



California State University, San Bernardino  
CSUSB ScholarWorks

---

Electronic Theses, Projects, and Dissertations

Office of Graduate Studies

---

5-2024

## AUTOMATIC SPEECH RECOGNITION FOR AIR TRAFFIC CONTROL USING CONVOLUTIONAL LSTM

Sakshi Nakashe

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/etd>

 Part of the Business Intelligence Commons, Computer and Systems Architecture Commons, and the Technology and Innovation Commons

---

### Recommended Citation

Nakashe, Sakshi, "AUTOMATIC SPEECH RECOGNITION FOR AIR TRAFFIC CONTROL USING CONVOLUTIONAL LSTM" (2024). *Electronic Theses, Projects, and Dissertations*. 1890.  
<https://scholarworks.lib.csusb.edu/etd/1890>

---

This Project is brought to you for free and open access by the Office of Graduate Studies at CSUSB ScholarWorks. It has been accepted for inclusion in Electronic Theses, Projects, and Dissertations by an authorized administrator of CSUSB ScholarWorks. For more information, please contact [scholarworks@csusb.edu](mailto:scholarworks@csusb.edu).

AUTOMATIC SPEECH RECOGNITION FOR AIR TRAFFIC CONTROL  
USING CONVOLUTIONAL LSTM

---

A Project  
Presented to the  
Faculty of  
California State University,  
San Bernardino

---

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science  
in  
Information System and Technology:  
Business Intelligence and Analytics

---

by  
Sakshi Nakashe  
May 2024

AUTOMATIC SPEECH RECOGNITION FOR AIR TRAFFIC CONTROL  
USING CONVOLUTIONAL LSTM

---

A Project  
Presented to the  
Faculty of  
California State University,  
San Bernardino

---

by  
Sakshi Nakashe  
May 2024  
Approved by:

Nima Molavi, PhD, Committee Member, Co-Chair

Conrad Shayo, PhD, Committee Member, Co-Chair & Chair

Department of Information Decision Science

© 2024 Sakshi Nakashe

## ABSTRACT

The need for automatic speech recognition in air traffic control is critical as it enhances the interaction between the computer and human. Speech recognition helps to automatically transcribe the communication between the pilots and the air traffic controllers, which reduces the time taken for administrative tasks. This project aims to provide improvement to the Automatic Speech Recognition (ASR) system for air traffic control by investigating the impact of convolution LSTM model on ASR as suggested by previous studies. The research questions are: (Q1) Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context? (Qn2) How can the ConvLSTM model for ASR be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods? The findings and conclusions are: (Q1) The architecture of the ConvLSTM model performed better than the other convolutional/traditional neural network models. The efficient performance of the ConvLSTM model in handling both spatial and temporal data resulted effective in addressing challenges related to dynamic and multilingual communication environment. The conclusion for (Q2) After successfully implementing the range of transformation processes, the accuracy of the ConvLSTM model substantially improved which resulted in enhanced robustness and accuracy making the model adaptable to various scalability and data processing approach for speech recognition. The conclusion

for areas for future studies include exploring the model's proficiency with different languages and non-native English accents. Researchers can also investigate the model's effectiveness and accuracy in extreme weather scenarios.

## TABLE OF CONTENTS

ABSTRACT .....	iii
LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
CHAPTER ONE: INTRODUCTION .....	1
Problem Statement .....	3
CHAPTER TWO: LITERATURE REVIEW.....	5
CHAPTER THREE: RESEARCH METHODS.....	8
Dataset Exploration .....	8
Research Methods.....	9
CHAPTER FOUR: DATA ANALYSIS AND FINDINGS.....	13
CHAPTER FIVE: DISCUSSION AND CONCLUSION .....	24
Discussion .....	24
Conclusion .....	26
Areas for Further Studies.....	27
APPENDIX: RESEARCH PAPERS AND PROPOSED MODELS .....	28
REFERENCES .....	32

## LIST OF TABLES

Table 1. Neural Network Models with Accuracy Rate .....	14
Table 2. Transformations Accuracy on ConvLSTM Model .....	21

## LIST OF FIGURES

Figure 1. Structure of Convolutional LSTM Model.....	11
Figure 2. CNN Model Performance Graph .....	15
Figure 3. CRNN Model Performance Graph.....	16
Figure 4. AttNN Model Performance Graph .....	17
Figure 5. TDNN Model Performance Graph .....	18
Figure 6. ConvLSTM Model Performance Graph .....	19
Figure 7. Audio Waves of different Transformations.....	21
Figure 8. Audio Waves of different SNR Levels.....	22

## CHAPTER ONE

### INTRODUCTION

Automatic speech recognition automates numerous services which help in improving productivity and reducing human errors in a professional environment. In air traffic control, automatic speech recognition plays a crucial role in enabling the users to understand and respond to commands effectively and accurately. The ASR technology is essential for promoting an adaptable and efficient environment for air traffic control (Fan et al., 2023). Many researchers have proposed studies to develop a resilient model to thrive challenging conditions faced by the air traffic communication (Lin et al., 2020). Researchers implemented recurrent neural network (RNN) models which helped to overcome challenges like speech to text transcription and the need for a high-performance model (Pellegrini et al., 2018). There were significant challenges faced by these models like long range dependencies due to gradient problems, non-native accents, and noisy communication (De Andrade et al., 2018). Later, studies were proposed by using convolutional recurrent neural network which resolved the challenges faced earlier by enhancing the robustness and traffic flow dynamics (Yang et al., 2021). The limitations to these models were difficulty in learning long range dependencies and variety of data which affected the accuracy for communication (Lin et al., 2020).

