International Conference on Smart Infrastructure and Construction 2019 (ICSIC):
Driving data-informed decision-making

ice Publishing

DeJong, Schooling and Viggiani
ISBN 978-0-7277-6466-9
https://doi.org/10.1680/icsic.64669.351

CSIC Cambridge Centre for Smart Infrastructure & Construction

A CONVOLUTIONAL NEURAL NETWORK APPROACH TO THE SEMI-SUPERVISED ACOUSTIC MONITORING OF INDUSTRIAL FACILITIES

*J. Bynum, G. Earle and D. Lattanzi**

George Mason University, 4400 University Drive, Fairfax, Virginia, USA
**corresponding author*

ABSTRACT Industrial manufacturing facilities require autonomous damage detection and monitoring to maintain functionality and reduce maintenance costs, particularly when human intervention is costly or dangerous. And as robotic manufacturing expands, the need for autonomous process monitoring has consequently expanded as well. In an industrial scenario, damage is typically incurred through machine wear from power transmission sources (e.g. bearing, gear, motor, belt, rail/track, and material wear), and performance degradations may occur nearly instantaneously, or slowly over time. The consequences of such damage can be localized to a single mechanical system, or it can cascade into catastrophic damage across a facility. In this work, a nondestructive method for identify and tracking processes and events within a manufacturing facility is presented, based on analysis of an autonomous acoustic monitoring system. The approach employs a deep convolutional neural network in combination with unsupervised similarity analysis to identify and track industrial processes based on their acoustic signatures within an image-like spectrogram. The approach is designed for flexibility and extensibility to a range of industrial scenarios, and requires only limited labeling of training data. The results of experimental testing indicate that the approach is capable of properly segmenting and tracking manufacturing processes manifested in acoustic signals across a range of spatial and frequency scales, and is capable of handling temporal distortions. Future work on fully unsupervised approaches are discussed as well.

1. Introduction

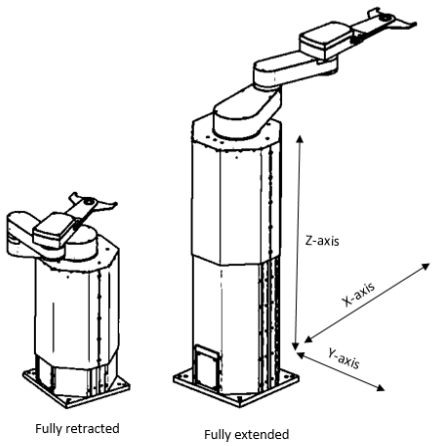
Stemming from production tolerances as well as extended operation requirements, manufacturing equipment must be persistently monitored and maintained. Conventional approaches rely on operator judgement; however, remote monitoring techniques are now increasingly applied. Semiconductor manufacturing facilities exemplify the need for remote monitoring procedures due to the inaccessibility of many of the related fabrication environments.

Robotic systems, such as SCARA-series (Selective Compliance Assembly Robot Arm) devices as shown in Figure 1, are employed in electronics manufacturing for material transitioning (Mathia, 2010). Primary SCARA robotic actuations can be generalized through a series of 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.

Primary motion classes are further subdivided into secondary classes, depending on processes. For example, a variety of actions including material loading and transition are possible, often combining sequential actuations. These actuations are also varying length as dictated by the specific process.

The underlying concept that motivated this research effort is that primary motion classes should have unique audio signatures that could be distinguished through pattern analysis algorithms such as machine learning classifiers. Once identified, processes could be tracked and monitored for wear and degradation, providing the basis for in-service remote monitoring.

Figure 1 SCARA and Cartesian reference frame used (Mathia, 2010)



Reprinted with permission from Cambridge Publishing

Rather than direct analysis of the time-series waveforms, audio signals were transformed into spectrograms and treated as pseudo-images. Analogous to pixels in images, the concept of an 'auxel' was used to represent the intensity value of a spectrogram at a specific time-frequency location. This image-like data format enabled representation of events with a

diverse range of spatial features, such as edges. Visually definable features in spectrograms, depicted as high-energy lines in the frequency domain, are comparable to object edges in images.

Figures 2 and 3 depict spectrogram representations of two primary motion classes. Figure 2 illustrates frequency content over time during a single y-axis motion. Figure 3 highlights visually definable, high-frequency features inherent to x-axis motion events. Visual indicators of z-axis motion actuations were not detectable within the spectrograms despite initial signal processing efforts to rectify this problem. Wide spectrum environmental noise, along with concurrent actuation (e.g. z-axis motion during y-axis travel), reduced potentially observable spectrogram features. As such, z-axis motions were not explicitly considered in this study due to their lower energy content.

