $\tau_{1/2}$ can be extracted from the curves and used to calculate R_m as shown in Eq. (5). The magnitude of the frequency responses of each panel are then plotted in (b). Generally, increasing panel damping yields a flatter magnitude response, whereas reducing panel damping introduces reverberant high-Q modes into the response.

The responses shown in Fig. 1(b) demonstrate that reducing material damping results in high-Q modes, which are detrimental to speech intelligibility because they can cause ringing, reverberation, and smearing of the audio signal.

1.2 Linearity of Panel Vibrations

It is shown by Fahy and Gardonio [25] that for flat plates, transverse deflection can produce non-linear vibrational behavior only if the deflection is significant. In this work, vibrations on the panel's surface will be induced by in-

Table 1. Properties of the materials used to construct the panels used in this experiment. Average R_m is calculated from Eq. (5) with $\tau_{1/2}$ extracted from the decay curves in Fig. 1(a).

Material	E (GPa)	v	ρ (kg/m^3)	h (mm)	Average R_m (Ns/m)
Aluminum	68.9	0.334	2,700	1	10,270
Acrylic	3.2	0.35	1,180	2	2,172
Gatorboard	1.5	0.35	222	3	241.5

cident plane waves and actuators on the panel's surface, which cause displacements on the order of tens of microns. This curvature is a fraction of the panel's minimum material thickness and dimensions, and it is well within linear vibrational limits.

Therefore, the displacement response of the panel at sensor location (x_i, y_i) can be modeled to induced vibrations from incident plane waves and actuators:

$$z_{(x_i,y_i)}[n] = s[n] \circledast h_1[n] + x[n] \circledast h_2[n], \qquad (6)$$

where $z_{(x_i,y_i)}[n]$ is the panel's displacement at position (x_i, y_i) at sample n, $s[n]$ is a signal being played by a source in the acoustic half-space in front of the panel, h_1 is the transfer function from the source's location to the panel's sensor, $x[n]$ is the signal being played by the affixed actuators, and h_2 is the transfer function from the actuator's location to the panel's sensor.

1.3 Duplex Mode Cancellation

Because $x[n]$ is directly coupled to the panel's surface and $s[n]$ is inefficiently coupled to the panel's surface, it may be necessary to remove non-zero $x[n]$. However, because $x[n]$ is directly known by the audio system and $h_2[n]$ can be determined for drivers at fixed panel locations, approaches such as spectral or time-domain subtraction, source separation, and artificial neural networks may be used to obtain an estimate of $s[n]\circledast h_1[n]$ from the structural sensor's audio stream.

In the following sections, the recorded speech signal is given as the convolution $s[n]\circledast h_1[n]$. The authors show that the audio-degrading effect of $h_1[n]$ creates only negligible effects on the ability of the recorded speech to be used with modern ASR systems.

2 DESCRIPTION OF EXPERIMENTS

2.1 Intelligibility and Transcription Experiment

The first experiment determined the accuracy with which an ASR system can transcribe audio from humans speaking near a panel when recorded by the structural sensors affixed to the panel. From Eq. (3), audio recorded using structural sensors on a panel will be subject to reverberation from high-Q modes, a challenge that traditional microphones or arrays do not face. A corpus of 500 sentences of speech recorded by structural sensors on the experiment panels were transcribed to compare the accuracy with those made with a reference microphone.

2.1.1 Impulse Response Acquisition

Each of the nine panels were placed in a semi-anechoic environment and equipped in their center with a PCB Piezotronics U352C66 accelerometer. A KEF LS50 loudspeaker was placed on-axis and half a meter away from the panel's surface, simulating a human speaking at this distance. The impulse response from the KEF to the panel was recorded using maximum length sequences to obtain an effective transfer function for this use case. Because the deflection of the panel from induced vibrations from the incident waves is well within the linear region of the panel, convolution with the panel's impulse response can be used to simulate how the panel's sensor records the panel's vibration induced under these conditions. This allows efficient testing with a large data set of speech samples. A similar measurement was taken for a calibrated PCB Piezotronics F130F20 free-field microphone used as a reference.

2.1.2 Testing Data

Recordings of Harvard Sentences were used to test how accurately an ASR system can transcribe audio recorded on a panel equipped with a structural sensor, in accordance with the "IEEE Recommended Practice for Speech Quality Measurements" [26]. It is possible that the harmonics of certain speech sounds optimally excite a set of the panel's resonant modes and negatively impact the intelligibility of words containing those sounds. The use of phonetically balanced Harvard Sentences ensures that the potential issues caused by these sounds are encapsulated in the experimental results. A male subject with diction training recorded 500 Harvard Sentences listed in [26]. These recordings could then be recorded by the panel using either the KEF loudspeaker or simulated as such using convolution. The recordings were scaled such to achieve an average of 71-dB sound pressure level at the panel's location, representative of speech at the half-meter test distance. [27].

