



Evaluation of Software and Hardware Settings for Audio-Based Analysis of Construction Operations

Chieh-Feng Cheng¹ · Abbas Rashidi² · Mark A. Davenport¹ · David V. Anderson¹

Received: 20 September 2018 / Revised: 10 February 2019 / Accepted: 18 February 2019
© Iran University of Science and Technology 2019

Abstract

Various activities of construction equipment are associated with distinctive sound patterns (e.g., excavating soil, breaking rocks, etc.). Considering this fact, it is possible to extract useful information about construction operations by recording the audio at a jobsite and then processing this data to determine what activities are being performed. Audio-based analysis of construction operations mainly depends on specific hardware and software settings to achieve satisfactory performance. This paper explores the impacts of these settings on the ultimate performance on the task of interest. To achieve this goal, an audio-based system has been developed to recognize the routine sounds of construction machinery. The next step evaluates three types of microphones (off-the-shelf, contact, and a multichannel microphone array) and two installation settings (microphones placed in machines' cabin and installed on the jobsite in relatively proximity to the machines). Two different jobsite conditions have been considered: (1) jobsites with single machines and (2) jobsites with multiple machines operating simultaneously. In terms of software settings, two different SVM classifiers (RBF and linear kernels) and two common frequency feature extraction techniques (STFT and CWT) were selected. Experimental data from several jobsites was gathered and the results depict an accuracy over 85% for the proposed audio-based recognition system. To better illustrate the practical value of the proposed system, a case study for calculating productivity rates of a sample piece of equipment is presented at the end.

Keywords Microphone arrays · Audio signal processing · Construction equipment · Activity recognition · Fourier transform

1 Introduction

Recognizing the activities of construction heavy equipment is a major step toward productivity analysis of a jobsite. Aside from productivity monitoring, recognizing the activities of construction machinery provides useful information for various other purposes including scheduling and cost estimation [1] jobsite layout analysis and decision making [2] and monitoring and optimization of fuel consumption [3]. Some manufacturers have started to integrate their proprietary monitoring devices with new equipment, but construction contractors are unlikely to have a fleet composed

entirely of new and single-manufacturer equipment for this to represent an immediate solution. The common alternative approaches for heavy equipment activity recognition require active and/or passive external sensors [4–14]. Active sensors include GPS and micro-electro-mechanical systems (MEMS) and devices such as accelerometers and gyroscopes, while passive sensors include image/video processing using computer vision algorithms [15].

During the past few years, another category of input data, audio, has been investigated by researchers for recognizing various activities of construction heavy equipment and machines. The idea is that construction heavy equipment generates distinct sound patterns while performing different activities that can provide useful information when properly recorded and processed [16–19]. The idea of analyzing sounds generated by various engineering systems and devices has appeared in several research and application areas such as speech recognition, audio-based navigation of robots, the use of sonar in the exploration and mapping,

✉ Abbas Rashidi
abbas.rashidi@utah.edu

¹ Georgia Institute of Technology, Atlanta, GA, USA

² Department of Civil and Environmental Engineering,
University of Utah, 110 Central Campus Drive, Room 2022,
Salt Lake City, UT 84112, USA

ultrasonic signal processing for condition-based maintenance (CBM) in manufacturing workplaces [20, 21], and audio signal processing for feature recognition in medical devices. Each of these applications requires their own hardware and software settings (types of microphone and other acoustic recording devices, layout of microphones, number and distance between the sound capturing devices, etc.)

Despite the widespread use of audio signal processing techniques for analysis and modeling of various engineering techniques, the application of these methods in construction management area is still in early stages of development. This paper summarizes recent studies conducted by the authors evaluating necessary hardware devices and computing techniques, mostly suitable for capturing and processing audio files in large-scale, noisy, and cluttered construction jobsites.

2 Literature Review: Acoustical Modeling of Engineering Systems

Bengtsson et al. [20] define audio-based feature identification as a three-module process (Fig. 1). First, the sound is recorded into a computer as an input to the processing module. The processing module consists of two steps: pre-processing to remove unwanted noise and to extract a period of interest and feature extraction to identify characteristic features of the audio sample and create a feature vector. Once the feature vector is created, the condition monitoring and diagnosis module consists of comparing the audio sample to an existing library and providing a condition diagnosis. If the software is in the learning mode, newly identified vectors are stored into the case library. Although this model describes ultrasound-based CBM systems, it is characteristic to most audio-based machine learning practices and can be applied to other fields.

Some typical applications of ultrasound-based CBM include: bearing inspection; testing gears/gearboxes; pumps; motors; steam trap inspection; valve testing; detection/trending of cavitation; compressor valve analysis; leak detection in pressure and vacuum systems such as boilers, heat exchangers, condensers, chillers, tanks, pipes, hatches, hydraulic systems, compressed air audits, specialty gas

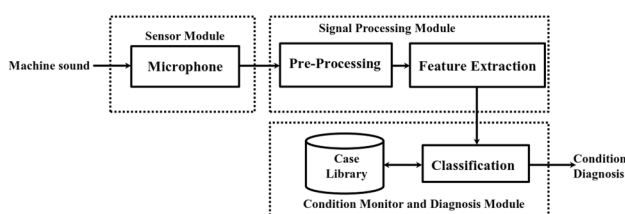


Fig. 1 Schematic for computer-based feature identification [20]

systems and underground leaks; and testing for arcing and corona in electrical apparatus [22].

In the medical setting, similar technologies have been developed to monitor patients with chronic diseases such as asthma and chronic obstructive pulmonary disease (COPD). Specifically, to evaluate the adherence to inhaler medication, which involves doses being taken in a consistent schedule and with a proper method. The inhaler compliance assessment (INCA) device was developed as an approach to evaluate inhaler adherence. This device is attached to a widely used variety of inhalers and when the patient opens the mouthpiece to take a dose, the INCA device takes an audio recording with a timestamp. One month of typical inhaler use yields 60 audio files corresponding to 60 doses of medication. Analyzing this data set takes an experienced pulmonary clinician an average of 30 min. This type of labor-intensive analysis is not feasible with a large group of patients. Thus, a computer algorithm has been designed to analyze audio data sets and provide an adherence score based on dose schedule and the pattern of blister, exhalation, and inhalation events [23]. This sort of study provides useful information for clinicians to understand why a specific inhaler is not effective and propose methods to enhance the patient's adherence [24].