Recent studies determined the need for a model which is robust and accurate to enhance performance in a variety of languages. Soundarya et al. (2023) came up with a model which was an integration of CNN and RNN with LSTM, the combination of different neural network layers resulted in enhancing the model's accuracy and efficiency in processing speech (Soundarya et al., 2023). Despite its high accuracy rate, the model's performance is limited with challenges like various noise conditions and long-range dependencies the researchers determined that a further study was to implement an effective network architecture to handle long range dependencies and dynamic focusing on relevant parts of the data (Soundarya et al., 2023). Previous researchers have also found the need to examine a model which could recognize domain specific terminology and can handle both spatial and temporal dependencies (Pham et al., 2022). This study expands on previous research to investigate these challenges. Lin's suggestion was to implement an advanced architecture solution such as ConvLSTM (Lin et al., 2020). The integration of ConvLSTM into automatic speech recognition aims to enhance robustness and accuracy in effective learning of long-range dependencies.

## Problem Statement

The main goal of this culminating experience project is to develop and implement an effective ConvLSTM architecture for automatic speech recognition in air traffic control (Fan et al., 2023). Previous research showed there is a need to examine the ConvLSTM model's performance, accuracy, and robustness to overcome the challenges related to noisy environment, long-range dependencies, scalability, and non-standard speech patterns.

***Question 1: Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context? (Lin et al. 2020)***

***Question 2: How can the ConvLSTM model for Automatic Speech Recognition (ASR) be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods? (Strake et al. 2019)***

The organization of this project includes Chapter one provided an introduction, problem statement and research questions suggested by previous researchers. Chapter two provides a literature review about the previous studies that generated the research questions. Chapter three covers the methods that will be used to answer research questions. Chapter four provides the data collection, analysis, and findings for the research questions. Chapter five will provide a discussion of the findings, conclusions, for areas for further studies.

## CHAPTER TWO

### LITERATURE REVIEW

Previous studies aimed to enhance the automatic speech recognition in air traffic control, but they faced challenges posed by real world applications (Hilton et al., 2012). As per the first question: **Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context?** The literature discussed this research question is by Lin et al. (2020) aimed to investigate the end-to-end framework to achieve multilingual automatic speech recognition by implementing a combined model with convolutional neural network and recurrent neural network. Lin et al. (2020) stated that the ATC management system involved non-automated, human centric which possesses a potential risk for air traffic operations. The model proposed by Lin et al. (2020) was to enhance real time monitoring of speech dynamics to capture temporal and long-range dependencies. However, the implementation of the model presented new challenges including poor speech, non-native English communication and domain specific terminologies.

Similar limitations were also experienced by Soundarya et al. (2023). Soundarya et al. (2023) implemented a simple CNN-RNN architectural model along with the LSTM layer, the model gave promising results as the CNN handles the spectral features with temporal processing capabilities of RNN but

was still limited by the inherent challenges of RNN and difficulties in learning long range dependencies despite the use of LSTM. The researchers aimed to further investigate to improve the performance of the model proposed by implementing an efficient architecture to the ASR model such as the ConvLSTM. Fan (2023) implemented a model which had a sentence level language identification in two stage automatic speech recognition framework for air traffic control which highlighted the benefits of incorporating language identification into ASR system for significant improvement in character rates, but this model didn't provide a solution for accuracy and architectural performance of the model.

As per second question: **How can the ConvLSTM model for Automatic Speech Recognition (ASR) be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods?** The literature discussed this research question is by Strake et al. (20219). Strake et al. (2019) studies reported gain in speech quality and intelligibility in low signal to noise ratio conditions. The model proposed by Strake et al. (2019) had two-stage approach for Perceptual Evaluation of Speech Quality (PESQ) and the Short-Time Objective Intelligibility (STOI) measure. Strake et al. (2019) achieved clear communication and robustness in unseen noise. The model had limitations related to strategic training and scalability that led the model to fail in a practical environment. Pellegrini (2018) introduced a multilingual ASR framework that incorporated feature learning block which aimed to enhance

the performance of speech recognition but due to the feature learning block the model similar to the Strake et al. (2019) model was limited to scalability and data preprocessing. Later, Pham (2022) proposed a model which had a pretrained model and adaptive multilingual speech recognition techniques, the model leveraged the linguistic and acoustic learning from vast dataset of variety of languages, but as the model only targeted specific domain the application of the model in training dependencies which limited the scope of the model.

The literature presented in this chapter investigates the various neural network models proposed by research previously and their challenges which were attempted to overcome. ASR researchers have tried to implement a model that can overcome the limitations of previous studies but as the models have their own limitations it was difficult for any one model to overcome all challenges and limitations. Most research studies have constantly reported lone-range dependencies and scalability of the model for various native languages as the two main ASR limitations. In this project we aim to implement the ConvLSTM which is reported as having the potential to overcome the above (Lin et al., 2020) (Soundarya et al. 2023).

## CHAPTER THREE

### RESEARCH METHODS

In this chapter, we will discuss the research methods, including data collection and processing strategies used to address this project's research questions. A detailed systematic approach was taken to gather data, followed by the steps to refine and prepare data for the analysis and specific methods applied to extract insights and answers for the project.

