Automatic Speech Recognition and Understanding Over Noisy Air Traffic Control VHF Channels in Singapore

Phuong Tuan Dat
Institute for Infocomm Research (I^2R)*
Agency for Science, Technology and Research (A*STAR)
Singapore
phuongtuandat2915@gmail.com

Luong Trung Tuan
Institute for Infocomm Research (I^2R)
Agency for Science, Technology and Research (A*STAR)
Singapore
luong-tt@i2r.a-star.edu.sg

Jayakrishnan Melur Madhathil
Institute for Infocomm Research (I^2R)
Agency for Science, Technology and Research (A*STAR)
Singapore
Jayakrishnan_Melur_Madhathil@i2r.a-star.edu.sg

Tran Huy Dat
Institute for Infocomm Research (I^2R)
Agency for Science, Technology and Research (A*STAR)
Singapore
hdtran@i2r.a-star.edu.sg

***Abstract*—In recent years, the demand for Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU) systems within the Air Traffic Control (ATC) domain has been increasing, especially systems that can be applied in practice. These systems are essential for reducing the workload of pilots and air traffic control officers (ATCOs) and ensuring the utmost accuracy in communication between pilots and ATCOs. However, ATC remains a low-resource and challenging domain. This paper presents our work on developing an ASR engine and an SLU system for ATC in Singapore, addressing these challenges. We introduce the Singapore Air Traffic Control (S-ATC) dataset, aimed at fostering research in this demanding field. We then discuss our contributions to constructing an efficient ASR system tailored for the ATC domain. Experimental results are provided to evaluate the effectiveness of combining an ASR system with a Natural Language Processing (NLP) model versus an End-to-End system for the SLU tasks in this specific domain. Additionally, we try to implement a model for Speaker Role Detection (SRD) task and propose ideas to enhance the efficiency of these systems in the ATC domain in the future.**

***Keywords*—Robust Automatic Speech Recognition, Natural Language Processing, Air Traffic Control Communications, Spoken Language Understanding, Signal Processing**

I. Introduction

Air Traffic Control (ATC) can be a highly stressful and voice-intensive domain due to the critical requirements of safety, reliability, and efficiency. The primary mode of communication between pilots and Air Traffic Control Officers (ATCOs) is via Very High Frequency (VHF) radio, a technology known for its dependable voice communication capabilities. Nonetheless, VHF radio technology is not without its limitations, particularly in terms of inherent noise. Furthermore, the quality of VHF radio signals is greatly influenced by external factors, such as weather conditions [1]. These issues can lead to incomplete reception of information by pilots, causing misunderstandings of vital instructions and significantly compromising aviation safety.

One of the challenges faced by ASR engines in the context of air traffic control is the diversity of pilot and ATCOs accents, as pilots come from various countries around the globe, whereas most ATCOs in this scenario are Singaporeans. Additionally, both pilots and ATCOs tend to speak rapidly due to the demands of their work [2]. The presence of numerous local terms, slang, and specialized jargon specific to the ATC domain further complicates the task, necessitating thorough research and sophisticated modeling to ensure accurate recognition and understanding.

To address these challenges, we introduce a high-quality ASR dataset specifically designed for the ATC domain. This dataset includes more label types for the slot-filling task compared to [3], and it can also support intent detection. Using our dataset, we implemented both an ASR engine and a Spoken Language Understanding (SLU) system. For the ASR task, we experimented with several methodologies, such as fine-tuning the OpenAI Whisper model [4], the Self-supervised Learning (SSL) model Wav2Vec 2.0 [5], and training an End-to-End system based on the WeNet framework [6] [7]. The SLU system was developed using two primary approaches: integrating an ASR model with a Natural Language Understanding (NLU) model or constructing an End-to-End SLU system that directly extracts high-level information from the conversation. Previous work [3] utilized ground truth text in NLU tasks; however, ground truth text is not available in practical applications. Thus, our research addresses this limitation by developing SLU systems that operate directly on audio input without relying on text transcriptions.

*Work done while interning at I^2R.

