Song Li
Department of Computer Information Science and Engineering,
University of Florida,
303 Weil Hall,
Gainesville, FL 32603;
Department of Computer Science,
Johns Hopkins University,
3400 N Charles Street,
Baltimore, MD 21218
e-mail: sli218@jh.edu

Mustafa Ozkan Yerebakan
Department of Industrial and Systems Engineering,
University of Florida,
303 Weil Hall,
Gainesville, FL 32603
e-mail: mu.yerebakan@ufl.edu

Yue Luo
Department of Industrial and Systems Engineering,
University of Florida,
303 Weil Hall,
Gainesville, FL 32603
e-mail: yueluo@ufl.edu

Ben Amaba
IBM,
Blue Lagoon Drive,
Miami, FL 33126
e-mail: baamaba@us.ibm.com

William Swope
IBM,
410 Robin Hood Cir Unit 102,
Naple, FL 34104
e-mail: bswope@us.ibm.com

Boyi Hu[1]
Department of Industrial and Systems Engineering,
University of Florida,
303 Weil Hall,
Gainesville, FL 32603
e-mail: boyihu@ise.ufl.edu

The Effect of Different Occupational Background Noises on Voice Recognition Accuracy

Voice recognition has become an integral part of our lives, commonly used in call centers and as part of virtual assistants. However, voice recognition is increasingly applied to more industrial uses. Each of these use cases has unique characteristics that may impact the effectiveness of voice recognition, which could impact industrial productivity, performance, or even safety. One of the most prominent among them is the unique background noises that are dominant in each industry. The existence of different machinery and different work layouts are primary contributors to this. Another important characteristic is the type of communication that is present in these settings. Daily communication often involves longer sentences uttered under relatively silent conditions, whereas communication in industrial settings is often short and conducted in loud conditions. In this study, we demonstrated the importance of taking these two elements into account by comparing the performances of two voice recognition algorithms under several background noise conditions: a regular Convolutional Neural Network (CNN)-based voice recognition algorithm to an Auto Speech Recognition (ASR)-based model with a denoising module. Our results indicate that there is a significant performance drop between the typical background noise use (white noise) and the rest of the background noises. Also, our custom ASR model with the denoising module outperformed the CNN-based model with an overall performance increase between 14–35% across all background noises. Both results give proof that specialized voice recognition algorithms need to be developed for these environments to reliably deploy them as control mechanisms. [DOI: 10.1115/1.4053521]

Keywords: human computer interfaces/interactions, machine learning for engineering applications

1 Introduction

Voice recognition has become a ubiquitous phenomenon in the past decade. The introduction of virtual assistants such as Siri and Alexa has propelled the use of voice recognition. These virtual assistants can execute search queries, dictate texts, order food and groceries, and even control lighting and temperature in houses, among many other applications. This breadth of applications has allowed voice recognition to be used in healthcare for accurately recording data [1], smart transportation systems [2], and smart homes [3]. The global voice recognition market size is predicted to grow by 17.2% between 2020 and 2025, reaching a size of $26.79 billion [4]. The fact that voice recognition allows for hands-free interaction with many systems provides ample opportunities making it a promising technology in many occupational scenarios, in which workers need constant interaction with the environment and machine while their other communication modalities (hands, visual, etc.) are heavily taxed.

Voice recognition to control industrial machinery has different requirements than the algorithms that are used in everyday settings. Rogowski [5] emphasizes the correct recognition of phrases as a crucial element in voice recognition for industrial settings. The main factor that affects the performance of the voice recognition algorithm is the existence of background noise, which significantly affected the accuracy of Rogowski's study. Occupational settings such as factory floors and construction sites for example have ambient noise sources (crushing, riveting, blasting, cutting torches, etc.) that are much different than homes and airports [6].

[1]Corresponding author.
Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received August 30, 2021; final manuscript received January 7, 2022; published online March 31, 2022. Assoc. Editor: Ying Liu.

Hence, applying current models to these settings will result in lower accuracy rates as they are developed for quasi-silent environments.

Another important element of voice recognition in industrial settings is the unique type of communication that is present compared to everyday life. Unlike the communication and information exchange that happens in our daily lives, communication in occupational settings is often short and imperative. The margin for error in occupational settings is much lower, and there is a finite set of commands. In this setting, the voice recognition software's main purpose is to detect these short phrases reliably and quickly with as few false positives as possible. This is not without its challenges though. High network latency that is caused by external factors such as electromagnetic interference (EMI) from nearby machines will affect the data quality. Therefore, these voice recognition models should be based on "keyword spotting," not the more common "full sentence recognition" approach. It is also important to consider that even though these words are rather short they have different sound compositions that combined with the existence of background noise and other sources of interference can translate into different levels of accuracy.

By considering the two aspects of voice recognition that are discussed, background noise and communication method, we believe a robust voice recognition algorithm can be developed that could be used in occupational settings. As interactions within occupational settings start to diversify with the increase of interactions between humans and machines [7,8], the need for reliable voice recognition software will continue to grow. Hence, the purpose of our study is to demonstrate the necessity of a specialized voice recognition algorithm by comparing it to a generic machine learning-based algorithm. Through the results of this study, we aim to guide occupational safety professionals and safety managers a guideline as to which sound levels and what types of background noise need to be dampened to implement voice-controlled systems safely and reliably. Additionally, we encourage researchers and professionals in the computer science field to consider developing voice recognition algorithms that are tailored to the background noise present in workplaces. We also highlight the importance of choosing the right set of command words as some words get confused for one another by the voice recognition algorithm, which may cause unintended consequences.

The manuscript is structured as follows: We first introduce the existing publicly available datasets, speech recognition algorithms, and denoising solutions and highlight their shortcomings. Then, we introduce our algorithm and explain how it is structured. This is followed by the comparison with the generic algorithm. Finally, we discuss the results and talk about future directions and the limitations of the study.

2 Relevant Work

Using voice control to interact with machines is not a new concept, with research dating back to 1995 [9]. However, with the rise of Industry 4.0, humans and machines are becoming more interconnected. Voice recognition research pertaining to this new paradigm is mostly focused on evaluating the feasibility of using voice recognition to interact with a number of systems. Longo and Padovano explored the possibility of using a virtual assistant akin to Siri and Alexa to safely operate machines and detect failures [10]. Rains also explored using voice recognition for safety, in this case as an emergency shut-off mechanism for tractors [11]. Two groups suggested using voice recognition as part of a cloud-based control system to interact with autonomous industrial systems [12] and semi-autonomous vehicles [13]. Voice-activated Human Robot Interaction (HRI) systems were also used in training novice users how to interact with robots [14].

The reliability of voice recognition software in industrial settings has just become a topic of interest in recent years, albeit the research on the topic is still very limited. Most studies that have investigated the effects of background noise were conducted in the context of voice-activated assistants in cars, [15–18], autonomous vehicles [19], virtual home assistants [20], and more recently cough detection [21]. Although the speech patterns inside these settings bear resemblance to the words we have selected in this study (short phrases such as place names or commands such as "call" or "open"), the background noise profile is completely different than the industrial settings we have included in this study. For example, the typical noise level for a chainsaw is 108 dB [22]. Other pieces of equipment we also included have noise levels between 90–100 dB. Compared to these levels, the noise level inside a car that is going 60 mph (which is not a typical speed in everyday urban settings) is 70 dB.

