Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015)

The Closed Runway Operation Prevention Device: Applying Automatic Speech Recognition Technology for Aviation Safety

Shuo Chen and Hunter Kopald
The MITRE Corporation
Center for Advanced Aviation System Development (CAASD)
McLean, Virginia
Contact email: chen@mitre.org

***Abstract*— The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) recently completed a field demonstration at Washington/Dulles International Airport (KIAD) of a proof-of-concept system called the Closed Runway Operation Prevention Device (CROPD) to validate the operational feasibility of employing an emerging technology—automatic speech recognition—in the Air Traffic Control (ATC) domain during live operations for safety improvement. Completed on behalf of the Federal Aviation Administration (FAA), the demonstration and subsequent analysis assessed the accuracy of speech recognition in detecting clearances to closed runways in Local Controller transmissions and the overall performance of an alerting mechanism dependent on speech recognition. The success of applying speech recognition technology in a live ATC environment depends on overcoming domain-specific challenges, such as rapid and/or slurred speech, poor field audio quality, and language ambiguity (e.g., the number sequence one-two can appear in a call sign, speed, wind advisory, or runway identifier), and stringent requirements on system accuracy. To address these challenges, MITRE CAASD employed a combination of tuning and configuration techniques to create the speech recognition component of the CROPD: dictionary customization, statistical language modeling, acoustic model adaptation, and robust parsing. Further, MITRE CAASD developed an application-specific analysis methodology, including performance metrics beyond the standard Word Error Rate (WER) measure of speech recognition performance, to better fit the application. This paper briefly outlines the challenges and considerations for applying speech recognition in the ATC domain and describes the CROPD as a particular application to exemplify how the challenges and considerations are addressed via tuning techniques used to adapt the speech recognition system. Performance results from the field test demonstration are presented to illustrate the value of these tuning techniques and identify where future research can target further improvement.**

Keywords-automatic speech recognition; air traffic control; voice communications; closed runway

I. INTRODUCTION

In the Tower/Surface domain, runway incursions are a major safety concern and, consequently, the primary metric for evaluating surface safety. A runway incursion is defined as the incorrect presence of an aircraft, vehicle, or pedestrian on a surface designated for the arrival or departure of aircraft [1]. However, despite existing mechanisms in place to prevent and reduce the severity of such incidents, severe runway incursions continue to occur. On behalf of the Federal Aviation Administration (FAA), the MITRE Corporation has been investigating the potential for applying automatic speech recognition technology to help prevent a specific type of runway incursion: a closed runway operation, defined as an aircraft landing on or taking off from a runway that is designated as closed. Automatic speech recognition offers a unique potential for this purpose because it can convert controller-pilot voice communications into information that can be used to produce a safety alert. Because controllers and pilots use voice communication to coordinate their intentions for future movements on the airport surface, automatic speech recognition can be used to infer that intent and feed an alert logic system that identifies the potential danger of the intent. In the case of preventing closed runway operations, automatic speech recognition can be used to detect Local Controller (LC) clearances to land, take off, or line up and wait on runways, and this intent information can be passed to a straightforward logic that triggers an alert if the runway is designated as closed. This system is called the Closed Runway Operation Prevention Device (CROPD).

There are several aspects of Air Traffic Control (ATC) speech transmissions that make applying automatic speech recognition difficult, but other aspects of ATC communications can be leveraged to improve speech recognition performance. Further, to develop and improve a system that uses automatic speech recognition on ATC communications, it is critical to define appropriate metrics. With the definition of performance metrics and an understanding of the ATC communications domain, a variety of speech recognition tuning techniques can be applied to develop a system that best achieves the performance required for the application to be effective.

This paper describes the development and evaluation of the speech recognition system that is central to the CROPD performance. In the context of the CROPD application, speech recognition performance measures are identified and several techniques for improving speech recognition performance on ATC communications are described and presented with

performance results from a field test demonstration conducted at Washington/Dulles International Airport (KIAD). These techniques are applicable to other applications of speech recognition on ATC communications, and the paper concludes with a discussion of the possible future applications of speech recognition based on the demonstrated (and expected near-future) performance.

II. BACKGROUND

This section provides an operational context for the CROPD and a background discussion on the application of automatic speech recognition to ATC communications.

A. Runway Safety and Closed Runway Operations

LCs at an airport are responsible for all activity on the airport's runways, particularly the use of the runways for operations (i.e., arrival or departure). The use of runways is typically dictated by the particular airport configuration at the time, but any open runway can be used for arrival or departure if authorized by the LC. Any unauthorized presence on a protected area of the surface is classified as a runway incursion, which is a metric used by the FAA for measuring runway safety. However, if the runway is closed, it may not be used for arrival or departure. Thus, one type of runway incursion occurs when an aircraft lands on or takes off from a runway that is designated as closed.

Runways may be closed for a variety of reasons, including runway inspections, snow removal, grass mowing, and construction. Because of the potential for equipment, vehicles or personnel to be present on a closed surface, aircraft operations are restricted or altogether prohibited on runways that are designated as closed. Closed runway operations are a particular type of runway incursion.

When a runway is closed, a series of procedures, outlined by the facility's Letters of Agreement (LOAs) and Standard Operating Procedures (SOPs), is followed to inform all relevant personnel of the runway status. Controllers in the Tower use flight strips or placards as a memory aid and automation systems that provide airport information to pilots—the Automatic Terminal Information Service (ATIS) and Notices to Airmen (NOTAMs)—are updated. The flight strips or placards are placed within the controller's typical field-of-view, intended to be passive memory aids that remind the controller that the runway is closed. Some Towers have additional surface safety system technology to help prevent closed runway operations by utilizing surveillance information. These systems, such as Airport Surface Detection Equipment Model X (ASDE-X), accept user input about runway status and use surveillance information about the location and movement of aircraft on the airport surface to determine when a closed runway operation may be about to occur. However, these surface safety systems require aircraft to reach certain kinetic parameters for an alert to be triggered, which can result in alerts issued too late for corrective action to be taken.