In addition to the challenges associated with ASR and SLU, Speaker Role Detection (SRD) is also one of the important tasks with significant practical applications in the ATC domain. Ground truth texts were used in one of the earlier ATC studies [3] for study in the SRD task in the difficult ATC domain. A BERT model was refined in [3] utilizing the UWB-ATCC[1] and LDC-ATCC[2] datasets. The F1 score for the ground truth texts is 0.84, with 0.85 for the "Pilot" label and 0.83 for the "ATC" label. While the outcomes show promise, in actual use, ground truth texts are not available, which means that the model presented in [3] is not yet suitable for practical situations. This paper presents several research findings on the application of text, audio, and a combination of both text and audio to address the SRD task in real-world conditions.

The subsequent sections in this paper are organized as follows: Section II describes the construction of the ATC dataset, and Section III discusses the experimental results of the ASR system using different methodologies. Section IV shows various experiments of Speaker Role Detection (SRD) task in the ATC domain along with several ways to build the SLU system in Section V. Finally, Section VI concludes our contributions to this paper.

II. The Singapore Air Traffic Control Corpus

In this section, we provide an overall description of the dataset utilized for our experiments. This dataset comprises approximately 440 hours of short utterances from dialogues in English between pilots and ATCOs. All audio files in the ATC dataset are sampled at 8 kHz and 16-bit PCM, and originate from SIN[3] airport in Singapore.

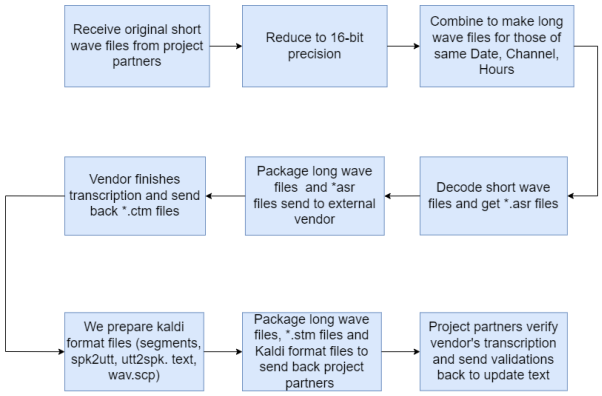

Figure 1. The pipeline of preparing the Singapore Air Traffic Control (S-ATC) dataset

Fig 1 illustrates the steps involved in creating the S-ATC dataset. The audio in the ATC dataset, collected from radio communication technology in the ATC domain from the airport in Singapore. The transcriptions are manually created to ensure the quality.

The data preprocessing pipeline implemented in this study has established the dependability of the dataset for the ASR task. As a result, the high-quality dataset enables experiments to be conducted with assurance, thereby ensuring the precision of the findings discussed in Section III.

A. The Air Traffic Control corpus for the ASR task

TABLE I. The statistics of each ASR corpus

	Training	Testing
# of Utterances	343,279	2,000
# of Hours	440	2.73

TABLE II. The statistics of two small testing sets

	Pilots testing set	ATCOs testing set
# of Utterances	983	1017
# of Hours	1.21	1.52

This dataset is divided into two main subsets: training and testing, as outlined in Table I. The testing set is further subdivided into two distinct categories: one subset exclusively containing audio from pilots, and the other subset consisting solely of audio from ATCOs, as detailed in Table II. We will conduct experiments on these two subsets to evaluate and compare the effectiveness of different ASR models in these distinct recording environments.

B. The Air Traffic Control corpus for the SLU task

For the SLU task, the ATC dataset comprises approximately 30 hours of around 27,000 short utterances and is divided into three subsets: training, validation, and testing. Table III provides the statistics for each subset of the SLU task.

TABLE III. The statistics of each SLU corpus

	Training	Validation	Testing
# of Utterances	20,412	3,205	3,346
# of Hours	25.35	3.34	3.63

TABLE IV. The statistics of various example labels for intent classification and speaker role detection task

	# of Utterances
ATC INSTRUCTION CLIMB	2,407
ATC INSTRUCTION DESCEND	1,397
ATC INSTRUCTION REDUCE SPEED	318
PILOT RESPONSE DESCEND	1,770
PILOT RESPONSE HEADING	657

In accordance with the instructions provided in Section II-A, this subset was extracted from the ATC dataset for the ASR task. It has been manually annotated with intent, slot, and speaker role labels for use in the SLU task.

