Speech control interface for Eurocontrol's LINK2000+ system

Claudiu-Mihai GEACAR*,[1], Dan-Cristian ION[1]

*Corresponding author
*,[1] Department of Aerospace Sciences, POLITEHNICA University of Bucharest
Splaiul Independenței 313, 060042, Bucharest, Romania
claudiugeacar@gmail.com

DOI: 10.13111/2066-8201.2012.4.2.5

Abstract: *This paper continues recent research of the authors, considering the use of speech recognition in air traffic control. It proposes the use of a voice control interface for Eurocontrol's LINK2000+ system, offering an alternative means to improve air transport safety and efficiency.*

Key Words: *speech recognition, air traffic control, Hidden Markov Models, control interface.*

1. INTRODUCTION

The present paper continues the theme of recent research of the authors [1], by detailing the implementation of a speech recognition system for the LINK2000+ system, in order to improve the existing air traffic control (ATC) communication procedures by reducing errors caused by channel congestion and English language proficiency differences between users.

These issues have been previously addressed and several solutions have been proposed and implemented. Out of these, the most important solutions are briefly described in the following paragraphs.

The solution to language issues was addressed by the implementation of improved radiotelephony rules, reducing the probability of confusing messages and ensuring that the phraseology is strictly defined. Some examples are:

- avoiding the use of similar words in a sequence, such as "climb to two thousand", which can easily be misunderstood as "climb two two thousand";
- avoiding ambiguous messages, such as "hold in position" and "holding position";
- reading back essential parts of the message so that the accuracy of the reception is confirmed by the transmitting user.

The first solution to channel congestion (the "saturation" of the radio frequency caused by a number of users exceeding the maximum number of users that can communicate on a given frequency) was to divide the congested sectors into smaller sectors, each having its own radio frequency, in order to reduce the number of aircraft handled by the ATC stations. This solution proved successful at first but, as air traffic volume increased, the division of sectors was unable to keep up with traffic growth, which in turn affected the safety and efficiency of ATC communications.

The current approach to channel congestion is the use of digital data links between ATC controllers and pilots for the transmission of routine ATC messages. Data links were initially implemented for transoceanic flights, where large distances made voice radio communication difficult.

After the initial implementations which had good results, Eurocontrol proposed the use of data link technology for routine messages in EU airspace, in order to decrease the load on

voice radio frequencies and pilot and controller workload. This solution is called LINK2000+ and is scheduled to become fully operational in all EU airspace by the end of 2016.

The LINK2000+ system consists of a set of standard uplink (ATC to pilot) and downlink (pilot to ATC) messages, with three functions:

- ATC communication management (ACM) – handles the handover from one ATC station to the next along the route;
- ATC clearances (ACL) – handles the routine ATC clearance messages, such as altitude and heading changes, information requests and so on;
- ATC microphone check (AMC) – an alternative means of communication in case of voice radio malfunction.

In addition to these predefined messages, the system provides the capability to send free text messages, similar to the SMS service on mobile phones. This function is not implemented (by means of voice control) in the voice recognition system described in this paper, for language model dimensions and data requirements.

The aircraft-mounted system uses the keyboard of the Flight Management System (FMS) for message parameter input and the FMS soft keys for message type selection. The ATC console system uses the console interface for message selection.

This input mode can prove rather difficult, especially in situations where a large number of messages is required (such as frequent route changes), as pilots have to keep their eyes off the instruments in order to use the LINK2000+ system, thus increasing the risk of incidents and accidents.

Fig. 1 – FMS control interface for aircrew LINK2000+ system – logon page (courtesy Eurocontrol)

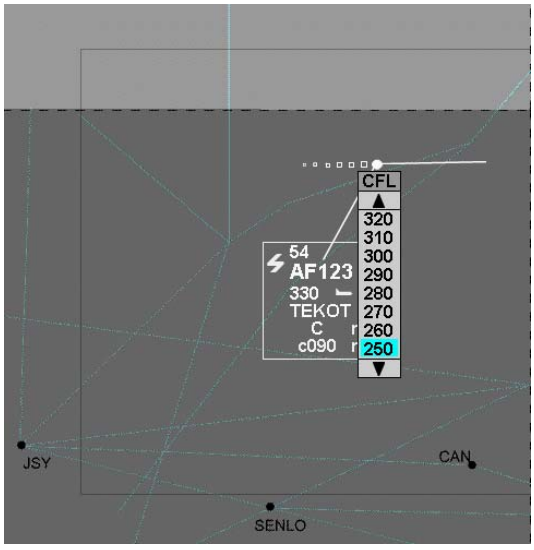

Fig. 2 – ATC control interface for the LINK2000+ system – altitude change request (courtesy Eurocontrol)

By implementing a speech recognition system at the source of messages and by using existing LINK2000+ technology and standard phraseology rules, the system aims to reduce the incidence of communication errors and to reduce pilot and controller workload for routine ATC procedures, ensuring safe and efficient communications while reducing overall operational costs.