Though not a frequently occurring type of incursion, closed runway operations do continue to occur, despite the presence of the existing prevention mechanisms. The FAA conceived the CROPD to help prevent these incursions using automatic speech recognition.

B. Applying Speech Recognition to ATC Communications

Voice communication between controller and pilot over a radio frequency is the primary means for the two parties to communicate the current and near-future state of the operational environment. Consequently, the controller-pilot voice communications contain a wealth of information that could be used by automation systems for a variety of purposes.

Research on ATC voice communication and applying automatic speech recognition in the ATC domain has included analysis of controller-pilot voice communications, call sign confusion by humans, and the role of voice communications in aviation accidents [2] [3] [4] [5].

There are five types of speech recognition applications in the broad Air Traffic Management (ATM) domain [6].

- Training – Automatic speech recognition is used to simulate pilot behavior (clearance readback and execution of commands) during controller training. [7]
- Human-in-the-Loop (HITL) Simulation Support – Similar to how it is used for controller training, automatic speech recognition can be used to simulate pilot behavior during HITL simulations for ATM research purposes.
- Safety Benefit in Live Operations – Automatic speech recognition can be used on live, real-world ATC communications to provide real-time safety alerts to controllers. Speech recognition is particularly suitable for live safety applications because of the fact that controller clearances (and subsequent pilot readback) almost always *precede* the movement of the aircraft. The CROPD falls in this category of application.
- Efficiency Benefit in Live Operations – Automatic speech recognition can also be used on live, real-world ATC communications to help automate routine tasks or to provide the automation system with information about human intent. [8]
- Research and Analysis – Automatic speech recognition can be performed on recorded audio to support post-operations analysis. The recognized speech can be associated with track data to provide a more complete flight record, as in the case of MITRE's National Voice Archive. [9]

As described in [6], the variety of applications for automatic speech recognition are differentiated by several important characteristics, such as the operational environment (i.e., real life or lab/simulated operations), processing timeframe (i.e., real-time or post-processed recognition), information needs, and the availability of external context information.

Some systems have been developed to take advantage of dynamic context information from other automation in the ATC environment (such as call signs) to improve the speech recognition performance for simulation [10] and training [7]

applications. Helmke et al. demonstrated both the value of context information to improving speech recognition and the value of using speech recognition to inform an Arrival Manager (AMAN) system [8].

Because of the differences in these characteristics, applications differ in both the level of speech recognition performance required for success and the level of speech recognition performance achievable. When speech recognition is applied in a lab/simulated environment, such as for controller training, the systems can take advantage of high quality audio through the controlled lab environment and context information available through the simulation platform. Further, the application itself defines what type of and how much information the speech recognition system needs to identify for the application to be successful. For example, a system to simulate pilot behavior needs to identify the call sign and all instruction information in order to execute aircraft maneuvers and deliver a correct readback.

The information needs of the system then drive the definition of metrics that can be used to evaluate speech recognition system performance. Because ATC communications commonly contain words that do not contribute to the meaning of the transmission, not all words are equally important to recognize. Furthermore, not all meaningful words are essential to capture a particular objective. Consequently, Word Error Rate (WER), the common metric for measuring speech recognition accuracy that attributes equal importance to all words in a spoken utterance, is not an appropriate performance measure for speech recognition on ATC communications. Rather, each transmission typically contains a series of 'concepts' (such as the call sign or a particular clearance), a subset of which are important to identify correctly, depending on the application [8]. As described in Section IV, the performance metric for the CROPD is the identification of controller intent to use a particular runway for arrival or departure. The efficacy of the entire system will ultimately be judged by how the speech recognition result is used in conjunction with other information. An alerting system, for example, would be judged by the number of missed and false alerts, but correct speech recognition performance does not directly correspond to correct system alert performance. The controller intent metric provides a more direct method of measuring speech recognition performance, independent of the other variables that may impact overall system performance.

Successfully capturing relevant concepts in ATC transmissions requires handling several characteristics of ATC voice communications that make automatic speech recognition more challenging, as well as taking advantage of several aspects that support the speech recognition process. One challenge to applying speech recognition in the ATC domain is audio quality of the transmissions, which is determined by the acoustic characteristics of the facility audio system, such as the voice switch. Further, acoustic characteristics of the speakers—rate of speech, pronunciation, annunciation, and differing accents—are more challenging because of the inconsistency in the phonemes present in the speech. Another challenge is that, while ATC communications are prescribed by the ATC Handbook, JO 7110.65, controllers do not always follow the phraseology exactly, making language model matching more difficult [11] [7]. Additional language modeling challenges come from differences in the types of transmissions required in each ATC domain (Tower, Approach Control, etc.) and the colloquialisms that vary by facility or region.

On the other hand, there are some aspects of ATC communications that are conducive to automatic speech recognition because they help tell the system what to 'expect.' The first is the prescribed phraseology, which, although not always followed precisely, can be leveraged to produce domain- and application-specific language models. For example, the fact that specific phrases are designated only for giving takeoff and landing clearances (e.g., "cleared for takeoff" and "cleared to land") reduces the potential variation in the speech and specifies what the system needs to detect. Further, the standard phraseology is generally designed to reduce ambiguous sounds for the benefit of the human listeners (such as 'niner' instead of 'nine' to reduce ambiguity with 'five'), which also benefits the speech recognition system. Additionally, by modeling previously observed ATC speech, systems can be tuned with custom pronunciation dictionaries, which prepare the speech recognition system for common pronunciation variations (such as "clear da lan"). Speech recognition systems can also be adapted to the acoustic characteristics of the audio, including speaker- or group-specific traits and impacts from the audio equipment. Finally, both static (e.g., the runway numbers at a particular airport) and dynamic (e.g., the call signs in the airspace at a given time) context information can further refine what the system expects to hear on a given transmission.