2.1.3 Evaluation Metrics

The speech transmission index (STI) was used as an objective metric. The STI was proposed by Houtgast and Steeneken to evaluate the intelligibility of speech through transmission channels using the system's modulation transfer function [28]. In this experiment, playing audio through a reference monitor into a semi-anechoic room and inducing vibrations on the surface of the panel serves as a channel, whereby the only reverberation or degradation of the audio signal should occur via the panel's resonances. Schroeder proposed a method for calculating a modulation transfer function from the system's impulse response, enabling indirect calculation of the STI [29]. STI values greater than 0.75 are generally regarded as excellent in quality.

Word error rate (WER) was used to directly compute the accuracy of the transcriptions returned by the ASR system. WER is a measure of Levenshtein distance, describing the rate at which errors occur when comparing a transcription to the known text. Errors include the erroneous insertion of words, deletion of words, or substitution of a correct word with an incorrect word. WER is given as a percentage by

$$\mathrm{WER} = \frac{\text{Insertions + Deletions + Substitutions}}{\text{Number of Words in Reference}} \times 100\%. \tag{7}$$

The impulse responses obtained as described in SEC. 2.1.1 were used to obtain STI scores for each of the panels in the semi-anechoic environment. The Harvard Sentence recordings described in SEC. 2.1.2 were convolved with these impulse responses to simulate a large data set of audio recorded from the structural sensors affixed to the experiment panels. These recordings were transcribed to text via IBM Watson's speech-to-text ASR service and were assigned a WER score when compared to true Harvard Sentence transcripts. WER is also reported for recordings made by the reference microphone to determine the error the ASR system introduces when transcribing the speech corpus under ideal conditions. Therefore, the WER reported for the panel microphones should be evaluated by the incremental increase in WER in comparison to the reference case. Results are discussed in SEC. 3.

2.2 Actuator Signal Cancellation Experiment

When an acoustically active surface is used to simultaneously record and reproduce audio, the signal recorded by the affixed structural sensors will contain a mixture of vibration induced by both the affixed actuators and the user's speech. The second experiment explores the use of signal processing to digitally remove the signal played by actuators from the audio stream. In many smart devices, interrupting a song that is playing or stopping an answer that a smart assistant provides is vital to the device's audio-based human-computer interaction.

Vibrations from affixed actuators more efficiently drive the panel's surface than induced vibration from incident plane waves. Therefore, the sensor will observe a larger contribution from the actuators than from the incident sound waves even if both signals contain the same power. This problem of simultaneous playback and recording also affects existing smart audio devices, although on these devices, microphones record both signals as acoustic pressure variations in air; therefore, neither has a coupling advantage to the microphone. A subtraction approach was used to show the feasibility of cancelling the vibrations due to the panel's audio stream.

2.2.1 Experimental Setup

In this experiment, simulation using impulse responses was not possible because cancelling a simulated vibrational contribution would be trivial or require assumptions about environmental and system noise. Instead, a KEF LS50 loudspeaker was placed in the far-field of a panel that is equipped with a structural sensor and an actuator. Harvard Sentence recordings were played via the KEF loudspeaker while audio was simultaneously being played through the actuators. The actuators played three different types of audio: white noise, classical music, and synthesized speech (such as from Amazon's Alexa assistant). This tested the effectiveness of the proposed method on wide-band signals, music, and speech.

2.2.2 Subtraction Method

Because the panel is operating in a linear deflection region, subtraction approaches can theoretically be used to directly cancel the audio from actuators, provided the transfer function from the actuator to the sensor, h_2 in Eq. (6), is known. Generally, this transfer function can be obtained at the time of device assembly because the actuator will never move once it is affixed to the panel's surface. The signal being played by the actuators, $x[n]$ in Eq. (6), is known because it is determined by device's the audio reproduction system. Therefore, Eq. (6) may be rewritten to isolate the unknown signal contribution from incident acoustic waves as

$$z_{(x_i, y_i)}[n] - x[n] \circledast h_2[n] = s[n] \circledast h_1[n]. \tag{8}$$

