university of groningen

faculty of science and engineering

Enhancing Automatic Speech Recognition in Vehicular Environments:

A Noise-Specific Fine-Tuning Approach

Dongwen Zhu

university of groningen

faculty of science and engineering

University of Groningen

Enhancing Automatic Speech Recognition in Vehicular Environments: A Noise-Specific Fine-Tuning Approach

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Matt Coler (Campus Fryslân, University of Groningen)

Dongwen Zhu (s5505925)

June 17, 2024

Contents

Page

Acknowledgements **4**

Abstract **5**

1 Introduction **6**

2 Background Literature **9**
2.1 Historical Developments of Speech Recognition 9
2.2 Current SOTA Techniques 10
2.3 Noise Robustness in ASR 12

3 Methodology **14**
3.1 Data 14
3.1.1 LibriSpeech 14
3.1.2 Vehicular Noise 15
3.1.3 Public Other Noise 16
3.1.4 Dataset Processing 17
3.2 Experimental Settings 17
3.2.1 Baseline Model 17
3.2.2 Fine-tuning 18
3.2.3 Evaluation 19

4 Results **22**
4.1 Model Performance 22
4.2 Answering Research Question 22
4.2.1 Research Question Analysis 22
4.2.2 Hypotheses Validation 24
4.3 Checking for Overfitting 25
4.4 Statistic Analyses 25

5 Discussion **27**
5.1 Limitations 27
5.2 Future Research 29
5.2.1 Development of datasets 29
5.2.2 Improvement of Evaluation Metrics 29
5.2.3 Future Practical Applications 30

6 Conclusion **31**

7 Ethics **32**

References **33**

Acknowledgments

Time goes too fast and my master's life finishes before I am well prepared to enter society.

It's a hard time for me to finish this thesis in such a short time and huge stress. Just one year ago, I had no fundamental knowledge of language and coding, and now I must complete this work entirely on my own.

Despite the difficulties in my studies and the constraints of time, my life in the Netherlands has been incredibly fulfilling and perhaps the best period of my life. I am deeply thankful to all my friends who have made me feel at home, helped me navigate through all challenges, and offered comfort when I felt stressed and sad. Expressing "thank you" and "I love you" does not come easily to me, but your support has been instrumental in my growth.

I would also like to express my sincere gratitude to my professors — Matt, Vass, Shekhar, Joshua, and Phat — for their invaluable guidance and support throughout my academic journey.

Abstract

The expansion of in-vehicle technologies has made it necessary for the development of advanced automatic speech recognition (ASR) systems that are capable of operating efficiently in noisy environments. This thesis explores the enhancement of ASR systems through fine-tuning for specific noise conditions, particularly focusing on vehicular noise environments. The research investigates whether ASR models fine-tuned with noise samples specific to a vehicular environment demonstrate superior performance compared to models that are generalized for noise robustness.

Using the "wav2vec2-base-960h" model pre-trained on the LibriSpeech corpus as the baseline model, this study conducts the fine-tuning experiments with two distinct noise datasets: Vehicular Noise Speech and Public Other Noise Speech. The performance of these three models - the baseline model, the model fine-tuned by vehicular noise, and the model fine-tuned by public other noise, is evaluated across three same noise conditions to ascertain their effectiveness in real-world scenarios. The results indicate that models fine-tuned on specific noise environments significantly outperform the general noise-robust model in their targeted settings.

This study contributes to the field by demonstrating the potential of environment-specific fine-tuning in enhancing ASR performance in noise-affected conditions. The findings could influence future ASR applications in vehicular systems, ensuring more reliable speech recognition and improving user interaction with in-vehicle electronics.

Key Words: ASR, vehicular environment, wav2vec 2.0, fine-tune, noise robustness

Chapter 1

Introduction

Voice is the most natural and common method of information transfer among humans; when used for human-computer interaction, voice interaction has significant advantages over manual control. In recent years, with the rapid development and application of neural networks, particularly deep learning, the field of speech recognition has made numerous advances, and the performance of speech recognition models has greatly improved. It has many applications, such as in voice assistants for computers and mobile phones, or for command control in wearable smart devices, smart homes, and in-vehicle electronics, all of which facilitate user interaction and control of electronic devices. Users can operate these devices simply by speaking commands. This is especially important for small devices that are inconvenient to operate manually due to the absence of a mouse, keyboard, or touchscreen. Additionally, for individuals with disabilities such as those who are physically unable to use their hands or are blind, speech recognition enables them to operate electronic devices. For example, Nuance has built the in-car voice interaction platform Dragon Drive to provide new fun for driving and meet the needs of future car travel. The in-car voice platform Dragon Drive is shown in Figure 1.

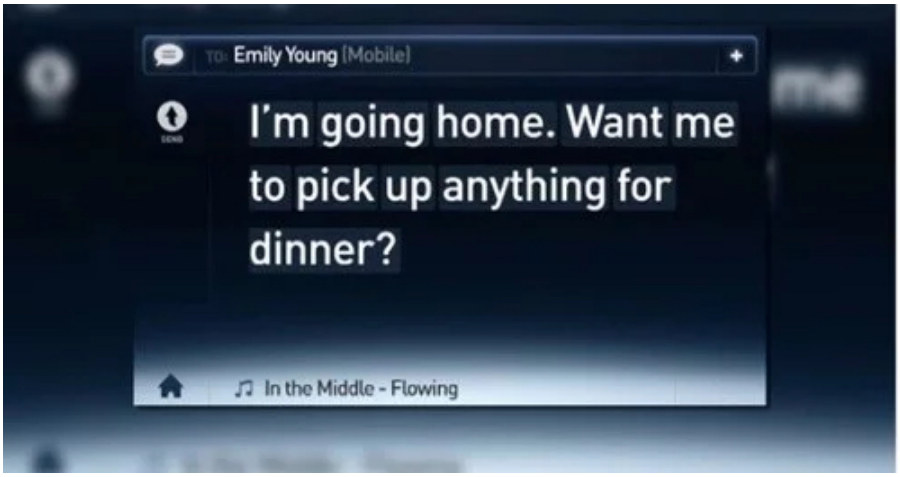

Figure 1: Car voice platform Dragon Drive

As people's material lives continue to improve, more individuals own cars and their expectations for vehicles are increasing, including demands for safety and convenience. Nowadays, many cars are equipped with in-vehicle electronics modules, including navigation systems and digital multimedia systems, which enable "human-vehicle" interaction. Drivers can safely send and receive messages, make phone calls, and navigate in real-time, making driving safer, more comfortable, and enjoyable. The voice interaction system, as a foundational component of various smart units in vehicles, allows drivers to control in-vehicle electronics more naturally and conveniently. It also reduces the distraction drivers face from operating these devices, thereby significantly enhancing driving safety.

Although many companies, such as Google, now provide high-performance speech recognition services, these are based on large-vocabulary continuous speech recognition technology, which involves large

models and complex computations (Lamel & Gauvain, 2022). Since speech recognition needs to run continuously, if used offline, the model must be directly integrated into the vehicle's electronics, which can heavily burden these systems. Moreover, although the in-vehicle smart system can accurately recognize and execute the driver's commands under ideal conditions, it can be interfered with by various types of noise in real life. For example, engine noise, conversations among passengers, and external environmental noises. These noise signals are complex and variable, often located in the low-frequency range and mixed with other speakers' signals. Commands issued by the driver can be contaminated during transmission, affecting the quality of the speech and thereby reducing the recognition rate of the in-vehicle voice interaction system, and even rendering it unusable.

Given these challenges, there is a compelling need to research and develop noise-resistant speech recognition models that are specifically tuned for vehicular environments. Such models must effectively handle both steady and non-steady noise types to ensure reliable performance.

Moreover, the field of ASR has seen substantial growth and improvement over the past decades. Current ASR systems are highly proficient under ideal conditions with well-represented languages. However, these conditions are seldom met in real-world scenarios where factors like diverse accents, various voice types, and background noises are prevalent. Despite this, most ASR systems are still trained on datasets that primarily consist of clear, accent-neutral speech, which fails to replicate real-world conditions adequately. Addressing these complex variables has become a focal point in the research community, aiming to develop more adaptable and inclusive ASR technologies.

This thesis examines the influence of noise on the effectiveness of ASR systems and investigates whether an ASR model fine-tuned to a specific dataset of data can achieve sufficient performance. It specifically examines ASR systems designed for vehicular environments and assesses their performance requirements within those contexts. Building on previous studies, such as the research by Schlotterbeck et al. (2022) which investigated the fine-tuning of an ASR model with classroom noise, this thesis extends that research by applying similar techniques across different noise conditions to evaluate their effectiveness. The goal is to determine whether fine-tuning an end-to-end ASR model to include environmental noise from a specific setting enhances its performance in that environment or if a more generalized approach to noise robustness could prove equally effective.

Research Questions:

Building on the foundational understanding of voice interaction and the advancement of speech recognition technologies discussed above, I now turn our attention to the specific challenges posed by environmental noises in ASR systems. The unique conditions of different noise environments necessitate models that can robustly interpret human speech, despite interference. Now let's delve into key research questions that aim to dissect the efficacy of ASR models fine-tuned for specific noise environments as opposed to those employing a generalized noise-robust approach. These questions are designed to rigorously test hypotheses about model performance in various noisy and noise-free contexts, reflecting real-world applications where such technologies are critical.

Q1. How does the performance of an ASR model fine-tuned for a specific noise environment compare to a general noise-robust ASR model when evaluated on data from the same noise environment?

Q2. How does the performance of an ASR model fine-tuned for a specific noise environment compare to a general noise-robust ASR model when evaluated on data from different noise environments?

Q3. How do two ASR models, each fine-tuned for specific but different noise environments, compare to a general noise-robust ASR model when evaluated on data that includes both environments?

Q4. How do two ASR models, each fine-tuned for specific noise environments, perform compared to a general noise-robust ASR model when evaluated on a clean (no noise) test dataset?

Hypotheses: To address these questions, the following hypotheses are proposed:

H1. The ASR model fine-tuned by a specific noise environment will perform better than the general noise-robust (no noise-robust) ASR model when evaluated on data from that particular noise environment.

H2. The ASR model fine-tuned by a specific noise environment will perform worse than the general noise-robust (no noise-robust) ASR model when evaluated on data from other noise environments.

