Combining convolutional neural networks with unsupervised learning for acoustic monitoring of robotic manufacturing facilities

Jeffrey Bynum\textsuperscript{1} and David Lattanzi\textsuperscript{2}

Abstract

For automated robotic manufacturing, a key aspect of monitoring is the identification and segmentation of core actuation processes captured in sensor logs. Once segmented, the behavior of an industrial system during a particular actuation can be tracked to detect signs of degradation. This study presents a technique for performing such an analysis through a combination of machine learning techniques designed to work with an acoustic monitoring system. A spectrogram-based convolutional neural network (CNN) is first trained to identify and segment primary motion classes from acoustic data. Unsupervised clustering and feature-space analysis are then employed to further separate the data into motion sub-classes beyond the capabilities of the CNN. This approach was evaluated on acoustic recordings of a Selective Compliance Assembly Robot Arm (SCARA) system. The developed CNN performed primary robotic motion segmentation with a maximum actuation identification accuracy of 87% when compared to validation data. The unsupervised clustering process had mixed success in distinguishing more fine-grained motion sub-classes due to strong variances in signal energy for some sub-classes. Further refinement is required for improved segmentation accuracy as well as automatic feature generation. The application of this process for life-cycle system monitoring is discussed as well.

Keywords

Acoustics, deep learning, convolutional neural network, manufacturing

Introduction

Industrial manufacturing equipment often requires extended operation for maintaining fabrication tolerances as well as production capacity.\textsuperscript{1} These systems demand persistent monitoring and maintenance. Conventional approaches rely on operator intuition; however, remote health monitoring techniques are increasingly applied due to caustic fabrication environments.

Industrial robots are complex mechanical systems that combine components such as motors, bearings, linear rails, and brakes. These systems typically perform a variety of actuations in 3D Cartesian space. These actuations produce measurably discernible acoustic signatures from various combinations of engaged mechanical hardware that should be identifiable through an acoustic monitoring system. Conceptually, changes in

\textsuperscript{1}\textit{Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA}

\textsuperscript{2}\textit{Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, VA, USA}

\textbf{Corresponding author:} David Lattanzi, Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, 4400 University Drive, MS 6C1, Fairfax, VA 22030, USA.

Email: dlattanz@gmu.edu
system performance stemming from mechanical degradation should be identifiable as well, with changes in acoustic harmonic content indicating progressive machine wear or damage. Frequently, only limited labeled data is available for algorithm training and model fitting, often with only weak labels that do not fully describe the complete range of system responses. This paper investigates how to develop a monitoring system under such conditions by combining deep supervised machine learning with unsupervised similarity analysis. This combined approach is then evaluated through a case study implementation.

Prior work

While monitoring industrial components using auditory characteristics has been studied, numerous practical challenges for monitoring systems in realistic environments persist. Several recent comprehensive literature surveys have outlined various challenges in remote monitoring.1–3 These surveys highlight the move toward data-driven approaches that minimize the need for statistical parameterization of monitored phenomena, as well as a reduction in the sensitivity of results to variations in feature engineering and extraction. Across a range of industrial applications, researchers have illustrated the benefits of employing customized deep and reinforcement learning strategies for robotic sensing and perception.4–7

With regards to acoustic monitoring, distinct audio features that can be used to classify events as well as damaged mechanical states are often difficult to determine. Signal features correlating with damage are often weak in amplitude compared to other recorded operations,8 and deviations reflecting damage are not fully understood. Expert knowledge is often required to generate features and categorize audible dynamics corresponding to mechanical wear. Moreover, unique features must be separable, distinctive, and representative for audio signature classification.1,9–12 There have been a variety of studies on generating such features,1,9,10,12–14 however such features are typically domain dependent. Audio-based features suitable for noisy environments are debated in literature. While features derived from mel-frequency cepstrum coefficients (MFCC) have been used in a variety of sound classification studies, other descriptors have outperformed MFCCs in high noise environments.10 Bach et al.15 suggested that other feature sets, such as amplitude modulation, can outperform MFCCs with noise corrupted signals in speech detection.

To mitigate this problem, modern machine learning methods can be employed.2 Many machine learning methods are at least somewhat insensitive to variations in feature quality. In particular, deep neural network learning models often mitigate the need for rigorous feature engineering at all, one of the advantages that motivated their development.16–18 Recently, a variety of studies have considered using a spectrogram representation of an acoustic signal as a neural network input, rather than the raw time series data. A spectrogram is generated through short-time Fourier transform procedures, and this data representation can capture evolving harmonic content over time across a broad range of the available frequency spectrum. The most effective deep learning method for such inputs has been shown to be a convolutional neural network (CNN).10 A CNN, as opposed to a conventional feed-forward network, is ideal for such applications due to its inherent shift-invariance. In a CNN, a series of convolutional kernels (filters) are learned, such that both spatial and temporal relationships are captured. In the context of a spectrogram, this allows a CNN to identify frequency response signatures that do not occur on a repeatable cycle. Segmentation of acoustic events using a spectrogram-based CNN has been shown effective across many application domains, however the effectiveness of such methods is reliant upon large and exhaustive data sets for CNN model training.10,19–22 Specific concerns and justifications for using convolutional neural networks (CNN) in environmental sound classification were presented in Piczak,23 most notably that the limited data sets available in such applications require alternative approaches when compared to state of the art speech recognition.