Section IV describes in detail how these characteristics are handled in the CROPD system via a variety of automatic speech recognition tuning techniques.

III. The Closed Runway Operation Prevention Device

This section describes the functional components of the CROPD, how it is intended to be used, and the role of its speech recognition system. More detail on the design of the system can be found in [12].

A. System Description

The purpose of the CROPD is to detect (and alert) when the LC gives a clearance indicating intent for an aircraft to land on or take off from a runway that is designated as closed. To determine if an alert should be generated, the two pieces of information that need to be compared are a) the clearance and associated runway spoken by the controller to the pilot, and b) the closed/open status of that runway.

The closed/open status of each runway is displayed on a small Graphical User Interface (GUI) in the Tower, which the controllers and/or supervisors would keep up to date as part of the checklist followed for closing and opening runways. Other sources of information could be used to provide the closed

runway status, such as ASDE-X or NOTAMs, but the current version of the CROPD has been designed to minimize the number of interfaces with other automation systems.

The automatic speech recognition component of the CROPD receives controller audio from the voice switch, but does not connect to other systems for context information. The speech recognition component feeds input to the alert logic, which compares the recognized clearance and runway to the closed runway status to trigger an alert. Fig. 1 illustrates the functional components of the CROPD.

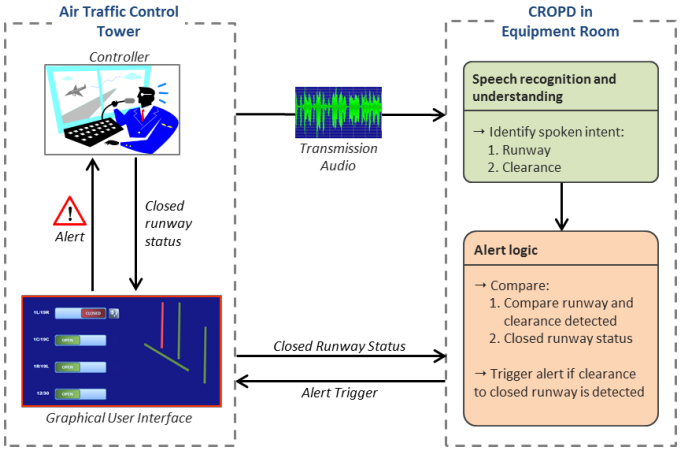

FIGURE 1. CROPD Functional Components

When the system detects a clearance to a closed runway, it triggers an auditory and visual alert in the tower, through the GUI and a set of loudspeakers. In response to an alert, the controllers in the tower will determine if the alert requires a response (i.e., if it is a true alert or a false/nuisance alert) and issue corrective instructions if necessary. Because the controller instruction almost always precedes the aircraft movement, and because the alert can be triggered almost immediately after the controller finishes the clearance, the CROPD alert will in theory give the controller the maximum possible amount of time to respond to the situation.

B. The Role of Speech Recognition

To identify controller intent to use a runway for an operation, the speech recognition component of the CROPD works to detect two specific pieces of information in a controller transmission: the clearance and the runway associated with that clearance. At a minimum, the clearance phrases include "cleared to land", "cleared for takeoff", and "line up and wait". If the CROPD is installed at a small airport with sufficient general aviation traffic, then other clearance phrases such as "cleared for the option" would also be included. The runway numbers are limited to the names of the runways at the particular airport. Ultimately, as described in Section IV, the speech recognition system is configured to "expect" the clearance phrases and runway numbers that are typically spoken at the airport.

ATC transmissions typically contain other concepts, such as call signs, wind information, and traffic advisories, but the speech recognition system for the CROPD only needs to detect the clearance phrase and associated runway. Recognizing the runway number can be particularly challenging because a transmission will contain many numbers as part of the other content. The name of another runway at the airport may be present in another context or as part of a traffic advisory. Table 1 presents an example clearance and the variety of concepts (or pieces of information) the system must differentiate.

TABLE 1. Example Clearance and Number Recognition

United one twenty-three, wind three three zero at six, runway one center cleared to land, traffic departs runway three zero.	
Recognized numbers	**Actual concept**
One twenty-three	ACID
Three three zero	Wind direction
Six	Wind speed
One (center)	Runway – clearance ✓
Three zero	Runway

The system may identify several different numbers, some of which may be part of mentioning another runway, but the system must determine which number is associated with the runway clearance. Because of the variety of numbers that may be present in a transmission, and because each physical runway has two names (one from each direction, such as 12 and 30), it is possible for the speech recognition system to identify that the LC spoke intent for an arrival or departure, but associate the wrong runway number with the clearance phrase. Consequently, there are five types of results needed to assess the speech recognition system's ability to identify controller intent: correct intent, incorrect intent, false intent, missed intent, and correct rejection of intent. Table 2 presents a matrix of the possible outcomes based on examples.

TABLE 2. Matrix of Speech Recognition Performance Outcomes

		Truth *Actual Transmission*	
		Intent: RWY30, CFT	No Intent
CROPD Result	Intent: RWY 30, CFT	Correct Intent	False Intent
	Intent: RWY 19L, CTL	Incorrect Intent	False Intent
	No Intent	Missed Intent	Correct Rejection

In the first example of the actual transmission (Truth) in Table 2, the controller gives a takeoff clearance ("cleared for takeoff" [CFT]) for runway three zero (RWY 30). In this case, the system identifying RWY 30, CFT is correct intent recognition. The system identifying a different intent in the transmission—RWY 19L, cleared to land (CTL)—is incorrect intent recognition. In other words, the system correctly identified intent to use a runway, but it identified the wrong

clearance phrase and wrong runway. Note that the system could also identify the correct clearance phrase but the wrong runway (for example, RWY 19L, CFT), or the correct runway but the wrong clearance phrase (for example, RWY 30, CTL). Thirdly, the system could not identify intent when intent is actually present, which is classified as missed intent (see Table 2).