H3. The two ASR models fine-tuned by a specific noise environment will perform better than the general noise-robust (no noise-robust) ASR model when evaluated on data from both environment noises.

H4. The two ASR models fine-tuned by a specific noise environment will perform worse than the general noise-robust (no noise-robust) ASR model when evaluated on data from a clean (no noise) test dataset.

The thesis is structured into six main chapters. The first chapter, Introduction, outlines the motivation and significance of enhancing ASR systems within vehicular environments, setting the stage with the objectives and scope of the study. In Chapter Two, Literature Review, reviews the background literature, highlighting key studies related to speech recognition development and ASR technologies. The third chapter, Methodology, details the experimental design including the datasets used—LibriSpeech, Vehicular Noise Speech, and Public Other Noises Speech—and the fine-tuning process of the baseline "wav2vec2-base-960h" model, along with the criteria for evaluating model performance. The Results, presented in Chapter Four, analyze the findings from the experiments, provide statistical validations, and discuss how these results address the initial research questions. Chapter Five, Discussion, interprets the results, acknowledges the study's limitations, and suggests future research directions that could further advance ASR technology. Finally, the sixth chapter, Conclusion, summarizes the key findings and their implications for future ASR technologies in vehicular environments, emphasizing the impact and potential of the study in enhancing user interaction and safety within noisy vehicular settings.

Chapter 2

Background Literature

In this literature review, I aim to explore the historical and current advancements in speech recognition technology, with a particular focus on the development and capabilities of the wav2vec 2.0 model. My methodology involved a comprehensive search of academic databases including IEEE Xplore, Google Scholar, and PubMed. The inclusion criteria focused on peer-reviewed articles, conference papers, and significant industry reports published in the last two decades, while exclusion criteria eliminated sources not directly relevant to ASR technologies or those that did not provide empirical data. This review first introduces the history of speech recognition, then delves into the development and advantages of various models leading up to wav2vec 2.0, and finally discusses noise robustness in ASR and the rationale for fine-tuning models to specific noisy environments.

2.1 Historical Developments of Speech Recognition

Today, speech recognition technology has entered a mature stage of development and has found widespread application in various fields. However, the successful rise of this technology was not achieved overnight. It has undergone more than half a century of updates and iterations, evolving from the early prototypes of speech recognition to a cutting-edge technology with vast development potential and application prospects. The research on speech recognition technology can be traced back to the 1950s. The specific development history is shown in Table 1.

Entering the 21st century, with continuous technological advancements, speech recognition technology truly entered its golden era of development. During this period, speech recognition technology began to serve human production and life, gradually forming a scale and entering the application market. In 2008, Google launched its first speech search software for Apple devices. The following year, Android 1.6 also included a text-to-speech function. In 2010, Apple collaborated with Nuance to develop the Siri voice assistant. In 2014, Microsoft introduced Cortana, and Amazon launched Alexa. By 2018, Amazon and Microsoft announced and completed the integration of their respective voice assistants, Alexa and Cortana.

Currently, speech recognition technology is no longer limited to feasibility in the development stage but focuses on improving recognition rates across different fields and specific environments (Irugalbandara, Naseem, Perera, Kiruthikan, & Logeeshan, 2023). Thus, speech recognition technology has entered a new phase of development, striving for both effectiveness and excellence. It has found extensive applications in diverse areas such as smart homes, voice payments, and clinical medicine. Gangmei, Singh, and Shougaijam (2021) proposed a voice interaction system for unmanned vending machines by integrating speech recognition with intelligent vending machines. This system enables a contactless, voice-activated shopping method, significantly reducing the risk of COVID-19 infection and transmission, thereby providing users with great convenience and safety. WANG, SHAN, and JING (2022) has introduced a voice ticketing feature in railway ticketing clients, achieving a recognition accuracy of up to 90%. This effectively addresses the difficulties elderly people face when purchasing tickets online. Alowais, Alghamdi, Alsuhebany, and et al. (2023) proposed replacing the traditional manual entry

Decade	Individual/ Company	Development
1950s	Bell Labs	Acoustic Spectrometer
(The Beginning of the		Audry System
Speech Recognition)	RCA Labs	Ten Monosyllabic Word Recognizer
(Le Prell & Clavier, 2017)	Fry and Denes et al.	Phoneme Recognizer
1960s	Martin	Time Normalization Method
(Significant Developments	Itakura et al.	Linear Prediction Coding (LPC)
in Speech Recognition)	Vintsyuk	Dynamic Programming (DP) Method
(Nwe, Foo, & De Silva, 2003)		
(Paulett & Langlotz, 2012)		
1970s	Sakoe	Dynamic Time-Warping (DTW)
(Breakthroughs Through	Linda et al.	Vector Quantization (VQ)
Theoretical Algorithms)	Philco-Ford	Real-time LPC Technology
(L. Zhang, 2020)		
1980s	Rabiner et al.	HMM Model
(Shift in Mainstream Research	Hopfield	Hopfield Neural Network Model
from Pattern Matching	Carnegie Mellon University	SPHINX Syetem
to Statistical Modeling)	BBN Corporation	SBYBLOS System
(Bou-Ghazale & Hansen, 1998)		
1990s	Microsoft Corporation	Whisper and Dragon Dictate Systems
(Integration into Society	IBM Corporation	Via voice System
Beyond Laboratories)		
(Lee, Hon, & Reddy, 1990)		

Table 1: History of Speech Recognition Development

method for medical records with speech recognition. Verified to be nearly three times more efficient than traditional methods, this approach not only substantially reduces the workload of medical personnel but also significantly enhances the operational efficiency of the entire medical system. Suresh, Sandra, Thajudheen, Hussain, and Amitha (2023) suggested applying speech recognition technology to assistive home systems for the visually impaired, using voice control to replace traditional input methods such as touchscreens or keyboards. This greatly improves the living conditions of special needs groups.

2.2 Current SOTA Techniques

As mentioned before, speech recognition technology originated in the United States at Bell Labs in the 1950s, where their research team pioneered the development of an isolated digit recognition system. In the 1970s, Soviet scientists were the first to propose using dynamic programming to solve the problem of unequal length in speech signals, developing the Dynamic Time Warping (DTW) algorithm (Vintsyuk, 1968) based on this, while simultaneously, the introduction of Linear Predictive Coding

(LPC) effectively resolved the feature extraction issues of speech signals, moving speech recognition from theory to practice.

During the 1980s, statistical model-based methods, represented by the Hidden Markov Model (HMM) approach (Rabiner, 1989), gradually became dominant in speech recognition research. However, as HMM transition probabilities are only related to the previous moment, this limited the use of contextual information and presented flaws in modeling long-term dependency in speech, leading to performance limitations as data volumes increased.

Further advancements in speech recognition benefited from the application of Deep Neural Networks (DNN). In 2006, G. E. Hinton, Osindero, and Teh (2006) used Restricted Boltzmann Machines (RBM) to initialize the nodes of neural networks, giving rise to Deep Belief Networks (DBN). DBNs employ an unsupervised greedy layer-by-layer approach (G. E. Hinton, 2002) that retains as much feature information of the modeling subject as possible while continually fitting to obtain weights. Due to their structure, which includes multiple layers of non-linear transformations, and because they do not require assumptions about the distribution of speech data, Yu and Deng (2011) introduced deep learning into acoustic modeling. They used more network layers to extract deeper features of speech and obtained longer structural information through frame splicing, which significantly enhanced the input length of recognizable speech, diversified the input features, and greatly improved text recognition accuracy by using DNNs to model the relationship between acoustic feature vectors and states. Additionally, Abdel-Hamid, Mohamed, Jiang, and Penn (2012) introduced Convolutional Neural Networks (CNNs) into DNN-HMMs, utilizing local convolution, weight sharing, and pooling to extract more complex and robust features from lower-level features to increase model stability.

For recognizing longer periods of speech information, Recurrent Neural Networks (RNNs) gradually became a focus of research. This model differs from other neural networks as each layer not only outputs to the next layer but also outputs a hidden state that participates in the next decision. However, RNN acoustic model training often uses Stochastic Gradient Descent (SGD), which can lead to issues like gradient vanishing (Bengio, Simard, & Frasconi, 1994), potentially causing the network to diverge or become rigid. To address this, Erdogan, Hershey, Watanabe, and Le Roux (2015) improved RNNs into Long Short-Term Memory networks (LSTM), utilizing input, output, and forget gates to control the flow of information, allowing gradients to propagate stably over relatively longer durations. LSTM networks typically consist of 3-5 LSTM layers. G. Hinton et al. (2012) introduced LSTM structural units into the hidden layers of DNNs, gaining the ability to remember longer sequences. The LSTM-DNN model performed excellently in noisy environments, subsequently leading to the development of the CNN-LSTM-DNN (CLDNN) architecture (Sainath, Vinyals, Senior, & Sak, 2015).

In the field of speech recognition, the emergence of wav2vec technology marked a significant turning point. In 2019, Facebook AI Research introduced the wav2vec model, which employs a self-supervised learning approach to directly learn useful feature representations from raw audio, without relying on traditional acoustic features such as Mel-frequency cepstral coefficients (MFCC). Wav2vec extracts deep features of speech effectively by pre-training on a large amount of unlabelled audio data and using contextual information to predict hidden units within audio segments. This method has demonstrated significant performance improvements in speech recognition tasks, especially in processing low-resource languages.

In 2020, Facebook AI enhanced this approach with the release of wav2vec 2.0, introduced by Baevski, Zhou, Mohamed, and Auli (2020), represents a transformative advancement in the field of self-supervised learning for speech recognition. Building upon the groundwork of earlier models such

as BERT and its predecessor, wav2vec, this framework enhances speech recognition capabilities significantly by utilizing a novel approach of learning from raw audio data before fine-tuning on transcribed speech. The model not only simplifies the learning process but also outperforms traditional semi-supervised methods with its efficient use of both labeled and unlabeled data.

The architecture of wav2vec 2.0 is notably robust, featuring a multi-layer convolutional neural network that processes raw audio into latent representations. These representations are then masked and refined by a Transformer network, which learns to identify the correct representations from a set of distractors through a contrastive task. This mechanism is inspired by the success of masked language modeling in NLP as seen Devlin, Chang, Lee, and Toutanova (2019).