Other researchers have addressed challenges in the comparative analysis of audio content through unsupervised machine learning. Such approaches seek to extract key signal signatures, or identify similarities between signals, in a manner that does not require explicit statistical training data for model fitting. This work has often taken the form of feature similarity analysis or data clustering through methods such as k-means clustering. In the context of acoustic waveforms, researchers have focused on creating unsupervised methods to autonomously identify similarities in audio recordings.24,25 MFCC features were extracted from partial spectrogram segments from the full-length waveform. The feature vectors from the partial segmentations were compared, resulting in a cosine similarity heatmap. The authors then used the mapping as both a visual tool and similarity comparison between waveforms. The study in Jang et al.26 documents how unsupervised spectrogram features could be extracted from an encoding layer in a deep autoencoder network and clustered to determine different audio sources. The two-fold approach suggested extracted spectrogram features allowed for separable classification. The concept of unsupervised analysis has been studied for robotic systems, for example with autonomous vehicles.27 Researchers have also explored approaches for combining unsupervised clustering with supervised
audio classification in application spaces such as emotion recognition and the analysis of animal vocalizations, though the concept has not previously been considered for industrial system monitoring.

Alias outlined several limitations for direct implementations of CNN architectures and spectrograms related to classifying audio content. Harmonic content is generally complex and spatially separable in the frequency axis, often requiring transforms for optimal kernel mapping. Limitations directly adapting image processing based CNNs to audio-based classification were discussed in Fanioudakis and Potamitis. Spectrogram content often contains a rectangular aspect ratio – different than traditional, square CNN inputs. This critical difference prevents the re-purposing of CNNs trained for image classification using techniques such as transfer learning.

Overall, a general survey of prior work indicated that deep neural networks, particularly CNN variants, are viable and effective for segmenting acoustic signals. Additionally, prior work reflects the value of using spectrogram representations as network inputs based on additional invariance to temporal distortions. Furthermore, studies indicate that features present in spectrograms contain viable information relevant for audio classification and analysis. However, the use of such approaches for acoustic-based remote health monitoring, particularly in combination with unsupervised analysis approaches, has not been extensively studied.

Research contribution

In this work, we propose and develop a new approach for the segmentation and analysis of robotic actuation signatures from acoustic recordings. Inspired by recent advancements in combining a CNN and unsupervised clustering for speech recognition, this approach combines a convolutional neural network designed for asymmetric spectrograms with unsupervised feature space analysis, in order to provide a more nuanced understanding of system behavior than can be accomplished with either method alone. The overall concept is to use deep learning to provide high-level segmentation on weakly labeled monitoring data, of the type commonly found in industrial applications. Unsupervised analysis via feature similarity and clustering are then used to provide insights into intra-class variances that would not be possible using only supervised learning.

The CNN architectures presented here are based on conventional architectures for square image inputs, however they are modified to accommodate the asymmetric nature of spectrogram inputs. No claims are made regarding the optimality of the presented CNN architectures. Rather, one goal of this work is to illustrate a generalizable approach to accommodating asymmetric spectrograms in industrial monitoring systems that employ deep learning.

As an illustrative example, this combination of supervised and unsupervised analysis is then used to capture and quantify instances of robotic motions embedded within recordings of a manufacturing facility. Recordings from a SCARA-series (Selective Compliance Assembly Robot Arm) were leveraged for model training and testing, as well as for assessment of the similarity analysis. Such robots are widely used in manufacturing for tasks such as material transitioning. Such an approach could be used to track system changes arising from progressive wear and damage across an array of similarly deployed systems. As such, this study also serves as a case study on the capabilities and challenges of acoustic monitoring in industrial settings.

The remainder of this study is organized as follows. Presented first is a discussion of the behavior of the robot used for model development, as well as a preliminary study using less complex network architectures that motivated the need for both a deep network and a convolutional architecture. This is followed by the development and delineation of both the deep learning and similarity analysis components of the methodology, and then by an experimental analysis of the methodology on the previously mentioned SCARA-series robot motion recordings. The paper concludes with a discussion of the findings and avenues for future work.

Preliminary study

For clarity, the behavior of the SCARA-series robot used for prototyping and development is presented.
Primary SCARA robotic actuations can be generalized as a series movements within a Cartesian coordinate system. As shown in Figure 1, these primary motions include y-axis (base movement), x-axis (arm extension/retraction), and z-axis (body extension/retraction) motions. Limited harmonic content emission and associated operator intuition with regards to z-axis motion limited the scope of this study to analyzing y- and x-axis actuations. Each motion class had a duration of approximately 1 s.

Primary motions can be further subdivided into separate subclasses. Each subclass describes a physically different actuation along the same axis of travel. For example, class (1a) describes motion from point A to point B whereas subclass (1b) describes motion from point B to point A. While these actuations are similar in nature, each motion generates slightly different mechanical responses and thus acoustic emissions. Similarly, x-axis actuations include combinations material loading and transition movements. However, many of these secondary actuation classes occur only sporadically. As a result, it was impractical to attempt to classify them through a supervised machine learning approach, given the lack of relevant labeled training data. A methodology to distinguish these secondary classes through feature-space analysis of segmentations is presented later.

These actuations were recorded using a dbx RTA-M microphone, at a sampling frequency of 48 kHz, in a enclosed manufacturing bay containing a single SCARA-type robotic tool. The microphone was omnidirectional, with a flat frequency response between 20 and 20 kHz. A 4-min audio recording containing a variety of actuations was used for training data. An additional 11-min audio recording, aligned with corresponding video data, was used as final validation data as well. Visual analysis of the resulting spectrograms of these recordings shows that primary motion audio signatures are distinct and frequent enough to be potentially distinguishable through supervised machine learning, however they occur over drastically different frequency ranges and with broad variations in power. For instance, y-axis motions are visually observable in a band ranging from 0 to 20 kHz, while x-axis motions are observable from 23 to 24 kHz (Figure 2).

Figure 2. Spectrogram representations of SCARA actuations captured in audio recordings.
aggregated as well. To reduce potential overfitting from class size imbalances, motion classes were capped to the minimum number of segmentations recorded (in this case $y$-axis motion) and sampled without replacement to reach the identical size. The process is analogously discussed in the subsequent CNN architecture development section.