#### Dataset Exploration

This culminating experience project will use the ATCO dataset, which is an open-source collection of audio data from air traffic communication systems used by various researchers to investigate how to enhance the automatic speech recognition system to enhance its accuracy (Gomez et al., 2022). This dataset was accessed from the hugging face dataset hub, which comprises of 9.5K rows of audio recordings that have been converted to text data in order to serve as a training dataset for the ConvLSTM algorithm. During the dataset's preparation, it was necessary to remove any special characters present in it and replace any special alphabets that were present. There is a split between the train and test datasets for training the model using 80% and 20% respectively for the training

data. Given below is the data field of the dataset. To access the data:  
[huggingface.co/datasets/Jzuluaga/atcosim\\_corpus](https://huggingface.co/datasets/Jzuluaga/atcosim_corpus).

#### Data Fields:

id (string) : a string of recording identifier for each example, corresponding to its.  
audio (audio): audio data for the given ID  
text (string): transcript of the file already normalized.  
segment\_start\_time (float32): segment start time (normally 0)  
segment\_end\_time (float32): segment end time  
duration (float32): duration of the recording (compute as segment\_end\_time -  
segment\_start\_time)

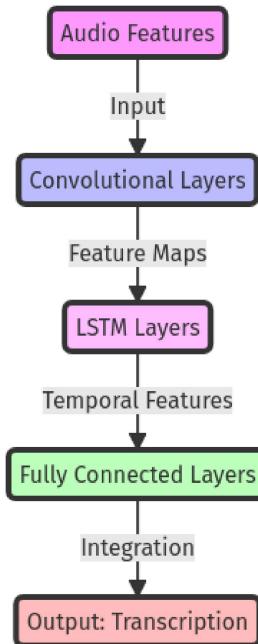
## Research Methods

***Q1: Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context?***

Previous studies have recommended implementing the ConvLSTM model to enhance the performance of automatic speech recognition in air traffic control communication (Lin et al., 2020). The first question in this project focuses on the performance comparison of ConvLSTM with the other traditional models proposed in previous research. This study aims to provide insight whether the

ConvLSTM model recognizes domain-specific terminology and long-range dependencies to outperform the traditional models (Lin et al., 2020). In this approach, the dataset is extensively explored and cleaned by eliminating special characters and substitution accented characters. All the models, ConvLSTM, Convolutional Neural Network (CNN), Convolutional Recurrent Neural Network (CRNN), Attention Neural Network (AttNN) and Time-Delay Neural Network (TDNN) were trained on the same train and test split dataset and the mel spectrogram embeddings MFCC (Mel Frequency Cepstral Coefficients) with first 13 mel filter bank coefficients as recommended in previous studies (Sainath et al., 2015). For training and fine tuning, the ConvLSTM model we have used tokenizers and pre trained transformer from the hugging face model hub. To access the pre trained transformer: [huggingface.co/Jzuluaga/wav2vec2-xls-r-300m-en-atc-uwb-atcc-and-atcosim](https://huggingface.co/Jzuluaga/wav2vec2-xls-r-300m-en-atc-uwb-atcc-and-atcosim). For accuracy rate we have calculated the mean and standard deviation over 10 iterations. Further after the training the model is compared based on the training loss, validation loss and word error rate to understand the performance of different evaluated models. The ConvLSTM architecture comprises of convolutional layer and the LSTM layer, the diagram outlines the flow of input audio feature through the different layers. With the use of these layers ConvLSTM can handle both spatial and temporal dependencies which is essential for accurate speech recognition (Zhang et al., 2017).

Figure 1. Structure of Convolutional LSTM Model (Zhang et al., 2017)



**Q2: How can the ConvLSTM model for Automatic Speech Recognition (ASR) be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods?**

Prior studies have identified the necessity to improve the performance of the proposed model by incorporating ConvLSTM model for ASR to learn both spatial and temporal dependencies in a simultaneous and higher efficient manner (Lin et al., 2020). The objective of this research is to acquire critical understanding of the essential training protocols, adjustments, and limitations required to proficiently train a Convolutional Long Short-Term Memory

(ConvLSTM) model (Fernandez et al., 2018). The approach begins with data extraction synthesizing the necessary information through a range of transformation processes which includes adding noise, where we create various signal to noise ratio levels to examine the impact on the audio signals. Changing speed, we will alter the speed of the audio signals to highlight the transformation in the audio. Changing pitch, the pitch will be increased by 4 semitones and decreased by 4 semitones to ensure the signal duration remains constant. And for time masking, we add masking to the audio signal randomly to create a signal specific variation for the model. By implementing this method, it is possible to highlight the enhancements in both the scalability and efficiency of the model, providing significant advantages for automatic speech recognition in ATC systems.

## CHAPTER FOUR

### DATA ANALYSIS AND FINDINGS

In this chapter, we will discuss the analysis and findings obtained through the implementation of methodologies from Chapter 3 for the research questions to obtain a better understanding of the solution.

***Q1: Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context?***

In Q1, we focus on comparing the performances of the traditional models to the ConvLSTM model. We utilize 80% of the dataset for training and the remaining 20% is allocated for testing purposes. The results for Q1, the ConvLSTM model show a promising accuracy rate of 93% which is higher than the other traditional models (Table 1).