The second example in Table 2 is if the controller gives a transmission but does not speak intent to use a runway for an operation ("No Intent" in the right-hand column). In this case, regardless of which runway and clearance phrase (e.g., RWY 30, CFT) is detected, the system performance is classified as false intent. If the speech recognition does not recognize intent in the transmission, performance is classified as correct rejection of intent.

Finally, to evaluate the performance of the CROPD overall, intent recognition results must be correlated with the actual alert performance. For a variety of reasons—number of runways, different types of runway closures, and that each runway has two names (one for each direction)—incorrect speech recognition results do not necessarily lead to incorrect CROPD alerts. For example, if the speech recognition detects a takeoff clearance for runway 12, but the controller actually gave the clearance for runway 30, then the system would still alert correctly, assuming the runway is closed to departures in both directions. CROPD system alert performance also depends on having the GUI set correctly.

While alert performance is the system performance measure that will directly impact safety and user acceptance of the system, an appropriate measure for evaluating the speech recognition component is the detection of controller intent to use a runway for arrival or departure. Better performance in recognizing controller intent in the transmission is the means for engineering a system that provides better alert performance to the user.

C. Field Test Demonstration

In the summer of 2014, at the request of and working collaboratively with the FAA, MITRE prepared for and executed a field test demonstration of the CROPD at KIAD. The objectives were to demonstrate the CROPD as part of the National Airspace System (NAS), evaluate the speech recognition performance on live controller audio, and to elicit feedback from operational personnel on the design and proposed use of the system. With respect to the analysis results presented in this paper, the field test demonstration serves as 1) the source of the audio data used for the analysis, and 2) validation that the performance results presented are applicable to the live operating environment.

When hosted in the field during live operations, the CROPD included a pre-processing component before the speech recognition component that performed automatic identification and delimitation of controller transmissions from the continuous stream of incoming audio. The component was built around the speech classifier components in Sphinx4, Carnegie Mellon University's open-source automatic speech recognition system. Because automatic speech identification and delimitation is itself a probabilistic process with inherent uncertainties and error, this component can contribute to overall system error. For clarity within the context of this paper, the results presented in subsequent sections focus only on the speech recognition performance of the CROPD, independent of any speech delimitation error. The tuning that could be implemented to reduce errors associated with speech segmentation is a separate topic not discussed within the scope of this paper.

The following three sections describe the performance measures, tuning techniques, and analysis results from the field test demonstration of the CROPD at KIAD.

IV. Performance Measures and Evaluation Data

This section describes the performance metrics used to measure performance of the CROPD's speech recognition component and summarizes the data set used to tune the speech recognition system and the data set used to evaluate its performance during live operations in the field.

A. Performance Metrics

As mentioned previously, WER is not an appropriate accuracy metric for applications in which some words are inherently more meaningful to the end objective of the overall system than others. In the case of the CROPD, runway and arrival/departure clearance phrases are more relevant than other phrases that may be present in the LC transmission, such as the aircraft call sign, weather advisories, courtesies, etc., to the deduction that the controller has expressed intent to use a particular runway for arriving or departing traffic. Thus, an alternate accuracy measure that only takes into account words that affect the overall system performance and does not penalize the system for failing to decipher words that are irrelevant to the application's purpose would provide a better evaluation of the system's performance with respect to the application objective.

In the case of the CROPD, its accuracy measures are based on intent, defined as the presence of both a clearance phrase, for arrival or departure, and an associated runway phrase in a single transmission, and closely mirror the five outcome types described in Table 2. Equations (1) through (5) describe how the intent-based performance metrics are calculated for the CROPD.

$$P_{\substack{True\ Intent \\ Detection}} = \frac{I_{correct\ runway\ and\ clearance}}{T_{with\ Intent}} \quad (1)$$

$$P_{\substack{Incorrect\ Intent \\ Detection}} = \frac{I_{incorrect\ runway,\ \ clearance,\ \ or\ both}}{T_{with\ Intent}} \quad (2)$$

$$P_{\substack{Missed\ Intent \\ Detection}} = \frac{NI_{with\ Intent}}{T_{with\ Intent}} \quad (3)$$

$$P_{\substack{Correct\ Non-Intent \\ Detection}} = \frac{NI_{without\ Intent}}{T_{without\ Intent}} \quad (4)$$

$$P_{\substack{Incorrect\ Non-Intent \\ Detection}} = 1 - P_{\substack{Correct\ Non-Intent \\ Detection}} \quad (5)$$

In the equations, $T_{with\ Intent}$ denotes the total number of transmissions with intent, $T_{without\ Intent}$ denotes the total number of transmissions without intent, $I_{correct\ runway\ and\ clearance}$ denotes the number of transmissions that were identified with intent and the correct runway and clearance, $I_{incorrect\ runway,\ clearance,\ or\ both}$ denotes the number of transmissions that were identified with intent but with an incorrect runway, clearance, or both, $NI_{with\ Intent}$ denotes the number of transmissions that were not identified with intent but actually contained intent, and $NI_{without\ Intent}$ denotes the number of transmissions that were not identified with intent and did not actually contain intent.