3 Literature Review: Automated Recognition and Monitoring of Construction Operations

Construction industry lags behind manufacturing industries in terms of adapting new technologies and automated solutions. According to the statistics provided by the Lean Construction Institute, this is the primary reason that the productive time in the construction industry is only 43%, while it is 88% in manufacturing. For many years, the construction industry has suffered from the lack of real-time performance monitoring, holistic project management, labor efficiency, and waste prevention tools. Fortunately, this pattern has started to change in recent years. Nowadays, sensing technologies are vastly used to collect necessary data (e.g., energy consumptions, spatial data, location tracking of building elements, etc.) from construction jobsites. Automation solutions and robotic techniques are becoming more popular among construction industry experts and building information modeling (BIM) revolutionized the entire design and construction phases of construction projects.

Although the potential for applications is boundless, there has been little prior work studying the implementation of audio signal processing techniques specifically in the area of construction engineering and management. The motivation behind an audio-based activity recognition framework for construction operations lies in the inherent

limitations of the existing techniques for detailed assessment of the construction activities. These limitations have hindered the successful adoption of these techniques and hence the industry still relies heavily on the experience of project managers on the jobsite to perform such assessments. In comparison with other active and passive activity recognition methods, working with audio signals provides the following advantages:

- Most active sensors (GPS, accelerometers, etc.) need to be directly mounted on the equipment. In addition, for each machine, at least one sensor is required. As explained in this work, this is not a limitation for an audio-based activity recognition method. Microphones can be installed in various locations at the jobsite and one microphone is usually able to cover the activities of many machines.
- Computer vision methods are very sensitive to environmental factors such as lighting conditions and occlusions. In addition, the limited field of view of cameras is another major drawback for computer vision methods (Fig. 2). Microphones and audio signals are more resilient against the above-mentioned limitations.
- The sound produced by each piece of equipment is independent of the operator. As the result, an operator can perform an activity in a variety of ways. An action recognition system based on location sensors and/or computer vision techniques would need to take these different scenarios into account, while an audio analysis always yields the same result.
- Compared to video data, audio files have lower data rates and are therefore computationally more efficient.

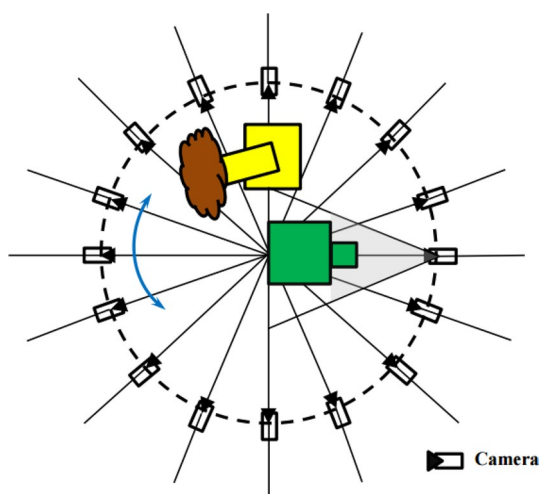


Fig. 2 Limited field of view of cameras is a major issue with using computer-based activity recognition algorithms at construction jobsites [16]

Choosing the optimal hardware setting (type and location of microphones) and software (selecting proper algorithms and tuning algorithmic parameters) is a crucial stage for acoustical modeling of real-world construction jobsites.

Using automated approaches such as the proposed audio-based activity detection and recognition system at construction jobsites would significantly reduce monitoring expenses and durations. In addition, it will serve as an assistant tool for construction managers and jobsite personnel. Automating various aspects of construction operations has another important impact: it would eliminate several labor-intensive and manual steps, which could eventually lead to involvement of people with limited physical activities and/or disabilities to be more engaged within the construction industry.

4 Hardware Setting: Different Types of Microphones

The primary devices used for audio recording are microphones. A microphone generally contains a diaphragm, either surface or moving, designed to capture electroacoustic waves and generate electronic signals. Typical microphone classifications involve distinguishing either by their pickup pattern or by the type of transducer they employ.

Per Ballou [25], the pickup pattern refers to how the microphone discriminates among different directions of the incoming sound. The most common microphone types are omnidirectional, bidirectional, and unidirectional or cardioid microphones. In an omnidirectional microphone, the pickup pattern is equal in all directions. Omnidirectional microphones are particularly useful when it is necessary to capture all audio sources in an environment. In a bidirectional microphone, the pickup pattern is equal in two opposite directions and negligible at 90° from these two directions. Bidirectional microphones are particularly useful when it is necessary to capture the conversations of two speakers positioned face to face. In a unidirectional microphone, the pickup pattern has a cardioid shape facing in just one direction. Of all three, unidirectional microphones are the most widely used, because they allow the user to focus on a specific source of interest. Pickup pattern is a key factor for acoustical modeling of construction jobsites, since in a multifaceted jobsite several pieces of construction equipment might work simultaneously, generating sound from multiple locations and directions. More details about various types of microphones are provided in Fig. 3 [25].