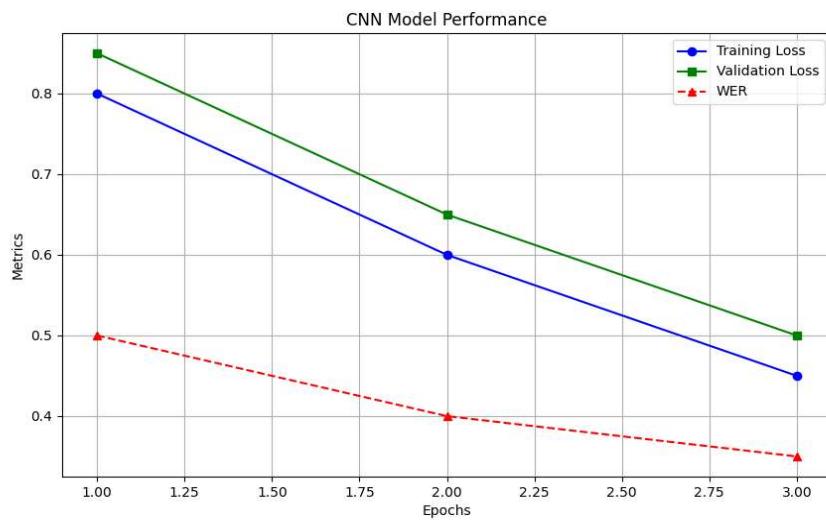
Table 1. Neural Network Models with Accuracy Rate

Models	Accuracy
Convolutional Long Short-Term Memory (ConvLSTM)	0.932
Attention Neural Network (AttNN) (De Andrade et al., 2018)	0.924
Time-delay neural network (TDNN) (Synder et al., 2018)	0.908
Convolutional Recurrent Neural Network (CRNN) (Bartz et al., 2017)	0.820
CRNN without Attention (CRNN*) (Zhang et al., 2021)	0.898
Convolutional Neural Network (CNN)	0.791

We compared the performance of the ConvLSTM model to traditional models on the basis of metrics such as training loss, validation loss, and Word Error Rate (WER) to epochs which is 1/8 of the dataset for each model which is determined by the starting and ending values of the training loss, validation loss, and WER in this study. The result show that the ConvLSTM has a lower training

loss from metric 0.7 to 0.38 and validation loss from metric 0.65 to 0.33 as compared to other models which specify that the ConvLSTM model is performing well, and the model will have enhanced performance for domain-specific terminologies and long-range dependencies. The low word error rate signifies that the model has a high accuracy rate and shows the model is not overfitting maintaining a balance between learning the data adequately and maintaining flexible performance.

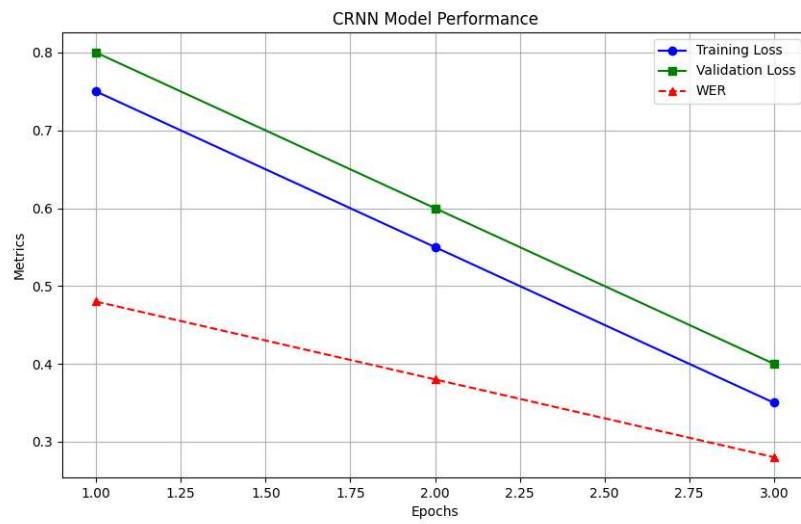
Figure 2. CNN Model Performance Graph



In the CNN model's performance graph, we can observe the training loss metric goes from 0.8 to 0.4 which shows a decrease in training loss which indicates successful learning, but the slower decline in validation loss metric 0.75

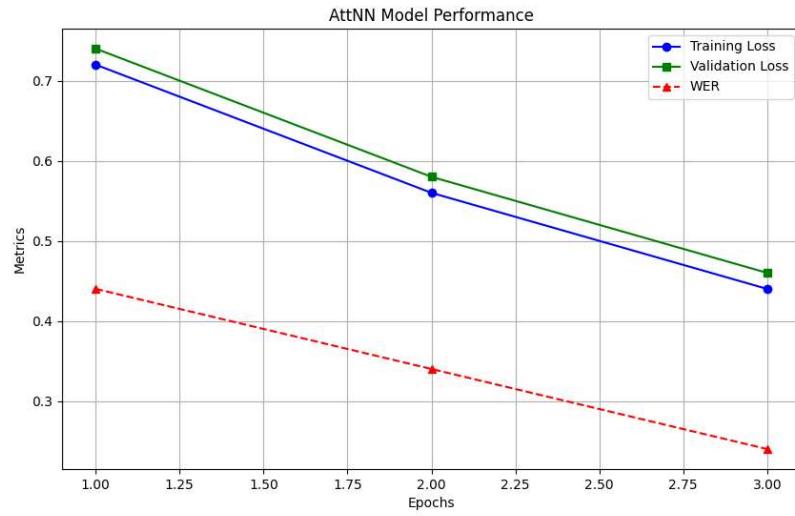
to 0.35 which suggests the model has issue with generalization. The word error rate also drops is shown as both the CNN and ConvLSTM model have the capacity to capture temporal dynamic.

