Challenges in using a Standard Speech Recognition Engine in Small Vocabulary Domain

Narayanan Srinivasan, S. R.Balasundaram

Abstract.This paper discusses the challenges and proposes recommendations on using a standard speech recognition engine for a small vocabulary Air Traffic Controller Pilot communication domain. With the given challenges in transcribing the Air Traffic Communication due to the inherent radio issues in cockpit and the con-troller room, gathering the corpus for training the speech recognition model is another important problem. Taking advantage of the maturity of today's speech recognition systems for the standard English words used in the communication, this paper focusses on the challenges in decoding the domain specific named entity words used in the communication.

Index terms – air traffic speech,contextual speech recognition, named entity recognition, non-trained speech,

I. INTRODUCTION

Automatic Speech Recognition (ASR) systems are being developed and deployed with great amount of accuracy for real world applications. Still, 100% accuracy is one of the biggest concerns in deploying them due to various chal-lenges in the deployed environment.

II. CHALLENGES IN AIR TRAFFIC SPEECH RECOGNITION

A. *Characteristics of Air Traffic Speech Recognition*

ATC communications have very specific characteristics, which are the following:

- Constrained domain-specific language (recurrent utterances, use of callsigns) and limited vocabulary;
- Important variability of speakers (potentially bad accents) with often no availability for training an ASR system;
- Noisy environment (cockpit, control rooms) and poor-quality transmission channels (radio communications);
- Stressed speech: rapid delivery, bad pronunciation, interrupted or overlapping utterances.
- In more technical terms, ASR for ATC communications face the following constraints
- Limited bandwidth (300 to 3300 Hz) of VHF voice communications; English language has major energies in higher parts, i.e. "th" (8 to 9 kHz);
- Technical instability at beginning of speech (VHF);
- Incomplete words (phonemes) at the beginning, i.e., "fthansa one two" due to the use of "Push to Talk" switches;
- Spontaneous speech with repetitions, hesitations: "mmh", "haa" (approximately 4 to 5% of the utterances);
- Mainly nonnative English speakers;
- Mix of official ICAO languages (English, French, Spanish, Russian), i.e. "Geneva" in approximately 16% of the French utterances;
- Pronunciation of navigation point names in the speaker's mother tongue language;

B. *ASR ATC problems addressed in this work*

Some of the ASR ATC challenges discussed earlier causes the named entities in the speech to be decoded incorrectly. That is defined as the area of interest of this research and solutions for addressing that problem is discussed and suggested as future work. The problem is further detailed as,

- Challenges in decoding waypoints, procedure names, named navigation aids from the speech conversation
- Challenges in decoding aircraft call sign value from the speech conversation

This work discusses the solution for the above challenges with the help of databases or information available within the avionics system. The databases which are available in the avionics system already has the text representation of the named entities which can be used to fill the gap in speech recognition of named entities.

C. *Challenges of Acoustic Model*

Acoustic models are in the process of being created relying upon training on the Vocalize corpus in order to adapt English phonemes to operational conditions and to replace French phonemes by the nearest English phonemes. The variability between nationalities and accents of speakers makes the creation of robust acoustic models become a real challenge. This issue is rather difficult to address due to the overall audio quality of the recordings (noise, saturation, …) and the amount of training data, which is far from enough. Collecting the corpus data for training the acoustic model is a time-consuming task when we try to adapt a general speech engine for a customized vocabulary.

.**Revised Manuscript Received on July 22, 2019**.
Narayanan Srinivasan, National Institute of Technology, Trichy, India.
S. R.Balasundaram, National Institute of Technology, Trichy, India.

D. *Quality Attributes to measure ATC ASR*

There are two primary quality attributes that need to be measured in ATC ASR. One is the word error rate (WER) and the performance of the ASR component. WER is very critical as it may lead to untoward consequences if the interpretation is incorrect either on the ATC system or in the Cockpit system as these are safety critical systems. Performance is also next critical parameter as the conversation between ATC and Cockpit happens where many instructions are given in short time. The references discussed earlier introduce an additional processing component in the existing ASR which is equally complex as the ASR itself and there is a marginal improvement were seen when the techniques proposed were applied.