A transducer is a device that converts a physical stimulus to an electrical signal output. Transducers can be carbon, crystal, and ceramic microphones, condenser microphones, dynamic microphones, and electret condenser microphones. Microphone selection by the type of transducer is another key consideration in a construction jobsite. Transducers






Microphone	Omnidirectional	Bidirectional	Directional	Supercardioid	Hypercardioid
Directional response characteristics					
Voltage output	$E = E_o$	$E = E_o \cos \theta$	$E = \frac{E_o}{2}(1 + \cos \theta)$	$E = \frac{E_o}{2}[(\sqrt{3} - 1) + (3\sqrt{3}) \cos \theta]$	$E = \frac{E_o}{4}(1 + 3 \cos \theta)$
Random energy efficiency (%)	100	33	33	27	25
Front response	1	1	∞	3.8	2
Back response					
Front random response	0.5	0.5	0.67	0.93	0.87
Total random response					
Front random response	1	1	7	14	7
Back random response					
Equivalent distance	1	1.7	1.7	1.9	2
Pickup angle (2θ) for 3 dB attenuation	-	90°	130°	116°	100°
Pickup angle (2θ) for 6 dB attenuation	-	120°	180°	156°	140°

Fig. 3 Performance characteristics of several microphones [25]

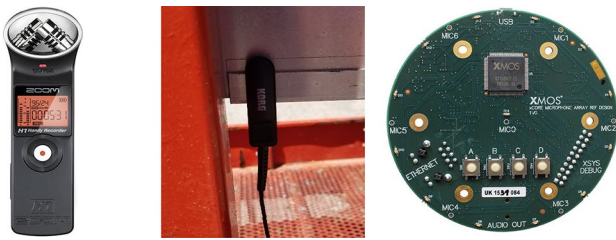


Fig. 4 Different types of microphones used in the project: Zoom H1 digital handy recorder (left) [28]; Korg CM-200 clip-on contact microphone attached to construction equipment (center); xCORE-200 microphone array evaluation board—top view—(right) [29]

must be robust enough to withstand inclement weather and other contingencies while maintaining stable audio signal recording characteristics.

For this research, three microphones of varying types have been selected (Fig. 4): (1) a Zoom H1 digital handy recorder, an off-the-shelf microphone; (2) a Korg CM-200 clip-on contact microphone; and (3) an xCore-200 multi-channel microphone array. Condenser microphones and a variant of these called the MEMS microphones are the type of microphones built into the Zoom H1 handy recorder and the XMOS microphone array, respectively. Per Yamaha [26], condenser microphones have good sensitivity to all frequencies, but are highly susceptible to structural vibration and humidity. Crystal and ceramic microphones are the type of microphones built into the Korg contact microphones [27]. Their name comes from the fact that these devices use piezoelectric crystals such as Rochelle salt, tourmaline, barium titanate, and quartz to convert pressure differences from the environment into a voltage output. Due to this pressure response characteristic, crystal microphones can be built to

pick up vibrations from contact with objects, as opposed to common microphones that pick up air-carried sound waves.

Contact microphones are built with a unidirectional, piezoelectric transducer, which is designed to be less susceptible to air-carried sound waves and more susceptible to surface-borne sound waves. The Korg CM-200 microphone, selected for data collection in this research, is commonly used in applications that involve capturing sound from a specific musical instrument (e.g., brass instrument, guitar, violin, and ukulele) when recording or practicing with an entire music band [27]. Contact microphones have the potential to be attached to heavy machinery while not being overly susceptible to the vibrations of the machine itself, unlike condenser microphones.

A microphone array is composed of at least two microphones working in tandem arranged in a linear, rectangular, or circular pattern. These microphones are usually omnidirectional; however, some microphone arrays are built using directional microphones or a combination of omnidirectional and directional microphones. Brandstein and Ward [30] compiled over 20 years of research in array-based microphone technology, which resulted in a thorough reference for immediate applicability of array technology in current systems and in improvement of existing devices. The applications for array-based microphones recorded in this compilation are based on beamforming techniques and include speech enhancement, speech recognition, source localization, noise reduction, echo cancellation, and separation of acoustic signals. Source localization and separation of acoustic signals have the potential to be very useful tools for locating heavy equipment and separating audio signals from other machinery working simultaneously in cluttered jobsites.

The XMOS xCORE-200 microphone array board, selected for data collection in this research, is a hardware and reference software platform equipped with 7 omnidirectional MEMS microphones with pulse density modulation (PDM) output (1 microphone is placed at the center at the board and the remaining 6 microphones are distributed equidistantly around the board edge, as shown in Fig. 4), a digital–analog converter (DAC), a processor with 16 32-bit logical cores, onboard low-jitter clock sources for multiple clocking options, 4 configurable buttons, 13 LED indicators, a USB 2.0 port, a RJ45 Ethernet port, a 3.5 mm audio jack, and other components. This particular microphone array board and its reference developer firmware are targeted, but are not strictly limited to, voice user interface (VUI) applications [29].

5 Research Methodology: Audio-Based Activity Recognition of Construction Heavy Equipment

To determine the software and hardware requirements for acoustical modeling of construction jobsites, the authors have developed an audio-based model for activity recognition of construction heavy equipment (Fig. 5). The five core components will be briefly described below. Hardware settings for data collection will be discussed thoroughly in the next section. More detailed information about the implemented system can be found at [31].

5.1 Denoising

Audio samples of heavy equipment mixed are almost certainly contaminated with noise from other sound sources found in a jobsite. To reduce the effect of this noise on later processing tasks, it is necessary to filter out as much noise as possible prior to subsequent processing. However, noise

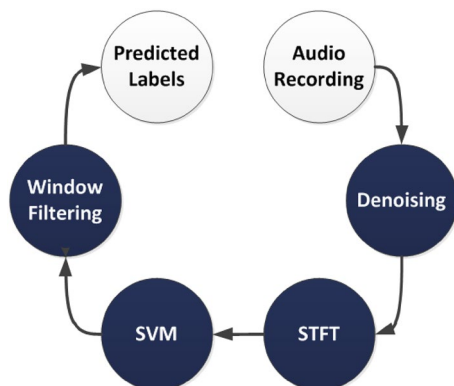


Fig. 5 The audio-based model for activity analysis of construction heavy equipment

filtering can also degrade the desired (equipment) signal and so the filtering must be balanced to reduce unwanted noise effectively while minimizing the distortion of the sound patterns of interest. Additionally, noise filtering must remain effective under the assumption that noise sources at a jobsite are not necessarily constant, since workers and equipment perform short tasks in an intermittent manner. Thus, a denoising algorithm for non-stationary environments developed by Rangachari and Loizou [32] was selected for implementation using MATLAB. This algorithm uses a noise estimation method for non-stationary environments that rapidly adapts depending on relative level of noise vs. signal of interest.

