MODULAR AND OFFSITE CONSTRUCTION SUMMIT
JULY 27-29, 2022 EDMONTON, CANADA

Keyword identification framework for speech communication on construction sites

Asif MANSOOR[1*], Shuai LIU[2], Ghulam Muhammad ALI[1], Ahmed BOUFERGUENE[3], Mohamed AL-HUSSEIN[3] and Imran HASSAN[4]

[1] *PhD candidate, Department of Civil and Environmental Engineering, University of Alberta*
[2] *PhD student, Department of Civil and Environmental Engineering, University of Alberta*
[3] *Professor, Department of Civil and Environmental Engineering, University of Alberta*
[4] *Student, Department of Computer Science, University of Turbat, Pakistan*
[*]*Corresponding author's e-mail: amansoor@ualberta.ca*

ABSTRACT

Worksite communication is a key to boosting teamwork and improving worker performance on the construction worksite. Communication among workers on the construction site mostly consists of speech communication. However, construction sites are typically noisy due to construction tasks like drilling and operation of heavy equipment. Meanwhile, workers on construction sites typically represent a range of different ethnic and linguistic backgrounds and have different speaking accents. This can make it difficult for the listener to understand the speaker clearly, leading to miscommunication and errors in decision making on the construction site. Technological advancements in recent years can be leveraged to mitigate this problem. In this paper, a keyword identification framework is developed for speech communication on the construction site. For this framework, 12 hours of raw audio data containing 18 crane signalman speech commands (referred to as "keywords") are collected. The crane signalman uses specific keywords to communicate with the crane operator and guide the crane operator in the crane operations. The 2-second audio clips (this being the approximate duration of each keyword) are extracted from the raw audio dataset, and construction site noise is added. Moreover, mel-frequency cepstral coefficients are extracted from the waveform audio dataset. The extracted mel-frequency cepstral coefficients, in turn, are used to train the 1-dimensional convolutional neural network. After training, the model is found to achieve a training accuracy of 97.3%, a validation accuracy of 96.1%, and a testing accuracy of 93.8%. The model is further deployed for real-time identification of keywords in speech, with the model achieving an accuracy of 95.3%. In light of these findings it can be concluded that the developed framework is suitable for real-time application in noisy construction sites for identifying specific keywords in speech.

KEYWORDS

Keyword identification; Mel-frequency cepstral coefficients; Convolutional neural network; Communication; Crane signalman

INTRODUCTION

Communication plays a significant role in establishing and maintaining effective working relationships in any industry. In current practice, workers on construction sites typically rely on face-to-face verbal, hand signalling, and two-way radio communication systems (Mansoor et al.,

2020). However, when there is an obstacle or significant distance between workers, face-to-face verbal or hand signalling communication may not be reliable or even feasible. In such cases, two-way radio communication is the best way to convey the message. In two-way radio communication, radio units are used to send and receive audio data (Carbonell et al., 2020). On the construction site, this approach is primarily used for communication between workers on the ground and heavy construction machinery operators, such as in crane operations, where a two-way radio communication system is typically used when it is difficult for the operator to see the signalman due to an obstacle in their line of sight (Stevenson, 2019). This communication approach requires a dedicated channel for the communication between operator and signalman that must be maintained at all times.

However, construction sites can be noisy due to construction activities such as drilling and operation of heavy equipment (Kwon et al., 2016), making it difficult for the listener to hear speech commands. Furthermore, construction workers typically represent a diverse range of different ethnic and linguistic backgrounds and have different accents, meaning that it may be difficult for the listener to understand the speaker in some cases, leading to misjudgments in decision-making, as well as safety and productivity issues (Bust et al., 2008).

There in an opportunity in the regard for the construction industry to benefit from recent developments in information technology. In particular, intelligent and automated systems of communication can be introduced to improve communication between heavy construction machinery operators and workers on the ground and thereby improve the safety and productivity of site operations. In this context, the present study aims to develop a framework for keyword identification in speech on construction sites. The developed framework can provide an intelligent, advanced, and more reliable communication system that can reduce the risk of miscommunication on construction sites.