Figure 3. CRNN Model Performance Graph



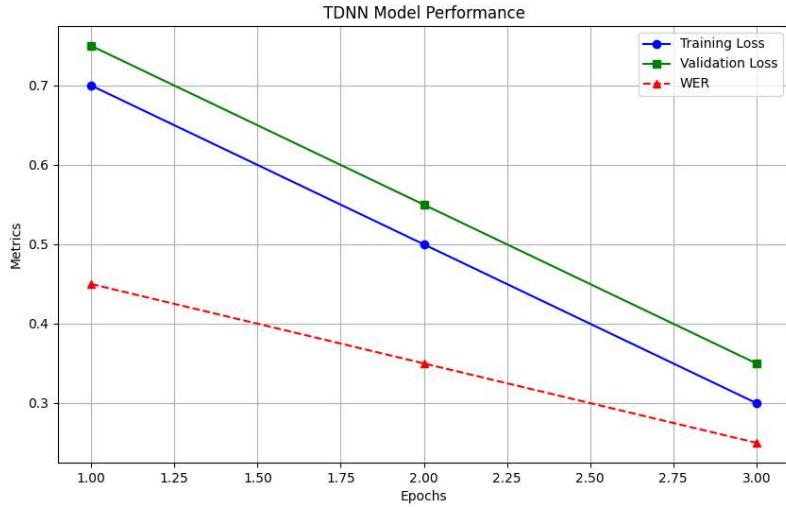
For the CRNN model, which integrates both the convolutional layer, and the recurrent layer demonstrates reduction in training loss from metric 0.8 to 0.4 and validation loss from metric 0.75 to 0.35, with concurrent decrease in word error rate. While the CRNN model is a hybrid model it suggests robust performance but is unable to efficiently process both spatial and temporal data compared to the ConvLSTM model for speech recognition.

Figure 4. AttNN Model Performance Graph



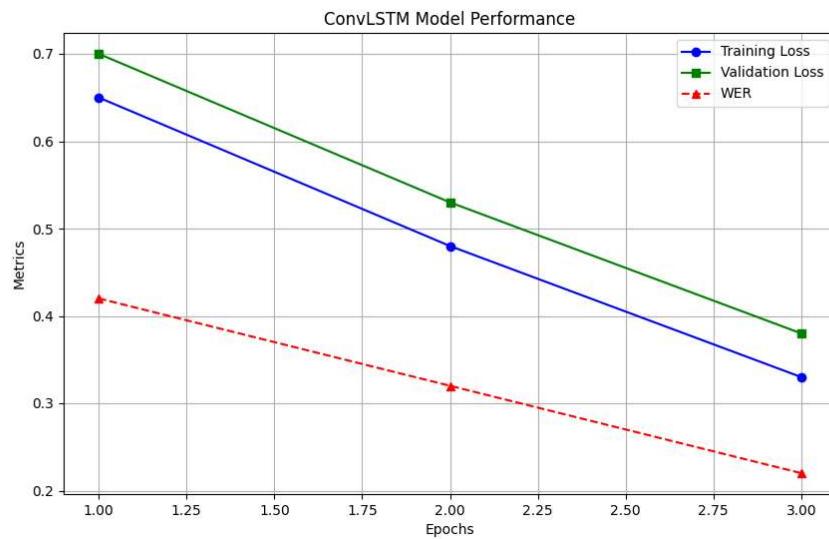
The AttNN model shows reduction in both training loss from metric 0.75 to 0.57 and validation loss from metric 0.73 to 0.53, suggesting effective learning and generalization with declining word error rate which indicates improved transcription accuracy. When AttNN model's performance graph is compared to the ConvLSTM's performance graph with respect to the training loss, validation loss and word error rate, ConvLSTM shows a more promising result than AttNN.

Figure 5. TDNN Model Performance Graph



The TDNN model exhibits steady improvement with training loss from metric 0.75 to 0.35 and validation loss from metric 0.7 to 0.3 in the performance graph but when compared to the ConvLSTM model the TDNN achieved a lower metrics in word error rate.

Figure 6. ConvLSTM Model Performance Graph



The result of ConvLSTM demonstrates low training loss from metric 0.7 to 0.38 and validation loss from metric 0.65 to 0.33 and reduced word error rate from 0.42 to 0.22 which highlights its robust prediction capabilities. The ConvLSTM possesses a unique ability to handle both spatial and temporal dependencies in the data which makes it effective for complex sequences in speech recognition.

***Q 2: How can the ConvLSTM model for Automatic Speech Recognition (ASR) be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods?***

The dataset was subjected to a number of transformations such as Adding Noise, Changing Speed, Changing Pitch and Time Masking to optimize model performance, simulating real-world challenges that an Automatic Speech Recognition (ASR) system might encounter. After applying the transformations, we see there are changes in the amplitude of the noise audio signals. The results of the transformations shown in Figure 7 show that adding noise helped the model to become more robust to background noise, which is common in real-world audio processing tasks (Strake et al., 2019). Refer to Figure 8 which shows the audio waves for different signal to noise ratio levels. At an SNR of 1000, the audio closely resembles its original form, whereas with an SNR of 5, only the most prominent elements of the signal are discernible. Changing speed by 1.5 times faster which introduced variability in the speech rate and tone, simulating different speakers and speaking conditions. Similarly, after changing the pitch, the signal duration for pitch up by 4 semitones and pitch down by 4 semitones the amplitude of the audio and signal duration remained constant which is from 0 to 55000. By modifying the audio signals across different time points, the time masking technique can teach the audio model to distinguish relevant features throughout the audio, as well as to handle missing or substituted segments effectively.