Figure 2 Y-axis process actuation spectrogram

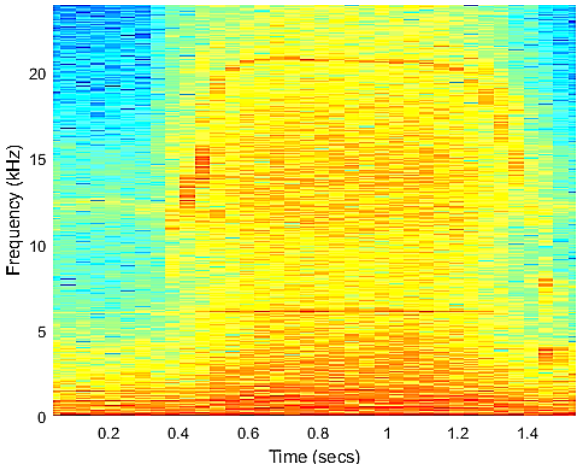
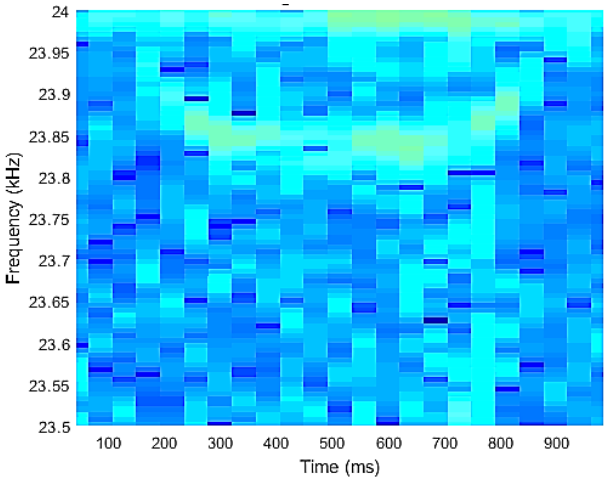


Figure 3 X-axis actuation spectrogram



2. Literature Survey

Surveying relevant literature, several important concepts emerge. Audio-based machine learning studies can be broadly separated into defined audio structures (e.g. music, speech, animal voicings) and chaotic audio structures (e.g. environmental or traffic noise). Deep neural network architectures are often employed due to complex feature spaces with different windowing procedures (Alías et al, 2016). Spectrogram analogies to images are commonly observed across an array of published work. Particularly relevant work that was relied upon on for development of preliminary methodologies is summarized below.

Alías et al. provided a broad literature review for audio-based, engineered feature techniques (2016). The authors documented sensing strategies for audio-based analysis under different environmental and test conditions. Windowing strategies, including a generic procedure for a sliding-window feature extraction algorithm, were presented. The authors documented how windowing and subsequent features were usually based on the minimum temporal scale of domain-specific events.

Environmental sound-event recognition contains different challenges compared to traditional speech recognition and audio classification tasks (Dennis, 2014). Prior knowledge for feature selection is often required due to the broad domain space in acoustic recognition research. Typical features in speech processing literature such as Mel-frequency Cepstrum Coefficients (MFCCs) can underperform in certain unstructured audio classification tasks; lower level spectral-temporal features, as noted by the author, can provide discriminating information between audio events with a lower dimensionality. Spectrogram based audio event classification, using image-based processing, has demonstrated high accuracy and robust feature identification. The author additionally notes challenges adapting novel audio classification methods to physical systems containing non-stationary noise and concurrent audio sources.

Yu et al. described the usefulness of spectrograms for classifying audio data (2018). Low-level audio features based on prior knowledge are often not fully representative since potentially discriminative information can be lost through engineered bias and preprocessing. Spectrograms contains a large amount of initial information present in audio data, while remaining discriminative and generalizable to other spectral-based problems.

Classification procedures were additionally dependent on temporal scale. The case for dimensionality reduction was outlined due to an enormous potential list of audio features. Importantly, the paper also introduced the concept of differentiation between audio structures. Speech and music audio sources contained mostly defined, periodic structures compared to environmental sounds. These complicated structures require more advanced techniques to extract relevant audial data, effectively. Environmental sound classification, more closely related to the problem domain in this report, had specific processing considerations, also discussed in (Piczak, 2014).

Specific concerns and justifications for using convolutional neural networks (CNN) in urban (environmental) sound classification were presented in (Piczak, 2014). The paper introduced a number of potential decisions and parameters required for spectrogram driven CNN as well as training data creation. The author briefly discussed the generation of training data of spectrograms of events; audio data split into overlapping segments could provide adequate information for training. Frequency invariance was adjusted based on kernel dimensionality.

Acoustic event detection, or classification of spectrogram features present in non-speech signals, was presented in (Espi et al., 2014). The work outlined how relevant features detected in spectrograms using convolutional architectures could outperform deep non-convolutional architectures.

Boddapati et al. described mapping existing image-based architectures to classifying audio signals (2017). The authors presented challenges to map preexisting image architectures to audio problems including the rectangular aspect ratio between time to frequency components within spectrograms. Different methods of visualization for classification were presented. A CNN/RNN based method was successfully introduced to classify non-structured audio events even under mixed source signals and noise.

Ruqiang and Gao reflected on the challenges present in acoustic emission for damage classification (2009). The authors noted discriminating information in audio-based damage classification is usually spectral based; however, these features contain lower amplitudes, week harmonics, and fewer defined spectral features. Preprocessing employed in other audio classification research can be potentially problematic due to eliminating useful frequency components not fully understood (usually high frequency bands). A novel, spectral (STFT) and wavelet-based approach introduced allowed for easier visualization in vibration-based damage signals.

Methods on multiple-stage environmental sound classification algorithms were introduced (Chachada, 2014). Authors note that environmental sound structure is fundamentally different from pattern-based audio such as speech. A useful taxonomy of audio features was diagrammed in their report. A k-nearest neighbor algorithm with a feedworward neural network architecture was used to successfully classify environmental sound sources.

Autonomous classification of audio events using neural network architectures was also studied in (Jang et al., 2014). While the published results are preliminary, a proposed audio source separation procedure using a combination of a deep network architecture and k-means clustering was employed. The study 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.

Cooper and Foote described a novel way to compare audio segments (2002). Spectral (MFCC) features were derived for windowed segments of music compositions. Termed 'self-similarity analysis,' the features at each window were compared using a cosine similarity heat-map. Different compositions were visually identifiable based on similarity metrics.