The resulting performance of these models is shown in Figure 4. Lower energy actuations ($x$-axis) were consistently confused with noise states (visualized in Figure 3(a)) and the addition of hidden layers provided only a slight improvement as described in Figure 3(b). The minimal improvement, despite increasing architecture layers, possibly indicates data overfitting rather than increased classification accuracy. These results suggested a more sophisticated architecture was required due to the complex relationships present in actuation signals and the broad variations in representations between $y$ and $x$-axis motions.

### Methodology

Similar to relevant prior literature, a convolutional neural network (CNN) was designed to perform segmentation of primary motion classes from an unlabeled spectrogram. From these segmentations, unsupervised feature similarity analysis and k-means clustering are then performed to identify secondary motions. This combined method offers unique advantages. Convolutional neural networks enable unsupervised
learning of critical feature representations, reducing systematic bias from manual feature engineering. Additionally, complexity inherent within robotic movement data as well as non-stationary environmental noise suggests that contextual information from spectrograms can be exploited. However, due to the limited labeled data available for actuation sub-classes, the training data could only be divided into primary groups, necessitating the need for the unsupervised feature analysis. However, this two-step analytical approach has the benefit of being generalizable to a wider range of SCARA implementations, as the primary motions are somewhat universal in their acoustic response signatures, whereas secondary class signatures are not.

**Spectrogram dataset development**

A review of the relevant literature indicated that signal pre-processing via filtering was not advisable; simplistic techniques to isolate noise and characteristic information can unintentionally decrease acoustic emission. Preliminary experiments also indicated that isolating and correcting for wide-spectrum environmental noise was infeasible and could potentially obscure valuable data. Furthermore, extensive signal pre-processing for initial segmentation would bias resulting approaches to the SCARA unit data employed for testing, limiting extensibility to other robotic systems.

Spectrograms were generated using a Blackman-Harris window of 4096 waveform data points, overlap of 2048 data points, and 8096 Discrete Fourier Transform (DFT) points. Input spectrograms were fixed with 4097 (height) × 25 (width) dimensions, with regards to the resulting temporal-frequency bins of the spectrograms. These dimensions stemmed from the shortest known event duration (1.1 s). Due to the fixed length of the spectrogram input and variability between sub-class y-axis motion duration, some training examples were unintentionally split into overlapping sections. While introducing redundant information potentially biases network training by overfitting, several other benefits may be introduced. The variation in training set may outweigh bias introduced in redundant overlapping sections, as mentioned in Piczak. Moreover, the CNN will realistically encounter similar, partially obscured spectrograms during segmentation in operation.

The training set consisted of 1360, 769, 1181 training examples of noise, y-axis, and x-axis motions, respectively. Training data sizes were kept constant. For example, training a 3-class classifier CNN between y-axis, x-axis, and noise used 769 examples as the maximum training set size; the other classes were randomly sampled, without replacement, to match the minimum training set amount. Keeping the training data count consistent further prevented overfitting bias. The final training dataset contained 2076 spectrogram samples with 231 samples (10%) withheld for validation.

Spectrogram normalization was first used to limit spectrogram variability. Across 3000 + test spectrograms, minimum, and maximum temporal-frequency bin values were calculated and used as boundaries. All training and test spectrogram temporal-frequency were rescaled from a minimum bound of \(-14.68\) and maximum bound of 5.89 to values between 0 and 255, making the spectrograms a more image-like data type and enabling a wider range of network activation functions. A similar mapping procedure was successfully implemented in Dennis.

**CNN architecture development**

To the authors’ knowledge no established CNN architecture exists for spectrogram training as reaffirmed in previously mentioned studies. Moreover, due to domain-specific requirements, other spectrogram specific architectures were unsuitable for direct application. Due to the high resolution and frequency dependent spacing in training spectrograms, an empirical study was conducted to evaluate general trends among candidate CNN architectures. A hyperparameter sensitivity study addressing convolutional kernel dimensions was performed; however, an exhaustive parameter search for an optimal architecture was considered outside the scope of the current work and requires future study.

A key consideration in the development of the CNN architecture was the highly asymmetric aspect ratio of the spectrogram inputs, a relatively understudied topic when compared to the exhaustive body of literature on developing CNNs for square aspect ratio image inputs. Prior work in developing CNN architectures suggested either extracting key excitation regions within spectrograms or scaling spectrogram inputs, in order to accommodate standardized CNN architectures with square aspect ratios. However, the dependence on coupled wide-band frequency and temporal content meant that this approach was infeasible, and necessitated the development of the customized CNN architectures presented in this work.

These architectures are based on the well-known LeNet and AlexNet CNN architectures due to their simplicity and flexibility. These CNN models use five successive convolutional layers with kernel mappings that range from \(11 \times 11\), \(5 \times 5\), and \(3 \times 3\). Due to initial difficulty with model convergence, likely due to the skew of the spectrogram inputs, a sixth convolutional layer was added. The dimensions for padding, stride, and receptive fields in the convolutional layers were adjusted to account for the asymmetric input. Other hyperparameters were tuned empirically. The
authors note that this approach was designed to explore applicability, but not necessarily the optimization, of convolutional nets for diagnostic analysis of spectrograms.

Three architectures were developed in order to understand how varying convolutional kernel dimensions impacted asymmetric spectrogram segmentation. The architecture construction does not vary dramatically among the three models, as the goal was to identify trends in symmetric/asymmetric filter sizes for the asymmetric spectrogram inputs. Padding, kernel, and layer dimensions were determined based on reducing dimensionality to a minimized, 1-D, fully connected softmax layer. Each architecture was comprised of six hidden, convolutional layers. While literature studies vary on optimal network size using CNNs for spectrogram analysis, 4–6 stacked hidden layers are common in spectrogram based classification, and motivated the CNN models designed in this work. Models with fewer than six layers suffered from convergence issues and are not considered further here. Initially, square kernel sizes varying between (5,5) and (3,3) were implemented for CNN architecture #1, and were varied empirically for subsequent architectures. Hyperparameters were identified empirically, using the AlexNet model as a baseline for exploring variations. The training rate for stochastic gradient descent with momentum (SGDM) was held to 0.001. The maximum epoch number was held at 1000 while minibatch size was kept constant at 64 samples. Max pooling layers were intentionally omitted with all tested architectures. While the down-sampling procedure provided feature translational invariance, the reduction in parameters also decreased the potentially learnable feature space and degraded performance. These behaviors were discussed in Sabour et al., Zhou et al. and Yu et al. To reduce the computation time, a stride corresponding to five temporal-frequency bins, or 0.264 s, was specified. While a CNN may benefit from over-fitting prevention methods such as dropout, these steps were omitted for simplicity; as stated in the future work section, such augmentations require a separate and thorough study. The complete list of CNN parameters is presented in Tables 1 to 3.