2.1 Speech Recognition Network. There are two main methods for developing voice recognition algorithms. One method is pattern-based and a second, more recent one is knowledge-based. Voice recognition algorithms that use pattern-based techniques rely on matching speech or voice patterns with existing templates [23]. Pattern-based templates have been the mainstay for voice recognition software for many years. The advent of artificial intelligence (AI) and machine learning methods has changed this, however. Compared to pattern-based methods, knowledge-based methods rely on breaking down the voice into features and then identifying which ones are the most relevant in recognizing different speech patterns [23]. Machine learning-based voice recognition research has attracted significant attention in recent years [24]. One study constructed a voice command database with background noises from three separate databases, Chime-3 and Aurora-2 which have noises from public spaces such as streets, sidewalks, airports, and pink noise samples from the Noisex−92 database [25]. This and similar studies have used machine learning to develop voice recognition algorithms that have high accuracy rates [26–28], yet all of them suffer from a common problem that affects their applicability, the dataset itself.

Deep learning has been recently applied to automatic speech recognition as cloud-based services that provide end-to-end speech recognition solutions. They often utilize multiple models together to solve problems. The main function of a speech recognition algorithm is to take the input of a sequence of signals and translate that into the target word. The most straightforward way to do this is to extend the Seq2seq model scientists have used in machine translation [29]. Seq2Seq is composed of two parts: an encoder and a decoder with an attention mechanism. It can use different models to be its encoder and decoder such as RNN and Transformer [30,31]. However, the traditional Seq2Seq model and other end-to-end deep learning models lack denoising ability in a noisy environment. Auto Speech Recognition (ASR) and the front-end denoise module are designed to solve the problem.

2.2 Dataset. There are several large, publicly available voice datasets. Mozilla's Common Voice dataset [32] contains over 500 h of recording with 20,000 different people's voices. Although the size of the dataset is remarkable, the recording lack background noise. The Speech Commands Dataset [33] from TensorFlow provides a dataset with 35 single word commands with thousands of utterances for each word. This database has some background noise present, both digitally generated and recorded, however, the author does not elaborate on how and why he selected the background noises. Microsoft Scalable Noisy Speech Dataset [34] collected 56 speakers' speech clips under a clean environment and provided 14 types of daily noises including air-conditioner, eating, shutting the door, and so on. Speech Commands Dataset and Microsoft Scalable Noisy Speech Dataset [33,34] both include background noises from daily living. However, a very limited number of datasets have addressed the ambient noise situation under industry, agriculture, and other occupational settings. This could be an issue as in occupational settings noise exposure may represent different profiles as domestic living scenarios [35]. This especially holds true when using machine learning models.

If the data that are used to train the model do not represent the sound profile of the environment the algorithm is deployed in, the performance of the model will be significantly affected when it is deployed.

One of the few studies that has researched the effect of background noise in industrial settings is a study done by Birch et al. [36]. Their word selection also consisted of short command words; up, down, left, and right. The background noises they selected were the operating noise of a drilling robot (just the movement), speech, drilling, and a sample of machine noises in the general vicinity. They reported a significant decrease in levels of noise above 61.2 dB due to ambient machine noise whereas other noise types did not have a significant effect. This finding matches with our results, as accuracy levels above 60 dB for both algorithms drop significantly. Their study is important in demonstrating the effect of ambient machine noise and recognizing the utility of voice recognition in the context of HRI, which has already been recognized by other researchers as well [10,37].

2.3 Denoising Solution. Sound enhancement algorithms have introduced ways to recognize speech under noisy circumstances. Among them, a few algorithms introduced have already gained traction in the speech recognition and sound enhancement community. RNNoise [38] is a low-complexity noise suppression algorithm that uses recurrent neural networks. The method estimates noise suppression gains in relevant critical bands. WaveNet [39] takes a series of convolutional filters with exponentially increasing dilation factors and predicts target fields. Speech enhancement generative adversarial network (SEGAN) [40] is an end-to-end model where input is directly the raw data. It combines two different models: a generative model that transfers the noisy speech into a clean one and a discriminative network that distinguishes whether inputs come from clean or enhanced speech. These models prove to be effective under white noise, noise with equal intensities, and background noise. However, little research has been done to test these denoise models on datasets with specific noise patterns such as datasets collected from occupational settings.

3 Methodology

3.1 Background Noise Selection. As mentioned earlier, our goal was to construct a database that would be representative of occupation settings that have harmful ambient noise. Considering the nature of speech that is present in those environments, we elected to construct a voice command database with a variety of background noises that would reflect different occupational contexts. For this, we first started surveying occupational settings that typically have higher levels of ambient noise. As high levels of ambient noise can have detrimental effects on a person's hearing [41], we focused on industries that had the highest reported levels of hearing loss. We also surveyed existing literature about occupational noise pollution to determine what ambient sounds we need to include. According to NIOSH's Occupational Hearing Loss Surveillance Data, the highest reported cases of hearing loss are in the mining/oil industries and the construction industries with 25% of workers reporting a hearing loss in both industries [42]. This is followed by manufacturing and agriculture/forestry industries. Hence, we chose these industries as our occupational settings. Subsequently, we surveyed the equipment used in these industries that are both common and have a prominent noise output. For the construction and mining/oil industries, we selected piling (or pile driving) as it is common in both offshore drilling operations for oil rig construction [43] and land-based construction [44]. Also, for the construction industry, we selected concrete mixing operations as concrete is universally used in many construction sites, and in many sites, the concrete is mixed onsite. For forestry/agriculture industries, we selected weeding and chainsaw operation as the representative ambient noise. Finally, we wanted to select an activity that is universally common across all industries; hence, we selected hammering. All noise samples are downloaded from FreeSound.[2] In total there are 8 different samples with white noise and noise-free samples included. It is worth noting that extra noise sources can be added, and new models can be trained easily via transfer learning techniques. In total, we have 6 different occupational background noises (Fig. 1): (1–2) high and low-frequency piling, (3) concrete mixer, (4) hammering, (5) weeder, and (6) chainsaw. As a control, we added a white noise background containing frequencies with equal intensities (Fig. 2).