Transfer function h_2 in Eq. (8) may be understood as a delay in sequence with a finite impulse response (FIR) filter such that

$$z_{(x_i, y_i)}[n] - x[n] \circledast h_2'[n - p_d] = s[n] \circledast h_1[n], \tag{9}$$

where $h_2'[n]$ contains the harmonic information from h_2 as an FIR filter with non-zero first tap and p_d represents the total delay from the time a sample is played via the actuator to when its response is recorded including propagation delay on the panel's surface and any hardware delays. A precise value for p_d is important if subtraction is to be done in discrete time, although spectral subtraction on a frame level may reduce sensitivity to slight drifting of the true value of p_d. In this experiment, cross correlation was used to determine a value for p_d for each panel. The audio stream recorded by the structural sensor on each panel was then stored in a buffer, whereas the expected contribution from the actuator signal was calculated and ultimately subtracted in either the sample or frequency domains using Eq. (9).

2.2.3 Evaluation Metrics

The signals recorded by the sensor contained contributions from both the actuators driving the panel in its function as a loudspeaker and the speech source that the authors are attempting to capture and transcribe. The contribution from panel as a loudspeaker is treated as a source of interference. Instead, the desired signal is the contribution from

Table 2. Average speech transmission index (STI) and word error rate (WER) scores for the each of the panel materials, and the standard deviation σ among small, medium, and large panel sizes. Higher damping is shown to improve both STI and WER score.

Material	Average STI Score	σ STI	Average WER (%)	σ WER (%)
Aluminum	0.983	0.007	10.1	0.47
Acrylic	0.922	0.01	11.5	0.42
Gatorboard	0.908	0.03	13.6	0.44
Reference Microphone	0.980		9.33	

the speech source in isolation. A signal-to-interference ratio (SIR) can therefore be reported as

$$\text{SIR (dB)} = 10 \log_{10} \left(\frac{P_s}{P_x} \right), \tag{10}$$

where P_s and P_x are the power of the signals in the recording from the incident acoustic waves and the induced panel vibrations from the loudspeaker actuators, respectively. The increase in SIR after the cancellation of the actuators' contribution describes the achievable amount of suppression and is reported in SEC. 3.

3 RESULTS

3.1 Intelligibility and Transcription Accuracy

The average STI and WER scores and their standard deviations σ for each panel material are tabulated in Table 2. For both metrics, the standard deviation is a small fraction of the overall scores, implying that panel size caused only a small effect on the results for the sizes tested. However, the panel's material did appear to cause a noticeable impact on the results. STI increases and WER decreases as the panel's damping increases. This result follows from the flattening of the frequency response and the reduction of reverberant high-Q modes as damping increases, shown in Figs. 1(a) and 1(b). In general, every material's STI average is above 0.9, meaning that any material used in this study captured excellent quality recordings according to the standard.

For the WER metric, no panel exceeded the WER reported when using the reference microphone by more than 4.5% even withstanding added reverberation in the lesser-damped panel materials. This experiment shows that the audio recorded through structural sensors affixed to panels is able to be transcribed with modern ASR systems without significant reduction of accuracy.

3.2 Cancellation

The spectrogram of acoustic waves containing a passage of speech recorded by the medium Gatorboard panel with no contribution from the affixed actuator is shown in Fig. 2. When the panel records signals that contain contributions from both incident waves and the affixed actuator, the spectrogram shown in Fig. 2 becomes the target spectrogram when applying the cancellation algorithm. Spectrograms showing this dialog snippet in a mixture with the white noise, classical music, and synthesized speech being played by the actuators are shown in Figs. 3–5. Quantitative results regarding post-cancellation SIR improvement among all panels are tabulated in Table 3.

In general, the subtraction has a large impact on the SIR of the audio stream. SIR increased an average of 51.1 dB among aluminum panels, 40.1 dB among acrylic panels, and 39.3 dB among gatorboard panels. This shows a similar trend to the WER metric results from Table 2, in that more highly damped panels show better reliably in cancelling the actuator's contribution to the audio stream. All SIR improvements reported in Table 3 show the feasibility of removing the highly-coupled actuator contribution to the audio stream.