5.2 Sound Source Detection (Steered Response Power and Beamforming)

The case of dealing with multiple machines operating simultaneously at a jobsite is more challenging due to overlap between the sounds patterns generated by various machines. As a result, it is necessary to separate generated sound patterns by each individual machine first. For multiple machine recordings, sounds from different machines come from different directions. A robust source detection and activity recognition algorithm must respond to sound from a specific direction and block most of the noise from outside the direction of interest. To achieve this goal, the first step is to estimate the arrival direction for the sounds of interest. Direction of arrival for different machines is estimated using steered response power (SRP). The idea of SRP is that the location of the signal source radiates more energy than all other locations [33].

Thus, we can identify directions of arrival by scanning over all feasible angles and choosing the one with the largest value power. In practice, we scan through a hemisphere and find the direction with largest power. After finding the range, elevation and azimuth value from SRP, beamforming is used to isolate the signal from the desired direction. Beamforming consists of a spatial filter that uses the inputs from multiple microphones to isolate (as much as possible) signals arriving from a particular direction [34, 35]. The illustration of ideal direction detection result is shown in Fig. 6. The top figure shows the positions for microphones and sound sources, and the bottom figure shows the directions detected by the SRP technique. In our pipeline, we choose delay-and-sum beamformer, since it is the simplest and still highly effective.

5.3 Short-Time Fourier Transform (STFT)

Once enhanced and isolated, each audio signal is then converted into a time–frequency domain representation using the STFT. This technique consists of dividing a temporal signal into short segments and computing the Fourier transform

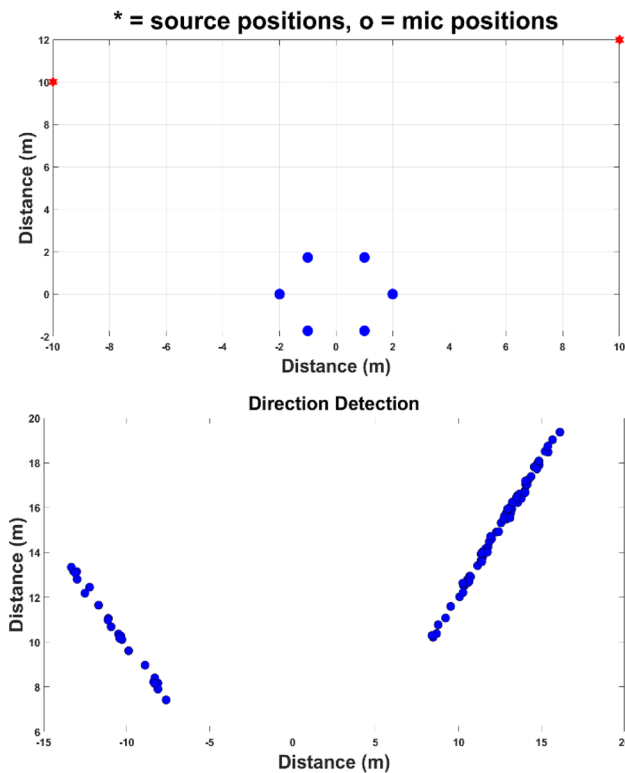


Fig. 6 Direction detection steps. The positions for microphone array and sound sources (top); two directions detected by steered response power method (bottom)

for each segment, thereby extracting sinusoidal frequency, magnitude, and phase content of each segment and representing these features as they change over time. MATLAB was used for STFT implementation using a Hanning window size of 512, a 1024-point discrete Fourier transform, and a 50% overlap. For the subsequent processing, only the magnitude content is retained.

5.4 Support Vector Machines (SVMs)

To identify activities being performed by construction equipment within an audio recording, an SVM classifier was trained to recognize each piece of equipment while performing major activities (class 1) and minor activities (class 2). The reasons for selecting SVM as the major machine learning platform was its popularity among signal processing research community as well as satisfactory results based on previous studies [36, 37]. SVM classifiers operate based on a simple principle: given an input of training data consisting of two classes, the system will generate a dividing hyperplane with maximum distance to the training samples (Fig. 7). Twice this distance is considered as the margin. Margin maximization reduces susceptibility to noise while employing the SVM to classify new data sets.

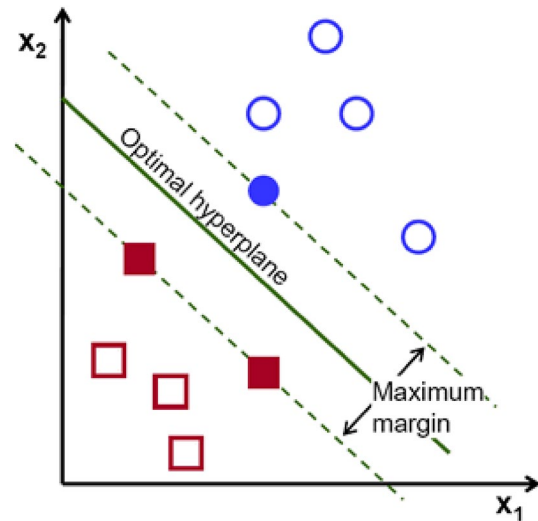


Fig. 7 A simple demonstration of optimal SVM hyperplane

For this project, the LIBSVM MATLAB package was used [36]. Four audio segments of the construction equipment performing a major activity were used to train class 1, and four audio segments of the construction equipment performing a minor activity were used to train class 2. Each of these segments was selected to be 2–6 s long and included only the STFT magnitude. To guarantee correct SVM parameter selection, tenfold cross-validation was used. It is well known that the performance of the SVM algorithm highly depends on the selected kernel function [37]. For this research, the linear and radial basis kernel functions have been selected, and experiments have been performed with both.

5.5 Window Filtering

Once an SVM classifier is generated for a specific construction equipment, it can be used for classification of the rest of the audio file. Nonetheless, direct implementation would potentially yield an output with predicted activities changing erratically from one time-frequency bin to the next. Therefore, a window filtering algorithm was implemented to smooth out the classified output. The window filtering parameters are a small window size, a large window size, and a threshold. Initially, if the SVM labels indicate that the percentage for a certain activity is greater than the threshold throughout the small window, the whole small window is labeled as that activity. Then, this is repeated using the small window labels for the large window size. The size of the window will vary in different cases, but in general the small window can be set as a quarter second and the large window can be a second or 2 s.