Architecture #1. The first tested architecture was based on CNNs developed for image processing, with square kernels applied to the asymmetric spectrogram inputs. Smaller initial kernel sizes (5 × 5) were used to understand the effects of rectangular kernel dimensions. While smaller than ImageNet (11 × 11), and GoogleNet (7 × 7) first layer kernel dimensions, the architecture was generally representative of smaller, local features embedded in the spectrogram’s time-spectral relationship.

Architecture #2. For the second architecture, a 9 × 7 input kernel was chosen due to the minimum spacing between features in an observed “cupping” phenomena present in x-axis motions (Figure 2(b)). This was considered the smallest asymmetric kernel dimension to completely capture the visual signature in training data.

Architecture #3. This architecture demonstrates the highest asymmetry tested. An initial 25 × 5 kernel dimension with large vertical strides attempted to capture more frequency dependence among temporal-frequency bins. However, the horizontal component of kernel filters should capture some temporal harmonic content.
The outputs of the softmax network layer provided the basis for classification. Overlapping segmentations shared between classes was potentially problematic. To handle class bounding box overlap between \( x \)- and \( y \)-actuations, any overlapping boundary times were averaged. While a simplified approach, the method allowed for a separation of independent segmentations. Other bounding box methods such as fuzzy classification are possible and are a potential avenue for future work.

**Unsupervised analysis**

**Feature engineering.** Once segmented into primary motion classes via CNN, the corresponding audio waveforms are further discriminated through feature engineering and unsupervised analysis. Derived from corresponding audio waveforms with varying lengths, features are calculated from time, statistical, and spectral domains. These features include peak amplitude, average amplitude, mean square, root-mean square, zero-crossing rate, variance, standard deviation, kurtosis, crest factor, skewness, and \( k \)-factor.

Frequency data between 4 and 24 kHz was empirically known to contain the most useful harmonic information to discriminate between actuation classes. However, comparing peak frequency content became immediately problematic due to motion complexity and diffuse noise concerns. As shown in Figure 5, relevant peak finding in the presence of wideband noise is almost impossible due to the variation in spectral content between sub-class samples. While optimal filtering

**Table 2. CNN architecture # 2: asymmetric convolutional kernels designed to captured known acoustic phenomena.**

| Layer | Type | Dimension | Depth | Stride (h,w) | Padding (h,w) |
|-------|------|-----------|-------|--------------|---------------|
| 1     | Input | 4097 × 25 |       |              |               |
| 2     | Conv  | 11 × 9    | 50    | (2,1)        | (3,4)         |
| 3     | ReLU  |           |       |              |               |
| 4     | Conv  | 9 × 7     | 50    | (2,1)        | (1,3)         |
| 5     | ReLU  |           |       |              |               |
| 6     | Conv  | 7 × 5     | 50    | (2,2)        | (1,2)         |
| 7     | ReLU  |           |       |              |               |
| 8     | Conv  | 5 × 3     | 75    | (2,2)        | (0,1)         |
| 9     | ReLU  |           |       |              |               |
| 10    | Conv  | 3 × 3     | 100   | (2,3)        | (1,1)         |
| 11    | ReLU  |           |       |              |               |
| 12    | Conv  | 3 × 3     | 100   | (2,3)        | (0,0)         |
| 13    | ReLU  |           |       |              |               |
| 14    | Fully connected | |       |              |               |
| 15    | Softmax | |       |              |               |
| 16    | Classification | |       |              |               |

**Table 3. CNN architecture # 3: asymmetric convolutional kernels designed to capture wideband acoustic phenomena.**

| Layer | Type | Dimension | Depth | Stride (h,w) | Padding (h,w) |
|-------|------|-----------|-------|--------------|---------------|
| 1     | Input | 4097 × 25 |       |              |               |
| 2     | Conv  | 25 × 5    | 50    | (2,1)        | (4,2)         |
| 3     | ReLU  |           |       |              |               |
| 4     | Conv  | 15 × 3    | 50    | (2,1)        | (3,2)         |
| 5     | ReLU  |           |       |              |               |
| 6     | Conv  | 5 × 3     | 75    | (2,2)        | (1,1)         |
| 7     | ReLU  |           |       |              |               |
| 8     | Conv  | 3 × 3     | 100   | (2,2)        | (0,0)         |
| 9     | ReLU  |           |       |              |               |
| 10    | Conv  | 3 × 3     | 100   | (3,2)        | (0,0)         |
| 11    | ReLU  |           |       |              |               |
| 12    | Conv  | 3 × 3     | 100   | (2,1)        | (0,0)         |
| 13    | ReLU  |           |       |              |               |
| 14    | Fully connected | |       |              |               |
| 15    | Softmax | |       |              |               |
| 16    | Classification | |       |              |               |
techniques would require further, comprehensive study, a widely adopted acoustic smoothing technique was implemented.\cite{44,45} Savitzky-Golay filtering, with a 3rd order model and 101 regressive points, was used to filter spectral content. The least-squares smoothing operation was chosen due to its application in retaining peak harmonic content and higher frequency dynamics with high noise concerns.\cite{46} While not rejecting as much noise as other autoregressive models, the method relatively retains peak widths adequately for this exploratory study.