2. SPEECH RECOGNITION AND DIGITAL COMMUNICATION TECHNOLOGIES FOR AIR TRAFFIC CONTROL

In a previous work [1], the authors introduced a concept for using voice recognition systems in air traffic control. The concept implied two possible solutions to replace standard voice radio communications in air traffic control:

- The broadcast of recognized speech, as digital messages using VHF Digital Link (VDL) networks, by the transmitting unit (aircraft or ground ATC unit) to all other receiving units in a given ATC sector;
- The use of a voice control interface for Eurocontrol's LINK2000+ system (set to be fully operative within EU airspace within 2016) for routine non-critical messages.

Both solutions make use of existing digital communication technologies (known as Controller-Pilot Data Link – CPDL) in order to reduce the channel congestion problem of present voice radio communication systems, adding further capabilities to ATC communication systems, such as message recall and recording and automatic translation of messages in the native tongue of the user.

Previous research of the authors focused on the development and evaluation of language models used by the speech recognition system, namely the development of generic language models (applicable to all user types – pilots, ATC controllers, ground personnel) and specific language models, especially tailored in order to be used by a specific category of users.

Results of that research showed that the optimum solution, based on modeling accuracy and computational and hardware load, is the implementation of user-specific speech recognition systems for each user category for the LINK2000+ system.

Furthermore, the implementation of a voice control interface for the LINK2000+ system allows for a more "natural" input, while enabling the pilots to keep their eyes on the instruments and other safety-critical equipment. This is true also for air traffic controllers, which can focus on aircraft position and collision avoidance instead of the current control interface of the system.

The following section provides general information related to Hidden Markov Models (HMMs), the leading standard in voice recognition algorithms.

3. HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

Hidden Markov Models (HMMs) were described in detail as a solution to the speech recognition by Rabiner and Juang [2]. The HMM method is a statistical modeling algorithm which addresses the issue of the non-stationary nature of speech and implies two key aspects of speech modeling:

- The analysis of spectral properties of individual sounds, with sampling periods of tens of milliseconds;
- The analysis of sound sequence characteristics related to changes in the human articulatory system, with sampling periods of hundreds of milliseconds.

The following proposition gives the formulation of the HMM problem ([2]):

Proposition. *Consider a N-state first order Markov chain. The system can be described as having one of the distinct states 1,…,N at any given discrete time t. The state of the system at time t is noted q_t. Now, the Markov chain can be described using a state transition matrix $A=[a_{ij}]$, where*

$$a_{ij} = \Pr\left(q_t = j \mid q_{t-1} = i\right), 1 \le i, j \le N, \tag{1}$$

with the constraints

$$a_{ij} \ge 0 \tag{2}$$

and

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \tag{3}$$

Assuming that q_0, the system state at t=0 is given by the initial state probability $\pi_i = \Pr\left(q_0 = i\right)$, then for any state sequence $\mathbf{q}=(q_0, q_1, \ldots, q_T)$, the probability of this sequence being generated by the Markov chain is

$$\Pr\left(\mathbf{q} \mid A, \pi\right) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \ldots a_{q_{T-1} q_T} \tag{4}$$

We will assume that $\mathbf{q}$ is not observable. Instead, we will assume that each observation $\mathbf{O}_t$ (the cepstrum – the coefficients of the Taylor series of the LPC spectrum of the speech signal) is generated by the system state q_t, $q_t \in \{1, 2, \ldots, N\}$. We also assume that the generation of $\mathbf{O}_t$ in any of the possible states i is stochastic and characterized by a probability set $B = \left\{b_i\left(\mathbf{O}_t\right)\right\}_{i=1}^{N}$, where

$$b_i\left(\mathbf{O}_t\right) = \Pr\left(\mathbf{O}_t \mid q_t = i\right). \tag{5}$$

*If the state sequence **q** that generated the observation sequence **O**=(**O**$_1$, **O**$_2$, …, **O**$_T$) is known, the probability of the sequence **O** being generated by the system is*

$$\Pr(\mathbf{O} \mid \mathbf{q}, B) = b_{q_1}(\mathbf{O}_1) b_{q_2}(\mathbf{O}_2) .. b_{q_T}(\mathbf{O}_T) \tag{6}$$

*The joint probability that **O** and **q** are generated by the system can be written as*

$$\Pr(\mathbf{O}, \mathbf{q} \mid \pi, A, B) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\mathbf{O}_t) \tag{7}$$

*The stochastic process, represented by the observation sequence **O**, is given by*

$$\Pr(\mathbf{O} \mid \pi, A, B) = \sum_{q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\mathbf{O}_t) \tag{8}$$

*which describes the probability of **O** being generated by the system without assuming the knowledge of the state sequence for which it was generated.*

The hidden Markov model is thus defined by $\lambda = (\pi, A, B)$, also known as model or model parameter set.