Wav2vec 2.0 has demonstrated exemplary performance across several benchmarks. On the LibriSpeech dataset, it achieved word error rates (WER) as low as 1.8/3.3 on clean/other test sets, significantly improving over previous methods. An exceptional achievement of wav2vec 2.0 is that its large pre-trained version reaches a WER as low as 4.8% on the clean test set of LibriSpeech with only 10 minutes of labelled training data, suggesting its potential to facilitate speech recognition technologies in languages and dialects with limited available data. This capability of training on minimal data aligns with the principles of human language acquisition, where exposure rather than explicit instruction drives learning, paralleling the findings in the broader field of machine learning and language processing, which allows the model to be pre-trained on a large dataset of multiple languages and fine-tuned on a small dataset of a low-resource language. The model can also be fine-tuned to work with different voices or in noisy environments.

The choice of wav2vec 2.0 for fine-tuning in this study is rooted in its robust architecture and its proven track record in handling complex audio processing tasks. Wav2vec 2.0 is designed to learn useful representations from raw audio data through a self-supervised learning mechanism before fine-tuning on transcribed speech. This model's ability to efficiently utilize both labeled and unlabeled data makes it especially potent for scenarios with limited annotated resources. Additionally, wav2vec 2.0's adaptability to varied acoustic environments and its capacity to improve through exposure to specific noise conditions align perfectly with the objectives of this research. By training on datasets augmented with vehicular and public noises, the model leverages its inherent strengths to enhance its noise robustness and recognition accuracy, making it an ideal choice for advancing ASR performance in real-world, noisy settings. The use of this model allows for a nuanced understanding and tackling of the challenges posed by different noise types, demonstrating its versatility and effectiveness in enhancing speech recognition technologies.

2.3 Noise Robustness in ASR

ASR systems are critical in various applications, from voice-activated assistants to automated transcription services. Their ability to convert spoken language into text accurately is important for effective communication between humans and machines. The performance of these systems, however, can significantly vary based on the quality and clarity of the input audio.

In ideal conditions, clean recordings, where the speech is unobscured by background noise, allow ASR systems to achieve their highest accuracy. The algorithms are optimized to detect and process clear speech signals effectively. However, real-world scenarios seldom provide such optimal conditions. Recordings often contain background noises from multiple sources, including street sounds, conversations, and mechanical noises. These noises can overlap with speech frequencies, masking the speech signals and significantly increasing error rates in speech recognition.

Research into ASR technology reveals that these systems' performance under different noise conditions can vary dramatically. For instance, vehicular environments present a particularly challenging noise scenario due to the complex composition of sounds—ranging from engine noises to road vibrations—that are spread out over a wide frequency range and often overlap with the frequencies of human speech. Traditional speech denoising methods described by Van Segbroeck and Narayanan (2013), which typically exploit the contrast in periodicity between speech and other sounds, are less effective in such settings.

Recent studies, such as those cited by researchers like Ochiai et al. (2024), suggest that addressing noisy speech recognition directly might be more effective than attempting to clean or enhance the audio before processing. This approach is particularly beneficial when dealing with non-stationary noises that fluctuate in volume and type.

To improve ASR performance in noisy conditions, recent advancements have focused on training, pre-training on noisy speech (Likhomanenko et al., 2020), fine-tuning on noisy speech (Schlotterbeck et al., 2022; Zhu et al., 2022; Prasad, Jyothi, & Velmurugan, 2021), and fine-tuning on noise-clean paired data (Maas et al., 2012). and applying attention (Higuchi et al., 2021). This involves training or adapting existing ASR models using datasets that encapsulate a variety of noise scenarios. By exposing the models to noisy data during training, they develop an enhanced ability to discern and interpret speech with background noise. This not only boosts the robustness of the systems but also their accuracy in less-than-ideal acoustic environments.

One effective strategy for fine-tuning involves using augmented noisy speech data, which artificially introduces various types of noise to clean speech during the training process. This method helps the model learn to recognize and differentiate speech from noise. Another approach is employing real-world noisy datasets, which provide a more realistic training environment for the models. For example, projects like the GitHub repository [1] on ASR for clean and noisy speech data illustrate how models such as Wav2Vec 2.0 can be fine-tuned with custom pre-processed noisy speech data to improve text recovery accuracy from noisy audio streams.

Fine-tuning on noisy speech not only improves the robustness of ASR systems but also prepares them to handle a broader range of acoustic scenarios. This is crucial for applications in diverse environments, from busy urban settings to bustling commercial areas, where noise levels can significantly impact the effectiveness of voice-activated systems.

[1] https://github.com/romtrost/ASR-for-clean-and-noisy-speech-data

Chapter 3

Methodology

I take fine-tuning as the method employed to address the research question. In order to get the result that whether a model fine-tuned on a specific noise environment offers advantages over a general noise-robust model, I developed two datasets from distinct noise environments. Using these datasets, I fine-tuned the base model into two different models, employing a methodology according to the same method used by Schlotterbeck et al. (2022). This chapter provides a detailed explanation of how the noise datasets were constructed and offers a deeper discussion of the experimental setup.

3.1 Data

One of the objectives of this paper is to simplify the process of optimizing models for specific acoustic settings. This involves minimizing the necessary costs and resources without compromising performance, allowing the process to be easily applied across multiple locations.

With the advancement of technology, the widespread use of in-vehicle voice interaction devices has replaced the manual control of primitive electronic vehicle devices, significantly enhancing driver concentration and ensuring driving safety. However, the interference from ambient noise in the vehicle environment during driving leads to a low accuracy rate in recognizing voice interaction commands, severely impacting the user experience of voice interactions. Therefore, suppressing noise interference and isolating the target speaker's voice from complex driving environments has become a focal point of research.

Due to time constraints, it is essential to remember the goal of minimizing the workload of the process; thus, the process of recording data should also be as simplified as possible. Directly recording training voices in noisy environments might yield the best model performance, but this process is complex and time-consuming. A simpler approach is to directly obtain the noise environment and then mix it with an existing available voice dataset, which is the method I employed in this project.

The first dataset was created using in-vehicle noise from the NoiseX-92 noise database, specifically Volvo car noise, and further enhanced with military vehicle noise (Leopard) to strengthen and refine the dataset. This compilation is referred to as the "Vehicular Noise Speech" dataset. And the second dataset was constructed using a collection of various noises collected from FreeSound [2], including sounds from airports, cafes, hospitals, metros, and vacuum noise. This compilation is referred to as the "Public Other Noise Speech" dataset.

3.1.1 LibriSpeech

In this thesis, to ensure an optimal comparison with earlier literature (Zhu et al., 2022),the well-known LibriSpeech corpus (Panayotov, Chen, Povey, & Khudanpur, 2015) was utilized. This speech corpus comprises audio and transcriptions sourced from English audiobooks. It is segmented into training,

[2] https://freesound.org/

testing, and development sets, with recordings categorized as “clean” or “other”, each covering different data subsets aimed at providing diverse data to support the training and evaluation of ASR systems.

The selection criterion is based on calculating the Word Error Rate (WER) of LibriSpeech recordings using the acoustic model from the Wall Street Journal (WSJ). Recordings from speakers with lower error rates are categorized as “clean”. Such recordings are typically of higher quality and feature accents closer to Standard American English.

The “clean” configuration is suitable for those focused on developing speech recognition systems in clear speech environments. The “other” configuration includes more challenging recordings (with higher WER), often containing more background noise or accent variations, making it suitable for applications requiring the model to perform well in complex or noisy environments.

The model employed in this study was pre-trained on the training set, which includes train-clean-100, train-clean-360, and train-other-500, these three subsets collectively provide the model with 960 hours of speech data, encompassing a variety of voice qualities. This diversity enables the model to better adapt to different acoustic environments. For fine-tuning, the development set (dev-clean) was used, and the test set (test-clean) was utilized for evaluation.

3.1.2 Vehicular Noise

In the process of driving, due to the complexity of the environment and noise sources, vehicle noise signals can be divided into interior noise, exterior noise, and vehicle body noise.

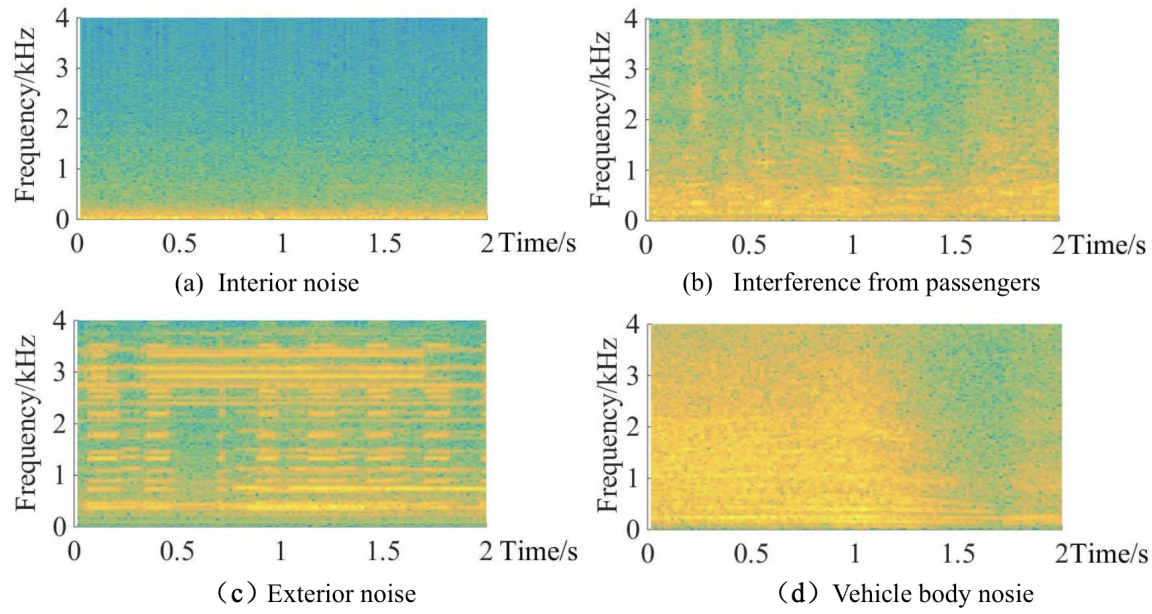

Figure 2: Feature analysis diagram of vehicle noise signal

1. Interior noise primarily includes noise generated by the internal systems of the vehicle, such as the air conditioning system, which, when operating, can interfere with the driver’s command recognition. This noise signal predominantly consists of low-frequency information. Additionally, conversations among passengers can overlap with the driver’s voice commands used to control the vehicle’s systems.