RELATED STUDIES

Speech is the primary means of communication among human beings; as such, speech recognition systems have received considerable interest among researchers in recent decades. However, due to reliability issues, the systems developed have not been widely implemented (Latif et al., 2021; Otter et al., 2020; Strehl et al., 2006). Nevertheless, the major advancements in machine learning and deep learning in recent years have led to accurate speech recognition with high reliability that has increased the practicability of speech recognition systems (Hinton et al., 2012; Meftah et al., 2018). Speech recognition systems are now being used for various applications, including (1) keyword identification/spotting (Lopez-Espejo et al., 2021; Michaely et al., 2017; Werchniak et al., 2021; Momeni et al., 2020); (2) automatic recognition of the content of words and phrases in order to direct computer tasks as an alternative to typing, facilitating human–machine interaction as a support for the disabled, supporting smart home functions, etc.; (3) emotion recognition, i.e., for recognizing the emotion of the speaker based on speech signals (Fragopanagos & Taylor, 2005; Petrushin, 2000); (4) in intelligent health care systems to provide information on patient health status (Zhou et al., 2001); (5) to assist in deciphering the speech of people with various accents (Biadsy, 2011); (6) in estimating a speaker's age (Bocklet et al., 2008) and gender (Vogt & André, 2006); and (7) for spoken language translation (Schultz & Waibel, 2001), spoken document retrieval (Chelba et al., 2008), and multilingual speech recognition (Kannan et al., 2019).

In the area of construction, Zhang et al. (2018a) used a speech recognition framework to analyze onsite conversations. Their framework used a naïve Bayes classifier to translate speech captured on site into text scripts, and to further classify the text scripts into construction activities and operations. The framework achieved an overall accuracy of 90.9%. In another study, Zhang et al., 2018b) developed a supervised machine learning-based sound identification framework, using it to identify six different sounds common to construction sites (concrete-grinding, hammering, concrete-pouring, drilling, excavator operation, and dozer operation). The framework achieved a maximum accuracy of 94.3%. Speech recognition has also been used to retrieve BIM data from BIM software (Shin & Issa, 2021) and to identify heavy construction equipment operating at a construction site (Cheng et al., 2017). To the authors' knowledge, though, the use of speech recognition systems to identify keywords in speech communication on construction sites has yet to be explored. The present study thus develops a keyword identification framework to facilitate communication by identifying keywords of interest in a noisy construction site. The framework is also capable of identify keywords in speech by workers with different accents.

RESEARCH METHODS

In this research, a crane signalman speech dataset is collected using a microphone attached to the crane signalman's helmet. The speech dataset is pre-processed by adding representative construction noises, and data augmentation is implemented by altering the pitch and adding random noise to the speech. This helps to generalize the dataset and reduces the likelihood of model overfitting, per Lei et al. (2019). The waveform speech is then converted into mel-frequency cepstral coefficients (MFCCs) in order to extract unique features from the speech dataset (Mahmood & Utka, 2021). MFCCs, it should be noted, are the most widely used feature extraction algorithm in the field of speech recognition. The purpose of using MFCCs is to reduce the complexity of the model and achieve higher accuracy (Mahmood & Utka, 2021). The extracted features obtained using MFCCs are then used to train the one-dimensional convolutional neural network (1-D CNN) classifier to identify the keywords in the speech. The 1-D CNN model is validated, tested, and deployed for real-time identification of keywords in speech as an output. An overview of the keyword identification framework is given in Figure 1.