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 from temporal distortions. Furthermore, studies indicate that features, present in both spectrogram and unmapped signal domains, contain viable information relevant for audio classification. However, the use of such approaches for remote system monitoring, particularly with respect to unsupervised learning approaches, has not previously been studied.

3. Methodology

A methodology to understand processes in a semi-conductor manufacturing environment is presented. A dbx RTA-M microphone captured audio process data in a manufacturing bay containing a single SCARA-type robotic tool, as shown in Figure 1. Four-minute audio recordings containing SCARA actions spanning various semiconductor manufacturing processes are captured. An eleven-minute audio recording, aligned with corresponding video data, is used as final validation data since ground-truth labels are established.

Two acoustic analysis methodologies are developed and evaluated in this study. The first is an unsupervised technique based on prior work in relevant domains. This technique uses feature extraction in tandem with k-means clustering to identify and segment robotic motion processes directly from segmented audio data. The resulting approach was not sufficiently accurate and thus a second semi-supervised technique was explored.

This second technique uses a combination of a CNN and feature space similarity analysis as well as k-means clustering to perform process identification and analysis. Several notable data science methods were explicitly omitted in this study. Preliminary experiments indicated that isolating and correcting for wide-spectrum environmental noise was infeasible. Furthermore, extensive signal preprocessing would bias resulting approaches for the current SCARA unit, limiting extensibility to other robotic tools.

3.1 Fully unsupervised process identification approach

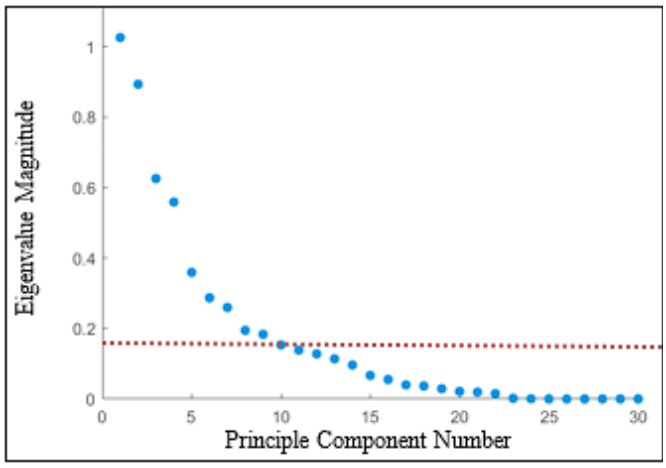
Expanding on previous research, a conventional unsupervised approach was initially developed in order to segment SCARA process motions from raw audio data. This advantageous approach supports identifying processes and sub-processes without leveraging statistical model fitting or supervised machine learning; supervised methods are susceptible to overfitting and difficult to generalize with larger and more diverse application domains. A sliding window algorithm with a varied window length and 50% overlap between adjacent

samples is first used to partition sections of the acoustic recording. The features of each window are then computed and concatenated into a feature vector, using the predetermined feature representations shown in Table 1. Dimensionality reduction is then performed on the feature vector via principal component analysis. Preliminary analysis indicated that ten principal components provided sufficient feature variance, as shown in Figure 4.

Table 1 Concatenated feature list

Engineered Features	
Peak amplitude	*Zero-crossing rate*
Average amplitude	*Skewness*
Mean square	*Kurtosis*
RMS	*Crest factor*
Variance	*K-factor*
Standard deviation	*Dominant frequency peak locations*

Figure 4 Eigenvalue magnitude versus principle component



Once the features are extracted and reduced to a lower dimensional representation, K-means clustering is used to separate and segment primary robotic motions into a specified number of clusters. Initially, two clusters were used in an attempt to coarsely segment robot actuations from idle/noise states. Additional clusters, with centroid amounts equaling potential process movements, were subsequently added to increase granularity of segmentation.

As will be shown in Section 4, the results of this approach did not provide sufficient segmentation accuracy, leading to the development of a semi-supervised segmentation approach.

3.2 Semi-supervised process identification approach

In response to the analysis of the fully unsupervised approach, a semi-supervised approach was additionally developed. Similar to relevant prior literature, a convolutional neural network (CNN) was designed to perform segmentation of the primary motion classes using image-like spectrograms. Unsupervised feature similarity analysis, based on features extracted from these segmentations, is then performed to identify secondary motions and identify process faults and irregularities.

Convolutional neural networks offer unique advantages to direct feature engineering methods. Critical feature representations are learned – rather than defined – reducing systematic bias. The approach requires a large dataset as well as a robust architecture to gain a relevant, learned feature set; however, due to complexity inherent within robotic movement data and environmental noise, conventional features may be inadequate to directly describe actuations during segmentation. *3.2.1 Dataset development:* A machine learning training dataset was created using collected audio recordings in conjunction with a synchronized video. Class labels were derived from both recorded video and from operator intuition. The reliance on training data had certain limitations. Building such datasets is time consuming, and the data set runs the risk of not properly representing all potential system states for statistical training. Second, visual indicators of lower energy motions (z-axis) were not observable in spectrogram representations and so were not included in the training set. Lastly, ground truth data for motion sub-classes was readily available fine-grained, CNN training segmentations. As a result, the large-scale training database consists of primary motion classes (x-axis, y-axis, and idle motion states) manually labeled by the authors. A smaller dataset containing labeled motion subclasses was used to evaluate the unsupervised similarity aspect of the overall approach.

Fixed signal length spectrograms were labeled as x-axis motion, y-axis motion and noise as depicted in Figures 2 and 3. Indicators for each motion class, based on visual observation, guided training spectrograms derivation. Events containing neither visual indicator are designated as noise.