2. Exterior noise is influenced by the surrounding environment during the vehicle's operation. For example, wind noise increases with speed, and noise from the interaction of tires with the ground or from gravel on the road constitutes road noise. Other environmental noises, such as the horns of other vehicles, also contribute to the complexity of exterior noise signals, which change according to the different environments the vehicle traverses.

3. Vehicle body noise mainly stems from the vehicle's hardware, including noise from the engine when it is operational and vibrations of the car body while driving. Noise from the vehicle's exhaust system and noise generated by airflows and acoustic excitation are characteristic of this category, primarily consisting of low-frequency sounds.

The spectrograms of various types of vehicle noise signals are shown in Figure 2. Image (a) displays the interior noise, from which it can be seen that the noise is mainly concentrated in the low-frequency range. The interference from passengers inside the vehicle is shown in Image (b). Image (c) represents the exterior noise, which primarily includes traffic noise during driving. Image (d) shows the vehicle body noise, mainly consisting of engine noise.

According to these vehicular noise characteristics, the noise data I chose to use in the study are from the NoiseX-92 noise library, specifically vehicle noises from a Volvo and a Leopard. The Volvo noise data was recorded inside a Volvo vehicle while driving on an asphalt road at 120 km/h in 4th gear during rainy conditions. The final recording obtained is a 235-second sample with a sampling rate of 19.98 KHz, stored in 16-bit format.

Similarly, the Leopard noise was recorded inside the Leopard 1 military vehicle. During the recording, the vehicle was moving at a speed of 70 km/h. The environment of the recording captured a sound level of 114 dBA. Like the Volvo noise, the final recording is a 235-second sample with a sampling rate of 19.98 KHz, stored in 16-bit format.

Incorporating noise from the Leopard military vehicle into training is essential because military settings are often noisy, and challenging voice recognition systems. Training with this noise improves the system's ability to distinguish speech in noisy environments and enhances overall robustness, enabling better performance in various challenging acoustic settings.

3.1.3 Public Other Noise

The second dataset builds on the approach used by Prasad et al. (2021), adhering to the methods described by Zhu et al. (2022) to ensure methodological consistency. This method of training or fine-tuning the model using datasets derived from different noisy backgrounds is referred to as 'multi-condition training' (Du et al., 2014). This training approach is designed to enhance the robustness of voice recognition systems by exposing them to a wide range of acoustic disturbances.

Prasad et al. (2021) compiled a diverse array of noise recordings from FreeSound.org [3], which included ambient sounds from public places like traffic and restaurants, as well as continuous machine hums and overlapping conversations known as babble noise. Similarly, I developed a comprehensive noise dataset from FreeSound, which encompasses a variety of realistic environments such as airports, cafes, hospitals, subway stations, and the distinct sound of vacuum cleaners [4].

[3] https://freesound.org/

[4] The noise used in this thesis can be downloaded directly from https://github.com/DongwenZhu/Noise_Robust-ASR/tree/main/PublicOtherNoise

3.1.4 Dataset Processing

For this study, the training utilized the LibriSpeech dev-clean subset and the evaluation employed the test-clean subset (Panayotov et al., 2015). The LibriSpeech dataset features mono channel recordings at a 16 KHz sample rate. Conversely, the vehicular noise recordings present a higher sample rate of 19.98 KHz, and the public other noise samples vary in format, with some at a 44.1kHz sample rate and possibly in stereo channels. As such, all noise recordings required resampling to 16 KHz and conversion to mono channel to align with the speech data [5].

Then each speech file from LibriSpeech was combined with a randomly selected noise file. A script [6] was developed to facilitate this process, incorporating various Signal-to-Noise Ratios (SNRs) ranging from -20 dB to +20 dB to replicate diverse auditory environments. This approach introduces considerable variability, closely mimicking the fluctuating background noise levels encountered in real-world settings, thereby enhancing the robustness of the system across different scenarios.

The procedure initiates by retrieving '.wav' files from specified directories, ensuring dataset diversity. Each clean speech file is paired with a noise file, with the SNR randomly selected from a predefined range. The add_noise' function standardizes the sample rate for consistency, adjusts the noise file's length to sync with the speech file by either repeating or trimming, and modifies the noise volume to achieve the targeted SNR by adjusting its RMS value in relation to the clean speech. This modified noise is then merged with the clean speech to generate the mixed audio. To avoid clipping, the audio's amplitude is reduced if it exceeds the format's representable range. The script also prevents repetitive selection of noise files and SNRs, promoting an even distribution of noise types and levels throughout the dataset. The resultant mixed files are named to reflect the speech and noise file IDs along with the SNR level, aiding in traceability.

These approaches resulted in two mixed datasets: Vehicular Noise Speech and Public Other Noises Speech, used in later experiments. The limitations and implications of these methodologies are elaborated upon in Chapter 5.

3.2 Experimental Settings

3.2.1 Baseline Model

The baseline model I used is "wav2vec2-base-960h" (Baevski et al., 2020) [7]. This model incorporates an advanced architecture and was pre-trained on 960 hours of English audio from the LibriSpeech corpus, which includes a diverse set of speakers and accents. This extensive training provides a robust foundation for the model to learn a wide variety of linguistic features. The "960h" in the model's name indicates that it was not only pre-trained but also fine-tuned on this significant set of labeled audio, optimizing its performance for English speech recognition tasks. The model utilizes 12 transformer layers, each with 768 hidden units and 8 attention heads, enabling it to model complex audio patterns and dependencies effectively.

Wav2vec technology began with the original wav2vec model developed by researchers at Facebook AI. This model marked a significant advancement in speech recognition by allowing systems to learn

[5]The full resampling code is available at `https://github.com/DongwenZhu/Noise_Robust-ASR/blob/main/16kHz.py`

[6]The complete script for mixing speech and noise can be accessed at `https://github.com/DongwenZhu/Noise_Robust-ASR/blob/main/mix_noisespeech_difflen_SNR.py`

[7]The baseline model can be found at `https://huggingface.co/facebook/wav2vec2-base-960h`

valuable representations of audio directly from raw waveforms without relying on annotated data. This dramatically reduced the dependency on costly labeled datasets and enabled more scalable and efficient development of robust speech recognition models.

Building on the strengths of the original model, wav2vec 2.0 introduced an enhanced architecture and learning process. This version features a two-stage training strategy: initial pre-training on large amounts of unlabeled data followed by fine-tuning on a smaller set of labeled data. The architecture of wav2vec 2.0 is fortified with transformer networks, which are highly effective in handling sequential data such as natural language and audio. These improvements allow wav2vec 2.0 to handle more complex patterns and longer dependencies in speech data, significantly enhancing its ability to understand and transcribe spoken language.

The capabilities of the wav2vec2-base-960h model ensure it excels in accurately transcribing English speech, outperforming many other models trained on traditional supervised learning methods. By learning from both labeled and unlabeled data, it achieves a higher degree of accuracy and adaptability, making it an ideal choice for applications requiring reliable speech recognition.

3.2.2 Fine-tuning

The Hugging Face's tutorial "Fine-tune a pretrained model"[8] outlines the steps necessary for fine-tuning a pre-trained model. The initial step is to prepare and uniform the dataset, focusing particularly on the 'audio' column to understand properties such as the sampling rate. In this experiment, it is essential to match the sampling rate of the audio files, both speech and noise files, to the model's expected rate of 16kHz.

Next, employ a feature extractor to convert the audio files into a format suitable for the model, like log-mel spectrograms, while a compatible tokenizer processes the transcription texts associated with these audio files.

Additionally, setting up an efficient data collator is crucial to handle and batch the pre-processed data properly, involving the correct padding of inputs and labels for model processing. A pre-trained model checkpoint is then loaded and configured for training, including the definition of the loss function and optimization parameters.

To organize the data for model training, it's useful to create a CSV file named `metadata.csv` [9]. This file should list each audio file and its transcription in the format shown in Table 2.

File_name	**Transcription**
train/84-121123-0012.wav	THE OLD MAN'S EYES REMAINED FIXED ON THE DOOR
test/1089-134691-0000.wav	HE COULD WAIT NO LONGER

Table 2: Example of `metadata.csv` content

Each line in the CSV file represents a data point, where the "file_name" indicates the path to the audio file and "transcription" provides the text that the audio file contains. This format helps in mapping each

[8]Hugging Face Fine-tune a pretrained model can be found at https://huggingface.co/docs/transformers/v4.27.2/en/training

[9]For the complete "metadata.csv" creation code, see: https://github.com/DongwenZhu/Noise_Robust-ASR/blob/main/create_metadata.py

audio file to its corresponding text, making it easier to train models that need to learn from audio-text pairs.

Training parameters such as learning rate, batch size, and the number of epochs are then defined. Specifically, I set the model to be trained for **59 epochs** with a batch size of **8**, and a **learning rate of 1e-57**. The training process is facilitated using Hugging Face's Trainer API, which manages loops, logging, and checkpoint saving. Checkpoints were saved every 1000 steps to choose the optimal number of steps.

Building upon this foundation, I fine-tuned [10] the baseline model on two distinct noise datasets that were explained earlier, the Vehicular Noise Speech dataset and the public Other Noise Speech dataset, resulting in three different models: the baseline model, the model fine-tuned on Vehicular Noise, and the model fine-tuned on Public Other Noise. Details of the relationship of the three models are provided in Figure 3.