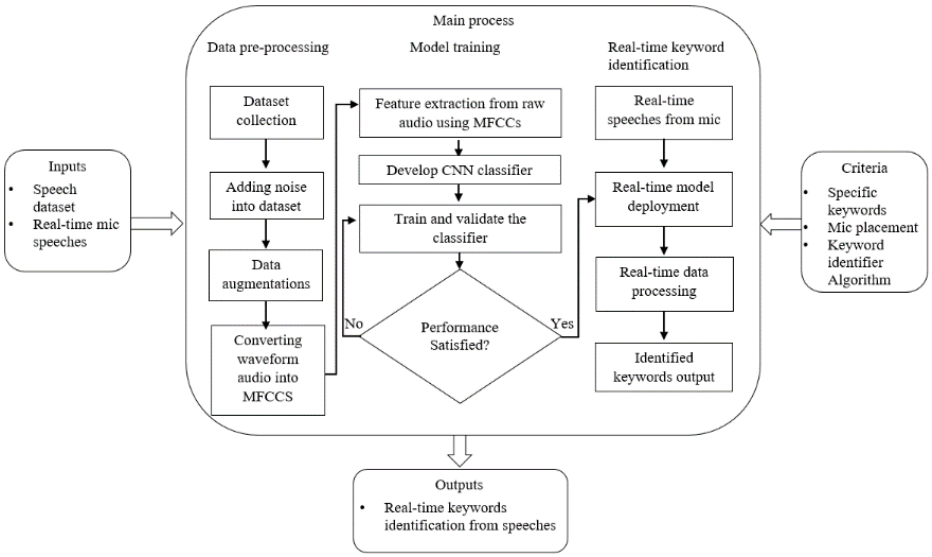

Figure 1. Overview of keyword identification framework

IMPLEMENTATION AND CASE STUDY

The framework is implemented and tested using the crane signalman speech commands used to guide the crane operator in the crane operations on the construction site (see Figure 2).

Boom up and lower the load	Boom down	Dog everything	Hoist	Move slowly	Swing
Boom down and raise the load	Boom up	Emergency stop	Lower	Stop	Travel
Use main hoist	Telescope out	Travel both tracks	Travel one track	Telescope in	Use whipline

Figure 2. Crane signalman speech command keywords

Dataset collection and pre-processing

With no existing dataset for crane signalman speech commands available, the speech command data is collected manually at a 16,000 Hz frequency using a microphone attached to the crane signalman's helmet. The dataset collected contains 12 hours of crane signalman speech commands made by 45 volunteers (30 male, 15 female) representing 13 different ethno-linguistic backgrounds and having different accents. The dataset is resampled into 2-second duration speech files for each command, referred to as a "keyword". Each sample is further normalized to adjust the range of speech, equalized to remove bumps from the speech, and compressed to modify the range of loudness of the speech. Furthermore, construction site-related noise, collected from the Mixkit (2022) and Zapsplat (2022) datasets, is added to the speech. The incorporation of construction site noises helps to generalize and reduce the likelihood of model overfitting. Data augmentation is then applied to artificially alter the pitch of the speech.

Model development

The set of 21,600 samples of 2-second audio commands representing a total of 12 hours of data is converted from waveform into MFCCs, which are capable of representing the amplitude spectrum of the sound wave in a compact vectorial form (De Pinto et al., 2020). In this technique, it should be noted, the audio file is divided into frames, usually using a fixed window size, in order to obtain statistically stationary waves and, in turn, frames. The frames having been obtained, discrete Fourier transform is applied, and only the logarithm of the amplitude spectrum is retained. The amplitude spectrum is normalized with a reduction of the mel frequency scale. This operation is executed for the purpose of identifying the frequencies (Logan, 2000). The interested reader may refer to Davis & Mermelstein (1980) and Huang et al. (2001), in which the MFCC calculations are thoroughly explained. To extract features from waveform audio, the main parameters used in MFCCs are the number of coefficients (referred to as the static features) that contain the information in a given audio frame, the fast Fourier transform length (which represents the number of samples in each window), the number of filters (which reflects the number of features extracted from the audio file), the frame stride, and the frame length. The values chosen for the developed framework are given in Table 1. These values are selected in a trial-and-error based on the performance of the model. Figure 3 shows examples of waveform audio and corresponding MFCCs.