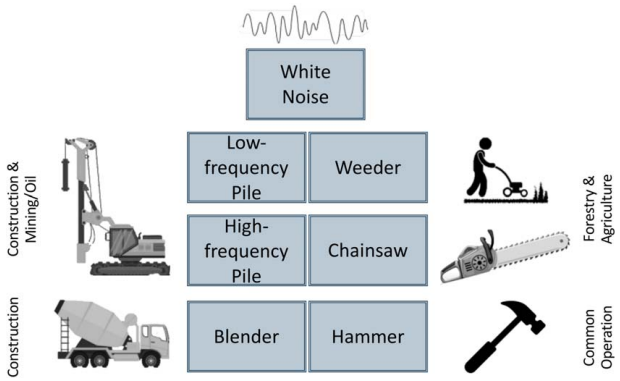

Fig. 1 Seven different occupational background noises

3.2 Command Word Dataset. The Speech Command Dataset provided by TensorFlow [32] is used as the basis of our voice command dataset (Fig. 2). It is one of the most widely used voice command datasets and has been used in previous studies extensively [45–48]. The dataset consists of repetition voice snippets of 35 different word commands from different participants. Each of the voice clips has approximately a 1-s duration. The speakers who gave the commands were native English speakers. The dataset also has a copy of the original dataset with white noise background. There are essentially two types of words in this dataset (Table 1). The first type of words is command words which indicate either a direction, action, quantity, or affirmation/negation. Examples of such words are up, down, one, two, stop, and yes. The second set of words are not command words but words that either have a similar pronunciation to the command words (three versus tree) or other words that cover different types of pronunciations, such as visual/sheila and happy/house. We believe by adding more words with different vowel sounds, we have increased the versatility of the algorithm. This is important as not everybody will pronounce the words in this database the same way, especially people with English as their second language [49]. We randomly split out our data in an 8/2 split; that is, each of the 36 trials will have 80% data in the training set and 20% in the testing set.

3.3 Voice and Noise Mixture. To test how noise under specific occupational settings will affect the performance of deep learning models, we needed to create new datasets that had the ambient noise overlaid on top of the clean dataset. We hypothesized two factors that will influence model performance: noise pattern and volume level. We determined an origin level for each noise which was between 50 and 60 dB. Then, noise levels that were 10 times (+10 dB) and 100 times (+20 dB) louder were tested. This was also done at the opposite trend (10 and 100 times less loud). Overall, the new dataset has 36 different background noise instances (seven types of background noise that claim at Sec. 3.1 × five volume levels, plus the original database).

[2]https://freesound.org

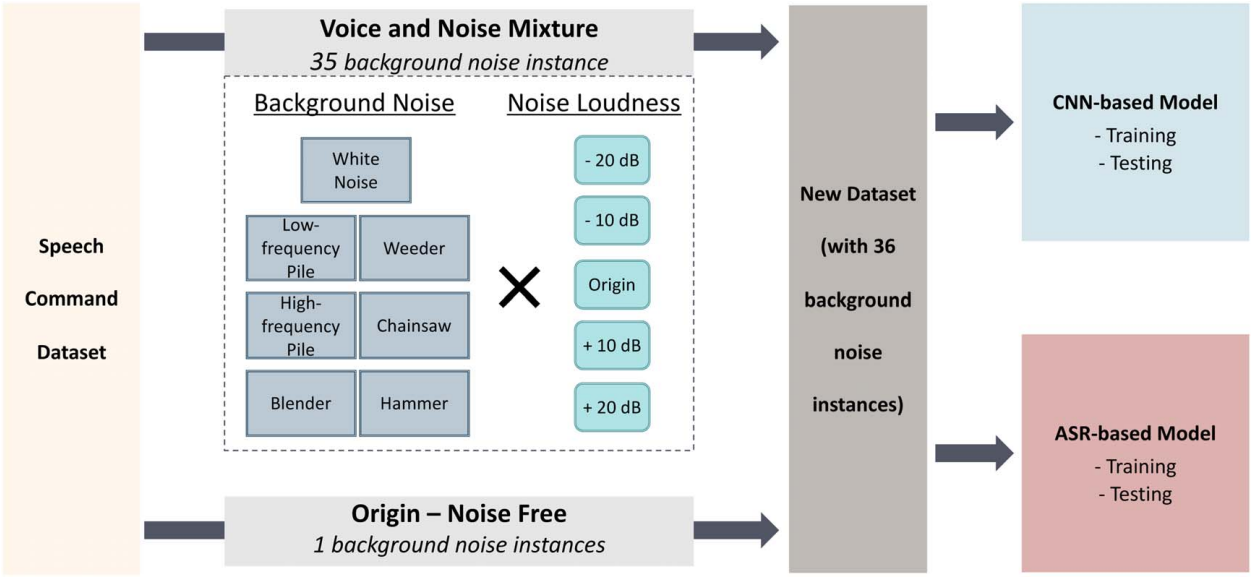

Fig. 2 The workflow of the experiment

Table 1 List of command words

Command words						
Backward	Bed	Bird	Cat	Dog	Down	Eight
Five	Follow	Forward	Four	Go	Happy	House
Learn	Left	Marvin	Nine	No	Off	On
One	Right	Seven	Sheila	Six	Stop	Three
Tree	Two	Up	Visual	Wow	Yes	Zero

3.4 Deep Learning Models Performance on New Datasets. We chose Convolutional Neural Network (CNN) as the baseline model of our new datasets as in the command set study, we have selected it as the baseline model. Additionally, in that study, CNN had a strong performance on classification tasks [32]. We built our own CNN using Tensorflow and PYTHON (Fig. 3). A noise-free reduction design was added to this model. We first applied a short-time Fourier transform (STFT) to transform the waveform into a spectrogram to expand features of the input audio before inputting it into the model. The model design was based on VGG16 that always has convolution layers of a 3×3 filter with a stride 1 and maxpool layer of a 2×2 filter of stride 2. The model was first trained and tested on all 36 backgrounds separately to evaluate how much performance will the noise affect the model using the CNN model. After that, we use the same model design to train the new dataset based on our new ASR-based model with all 36 background noises together and see if the test result has a performance increase. We developed our ASR model based on the Kaldi, the most widely used open-source ASR development framework. We use a similar ASR design to DeepSpeech2 [50]. It has a straightforward design of CNN layer, long short-term memory (LSTM) layer, Fully connected Layer and Softmax, respectively (Fig. 4) We also created a decoder to decode and get the final output. We wanted to show that End-to-End models perform better on our new datasets than baseline methods such as CNN. Modern E2E models may have too many blackbox functions that are hard to explain. So, we chose this naive and understandable model design.