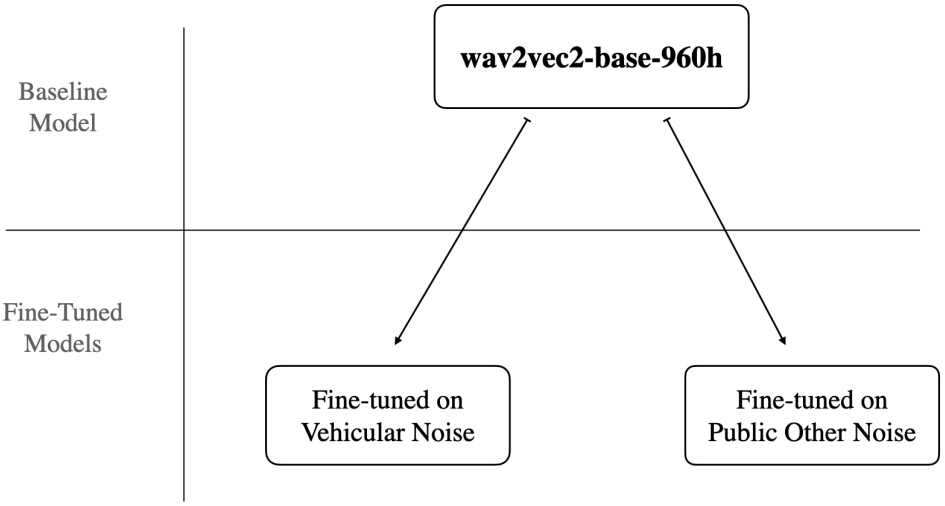

Figure 3: Fine-Tuned

3.2.3 Evaluation

To evaluate the effectiveness of speech recognition systems, I use two primary metrics: Word Error Rate (WER) and Character Error Rate (CER). Both metrics serve to quantify how accurately a model transcribes spoken language into text.

The Word Error Rate (WER) measures the performance by calculating the ratio of the total number of errors (which include substitutions, insertions, and deletions) to the number of words in the reference text. Essentially, it indicates how many words out of every hundred were incorrectly transcribed by the speech recognition system. The Character Error Rate (CER) functions similarly but operates at the character level, assessing the number of letter errors against the total number of characters in the

[10]The complete fine-tuning code can be found at `https://github.com/DongwenZhu/Noise_Robust-ASR/blob/main/finetune.py`

original script. CER provides a more detailed analysis of transcription accuracy because it considers each individual character, making it particularly useful for languages where character accuracy is more indicative of overall performance.

$$WER = \frac{S + I + D}{N}$$

This formula calculates the Word Error Rate by dividing the sum of substitutions (S), insertions (I), and deletions (D) by the total number of words (N) in the reference.

$$CER = \frac{S_c + I_c + D_c}{N_c}$$

This formula calculates the Character Error Rate, similarly to WER, but at the character level. It divides the sum of character substitutions (S_c), insertions (I_c), and deletions (D_c) by the total number of characters (N_c) in the reference.

It is important to recognize that these metrics do not effectively measure how well the meaning is conveyed, since some words carry more significance than others (Wang, Acero, & Chelba, 2003). This scenario could result in a low Word Error Rate (WER) yet produce a transcription that lacks critical information, rendering it unintelligible. While I chose these metrics to align with previous research for consistency, it's crucial to acknowledge that they do not reflect various performance aspects, such as the impact of specific noises on accuracy.

In my thesis, I analyzed three speech recognition models [11]: the baseline model, the model fine-tuned on Vehicular Noise, and the model fine-tuned on Public Other Noise. The purpose of my experiment was to demonstrate whether that the model fine-tuned on Vehicular Noise performs better in environments with vehicular noise, using the model fine-tuned on Public Other Noise as a control group for comparison. These models were tested using both noisy datasets and a clean speech dataset consisting of 120 randomly selected speech files from the LibriSpeech test-clean set. Details of the experimental setup are provided in Figure 4.

Each model was evaluated under three conditions:

1. Targeted Noise: Testing the vehicular model in vehicular noise settings and the public noise model in various public noise settings to assess performance in intended conditions.
2. Non-targeted Noise: Assessing how each model performs in noise conditions it was not specifically tuned for.
3. Clean Condition: Evaluating model performance in the absence of noise to gauge any loss of general ASR capability due to fine-tuning.

The first experimental setting, Targeted Noise, involves testing each model in the noise environment for which it was specifically fine-tuned—this directly addresses Hypothesis 1 (H1), which posits that models fine-tuned for specific noise environments will outperform the general noise-robust model in those same environments. If the fine-tuned models show superior performance in their respective environments under this condition, H1 is supported. The second setting, Non-targeted Noise, assesses

[11] The complete evaluation code can be found at `https://github.com/DongwenZhu/Noise_Robust-ASR/blob/main/eval.py`

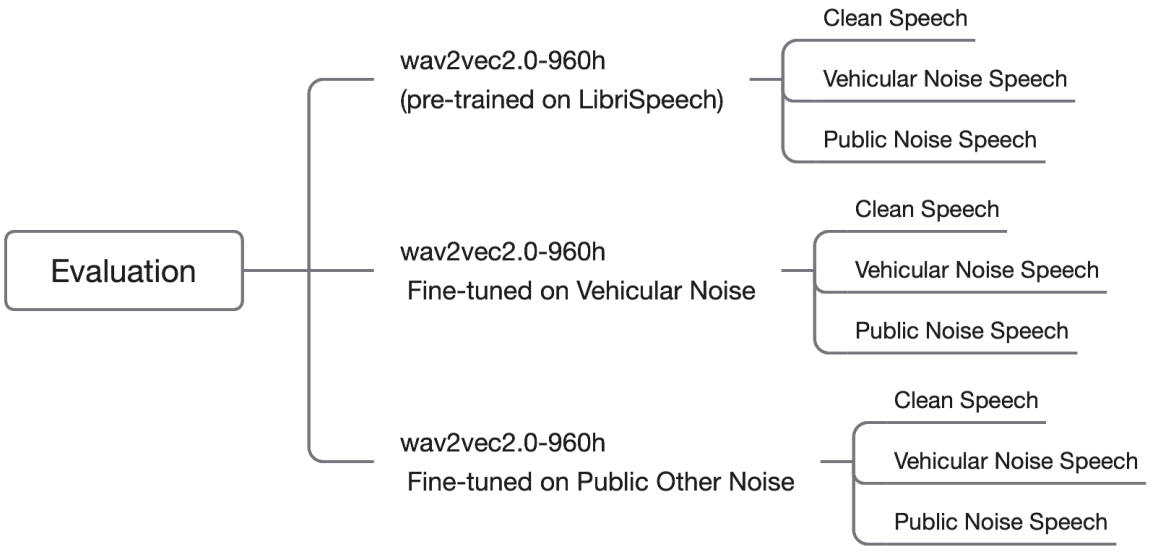

Figure 4: Evaluation

model performance in noise environments for which they were not specifically fine-tuned. This setting is critical for testing H2, which suggests that each fine-tuned model will perform worse in unfamiliar noise environments compared to the general noise-robust model. A drop in performance under this condition would support H2. The third condition, Clean Condition, evaluates how each model performs in an environment without any noise, aimed at testing H3 and H4. H3 suggests that models fine-tuned on specific noises will perform better than a general model when evaluated on data from both environments' noises—indicative of their ability to handle diverse noise settings effectively if supported. Conversely, H4 posits that these fine-tuned models will underperform in clean, noise-free conditions, highlighting a potential overfitting to noisy data—if the general model performs better in clean conditions, then H4 is supported, showing a loss of general ASR capabilities in the fine-tuned models.

Chapter 4

Results

This chapter presents the outcomes of the experiments conducted to assess the performance of the ASR models fine-tuned for specific noise environments. Following a rigorous methodology outlined in the previous sections, the models—including the baseline, the model fine-tuned on Vehicular noise, and the model fine-tuned on Public Other noise—were evaluated across two specific noise datasets and one clean dataset. The results, encapsulated in comprehensive statistical analyses and visual representations, aim to illuminate the effectiveness of these models under varying acoustic conditions.

4.1 Model Performance

After fine-tuning the models, I utilized the evaluation code to conduct assessments on all three datasets. This process was executed according to the steps outlined in the evaluation section's diagrams. Each of the three models—the baseline model, the model fine-tuned on Vehicular Noise, and the model fine-tuned on Public Other Noise—was evaluated using two noise-specific speech datasets and one clean dataset. The results of these evaluations are presented in the Table 3, and the Word Error Rate (WER) for each model is illustrated in a line chart presented in Figure 5. This chapter will also explore potential explanations for the findings based on the observed performance differences among the datasets.

Models	**VehicularNoise Speech**		**PublicOtherNoise Speech**		**LibriSpeech**	
	WER	**CER**	**WER**	**CER**	**WER**	**CER**
Vehicular Model	**11.26%**	**5.45%**	59.08%	42.59%	4.09%	1.13%
Public Other Model	16.05%	8.81%	**43.49%**	**30.68%**	4.04%	1.13%
Baseline Model	19.77%	11.96%	56.71%	45.83%	**3.39%**	**0.96%**

Table 3: Model Performance under Different Noise Conditions

4.2 Answering Research Question

In this section, I address the central research questions posed in Chapter 1 concerning the performance of an ASR model fine-tuned with noise samples specific to a vehicuar environment versus a general noise-robust ASR model across different noise conditions.

4.2.1 Research Question Analysis

Q1. How does the performance of an ASR model fine-tuned for a specific noise environment compare to a general noise-robust ASR model when evaluated on data from the same noise environment?

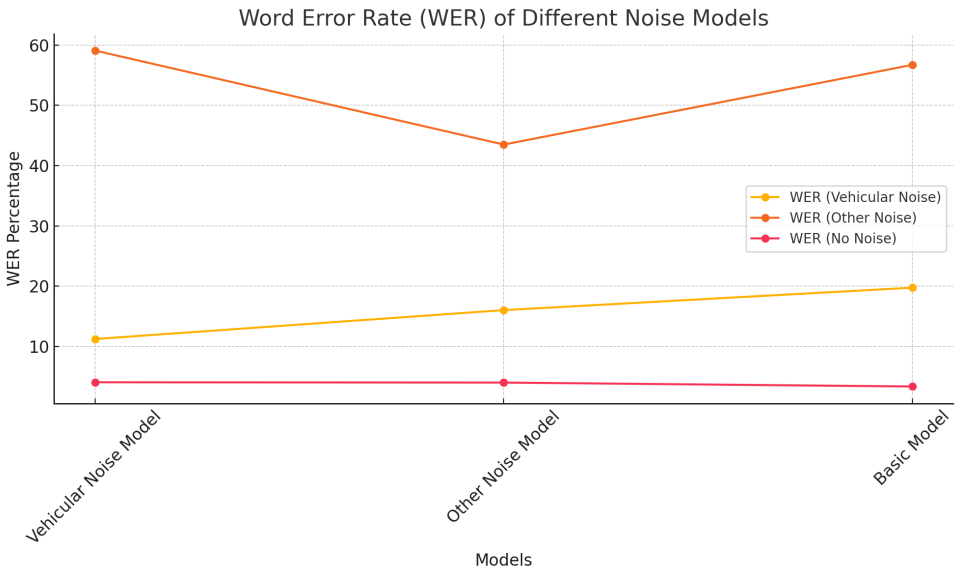

Figure 5: WER across Different Noise Models

Q2. How does the performance of an ASR model fine-tuned for a specific noise environment compare to a general noise-robust ASR model when evaluated on data from different noise environments?