Table 1. Parameter of MFCCs

Parameters	
Number of coefficients	40
Fast Fourier transform length	512
Number of filters	40
Frame stride	0.02
Frame length	0.02

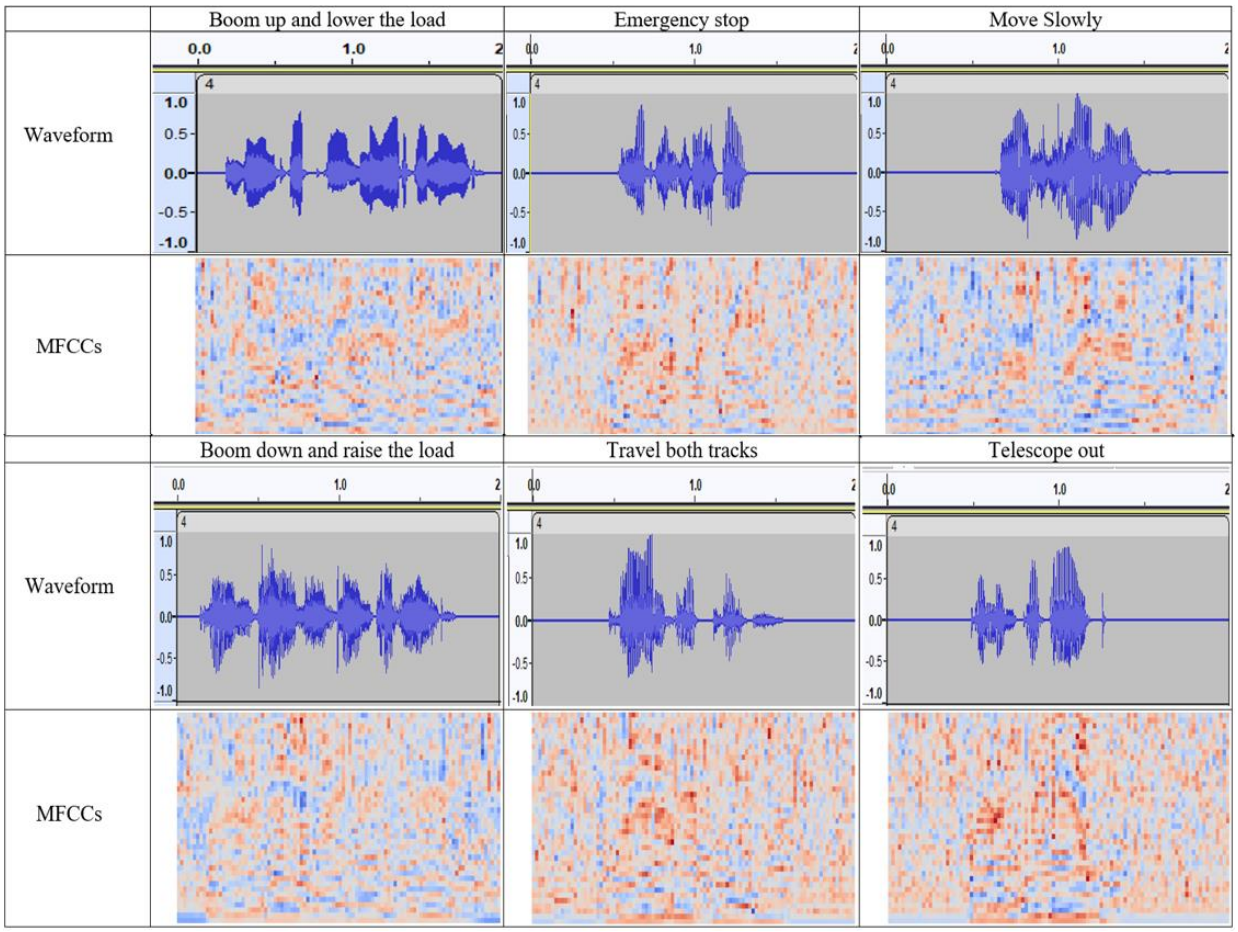

Figure 3. Waveform audio sample of keywords and corresponding MFCCs

The features extracted from the MFCCs are then fed into 1-D CNN classifiers, which a classifier can operate on vectors of the features for each audio file provided as input. Here, the values represent the compact numerical form of the audio frames of 2-second duration. The compact numerical form of the audio frame is input to the 1-D CNN, the architecture of which includes four convolutional layers that are responsible for extracting and learning features from the input data. Each convolutional layer is followed by an activation function to add non-linearity to the output neuron. For this work, the rectified linear unit (ReLU) activation function is used in the convolutional layers. The ReLU activation function, it should be noted, is an identity line for which $y = x$ for all positive lines and $y = 0$ for all negative values. Each convolutional layer and activation function is followed by a pooling or subsampling layer and dropout layer. The pooling layer helps the model to focus on the principal characteristics of each portion of speech data, making the potion of data them invariant by their position, while the dropout layer activates and deactivates the neurons with respect to their weights. (This technique helps to better generalize the predictive capabilities of the model.) The output from the dropout layer is then flattened to make it compatible with the subsequent layers. Finally, a softmax activation function is applied to one dense layer (i.e., fully connected layer) in order to estimate the probability distribution of each of the classes properly encoded in the model.

RESULTS

The dataset is randomly split into training, validation, and test sets. The proportion of the training set is kept at 80% (17,280 samples) while the validation and test sets are kept at 10% (2,160 samples) each. The reason for using more samples in the training set is to allow the 1-D CNN model to learn more features from the dataset (as this will lead to a more accurate model for identifying keywords in the validation and test sets). The 1-D CNN model is trained on 100 epochs while keeping the model learning rate to 0.005 and the dropout rate to 0.25. The number of epochs, values of learning, and dropout rate are determined experimentally to boost the accuracy of the validation and testing. The model is found to achieve average accuracy of 97.3% and 96.1% and losses of 0.12 and 0.18 in the training and validation processes, respectively. Moreover, the model is found to achieve an accuracy of 93.8% in the test set. The accuracy and loss are measured using Equations 1 and 2.