6 Experimental Setup: Audio-Based Activity Recognition of Single Machines

A selection of various construction heavy equipment has been chosen to help assess how well the audio-based activity analysis system performs under different hardware and software settings. Using four recording devices, audio data from heavy machinery performing routine activities was collected. The selected microphones were two off-the-shelf Zoom H1 handy recorders (regular microphone), one on a tripod located on-site and one onboard the machine; one XMOS xCORE-200 USB microphone array connected to a laptop; and one KORG CM-200 contact microphone connected to a flat surface in the cabin of the equipment. The XMOS microphone array was interfaced to a Windows PC using the manufacturer's USB Audio Class 2.0 Evaluation Driver for Windows and multichannel audio was recorded through Audacity, an open source program for recording and processing audio. The setup process for data collection is depicted in Fig. 8.

7 Experimental Setup: Audio-Based Activity Recognition of Multiple Machines Working Simultaneously on a Jobsite

The experimental setup for the single machine case is almost the same as the multiple machine case. The only difference is that we only use the XMOS xCORE-200 USB microphone array and regular on-site microphone in the multiple machine cases. Also, we record each machine separately to be used as our training library in the multiple machine cases. Recordings from the microphone array can provide us with phase information, which can help us track the location of sound sources as described above. With the spatial

information, the proposed activity recognition framework can work better in the multiple machine cases.

In addition to audio data recording, a video sample was taken using a digital video camera to serve as a reference for manual labeling of major and minor activities (Fig. 8). Examples of major activities include digging, loading, dumping, and crushing rock, while examples of minor activities include swinging, maneuvering, and extending arm. Once the devices were in place and recording, a loud sound was produced with an air horn. This signal was to be used as a synchronization point for the data. Manual labels served as ground truth data to assess the activity recognition accuracy.

The results for one sample machine are depicted in Fig. 9. The machine, an Ingersoll Rand SD25 compactor, was recorded using a Zoom H1 digital handy recorder placed relatively close to the machine on the jobsite. The presented results are based on using an SVM with a linear kernel. The top part of the figure presents the normalized frequency for the machine's audio signal, while the middle and bottom plots show the actual (black) versus predicted (blue) activity labels and over the audio recording time period. As indicated in these figures, there is an excellent correlation between the actual and predicted results generated for the Ingersoll Rand compactor.

To quantify the accuracy of obtained results, the mean squared error (MSE) or mean squared deviation (MSD) values of case study machines have been calculated. MSE is the measure of quality of an estimator—it is always non-negative, and values closer to zero are better:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{X}_i)^2, \quad (1)$$

where X_i is the predicted label, \hat{X}_i is the actual label, and n is the total number of labels. Obtained results, which are very promising, are summarized in Table 1.



Fig. 8 Setup process for audio collections using microphone array and off-the-shelf microphone on-site (left), and contact microphone and off-the shelf microphone onboard (middle); simultaneous video recording for generating the ground truth data (right)

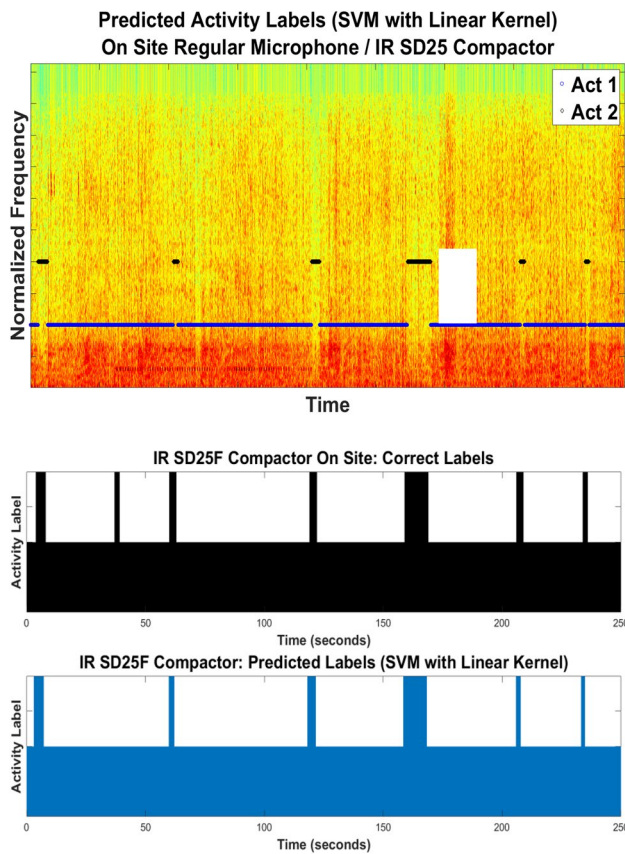


Fig. 9 Results for Ingersoll Rand SD25 Compactor

Table 1 Comparison between actual and predicted activity labels through calculating MSE values

Machine	Mean squared error (MSE)
JD 700J Dozer	0.1205
JD 670G Grader	0.204
JCB 3CX Excavator	0.253
Komatsu PC200 Excavator	0.291
Concrete Mixer	0.147

8 Experimental Setup: Comparison between Different Frequency Feature Extraction Techniques

Selecting optimum frequency feature extraction method is an important decision for optimizing the performance of the proposed audio-based activity recognition package. In this study, two common feature extraction techniques, the short-time Fourier transform (STFT) and the continuous wavelet transform (CWT) have been selected and compared. STFT is based on dividing a long-time signal into

short segments and computing the Fourier transform for each segment. CWT is a derivation of the Fourier transform that was primarily designed to locate frequency features in time or space. The Fourier transform is a powerful tool for frequency analysis; however, it does not characterize rapid frequency changes efficiently, because it represents data as a sum of sine waves, which extend to infinity. A wavelet is a rapidly decaying, wave-like oscillation that has zero mean. More detailed information regarding each feature extraction method could be found at relevant references such as Mathworks, Inc. report, 2017 [38].