Peak estimates were difficult to directly compare due to spectral distribution even after filtering. As shown in Figure 6, even smoothed frequency responses demonstrated visual indicators with differing peak content including peak locations, magnitudes and peak number. Typical peak finding procedures were insufficient to capture seemingly discriminate frequency content through energy signatures. After numerous attempts, peak amplitude(s) features were intentionally omitted. Spectral moments were subsequently used as a metric for evaluation and considered more robust despite corrupting noise. Analogous to the statistical features from the time domain response, kurtosis, crest factor, k-factor, and skewness were applied on the FFT response data between 4 and 24 kHz. These new features described statistical moments in the spectral domain. The procedure demonstrated discriminating

![Figure 5](image-url)  
**Figure 5.** Frequency response and spectral distribution of a segmented acoustic waveform (Class 1a actuation) without smoothing.

![Figure 6](image-url)  
**Figure 6.** Illustration of differences in spectral responses between y-axis sub-classes after signal smoothing.
information in spectral distributions between unfiltered (Figure 5(b)) and filtered (Figure 7(a)) segmentations. General differences between two subgroups’ spectral content can be captured from these filtered distributions; Figure 7(a) and (b) provides the reader with a visual representation of these differences. A final concatenated feature set is shown in Table 4. Principal components were then computed to find features of maximum variance for clustering and similarity analysis. Principal components were chosen to represent at least 95% of the variability present in the feature set.

After feature extraction, several unsupervised methods were evaluated for assessing similarity between segmented sub-classes. Seven y-axis and eight x-axis actuation sub-classes were observed in the labeled ground-truth dataset. The cosine similarity, or the inner product between sub-class feature spaces, was calculated as a metric invariant to the dimensionality of the feature space (equation (1)).

$$\theta_{\mathbf{x}_i, \mathbf{x}_j} = \arccos\left(\frac{\mathbf{\sigma}_i \cdot \mathbf{\sigma}_j}{\|\mathbf{\sigma}_i\| \|\mathbf{\sigma}_j\|}\right)$$  \hspace{1cm} (1)

The parameter $\theta_{\mathbf{x}_i, \mathbf{x}_j}$ describes the angle between feature space vectors from segmentations $i$ and $j$, respectively. Plotting these cosine-similarities allows for a heat-map visualization between feature vectors from each segmentation.

K-means clustering was then used to further delineate primary motion segmentations based on the extracted feature set. While other works discuss the k-means algorithm in depth, the clustering procedure attempts to minimize distances between translatable centroid locations for a set of feature vectors describing a range of actuation events. This distance is computed as:

$$D(\mathbf{x}_i, \mathbf{x}_c) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_c}{\sqrt{(\mathbf{x}_i \cdot \mathbf{x}_i)(\mathbf{x}_c \cdot \mathbf{x}_c)}}$$  \hspace{1cm} (2)

The angular distance, $D(\mathbf{x}_i, \mathbf{x}_c)$, between the feature vector and the iteratively computed centroid clusters describes one minus the angle between the feature vector $\mathbf{x}_i$ and the centroid $\mathbf{x}_c$. Cluster membership is assigned based on minimum cluster distance.

### Results

**Supervised classification**

Due to the sliding window nature of the spectrogram inputs, a classification accuracy metric based on bounding box area was employed, based on the work in Fanioudakis and Potamitis. Overlapping false-
positive and false-negative areas were calculated from differences between labeled actuation locations in a validation spectrogram and CNN classification results. Waveform segments shared by known actuation durations and classification were recorded as overlapping area percentage. Actuation data not captured by classification resulted in a false-negative area percentage. Audio segments incorrectly classified as actuations resulted in a false-positive area percentage. The results from each architecture were presented in Tables 5 and 6.

Tables 5 and 6 highlight architecture differences in discerning x-axis and y-axis motions. Architecture #3 demonstrated the highest overlapping area percentage and F1 score in y-axis classification, while architecture #2 had the lowest classification accuracy for y-axis motions. Architecture #2 falsely classified 8.7% of spectrogram data as y-axis motions while missing nearly 45% of known actuation durations. Architecture #3 additionally held the lowest false positive and false negative area during y-axis classification. Results for x-axis motions differed slightly. While architecture #1 was still clearly inferior, architecture #2 showed improved performance in capturing x-axis actuations. However, the increase in accuracy was accompanied by the highest false positive rate. Architecture #3 again had the highest F1 score.

The results suggest that classification accuracy is somewhat dependent on kernel dimensions. Highly asymmetric spectrogram input kernels demonstrated the best compromise in accuracy for y and x-axis classification. The result in Table 5 suggests that harmonic complexity in y-axis motions was better captured with skewed kernels with dominating frequency content. Smaller kernel dimensions demonstrated a lower classification accuracy with respect to discriminating known y-axis actuations. However, this trend is not followed with x-axis actuations – harmonic content may be captured with smaller kernel dimensions.

Figure 8 describes the difficulty in delineating feature edges with bounding box methods. While the method cleanly identified 87.1% of y-axis motions (architecture #3), y-axis segmentations corrupted by noise artifacts degraded classification accuracy, particularly near the boundaries of actuation initiation and termination.

The trained CNN architectures had difficulty distinguishing differences between x-axis motions from y-axis motions at the boundaries of event transitions. This may be due to several phenomena including a similar feature-space for both events. This may have led to segmentation bias during y–x transitions; spectrograms with overlapping x/y transitions were biased toward a probability of y-axis motions (architecture #3) or toward x-axis motions (architecture #2). Sharp, wideband noise arising from neighboring mechanical phenomena during production processes were observed in both x- and y-axis segmentations. These artifacts were observed to degrade classifier accuracy as well, as was to be expected.