$$Accuracy = \frac{Correctly\ identified\ keywords\ in\ speech}{Total\ number\ of\ keywords\ in\ speech} \times 100 \qquad (1)$$

$$cross\ entropy\ loss = \frac{-1}{N} \times \sum_{x=1}^{N} \sum_{y=1}^{M} Z_{xy} \times \log\left(p_{xy}\right) \qquad (2)$$

where N is the number of samples and M is the number of classes; Z_{xy} specifies whether or not sample *x* belongs to class *y*; and p_{xy} represents the probability of sample *x* belonging to class *y*. The loss has no upper limit and falls within the range [0,∞], where a value of loss near 0 indicates high accuracy.

The model is then deployed for real-time keyword identification from live speech using a microphone. Based on 650 iterations, the model is found to achieve an overall accuracy of 95.3% in real time. This result demonstrates that the developed model is capable of accurately identifying keywords in speech in the context of a construction site environment. As such, the model can be considered suitable for use as an additional layer of communication on noisy construction site.

CONCLUSION

In this study, a keyword identification framework is developed that is capable of identifying, in real time, 18 different crane signalman speech commands (i.e., "keywords"). To begin with, a dataset of 12 hours of crane signalman speech commands is collected manually using a microphone attached to the signalman's helmet. Construction site noise is then added to generalize the dataset. Short audio clips of 2-second duration (i.e., the approximate duration of a keyword/command) are then separated from the dataset, and features are extracted from the audio dataset using MFCCs. The extracted features are used as an input to train the 1-D CNN model, which is found to achieve training, validation, and testing accuracies of 97.3%, 96.1%, and 93.8%, respectively. The model is further validated in the real-time identification of keywords in live speech, achieving an accuracy of 95.3%. In future work, more data will be added, and the model will be further optimized to improve its accuracy in performing real-time keyword identification.

REFERENCES

Biadsy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. Doctoral dissertation, Columbia University, New York City, NY, USA.

Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., & Noth, E. (2008). Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1605–1608.

Bust, P. D., Gibb, A. G., & Pink, S. (2008). Managing construction health and safety: Migrant workers and communicating safety messages. *Safety Science*, *46*(4), 585–602.

Carbonell, M. L. B., Carpio, J. M. R., Medina, J. C. C., Perote, J. P., Tamayo, T. J. J., & Mappatao, G. P. (2020). Development of a stand-alone and scalable weather monitoring system using two-way VHF radios. *Indonesian Journal of Electrical Engineering and Computer Science*, *20*(1), 475–484.