To conduct the comparison study between STFT and CWT, both techniques were implemented using recordings taken at local construction jobsites for various types of equipment. Two configurations for CWT (8 octaves/32 scales per octave; and 10 octaves and 24 scales per octave) and one configuration for STFT (1024 frequency points, 512 sample windows, and 256 overlapped samples) have been employed in this research.

9 Results and Discussion

Tables 2, 3, 4, 5 and 6 and Figs. 9, 10, 11 and 12 summarize comparison results for several construction equipment under different hardware and software configurations. Through careful analysis of these results, we arrive at the following conclusions.

- For onboard microphone placement, contact microphones generate slightly more accurate results and, thus, are the better choice (Table 2).
- A comparison for the current case study results shows that regular microphones placed on the construction site

Table 2 Comparison between contact and regular microphones (onboard/SVM using the RBF kernel)

Machine	Contact microphone		Regular microphone	
	Major activity (%)	Minor activity (%)	Major activity (%)	Minor activity (%)
JD50D Compact Backhoe	72.08	74.02	71.14	82.78
Ingersoll Rand Compactor	80.34	7.83	80.07	1.98
CAT 320E Excavator	80.78	38.26	71.20	29.71
Komatsu PC200 Excavator	71.05	48.56	70.16	55.59
JD 700J Dozer	76.16	64.46	78.79	60.15
Hitachi 50U Excavator	49.58	57.96	53.15	54.17
Concrete Mixer 2	68.79	65.68	77.16	70.23

Table 3 Comparison between contact (onboard) and regular (on-site) microphones

Machine	Contact microphone (onboard)		Regular microphone (on-site)	
	Major activity (%)	Minor activity (%)	Major activity (%)	Minor activity (%)
CAT 320D Backhoe Excavator	80.78	38.26	81.09	71.48
JD 333E Compact Loader	73.42	95.68	86.54	90.11
JD50D Compact Backhoe	72.08	74.02	86.65	57.74
Ingersoll Rand Compactor	80.34	7.83	81.58	32.03
CAT 320E Excavator	80.78	38.26	81.09	71.48
Komatsu PC200 Excavator	71.05	48.56	82.17	67.89
JD 700J Dozer	76.16	64.46	84.69	72.45
Concrete Mixer 2	68.79	65.68	83.23	61.59

Table 4 Comparison between on-site and onboard regular microphones

Machine	On-site		Onboard	
	Major activity (%)	Minor activity (%)	Major activity (%)	Minor activity (%)
JD50D Compact Backhoe	86.65	57.74	71.14	82.78
Ingersoll Rand Compactor	81.58	32.03	80.07	1.98
JD 333E Compact Loader	86.54	90.11	80.67	81.47
CAT 320E Excavator	81.09	71.48	71.20	29.71
Komatsu PC200 Excavator	82.17	67.89	70.16	55.59
JD 700J Dozer	84.69	72.45	78.79	60.15
Hitachi 50U Excavator	79.89	55.97	53.15	54.17
Concrete Mixer 2	83.23	61.59	77.16	70.23

yield better accuracies than contact microphones attached to flat surfaces in the cabin. One clear reason for this phenomenon is that contact microphones onboard are affected by engine noise and the vibrations; nonetheless, further investigations using several other data sets and

Table 5 Comparison between array and on-site regular microphones on multiple machine recordings

Multiple machine cases	Microphone array		On-site regular microphone	
	Major activity (%)	Minor activity (%)	Major activity (%)	Minor activity (%)
Bobcat Loader and Bobcat Excavator				
Bobcat Loader	72.37	52.74	65.34	62.78
Bobcat Excavator	74.29	60.17	68.16	51.17
Bomag Compactor and JD 544K Loader				
Bomag Compactor	81.12	49.07	70.22	57.88
JD 544K Loader	76.14	55.47	59.74	41.81
CAT Bulldozer and CAT CS44 Compactor				
CAT Bulldozer	80.42	81.33	71.54	65.14
CAT CS44 Compactor	79.08	77.29	67.44	50.08
CAT C5K Bulldozer and CAT 305E Excavator				
CAT C5K Bulldozer	78.57	61.90	57.82	39.97
CAT 305E Excavator	77.69	64.18	62.26	47.64

Table 6 Calculated cycle time and productivity rates for the selected case study machine: JD 50G Backhoe

JD 50G Backhoe (bucket size: 24 inches—approximately 0.15 m ³)	
Cycle time (s)	
Actual	18.6 s
Predicted	16.97 s
% Error	9.6%
Number of cycles per hour (assuming 45 min efficient time per hour)	
Actual	144
Predicted	159
% Error	10.42%
Productivity rate (m ³ /h)	
Based on bucket fill factor = 0.75	
Actual	15.85 m ³
Predicted	17.88 m ³
% Error	12.81%

- under different jobsite conditions might be required to fully substantiate this conclusion (Tables 3, 4).
- As shown in Fig. 10, results show that the radial basis function kernel outperforms the linear kernel in most cases, particularly when it comes to recognizing the major activities.
 - No significant differences can be found between regular microphones and microphone arrays when it comes to recognizing the activities of single machines through microphones placed on-site. This is due to there being only one source of audio, and therefore only one direction of incoming audio. It is suggested that a noticeable

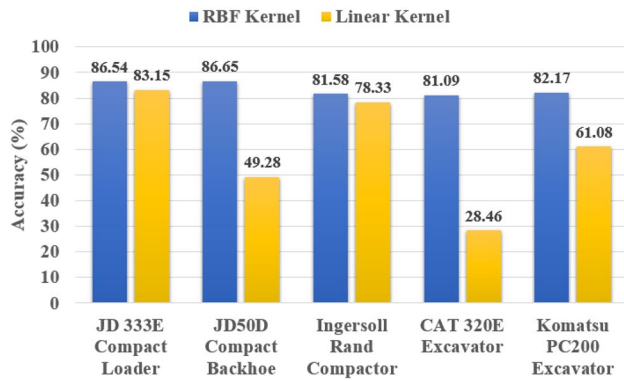


Fig. 10 Summary of the results for implementing RBF and linear kernels for activity recognitions of machines (major activities)