3.5 Specific Command Word Performance. After we determined the accuracy levels of the models dependent on the background noise type, we wanted to know the performance for specific command words within the dataset. We selected four pairs

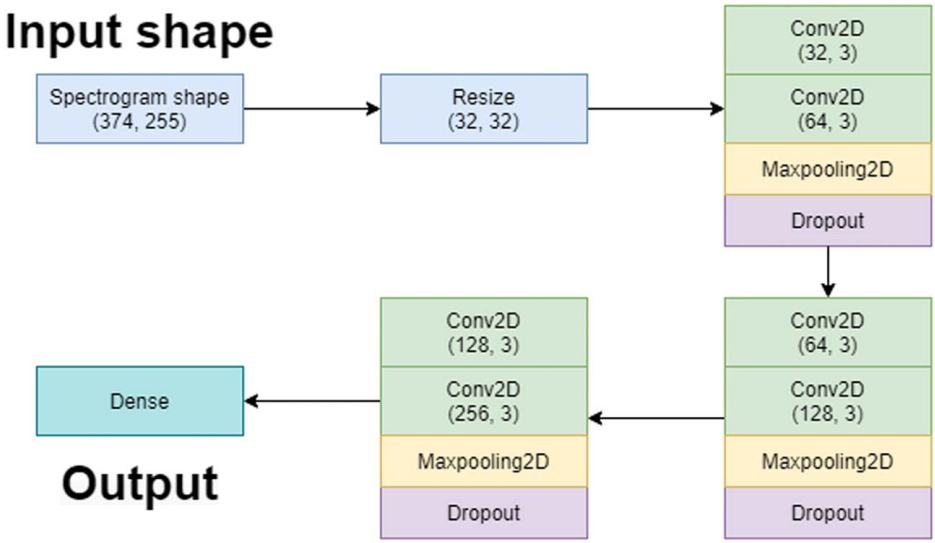

Fig. 3 CNN model structure

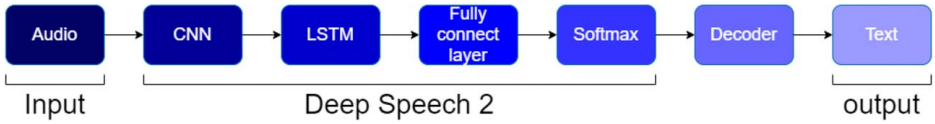

Fig. 4 ASR model structure

Table 2 Recognition accuracy of the CNN-based model

	−20 dB	−10 dB	Origin	+10 dB	+20 dB
Noise free	0.89				
White noise	0.81	0.62	0.49	0.37	0.24
Low-frequency pile	0.78	0.51	0.46	0.31	0.18
High-frequency pile	0.77	0.5	0.42	0.29	0.18
Blender	0.69	0.42	0.34	0.25	0.15
Weeder	0.68	0.41	0.33	0.24	0.15
Chainsaw	0.71	0.44	0.36	0.26	0.16
Hammer	0.79	0.54	0.48	0.33	0.21

of opposite command words: up and down, left, and right, stop and go, and finally yes and no. In order not to influence the performance in any other fashion, we kept the type of background noise (blender) and the level of noise constant and constructed two confusion matrices for each model. Confusion matrices are used to compare directly between to different words about their True Positives, False Positives, True Negatives, and False Negatives. They can clearly show which two commands are mixing up together. Hence by looking at the confusion matrices, we determined which command words should be used together in a real-world scenario.

4 Results

Tables 1 and 2 present the comparison between the CNN model and our ASR model using a pretrained DeepSPeech2 model on the Kaldi framework. The volume of the original sound was controlled between 50–60 dB and four different loudness levels were tested: +10 dB, +20 dB and −10 dB, −20 dB, respectively. Both models performed at higher levels of accuracy (89% and 91% for CNN and our model respectively) under noise-free conditions.

Other background noise types had more of an effect compared to white noise in all instances, with the weeder (Fig. 7) and blender noise (Fig. 8) having a 68% and 69% accuracy at the lowest noise level for CNN and 76% and 81% for ASR, compared to the 81% and 88% accuracy levels, respectively, for white noise at the same level. The −10 dB level also represented a precipitous drop for all selected background noises, with a 27% drop in accuracy level across all noise types for the CNN, but the drop was not as severe for ASR values. We have also depicted the declining trend of both ASR and CNN models for all background noises in Figs. 5–8.

Figures 9 and 10 show the confusion matrices for the CNN and ASR models, respectively. For the CNN model, the top three words with the highest instances of correct identification were for "down," "no," and "up" with correct instances ranging from 50 to 65. The ASR model's top three was also the same as the CNN model's however the best-performing word was "no" compared to "down" in CNN. The word that showed the poorest performance was "right," with only 15 instances of correct recognition of the CNN model and 34 for the ASR model. There was a non-negligible number of mischaracterizations for the top-performing words though. For example, for the ASR model, compared to the high number of characterizations for "no," there were 29 instances where "no" was recognized as "go." The same phenomenon was true for the CNN model, with 23 "go" recognitions compared to 55 correct "no" recognitions. In words with lower performances, this disparity was even more pronounced. The word "right" had 12 instances where it was recognized as "yes" of the CNN model compared to 15 correct instances. The ratio for the ASR model was better, with 16 "yes" recognitions to 32 correct ones.

5 Discussion

The purpose of this study is to demonstrate the need to develop dedicated voice recognition algorithms for ambient noise conditions that are typically found in occupations settings that have high levels of noise as typical voice recognition algorithms perform poorly in these conditions. As such in our study, we have investigated two things, the effect of different types of background noise on model performance and comparing the performance of a rough CNN-based voice recognition model to our model which was an ASR module with a denoising module. The discussion of the effect of different background noises in industrial settings is rare in the voice recognition domain with many studies focusing on civilian settings. Our study represents one of the first attempts to