The results suggest a supervised approach to actuation classification from acoustic signals is feasible, given sufficient training data and accounting for the necessary asymmetry of the convolutional layers. An exhaustive hyperparameter search to determine optimal convolutional kernels and network topology would ultimately lead to increased classification accuracy than what is presented here. Unforeseen noise artifacts and other

| Table 5. Validation on ground truth spectrogram (y-axis motions). |
|-----------------|-----------------|-----------------|-----------------|
| Architecture   | Overlapping area (%) | False-pos. area (%) | False-neg. area (%) | F1 score |
| 1               | 67.8             | 1.4              | 6.4              | 0.946    |
| 2               | 55.3             | 2.5              | 8.7              | 0.908    |
| 3               | 87.1             | 1.2              | 2.6              | 0.979    |

| Table 6. Validation on ground truth spectrogram (x-axis motions). |
|-----------------|-----------------|-----------------|-----------------|
| Architecture   | Overlapping area (%) | False pos. area (%) | False neg. area (%) | F1 score |
| 1               | 67.3             | 12.0             | 12.3             | 0.840    |
| 2               | 91.4             | 20.2             | 3.5              | 0.885    |
| 3               | 84.2             | 14.6             | 6.3              | 0.890    |

Figure 8. Example of poor classification around actuation transitions with wideband noise.
factors not experienced in the CNN training set may have additionally contributed to an architecture not fully representative of SCARA sub-classes.

Due to its more consistent performance, CNN architecture #3 was used for the analysis of the unsupervised methodology. The trade-off in x-axis classification overlap was acceptable given the higher y-axis classification accuracy along with lower false positive and false negative percentages among certain classification tests.

**Unsupervised clustering**

**Analysis of y-axis motions.** Cosine similarity analysis was conducted between each segmentation for y-axis motions, illustrated as a heat-map visualization of similarity in Figure 9. A numeric score of zero (dark blue color) indicates perfect similarity. As dissimilarity between segmentations increases, the numeric similarity score increases toward 1 and red shading.

An evaluation of this similarity yields several insights. Actuation sub-classes 1(a) and 1(b) are mechanically similar and dissimilar from other actuations, and this is clearly evidenced in the similarity results. Actuation sub-class 2 is relatively similar to other motions in its category; however, it shares some similarity with subclass 4. Subclass actuation 5 is distinctly dissimilar from the other actuation sub-classes. Subclass 3(a) and 3(b) are not clearly distinguished through cosine similarity. Notably, several outliers are present in each subclass, potentially due to incorrect segmentations or noise-corrupted results.

After similarity analysis, k-means clustering was employed to separate primary classes, as a point of comparison with the cosine similarity results. As discussed previously, principal component analysis was conducted on segmentation features before applying k-means clustering, with six principal components comprising 96.2% variability in the feature set. The first and second principal components were subsequently plotted with respect to their corresponding subclass labels in Figure 10.

When looking at a three cluster segmentation, certain actuation sub-classes were naturally clustered with their inherent segmentation labels. While perfectly separated clusters were not observed, certain trends became evident. Figure 11 describes the qualitative assessment of clusters. Firstly, actuations 1(a) and 1(b) shared a cluster. Corresponding with the cosine similarity visualization, these subclass actuations share similar feature
spaces. Furthermore, these subclasses shared relatively dissimilar principal component spaces compared to groups 2–5. Sub-class 5 was partially separable from other subclasses. Other subclasses such as 3 and 5 are relatively tightly clustered with respect to their neighboring segmentations. Importantly, it was observed that visually separable clusters may be impossible with the chosen engineered feature set, reinforcing the results of the cosine similarity analysis.

Increasing the number of clusters to 7, the number of known actuation sub-classes, did not improve results (Figure 12). Because some motion subclasses share similar principal component space (and are mechanically similar as well), it was not possible to separate them.

As with the similarity analysis, the mixed accuracy suggests other contributing factors. Noise artifacts may have contributed to potential feature set outliers. Both PCA and k-means are sensitive to noise, suggesting a more robust algorithm could better reject variations present in segmentation. Other feature sets as well as fuzzier classification attempts (for instance a Gaussian mixture model) may address this issue. However, the results show the potential to isolate some key actuation sub-classes in an unsupervised regime.

Analysis of x-axis motions. X-axis segmentations were subsequently analyzed (Figure 13). Compared to y-axis segmentations, most x-axis segmentations were not as easily separable into the eight known actuation subclasses. The majority of blue-shaded regions and high similarity scores suggests that the x-axis motion features do not contain enough discriminating information. Considering that they are lower power events and manifest in a smaller frequency band than the y-axis motions, this performance was anticipated. Some outlier segmentations were identified due to their higher dissimilarity scores, particularly in sub-classes 1 and 2.

Similar to the y-axis clustering, x-axis actuations were grouped using the k-means algorithm with angular distance metrics. The two highest principal components, depicted in Figure 14, depict the general inseparability of these segmentations, as well as the fact that X-actuations are predominately determined by a dominant principal component. Clustering x-axis segmentations reflect the inseparability of extracted features leading to over-segmentation, see Figure 15.

Sources of error

Several potential error sources were noted which may have negatively impacted results. Without an exhaustive hyperparameter study, sub-optimal neural network architectures may have contributed to reduced classification accuracy. Furthermore, deep-learning techniques are notoriously data hungry – and the relatively limited training set may have unintentionally contributed to segmentation error. Omitting overfitting concerns, unintentional training set bias was possible. The proposed methodology is also dependent on the bounding box approach to handling overlapping segmentations, which could potentially be improved. Noise in the acoustic recordings also likely impacted classification results. Lastly, user-defined ground-truth labels may have errors, which would have explained some of the strong outliers within the x-axis similarity analysis.

