Air Traffic Control Speech Recognition

Shuo Chen, Hunter Kopald, Weiye Ma, Robert Tarakan, Yuan-Jun Wei

The MITRE Corporation

chen@mitre.org, hkopald@mitre.org, wma@mitre.org, rtarakan@mitre.org, ywei@mitre.org

Abstract

To help the Federal Aviation Administration (FAA) better understand air traffic operations in the National Airspace System (NAS), The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) has been developing and using voice data processing capabilities to analyze air traffic control (ATC) radio/voice communications. These analyses provide new insights to help the FAA make safety and efficiency improvements. This paper describes capabilities we have developed and use to analyze NAS ATC operations, focusing on the influence of the ATC domain on our technical approaches to voice processing.

Index Terms: speech recognition, air traffic control, aviation

1. Introduction

Over the last several years, to help the Federal Aviation Administration (FAA) better understand air traffic operations in the National Airspace System (NAS), The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) has been developing and using voice data processing capabilities to analyze air traffic control radio/voice communications. These analyses provide new insights such as the exact clearance issued by a controller, which help inform operational understanding, such as the use of a certain procedure or the reason for or occurrence of a safety event. Air Traffic Control (ATC) voice communications have a unique set of characteristics that are relevant to applying natural language processing technologies such as automatic speech recognition.

This paper describes some of the voice/speech capabilities we have developed and use together as a system specifically to analyze NAS ATC operations. We aim to provide enough detail for other researchers applying voice/speech capabilities to ATC voice communications to consider how similar techniques might work for them. We also aim to provide enough detail for other speech researchers to consider how their techniques or ideas might be a good fit for the ATC voice domain. Finally, we will describe some remaining technical challenges to start a dialog on how to address these challenges and to provide specific use cases for future research across the speech processing community.

2. Background

Voice communication by radio remains the primary method by which air traffic controllers communicate with pilots in the airspace and on the airport surface. These ATC voice communications are critical to a complete understanding of the NAS through post-operations analysis and, if processed quickly and accurately, can be used to improve operations in real time, e.g., to detect safety risk situations and notify the controller [1] [2]. Helmke et al. have demonstrated the use of automatic speech recognition to improve system efficiency and decrease controller workload [3] [4].

FAA and MITRE CAASD have been researching and using automatic speech recognition (ASR) and related capabilities for several purposes within the ATC domain, across both live operations and simulation and training. The research and capabilities described in this paper are focused on recordings of live operational voice communications – i.e., between controllers and pilots in actual operations in the NAS.

The characteristics of the FAA's ATC domain influence the technical approaches we describe in this paper. Note that while the high-level domain of ATC voice is the same, there are differences between Air Navigation Service Providers (ANSP) that influence what research and applications are most valuable and the techniques that can be employed to process speech data effectively. We focus in this paper on characteristics specific to the United States NAS.

2.1. Acoustic Characteristics

The ATC speech signal has domain-specific characteristics. The audio of the U/VHF AM radio communication is noisy, with limited frequency bandwidth. The FAA systems that record the audio can mix audio from multiple sources (e.g., pilot, controller, intercom, and other audio that is being monitored by a controller) and discard the push-to-talk information delimiting individual transmissions. The audio we receive from the FAA is provided as 16 bit/8kHz mp4 format, after multiple lossy audio data compression/decompression cycles. In addition to the audio quality, we also encounter pronunciation variations, with challenges such as fast speech, coarticulation, and incomplete articulation. The concise style of ATC also results in some utterances less than two seconds in duration.

2.2. Language Characteristics

ATC voice communications have a particular phraseology and lexicon that have influenced our approaches for speech processing and understanding. ATC vocabulary is limited compared to general English speech (20,000:134,000), with numbers spoken most frequently (one third of all ATC speech consists of numbers.). To a large extent, FAA ATC vocabulary and phraseology are standardized, although FAA rules allow free form speech when needed [5], and controllers and pilots sometimes use local colloquialisms or otherwise vary from standard phraseology. One notable characteristic of FAA phraseology is that numbers are to be spoken as individual-digits in some circumstances and in group-form in other cases, such as commercial aircraft callsigns. Controllers and pilots also typically speak much faster than general

English speech and with minimal silence between speakers. ATC has domain specific terminology, such as over six thousand callsigns from major airlines, hundreds of facility names, tens of thousands of waypoints, and thousands of arrival, approach, and departure procedures in the NAS.