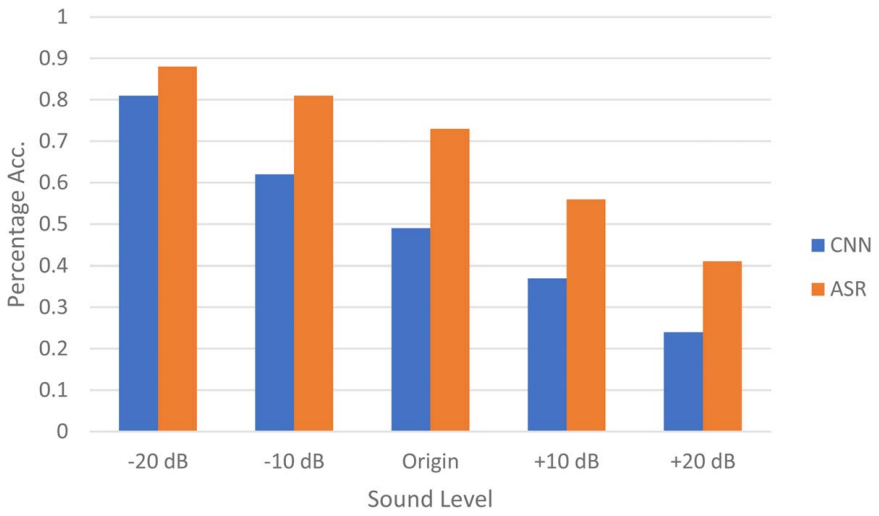

Fig. 5 Recognition accuracy for white noise

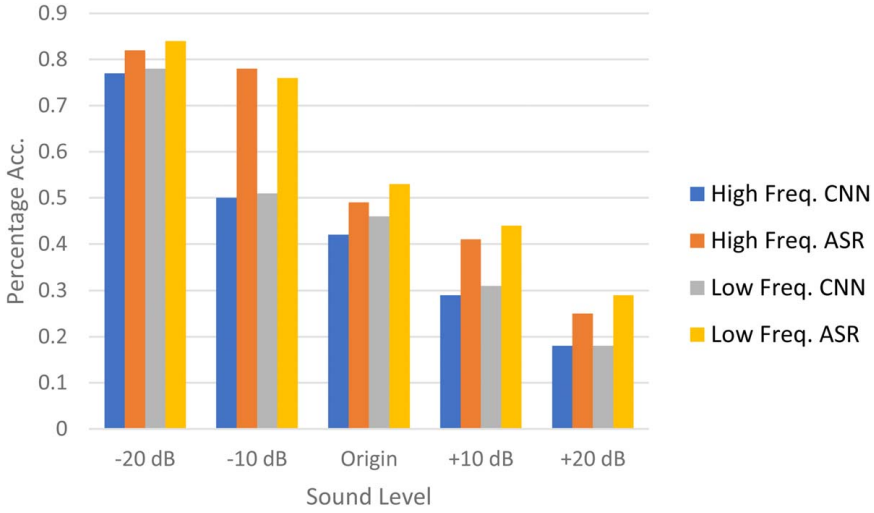

Fig. 6 Recognition accuracy for high- and low-frequency pile driver

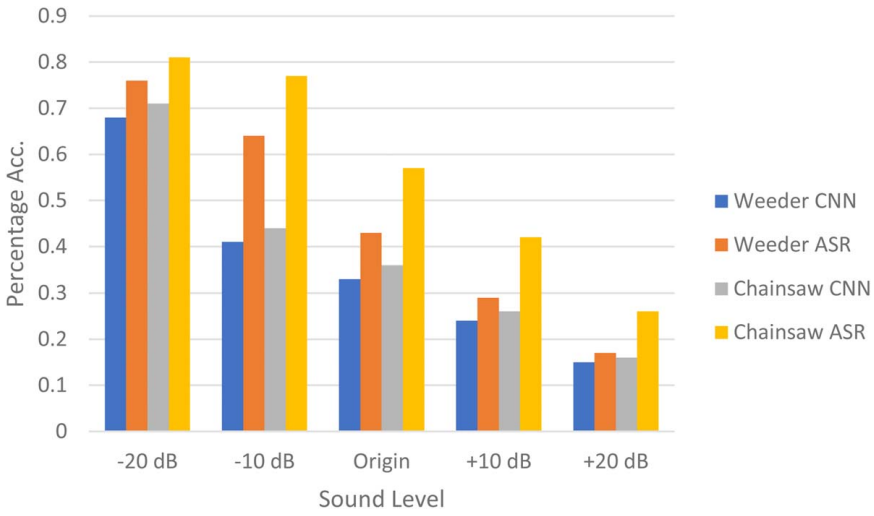

Fig. 7 Recognition accuracy for weeder and chainsaw

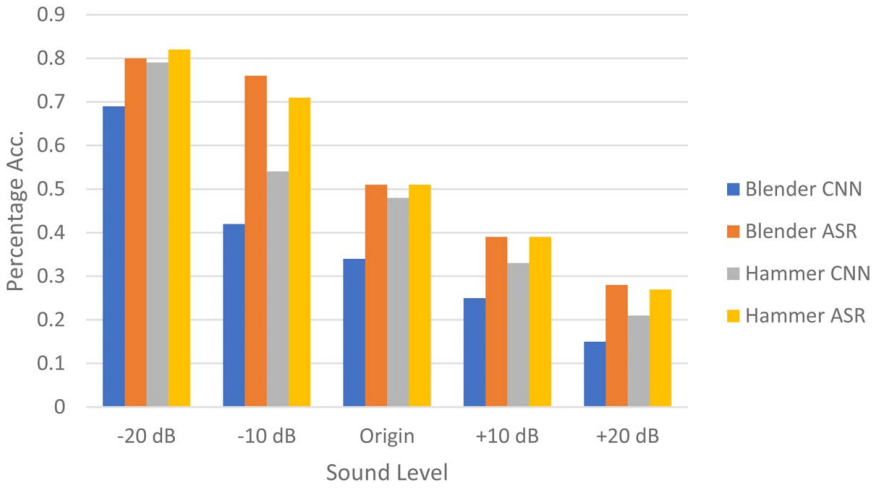

Fig. 8 Recognition accuracy for blender and hammer

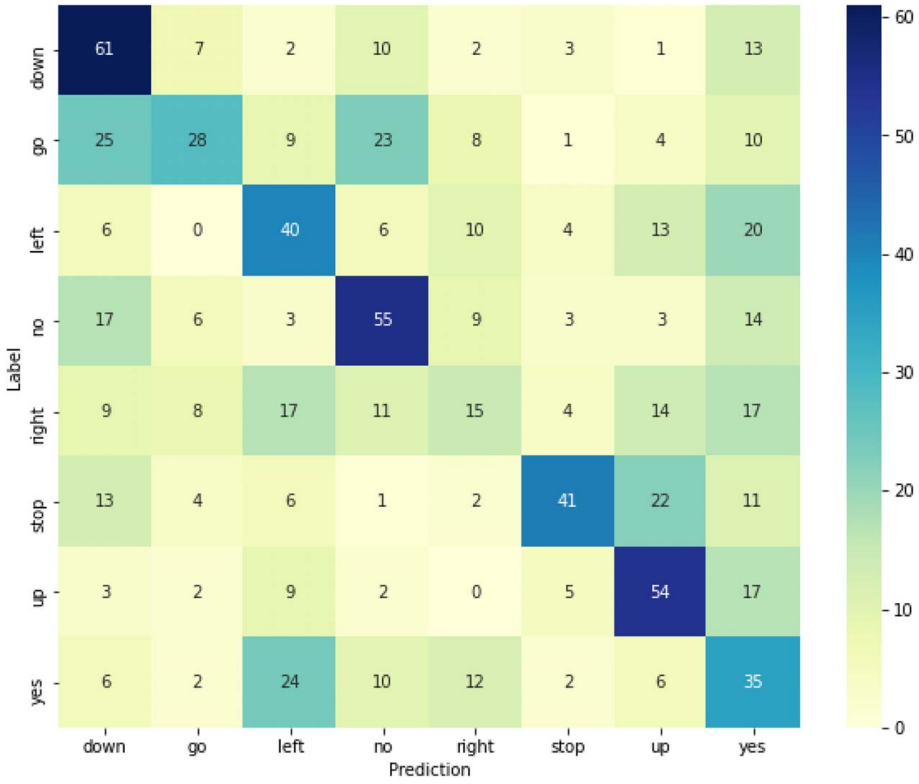

Fig. 9 Confusion matrix for the CNN model (blender, no dB adjustment)

display how different industrial background noises will need different voice recognition algorithms.