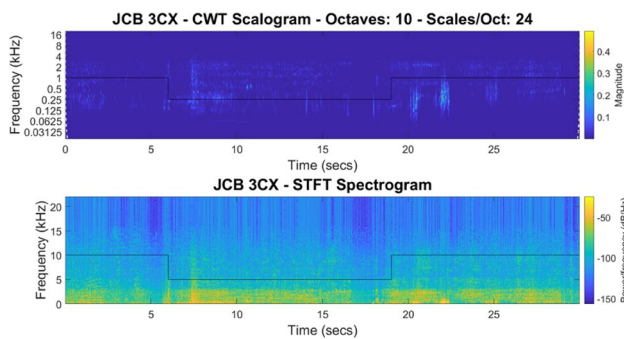


Fig. 11 Sample of comparison results between CWT (10/24) vs. STFT spectrograms—machine: JCB 3CX

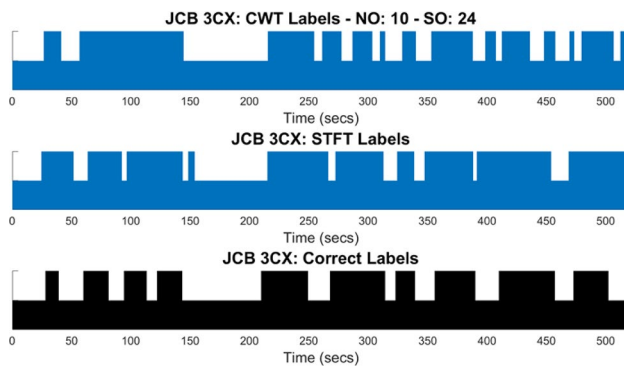


Fig. 12 Labeling comparison between CWT (10/24) vs. STFT spectrograms—machine: JCB 3CX

difference between a regular microphone and a microphone array will be seen in the recording of multiple machines, and the multiple directions of generated audio from said machines. Microphone arrays will become useful to detect multiple audio directions from multiple audio sources.

- In Table 5, we can find in the multiple machine case, microphone array can provide us with better recognition

results compared to regular on-site microphone. The reason is that with direction detection step, we can eliminate most of the noise outside the direction of interest. When using the microphone array, we improve the performance of activity recognition using recordings that only contain one machine sound as our training library. After the direction detection step, we used the trained library to perform the activity recognition in the mixed recording. However, when using a single on-site microphone, we can only do the activity recognition on the mixed recording since we are lacking spatial information.

- In general, it can be observed that CWT scalograms yield a better time–frequency magnitude representation than STFT spectrograms (samples of generated results for a case study machine (JCB 3CX) is presented in Fig. 11). Additionally, it has been illustrated that using a smaller number of octaves with higher resolution within the octaves is preferable when lower-frequency features are not of interest. Using a CWT with eight octaves outperformed using a CWT with ten octaves on processing time and labeling accuracy.

10 Case Study

To better understand the potential application of the proposed system in construction projects, a case study is presented here: An earthmoving jobsite with a John-Deere 50G Backhoe was considered as the testbed. The Backhoe was equipped with a 24” bucket with a nominal capacity of 0.15 m³. The machine performed the excavation tasks in a cyclic form. Each operation cycle consists of a series of activities: excavating, loading, swinging, and dumping. The entire operation was carefully monitored manually and by recording generated audio files for one complete shift (4 working hours). The recorded audio files were processed by the proposed system and different activities as well as average cycle times have been calculated. By calculating cycle times and since the bucket size is known (0.15 m³), the average productivity rate of the whole operation could be calculated using the following equation:

$$\text{Productivity rate} = \text{Output per cycle} \times \text{number of cycles per hour} \quad (2)$$

Results of recognizing activities and average productivity rates based on generated sound patterns as well as actual observations are summarized in Table 6.

Obtained results illustrated that is possible to accurately predict productivity rates of the case study machine with less than 15% error. Accurately predicting productivity rates of construction equipment is an important task for construction engineers and jobsite personnel and the results could be further employed for different applications such as automated

calculation of accomplished work, predicting necessary budget (and time) for remaining work, earned value analysis, etc.

11 Summary and Conclusion

This work provides an innovative audio-based solution for activity analysis of construction heavy equipment and acoustical modeling of construction jobsites. An audio-based activity recognition model has been developed and tested under various hardware and software settings for the case involving single machines. The methods of the experiment apply the three types of microphones discussed earlier (off-the-shelf, contact, and microphone arrays) and two placement settings (on-board vs. on-site). The comparison results have been presented in the previous sections of this paper. Additionally, the results were also processed with two major SVM kernel types (linear and RBF) to determine which kernel yielded the best results. The RBF kernel performed better in most cases. As an extension of the current research, plans for future research include expanding into the following items:

- Implementing and testing the audio-based model with more diverse data sets and conditions.
- Determining what hardware and software settings would be required to implement this system on a jobsite with multiple machines working simultaneously.
- Developing more robust algorithms that would recognize more detailed tasks performed by heavy equipment and machinery. That is, separating major and minor activities into sub-activities.

Acknowledgements The presented research has been funded by the U.S. National Science Foundation (NSF) under Grants CMMI-1606034 and CMMI-1537261. The authors gratefully acknowledge NSF's support. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. The authors also would like to thank Mr. Chris Sabillon, MS student at Georgia Southern University, for assisting with data collections and processing tasks.