Most ATC voice transmissions contain a callsign to identify the flight and a command, advisory, or query of some kind issued by the controller to the pilot, or readback, request, report, or query by the pilot to the controller. However, each facility typically issues a subset of all types of communications—for example, controllers managing traffic on the airport surface issue a different set of commands than controllers managing flights at cruising altitudes. Facilities also have a specific set of resources (e.g., runways at an airport, or routes and waypoints in airspace) they use, which further constraints the words and phrases likely to be spoken by controllers and pilots.

2.3. Analysis Needs

Our approaches to recognizing US ATC voice communications have also been informed by our access to a large quantity and variety of FAA voice data. MITRE CAASD has access to FAA voice recordings across 130 FAA ATC facilities, covering the majority of ATC voice transmissions in the NAS. This variety of data has presented some challenges that we had not encountered while developing ASR for ATC speech for human-in-the-loop simulation, controller training, and real-time safety applications. For example, it is not practical for us to perform the same degree of manual channel/facility/sector/position-specific customization that we were able to do with single-facility audio.

A second challenge is that unlike our previous applications, the large-scale processing capability is intended to support a larger range of analysis use cases, many that we cannot anticipate in advance. Some users measure routine situations while others wish to identify the rare events. Identifying speech associated with safety risks can be especially tricky because these risky situations are rare in modern ATC. Given that so much of ASR is based on statistical models that are intentionally biased toward recognizing probable speech, recognizing this less common speech can be particularly challenging.

To use ATC voice information for analysis or real-time applications, we need to be able to extract relevant semantic content from the voice communications. To account for phraseology variations, shorthand, and ASR error, we have developed a semantic parsing capability that uses non-speech context information to resolve ambiguities.

Section 3 describes our ATC domain-specific techniques for voice data processing.

3. ATC Domain-Specific Techniques for Voice Data Processing

This section briefly describes our ATC speech corpus and then describes the three main components we need for voice data processing: segmentation and speaker role identification, automatic speech recognition, and semantic parsing.

3.1. ATC Speech Corpus

Through several years of FAA research, we have accumulated a human-transcribed ATC speech corpus of over one thousand hours silence-reduced audio, with data from over 70 FAA ATC facilities. This corpus enables us to train speech models purely from NAS ATC speech and provides enough variety of NAS operations that we can use a single model to provide transcripts for a wide variety of analyses. However, facility- or region-specific customization still provides additional accuracy improvements. All audio transcriptions and ATC vocabularies for the corpus have been subjected to a rigorous quality control process to ensure ground truth correctness. Note that this corpus does not contain any data from non-FAA facilities, such as the corpora developed within Europe [6] [7].

3.2. Segmentation and Speaker Role Identification (SRID)

3.2.1. Automatic Speech Recognition approach for SRID

The first step of audio processing is to separate controller speech from pilot speech. We call this step Speaker-role Identification (SRID). We initially tested the traditional approach of using bandwidth separation. While it worked well when there is sufficient silence between transmissions, the approach usually failed where there is little to no pause between speaker turns, which is common in ATC voice communication. The existence of controller-to-controller speech also complicated the task of separating controller speech from pilot speech. Finally, for a particular channel and time period, we can customize the segmentation system to perform reasonably well. However, this level of customization is not practical on a large scale.

Thus, we recharacterized SRID as a speech-to-text task with a two-word vocabulary: controller and pilot. The resulting ASR solution consisted of an acoustic model (AM) and a statistical language model (SLM). To train the AM and SLM, we mapped our word corpus into a speaker role based corpus. For example, the text *american twenty two turn left* from a controller transmission was mapped to a string of controller role labels “CCCCC”, with one label per word. Similarly, the text *turn left american twenty two* from a pilot transmission was mapped to a string of pilot role labels “PPPPP”.

Transcribing a sentence with multiple labels allows us to use any ASR training procedure to train SRID AM and SLM as with a regular ASR. Use of ASR AM training procedures allows us to jointly train the controller and pilot acoustic model together with the silence model and noise model. SRID SLM trigrams are created from role label transcriptions. We use the same ASR decoder as in our text-to-speech component to convert audio into a sequence of speaker role labels with time marks. For our example, the output for controller and pilot mix audio looks like “CCCCPPPPPP”. During ASR decoding, the decoder jointly carries out classification at a frame level and clustering at an utterance level. Output sequence is optimal in terms of number of speaker role words and the time mark of each word. The segmentation is created by merging same role labels and their time marks. Our approach does not assume a max number of segments. When an input utterance only contains one speaker, we expect output string consist of only label “C” or “P”. We use single speaker utterances to measure frame level classification error rate (FER), which is the number of frames misclassified divided by the total number of frames. Note that this method alone does not separate consecutive controller-controller or pilot-pilot transmissions.