As we expected before the study, the type and the level of background noise had a significant impact on model performance. Both models performed well under noise-free conditions with high levels of accuracy. Our accuracy levels match existing studies that have used neural networks to build voice recognition algorithms. Hamza et al. for their voice recognition model with Gaussian Mixture Models reported 100% accuracy in their model [51]. In a similar study, Song et al. reported 95% accuracy in a quiet office

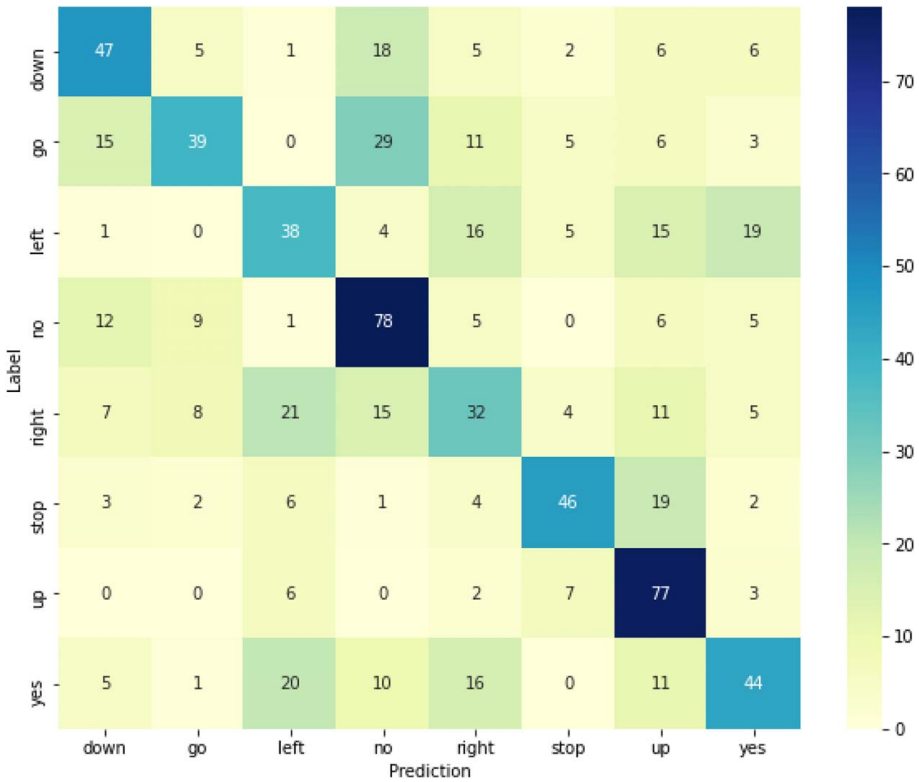

Fig. 10 Confusion matrix for the ASR model (blender, no dB adjustment)

Table 3 Recognition accuracy of the ASR-based model

	−20 dB	−10 dB	Origin	+10 dB	+20 dB
Noise free	0.91				
White noise	0.88	0.81	0.73	0.56	0.41
Low-frequency pile	0.84	0.76	0.53	0.44	0.29
High-frequency pile	0.82	0.78	0.49	0.41	0.25
Blender	0.8	0.76	0.51	0.39	0.28
Weeder	0.76	0.64	0.43	0.29	0.17
Chainsaw	0.81	0.77	0.57	0.42	0.26
Hammer	0.82	0.71	0.51	0.39	0.27

Table 4 Percentage performance increase for ASR model compared to CNN

	−20 dB	−10 dB	Origin	+10 dB	+20 dB	Average
White noise	7.95	23.46	32.88	33.93	41.46	27.94
Low-frequency pile	7.14	32.89	13.21	29.55	37.93	24.14
High-frequency pile	6.10	35.90	14.29	29.27	28.00	22.71
Blender	13.75	44.74	33.33	35.90	46.43	34.83
Weeder	10.53	35.94	23.26	17.24	11.76	19.75
Chainsaw	12.35	42.86	36.84	38.10	38.46	33.72
Hammer	3.66	23.94	5.88	15.38	22.22	14.22

setting with their Hidden Markov Model-based model [52]. The introduction of white noise did affect the accuracy of the models but not as much, at least in lower noise levels. Our industrial background noises, however, had a more significant effect on model accuracy even at the lowest sound level we selected. This effect was even more pronounced as the sound levels were increased. This means that using white noise to simulate noise in occupational settings is not a feasible approach. Another important point is that in higher levels of noise both models performed poorly. Considering typical noise levels in the industries we included, this problem needs to be alleviated before any voice-activated systems are implemented.

An additional issue that is present is the accuracy difference between the background noises that we included in our study, regardless of the noise level. At the −20 dB noise level for the CNN-based model, there is a 10% difference between the highest (hammer) and the lowest (weeder). This difference is also present for the ASR-based model but not as pronounced. Although the accuracy values for both models seem to converge at higher noise levels, the difference between the lowest and the highest accuracy values does not get below 5%. We surmise that the different sound profiles of each background noise contribute to this difference. Taking the two extremes as an example, weeder noise and hammering noise, the weeder noise sample has higher frequency levels, and the noise is produced continuously, whereas the hammer noise has a lower frequency level, and it is produced in fixed intervals. All these noise profiles have practical representations and cannot be replaced with simple white noise during the model training phase. As such, voice recognition algorithms and denoising solutions need to be tailored to the specific noise profile of the environment they will be applied to.

We also identified from the results of the confusion matrices the importance of selecting command words that will not negatively affect each other in terms of recognition accuracy. If the algorithm cannot distinguish between the commands "no" and "go," for example, this might result in the machine activating in such a manner that could pose danger to the worker. From our observations, we determined that the biggest contributor to misrecognition is the similarity of vowel sounds in words. Pairs such as left-yes, up-stop, and no-go had either close to or higher than 20 instances of misrecognition. Hence, extra attention needs to be given to selecting words that have sufficiently different sounds from each other. On top of words affecting each other, the voice recognition accuracy of the words themselves needs to be considered while populating selecting command words. If one word consistently underperforms in a set of conditions, alternative phrases need to be considered, and this process should happen before the algorithm is fully deployed.

The ASR model is built using the Kaldi speech recognition toolkit. More features have been extracted from the audio file using the toolkit automatically and the data can be trained on pretrained ASR models. The pretrained model we chose is a deep neural network (DNN) model called DeepSpeech2 [50]. The advantage of the ASR model compared to normal deep learning models like CNN is that it uses end-to-end speech recognition to substitute multiple complicated feature engineering modules such as alignment, clustering, and Hidden Markov Models (HMM) that let the model structure be more stable and reasonable and can be trained bigger. As a result, compared to our CNN model, which only has 1.6 million parameters, the DeepSpeech2 has 35 million parameters. More parameters allow the model to extract more features even under noisy environments. The results of the ASR model have a large improvement compared to the CNN model, especially under large noise conditions. This demonstrates that even with a standard denoising model, model accuracy increases regardless of background noise type. It is important to note the +20 dB level proved challenging for both models. As industrial machines tend to produce sound levels that are higher than 70 dB, this accuracy issue needs to be resolved before any type of voice-activated machine control measure is implemented.