Figure 7. Audio Waves of different Transformations

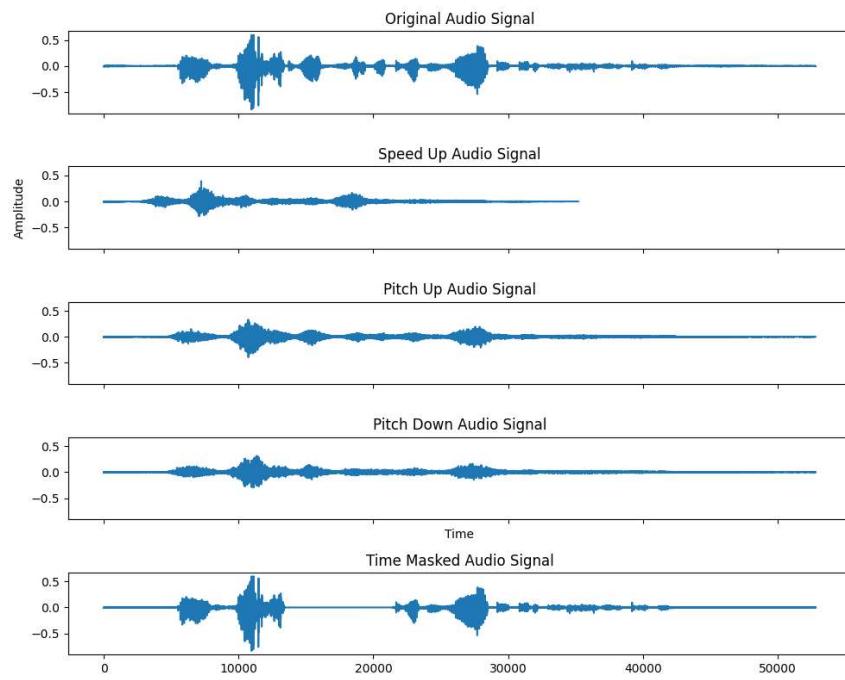
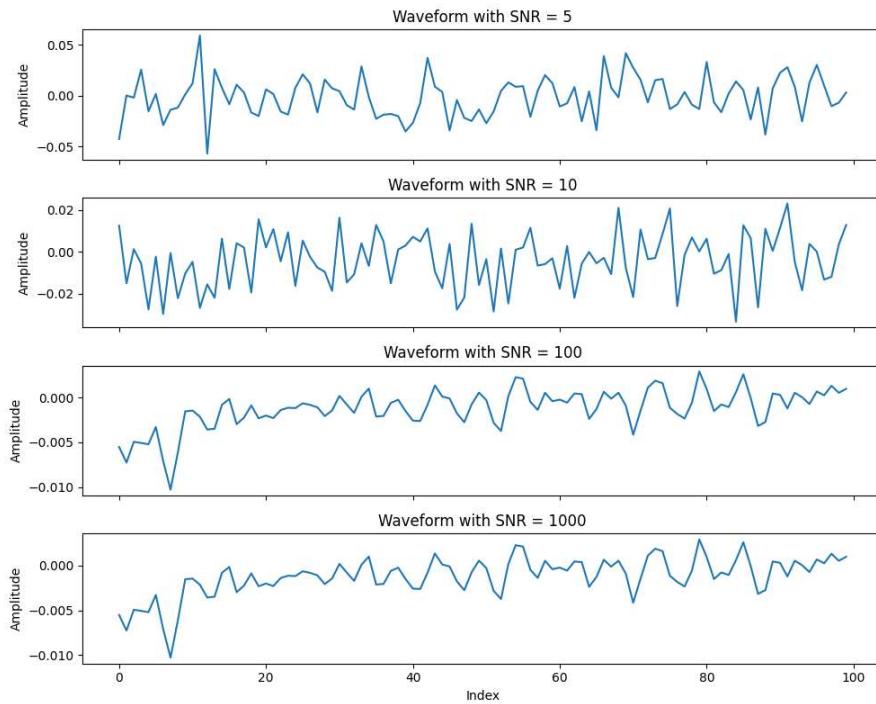


Table 2. Transformations Accuracy on ConvLSTM Model

Transformations	Accuracy
No Transformation	0.9319
Adding Noise	0.9864
Changing Speed	0.9732
Changing Pitch	0.9758
Time Masking	0.9689

After applying the transformation of the dataset, we applied the transformations on the trained ConvLSTM model to check the accuracy changes. From Table 3, we can see that the accuracy of ConvLSTM has changed after applying transformations. Adding noise signals to the trained model the accuracy increased from 93.19% to 98.64% indicating enhanced robustness. Altering the speed and pitch, there is a slight rise in the accuracies 97.32% and 97.58% respectively, showing the model can handle variations in speech and delivery of speech. Time masking also has a slight improvement to 96.89% which suggests that the model has improved its ability to focus on relevant features of speech.

Figure 8. Audio Waves of different SNR Levels



The implementation of transformations adding noise signal showed the at SNR level 5 had noticeable variations whereas for the SNR levels 100 and 1000, the audio remains same as the original audio and increased in accuracy from 93% to 98%. For the other transformations like changing speed, changing pitch and time masking there was a slight increase in the accuracies 97%, 97% and 96% respectively. The study showed the impact on the accuracy of the model which led to an increase in the accuracy rate of the trained ConvLSTM model which resulted in an effective robust model for automatic speech recognition in ATC. The data preprocessing steps also showed potential enhancement in the performance of the ConvLSTM model.