3.2.2. Testing Results: Multi facility SRID system

We randomly partitioned our corpus into training and testing. For facilities that do not have enough data to build reasonable test set, we put all data from those facilities into the training partition. The training set consists of 780 hours of silence reduced speech from 81 facilities and the test set consists of 53 hours of silence reduced speech from 51 of those 81 facilities.

The SRID AM used in this experiment is a 6-hidden layer Time Delayed Neural Network (TDNN) with 256 nodes with MFCC (Mel Frequency Cepstral Coefficient) of 40 basic parameters. We used the same training procedure that we used for our ASR AM described in the next section. The decoding SLM is a trigram with speaker role labels.

The result is a FER of 3.3% with a real-time factor that is only a fraction of our ASR component. These results suggest that with this approach we can use a single SRID model to segment all audio channels. Going forward, we plan to use un-supervised training and to explore how language can be used to improve SRID and segmentation.

3.3. Automatic Speech Recognition

3.3.1. Approach

We have created an ATC domain-specific Kaldi recipe exclusively using ATC speech [8] [9]. We have also created a tool to enable us to easily add additional unseen vocabulary and facilities to resolve ATC-specific issues mentioned above.

The lexicon for the speech engine is an extension of the Carnegie Mellon University (CMU) lexicon [10] where we have established word to phoneme models using a grapheme-to-phoneme (G2P) tool. If a given word does not exist in the CMU lexicon, we run G2P to generate its pronunciation and check the pronunciation with ATC subject matter experts. Our core lexicon contains about 20,000 active words and phrases.

Using the Stanford Research Institute Language Modeling (SRILM) tool [11], we implemented two 4-gram language models with two different sizes: the smaller language model is for normal decoding and the bigger one is for rescoring to increase speech recognition accuracy. Controller-to-controller conversations have been incorporated into the training text with lower weight to provide interpolation as our ATC transcriptions have limited language coverage. We developed the post language modeling adaption capability by retraining language models with additional facilities and vocabularies to extend the baseline language models without training new acoustic models. This is especially useful for facilities and waypoints where transcribed audio is not available.

We perform AM training after the lexicon and language models are in place. The raw audio recordings are converted to 8kHz, 16 bits, linear PCM format for MFCC feature extraction with 40 parameters. Mono-phone to tri-phone Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) training followed by alignment to generate the phonetic labeling data for supervised DNN training. We incorporated the Kaldi TDNN chain Lattice-Free Maximum Mutual Information (LF-MMI) structure to train the DNN-HMM hybrid acoustic models [12]. The training process includes steps of volume perturbation to reduce the audio amplitude influence, with one third frame sub-sampling for optimizing GPU calculation and improving real time factor. The AM training data includes over 700,000 transcribed utterances/segments and the testing data contains over 70,000 transcribed utterances/segments.

3.3.2. Testing Results

We have achieved 5% Word Error Rate (WER) for controller speech and 13% WER for pilot speech, and over 97% in accuracy for key ATC vocabulary, such as numbers and airline callsigns, on controller speech. Our unpublished observations are that the Kaldi speech recognition engine trained exclusively with ATC voice data yielded at least a 50% relative improvement over COTS telephony speech recognizers. Yet additional room for improvement of recognition accuracy remains, especially for pilot speech, with additional training data, facility- or region-specific language models, and exploring other state-of-the-art DNN technology, such as transfer-learning and/or end-to-end speech recognition.

3.4. Semantic Processing Techniques

3.4.1. Multi-Hypothesis Semantic Extraction with Context

We use a rules-based semantic parse algorithm for extracting ATC concepts such as controller-issued commands and advisories, pilot-spoken readbacks, and aircraft callsigns. We have improved on our previous iteration described in [1] by combining the parsing of callsigns with the parsing of other ATC content (which improves alphanumeric phrase disambiguation) and implementing a branching parse that can generate multiple semantic interpretations for the same transcript. Having multiple semantic interpretations is helpful when speech recognition or other errors cause ambiguity.