In the previous dataset within the ATC domain [3], the labels for the named entity recognition (NER) task included Command, Value, and Callsign. In the current dataset, we have introduced additional labels such as Airport/City, Greetings, Waypoint, and Unit for the NER task. Furthermore, the dataset we present is also annotated for intent classification and speaker role detection tasks. We have combined the intent

[1]This corpus is public in: https://catalog.ldc.upenn.edu/ LDC94S14A.
[2]This corpus is public in: https://s.net.vn/c38y.
[3]Singapore Changi Airport

label and speaker role into a single category that encompasses 40 distinct categories, which are determined by common commands from ATCOs or responses from pilots regarding flight paths, aircraft movement directions, and other related information.

For the speaker role detection task, the label "ATC" denotes that the speaker is an air traffic control officer, whereas the label "PILOT" is used for audio originating from pilots.

III. Automatic Speech Recognition

Automatic Speech Recognition (ASR) refers to the process of converting speech signals into corresponding text transcripts. This section outlines the foundational methodology employed in developing and training an ASR model within the ATC domain. The primary objective of this research is to create an ASR engine that can be effectively utilized in the aviation sector. Our methodology encompasses several key approaches: (1) training a hybrid model [8], (2) fine-tuning a pre-trained ASR model along with a Self-supervised Learning (SSL) model, and (3) constructing an End-to-End (E2E) model.

A. System Overview

The hybrid model consists of two components: an Acoustic Model (AM) and a Language Model (LM). The AM utilizes a CNN-TDNNF architecture [9] to predict posterior distributions over the tied Hidden Markov Model (HMM) states corresponding to context-dependent phonemes. These posterior probabilities are then combined with a language model developed using the SRILM toolkit [10].

With its remarkable performance, the sequence-to-sequence attention-based encoder-decoder network (Transformer) has been successfully applied in speech recognition [11] [12]. The encoder employs a conformer architecture that integrates convolutional layers with self-attention mechanisms to effectively capture both local and global contexts. Meanwhile, the bi-transformer decoder processes the encoded speech in a bidirectional manner, improving transcription accuracy by utilizing both past and future contexts.

We also fine-tune the Wav2Vec2-Conformer [13] and the OpenAI Whisper model [4]. These efforts are aimed at improving ASR performance and adaptability within the challenging ATC domain.

B. Experimental Setup

Our experiments are conducted using the ATC dataset for the ASR task, as outlined in Section II-A. The detailed configurations for each model type utilized in the ASR task are presented in the following sections.

1) Hybrid HMM-Deep Learning Model: Our methodology utilizes the Kaldi toolkit [14] to train our HMM-DNN-based ASR systems, its architecture is shown in Fig 2. The input to the 13 layers of TDNN-F comprises 40-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features and 100-dimensional features derived from i-vectors. Each hidden layer has 1280 dimensions, with a linear bottleneck of 128 dimensions.

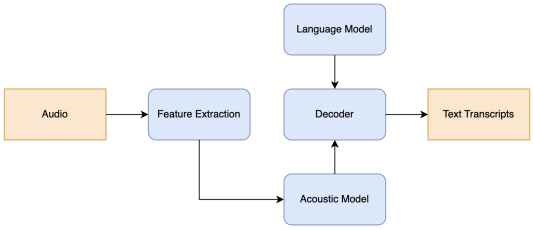

Figure 2. The architecture of the hybrid model

2) End-to-End Model: In our research, we train an End-to-End ASR system, the architecture of which is described in Fig 3, based on the WeNet framework [6] [7]. The input features consist of 83 dimensions, including 80-dimensional log-Mel filterbank coefficients extracted every 10 milliseconds, along with 3-dimensional pitch features. During both training and decoding, we apply global cepstral mean and variance normalization to the feature vectors. The feature encoder comprises 6 stacked multi-head attention blocks built on 2 CNN layers, designed to capture high-level or abstract features and provide more discriminative attributes to support acoustic modeling. Each CNN layer contains 256 filters, each with a kernel size of 3x3 and a stride of 2x2.