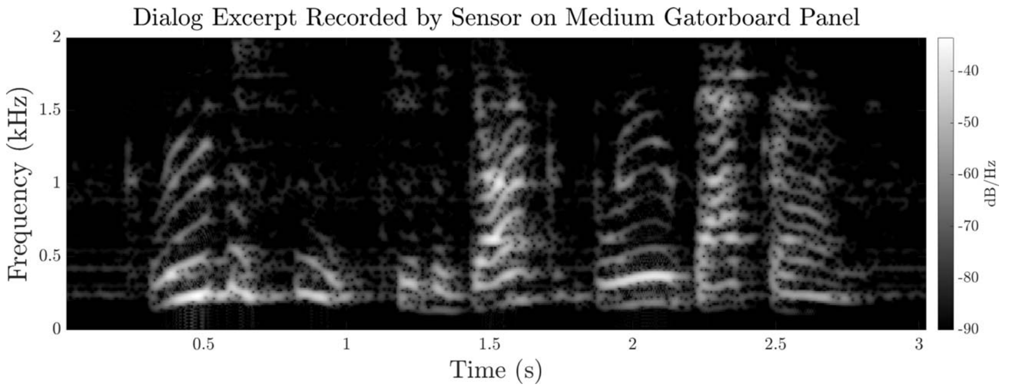


Fig. 2. Spectrogram of dialog snippet in isolation recorded by the medium Gatorboard panel, which is the target for the resulting post-cancellation spectrograms from cancelling the different types of actuator signal. Cancellation results for the medium Gatorboard panel are shown in Figs. 3–5. The word error rate (WER) when transcribing this snippet with IBM Watson is 0.00%.

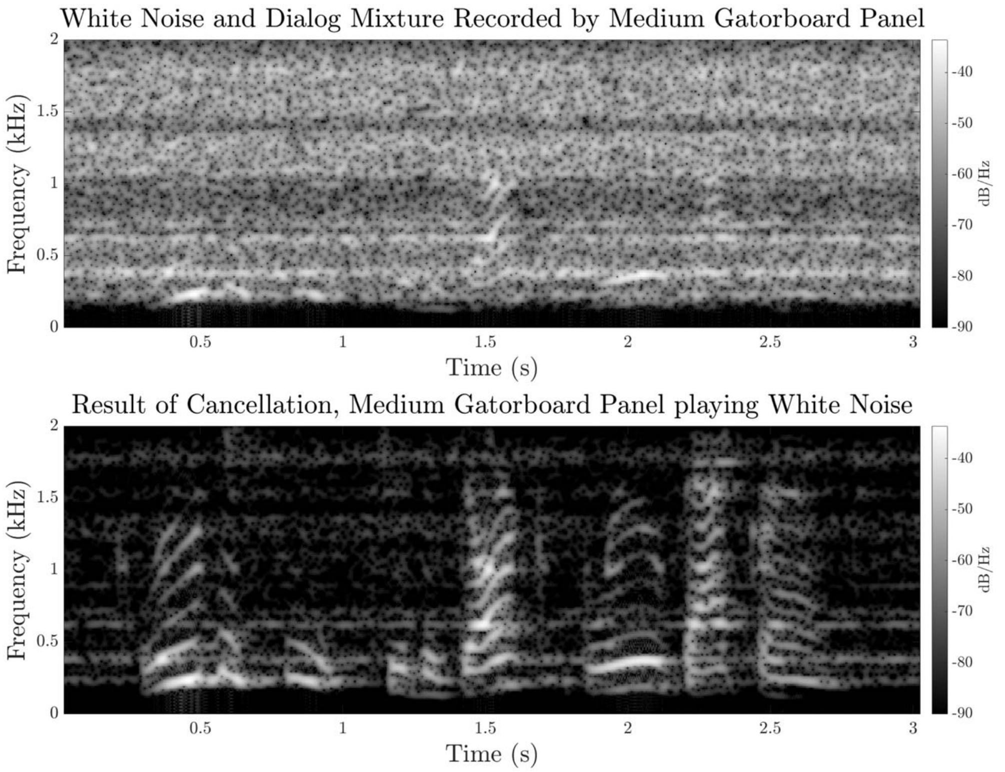


Fig. 3. Spectrogram of recorded dialog while white noise was played by actuators before and after cancellation. Before cancellation, a word error rate (WER) of 100% is reported, whereas a WER of 0.00% is reported after cancellation.