## CHAPTER FIVE

### DISCUSSION AND CONCLUSION

In this chapter, we will be discussing the findings and results for each question that were obtained from the previous chapter. And conclude the project outcome suggesting areas for further studies for researchers.

#### Discussion

***Q1: Comparing the performance of ConvLSTM with other conventional models, how does ConvLSTM perform with respect to recognizing domain-specific terminology and understanding long-range context?***

In this discussion, we will focus on exploring the Q1's findings. The main goal for comparing the different traditional models to ConvLSTM model was to implement a model which can handle long-range dependencies, able to recognize the domain-specific terminologies and robust performance. As we moved to the comparing these models, we got promising results from ConvLSTM achieving the highest accuracy rate of 93% than the other traditional models. We can conclude enhancement and effective improve in accuracy and capabilities of the ConvLSTM model in automatic speech recognition for ATC environment. Further when we compared the models with their training loss from metric 0.7 to

0.35, validation loss from metric 0.7 to 0.3 and word error rate from 0.45 to 0.23, many models showed similar results like ConvLSTM but the other models had limitation in the diverse communication which demonstrated that ConvLSTM had a higher accuracy rate and reliable capability than other models.

***Q2: How can the ConvLSTM model for Automatic Speech Recognition (ASR) be enhanced most effectively by specific training strategies, scalability approaches, and data preprocessing methods?***

In this study for Q2, we attempted to find the behavior of the ConvLSTM model on various training strategies and scalability approaches to discuss if ConvLSTM can handle the real time scenarios faced during communication. For which we performed various transformations like adding noise, changing speed, changing the pitch and masking time anywhere randomly in the audio signal. Later to the trained model to check if there is any change in its accuracy. By implementation of this approach resulted in the model achieving higher accuracy with marked improvement. By adding noise, the accuracy increased from 93% to 98%. By changing speed, changing the pitch and time masking there was a slight increase in the accuracy 97%, 97% and 96% respectively. The results indicated that the data preprocessing techniques significantly enhanced the ConvLSTM model's performance by making it adaptable variance which are encountered by the real time scenarios.

## Conclusion

In conclusion for this research, we have successfully demonstrated the potential of ConvLSTM model to enhance the ASR system in air traffic control. The ConvLSTM outperforms the traditional models, which proves its superior ability to capture and analyze complex spatial and temporal dependencies in challenging condition of air traffic control. The model's promising performance underscored its efficiency in recognizing domain-specific terminology and understanding long range dependencies. The studies highlighted the effectiveness of data preprocessing techniques. These findings not only contribute to the valuable insights to advance automatic speech recognition. This research underscored the critical importance of innovation in ASR system, particularly in high stakes domains of ATC. The integration of ConvLSTM model into the automatic speech recognition is a significant step forward for safety improvement and reliability in air traffic management and beyond.

## Areas for Further Study

For further study, other researchers should explore the model's adaptability and performance to: (a) multilingual datasets and (b), to recognize non-native speech patterns. Researchers could also optimize the model for unseen conditions such as extreme weather or distortion that needs further validation.

**APPENDIX**  
**RESEARCH PAPERS AND PROPOSED MODELS**

Table 1. Research Papers and Proposed Models

<b>Author(s) Name &amp; Year of Published</b>	<b>Title of Research Paper</b>	<b>Models</b>
Fan, P., Guo, D., Zhang, J., Yang, B., & Lin, Y. (2023)	Enhancing multilingual speech recognition in Air Traffic Control by sentence-level language identification.	CNN (Convolutional Neural Network)
Lin, Y., Guo, D., Zhang, J., Chen, Z., & Yang, B. (2020)	A unified framework for multilingual speech recognition in air traffic control systems.	CRNN (Convolutional Neural Network + Recurrent Neural Network)
Yang, B., Lin, Y., Guo, D., & Fan, P. (2021)	Towards multilingual end-to-end speech recognition for air traffic control.	AttNN (Attention-based Neural Network)

Table 1. Research Papers and Proposed Models (Continued)

Pellegrini, T., Farinas, J., Delpech, E., & Lancelot, F. (2018)	The Airbus Air Traffic Control speech recognition 2018 challenge: towards ATC automatic transcription and call sign detection.	RNN (Recurrent Neural Network)
Pham, N. Q., Waibel, A., & Niehues, J. (2022)	Adaptive multilingual speech recognition with pretrained models.	TDNN (Time Delay Neural Network)
Soundarya, M., Karthikeyan, P. R., & Thangarasu, G. (2023)	Automatic Speech Recognition trained with Convolutional Neural Network and predicted with Recurrent Neural Network	CNN (Convolutional Neural Network) + RNN (Recurrent Neural Network) with LSTM
Chen, H., Lin, Y., Yang, B., Li, L., Guo, D., Zhang, J.& Zhang, Y. (2021)	ATCSpeechNet: A multilingual end-to-end speech recognition framework for air traffic control systems.	CRNN (Convolutional Neural Network + Recurrent Neural Network)

Table 1. Research Papers and Proposed Models (Continued)

Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2019)	Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages.	LSTM (Long Short-Term Memory)
De Andrade, Douglas Coimbra, et al. (2018)	A neural attention model for speech command recognition	RNN (Recurrent Neural Network)