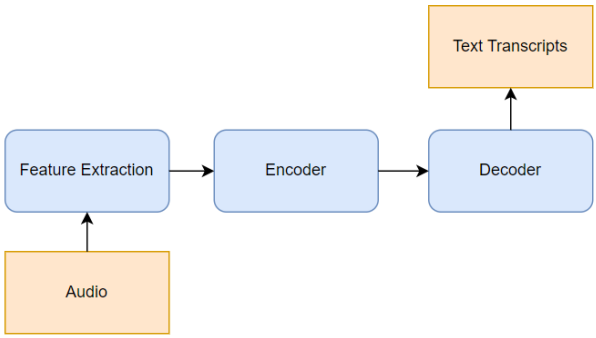

Figure 3. The architecture of the End-to-End model

3) Fine-Tuned Models: We fine-tuned the Wav2Vec2-Conformer[4] model as the encoder and evaluated two decoding strategies: one utilizing a 5-gram language model (LM) for decoding with KenLM, and the other without any language model.

To fine-tune the Whisper model, we first download a robust pre-trained version of Whisper[5] from Hugging Face. This pre-trained model is subsequently fine-tuned for the ASR task. Each fine-tuning process is conducted on an NVIDIA A40 GPU for a total of 20,000 steps.

During the experimentation, we maintain a consistent learning rate of $\gamma = 5 \times 10^{-5}$, utilizing a linear learning rate scheduler.

[4]We use the pre-trained version of facebook/wav2vec2-conformer-rope-large-960h for all the experiments.

[5]We use the pre-trained version of openai/whisper-large-v3 with 1550M parameters for all the experiments.

C. Experimental Results

TABLE V. PERFORMANCE OF ASR SYSTEM (WER)

	Pilots testing set	ATCOs testing set	Entire testing set
Hybrid model	8.07	3.71	9.90
E2E model	8.14	3.10	12.30
Wav2vec2 (w/o LM)	8.24	3.46	11.43
Wav2vec2 (w LM)	7.02	2.77	9.90
Whisper	11.21	3.6	11.6

To evaluate the accuracy performance of our system, we designated a test set consisting of pilot-ATCO speech data from all three Air Traffic Control (ATC) domains. We ensured that the test data was recorded at various times and on different days to achieve a diverse representation.

Performance evaluations are presented in Table V, where we computed the Word Error Rate (WER) across three test sets: the *Pilots testing set*, the *ATCOs testing set*, and the *Entire testing set* for each implementation. WER is calculated as the total number of transcription errors (insertions, substitutions, and deletions) relative to the total number of words in the ground truth.

The conventional hybrid model achieved an average Word Error Rate (WER) of 3.71% for ATCOs and 8.07% for pilots. Notably, the End-to-End (E2E) model based on the Wav2Vec2 architecture with a 5-gram language model (LM) using a KenLM decoder consistently produced better results, achieving an average WER of 2.77% for ATCOs and 7.02% for pilots. This improvement can be attributed to the increased variability in the ATC context, including differences in speaking rates and varying phoneme durations, which present challenges for alignment in hybrid models.

Furthermore, significant differences in performance were observed between the ASR systems for ATCOs and pilots. The Word Error Rate (WER) for pilots was consistently at least 5 percent higher than that for ATCOs, primarily due to the differences in recording environments. While the audio from ATCOs is recorded directly, the audio from pilots is collected via VHF radio, which contributes to the increased error rate.

The performance of the Whisper model [4] was found to be suboptimal compared to other ASR models, primarily due to discrepancies in sampling rates between the model's training data and the ATC dataset used in this study. The ATC dataset, which is essential for evaluating ASR systems in aviation contexts, is sampled at 8 kHz. In contrast, the Whisper model [4] can only be fine-tuned on data sampled at 16 kHz. As a result, the ATC dataset had to be up-sampled from 8 kHz to 16 kHz for training with the Whisper model. This up-sampling process likely introduced artifacts and distortions, thereby reducing the transcription accuracy of the Whisper model [4].

IV. SPEAKER ROLE DETECTION

The Speaker Role Detection (SRD) task is crucial for analyzing the journey of a plane within the ATC domain. This task is typically viewed as either a text classification task or an audio classification task, relying solely on audio or text data. In [3], the applicability of ground truth texts in practical scenarios is limited, and transcribed texts may contain errors that impact the model's accuracy. Therefore, we employ both audio and text concurrently to complement one another and improve the model's overall accuracy. This section presents an overview of the system and the model designed for the SRD task in the ATC domain.