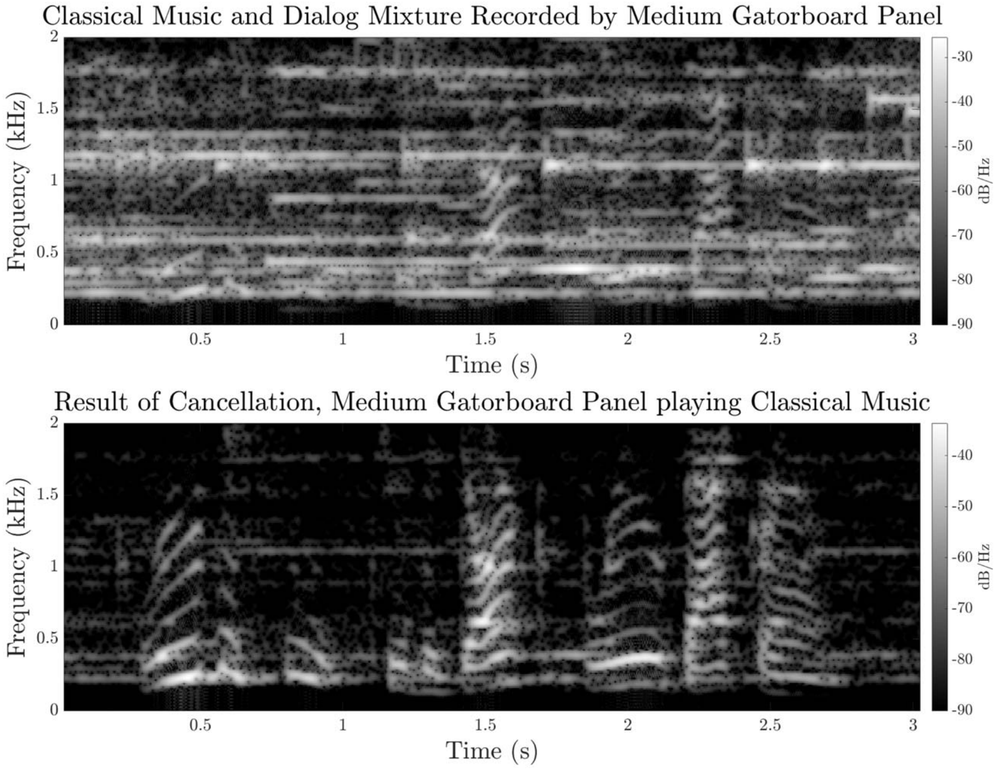


Fig. 4. Spectrogram of recorded dialog while classical music was played by actuators before and after cancellation. Before cancellation, a word error rate (WER) of 100% is reported, whereas a WER of 0.00% is reported after cancellation.

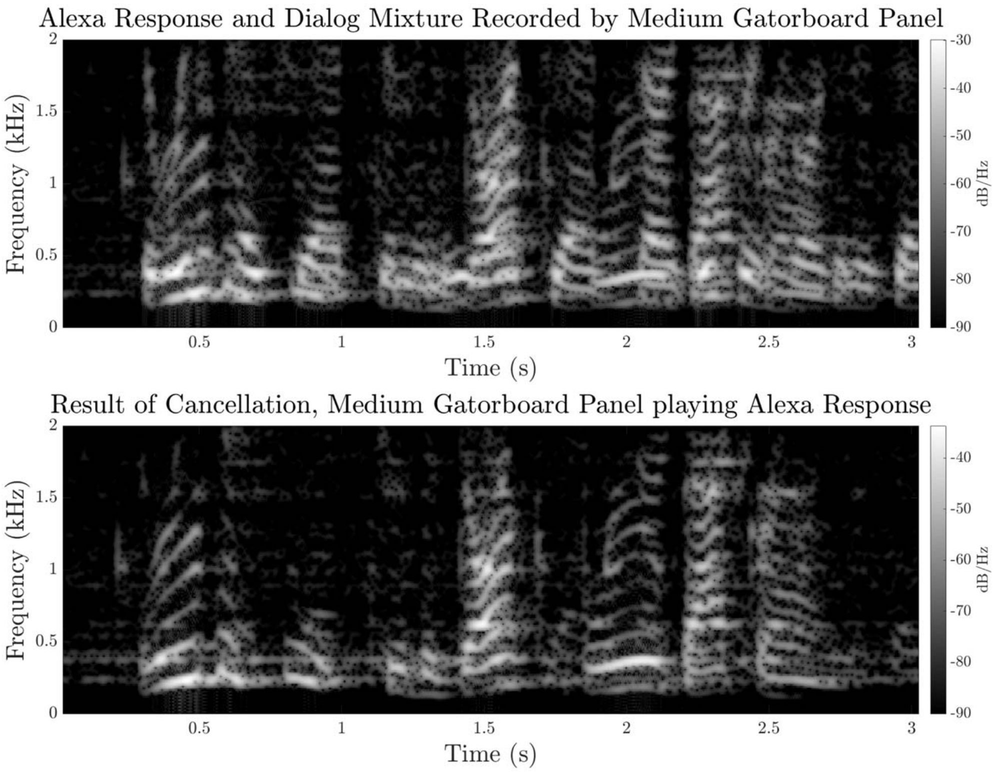


Fig. 5. Spectrogram of recorded dialog while a synthesized speech passage was played by actuators before and after cancellation. Before cancellation, a word error rate (WER) of 122% is reported, whereas a WER of 0.00% is reported after cancellation.