Discussion

This work presents three potential CNN architectures to explore spectrogram classification; however,
numerous architectures are possible or more optimal, particularly if a larger set of training data is available. It is worth emphasizing that the rapid advancement of deep neural network research makes determining an optimal architecture difficult. Rather, the goal here is to illustrate a generalized approach that can be adopted and optimized as the field of research evolves. The architectures introduced demonstrated adequate performance, and more importantly, feasibility. However, no claim to computational optimality is made here. A more in depth study on the impacts of asymmetric convolutional kernels is also needed. A fuzzier segmentation algorithm may also be required in high noise spectrogram cases where class labels are ambiguous, or where actuations occur nearly simultaneously. Bounding boxes may have introduced unintentional bias between neighboring y- and x-axis segmentations and the results suggest refining this aspect of the approach. Non-stationary noise present in the manufacturing environment tended to corrupt spectral features significantly, including peak definition and overall energy. Optimal filtering may present one potential improvement, and smoothing the FFT responses yielded an empirical improvement for other spectral features such as spectral moments.

Further refinements in feature generation are similarly required. The selected feature set was based on features sets that have had demonstrated success in other applications, but their use may require additional work for domain specific implementation. Automatic and more nonlinear feature set generation approaches could be employed, such as those provided by a convolutional autoencoder.

The presented approach has a few key limitations worth noting. The CNN architectures presented here were designed for spectrograms with heavily skewed aspect ratios (4097 × 25). Spectrogram inputs that approach a 1:1 aspect ratio may lead to dramatically different system behavior or CNN architectures. Additionally, the unsupervised analysis performed better on signals that manifested across a wide band of frequencies due to the featurization methods employed. This is best shown through the illustrative examples, where the performance for y-axis motions was shown to be superior when compared to x-axis motions. Narrow band signal manifestations would likely require additional signal processing or different featurization methods.

**Conclusion**

This study demonstrated that a combination of deep supervised and unsupervised machine learning techniques could identify and classify robotic sound events in a realistic manufacturing environment. A supervised convolutional neural network architecture was developed to isolate primary actuation instances from an arbitrary length spectrogram. Using ground-truth, labeled spectrogram data, 87.1% of y-axis and 84.2% of x-axis motions were successfully segmented. Asymmetric convolutional kernels yielded the most consistent results.

Once the primary actuations were segmented, an unsupervised similarity analysis based on cosine similarity or k-means clustering was able to distinguish primary y-axis actuations into smaller sub-classes, though this was not possible for low-power x-axis motions. Despite limitations, including challenges with mechanically similar actuation sub-classes, noise, and possible mislabeled data, this approach showed reasonable efficacy.
Future work will look at how to improve the feature selection process through the use of an autoencoder. This will be followed by a study on how this process can be used to track and quantify progressive wear and deterioration in a mechanical system that cannot be detected through conventional outlier-based damage detection strategies.

Acknowledgements
The authors would also like to thank members of the Lattanzi Research Group for their contribution and input to this project.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Portions of this work were supported by a grant from the Office of Naval Research (No. N00014-18-1-2014). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

ORCID iD
David Lattanzi https://orcid.org/0000-0001-9247-0680

References
1. Liu R, Yang B, Zio E, et al. Artificial intelligence for fault diagnosis of rotating machinery: a review. Mech Syst Signal Process 2018; 108: 33–57.
2. Zhao R, Yan R, Chen Z, et al. Deep learning and its applications to machine health monitoring. Mech Syst Signal Process 2018; 115: 213–237.
3. Khan S and Yairi T. A review on the application of deep learning in system health management. Mech Syst Signal Process 2018; 107: 241–265.
4. Kim JI, Kim D, Krebs M, et al. Force sensitive robotic end-effector using embedded fiber optics and deep learning characterization for dexterous remote manipulation. IEEE Robot Automat Lett 2019; 4(4): 3481–3488.
5. She Q, Feng F, Hao X, et al. Openloris-object: a robotic vision dataset and benchmark for lifelong deep learning. In: 2020 IEEE international conference on robotics and automation (ICRA), 31 May–31 August 2020, pp 4767–4773. New York: IEEE.
6. Wen Y, Si J, Brandt A, et al. Online reinforcement learning control for the personalization of a robotic knee prosthesis. IEEE Trans Cybernet 2019; 50(6): 2346–2356.
7. Singh A, Yang L, Hartikainen K, et al. End-to-end robotic reinforcement learning without reward engineering. arXiv preprint 190407854, 2019.
8. Yan R and Gao RX. Multi-scale enveloping spectrogram for vibration analysis in bearing defect diagnosis. Tribol Int 2009; 42(2): 293–302.
9. Wu Y, Mao H and Yi Z. Audio classification using attention-augmented convolutional neural network. Knowl Based Syst 2018; 161: 90–100.
10. Dennis J. Sound event recognition in unstructured environments using spectrogram image processing. PhD Thesis. Singapore: School of Computer Engineering, Nanyang Technological University, 2014.
11. Zahorian SA and Nossair ZB. A partitioned neural network approach for vowel classification using smoothed time/frequency features. IEEE Trans Speech Audio Process 1999; 7(4): 414–425.
12. Mao W, Feng W and Liang X. A novel deep output kernel learning method for bearing fault structural diagnosis. Mech Syst Signal Process 2019; 117: 293–318.
13. Alias F, Socoró JC and Sevillano X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl Sci 2016; 6(5): 143.
14. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. CUIDADO IST Project Report, 2004, pp. 1–25.
15. Bach J, Anemüller J and Kollmeier B. Robust speech detection in real acoustic backgrounds with perceptually motivated features. Speech Commun 2011; 53: 690–706.
16. Aldrich C. Unsupervised process monitoring and fault diagnosis with machine learning methods. London: Springer, 2013.
17. Sohn H and Farrar CR. Damage diagnosis using time series analysis of vibration signals. Smart Mater Struct 2001; 10(3): 446–451.
18. Lei Y, Jia F, Lin J, et al. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Trans Ind Electron 2016; 63(5): 3137–3147.
19. Zhang H, McLoughlin I and Song Y. Robust sound event recognition using convolutional neural networks. In: Acoustics, speech and signal processing (ICASSP), 2015 IEEE international conference, Brisbane, Australia, 19–24 April 2015, pp.559–563. IEEE.
20. Boddapati V, Petef A, Rasmusson J, et al. Classifying environmental sounds using image recognition networks. Proc Comput Sci 2017; 112: 2048–2056.
21. Faniodakis L and Potamitis I. Deep networks tag the location of bird vocalisations on audio spectrograms. arXiv preprint, 2017.
22. McLoughlin I, Zhang H, Xie Z, et al. Robust sound event classification using deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2015; 23(3): 540–552.
23. Piczak KJ. Environmental sound classification with convolutional neural networks. In: IEEE 25th international workshop on machine learning for signal processing (MLSP), Boston, MA, 17–20 September 2015, pp.1–6. IEEE.
24. Cooper ML and Foote J. Robust sound event classification using deep neural networks. In: Automatic music summarization via similarity analysis in ISMIR, Paris, France, 2002.
25. Cooper M and Foote J. Summarizing popular music via structural similarity analysis. In: *Applications of signal processing to audio and acoustics: IEEE workshop*, New Paltz, New York, 19–22 October 2003, pp.127–130. IEEE.