For all the experiments in this section, we utilized the ATC corpus for the SLU task with the aim of detecting speaker roles, as mentioned in Section II-B. To construct this dataset, a long audio segment containing the conversation between the pilot and Air Traffic Control Officers (ATCOs) had to be segmented into smaller audio clips, and each clip was labeled to identify the speaker. In real-world situations, segmenting conversations can be challenging due to the spontaneous nature of discussions, which may involve overlapping voices. However, in the ATC settings, similar to voice communication in Air Traffic Control (ATC) as mentioned in [15], communication over VHF radio simplifies this task. The requirement for the speaker to press a button to talk creates natural pauses in the conversation, resulting in blank spaces in the audio spectrogram and waveform, as illustrated in Fig. 4.

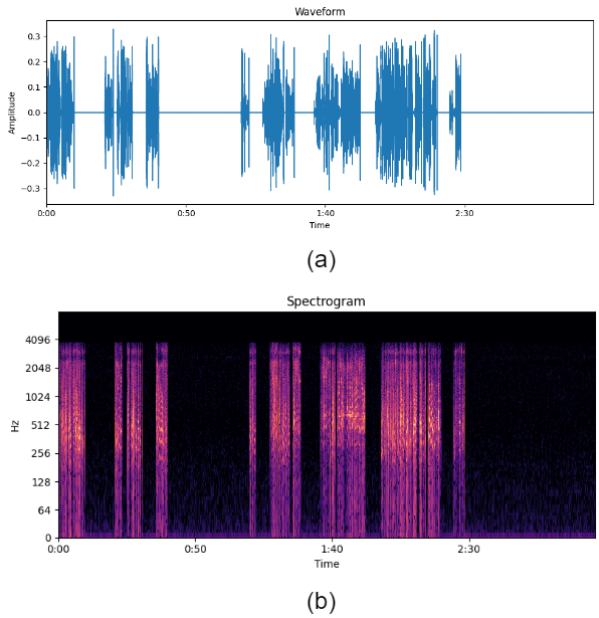

Figure 4. Analyze an example audio: (a) Waveform, (b) Spectrogram.

Therefore, in this scenario, a straightforward Voice Activity Detection (VAD) model can be employed to effectively segment the conversation into individual utterances, providing suitable input for the ASR models that will be discussed later. This approach also streamlines the labeling process for the SRD task. While this paper will not delve extensively into the VAD model, it will focus on enhancing a model for the SRD task. Once the conversation is segmented into multiple small audio clips, labels for the SRD task are manually assigned to ensure the quality of the labels.

A. System overview

To ensure optimal results for the Speaker Role Detection (SRD) task, the quality of both the input text and audio must be of the highest standard. The architecture of the entire system is illustrated in Fig. 5. For the ASR model, we utilized the Wav2Vec2-Conformer model [13], which was fine-tuned with about 440 hours of the ATC dataset, achieving a 7% Word Error Rate (WER) on the test set. Following this, the SRD model was trained using both audio and transcribed text. The detailed architecture of the SRD model will be presented in the subsequent sections.

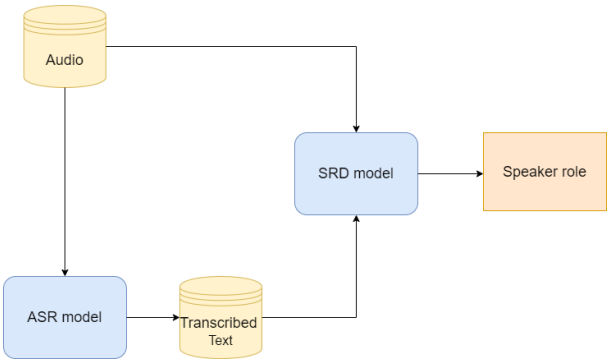

Figure 5. The architecture of the entire model

B. Architecture

1) Text-based Speaker Role Detection: Text classification involves assigning a label or class to a sequence of words [16]. One of the most widely recognized tasks in text classification is sentiment analysis, where the labels are positive, negative, or neutral. Many state-of-the-art systems for this task are based on the well-known Transformer [17] architecture. In this paper, we also utilize the Transformer [17] architecture to classify input text into two labels: "PILOT" and "ATC".

Similar to [3], we fine-tune a pre-trained model for sequence classification, specifically the BERT[6] model [18], using the ATC dataset. However, unlike [3], our experiments incorporate both ground truth texts and transcribed texts from an ASR model. The results from these two input types will be compared to evaluate whether a Natural Language Understanding (NLU) system can be practically applied, given that ground truth texts are often unavailable in real-world scenarios.