B. Tuning Data Set

To build the CROPD demonstration system, MITRE assembled and tuned its speech recognition component using recorded operational audio from KIAD. The audio was collected by the FAA over several weeks in the spring of 2013, from an analog patch panel interface on the Enhanced Terminal Voice Switch (ETVS) in the KIAD equipment room. Portable digital recorders were used to digitize and store the audio at 8 kHz with 16 bits per sample. All LC radio transmissions from the three LC frequencies at KIAD were recorded during the audio collection process. In total, over 360 hours of audio were collected. A subset of this audio data, 144 hours, was selected as a representative data set for transcription, analysis, and tuning of the CROPD speech recognition component. When analyzed and transcribed, the 144 hours of audio yielded 12,834 distinct radio transmissions and corresponding transcriptions. One sixth of this data set was reserved for evaluation and benchmarking during the tuning process, and the remaining data were used for tuning and automatic training.

C. Field Demonstration Data

From the field test demonstration at KIAD, a subset from the audio and system logs archived by the system during its demonstration was randomly selected for performance analysis. Out of over 1,100 hours of archived audio data, 125 hours were selected as the evaluation data set and manually transcribed and analyzed. The audio yielded 13,804 distinct controller transmissions and corresponding transcriptions. Within this data set, just over forty percent of the data, 5,822 transmissions, contained intent to use a runway (as previously defined). The human-generated transcriptions and identifications of intent in the audio set were designated as the ground truth and used to validate the speech recognition results from the CROPD system.

V. Speech Recognition Tuning

The CROPD is built around the Loquendo automatic speech recognition engine, a commercial speech recognition system. Before the CROPD was fielded at KIAD for evaluation in a live ATC environment, the speech recognition component was configured and tuned using the tuning data set described earlier. This configuration and tuning process was required because the performance accuracy of a speech recognition system is largely dependent on how well the target audio being recognized matches the system's internal models of speech and language patterns. Most commercial speech recognition systems are sold with generalized internal models that are optimized for recognition of conversational speech from a large population of speakers. A number of factors—environment and acoustic characteristics, speaker accents, speech rate, non-standard vocabulary, and context ambiguity—that are specific to ATC applications make these generalized models less optimal for application in the ATC domain. This section outlines tuning techniques for improving recognition accuracy, specifically for use in the ATC domain, and describes how each was implemented in the CROPD with the observed benefit from the implementation of each technique.

A. Language Modeling

Language modeling, in the context of speech recognition, defines the universe of word sequences that the automatic speech recognition system can recognize and models the probability or likelihood of occurrence of that word sequence. There are typically two methods of language model definition: (1) finite-state grammars and (2) statistical language models (SLMs). The first is a manually-defined list of word sequence rules, frequently referred to as a grammar, which encapsulates the logical and conceivable phrases that may be encountered during the recognition process. Fig. 2 depicts an example of a grammar rule for runways 1L/19R, 1C/19C, and 1R/19L. A full grammar could comprise of multiple rules such as the one depicted as well as a root rule that specifies the interconnection of rules.

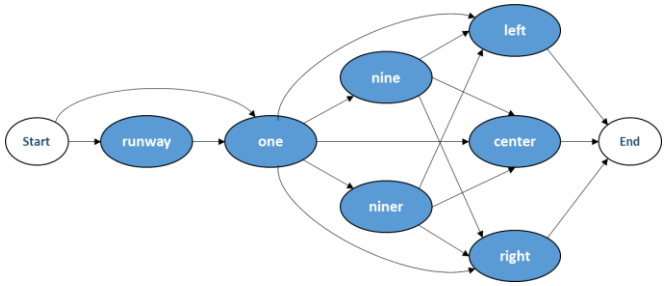

FIGURE 2. Grammar Rule for Runways 1L/19R, 1C/19C, and 1R/19L

Defined well, these language models can yield near-perfect recognition in applications where speakers adhere strictly to the expected phraseology. However, grammars have low tolerance for speech that deviates from its predefined patterns. Previously unseen patterns can lead to a dramatic drop in accuracy.

The SLM method of definition results in a machine-generated probability model of word sequence occurrences created through analysis of a set of transcriptions (also known as a corpus) from the target speech environment, perhaps mixed with a more general training corpus. SLMs bias recognition toward words and combinations of words that have been previously observed in the automatic training corpus, but do not enforce strict conformance to a finite set of predefined word combinations as grammars do. This behavior makes SLMs more robust to speech variation and disfluencies such as hesitation, stuttering, and coughs but can also make their results less accurate or illogical, in poor recognition conditions such as significant background noise, unforeseen acoustic variations, and atypical speaker accents.

The CROPD uses both types of language models. The finite-state grammar was manually crafted to contain only word sequences relevant to the CROPD application, which are the commonly used variations of clearance and runway phrases. These variations included word substitutions such as "clear" for "cleared", additions such as "immediate", and omissions such as absence of the word "runway". Table 3 depicts examples of the variations included for clearance and runway phrases. Note that letters and words in the square brackets are optional alternatives.

TABLE 3. Examples of Variations on Clearance and Runway Phrases

Original Phrase	Possible Variations
cleared for takeoff	clear for takeoff clear[ed] to takeoff clear[ed] for immediate takeoff
runway one niner left	one nine[r] left
runway three zero	[runway] tree zero

The SLM was created and trained on all transcriptions in the tuning data set described earlier. Note, the tuning data set contained transcriptions from all positions at KIAD and time periods throughout the day, offering a reasonable distribution of the facility's runway usage during the data collection period. Selection of a representative data set for automatic SLM training is important, especially with small data sets, because this data determines the language model's bias during recognition. If the training data was selected from time periods during which the facility used one particular runway configuration and set of runway names, these runway names could be over-represented in the SLM while other runway configurations and runway names would be under-represented. During recognition time, this unrealistic bias could lead to false recognition of the over-represented runway names and missed recognition of the under-represented runway names.