Accuracy and speed are the two most common metrics for measuring speech recognition system performance. Word Error Rate (WER) is usually used for measuring accuracy, whereas speed is usually rated with Real Time Factor (RTF).

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

Equation 1

Other measures of performance include Concept Error Rate (CER), Single Word Error Rate (SWER) and Command Success Rate (CSR).

The other measure used in this work is the number of files which were/not de-coded correctly in for the test data set.

III. RELATED WORK OF ASR IN COCKPIT SPEECH PROCESSING

In Air Traffic control, which is a limited vocabulary system, where the environment noise and speed at which the speech is delivered is high, achieving 100% accuracy is a real challenge for any speech recognition system. There are quite a few published papers where various techniques of applying contextual information for improving the accuracy has been proposed.

Van Nhan Nguyen et all in 2015 [1] have listed the challenges and what techniques can be adopted to resolve the challenges for deploying an ASR for Air Traffic Control system. In this paper, the author explained the challenges call signdetection, poor input signal quality, the problem of ambiguityand the use of non-standard phraseology which dramaticallyreduce the recognition rate and the performance of speechrecognition systems. The author further details out the current state of art ASR systems which are available in general and for Air Traffic Control system. The author concludes that Combining state-of-the art ASR approaches with contextualinformation to include syntactic, semantic and pragmaticanalysis in the recognition process, and the identification ofdialects, accents and languages holds great promise for theapplication of automatic speech recognition in the air trafficcontrol domain.

In their work, "N-best List Re-ranking Using Syntactic Score" [2] the author has proposed a N-best list rescoring using Syntactic score to improve the accuracy of the speech engine. This rescoring is done based on the syntactic score which is computed using syntactic rules. This approach uses context dependent re generation of language model.The proposed model outperformed the traditional n-gram model and showed 18.21% improvement in terms of Word Error Rate (WER) with Air Traffic Control Speech Corpus.

In another work,"N-best List Re-ranking Using Semantic Relatedness and Syntactic Score", [3] the authors proposed a method which combines the semantic relatedness, syntactic score and speech decoder'sconfidence score features to perform n-best list re-ranking.Thisapproach also shows 19.93% WERimprovement compared with traditional n-gram model on the ATCSC corpus.

Youssef Oualil et all in 2015 [23]have explained how to integrate real time contextual information in ATC ASR. In this paper a weighted Levenshtein distance is calculated for the contextual words and the speech recognized words and then to pick appropriate words as output.

In another work, the same author presents a multi-modal ASRU system [24]which dynamically integrates partial temporal and situational ATC context information to improve its performance. This is done either by 1) extracting word sequences which carry relevant ATC information from ASR N-best lists and then perform a context-based rescoring on the extracted ATC segments or 2) by a partial adaptation of the language model.

IV. EXPERIMENT SETUP

The experiment was conducted using CMU Sphinx as the speech recognition engine and ATCOSIM corpus as the data set.

The CMU Sphinx toolkit is a leading speech recognition toolkit with various tools used to build speech applications. CMUSphinx contains several packages for different tasks and applications.

The ATCOSIM Air Traffic Control Simulation Speech corpus is a speech database of air traffic control (ATC) operator speech, provided by Graz University of Technology (TUG) and Euro control Experimental Centre (EEC). It consists of ten hours of speech data, which were recorded during ATC real-time simulations using a close-talk headset microphone. The utterances are in English language and pronounced by ten non-native speakers. The database includes orthographic transcriptions and additional information on speakers and recording sessions.

To experiment how the CMU Sphinx speech engine works for ASR for ATC systems, two data sets from ATCOSIM corpus were tried. Both data sets were spoken by the same woman speaker. One of the data set "DataSet1" is used for adapting the engine for the ATC specific vocabulary and language model. Both the datasets have 163 audio file which are in 16-bit PCM format.

Experiments were conducted with different techniques and their impact to accuracy is analyzed. The techniques that were used,

Table 1

Speech Engine Component	Remarks
Acoustic Model	Adapted with DataSet1 and not adapted with DataSet2
Language Model	Generated with DataSet1
Lexical Model	Generated with DataSet1

The following sections details the number of ways the experiment was conducted.