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, 2014). Moreover, the CNN will realistically encounter similar, partially obscured spectrograms during segmentation in operation.

The final training set consisted of 1360, 777, 1181 training examples of noise, y-axis, and x-axis motions, respectively. Training example sizes were kept constant. For example, training a 3-class classifier CNN between y-axis, x-axis, and noise used 777 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. *3.2.2 CNN architecture development:* 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. Spectrogram normalization was first used to limit spectrogram variability. Across 3000+ test spectrograms, minimum and maximum auxel values were calculated and used as boundaries. All training and test spectrogram auxels were rescaled from a minimum bound of -13.2699 and maximum bound of 5.7783 to values between 0 – 255, making the spectrograms a more image-like data type.

After development of the labeled dataset and spectrogram normalization, an optimal CNN architecture was empirically determined through a hyperparameter study. While initial architectures were adapted from the field of image analysis (MathWorks, 2019), it was determined that these architectures were not well suited to spectrogram segmentation, and to the authors' knowledge no established CNN architecture exists for spectrogram training. A more exhaustive study was thus performed, using both classifier accuracy and convergence rates as metrics for understanding the relative merits of tested CNN architectures.

Padding, kernel, and layer dimensions were determined based on reducing dimensionality to a minimized, 1-D, fully connected softmax layer; these dimensions were iteratively changed based on minibatch accuracy as well as empirical performance. Smaller kernel sizes with lower dimensional padding sizes were shown to have better convergence than excessively large filters. An 11x11 input kernel was chosen due to the minimum spacing between auxel-based features in the "cupping" phenomena present in x-axis motions. A minibatch size of 64 and a training rate of 0.001 were additionally specified. The complete list of CNN parameters is presented in Table 2.

Table 2 CNN architecture

Layer	Type	Dimension	Filter Number	Stride (h,w)	Padding (t,b,l,r)
1	Input	4097x25x1	-	-	-
2	Conv	11x11	80	[1,1]	[5,5,5,5]
3	ReLU	-	-	-	-
4	Conv	7x7	50	[1,1]	[3,3,3,3]
5	ReLU	-	-	-	-
6	Conv	5x5	50	[2,2]	[1,1,1,1]
7	ReLU	-	-	-	-
8	Conv	3x3	100	[1,1]	[0,0,0,0]
9	ReLU	-	-	-	-
10	Fully connected	-	-	-	
11	Softmax	-	-	-	
12	Classification output	-	-	-	

3.2.3 CNN architecture development: Once segmented into primary motion classes via CNN, the corresponding audio waveforms are further discriminated through feature engineering and unsupervised analysis, similar to what was developed for the fully unsupervised approach. An augmented feature set based on the initial unsupervised feature set was used for clustering, shown in Table 3. Principal components were again computed to find features of maximum variance for clustering and similarity analysis.

Table 3 Engineered features on CNN segmentations

Features from raw data		Features from filtered FFT response
Peak amplitude	Zero-crossing rate	Skewness
Average amplitude	Skewness	Kurtosis
Mean square	Kurtosis	Crest factor
RMS	Crest factor	K-factor
Variance	K-factor	
Standard deviation		

Frequency content features were revisited to better distinguish motion sub-classes. Frequency data between 4 kHz – 24 kHz held the most useful discriminating information between motions. As shown in Figure 5, relevant peak finding in the presence of wideband noise is almost impossible due to the variation between sub-class samples. The Smoothing Fast Fourier Transform (FFT) response with Savitzky-Golay filtering yielded improvements, allowing comparative analysis between samples while reducing overall peak amplitudes, as shown in Figure 6. Additional statistical features were then extracted from this filtered FFT response data. 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 kHz – 24 kHz. These new features described the statistical distribution in the frequency domain. Adding these parameters immediately showed a variance increase between classes.

Figure 5 FFT without smoothing

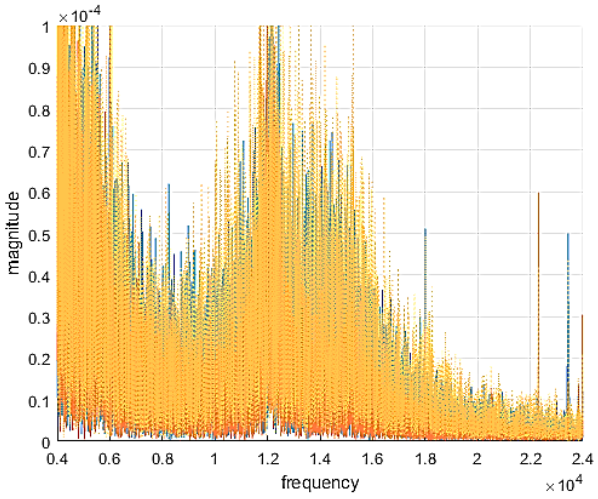
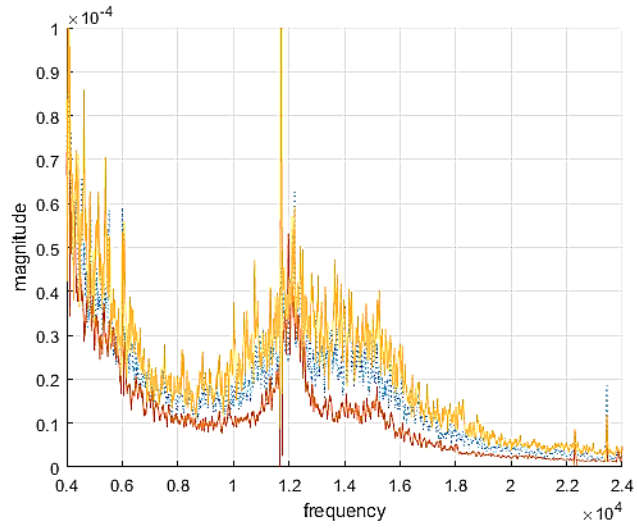