The HMM method requires the solving of three problems:

- The evaluation problem: Efficiently evaluate the probability that **O** is generated by λ, given λ and the observation sequence **O**;
- The estimation problem: Given **O**, estimate the parameters of λ;
- The decoding problem: Finding the most probable state sequence **q** that generated the given observation sequence **O**.

For the speech recognition system described in this paper, the speech recognition problem can be formulated as follows: given the sequence of cepstral coefficients (which describe the sound characteristics), determine the most probable phoneme sequence (based on individual phoneme models contained in the language models) that generated the unknown sound sequence. The phoneme sequence obtained in the previous step is then used to determine the word sequence (using pronunciation rules) and finally the sentences that make up the messages.

In the next section we will describe the steps necessary for building the speech recognition system.

4. THE SPEECH RECOGNITION SYSTEM PROPOSED FOR THE LINK2000+ SYSTEM

As mentioned earlier, previous research has demonstrated that the optimum solution is the development of a specific speech recognition system for each user category of the LINK2000+ system, in this case a pilot speech recognition system. The speech recognition system described in this section and proposed for LINK 2000+ is built using HTK [3] (HMM ToolKit, developed by Cambridge University Engineering Department).

The first step in building the system is the definition of the dictionary (the words that the system will be able to recognize). Based on [4] and [5], the dictionary contains 134 distinct words, covering all possible standard LINK2000+ messages (excluding automated messages, for which no user input is required and free text messages, for which the Flight Management System/FMS keyboard will be used). The messages implemented are shown below, along

with the associated LINK2000+ message ID (DM meaning Downlink Message – pilot to ATC).

DM0 WILCO
DM1 UNABLE
DM2 STANDBY
DM89 MONITORING [ATC unit designator] [fequency]
DM2 STANDBY
DM6 REQUEST [flight level]
DM9 REQUEST CLIMB TO [flight level]
DM10 REQUEST DESCENT TO [flight level]
DM22 REQUEST DIRECT TO [position]
DM65 DUE TO WEATHER
DM66 DUE TO AIRCRAFT PERFORMANCE
DM18 REQUEST [speed]
DM27 REQUEST WEATHER DEVIATION UP TO [distance] [direction] OF ROUTE
DM3 ROGER
DM4 AFFIRM
DM5 NEGATIVE
DM32 PRESENT LEVEL [flight level]
DM81 WE CAN ACCEPT [flight level] AT [time]
DM82 WE CANNOT ACCEPT [flight level]
DM106 PREFERRED LEVEL [flight level]
DM109 TOP OF DESCENT [time]

The dictionary also contains the pronunciation for each individual word, based on the English language pronunciation rules [6] and the standard ICAO pronunciation rules.

The next step is defining the system's grammar. For this, the words are divided into classes, then the messages are also divided into message classes based on the previously defined word classes. Finally, a generic message mask is defined using the message classes mentioned before, which tells the system what word sequences it should expect.

Below is the grammar definition used for the proposed system:

```
$digit = ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINER | ZERO;
$letter = ALPHA | BRAVO | CHARLIE | DELTA | ECHO | FOXTROT | GOLF | HOTEL | INDIA |
 JULIETT | KILO | LIMA | MIKE | NOVEMBER | OSCAR |PAPA | QUEBEC | ROMEO | SIERRA |
TANGO | UNIFORM | VICTOR | WHISKEY | XRAY | YANKEE | ZULU;
$single = WILCO | UNABLE | STANDBY | ROGER | AFFIRM | NEGATIVE;
$dues = DUE TO WEATHER | DUE TO AIRCRAFT PERFORMANCE;
$direction = NORTH | SOUTH | EAST | WEST;
$monitor = MONITORING ( { $digit } { $letter } { $digit } < $digit > [ DECIMAL < $digit > ] );
$requestalt = REQUEST < $digit > [ $dues ];
$requestclb = REQUEST CLIMB TO < $digit > [ $dues ];
$requestdesc = REQUEST DESCENT TO < $digit > [ $dues ];
$requestdir = REQUEST DIRECT TO { $letter } { $digit } [ $dues ];
$requestspd = REQUEST < $digit > KNOTS [ $dues ];
$requestwdev = REQUEST WEATHER DEVIATION UP TO < $digit > $direction OF ROUTE;
$presentlev = PRESENT LEVEL < $digit >;
$acceptlev = WE CAN ACCEPT < $digit > AT < $digit >;
$noacceptlev = WE CANNOT ACCEPT < $digit > [ $dues ];
$preflvl = PREFERRED LEVEL < $digit >;
```

```
$topdesc = TOP OF DESCENT < $digit >;
$message =  $single | $monitor | $requestalt | $requestclb | $requestdesc | $requestdir | $requestspd | $requestwdev | $presentlev | $acceptlev | $noacceptlev | $preflvl | $topdesc;
(SENT-START $message SENT-END)
```

The last line is the generic message mask, while the lines above contain word and message class definitions.