First, the impact of using a domain specific Acoustic, Language and Lexical model is conducted and analyzed by varying the number of training data sets.

Second, the output is analyzed for errors which are of interest to this work. It was evident that named entities were not decoded correctly which is the problem domain for this work.

Third, the non-adapted dataset "dataset2" was decoded using the same setup and the output was analyzed. The engine performed well except once again for the named entity words.

Sample data set (manually decoded by hearing out the audio) which are used in this experiment are,

Table 2

Audio File Name	Audio File content
sm1_01_001.txt	eight one zero turn right to trasadingen
sm1_01_002.txt	lufthansa five three one eight contact Zurich one three four decimal six
sm1_01_003.txt	eight one zero contact Zurich one three decimal four
sm1_01_004.txt	sabena four eight one rhein identified
sm1_01_005.txt	transwedeone zero one rheinidentified set course trasadingen

V. TECHNIQUES EXPERIMENTED

A. *Impact of Adaptation*

The first task of the experiment is to measure the accuracy improvement when the adapted speech engine is used than the non-adapted fresh engine downloaded as such from CMU Sphinx website. This approach is different from the conventional way of deploying speech recognition systems where the engine is trained for the application specific acoustic model for maximum possible words.

It makes common sense to use the adaptation technique rather than training the engine for the specific application, as adaptation is much easier and takes less effort.

The default CMU Sphinx speech acoustic model is adapted with "DataSet1" and the Word Error Rate parameter was measured. The results are tabulated in the table below,

Table 3

Data Set Count for Adaption	Number of audios correctly decoded	Number of audios decoded incorrectly
30	57	106
100	**75**	**88**

B. *Impact of Language model*

Next, the impact of Language model to the speech engine is determined by using incremental data set to generate the language model and use the language model for speech recognition.The results were,

Table 4

Data Set Count for Language model	Number of audios correctly decoded	Number of audios decoded incorrectly
50	37	126
100	50	113
163	64	99

The output is encouraging, as increasing the data set for language model increased the overall accuracy.

C. *Impact of Language model and Adaptation*

As we understood that, increasing the data set for adaptation and language model creating improved the overall accuracy, we wanted to try the combination to see how much improvement is obtained. Hence, the experiments were con-ducted with the adapted engine and the generated language model. The results were,

Table 5

Data Set Count for the combination method	Number of audios correctly decoded	Number of audios decoded incorrectly
100	76	87
163	96	67

1.1 Impact of number of grams Language model

Another technique which was experimented is to understand the impact of number of grams in the language model.

Table 6

Number of grams	Number of audios correctly decoded	Number of audios decoded incorrectly
2	87	76
3	106	57
5	106	57

D. *Analysis of decoder errors*

The errors were analyzed based on Word Error Rate values. The errors were classified into 4 groups based on Word Error rates.

Word Error Rate < 0.5
Word Error Rate between 0.5 and 1.0
Word Error Rate between 1.0 and 1.5
Word Error Rate > 1.5

Experiment	0s	< 0.5s	0.5s to 1.0s	1.0s to 1.5s	> 1.5s
Adapted with 100 audios + Language model with 163 audios	96	57	9	1	0
Language model with 163 audios	64	82	17	0	0
3Gram	106	46	11	0	0

From the above analysis, it is understood that when a 5gram or 3-gram model is used, errors were minimum and the number of correct decoding (0s) is the maximum (~100).

E. *Experiment results with adapted data set 1*

The experiment was conducted with 163 audio files which were used for the adaptation technique and used to create the language model. The results were,

110 audio files got decoded correctly
53 files have minor errors in the decoding

When the 53 files were analyzed for errors, the observation is one or two words were not decoded in a single sentence. Only errors of interest for this paper were listed here and there were other errors which are not in scope of the proposal described in this paper.