2) Audio-based Speaker Role Detection: We approach the Speaker Role Detection (SRD) task as an audio classification problem, in addition to using text. Within the ATC industry, there are notable differences in the communication contexts between shore-based controllers and pilots. As such, these two categories of audio have unique properties that deep learning models can use to learn from. However, outside influences like the sound of the air, the weather, and machinery noise can negatively impact audio quality, which in turn affects the model's ability to learn.

[6]We use the pre-trained version of bert-base-uncased with 110 million parameters for all the experiments.

We fine-tune Audio Spectrogram Transformer (AST) model [19], which is the state-of-the-art model in audio classification task using Google SpeechCommand dataset [20], with our corpora mentioned in Section II. The Vision-Transformer model [21] serves as the foundation for this model, which was pre-trained using the ImageNet [22] and AudioSet [23] datasets.

3) Audio-Text-based Speaker Role Detection: Owing to the drawbacks and shortcomings of relying solely on texts or audio for the SRD assignment, we created a system that employs an ensemble algorithm to integrate the previously described text- and audio-based models. Fig. 6 describes the architecture of the combination between the AST model [19] and the fine-tuned BERT model [18].

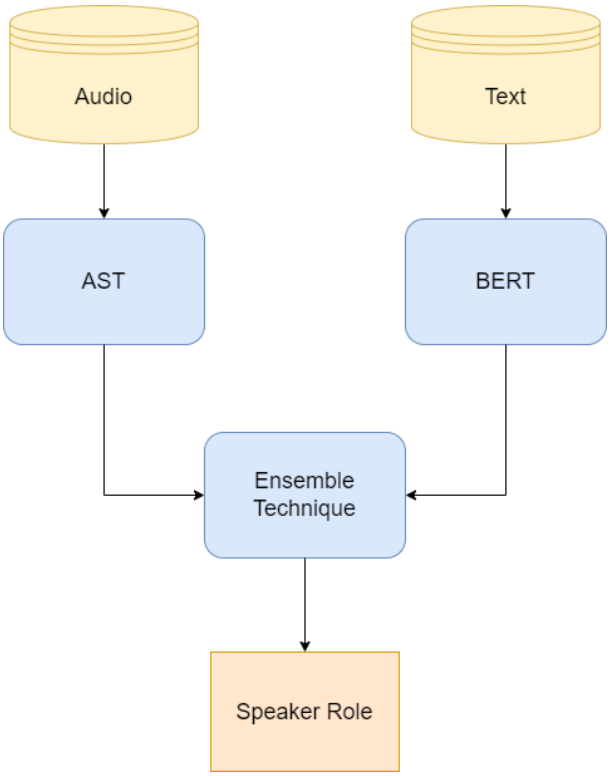

Figure 6. The architecture of the combined model

Using audio and textual data from the ATC dataset, respectively, the AST [19] and BERT [18] models were individually fine-tuned. Then, using ensemble techniques—more especially, the soft voting technique—the predictions from these two models are integrated. We take the output probabilities from each model and average them for each label. The final forecast is given back to the label with the highest average probability.

C. Experimental Results

The experimental findings of the models discussed above are shown in Table VI, arranged according to F1 score. It is evident that the combined model's outcomes outperform those of the two separate models. The capacity of the model to learn features from both texts and audio is responsible for

this improvement. These findings are similar to those in [3]; however, because this method does not rely on ground truth texts, it may be used for the first time in real-world ATC circumstances.

TABLE VI. THE EXPERIMENTAL SRD RESULTS OF EACH MODEL (F1 SCORE).

	Pilot	ATC	Average
AST fine-tuned	0.80	0.84	0.82
BERT fine-tuned (w ground truth text)	0.89	0.90	0.90
BERT fine-tuned (w/o ground truth text)	0.79	0.80	0.80
Combined model	0.84	0.85	0.85

V. SPOKEN LANGUAGE UNDERSTANDING

In addition to accurately transcribing dialogues between pilots and ATCOs, it is essential to extract critical information for practical applications. This includes tasks such as entity highlighting (commonly referred to as intent classification and slot-filling) and speaker role detection. To address these requirements, we have explored various methods for constructing a Spoken Language Understanding (SLU) system within the air traffic control domain. We compare the performance of extracting information from text (Natural Language Understanding) and directly from audio (Spoken Language Understanding) to identify the most effective approach for this task. In addition to the named entity recognition task, which is similar to previous work [3], we conduct intent classification and speaker role detection simultaneously, as our dataset contains shared labels for these two tasks, as outlined in Section II-B.