Figure 6 FFT with Savitzky-Golay smoothing



3.2.4. Similarity Analysis:

Using the features calculated in the preceeding section, the similarity between descriptor vectors can then be computed to to provide a metric for understanding separable differences between segmentations. For this work, the cosine similarity was employed. For feature vectors A and B, the angle θ_{AB} between them represents the similarity of the compared vectors in feature space (Eq. 1). As the parameter θ_{AB} diverges from 0, the feature vectors increase dissimilarity.

$$\theta_{AB} = arccos\frac{A \cdot B}{\|A\|\|B\|} \qquad (1)$$

K-means clustering can also be used to groups segmentations based on the extracted feature set.

4. Experimental analysis

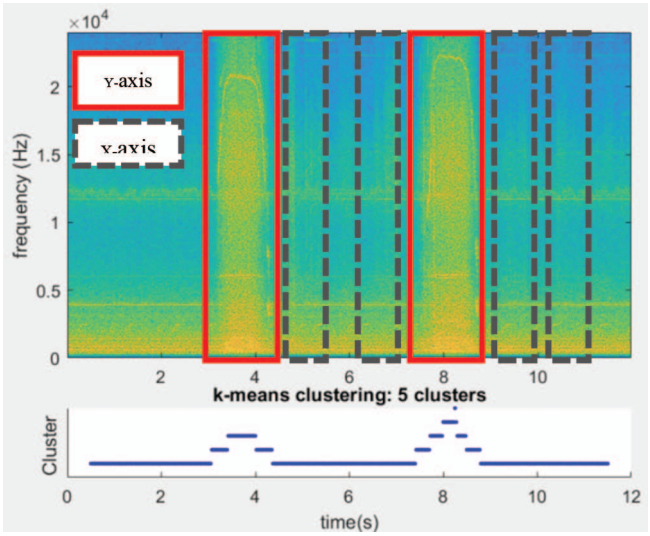
4.1 Analysis of unsupervised approach

The first set of analyses evaluated the ability to segment acoustic responses into two unsupervised clusters, ideally separating robotic motions from idle/noise states. Performance was measured based on comparisons against manual segmentations of the corresponding spectrograms. While this approach showed poor results with sliding window sizes between 0.05s and 0.5s, concatenating features from different window sizes showed improved performance. This concatenated feature set demonstrated relatively high clustering accuracy in segmenting y-axis motions from noise states (Figure 7).

While the fully unsupervised method was able to consistently distinguish y-axis motions, this approach demonstrated several limitations. The clustering method was unable to discern lower power events (such as x-, and z-axis motions) from noise. These actuation events were incorrectly binned with the noise state – effectively halting any further segmentation. Due to the relative feature signature between low signal energy and high energy events (y-axis motions), the results suggest that y-axis motion features dominated all clustering results. This effect can be seen as centroid sizes increased away from actuation/noise states in Figure 7.

Y-axis segmentations were consistently truncated at the beginning and ending (edges) of actuation (Figure 7). It is most likely a result of feature averaging when larger sliding windows are used. Rather than grouping the transitions as part of y-axis segmentations, actuation transitions were effectively averaged between lower intensity features, from noise and secondary states, and dominating y-axis motion features.

Figure 7 K-means clustering (5 centroids)



Increasing the number of clusters did not result in capturing additional motion classes. Instead, only two y-axis motions and one idle state was distinguished. Instead of capturing multiple event instances, cluster centroids separated y-axis actuation transitions (Figure 7). This result confirmed that feature bias introduced with sliding windows was improperly segmenting events.

Another key limitation regarded other spurious events such as high power environmental noise corrupting in audio samples. These wide-spectrum audio occurrences were sometimes

falsely clustered along with y-axis motions – suggesting that relative energy signatures between y-axis and wide-spectrum environmental noise events overpowered any relevant features defining lower energy actuations, such as x-axis motions.

Overall, the analysis of the unsupervised segmentation approach indicated that bias introduced by windowing procedures during feature extraction negatively impacted direct clustering of robotic motions in acoustic data. Sliding windows tended to average features between high energy and low energy actuations. Lower power (as x-, z-axis) motions were indistinguishable from noise due to the energy content present in higher power y-axis motions and spurious noise events biasing the clustering algorithm. These results suggest that a purely unsupervised segmentation and analysis process is unlikely to be successful, though unsupervised deep network approaches with autoencoder layers may prove more successful.