4 CONCLUSION

It is known that using structural sensors affixed to a panel's surface will introduce reverberation and degrade the quality of the recorded signal when compared to a reference microphone. However, the purpose of input audio streams in smart devices is often for transcription by ASR systems so that devices can make human-computer–interaction decisions and not for studio-quality recordings. As such, the results in Table 2 support the use of surface audio systems for smart displays and devices, because recordings taken from structural sensors on the surface of the panel yielded only a negligible reduction in transcription accuracy when transcribed by an ASR system.

Surface audio systems must be able to address the problem of crosstalk. When a surface audio system is simultaneously playing audio via its actuators and recording incident acoustic waves, the resulting audio stream will contain contributions from both. Unlike smart devices with traditional microphones, the actuators on the surface of the panel are strongly coupled to the panel's surface and therefore are able to induce vibration on the panel's surface more efficiently than incident acoustic waves. Experimental measures show significant SIR improvement when subtracting

Table 3. SIR improvement in the sensor's audio stream after cancelling the contribution from the actuators for each panel. Significant SIR improvements are seen for all three material types.

Panel Type	SIR Increase (dB) White Noise	SIR Increase (dB) Classical Music	SIR Increase (dB) Alexa Speech	SIR Increase (dB) Material Average
Large Aluminum	55.3	64.1	55.5	51.1
Medium Aluminum	53.4	53.0	44.1	
Small Aluminum	39.4	48.9	46.3	
Large Acrylic	67.6	57.2	66.6	40.1
Medium Acrylic	33.2	28.0	23.2	
Small Acrylic	29.3	30.6	25.2	
Large Gatorboard	53.8	43.2	40.9	39.3
Medium Gatorboard	29.9	31.4	25.0	
Small Gatorboard	42.3	41.7	45.3	

the actuator's contribution to the audio stream, with an average increase of over 50 dB for heavily damped panels.

The apparent correlation between the panel's damping, R_m, and the resulting STI, WER, and SIR improvement scores for audio captured by its affixed structural sensor introduces an important trade-off. In general, flat panel loudspeakers that are made out of more compliant materials can move more air with less energy, thus making them louder and more efficient as speakers. However, it would appear that the physical parameters of a panel that make it work well as a speaker would result in a lower-fidelity microphone. However, even a very compliant material such as Gatorboard only experienced a 3% reduction in transcription accuracy when compared to aluminum panels, well within the margin of error, despite its damping constant being roughly forty times smaller. More robust methods for cancelling the actuator's contribution to the recorded audio stream, such as artificial neural networks, are left to future work. Though further exploration into this trade-off is needed, experimental results support that even compliant materials are viable for duplex surface audio systems.

The combination of transcription accuracy and crosstalk cancellation implies that surface audio systems are a viable alternative to the audio systems on modern smart devices. Applications of signal processing may give surface audio systems advantages over MEMS microphones arrays on current smart devices. Sensors may be placed more than the standard 1–4 cm in smart speakers on the market. Increasing the distance between sensors can potentially improve the precision of source localization and beamforming tasks. Fuller [13] also explains that different angles of incidence to a panel correspond to unique vibration responses on the panel's surface, which can be leveraged by machine learning algorithms to estimate the direction of arrival of an incident acoustic wave.

Using surface vibrations to perform beamforming, source localization, direction-of-arrival estimation, and other common tasks performed by smart devices will be an important next step in understanding the viability of using surface audio systems in smart devices. The methods described in this paper were evaluated under test conditions typical of conversational speech sound levels with high signal-to-noise plus interference ratios (greater than 30 dB in these tests). However, these methods may also be employed following the use of speech enhancement methods employed for more challenging speech capture scenarios that may require background suppression with directional microphone arrays or other noise suppression methods.

The results from this study demonstrate the viability of surface audio systems for modern smart devices. Using these systems can help smart device manufacturers make devices more durable, waterproof, and efficient and give them tools for designing better-sounding display devices without changing their form factor.