The errors which related to contextual substitution are listed in the table below,

Table 7

Original Text	Decoded Text
EIGHT ONE ZERO TURN RIGHT TO TRASADINGEN	JET FLIGHT ONE ZERO TURN RIGHT TO TRASADINGEN
TRANSAVIA THREE EIGHT ONE ZURICH RADAR ONE THREE FOUR DECIMAL SIX	TRANSAVIA THREE EIGHT ONE THREE TWO ZERO ONE THREE FOUR DECIMAL SIX
SWISSAIR NINE THREE FIVE TWO CLIMB FLIGHT LEVEL THREE FIVE ZERO SET COURSE TO GOTIL	SWISSAIR NINE THREE FIVE TWO CLIMB FLIGHT LEVEL THREE FIVE ZERO SET COURSE TO FOUR TO
SABENA EIGHT THREE EIGHT RHEIN IDENTIFIED DIRECT TO GOTIL	SABENA EIGHT THREE EIGHT RHEIN IDENTIFIED DIRECT TO COURSE TO
SPEEDBIRD ONE FIVE SIX CONTACT RHEIN ONE TWO SEVEN THREE SEVEN	PEEDBIRD ONE FIVE SIX CONTACT RHEIN ONE TWO SEVEN THREE SEVEN
TRANSWEDE ONE ZERO ONE PROCEED DIRECT TO HOCHWALD THEREAFTER ST PREX	TRANSWEDE ONE ZERO ONE PROCEED DIRECT TO FIVE TWO SIERRA ALFA SO MAKE
MALAYSIAN TWO CONTACT RHEIN ONE TWO SEVEN THREE SEVEN	MALAYSIAN TWO CONTACT RHEIN ONE TWO SEVEN CHARLIE SEVEN
SOBELAIR TWO FIVE SEVEN RHEIN IDENTIFIED	SABENA TWO FIVE FIVE SEVEN RHEIN IDENTIFIED
OLYMPIC ONE FOUR FOUR SET COURSE DIRECT TO TANGO	OLYMPIC ONE FOUR FOUR SET GULF LEFT TO TANGO
SOBELAIR TWO FIVE FIVE SEVEN REQUEST HEADING	SO, LEFT TO FIVE FIVE SEVEN REQUEST HEADING
SOBELAIR TWO FIVE FIVE SEVEN TURN LEFT HEADING ONE ONE ZERO SEPARATION	SABENA TWO FIVE FIVE SEVEN TURN LEFT HEADING ONE ONE ZERO SEPARATION
SOBELAIR TWO FIVE FIVE SEVEN MAINTAIN HEADING CONTACT RHEIN ONE THREE TWO DECIMAL FOUR	SABENA TWO FIVE FIVE SEVEN MAINTAIN HEADING CONTACT RHEIN ONE THREE TWO DECIMAL FOUR
SABENA SEVEN EIGHT ONE SIX TURN LEFT TO GOTIL	SABENA SEVEN EIGHT ONE SIX TURN LEFT TO FOUR TO
GULF AIR ZERO ZEROSEVEN RHEIN IDENTIFIED	GUTEN SO SO SEVEN RHEIN IDENTIFIED
DELTA INDIA MIKE LIMA LIMA RHEIN RADAR IDENTIFIED	JET INDIA MIKE LIMA LIMA RHEIN RADAR IDENTIFIED
ALITALIA ONE ONE NINE ZURICH ONE THREE FOUR DECIMAL SIX	ALITALIA ONE ONE NINE UNTIL IT ONE THREE FOUR DECIMAL SIX
MIDLAND SEVEN TWO ZERO SET COURSE TO COSTA	MIDLAND SEVEN TWO ZERO SET COSTACOSTA
CONDOR TWO ONE EIGHT RHEIN RADAR IDENTIFIED	ONEONE TWO ONE EIGHT RHEIN RADAR IDENTIFIED

A closer look at the words reveal that the words in each pair matches with the other word in the pair by one or more phonetic spelling.