With the language models, the speech recognition component of the CROPD was able to identify clearance and runway phrases that were present in the evaluation data set fairly well. However, in some instances, recognition accuracy was offset by a high percentage of false detections—that is, identification of phrases of interest when none were actually present in the speech. The two language models were a good complement to each other and each introduced performance benefit that would not have been present with only the other. An initial benchmark of the system taken after both language models were in place showed that true intent detection was just less than 70 percent and false intent detection was over 50 percent.

B. Acoustic Modeling

Acoustic models define the statistical signatures that identify sub-components of a spoken language, known as phonemes, and specify the combinations of these phonemes that form words in the language. Most commercial speech recognition systems come pre-configured with a default acoustic model, generalized for performance on conversational speech from a large speaker population. Because controller speech can differ significantly from conversational speech in the general population, these acoustic models are not optimal for use in the ATC domain.

There are two standard methods for modifying and tuning the acoustic model: (1) pronunciation dictionaries and (2) acoustic model adaptation (AMA). The first method is a means of supplementing the existing acoustic model's word pronunciation definitions, also known as lexicons, with custom word pronunciations specific to the application. These custom pronunciation dictionaries can be crafted to target known pronunciation variations of existing words and to introduce new words and phonetic sequences. This capability is useful in the ATC domain, where there are many recurring, non-standard words such as fix names, route names, and aircraft call signs. Furthermore, rapid rate of speech and repetitive phraseology frequently lead to coarticulation[1] and assimilation[2] in ATC speech, resulting in truncation and distortion of well-known word pronunciations; a pronunciation dictionary can be modified in response to these observations. For example, in some of the speech samples analyzed for the CROPD application, the phrase "cleared to land" was sometimes truncated to sound like "clea da lan" or "cland" and the phrase "cleared for takeoff" was distorted to "clea fa taguv". Although custom user dictionaries can be a powerful tuning tool, entries must be judiciously selected because each entry expands the search criteria, increasing sensitivity of the system to the defined lexicon. Over-expansion can lead to unintentional overlap with other word lexicons resulting in erroneous detection of a confusable alternative rather than correct identification of the actual word.

The second method of acoustic model tuning, AMA is an automated means of modifying a general acoustic model to better match observed acoustic channel characteristics and speech patterns within a particular set of speakers. Similar to SLMs, AMA relies on analysis of representative data—both audio and transcriptions—from the target speech environment and speaker set to adjust the default acoustic signatures of phonemes in the acoustic model. This method can be utilized to account for speaker-dependent characteristics, such as speech rate and accents, with a fairly limited training data set. With a large training data set, this method can also account for non-speaker specific characteristics, such as audio channel properties and bandwidth limitations.

The CROPD uses both a custom pronunciation dictionary and AMA. The custom dictionary was manually created with unique ATC pronunciations (such as "niner") as well as custom words that defined entire clearance phrases (such as "cleared to land") and runway phrases (such as "one niner left") as single

[1] Coarticulation is the action of a speaker to anticipate the next sound in speech, carrying over and merging speech sounds from preceding or subsequent speech segments to the current sound being vocalized [14].

[2] Assimilation is the change performed on an isolated speech sound to make it more like a neighboring speech segment [14].

word entries. The latter variation on phrases was added to allow for better handling of coarticulation and assimilation observed between words in these frequently spoken phrases. When added to the speech recognition configuration alongside the two language models, the custom pronunciation dictionary dramatically boosted recognition accuracy for several keyword phrases. However this performance gain was also accompanied by an increase in the number of false detections across all the keyword phrases. A benchmark of the system taken after the dictionary was implemented alongside the language models showed that true intent detection increased to 85 percent but false intent detection increased to 61 percent.

Acoustic model adaptation was implemented next and trained on all audio and corresponding transcriptions in the tuning data set. The goal of implementing this technique was to shift the default acoustic model closer to the speaker set at KIAD and account for some of the acoustic characteristics unique to the facility voice switch. The addition of the acoustic model adaptation improved the correct recognition rate minimally and significantly decreased the false detection rate observed on keyword phrases. A third benchmark of the system taken after language modeling and acoustic modeling were in place showed true intent detection increased to 89 percent and false intent detection dropped to 18 percent.

C. Semantic Interpretation

Semantic interpretation is a form of text processing that can be used to derive logical concepts from the potentially error-filled raw text output, or hypotheses, of the speech recognition system. The process of translating words or groups of words to logical concepts can vary across applications, even within the same domain, and is dependent on the quality of the transcriptions being processed, the complexity of the concepts being derived, as well as the relation between concepts within the same transcription. In addition to deducing logical concepts, semantic interpretation techniques can also reduce errors by skipping over words that are not relevant.

In the CROPD, the semantic interpretation component is responsible for sifting through the hypotheses from the two language models for relevant runways and clearances of interest, resolving conflicts between recognition hypotheses if differing clearance or runway phrases were recognized within the same time segment, correlating the correct runway to the clearance if multiple runways were identified, and arriving at a final intent-to-use-a-runway hypothesis, if one was present.

The semantic interpretation component accepts word-level confidence scores, which are machine-generated measures of accuracy likelihood from the speech recognition system, per-word start and stop times, and the recognized transcription hypotheses for each interpretation task. During the mediation process, the algorithm prioritizes recognized phrases based on phrase length (that is, the number of words in the phrase), phrase-level confidence score, phrase proximity to other phrases, and identifying text qualifiers (such as the word "runway"), which if present could disambiguate the meaning of numerical sequences. The addition of the semantic interpretation component to the CROPD did not improve recognition of the individual keyword runway or clearance phrases but did improve overall intent detection performance of the system because it introduced a more intelligent method of disambiguating runway concepts from other confusable numerical phrases and correlating the correct runway with a clearance when multiple runways were mentioned and recognized. A final benchmark of the system taken after language modeling, acoustic modeling, and semantic interpretation were all implemented showed true intent detection to be 95 percent and false intent detection less than 17 percent.