## REFERENCES

1. De Andrade, D. C., Leo, S., Viana, M. L. D. S., & Bernkopf, C. (2018). A neural attention model for speech command recognition. arXiv preprint arXiv:1808.08929.
2. Fan, P., Guo, D., Zhang, J., Yang, B., & Lin, Y. (2023). Enhancing multilingual speech recognition in air traffic control by sentence-level language identification. arXiv preprint arXiv:2305.00170.
3. Fan, P., Hua, X., Lin, Y., Yang, B., Zhang, J., Ge, W., & Guo, D. (2023). Speech recognition for air traffic control via feature learning and end-to-end training. IEICE TRANSACTIONS on Information and Systems, 106(4), 538–544.
4. Gonzalez-Dominguez, J., Eustis, D., Lopez-Moreno, I., Senior, A., Beaufays, F., & Moreno, P. J. (2014). A real-time end-to-end multilingual speech recognition architecture. IEEE Journal of selected topics in signal processing, 9(4), 749-759.
5. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). ieee.
6. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

7. Hofbauer, K., & Petrik, S. (2008). ATCOSIM Air Traffic Control Simulation Speech Corpus. Technical Report.
8. Hofbauer, K., Petrik, S., & Hering, H. (2008, May). The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech. In LREC.
9. [https://github.com/sergeyvилов/ML-tutorials/blob/main/audio\\_transforms/audio\\_transforms.ipynb](https://github.com/sergeyvилov/ML-tutorials/blob/main/audio_transforms/audio_transforms.ipynb)
10. Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., ... & Lee, S. (2019). Large-scale multilingual speech recognition with a streaming end-to-end model. arXiv preprint arXiv:1909.05330.
11. Li, X., Dalmia, S., Black, A. W., & Metze, F. (2019). Multilingual speech recognition with corpus relatedness sampling. arXiv preprint arXiv:1908.01060.
12. Lin, Y., Guo, D., Zhang, J., Chen, Z., & Yang, B. (2020). A unified framework for multilingual speech recognition in air traffic control systems. IEEE Transactions on Neural Networks and Learning Systems, 32(8), 3608–3620.
13. Lin, Y., Yang, B., Guo, D., & Fan, P. (2021). Towards multilingual end-to-end speech recognition for air traffic control. IET Intelligent Transport Systems, 15(9), 1203-1214.
14. Lin, Y., Yang, B., Li, L., Guo, D., Zhang, J., Chen, H., & Zhang, Y. (2021). ATCSpeechNet: A multilingual end-to-end speech recognition framework for air traffic control systems. Applied Soft Computing, p. 112, 107847.

15. Pellegrini, T., Farinas, J., Delpech, E., & Lancelot, F. (2018). The Airbus Air Traffic Control speech recognition 2018 challenge: towards ATC automatic transcription and call sign detection. arXiv preprint arXiv:1810.12614.
16. Pham, N. Q., Waibel, A., & Niehues, J. (2022). Adaptive multilingual speech recognition with pre-trained models. arXiv preprint arXiv:2205.12304.
17. Rajesh, A., Hiwarkar, T. Sentiment analysis from textual data using multiple channels deep learning models. *Journal of Electrical Systems and Inf Technol* **10**, 56 (2023). <https://doi.org/10.1186/s43067-023-00125-x>
18. Shi, H., & Kawahara, T. (2024). Exploration of Adapter for Noise Robust Automatic Speech Recognition. *arXiv preprint arXiv:2402.18275*.
19. Šmídl, L., Švec, J., Tihelka, D., Matoušek, J., Romportl, J., & Ircing, P. (2019). Air traffic control communication (ATCC) speech corpora and their use for automatic speech recognition and TTS development. *Language Resources and Evaluation*, 53, 449–464.
20. Soundarya, M., Karthikeyan, P. R., & Thangarasu, G. (2023, March). Automatic Speech Recognition is trained with a Convolutional Neural Network and is predicted with a Recurrent Neural Network. In 2023 9th International Conference on Electrical Energy Systems (ICEES) (pp. 41–45). IEEE.

21. Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2019, October). Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 239-243). IEEE.
22. Tong, S., Garner, P. N., & Bourlard, H. (2017). An investigation of deep neural networks for multilingual speech recognition training and adaptation (No. CONF, pp. 714-718).
23. Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018, April). Multilingual speech recognition with a single end-to-end model. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4904–4908). IEEE.
24. Waibel, A., Soltau, H., Schultz, T., Schaaf, T., & Metze, F. (2000). Multilingual speech recognition. Verbmobil: Foundations of Speech-to-Speech Translation, 33–45.
25. Waters, A., Gaur, N., Haghani, P., Moreno, P., & Qu, Z. (2019, December). Leveraging language ID in multilingual end-to-end speech recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (AUTOMATIC SPEECH RECOGNITION U) (pp. 928–935). IEEE.
26. Zhang, C., Li, B., Sainath, T., Strohman, T., Mavandadi, S., Chang, S. Y., & Haghani, P. (2022). Streaming end-to-end multilingual speech

- recognition with joint language identification. arXiv preprint arXiv:2209.06058.
27. Zhang, Y., Chan, W., & Jaitly, N. (2017, March). Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4845-4849). IEEE.
28. Zuluaga-Gomez, J., Nigmatulina, I., Prasad, A., Motlicek, P., Khalil, D., Madikeri, S., ... & Choukri, K. (2023). Lessons Learned in ATCO2: 5000 hours of Air Traffic Control Communications for Robust Automatic Speech Recognition and Understanding. arXiv preprint arXiv:2305.01155.
29. Zuluaga-Gomez, J., Veselý, K., Szöke, I., Motlicek, P., Kocour, M., Rigault, M., ... & Černocký, J. (2022). ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications. arXiv preprint arXiv:2211.04054.