To differentiate interpretations and highlight more likely ones, we added a ranking function designed on the premise that almost all words spoken over the radio should contribute meaningfully to an ATC semantic concept. Thus, the function favors interpretations that maximize the number of words used by concepts; it intentionally does not attempt to maximize the number of final concepts parsed.

The current parser extracts over 60 ATC concepts, and we continue to expand the number of concepts extracted and fine tune the existing concept coverage on a case-by-case basis as identified by end user analysts.

We use site adaptation data to automatically customize parsing rules for each ATC facility. For example, the parser is dynamically configured with the appropriate runway identifiers for each airport. This configuration is performed automatically and is easily scalable because the site adaptation data is available in a structured, digital form, and the vocabulary consists of words found in common English speech, often with similar or phonetic pronunciations.

3.4.2. Situational Context Review of Aircraft Identifiers

Aircraft identifiers (ACID) remain one of the most important semantic concepts in a controller or pilot transmission because it allows the transmission to be linked (or fused) to flight information from other aviation data sources, such as flight plans and surveillance data.

To improve accuracy on ACIDs, we apply post-parsing processing to mitigate errors that might have been introduced by diarization, speech recognition, or even the speaker. Specifically, we compare all parsed ACID candidates, and any unused alphanumeric sequences in the transmission transcript

that could be ACID candidates, against an automatically generated list of aircraft likely to be present in the airspace at the time of the radio transmission. While an exact match to a context list ACID is ideal, the review algorithm will also accept partial matches with a sufficiently high alignment percentage and map the parsed ACID to the partially matched ACID in the context list. Following this review, the algorithm assigns the final ACID an integer rank between one and nine that categorizes the amount of effort required to map the parsed ACID to the final reviewed ACID. For example, if the parser successfully extracts *DAL1579* from the transcript *delta fifteen seventy nine normal speed* and the review algorithm finds *DAL1579* in the context list, the final ACID *DAL1579* is assigned rank one. Alternately, if the parsed ACID does not perfectly match any ACID in the context list (e.g., because of a word error in the transcript), the review algorithm attempts to find the closest match and assigns a corresponding rank.

This ranking system has been particularly useful when determining which ACIDs to use when fusing with other surveillance data sources. In general, rank one ACIDs have near perfect accuracy and make up between 55 and 65 percent of our extracted ACIDs. The remaining 45 to 35 percent is made up of the remaining ranks, with decreasing accuracy as the rank values increase.

There is still room for improvement in our aircraft identifier extraction algorithm, particularly in the area of general aviation and military aircraft, which follow a different naming convention than for commercial aviation.

4. Challenges and Future Research

Enhancements to improve accuracy and add new features to our voice data processing capability make the voice data useful for new aviation applications and more useful for existing applications. This section describes two challenges we have yet to fully address for our ATC voice processing needs. We aim to provide enough detail for readers to offer techniques or ideas for addressing these challenges.

4.1. Controller-Pilot Dialogue

Our current system treats each transmission in isolation, but in many cases information from preceding transmissions is needed to disambiguate or otherwise make sense of a given transmission. Many ATC communications consist of a simple controller-instruction followed by a fully articulated pilot-readback, which are easy to string together to create a *story* for a flight, but some pilot requests, pilot readbacks, and controller queries can involve multiple back-and-forth transmissions and abbreviated speech, with each transmission building on the context from earlier communications.

Most likely the best solution would combine multiple approaches. One approach is to use language across transmissions to stitch communications together. For example:

controller: *acey two twenty turn left heading two four zero*

pilot: *two forty two twenty*

In this exchange, the pilot transmission makes sense if you know the callsign and instruction in the previous transmission; otherwise, the pilot transmission is ambiguous.

A rules-based approach is challenging, particularly due to shorthand phraseology and speech recognition error. We have tested incorporating the timing between the transmissions—e.g., if a pilot transmission follows a controller transmission within two seconds, associate it with the callsign detected in the controller transmission. This approach works well for most readback situations, but we have not yet implemented it. Machine learning seems to be a good fit for the language approach, but we have not yet succeeded in this area.

Another approach to this problem is for the audio processing system to identify the voice(s) associated with each flight. A pilot will typically have several transmissions (at least four and up to a two or three dozen) to and from a given controller before being transferred to the next controller's radio frequency. If audio processing could recognize the same voice across multiple transmissions, that information could be used to associate transmissions with a flight, even if the identifying callsign is never spoken. One challenge is that the recording system applies automatic gain control to all pilot audio, making it more difficult to distinguish between different aircraft. Another challenge is that the system would need to establish a speaker model dynamically with very few audio samples. Finally, flight deck voice communications are usually handled by one pilot or the other, but sometimes both pilots speak over a short period of time.