D. Confidence Thresholding

As mentioned briefly earlier, speech recognition systems commonly produce confidence scores as a measure of the likelihood of accuracy in its speech-to-text translation. Most commercial speech recognition systems provide both word-level and hypothesis-level confidence scores as a means of discriminating between multiple hypotheses and between correct recognition and erroneous identification of specific words. High confidence scores from the system indicate that it believes its recognition hypothesis is most likely accurate, while low confidence scores could indicate that segments of uncertainty or error exist within the translation.

Confidence thresholding is a methodology that exploits the availability of these system-generated scores to bias the overall system towards a specific balance of missed and false detections. This process does not affect the accuracy of the underlying speech recognition system; rather, it advises the system on whether to accept a recognition based on its own certainty of accuracy. Confidence thresholding establishes the minimum system confidence, or confidence threshold, that must be met in order for a hypothesis of true detection to be accepted. In the signal detection theory framework, this confidence threshold is the decision criterion that determines the response bias or response criterion [13]. If the confidence threshold is set with a liberal response bias, then it leans the overall system toward accepting false detections so that missed detections are minimized. If the confidence threshold is set with a conservative response bias, then it leans the overall system toward accepting missed detections so that false detections are minimized. The actual selection of the confidence threshold, and thus the response bias, is usually determined by the application context and the application users.

In the CROPD, the capability to adjust the confidence threshold on the system is implemented at two different levels. The first threshold is within the semantic interpretation algorithm, during the mediation between clearance and runway phrases from across multiple recognition hypotheses. This threshold eliminates phrases with low likelihoods of accuracy early in the selection process, before they introduce conflict with much higher-scoring phrase hypotheses. The second threshold is implemented on the outcome of the semantic interpretation algorithm—the final intent determination. This threshold discards final intent determinations that have a low aggregate confidence across all the phrases and words used to

determine the intent. The confidence threshold is set to balance the trade-off between missed alerts and false alerts.

VI. Analysis and Results

All of the speech recognition tuning techniques described in the previous section were implemented using the tuning data set from KIAD and then benchmarked using a subset of the tuning data that was set aside for evaluation and not used for tuning nor automatic training. Thus the benchmark results from the tuning process were generated on previously "unseen" data and acted as predictors of how the system would perform in the field, when it was exposed to new data during live operations. This section describes the observed speech recognition performance of the tuned CROPD on the demonstration data set described earlier, which was processed during the field demonstration at KIAD.

A. Speech Recognition Performance

For the 13,804 transmissions that were selected for analysis from the collected demonstration data, the CROPD was able to correctly detect over 92 percent of the transmissions containing intent to use a runway. Of the remaining transmissions without intent, less than 10 percent were incorrectly identified as containing intent (i.e., false intent). Table 4 depicts all five performance measures calculated per the metric definitions specified earlier.

TABLE 4. Performance Measures

Transmissions with Intent			**Transmissions without Intent**	
True Intent	***Incorrect Intent***	***Missed Intent***	***Correct Non-Intent***	***False Intent***
92.70%	4.77%	2.53%	90.33%	9.67%

The performance percentages listed above were calculated with the overall system confidence threshold set to zero. Fig. 3 depicts the receiver-operating characteristic (ROC) curve capturing the full range of confidence score threshold alternatives and their subsequent impact on the balance between correct and false intent detection.

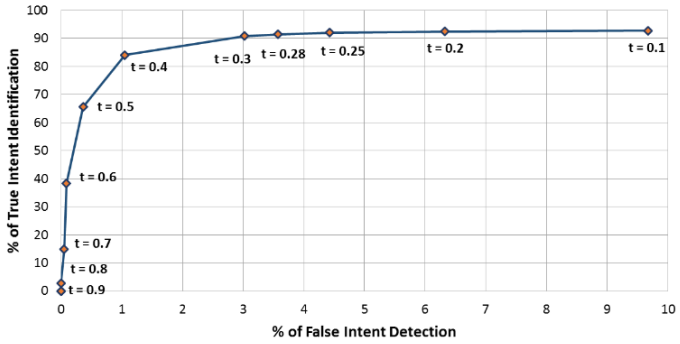

FIGURE 3. ROC Curve for Different Confidence Thresholds

In Fig. 3, 0.9 is the highest confidence score threshold depicted. At this threshold, zero false intent detections were observed but also zero true intents were accepted by the system. At the opposite end of the scale, 0.1 is the lowest confidence score threshold depicted. At this threshold, over 92 percent of intent present were accepted but simultaneously almost 10 percent of transmissions not containing intent triggered a false intent detection.

The selection of a "best" confidence score threshold is dependent on the desired tradeoff between missed and false detection, which is informed by operational preference. The ROC curve can aid in this selection by elucidating the tradeoff between missed and incorrect intent detection. For example, based on the ROC curve, if the requirement is for at least 90 percent correct intent detection (i.e. no more than 10 percent missed intent detection), a confidence score threshold can be set at t=0.3 to limit the incorrect intent detection (i.e. false intent detection) to three percent.

B. Observations

The performance results documented here suggest that speech is a feasible and usable information source for deriving intent that can provide potential benefit in a safety application. However the results also indicate that there is still room for improvement in the speech recognition component. Analysis of the error cases in the evaluation data set revealed some patterns that highlight areas for future research and tuning.

In Table 4, incorrect intent detections, in which intent was detected but either the clearance type, the runway associated with the clearance, or both were incorrect and did not match the true intent details, comprise nearly 75 percent of the total error in intent detections for transmissions containing intent. Delving into these error cases showed that incorrect runway identification or association with the intent caused the majority of these errors; further, in many of these cases, the correct runway to associate with the intent was recognized by the speech recognition, but disregarded by the semantic interpretation algorithm for another recognized runway. This analysis indicates that the semantic interpretation algorithm could be improved to better correlate clearances with their related runways.