5 ACKNOWLEDGMENT

This work was supported by NSF Award 2104758.

6 REFERENCES

[1] Strategy Analytics, "Strategy Analytics: The Smart Speaker Market's Recovery is in Full Swing as Shipments in 2Q21 Surged to Record Levels," https://news.strategyanalytics.com/press-releases/press-release-details/2021/Strategy-Analytics-The-Smart-Speaker-Markets-Recovery-is-in-Full-Swing-as-Shipments-in-2Q21-Surged-to-Record-Levels/default.aspx (2021 Sep.).

[2] B. Kinsella, "Streaming Music, Questions, Weather, Timers and Alarms Remain Smart Speaker Killer Apps, Third-party Voice App Usage not Growing," https://voicebot.ai/2020/05/03/streaming-music-questions-weather-timers-and-alarms-remain-smart-speaker-killer-apps-third-party-voice-app-usage-not-growing/ (2020 May).

[3] Y. Choi, C. Oh, K. Park, and S. Lee, "Organic Light Emitting Display Device Including a Sound Generating Apparatus," US Patent 10,847,585 (2020 Nov.).

[4] T.-H. Kim and G.-C. Park, "Display Device," US Patent 2019/0163234 A1 (2019 May).

[5] S. Lee, K. Park, Y. Choi, K. Kim, and M. Bae, "Actuator Fixing Device and Panel Vibration Type Sound-Generating Display Device Including the Same," US Patent 10,412,500 B2 (2019 Jul.).

[6] M. C. Heilemann, D. A. Anderson, S. Roessner, and M. 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 (2021 Jan.). https://doi.org/10.17743/jaes.2020.0057.

[7] G. Bank and N. Harris, "The Distributed Mode Loudspeaker-Theory and Practice," in *Proceedings of the AES UK 13th Conference: Microphones & Loudspeakers* (1998 Mar.), paper MAL-18.

[8] S. Lee, K. Park, K. Jang, and C. Oh, "16-3: Study on Enhancement of the Sound Quality by Improvement of Panel Vibration in OLED TV," *SID Symp. Dig. Tech. Papers*, vol. 49, no. 1, pp. 185–187 (2018 May). https://doi.org/10.1002/sdtp.12515.

[9] S. Roessner, M. Heilemann, and M. F. Bocko, "Evaluating Listener Preference of Flat-Panel Loudspeakers," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct), paper 10230. http://www.aes.org/e-lib/browse.cfm?elib=20603.

[10] R. Haeb-Umbach, S. Watanabe, T. Nakatani, et al., "Speech Processing for Digital Home Assistants: Combining Signal Processing With Deep-Learning Techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124 (2019 Nov.). https://doi.org/10.1109/MSP.2019.2918706.

[11] K. Kumatani, J. McDonough, and B. Raj, "Microphone Array Processing for Distant Speech Recognition: From Close-Talking Microphones to Far-Field Sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140 (2012 Nov.). https://doi.org/10.1109/MSP.2012.2205285.

[12] L. Meirovitch, *Dynamics and Control of Structures* (Wiley, New York, NY, 1990).

[13] S. J. Elliott, C. R. Fuller and P. A. Nelson, *Active Control of Vibration* (Academic Press, London, UK, 1996).

[14] S. Rubenstein, W. Saunders, G. K. Ellis, H. Robertshaw, and W. Baumann, "Demonstration of a LQG Vibration Controller for a Simply-Supported Plate," in Proceedings of the Conference on Recent Advances in Active Control of Sound and Vibration, pp. 15–17 (Blacksburg, VA) (1991 Apr.).

[15] L. D. Lafleur, F. D. Shields, and J. E. Hendrix, "Acoustically Active Surfaces Using Piezorubber," *J. Acoust. Soc. Am.*, vol. 90, no. 3, pp. 1230–1237 (1991 Sep.). https://doi.org/10.1121/1.402384.

[16] F. D. Shields and L. D. Lafleur, "Smart Acoustically Active Surfaces," *J. Acoust. Soc. Am.*, vol. 102, no. 3, pp. 1559–1566 (1997 Sep.). https://doi.org/10.1121/1.420102.

[17] J. J. Gamboa-Montero, F. Alonso-Martin, J. C. Castillo, M. Malfaz, and M. A. Salichs, "Detecting, Locating and Recognising Human Touches in Social Robots With Contact Microphones," *Eng. Appl. Artificial Intelligence*, vol. 92, paper 103670 (2020 Jun.). https://doi.org/10.1016/j.engappai.2020.103670.