References

- Rashidi A, Nejad HR, Maghiar M (2014) Productivity estimation of bulldozers using generalized linear mixed models. *KSCE J Civ Eng* 18(6):1580–1589
- Pradhananga N, Teizer J (2013) Automatic spatio-temporal analysis of construction site equipment operations using GPS data. *Autom Constr* 29:107–122
- Akhavian R, Behzadan AH (2013) Simulation-based evaluation of fuel consumption in heavy construction projects by monitoring equipment idle times. In: 2013 Winter simulations conference (WSC), Washington, DC. <https://doi.org/10.1109/WSC.2013.6721677>
- Ahn CR, Lee S, Peña-Mora F (2015) Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet. *J Comput Civ Eng* 29(2):04014042
- Akhavian R, Behzadan AH (2013) Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining. *J Constr Eng Manag* 139(11):04013021
- Akhavian R, Behzadan AH (2015) Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. *Adv Eng Inform* 29(4):867–877
- Brilakis I, Fathi H, Rashidi A (2011) Progressive 3D reconstruction of infrastructure with videogrammetry. *Autom Constr* 20(7):884–895
- Golparvar-Fard M, Heydarian A, Niebles JC (2013) Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Adv Eng Inform* 27(4):652–663
- Gong J, Caldas CH, Gordon C (2011) Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Adv Eng Inform* 25(4):771–782
- Heydarian A, Golparvar-Fard M, Niebles JC (2012) Automated visual recognition of construction equipment actions using spatio-temporal features and multiple binary support vector machines. In: Construction Research Congress, West Lafayette. <https://doi.org/10.1061/9780784412329.090>
- Kim JY, Caldas CH (2013) Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects. In: 30th International Association for Automation and Robotics in Construction, Montréal. <https://doi.org/10.22260/ISARC2013/0072>
- Rashidi A, Fathi H, Brilakis I (2011) Innovative stereo vision-based approach to generate dense depth map of transportation infrastructure. *Transp Res Rec* 2215(1):93–99
- Rezaazadeh Azar E, McCabe B (2012) Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Autom Constr* 24:194–202
- Zhu Z, Park MW, Koch C, Soltani M, Hammad A, Davari K (2016) Predicting movements of onsite workers and mobile equipment for enhancing construction site safety. *Autom Constr* 68:95–101
- National Instruments (2016) What determines if a transducer is active or passive?. <http://bit.ly/ICSlyS3>. Accessed 30 Sept 2016
- Cheng CF, Rashidi A, Davenport MA, Anderson DV (2017) Activity analysis of construction equipment using audio signals and support vector machines. *Autom Constr* 81:240–253
- Cho C, Lee YC, Zhang T (2017) Sound recognition techniques for multi-layered construction activities and events. In: ASCE international workshop on computing in civil engineering, Seattle. <https://doi.org/10.1061/9780784480847.041>
- Yang S, Cao J, Wang J (2015) Acoustics recognition of construction equipments based on LPCC features and SVM. In: 34th Chinese in control conference (CCC), Hangzhou. <https://doi.org/10.1109/ChiCC.2015.7260254>
- Cao J, Wang W, Wang J, Wang R (2016) Excavation equipment recognition based on novel acoustic statistical features. *IEEE Trans Cybern* 47(12):4392–4404
- Bengtsson M, Olsson E, Funk P, Jackson M (2004) Technical design of condition-based maintenance system—a case study using sound analysis and case-based reasoning. In: 8th International conference of maintenance and reliability (MARCON), Knoxville
- Greenemeier L (2008) A positioning system that goes where GPS can't. *Scientific American*. <http://bit.ly/2hsVPNJ>

22. Naik RP (2009) Ultrasonic: a new method for condition monitoring. National Thermal Power Corporation, Raipur. <http://www.reliableplant.com/Read/20554/ultrasonic-condition-monitoring>. Accessed 26 March 2016
23. Holmes MS, D'Racy S, Costello RW, Rielly RB (2014) Acoustic analysis of inhaler sounds from community-dwelling asthmatic patients for automatic assessment of adherence. *IEEE J Transl Eng Health Med* 2:1–10. <https://doi.org/10.1109/JTEHM.2014.2310480>
24. Sulaiman I, Cushen B, Greene G, Seheult J, Seow D, Rawat F, MacHale E, Mokoka M, Moran CN, Bhreathnach S, MacHale A, Tappuni P, Deering S, Jackson B, McCarthy M, Mellon H, Doyle L, Boland F, Reilly F, R.B. and Costello RW (2017) Objective assessment of adherence to inhalers by patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 195(10):1333–1343
25. Ballou G (2015) *Handbook for sound engineers*, 5th edn. Focal Press, New York
26. Yamaha (2016) Which types of microphone are used with PA systems? <http://bit.ly/2cMoi0Y>
27. Korg (2016) CM-200 contact microphone. <http://bit.ly/2qAhUOf>. Accessed 8 Oct 2016
28. Zoom Corporation (2016) <https://www.zoom.co.jp/products/handy-recorder/h1-handy-recorder>. Accessed 13 May 2017
29. XMOS Ltd (2016) xCORE Microphone Array Hardware Manual
30. Brandstein M, Ward D (2001) *Microphone arrays: signal processing techniques and applications*. Springer, Berlin. <https://doi.org/10.1007/978-3-662-04619-7>
31. Cheng CF, Rashidi A, Davenport MA, Anderson DV (2016) Audio signal processing for activity recognition of construction heavy equipment. In: 33rd International symposium on automation and robotics in construction (ISARC 2016), Auburn. <https://doi.org/10.22260/ISARC2016/0078>
32. Rangachari S, Loizou PC (2006) A noise-estimation algorithm for highly non-stationary environments. *Speech Commun* 48(2):220–231
33. Dmochowski JP, Benesty J, Affes S (2007) A generalized steered response power method for computationally viable source localization. *IEEE Trans Audio Speech Lang Process* 15(8):2510–2526
34. Gannot S, Burshtein D, Weinstein E (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans Signal Process* 49(8):1614–1626
35. Johnson DH, Dudgeon DE (1993) *Array signal processing: concepts and techniques*. Prentice Hall, Englewood Cliffs
36. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST)* 2(3):1–27
37. Rashidi A, Sigari MH, Maghiar M, Citrin D (2016) An analogy between various machine-learning techniques for detecting construction materials in digital images. *KSCE J Civ Eng* 20(4):1178–1188
38. MathWorks, Inc. (2017) Time–frequency analysis with the continuous wavelet transform. <http://bit.ly/2yyMM7j>. Accessed Oct 2017