26. Jang G, Kim HG and Oh YH. Audio source separation using a deep autoencoder. arXiv preprint, 2014.

27. Marchegiani L and Posner I. Leveraging the urban soundscape: auditory perception for smart vehicles. In: *2017 IEEE international conference on robotics and automation (ICRA)*, Singapore, 29 May–3 June 2017, pp. 6547–6554. IEEE.

28. Nanni L, Rigo A, Lumini A, et al. Spectrogram classification using a deep autoencoder. *Advances in Mechanical Engineering* 2019; 7: 7717–7727.

29. Sajjad M, Kwon S and Mustaqeem. Clustering-based speech emotion recognition using k-means clustering and spectrograms. *Entropy* 2019; 21(5): 479.

30. Hajarolasvadi N and Demirel H. 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. *Entropy* 2020; 8: 79861–79875.

31. Merchan F, Echevers G, Poveda H, et al. Spectrogram classification using dissimilarity space. *Appl Sci* 2020; 10(12): 4176.

32. Mathworks. Alexnet convolutional neural network. *IEEE Trans Multi-Signal Multi-System* 2018; 20(3): 513–524.

33. Mathia K. *Robotics for electronics manufacturing*. Cambridge: Cambridge University Press, 2010.

34. Takahashi N, Gygli M and Van Gool L. Aenet: learning deep audio features for video analysis. *IEEE Trans Multimedia* 2018; 20(3): 513–524.

35. Wang C, Santoso A, Mathulaprangsan S, et al. Recognition and retrieval of sound events using sparse coding convolutional neural network. In: *2017 IEEE international conference on multimedia and expo (ICME)*, Hong Kong, 10–14 July 2017, pp.589–594. IEEE.

36. Monteiro R, Bastos-Filho C, Cerrada M, et al. Convolutional neural networks using fourier transform spectrogram to classify the severity of gear tooth breakage. In: *2018 International conference on sensing, diagnostics, prognostics, and control (SDPC)*, Xian, China, 15–17 August 2018, pp.490–496. IEEE.

37. Nayyar RK, Nair S, Patil O, et al. Content-based autotagging of audios using deep learning. In: *2017 International conference on big data, IoT and data science (BID)*, Pune, India, 20–22 December 2017, pp.30–36. IEEE.

38. Khamparia A, Gupta D, Nguyen NG, et al. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access* 2019; 7: 7717–7727.

39. Ren Z, Qian K, Wang Y, et al. Deep scalogram representations for acoustic scene classification. *IEEE/CAA J Automat Sin* 2018; 5(3): 662–669.

40. Sabour S, Frost N and Hinton GE. Dynamic routing between capsules. In Guyon I, Luxburg UV, Bengio S, et al. (eds.) *Advances in neural information processing systems 30*. Long Beach, CA, Neural Information Processing Systems Foundation, Curran Associates, Inc., 2017. pp. 3856–3866.

41. Zhou J, Xu W and Chellali R. Analysing the effects of pooling combinations on invariance to position and deformation in convolutional neural networks. In: *2017 IEEE international conference on cyber and bionic systems (CBS)*, Beijing, China, 17–19 October 2017, pp.226–230. IEEE.

42. Yu D, Wang H, Chen P, et al. Mixed pooling for convolutional neural networks. In: *International conference on rough sets and knowledge technology*, Beijing, China, 24–26 October 2014, pp.364–375.

43. Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: a review. *Neurocomputing* 2016; 187: 27–48.

44. Zhang D, Entezami M, Stewart E, et al. Adaptive fault feature extraction from wayside acoustic signals from train bearings. *J Sound Vibrat* 2018; 425: 221–238.

45. De Oliveira MA, Araujo NVS, Da Silva RN, et al. Use of savitzky–golay filter for performances improvement of shm systems based on neural networks and distributed pzt sensors. *Sensors* 2018; 18(1): 152.

46. Orfanidis SJ. *Introduction to signal processing*. Hoboken, NJ: Pearson Education, Inc., 1996–2009.

47. Park S, Lee JJ, Yun CB, et al. Electro-mechanical impedance-based wireless structural health monitoring using pca-data compression and k-means clustering algorithms. *J Intellig Mater Syst Struct* 2008; 19(4): 509–520.

48. Banerjee S, Choudhary A and Pal S. Empirical evaluation of k-means, bisecting k-means, fuzzy c-means and genetic k-means clustering algorithms. In: *2015 IEEE international WIE conference on electrical and computer engineering (WIECON-ECE)*, Dhaka, Bangladesh, pp.168–172. IEEE.

49. Alamdari MM, Rakotoarivelovabe and Khoa NLD. A spectral-based clustering for structural health monitoring of the sydney harbour bridge. *Mech Syst Signal Process* 2017; 87: 384–400.