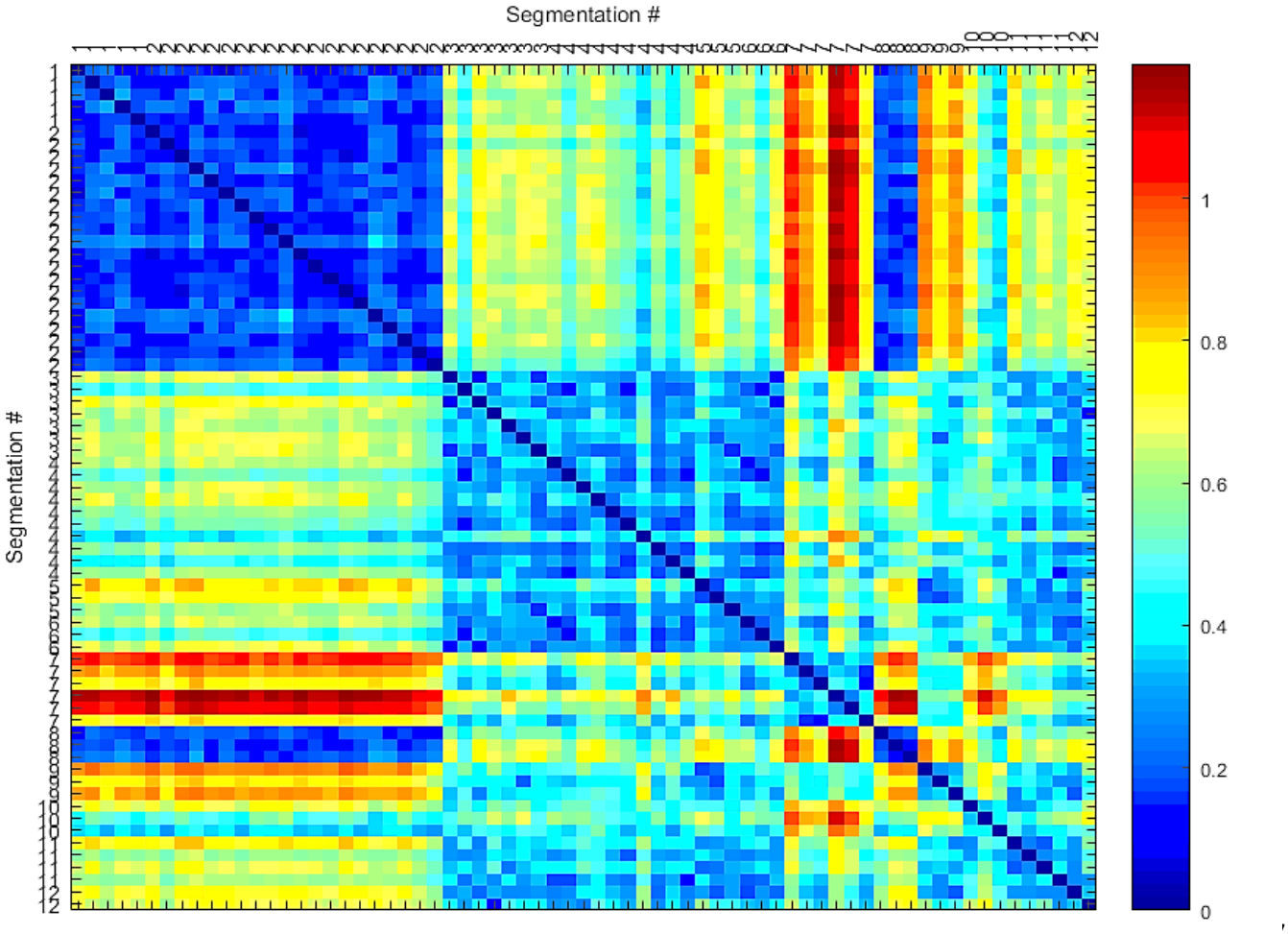
4.2 Analysis of semi-supervised approach

4.2.1 CNN performance: Using the labeled spectrogram database, a CNN was trained to perform segmentation of primary motion classes from noise. A hold-out set of ground-truth spectrograms that included both primary and secondary motion class labels were then input into the CNN for segmentation testing and evaluation. 77 y-axis motions and 148 x-axis motions were known to occur within this data set. The softmax layer within the CNN computed class probabilities. If adjacent segmentations shared overlapping bounding boxes and were of the same class, they were combined.

Three-class segmentation (y-axis, x-axis, noise) showed consistently poor classification accuracy. While y-axis motions were cleanly segmented, CNN activations of x-axis motion were seemingly biased by wide-spectrum noise rather than the distinct 'cup' discriminator observed in corresponding spectrograms. Wide-spectrum noise was consistently misclassified as x-axis motion and vice versa. Several error sources are possible including architecture bias and lack of adequate training data.

Binary segmentation of the y-axis motions resulted in notable accuracy. Compared to 77 labeled y-axis motions, 76 were properly labeled by the CNN. Four false positive and nine improper segmentations were observed. Improper segmentations were visually assessed to contain some additional high power, wide-spectrum noise for idle to y-axis and y-axis to idle actuation transitions. With removal of these outliers, this method showed roughly 87% accuracy on a ground-truth, labeled spectrogram. Some false positives can be further removed by thresholding bounding box sizes; for example, if a bounding box was smaller than an expected

Figure 8 Cosine similarity heat-map of y-axis segmentations



window size, the segmentation was ignored. *4.2.2 Unsupervised Sub-Class Segmentation Analysis:* Following the methodology laid out previously, features were extracted from each CNN segmentation. The cosine similarity was then computed between each segmentation feature vector, shown as a similarity heatmap (correlation plot) in Figure 8. Blue intensities correspond to similar feature vectors (values close to 0), with while red intensity indicates dissimilar segmentations. Segmentations were grouped with respect to their true labels for visual clarity. Sub-classes 1, 2, and 8 showed high similarity, indicated by the near-zero similarity scores. Sub-classes 7 and 10 generally had the highest dissimilarity with respect to other classes. Motion sub-classes 3, 4, 5, 6, 11, and 12 had the least definable feature sets, being similar to themselves and dissimilar to other classes. Sub-classes 5 and 9 had similarly ambiguous group definitions; however, these two groups held relatively high similarity within each group.