Chelba, C., Hazen, T. J., & Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, *25*(3), 39–49.

Cheng, C. F., Rashidi, A., Davenport, M. A., & Anderson, D. V. (2017). Activity analysis of construction equipment using audio signals and support vector machines. *Automation in Construction*, *81*, 240–253.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

De Pinto, M. G., Polignano, M., Lops, P., & Semeraro, G. (2020). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. *Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems*.

Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, *18*(4), 389–405.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97.

Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., & Lee, S. (2019). Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*.

Kwon, N., Park, M., Lee, H. S., Ahn, J., & Shin, M. (2016). Construction noise management using active noise control techniques. *Journal of Construction Engineering and Management*, *142*(7), 04016014.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H. S., & Cambria, E. (2021). A survey on deep reinforcement learning for audio-based applications. *arXiv preprint arXiv:2101.00240*.

Lei, C., Hu, B., Wang, D., Zhang, S., & Chen, Z. (2019). A preliminary study on data augmentation of deep learning for image classification. *Proceedings of the 11th Asia-Pacific Symposium on Internetware*.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. *Proceedings of the International Symposium on Music Information Retrieval*.

López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep spoken keyword spotting: An overview. *IEEE Access*.

Mahmood, A., & Utku, K. Ö. S. E. (2021). Speech recognition based on convolutional neural networks and MFCC algorithm. *Advances in Artificial Intelligence Research*, *1*(1), 6–12.

Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M. (2020). Conceptual Framework for Safety Improvement in Mobile Cranes. In *Construction Research Congress 2020: Computer Applications* (pp. 964-971). Reston, VA: American Society of Civil Engineers.

Meftah, A. H., Alotaibi, Y. A., & Selouani, S. A. (2018). Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis. *IEEE Access*, *6*, 72845–72861.

Michaely, A. H., Zhang, X., Simko, G., Parada, C., & Aleksic, P. (2017). Keyword spotting for Google assistant using contextual speech recognition. *Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 272–278.

Mixkit (2022). Construction sound effects. Retrieved March 3, 2022, from https://mixkit.co/free-sound-effects/construction/

Momeni, L., Afouras, T., Stafylakis, T., Albanie, S., & Zisserman, A. (2020). Seeing wake words: Audio-visual keyword spotting. *arXiv preprint arXiv:2009.01225*.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624.

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. *Proceedings of the Sixth International Conference on Spoken Language Processing*.

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, *35*(1–2), 31–51.

Shin, S., & Issa, R. R. (2021). BIMASR: Framework for voice-based BIM information retrieval. *Journal of Construction Engineering and Management*, *147*(10), 04021124.

Stevenson Crane (2019). *Crane signals 101 – Can you hear me now?* Retrieved June 17, 2019, from https://stevensoncrane.com/crane-signals-101-can-hear-now/

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., & Littman, M. L. (2006). PAC model-free reinforcement learning. *Proceedings of the 23rd International Conference on Machine Learning*, 881–888.

Vogt, T., & André, E. (2006). Improving automatic emotion recognition from speech via gender differentiaion. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Werchniak, A., Chicote, R. B., Mishchenko, Y., Droppo, J., Condal, J., Liu, P., & Shah, A. (2021). Exploring the application of synthetic audio in training keyword spotters. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7993–7996.

Zapsplat (2022). Construction site sound effects. Retrieved March 2, 2022, from https://www.zapsplat.com/sound-effect-category/construction-site/

Zhang, T., Lee, Y. C., Scarpiniti, M., & Uncini, A. (2018b). A supervised machine learning-based sound identification for construction activity monitoring and performance evaluation. *Proceedings of the Construction Research Congress*, 358–366.

Zhang, T., Lee, Y. C., Zhu, Y., & Hernando, J. (2018a). A conversation analysis framework using speech recognition and naïve bayes classification for construction process monitoring. *Proceedings of the Construction Research Congress*, 572–580.

Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 201–216.