Q3. How do two ASR models, each fine-tuned for specific but different noise environments, compare to a general noise-robust ASR model when evaluated on data that includes both environments?

Q4. How do two ASR models, each fine-tuned for specific noise environments, perform compared to a general noise-robust ASR model when evaluated on a clean (no noise) test dataset?

To answer these questions, I conducted experiments comparing two fine-tuned models—one targeting vehicular noise and the other optimized for a variety of public noises—with a baseline general noise-robust model. The findings are summarized below:

- Vehicular Noise Environment: The model fine-tuned for vehicular noise demonstrated superior performance in its targeted environment, with a Word Error Rate (WER) of 11.26% and a Character Error Rate (CER) of 5.45%, substantially outperforming the general model and the public noise model in these conditions.

- Public Other Noise Environment: In contrast, the model tailored for public noises performed best in its specific conditions, achieving the lowest WER and CER among the models in such environments. It shows that specificity in training correlates with enhanced performance in corresponding environments.

- General Noise-Robust Model Performance: The baseline model, designed to handle a broad range of noises, showed the most consistent performance across diverse environments but did not excel in any specific noise condition as the specialized models did.

4.2.2 Hypotheses Validation

The evidence gathered from the different noise environments and the performance of each model type is critically examined below:

H1. The ASR model fine-tuned by a specific noise environment will perform better than the general noise-robust (no noise-robust) ASR model when evaluated on data from that particular noise environment.

It was hypothesized that the ASR model fine-tuned on a specific noise environment (vehicular noise) would perform better than a general noise-robust model when evaluated in that particular environment, which is the main research objective of this thesis. The results strongly support this hypothesis, as the vehicular model achieved a WER of 11.26% and a CER of 5.45% in vehicular noise conditions, which is significantly lower compared to the baseline model's WER of 19.77% and CER of 11.96%. This marked improvement underscores the effectiveness of targeted noise training in enhancing model accuracy in specific settings.

H2. The ASR model fine-tuned by a specific noise environment will perform worse than the general noise-robust (no noise-robust) ASR model when evaluated on data from other noise environments.

The testing corroborates this, as seen when the vehicular model was subjected to public noise environments, yielding a WER of 59.08% and a CER of 42.59%, substantially higher than the baseline model's performance (WER of 56.71% and CER of 45.83%) in the same conditions. This outcome illustrates the limitations of specialized training when applied to non-targeted noise environments.

H3. The two ASR models fine-tuned by a specific noise environment will perform better than the general noise-robust (no noise-robust) ASR model when evaluated on data from both environment noises.

The expectation was that the two ASR models fine-tuned on specific environments would outperform the general noise-robust model when evaluated across both environment noises. The data partially support this hypothesis. In their respective targeted environments, both models significantly outperformed the baseline model. However, in non-targeted noise settings, their performance declined, indicating that while specialization improves performance in familiar settings, it does not universally enhance performance across diverse noise conditions.

H4. The two ASR models fine-tuned by a specific noise environment will perform worse than the general noise-robust (no noise-robust) ASR model when evaluated on data from a clean (no noise) test dataset.

The results confirm this hypothesis, as the baseline model exhibited the lowest WER (3.39%) and CER (0.96%) in clean conditions, compared to the vehicular model (WER 4.09%, CER 1.13%) and the public other noise model (WER 4.04%, CER 1.13%). This suggests that general models, while not excelling in noisy environments, maintain superior performance in clean settings where the complexities of specific noise types are absent.

The validation of these hypotheses highlights the intricate balance between model specialization and generalization. While specialized models excel in their respective noise environments, their performance can degrade in unfamiliar settings, emphasizing the need for adaptive or hybrid models

that can leverage both specific and general training methodologies to achieve optimal performance across varied acoustic scenarios.

It's important to address that the model fine-tuned on the Public Other Noise dataset demonstrated less robustness to noise overall. It obtained the lowest result 43.49%, which is already 4 times than the other two models. One significant reason for this could be the methodology where noises were recycled across different speech files at varying SNRs, rather than utilizing distinct noises for each file. This approach aligns with the procedures used in previous studies, such as those by Zhu et al. (2022), to ensure consistency with existing research. Although data augmentation techniques, such as reusing noise samples at different SNRs, have been shown to enhance model performance effectively (Sivasankaran, Vincent, & Illina, 2017), employing a more varied set of original noises could potentially lead to a broader and more generalizable noise representation within the model. This decision raises questions about the quality and comparability of the training data sets and models, suggesting that a more diverse noise dataset might improve general noise robustness.

4.3 Checking for Overfitting

Overfitting occurs when a model is trained too closely to a specific dataset, capturing noise and outliers as if they were representative of the general trend. This usually results in high performance on training data but poor performance on new, unseen data. To ensure that our models are both robust and generalized well beyond the training data, it is crucial to monitor for signs of overfitting during the training process.

As mentioned in Chapter 3, the models were fine-tuned for 10,000 steps, 59 epochs, and checkpoints were saved every 1,000 steps, allowing for a periodic evaluation of the models' performance across these intervals. Performance metrics WER and CER were regularly checked at each checkpoint. This periodic evaluation helps to track whether the performance improvements plateau, continue to improve, or start to worsen, which can indicate overfitting.

The Figure 6 is the performance of the Wav2vec 2.0 model fine-tuned on the vehicular noise dataset, and the model fine-tuned on the public other noise dataset to check whether overfitting is occurring. For the Vehicular Noise Speech model, there is a sharp decrease in WER between 2,000 to 4,000 training steps, after which the WER begins to plateau, albeit with a minor but steady decrease up to 10,000 steps. This trend suggests that while the model quickly learns to adapt to vehicular noise, the learning rate stabilizes without a subsequent increase in error, indicating no significant overfitting. The Public Other Noise Speech model shows a gradual decrease in WER until around 6,000 steps, after which the rate of decrease slows, and the WER slightly fluctuates but generally stabilizes, demonstrating that it's not overfitting.

4.4 Statistic Analyses

In the analysis of the effects of model type and noise conditions on the WER, an Analysis of Variance (ANOVA) was conducted. The results, as summarized in Table 4, highlight significant findings. The ANOVA results convincingly demonstrated that the type of noise condition profoundly influences WER ($F(2,4) = 60.079$, $p = 0.001$), emphasizing the critical role of specific noise environments in enhancing model performance. This finding strongly supports H1, which posited that ASR models fine-tuned for specific noise environments would outperform the general noise-robust model in those same environments, and underscores the importance of targeted model tuning for improving ASR

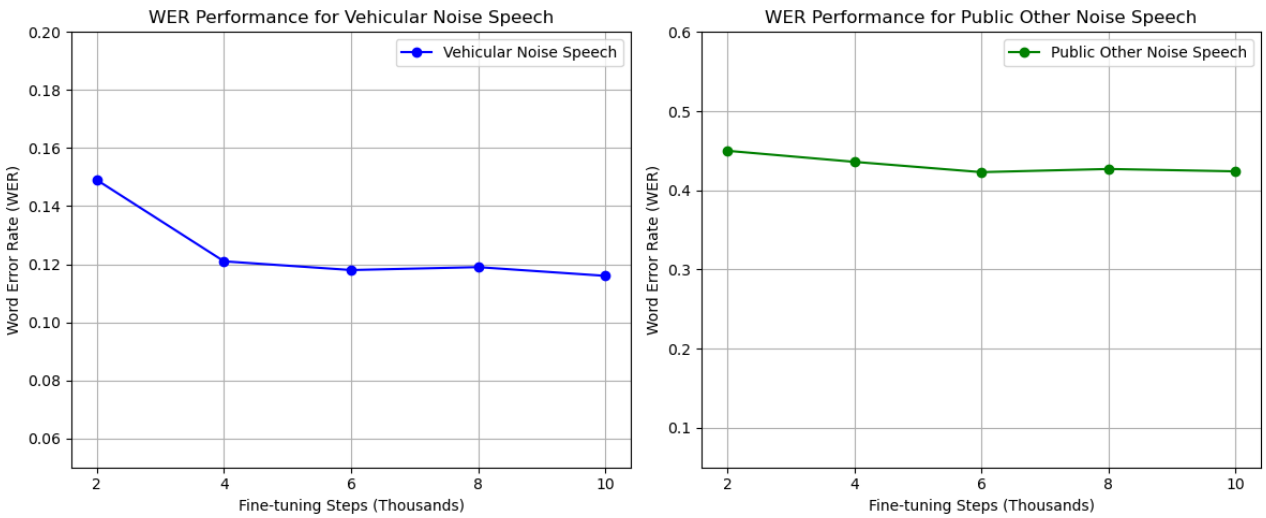

Figure 6: Performance throughout training

accuracy in expected noisy conditions.

Contrastingly, variations in model types did not show a statistically significant effect on WER ($F(2,4) = 0.695$, $p = 0.551$). While this finding indicates a lack of significant differences in overall model types, it partially supports H2 by suggesting that no single model type uniformly outperforms others across various noise settings, highlighting the challenge in generalizing performance across non-targeted noise environments. Moreover, this outcome suggests that while fine-tuned models do not always excel in noise conditions they were not specifically optimized for, there is a positive trend toward refining ASR models with even minor advancements in model design. This insight aligns with the view in H3, which expected fine-tuned models to perform better in mixed noise environments but acknowledges that the performance boost is not universal.

This positive trajectory underlines the importance of continued efforts in environmental adaptation strategies and the exploration of new model architectures. Emphasizing these aspects could lead to substantial advancements in ASR technology, reinforcing the need for specialized models that are adept at handling diverse and challenging acoustic environments.