A. Natural Language Understanding

In line with the method employed in [3], we fine-tune a pre-trained BERT model [18] for ATC tasks. However, unlike [3], our experiments utilize not only ground truth text but also transcribed text obtained from an ASR model. The results from these two input types will be compared to assess whether a Natural Language Understanding (NLU) system can be effectively applied, considering that ground truth text is not available in real-world scenarios.

TABLE VII. EXPERIMENTAL RESULTS BETWEEN INPUT GROUND TRUTH TEXT AND TRANSCRIBED TEXT

Input	Slot F1 score	Intent Accuracy
Ground truth text	0.94	0.90
Transcribed text	0.75	0.87

Table VII presents the experimental results of slot-filling and intent classification comparing ground truth text with transcribed text. In this case, we chose the output text from the Wav2Vec 2.0 model [5] because the results are detailed in Section III. With a high-quality transcription achieving a 14% Word Error Rate (WER), the output of the Wav2Vec 2.0 model [5] retains much of its content compared to the ground truth text. This retention of content enables the BERT model [18] to perform effectively in the intent detection task. However, certain errors in the ASR system's output still impact the results of named entity recognition, which necessitates precise accuracy for every word in the sentence, as illustrated in Table VII. To enhance the effectiveness of the BERT model [18] in the named entity recognition task, we propose an alternative approach for constructing the SLU system, which will be discussed in the following section.

B. Spoken Language Understanding

To address the inefficient performance of the NLU system caused by the poor quality of transcribed texts, we experimented with an End-to-End (E2E) SLU system. This approach omits the transcription stage to minimize errors and improve overall accuracy.

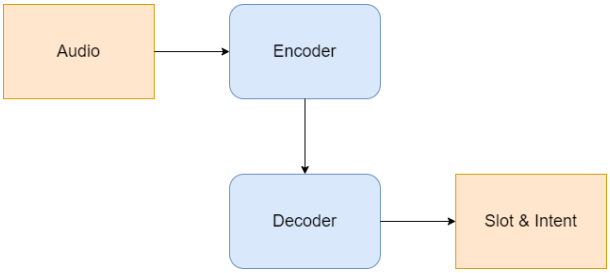

Figure 7. The architecture of the E2E SLU system

This SLU system extracts high-level information directly from the audio. As illustrated in Figure 7, the system integrates a Self-supervised Learning (SSL) model as the feature extractor, an Encoder-Decoder architecture, and a pre-trained Language Model (LM) to comprehend the audio. For our experiments, we utilized the Wav2Vec 2.0 model [5] as the SSL model, integrated with the Conformer [24] - Transformer [17] architecture, and the BERT model [18] as the LM model. The aforementioned system is implemented using the ESPNET framework [25].

Initially, we trained the Wav2Vec 2.0 model [5] with the ATC dataset for the ASR task over 70 epochs. Next, we trained the entire SLU system for 100 epochs using the same dataset, while keeping the parameters of the pre-trained Wav2Vec 2.0 [5] and BERT model [18] frozen.

TABLE VIII. EXPERIMENTAL RESULTS OF THE E2E SLU SYSTEM

Input	Slot F1 score	Intent Accuracy
Audio	0.96	0.89

As demonstrated in Table VIII, the results for both the named entity recognition and intent detection tasks have improved compared to the NLU system described in Section V-A. This improvement arises from the elimination of the intermediary step of converting speech to text, thereby minimizing errors introduced by models during these intermediate transformations. This system can be practically applied, differing from previous research [3] [26] [27] methodologies

by not using ground truth text and effectively addressing errors in transcribed text that affect sentence meaning.

VI. Conclusion

This paper introduces a dataset and the application of Automatic Speech Recognition (ASR) systems and Spoken Language Understanding (SLU) systems in the specific domain of Air Traffic Control (ATC). ATC is considered a challenging and low-resource domain for applying ASR systems. However, we provide several contributions aimed at addressing these challenges and exploring new methods to reduce the workload of pilots and ATCOs by leveraging recent advances in ASR technology.