A more detailed analysis of the 10 percent of false intent detections showed that a notable fraction of the error cases were caused by the incorrect matching of a runway concept to a correctly recognized numerical phrase. For example, number 30 actually appeared as a part of a call sign, an altimeter reading, a radio frequency, a wind advisory, or a flight level in a number of error cases but it was incorrectly tagged as a runway by the semantic interpretation algorithm. Similarly, the number 12 frequently appeared as a part of a radio frequency (although it also appeared in call signs, altimeter readings, or wind advisories to a lesser degree); the semantic interpretation algorithm would incorrectly tag the correctly recognized numerical phrase as a runway. This result suggests that the semantic interpretation algorithm could be improved, but it also suggests that additional keyword phrases may need to be recognized in order to provide more intelligent context cues for the semantic interpretation algorithm.

VII. Implications and Future Extensions

The metrics presented in the previous section are appropriate for measuring improvement in speech recognition performance, but they do not directly translate to system alert performance, which ultimately dictates user experience. If the CROPD is to be implemented in the field, the ultimate system alert performance must first be validated as acceptable to operational users.

The speech recognition performance observed during the field test demonstration at KIAD demonstrates both the feasibility of applying automatic speech recognition on live ATC transmissions, and the value of the various tuning techniques on improving performance. While there is still room for speech recognition performance improvements, those next steps would involve algorithm refinements targeted at specific situations to improve performance in small incremental steps. Further, the tuning techniques used are transferrable to speech recognition in other ATC domains, such as Approach Control operations.

The performance demonstrated also lends support to the concept of using speech recognition for other applications in the ATC domain. Future applications may involve using the same detected clearances for different purposes, or they may involve detecting other controller instructions, such as “cross” or “hold short”. Speech recognition could be extended to Ground Controller clearances, perhaps for the purpose of providing automation systems with controller-issued taxi route clearances. Finally, speech recognition could potentially be applied to pilot transmissions as well, for the purpose of detecting readback errors or for supplementing the information derived from the controller clearance.

Depending on the application, the level of speech recognition performance required for application success may be even higher than what has been demonstrated so far. One technique that has been shown to improve speech recognition performance is the inclusion of dynamic context information. In a Tower/Surface environment, this information could come from a tower automation system, such as ASDE-X or a terminal radar system. Further, the speech recognition results could be fed back into the surface safety logic to improve safety alert performance.

Ultimately, although the CROPD is a relatively simple, isolated application of speech recognition in the Tower/Surface domain, its development helps to lay the foundation for other, more sophisticated applications of this transformative technology to enhance ATM system performance. As speech recognition tuning techniques evolve to produce improved performance, particularly with the integration of dynamic context information, automatic speech recognition can be used more extensively on ATC communications to provide benefits to both safety and efficiency.

Acknowledgements

The authors would like to thank Elida Smith, Rob Tarakan, Juliana Goh, Craig Johnson, Katie Shepley, and Suzanne Porter at CAASD for their guidance in this project. Additionally, the authors would like to extend special thanks to Herb King, Jonathan Gray, and Ron Singletary at the FAA for supporting this project.

References

[1] International Civil Aviation Organization, *Manual on the Prevention of Runway Incursion*, Doc 9870 AN/463, 2007.

[2] K. Ward, *A speech act model of air traffic control dialogue*, Oregon Graduate Institute of Science & Technology, Ph.D. Thesis, 1992.

[3] H. Said, M. Guillemette, J. Gillespie, C. Couchman and R. Stilwell, *Pilots & Air Traffic Control Phraseology Study*, International Air Transport Association, 2011.

[4] Great Britain Civil Aviation Authority and National Air Traffic Services (Great Britain), "Aircraft Call Sign Confusion Evaluation Safety Study (ACCESS)," CAP 704, Civil Aviation Authority, 2000.

[5] S. Cushing, Fatal Words: Communication Clashes and Aircraft Crashes, Chicago, IL: University of Chicago Press, 1997.

[6] H. D. Kopald, A. Chanen,, S. Chen, E. C. Smith and R. M. Tarakan, "Applying Automatic Speech Recognition Technology to Air Traffic Management," in *32nd Digital Avionics Systems Conference*, Syracuse, NY, 2013.

[7] R. Tarakan, K. Baldwin and R. Rozen, "An automated simulation pilot capability to support advanced air traffic controller training," in *26th Congress of the International Council of the Aeronautical Sciences*, Anchorage, AK, 2008.

[8] H. Helmke, H. Ehr, M. Kleinert, F. Faubel and D. Klakow, "Increased Acceptance of Controller Assistance by Automatic Speech Recognition," in *Tenth USA/Europe Air Traffic Management Research and Development Seminar*, Chicago, IL, 2013.

[9] The MITRE Corporation, *CAASD Product 4-2.1.C.1.6; Voice Data; F075-L14-030*, McLean, VA: Center for Advanced Aviation System Development, 2014.

[10] D. Schäfer, "Context sensitive speech recognition in the air traffic control simulation," in *4th USA/Europe Air Traffic Management R&D Seminar*, Santa Fe, NM, 2001.

[11] Federal Aviation Administration, Air Traffic Organization, *Air Traffic Control, Order JO7110.65U*, Washington, D.C.: U.S. Department of Transportation, 2012.

[12] H. Kopald and S. Chen, "Design and Evaluation of the Closed Runway Operation Prevention Device," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Chicago, IL, 2014.

[13] C. D. Wickens and J. G. Hollands, Engineering Psychology and Human Performance, Third Edition, Upper Saddle River, NJ: Prentice Hall, 2000.

[14] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, 2009.