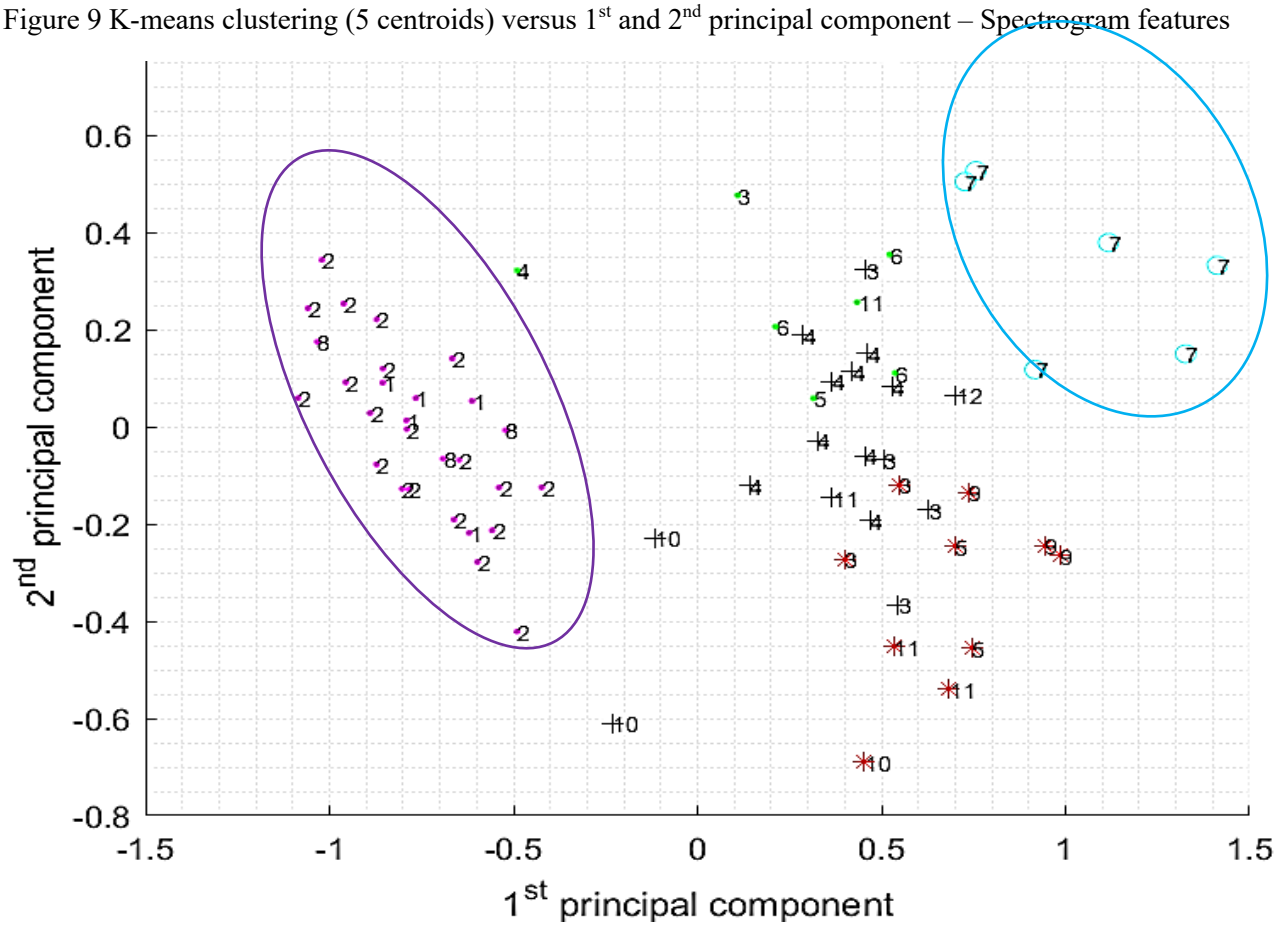
These results have several notable implications. First, motion subclasses 1, 2, and 8 describe are seemingly indistinguishable. The result is confirmed with visually similar spectrograms and subsequent k-means clustering.

Other segmentations shared strong dissimilarities with respect to other groups. For example, sub-class labels 3 & 4 are visually dissimilar from sub-classes 1 & 2. Moreover, other classes such as sub-class 7, shows strong dissimilarity with nearly every other segmentation set.

Figure 9 shows a plot of the 1st and 2nd principal components of maximum variance for the y-axis segmentations, with true labels overlaid for each instance. Several sets are noticeably separable and distinct. Certain sets including {1, 2, 8} as well as {7} and {10} show clearly discriminating characteristics from other groups – reinforcing the results of the cosine similarity analysis. Another group {3, 4, 5, 6, 9, 11, 12} has a less distinct cluster region, again as was the case with the cosine similarity results.

K-means clustering was then recomputed with the derived features extracted from the CNN segmentations. Principal component analysis reduced the dimension of the dataset to 10 key features. The results are shown in Figure 9. An ellipse is drawn over key segmentations for clarity of discussion. While the number of centroids should theoretically correspond to the number of known labels, empirical tests showed spurious over-segmentation tended to occur above five defined clusters. This was due to several factors, notably that features between certain sub-class sets were similar enough to prevent distinct separation. Certain labels also had a limited number of occurences, such as sets {10, 12}, which implicitly reduced their significance in the k-means algorithm.

Figure 9 K-means clustering (5 centroids) versus 1st and 2nd principal component – Spectrogram features



Several predominant clusters were denoted by k-means segmentation and validated against known labels and cosine similarity analyses. Sets {1, 2, 8} are observed to be most dissimilar from set {7} which was again confirmed with clustering. Other segmentations were less well defined but shared similar features and are sufficiently discriminated from other sets. For example, most labels in sets {5, 9} are classified as an independent and distinct.