Source of Variation	df	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	2	45.835	22.927	0.695	0.551
Noise Condition	2	3965.152	1982.576	60.079	0.001
Residual	4	131.997	33.000	-	-

Table 4: ANOVA Results on WER

The experimental outcomes validate the thesis's premise that an ASR model fine-tuned on specific noise types outperforms a general noise-robust model in those specific conditions. However, the trade-off in specialization is the reduced flexibility and performance across varied acoustic environments not represented in the training data. This insight is crucial for the development of future ASR systems, suggesting a potential pivot towards hybrid models that combine the robustness of general noise training with the precision of environment-specific tuning.

Chapter 5

Discussion

This discussion aims to delve into the complexities and implications of the research findings detailed in this thesis, by analyzing this study's limitations, future directions, and applications, exploring how fine-tuning ASR models for specific noise profiles not only advances the theoretical understanding but also paves the way for practical applications that can significantly improve user interactions within noisy environments. As unpacking the results and the broader implications, this section critically examines both the successes and challenges encountered in striving to enhance ASR technology, thereby guiding future initiatives in this rapidly evolving field.

5.1 Limitations

In this research, I have encountered several significant limitations that could impact the performance and application of our models under real-world conditions. I aim to outline these limitations to better understand the potential challenges and areas for future improvement.

The first significant limitation is related to the datasets used and the realism of speech conditions. The datasets used for fine-tuning and testing these models mostly consist of English speech from audiobooks, which feature well-articulated and clean audio. This type of speech differs significantly from the diverse and spontaneous speech patterns found in real-world settings, including overlapping conversations, slang, and various non-standard dialects. Such a limited scope in dataset diversity might hinder the models' robustness and limit their effectiveness across different linguistic contexts, particularly in real-life applications like voice-activated assistants in public places or devices used by individuals with pronounced accents or dialect variations (Basak et al., 2023). Additionally, the approach to simulating noisy conditions in these models involves synthetically mixing clean speech recordings with artificial noise. While this method is standard in research for controlling variables, it fails to accurately mimic the complex and dynamic interplay of overlapping conversations and fluctuating noise levels typical in natural environments. This could potentially lead to an overestimation of the model's performance when deployed in real-world conditions, where noise and speech interactions are far less predictable (Szymański et al., 2020).

The second limitation is the conflict between Model Specificity and Noise Complexity. The research demonstrates the effectiveness of models fine-tuned for specific noise environments, such as vehicular or public noise; however, this approach encounters significant limitations when applied outside these targeted settings. For instance, a model optimized for vehicular noise shows exceptional performance within that specific environment but performs poorly in different acoustic settings, such as public noise in cafes or subway stations. This indicates a critical lack of adaptability, suggesting that these models may not be versatile enough for real-world scenarios where noise conditions are variable and unpredictable. Moreover, the models are typically tested against specific types of noise, such as clear vehicle noise. However, real-life settings often feature more complex and variable noise environments that combine multiple noise sources, including music, people talking, and traffic. These unpredictable noise conditions can severely test the models' effectiveness, leading to potential

failures in more complex acoustic environments that are common in normal life. Consequently, developing multiple specialized models to cater to different noise conditions poses a significant practical challenge, particularly in systems with limited memory or processing capabilities. The reliance on multiple specialized models exacerbates the issue, as each model's performance declines sharply when confronted with non-targeted noise conditions. This decline not only limits their generalizability but also reduces their utility across broader applications, underlining a fundamental conflict between model specificity and noise complexity in real-world applications.

The third limitation combines concerns about the generalizability of the research results due to the methodology used in dataset selection and testing and the evaluation metrics. Firstly, the generalizability of the study's results is further compromised by the choice of using splits from the same speech corpus for both training and testing the models. This approach, as discussed by Szymański et al. (2020), while common for maintaining consistency with prior research, limits the robustness of the findings across different linguistic environments. A more effective approach could have involved using a test set from a different corpus, such as Mozilla's Common Voice [12], or comparing model performance across various corpora. However, constraints such as the limited time available for completing this master's thesis prevented the inclusion of more complex evaluations, like testing the models against different speech datasets mixed with various noise datasets, which could have provided a more comprehensive understanding of the model's effectiveness in diverse real-world situations.

Moreover, the reliance on standard evaluation metrics like Word Error Rate (WER) and Character Error Rate (CER) potentially overlooks crucial aspects of practical usability in ASR systems. These metrics, while useful for benchmarking basic transcription accuracy, fail to measure the semantic accuracy or contextual appropriateness of the transcriptions. This oversight can be particularly critical in applications where the precise understanding of context or intent is necessary, such as in voice-activated vehicle controls, where misinterpreted commands due to semantic inaccuracies could lead to significant consequences. For instance, in the provided examples showed in the Table 5, discrepancies between the transcription and the recognized text demonstrate the limitations of relying solely on WER and CER. In the first row, the original transcription reads "HE COULD WAIT NO LONGER", but the recognized text states "HE WOULD AWAY FOR THE COBE". Although the WER or CER might suggest a low error rate due to the similar number of words and characters, the actual semantic content differs drastically. This points to a significant flaw: these metrics do not account for whether the ASR system preserves the meaning or intent of the spoken words. Furthermore, in the second example where "THE UNIVERSITY" is recognized as "IT HUM ADVERSITY", the issue becomes more apparent. In contexts such as navigation function, such inaccuracies are not just transcription errors but could lead to misunderstandings or misdirections that could have serious implications.

File_name	**Transcription**	**Recognition_text**
test/1089-134691-0000.wav	HE COULD WAIT NO LONGER	HE WOULD AWAY FOR THE COBE
test/1089-134691-0003.wav	THE UNIVERSITY	IT HUM ADVERSITY

Table 5: Example of Transcription Discrepancies

In conclusion, while my study provides valuable insights into the design and tuning of speech recognition models under controlled conditions, it also highlights critical areas where further research and development are needed. Addressing these limitations could lead to more robust and adaptable ASR

[12] https://commonvoice.mozilla.org/

systems that better serve the diverse needs and challenges of real-world applications. This will require not only broader and more varied datasets but also advancements in evaluation metrics and testing methodologies.

5.2 Future Research

The journey through the research presented in this thesis has been both enlightening and challenging. Due to constraints in time and computational resources, certain aspects of this study could not be explored as deeply as desired. These limitations, while frustrating, have opened the door to numerous potential areas for further exploration and improvement. The field of ASR is vast and constantly evolving, and each piece of research contributes incrementally to our collective knowledge. In this section, I propose several directions for future research that could address the limitations identified in this thesis. It is my hope that these suggestions will inspire and guide future efforts to enhance the robustness and applicability of ASR systems, making them more effective in the complex and unpredictable environments of the real world.

5.2.1 Development of datasets

The current datasets, derived predominantly from audiobooks, provide high clarity and articulation but do not capture the complexities of natural speech environments encountered daily. To address this gap, future research could explore the inclusion of more diverse conversational datasets. For example, recordings from street interviews can capture varied speech dynamics and background noises typical in urban settings, while capturing conversations in cafes or transportation hubs could offer insights into speech patterns amidst mild to moderate ambient sounds (Z. Zhang et al., 2018). Discussions on social media platforms could also be invaluable, providing examples of colloquial and abbreviated language usage which are common in informal digital communications.

Moreover, the practice of training fine-tuned models on datasets that artificially mix clean speech with pure noise samples is inadequate for simulating the nuanced environments where noise and speech coexist. Future studies should consider utilizing direct recordings from noisy environments—such as busy marketplaces or public transit—where noise is not just a backdrop but a part of the interactive context.

Additionally, the variation in speech due to linguistic diversity is profound and impacts ASR performance significantly. Speech varies extensively across different languages and even within the dialects of a single language. Future research should focus on creating region-specific ASR models. For instance, models fine-tuned on the tonal variations and speech cadences unique to English could be vastly different from those tailored to Korean's rhythmic and melodic qualities (Yoo et al., 2024). Developing such localized models could drastically improve ASR applicability and user satisfaction by catering specifically to the linguistic nuances of each region.

5.2.2 Improvement of Evaluation Metrics

In the evolution of ASR technologies, enhancing the methods by which these systems are evaluated is crucial for ensuring their effectiveness and reliability. Currently, the predominant metrics used, Word Error Rate (WER) and Character Error Rate (CER), focus largely on transcription accuracy without assessing semantic correctness or contextual appropriateness. This limitation can be particularly problematic in applications requiring precise command interpretation, such as voice-activated vehicle

controls or smart home devices. For instance, accurately differentiating commands like “turn on the lights” from “turn on the right light” is essential for user safety and satisfaction. Thus, future research should aim to develop evaluation metrics that not only assess transcription accuracy but also examine the semantic coherence and context-awareness of ASR outputs (Kim et al., 2021). Implementing such metrics will improve the functional utility of ASR systems, ensuring they perform intended actions correctly in complex real-world environments.

Furthermore, the evaluation results may not be very accurate as they often rely on very similar types of data. In this thesis, both the training and testing datasets for these three models are derived from the LibriSpeech dataset, and mixed with their respective noise datasets in a consistent manner. To enhance the reliability of these evaluations, it’s crucial to utilize a broad range of speech datasets that encompass multiple languages and accents. Utilizing a diverse dataset like Mozilla’s Common Voice, which was previously mentioned in the limitations section, would enable researchers to assess how ASR systems perform across various linguistic scenarios.

By expanding evaluation metrics to include semantic and contextual accuracy and embracing diverse linguistic datasets for system testing, future research can significantly enhance the reliability and applicability of ASR systems.

5.2.3 Future Practical Applications

The success of fine-tuning ASR models for specific vehicular noise environments suggests several promising directions for enhancing in-car communication systems. Future applications could focus on integrating these specialized models into vehicle infotainment and navigation systems to improve reliability and user experience. For instance, voice-activated controls could be optimized to recognize commands accurately in a variety of driving conditions, from quiet highways to noisy urban settings. This would not only enhance safety by reducing driver distraction but also improve the functionality of voice commands under diverse acoustic challenges.

Moreover, the research highlights the potential for deploying these models in new areas of automotive technology. Autonomous vehicles, for example, could benefit greatly from improved speech recognition capabilities, enabling more intuitive interaction between the vehicle and its passengers. This could be crucial as autonomous technologies rely heavily on seamless human-machine communication for operation adjustments and emergency interactions.