4.2. Parsing, Semantic Tagging and Categorization of Additional ATC Concepts

While most analyses can use the derived data that the rules-based parser currently provides, some of the most interesting new applications involve adding to parsing, semantic tagging, and categorization capabilities. We have had good success using Deep Neural Networks (DNN) for classification of speech utterances that often do not conform to the formal ATC phraseology or where there is no documented formal ATC phraseology. We plan to continue exploring the use of DNN to improve the voice data processing capability. While the rules-based approach provides pretty good accuracy on reasonably well-formed speech, it cannot adapt to significant phraseology variations or ASR errors.

A machine learning approach should be more robust to text variations. We have had success using DNNs to classify transcripts into a small number of simple categories but have not yet demonstrated a machine learning-based semantic parsing capability that extracts most relevant ATC concepts. We have used results from our rules-based parser to supplement a manually annotated training set for a DNN. Our tests have demonstrated DNN accuracy similar to, but not better than, the accuracy of the rules-based parser.

Further, the ultimate parsing capability could identify semantic concepts that have not been pre-specified.

5. Summary

We have helped the FAA better understand the NAS by analyzing NAS voice communications using our ATC voice data processing capabilities built with an ATC speech corpus and domain-specific parsing techniques. The processing capabilities are designed to accommodate the unique characteristic of ATC voice communications and to operate on a large scale, handling the majority of voice communications in the NAS. We continue to mature and enhance the processing capabilities to enable more accurate voice analysis to make analyses easier and more insightful.

6. Acknowledgements

The authors would like to thank the FAA for sponsoring much of the work described in this paper.

7. References

[1] S. Chen, H. Kopald, R. Tarakan, G. Anand and K. Meyer, "Characterizing National Airspace System Operations Using Automated Voice Data Processing," in *Thirteenth USA/Europe Air Traffic Management Research and Development Seminar*, Vienna, Austria, 2019.

[2] S. Chen, H. Kopald, D. R. S. Chong, D. Y.-J. Wei and Z. Levonian, "Read Back Error Detection using Automatic Speech Recognition," in *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*, Seattle, Washington, 2017.

[3] H. Helmke, O. Ohneiser, J. Buxbam and C. Kern, "Increasing ATM Efficiency with Assistant Based Speech Recognition," in *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*, Seattle, Washington, 2017.

[4] H. Helmke, O. Ohneiser, T. Mühlhausen and M. Wies, "Reducing Controller Workload with Automatic Speech Recognition," in *35th Digital Avionics Systems Conference*, Sacramento, California, 2016.

[5] Federal Aviation Administration, *Air Traffic Control 7110.65Y,* Washington, DC: Air Traffic Organization Policy, 2019.

[6] E. Delpech, M. Laignelet, C. Pimm, M. Trzos, A. Arnold and D. Pronto, "A Real-life French-accented Corpus of Air Traffic Control Communications," in *Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

[7] K. Hofbauer, S. Petrik and H. Hering, "The ATCOSIM Corpus of Non-Prompted Clear Air Traffic Control Speech," in *Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[8] D. Povey, "Kaldi," [Online]. Available: http://www.kaldi-asr.org/. [Accessed 10 March 2021].

[9] H. Hadian, H. Sameti, D. Povey and S. Khudanpur, "End-to-end Speech Recognition Using Lattice-free MMI," in *Nineteenth Annual Conference of the International Speech Communication Association*, Hyderabad, India, 2018.

[10] Carnegie Mellon University, "CMU Lexicon," [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict. [Accessed 10 March 2021].

[11] Stanford Research Institute, "SRILM Toolkit," [Online]. Available: http://www.speech.sri.com/projects/srilm/. [Accessed 10 March 2021].

[12] D. Povey, "'Chain' Models," [Online]. Available: http://kaldi-asr.org/doc/chain.html. [Accessed 10 March 2021].

[13] A. Srinivasamurthy, P. Motlicek, M. Singh, Y. Oualil, M. Kleinert, H. Ehr and H. Helmke, "Iterative Learning of Speech Recognition Models for Air Traffic Control," in *Nineteenth Annual Conference of the International Speech Communication Association*, Hyderabad, India, 2018.