Overall, this semi-supervised approach overcame many of the limitations of the fully unsupervised segmentation approach. Errors in clusters are likely due to the dimensionality reduction inherent in feature engineering, as well as the high degree of similarity between many y-axis secondary motion classes. While all sub-classes were not independently clustered, several key groups were distinguished – indicating the method's feasibility for further study. More sophisticated clustering approaches may be required if sub-classes consistently share similar feature subspaces. Unfortunately, other primary motion classes, such as x-axis motions, showed poor segmentation accuracy with the CNN, though these results may be improved through an expanded labeled data set, which was not available at the time of this report.

5. Conclusion

Research was conducted to supplement remote acoustic monitoring procedures in a semiconductor device manufacturing environment. Specifically, research into analytical approaches to identify and isolate robotic actuation instances for a SCARA series system based on acoustic monitoring devices was conducted. Two approaches were developed and tested: a fully unsupervised approach and a semi-supervised approach.

Building on conventional, fully-unsupervised machine learning approaches, engineered signal features were extracted from windowed segments of acoustic audio data. K-means clustering then attempted to separate these segments into defined actuation groups based on the characteristics of the engineered features. The results of this approach were mixed due to an inability to consistently cluster lower energy actuations. Higher energy motions were distinguishable from idle states using this method; however, transitional periods between y-axis motion and idle/noise states were not distinctly captured. Overall, analysis of this approach suggested that unsupervised procedures were heavily biased by windowing, noise, the nature of the engineered feature sets, and the relative energy differences between motion classes.

A semi-supervised architecture was subsequently explored to see if a combination of supervised and unsupervised methodologies would increase identification accuracy. This method would additionally remove dependency on fully supervised neural network architectures for highly segmented data; instead, primary and sub-class segmentation would be introduced in separate procedures limiting highly granular dataset generation. This method, while not fully unsupervised was generalizable to other processes and considered valid. A deep convolutional neural network was first constructed and trained using normalized spectrograms labeled with primary motion class data (x, y, noise) as inputs. While mixed results were observed for x-axis activations, y-axis motions were segmented with high accuracy. The poor performance of the x-axis segmentations was most likely due the limited labeled data and visually definable spectrogram features available for CNN training.

Once segmented by the CNN, a new set of features incorporating frequency-based statistical features was extracted for unsupervised clustering and similarity analyses. Both cosine similarity and k-means clustering methods were able to distinguish and separate many y-axis actuation sub-classes based on the segmentations provided by the CNN. Errors were due to having a limited number of instances of some data sub-classes, as well as high similarity between certain actuation sub-classes. Overall, a semi-supervised approach to acoustic monitoring is recommended over fully unsupervised techniques, despite the need for additional labeled data. Regardless of the analytical approach, the range of frequency domain responses for primary actuation classes poses the biggest challenge to accurate segmentation.

Avenues for future work most notably include the development of a larger and more diverse training data set, which will likely improve the performance of both the CNN and clustering methods, particularly with respect to x-axis motion segmentations. Segmentation and identification of lower-energy actuations such as z-axis motions will most likely require other monitoring methods, as these actuations are not clearly and consistently manifested within an acoustic spectrogram. Data fusion methods, either with video or embedded sensor data, are recommended as an avenue to improve all aspects of this methodology. Approaches that fuse multiple acoustic recordings together could also improve the consistency and broader capabilities of any monitoring approach.

6. Future Work

To improve the current methodology, several approaches will be further explored. More robust, generalizable spectral features are required for increased accuracy, as well as lower energy actuation segmentations. Additional autoencoder layers may introduce more robust features due to learned data representations. Data fusion provides another opportunity to concatenate features from other signal sources to increase data richness.

7. Acknowledgements

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

8. References

Mathia K (2010) *Robotics for Electronics Manufacturing*. Cambridge University Press, Cambridge, UK, pp. 26-28.

Alías F *et al.* (2016) A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Applied Sciences*, 143(6): 10.3390/app6050143.

Dennis J (2014) *Sound Event Recognition in Unstructured Environments Using Spectrogram Image Processing*. Doctoral Dissertation, School of Computer Engineering, Nanyang Technological University, Singapore.

Yu W *et al.* Y (2018) Audio Classification Using Attention-Augmented Convolutional Neural Network*, Knowledge-Based Systems* 161, pp. 90-100.

Boddapati, V *et al*. (2017) Classifying Environmental Sounds Using Image Recognition Networks. *Procedia Computer Science* 112: 2048-2056.

Ruqiang Y and Robert XG (2009) Multi-scale Enveloping Spectrogram for Vibration Analysis in Bearing Defect Diagnosis, *Tribology International* 42(2), pp. 293-302.

Espi M *et al.* T (2015) Exploiting Spectro-Temporal Locality in Deep Learning Based Acoustic Event Detection. *EURASIP - Journal on Audio, Speech, and Music Processing* 26(1).

Jang G *et al.* (2014) Audio Source Separation Using a Deep Autoencoder, *arXiv preprint:* arXiv:1412.7193.

Piczak KJ (2015) Environmental Sound Classification with Convolutional Neural Networks. *Proceedings in IEEE 25th International Workshop Machine Learning for Signal Processing (MLSP)*, pp. 1-6.

Chachada S and Kuo J (2014) Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing* (3).

Cooper ML and Foote J (2002) Automatic Music Summarization via Similarity Analysis. *ISMIR*.

MathWorks Inc (2019) Convolutional Neural Networks. https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html (accessed 11/04/2019).