Summarizing the error words only, the following are the incorrect words and the corresponding correct words.These words were further classified to determine the type of the word and listed in table below,

Table 8

Actual Text	Decoded Text	Type of word
EIGHT	JET FLIGHT	Call Sign
GOTIL	FOUR TO	WayPoint
GOTIL	COURSE TO	Waypoint
SPEEDBIRD	PEEDBIRD	Call Sign
THREE	CHARLIE	Taxiway
SOBELAIR	SABENA	Call Sign
GULF AIR	GUTEN	Call Sign
ZERO ZERO	SO SO	Number
DELTA	JET	Call Sign
CONDOR	ONE	Call Sign

This method of adding another processing step after speech decoding is optimized when compared with the heavy contextual engine as cited in the related work

F. *Experiment results with a non-adapted data set 2*

The next step in the experiment is to determine the Word Error Rate for another set of 163 audio samples which are different from the data set 1. These audio samples are neither used for adaptation or for creating the language model. The only similarity of this data set 2 with data set 1 is, both are from the same speaker.

The intent of this experiment is to prove that the experiment is not biased with a trained data set, but to see how the speech engine works for a non-trained but similar data set.

In this experiment,

76 files were decoded correctly
87 files were not decoded correctly

Once again, a closer look at the analysis was done to understand which words were not decoded correctly in the error files

Actual Text	Decoded Text	Type of word
HAPAG LLOYD SIX	HAPAG LLOYD TWO	Call Sign
MALAYSIAN TWO	MALAYSIAN KILO	Call Sign
SOBELAIR TWO FIVE FIVE	TO TWO FIVE FIVE	Call Sign
LEISURE SIX ZERO	TURKISH SIX AIR	Call Sign
SET COURSE DIRECT	SET COSTA MIKE	Action
MORNING SABENA	ONE EIGHT SABENA	Call Sign
GEORGIA AIR ZERO	GEORGIA AIR SABENA	Call Sign
TWO NINE ZERO	TWO NINE ALFA FOUR	Call Sign
NETHERLANDS AIR FORCE	MIDLAND SET COURSE	Call Sign
SOBELAIR TWO FIVE	SABENA TWO FIVE	Call Sign
GERMAN AIR FORCE FIVE EIGHT FIVE	TO COSTA FIVE EIGHT FIVE	Call Sign
SOBELAIR	SO READ	Call Sign
GOTIL	FOUR TO	Waypoint

VI. CONCLUSION AND FUTURE WORK

The work which is explained above provides the analysis of the CMU Sphinx Speech recognition engine in decoding Air Traffic Controller Pilot communication using the standard English speech recognition models. It is evident that using the adaptation technique than training decodes the standard English words and does not decode correctly the domain specific words or the domain specific named entity words. This also gives the way to experiment different techniques in speech post processing to decode the domain specific words which are incorrectly decoded in this experiment. This would reduce the time taken to build the speech recognition model but get good accuracy for the problem domain.

REFERENCES

1. Van Nhan Nguyen, HaraldHolone “Possibilities, Challenges and the State of the Art of Automatic Speech Recognition in Air Traffic Control” in International Journal of Com-puter, Electrical, Automation, Control and Information Engineering Vol:9, No:8, 2015
2. Van Nhan Nguyen and HaraldHolone “N-best List Re-ranking Using Syntactic Score: A Solution for Improving Speech Recognition Accuracy in Air Traffic Control” in 2016 16th International Conference on Control, Automation and Systems (ICCAS 2016)
3. Van Nhan Nguyen and HaraldHolone “N-best List Re-ranking Using Semantic Relat-edness and Syntactic Score: An Approach for Improving Speech Recognition Accuracy in Air Traffic Control” in 016 16th International Conference on Control, Automation and Systems (ICCAS 2016)
4. Rima Shah, Dheeraj Kumar Singh “Analysis and Comparative Study on Phonetic Matching Techniques in International Journal of Computer Applications” (0975 – 8887) Volume 87 – No.9, February 2014
5. VimalP. Parmar, CK Kumbharana “Study Existing Various Phonetic Algorithms and De-signing and Development of a working model for the New Developed Algorithm and Comparison by implementing it with Existing Algorithm” in International Journal of Computer Applications (0975 – 8887) Volume 98– No.19, July 2014
6. George E. Dahl, Dong Yu, Li Dengand Alex Acero “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition” in IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012
7. Shuo Chen, Hunter Kopald, Dr. Ronald S. Chong, Dr. Yuan-Jun Wei, Zachary Levonian “Read Back Error Detection using Automatic Speech Recognition” in Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017)
8. Claudiu, MihaiGeacăr “REDUCING PILOT / ATC COMMUNICATION ERRORS USING VOICE RECOGNITION” IN 27TH INTERNATIONAL CONGRESS OF THE AERONAUTICAL SCIENCES (ICAS 2010)
9. Mira Pavlinović, Damir Boras, and IvanaFrancetić “First Steps in Designing Air Traffic Control Communication Language Technology System - Compiling Spoken Corpus of Radiotelephony