Various experiments were conducted to demonstrate how each of the applied methodologies influences the performance of ASR and SLU systems. We observed significant improvements in recognizing named entities and detecting the intent of speech by employing an End-to-End SLU system, as opposed to a combination of ASR and NLP models. This system is capable of detecting not only callsigns, commands, and values, but also units, waypoints, greetings, and the names of cities or airports from both speech and textual input. This capability holds particular significance for the ATC community, as this high-level information can assist pilots in reducing their overall workload. In addition, this paper also presents research on the application of deep learning models to address the Speaker Role Detection (SRD) task in the ATC domain. Detecting speakers in the ATC domain aids professionals in the field in analyzing issues and incidents that occur during flight operations.

Finally, we believe that the experiments conducted in this research will significantly benefit deeper applications and other specialized fields in the future, such as the application of Large Language Models (LLMs) [28] or Graph Neural Networks (GNNs) [29] to End-to-End SLU systems to enhance their efficiency. We hope that this research can be applied in practice within the field of air traffic management, contributing to improved service quality and enhanced safety in the aviation industry.

VII. Acknowledgement

This research / project is supported by the National Research Foundation, Singapore, and the Civil Aviation Authority of Singapore, under its “Automatic Speech Recognition and Understanding for improvements in ATM operations” (Award No.:ATP_ASRU_I2R). Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Civil Aviation Authority of Singapore.

References

[1] Yu Song Meng, Yee Hui Lee, and Boon Ng, “Further study of rainfall effect on vhf forested radio-wave propagation with four-layered model,” *Progress In Electromagnetics Research*, vol. 99, pp. 149–161, 01 2009.

[2] Yi Lin, Dongyue Guo, Jianwei Zhang, Zhengmao Chen, and Bo Yang, “A unified framework for multilingual speech recognition in air traffic control systems,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3608–3620, 2021.

[3] Juan Pablo Zuluaga, Karel Veselý, Igor Szöke, Petr Motlícek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolcárek, Allan Tart, and Jan Honza ernocký, “Atco2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications,” *ArXiv*, vol. abs/2211.04054, 2022.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” 2020.

[6] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” 2021.

[7] Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu, “Wenet 2.0: More productive end-to-end speech recognition toolkit,” 2022.

[8] Artur M. Schweidtmann, Dongda Zhang, and Moritz von Stosch, “A review and perspective on hybrid modeling methodologies,” *Digital Chemical Engineering*, vol. 10, pp. 100136, 2024.

[9] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. Interspeech 2015*, 2015, pp. 3214–3218.

[10] Andreas Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 901–904.

[11] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang, “A comparative study on transformer vs rnn in speech applications,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Dec. 2019, IEEE.

[12] Linhao Dong, Shuang Xu, and Bo Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[13] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” 2022.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel, “The kaldi speech recognition toolkit,” *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

[15] Michael Neffe, Tuan Pham, Horst Hering, and Gernot Kubin, *Speaker Segmentation for Air Traffic Control*, pp. 177–191, 09 2007.

[16] Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang, “A survey on recent advances in sequence labeling from deep learning models,” 2020.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.

[19] Yuan Gong, Yu-An Chung, and James Glass, “Ast: Audio spectrogram transformer,” 2021.

[20] Pete Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” 2018.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009*

IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[24] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[25] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "Espnet: End-to-end speech processing toolkit," 2018.

[26] Hartmut Helmke, Michael Slotty, Michael Poiger, Damian Herrer, Oliver Ohneiser, Nathan Vink, Aneta Cerna, Petri Hartikainen, Billy Josefsson, David Langr, Raquel Lasheras, Gabriela Marin, Odd Mevatne, Sylvain Moos, Mats Nilsson, and Mario Boyero, "Ontology for transcription of atc speech commands of sesar 2020 solution pj.16-04," 09 2018, pp. 1–10.

[27] Juan Zuluaga-Gomez, Seyyed Saeed Sarfjoo, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlicek, Karel Ondrej, Oliver Ohneiser, and Hartmut Helmke, "Bertraffic: Bert-based joint speaker role and speaker change detection for air traffic control communications," 2022.

[28] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao, "Large language models: A survey," 2024.

[29] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.