[18] S. Kita and Y. Kajikawa, "Fundamental Study on Sound Source Localization Inside a Structure Using a Deep Neural Network and Computer-Aided Engineering," *J. Sound Vibr.*, vol. 513, paper 116400 (2021 Nov.). https://doi.org/10.1016/j.jsv.2021.116400.

[19] P. Ladefoged and K. Johnson, *A Course in Phonetics* (Wadsworth Cengage Learning, Boston, MA, 2011).

[20] K. Arcas and A. Chaigne, "On the Quality of Plate Reverberation," *Appl. Acoust.*, vol. 71, no. 2, pp. 147–156 (2010 Feb.). https://doi.org/10.1016/j.apacoust.2009.07.013.

[21] S. A. Gelfand, "Room Reverberation Effects on Recognition of Some Consonant Features," *J. Acoust. Soc. Am.*, vol. 62, no. S1, pp. S79–S80 (1977 Dec.). https://doi.org/10.1121/1.2016388.

[22] D. Anderson and M. F. Bocko, "Measuring Speech Intelligibility Loss in Single-Driver Panel Loudspeakers," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), e-Brief 228. http://www.aes.org/e-lib/browse.cfm?elib=17904.

[23] L. Cremer, M. Heckl, and B. A. T. Petersson, *Structure-Borne Sound: Structural Vibrations and Sound Radiation at Audio Frequencies* (Springer, Berlin, Germany, 2005).

[24] A. K. Mitchell and C. R. Hazell, "A Simple Frequency Formula for Clamped Rectangular Plates," *J. Sound Vibr.*, vol. 118, no. 2, pp. 271–281 (1987 Oct.). https://doi.org/10.1016/0022-460X(87)90525-6.

[25] F. J. Fahy and P. Gardonio, *Sound and Structural Vibration: Radiation, Transmission and Response* (Academic Press, Cambridge, MA, 2007), 2nd ed.

[26] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246 (1969 Sep.). https://doi.org/10.1109/TAU.1969.1162058.

[27] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119 (1947 Jan.).

[28] T. Houtgast and H. J. Steeneken, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," *J. Acoust. Soc. Am.*, vol. 54, no. 2, p. 557 (1973 Aug.). https://doi.org/10.1121/1.1913632.

[29] M. R. Schroeder, "Modulation Transfer Functions: Definition and Measurement," *Acta Acust. united Acust.*, vol. 49, no. 3, pp. 179–182 (1981 Nov.).

THE AUTHORS

Tre DiPassio

Michael C. Heilemann

Mark F. Bocko

Tre DiPassio is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Rochester. He received his B.S. and M.S. in Electrical Engineering focused on image and speech processing from the Rochester Institute of Technology in 2018. He received his M.S. in Electrical and Computer Engineering focused on audio signal processing from the University of Rochester in 2019. He won the Acoustical Society's International Student Challenge in Acoustic Signal Processing in 2019. His research interests include vibro-acoustics, smart audio devices, and audio signal processing.

•

Michael C. Heilemann received his M.S. and Ph.D. in Electrical Engineering from the University of Rochester (UR) in 2015 and 2018, respectively. In 2017, he was named the Harman Scholar by the Audio Engineering Society Educational Foundation. After completing his doctoral studies, he joined the faculty at the UR in the Department of Electrical and Computer Engineering. His research interests are in the areas of electroacoustics and spatial audio. Professor Heilemann advises the capstone projects for students majoring in Audio and Music Engineering and teaches courses on signal processing and acoustics.

•

Mark F. Bocko earned his Ph.D. in Physics from the University of Rochester in 1984. After a brief post-doctoral appointment, he joined the Rochester Electrical and Computer Engineering (ECE) Department in 1985. His research has spanned multiple areas with its current focus on audio and acoustic signal processing. He is also the Director of the Center for Emerging and Innovative Sciences (CEIS), a New York State Office of Science, Technology and Academic Research (NYSTAR)–supported Center for Advanced Technology at the University of Rochester. Professor Bocko has taught courses on solid state devices, microwaves, circuits and systems, audio signal processing, and acoustics. He has won five teaching awards at the University of Rochester and was named the Mercer Brugler Distinguished Teaching Professor at the University from 2008 to 2011. He was named Distinguished Professor of Electrical and Computer Engineering in 2013 and served as Chair of the ECE Department from 2004 to 2010 and again from 2012 to 2020.