6 Limitations and Future Work

There are several limitations in the study that need to be mentioned. To start, we used pre-recorded samples as our background noises for the model. Although this is convenient in terms of testing multiple types of background noises at once, they do not completely encapsulate the occupational settings they occupy. For example, there could be both a pile driver and a blender working at the same time at a construction site. The same thing could be said for chainsaws and weeders for the forestry industry. The combination of these background noises would give a more accurate representation of the ambient noise that is present. A related limitation to this is our use of singular noise sources for model training. As mentioned, there are multiple noise sources in every workplace, and using a representative noise sample will undoubtedly affect performance. To counteract both limitations, we plan to record samples

at their respective occupational settings in our future model development. Another limitation is our limited use of different machine learning methods. As the focus of our study was to highlight the need to use industry-specific background noises, we wanted to keep the methodology consistent. However, in future studies, we will explore the different ML techniques and determine whether they have different performance levels.

7 Conclusion

In this study, we aimed to demonstrate the need to develop voice recognition algorithms that are designed for specific occupational settings. We focus on two aspects of voice recognition, the type of background noise representing different occupational settings and the effect of a denoising module on a machine learning-based voice recognition algorithm. We identified occupational settings that tend to have higher levels of ambient noise and selected the most ubiquitous items that belong to that setting. Then, we selected the phrases that we are going to use and the background noise levels. We compared the accuracy rates of a CNN-based voice control algorithm and an ASR-based model with a standard denoising module. After this, we selected a subset of the command word dataset and investigated the accuracy rates of specific words using confusion matrices for the two models. We found that using white noise to train algorithms to represent background noises from occupational settings would not be accurate, as the background noises we selected affected the accuracy rates more than white noise. We also found that even a standard denoising module had a positive effect on the accuracy level, with our ASR module that had the denoising module outperforming the CNN-based model on average by 14–35% across all background noises. The improvement percentages did vary across different noise levels/background noise combinations which warrant further investigation. The confusion matrices showed differences between words that were sometimes quite significant, and it the effect on accuracy some words could have over other, similar sounding command words. These results indicated that to apply voice recognition-based controls in occupational settings, the noise level, the noise profile of the environment, the machine that is controlled by voice recognition, and the selection of command words need to be considered to create algorithms that are tailored to them.

Funding Data

- This work has been supported by the NRI: FND: Investigating the Safety Challenges of Co-drones in Future Construction Workplaces (NSF Grant No. 2024656) and FW-HTF-RL: Collaborative Research: The Future of Remanufacturing: Human–Robot Collaboration for Disassembly of End-of-Use Products (NSF Grant No. 2026276).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request. The data and information that support the findings of this article are freely available online.[3] The authors attest that all data for this study are included in the paper.

[3]https://www.ise.ufl.edu/hu/

References

[1] Uddin, M. M., Huynh, N., Vidal, J. M., Taaffe, K. M., Fredendall, L. D., and Greenstein, J. S., 2015, "Evaluation of Google's Voice Recognition and Sentence Classification for Health Care Applications," Eng. Manage. J., **27**(3), pp. 152–162.

[2] Cevher, D., Zepf, S., and Klinger, R., 2019, "Towards Multimodal Emotion Recognition in German Speech Events in Cars Using Transfer Learning." arXiv preprint arXiv:1909.02764.

[3] Mittal, Y., Toshniwal, P., Sharma, S., Singhal, D., Gupta, R., and Mittal, V. K., 2015, "A Voice-Controlled Multi-Functional Smart Home Automation System," Proceedings of the 2015 Annual IEEE India Conference (INDICON), IEEE, pp. 1–6.

[4] Meticulous Market Research, 2019, "Speech and Voice Recognition Market by Type (SPEECH and Voice Recognition), End User (Automotive, Healthcare, BFSI, EDUCATION, Legal), Technology (Artificial Intelligence and NON-ARTIFICIAL Intelligence), and Geography—Global Forecast to 2025." Speech and Voice Recognition Market | Meticulous Market Research Pvt. Ltd. https://www.meticulousresearch.com/product/speech-and-voice-recognition-market-5038

[5] Rogowski, A., 2012, "Industrially Oriented Voice Control System," Robot. Comput.-Integr. Manuf., **28**(3), pp. 303–315.

[6] Brauer, R. L., 2016, *Safety and Health for Engineers*, John Wiley & Sons, Hoboken, NJ.

[7] Tilley, J., 2020, "Automation, Robotics, and the Factory of the Future," McKinsey & Company, https://www.mckinsey.com/business-functions/operations/our-insights/automation-robotics-and-the-factory-of-the-future, Accessed October 20.

[8] Longo, F., Nicoletti, L., and Padovano, A., 2017, "Smart Operators in Industry 4.0: A Human-Centered Approach to Enhance Operators' Capabilities and Competencies Within the New Smart Factory Context," Comput. Ind. Eng., **113**, pp. 144–159.

[9] Cohen, P. R., and Oviatt, S. L., 1995, "The Role of Voice Input for Human-Machine Communication," Proc. Natl. Acad. Sci. U. S. A., **92**(22), pp. 9921–9927.

[10] Longo, F., and Padovano, A., 2020, "Voice-Enabled Assistants of the Operator 4.0 in the Social Smart Factory: Prospective Role and Challenges for an Advanced Human–Machine Interaction," Manuf. Lett., **26**, pp. 12–16.

[11] Rains, G. C., 2014, "Emergency Tractor Shut-Off Using a Voice Command System," 2014 Montreal, Quebec Canada, American Society of Agricultural and Biological Engineers, July 13–16, p. 1.

[12] Valenzuela, V. E., Lauria, V. F., Lucena, P. P., Jazdi, N., and Göhner, P., 2013, "Voice-Activated System to Remotely Control Industrial and Building Automation Systems Using Cloud Computing," Proceedings of the 2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA), Cagliari, Italy, Sept. 10–13, IEEE, pp. 1–4.

[13] Solorio, J. A., Garcia-Bravo, J. M., and Newell, B. A., 2018, "Voice Activated Semi-autonomous Vehicle Using Off the Shelf Home Automation Hardware," IEEE Internet Things J., **5**(6), pp. 5046–5054.

[14] Pleva, M., Juhar, J., Ondas, S., Hudson, C. R., Bethel, C. L., and Carruth, D. W., 2019, "Novice User Experiences With a Voice-Enabled Human–Robot Interaction Tool," Proceedings of the 2019 29th International Conference Radioelektronika (Radioelektronika), Pardubice, Czech Republic, Apr. 16–18, IEEE, pp. 1–5.

[15] Lee, S. J., Kang, B. O., Jung, H. Y., Lee, Y., and Kim, H. S., 2010, "Statistical Model-Based Noise Reduction Approach for Car Interior Applications to Speech Recognition," ETRI J., **32**(5), pp. 801–809.

[16] Sokol, N., Chen, E. Y., and Donmez, B., 2017, "Voice-Controlled In-Vehicle Systems: Effects of Voice-Recognition Accuracy in the Presence of Background Noise," Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: Driving Assessment 2017, Manchester Village, VT, June 26–29, pp. 158–164.

[17] Czap, L., and Pinter, J. M., 2018, "Noise Reduction for Voice-Activated Car Commands," *Vehicle and Automotive Engineering*, Springer, Cham, pp. 351–358.