Communication" in INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS Issue 3, Volume 7, 2013
10. Oliver Ohneiser, HartmutHelmke, HeikoEhr, HejarGürlük, Michael Hössl, Thorsten Mühlhausen "Air Traffic Controller Support by Speech Recognition" in Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, Kraków, Poland 19-23 July 2014
11. VatsalaMathapati, AnjaneyKoujalagi, Naveen Kumar C "Sphinx 4 Speech Recognition in ATC" in International Journal of Advanced Engineering Research and Science (IJAERS) Vol-3, Issue-4 , April-2016]
12. YoheiFusayasu, Katsuyuki Tanaka, Tetsuya Takiguchi, YasuoAriki "Word-Error Cor-rection of Continuous Speech Recognition Based on Normalized Relevance Distance" in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelli-gence (IJCAI 2015)
13. Youssef Bassil, Mohammad Alwani "Post-Editing Error Correction Algorithm for Speech Recognition using Bing Spelling Suggestion" in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No.2, 2012
14. N USHA RANI∗ and P N GIRIJA "Error analysis to improve the speech recognition ac-curacy on Telugu language" Sa¯dhana¯ Vol. 37, Part 6, December 2012, pp. 747–761.c Indian Academy of Sciences
15. Youssef Bassil, Paul Semaan "ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset in JOURNAL OF COMPUTING", VOLUME 4, ISSUE 1, JANUARY 2012, ISSN 2151-9617
16. MinwooJeong, Sangkeun Jung, Gary Geunbae Lee "Speech Recognition Error Correction Using Maximum Entropy Language Model" in Department of Computer Science and Engineering, Pohang University of Science & Technology (POSTECH), San 31, Hyoja-Dong, Pohang, 790-784, Republic of Korea
17. KonradHofbauer, Stefan Petrik, Horst Hering "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech"
18. "Enhanced Speech Tracking of Air Traffic Control Communications", Technical University of Crete
19. José Manuel Cordero, Manuel Dorado, José Miguel de Pablo "Automated Speech Recognition in ATC Environment" in ATACCS'2012 | RESEARCH PAPERS
20. Woo KyeongSeong, Ji Hun Park, and Hong Kook Kim "Dysarthric Speech Recognition Error Correction Using Weighted Finite State Transducers Based on Context–Dependent Pronunciation Variation"
21. Anna Schmidty, Youssef Oualily, Oliver Ohneiserz, Matthias Kleinertz, Marc Schuldery, ArifKhany, HartmutHelmkez, Dietrich Klakowy "CONTEXT-BASED RECOGNITION NETWORK ADAPTATION FOR IMPROVING ON-LINE ASR IN AIR TRAFFIC CONTROL"
22. Youssef Oualil, Dietrich Klakow, GyorgySzaszak, Ajay Srinivasamurthy, HartmutHelmke, PetrMotlicek "A CONTEXT-AWARE SPEECH RECOGNITION AND UNDERSTANDING SYSTEM FOR AIR TRAFFIC CONTROL DOMAIN"
23. Youssef Oualil, Marc Schulder, HartmutHelmke, Anna Schmidt, Dietrich Klakow "Re-al-Time Integration of Dynamic Context Information for Improving Automatic Speech Recognition"
24. Youssef Oualil, Marc Schulder, HartmutHelmke, Anna Schmidt, Dietrich Klakow "Re-al-Time Integration of Dynamic Context Information for Improving Automatic Speech Recognition"