Additionally, considering the broader implications of this research, developers might explore the application of fine-tuned ASR models in other types of vehicles such as buses and trains where background noise varies significantly. The integration of robust voice interaction systems in public transport could enhance accessibility, offering a hands-free control system for information retrieval and transaction processes, thus improving the passenger experience.

The next phase of research should also investigate the integration of multi-modal feedback systems that combine voice with visual or tactile responses to enrich the user interface in vehicles. This could mitigate some of the limitations identified in voice-only systems, particularly in high-noise scenarios.

These practical applications underscore the importance of continuing to advance ASR technology in specific noise environments, ensuring that the benefits of this research extend beyond theoretical models to real-world implementations in vehicular and other transportation systems .

Chapter 6

Conclusion

This study has conclusively demonstrated that fine-tuning end-to-end automatic speech recognition (ASR) models for specific noise environments substantially enhances their robustness and accuracy.

Vehicular environments, especially private cars, are increasingly equipped with smart systems that rely on voice commands for operation. However, these environments are fraught with various noises—from engine roars and tire friction to external traffic and environmental sounds—posing significant challenges for ASR performance. The adaptation of the "wav2vec2-base-960h" model to distinct noise conditions, particularly vehicular noise and various public noises, has yielded significant improvements. For instance, in vehicular noise environments, the model achieved a reduction in Word Error Rate (WER) from 19.77% to 11.26% and in Character Error Rate (CER) from 11.96% to 5.45%. Similarly, in other noise environments, the WER improved from 56.71% to 43.49%, and the CER from 45.83% to 30.68%.

The experimental results support the initial hypotheses.The ASR model fine-tuned for specific noise conditions significantly excels in its targeted environment, aligning with the substantial improvements observed in both WER and CER. However, this specialization resulted in diminished performance when the models were evaluated in non-targeted, different noise settings, underscoring the limitations of fine-tuning. While the combination of models did not consistently outperform the general noise-robust model across mixed environments, both exhibited slight performance declines in clean, noise-free conditions, indicating a trade-off between specialized effectiveness and general versatility. These insights emphasize the need for balanced approaches in ASR technology, integrating both specific enhancements and broad robustness.

In conclusion, these enhancements confirm the potential of targeted fine-tuning in ASR systems. Notably, these improvements were achieved without compromising the model's performance in no-noise conditions, where both fine-tuned models nearly matched the baseline model's accuracy. This underscores the effectiveness of environment-specific fine-tuning in not only enhancing ASR performance in noisy settings but also in maintaining general capabilities in quieter environments. This thesis advances the field of speech recognition by elucidating how specific fine-tuning strategies can significantly improve the performance of ASR systems in challenging acoustic settings. The findings not only enhance our understanding of ASR system adaptability but also pave the way for future research aimed at optimizing ASR technologies for both general and specific applications, thereby improving reliability and user interaction in smart vehicular systems and public communication devices.

Chapter 7

Ethics

This thesis only utilizes open-source databases, audio samples, and models to ensure transparency, reproducibility, and ethical integrity. The LibriSpeech corpus and NoiseX-92 noise database are public resources that support unrestricted academic usage. Additionally, public other noise samples were sourced from FreeSound, a platform that allows free usage of sounds for even commercial purposes without the need to seek permission from the authors. The use of the open-source wav2vec 2.0 model by Facebook AI further underscores our commitment to accessible and collaborative scientific inquiry.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4277-4280). doi: 10.1109/ICASSP.2012.6288864

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., & et al. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(689). doi: 10.1186/s12909-023-04698-z

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, June). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv e-prints*, arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477

Basak, S., Agrawal, H., Jena, S., Gite, S., Bachute, M., Pradhan, B., & Assiri, M. (2023). Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *Computer Modeling in Engineering & Sciences*, *135*(2), 1053–1089. doi: 10.32604/cmes.2022.021755

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166. doi: 10.1109/72.279181

Bou-Ghazale, S., & Hansen, J. (1998). Hmm-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. *IEEE Transactions on Speech and Audio Processing*, *6*(3), 201-216. doi: 10.1109/89.668815

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: pre-training of deep bidirectional transformers for language understanding.* Association for Computational Linguistics. doi: 10.18653/V1/N19-1423

Du, J., Wang, Q., Gao, T., Xu, Y., Dai, L.-R., & Lee, C.-H. (2014). Robust speech recognition with speech enhanced deep neural networks. In *Proc. interspeech 2014* (pp. 616–620). doi: 10.21437/Interspeech.2014-148

Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 708-712). doi: 10.1109/ICASSP.2015.7178061

Gangmei, M. G. C., Singh, S. S., & Shougaijam, B. (2021). Design of low cost voice operated vending machine using v3 module. In *2021 ieee 2nd international conference on applied electromagnetics, signal processing, & communication (aespc)* (pp. 1–4). doi: 10.1109/AESPC52704.2021.9708486

Higuchi, Y., Tawara, N., Ogawa, A., Iwata, T., Kobayashi, T., & Ogawa, T. (2021). Noise-robust attention learning for end-to-end speech recognition. In *2020 28th european signal processing conference (eusipco)* (p. 311-315). doi: 10.23919/Eusipco47968.2020.9287488

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. doi: 10.1109/MSP.2012.2205597

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771-1800. doi: 10.1162/089976602760128018

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554. doi: 10.1162/neco.2006.18.7.1527

Irugalbandara, C., Naseem, A., Perera, S., Kiruthikan, S., & Logeeshan, V. (2023). A secure and smart home automation system with speech recognition and power measurement capabilities. *Sensors*, *23*(5784). doi: 10.3390/s23135784

Kim, S., Arora, A., Le, D., Yeh, C., Fuegen, C., Kalinli, O., & Seltzer, M. L. (2021). Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. *CoRR*, *abs/2104.02138*. doi: 10.48550/arXiv.2104.02138

Lamel, L., & Gauvain, J.-L. (2022, 06). 770Speech Recognition. In *The Oxford Handbook of Computational Linguistics.* Oxford University Press. doi: 10.1093/oxfordhb/9780199573691.013.37

Le Prell, C. G., & Clavier, O. H. (2017). Effects of noise on speech recognition: Challenges for communication by service members. *Hearing Research*, *349*, 76-89. (Noise in the Military) doi: https://doi.org/10.1016/j.heares.2016.10.004

Lee, K.-F., Hon, H.-W., & Reddy, R. (1990). An overview of the sphinx speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1), 35-45. doi: 10.1109/29.45616

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., & Synnaeve, G. (2020). Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint*. doi: 10.48550/arXiv.2010.11745

Maas, A. L., Le, Q. V., O'Neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. (2012). Recurrent neural networks for noise reduction in robust asr. In *Interspeech.* doi: 10.21437/Interspeech.2012-6

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech Communication*, *41*(4), 603-623. doi: 10.1016/S0167-6393(03)00099-2

Ochiai, T., Iwamoto, K., Delcroix, M., Ikeshita, R., Sato, H., Araki, S., & Katagiri, S. (2024). Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance. *ArXiv*, *abs/2404.14860*. doi: 10.48550/arXiv.2404.14860

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 5206-5210). doi: 10.1109/ICASSP.2015.7178964

Paulett, J. M., & Langlotz, C. P. (2012). Improving language models for radiology speech recognition. *Journal of Biomedical Informatics*(7), 12. doi: 10.1016/j.jbi.2008.08.001

Prasad, A., Jyothi, P., & Velmurugan, R. (2021). An investigation of end-to-end models for robust speech recognition. In *Icassp 2021 - 2021 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 6893-6897). doi: 10.1109/ICASSP39728.2021.9414027

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286. doi: 10.1109/5.18626

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4580-4584). doi: 10.1109/ICASSP.2015.7178838

Schlotterbeck, D., Jiménez, A., Araya, R., Caballero, D., Uribe, P., & Van der Molen Moris, J. (2022). "teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data. In *International conference on artificial intelligence in education* (pp. 269–280). doi: 10.1007/978-3-031-11644-5_22

Sivasankaran, S., Vincent, E., & Illina, I. (2017). Discriminative importance weighting of augmented training data for acoustic model training. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4885-4889). doi: 10.1109/ICASSP.2017.7953085

Suresh, S., Sandra, S., Thajudheen, A., Hussain, S., & Amitha, R. (2023). An intelligent voice assistance system for the visually impaired people. *International Journal of Engineering Research & Technology (IJERT)*, *11*(04). doi: 10.17577/IJERTCONV11IS04022

Szymański, P., Zelasko, P., Morzy, M., Szymczak, A., Zyla-Hoppe, M., Banaszczak, J., ... Carmiel, Y. (2020). WER we are and WER we think we are. *CoRR*, *abs/2010.03432*. doi: 10.48550/arXiv.2010.03432

Van Segbroeck, M., & Narayanan, S. S. (2013). A robust frontend for asr: Combining denoising, noise masking and feature normalization. In *2013 ieee international conference on acoustics, speech and signal processing* (p. 7097-7101). doi: 10.1109/ICASSP.2013.6639039

Vintsyuk, T. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, *4*, 11-18. doi: 10.1007/BF01074755

WANG, X., SHAN, X., & JING, H. (2022). Speech recognition method for railway ticket purchase based on the fusion of multiple language models. *Railway Transport and Economy*, *44*(3), 23–30. doi: 10.16668/j.cnki.issn.1003-1421.2022.03.04

Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 ieee workshop on automatic speech recognition and understanding (ieee cat. no.03ex721)* (p. 577-582). doi: 10.1109/ASRU.2003.1318504

Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., Kang, J., ... et al. (2024). Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*. doi: 10.48550/arXiv.2404.01954

Yu, D., & Deng, L. (2011). Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine*, *28*(1), 145-154. doi: 10.1109/MSP.2010.939038

Zhang, L. (2020). Research on voiceprint recognition technology based on deep learning. *Hangzhou Dianzi University*. doi: 10.27075/d.cnki.ghzdc.2020.000345

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., Jin, W., & Schuller, B. (2018, apr). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Trans. Intell. Syst. Technol.*, *9*(5). doi: 10.1145/3178115

Zhu, Q.-S., Zhang, J., Zhang, Z.-Q., Wu, M.-H., Fang, X., & Dai, L.-R. (2022). A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *Icassp 2022 - 2022 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 3174-3178). doi: 10.1109/ICASSP43922.2022.9747379