[18] Tamoto, A., and Itou, K., 2019, "Voice Authentication by Text Dependent Single Utterance for In-Car Environment," Proceedings of the Tenth International Symposium on Information and Communication Technology, Hanoi and Halong, Vietnam, Dec. 4–6, pp. 336–341.

[19] Sachdev, S., Macwan, J., Patel, C., and Doshi, N., 2019, "Voice-Controlled Autonomous Vehicle Using IoT," Procedia Comput. Sci., **160**, pp. 712–717.

[20] Susanto, D., Mujaahid, F., Syahputra, R., and Putra, K. T., 2019, "Open Source System for Smart Home Devices Based on Smartphone Virtual Assistant," J. Electr. Eng. UMY, **3**(1), pp. 1–7.

[21] Orlandic, L., Teijeiro, T., and Atienza, D., 2021, "The COUGHVID Crowdsourcing Dataset, a Corpus for the Study of Large-Scale Cough Analysis Algorithms," Sci. Data, **8**(1), pp. 1–10.

[22] Davis, G., 1978, "Noise and Vibration Hazards in Chainsaw Operations: A Review," Aust. For., **41**(3), pp. 153–159.

[23] Ghai, W., and Singh, N., 2012, "Literature Review on Automatic Speech Recognition," Int. J. Comput. Appl., **41**(8), pp. 42–50.

[24] Deng, L., and Li, X., 2013, ""Machine Learning Paradigms for Speech Recognition: An Overview," IEEE/ACM Trans. Audio Speech Lang. Process., **21**(5), pp. 1060–1089.

[25] Ouisaadane, A., Safi, S., and Frikel, M., 2019, "English Spoken Digits Database Under Noise Conditions for Research: SDDN," Proceedings of the 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, Apr. 3–5, IEEE, pp. 1–5.

[26] Bach, J.-H., Kollmeier, B., and Anemüller, J., 2010, "Modulation-Based Detection of Speech in Real Background Noise: Generalization to Novel Background Classes," Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, Mar. 15–19, IEEE, pp. 41–44.
[27] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H., 2014, "Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks," Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, Sept. 14–18, pp. 2670–2674.
[28] Krishna, G., Tran, C., Yu, J., and Tewfik, A. H., 2019, "Speech Recognition with no Speech or with Noisy Speech," Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12–17, IEEE, pp. 1090–1094.
[29] Chan, W., Jaitly, N., Le, Q., and Vinyals, O., 2016, "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition," Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 20–25, IEEE, pp. 4960–4964.
[30] Sutskever, I., Vinyals, O., and Le, Q. V., 2014, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems, Montreal, Canada, Dec. 8–13, pp. 3104–3112.
[31] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., 2014, "Learning Phrase Representations Using RNN Encoder-Decoder For Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1724–1734.
[32] Mozilla, 2017, "Common Voice by Mozilla." Common Voice. https://voice.mozilla.org/en
[33] Warden, P., 2018, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition." arXiv preprint arXiv:1804.03209.
[34] Reddy, C. K., Beyrami, E., Pool, J., Cutler, R., Srinivasan, S., and Gehrke, J., 2019, "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," Interspeech.
[35] Flamme, G. A., Stephenson, M. R., Deiters, K., Tatro, A., Van Gessel, D., Geda, K., Wyllys, K., and McGregor, K., 2012, "Typical Noise Exposure in Daily Life," Int. J. Audiol., **51**(1), pp. S3–S11.
[36] Birch, B., Griffiths, C. A., and Morgan, A., 2021, "Environmental Effects on Reliability and Accuracy of MFCC Based Voice Recognition for Industrial Human-Robot-Interaction," Proc. Inst. Mech. Eng. B: J. Eng. Manuf., **235**.
[37] Bingol, M. C., and Aydogmus, O., 2020, "Performing Predefined Tasks Using the Human–Robot Interaction on Speech Recognition for an Industrial Robot," Eng. Appl. Artif. Intell., **95**, p. 103903.
[38] Valin, J.-M., 2018, "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," Proceedings of the 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), Vancouver Canada, Aug. 29–31, IEEE, pp. 1–5.
[39] Rethage, D., Pons, J., and Serra, X., 2018, "A Wavenet for Speech Denoising," Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, Apr. 15–20, IEEE, pp. 5069–5073.
[40] Pascual, S., Bonafonte, A., and Serra, J., 2017, "SEGAN: Speech Enhancement Generative Adversarial Network." arXiv preprint arXiv:1703.09452.
[41] Rabinowitz, P. M., Galusha, D., Dixon-Ernst, C., Slade, M. D., and Cullen, M. R., 2007, "Do Ambient Noise Exposure Levels Predict Hearing Loss in a Modern Industrial Cohort?," Occup. Environ. Med., **64**(1), pp. 53–59.
[42] NIOSH, 2019, Overall Statistics—All U.S. Industries—Ohl." Centers for Disease Control and Prevention, August 27. https://www.cdc.gov/niosh/topics/ohl/overall.html
[43] Bailey, H., Senior, B., Simmons, D., Rusin, J., Picken, G., and Thompson, P. M., 2010, "Assessing Underwater Noise Levels During Pile-Driving at an Offshore Windfarm and its Potential Effects on Marine Mammals," Mar. Pollut. Bull., **60**(6), pp. 888–897.
[44] Fleming, K., Weltman, A., Randolph, M., and Elson, K., 2008, *Piling Engineering*, CRC Press.
[45] Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T., and Dureau, J., 2019, "Federated Learning for Keyword Spotting," Proceedings of the InICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12–17, IEEE, May 12, 2019, pp. 6341–6345.
[46] Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., and Bengio, Y., 2019, Speech Model Pre-Training for End-to-End Spoken Language Understanding, Proceedings of Interspeech 2019, pp. 814–818. .
[47] de Andrade, D. C., Leo, S., Viana, M. L., and Bernkopf, C., 2018, A Neural Attention Model for Speech Command Recognition. arXiv preprint arXiv:1808.08929.
[48] Kim, T., Lee, J., and Nam, J., 2019, "Comparison and Analysis of Sample CNN Architectures for Audio Classification," IEEE J. Sel. Top. Signal Process., **13**(2), pp. 285–297.
[49] Coniam, D., 1999, "Voice Recognition Software Accuracy With Second Language Speakers of English," System, **27**(1), pp. 49–64.
[50] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., and Casper, J., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," International Conference on Machine Learning, New York City, June 19–24, pp. 173–182, PMLR, 2016.
[51] Hamza, M., Khodadadi, T., and Palaniappan, S., 2020, "A Novel Automatic Voice Recognition System Based on Text-Independent in a Noisy Environment," Int. J. Electr. Comput. Eng., **10**(4), pp. 3643–3650. .
[52] Song, J., Chen, B., Jiang, K., Yang, M., and Xiao, X., 2019, "The Software System Implementation of Speech Command Recognizer Under Intensive Background Noise," IOP Conference Series: Materials Science and Engineering, Changsha, China, Apr. 19–21, IOP Publishing, Vol. 563, no. 5, p. 052090.