The third step is the generation and recording of training data. The data consists of sample sentences, conforming to the previously defined grammar and pronunciation rules. In order to obtain a satisfactory statistical distribution of words, the training data for the speech recognition system consists of 100 distinct sentences, covering all message types that the system can handle and all words in the dictionary.

Next, a transcription of the sentences, called a *master label file* is created. This file contains a list of the filenames associated with each sentence and the corresponding word-level transcription. Separately, a phoneme (individual phonetic part of each word)-level transcription is generated, based on the associated pronunciation of each word contained in the dictionary.

An example of a word-level transcription of the sentence "Monitoring EFGH 567.89" ("Monitoring Echo Foxtrot Golf Hotel five six seven decimal eight niner") is:

```
"*/S008.lab"
MONITORING
ECHO
FOXTROT
GOLF
HOTEL
FIVE
SIX
SEVEN
DECIMAL
EIGHT
NINER
.
```

The training data is also processed in order to obtain the cepstral coefficients for each individual sentence, which will be used in the next step, which is the generation of individual HMMs for each phoneme.

The grammar definitions, dictionary, training data and its associated phoneme-level transcription are used to generate the individual model parameters. For this, a prototype HMM is defined. This prototype contains five states, out of which only three are emitting (the entry and exit states are considered by default to be non-emitting). Using the Baum-Welch algorithm [7], the estimation problem of the HMM algorithm is solved, resulting in a set of parameters defining the model for each phoneme in the system's dictionary.

In order to handle the different energy levels of the pauses between words, a "virtual" phoneme noted "sp" (short pause) is also created. This has a reduced model, which is based on the model "sil" (silence) used for the beginning and end of the sentences. The "sp" model contains three states (with only one emitting), the second (middle) one being linked to the second emitting state of the "sil" model. The training data is then used to re-estimate the model parameters.

The next step in the building of the speech recognition system is the generation of tri-phone (groups of three phonemes) models. For this, all possible tri-phone combinations are generated. Next, the individual model for each tri-phone is generated by copying the model

for the middle phoneme and then linking it to the other two models using the transition matrix coefficients. As a final step, the model parameters are re-estimated using the training data. In order to reduce the size of the overall model, acoustically-similar states of the models are linked by creating so-called phonetic classes, based on acoustical similarity criteria (like the position of a certain phoneme in the tri-phone).

The final step is the re-estimation of the new model parameters using the training data, completing the voice recognition system.

5. TEST RESULTS AND FINAL REMARKS

The speech recognition system is tested using a set of recorded sentences, similar (in terms of grammar and pronunciation rules) to the ones used for the training process.

The test set contains a number of 20 sentences, which are recorded and processed similar to the training data (generating the cepstral coefficients and the word-level transcriptions).

The speech recognition system uses the Viterbi algorithm [8] in order to determine the tri-phone sequence which has the highest probability to generate the cepstral coefficient vectors corresponding to the test data. Based on this tri-phone sequence, the system generates its own word-level transcription, which is then compared to the test data transcription.

The resulting accuracy was 84%, which is satisfactory given the reduced amount of training data, but low for use in aviation applications.

This result created the need for further fine-tuning of the speech recognition system and the evaluation of the impact of its parameters on the overall recognition accuracy.

The first parameter that was changed was the number of states. Given the fact that the initial speech recognition system was built based on systems with similar complexity, using a five-state model, it was rebuilt (based on the same algorithm and training data) using a seven-state model. The resulting accuracy, using the same test data, was 90%, which is a major impact on recognition performance and is similar to the accuracy of most commercially-available speech recognition systems. Further research using different numbers of states, resulted in no improvement in terms of accuracy, which lead to the conclusion that the seven-state model is optimum for this particular application.

Also, the analysis of the system-generated transcription of the training data showed that the lowest accuracy levels were obtained for the words denoting letters (such as “ALPHA”, “BRAVO” and so on). This was found to be consistent with the reduced occurrence of such words in the training data, resulting in statistical imbalance of the data and poor modeling performance.

In order to correct this deficiency, 36 additional sentences were added to the training data, sentences containing messages which required letters (such as flight paths and points).

After the models were re-generated, the accuracy obtained was 94%, which is a more than satisfactory result, given the fact that the training of the system would take a very little amount of time, making it feasible for real-life implementations. Still, further research is required for the implementation of an error-handling procedure (based, for example, on a figure-of-merit approach, in which sentences which have a recognition probability less than a given threshold are rejected and the user is prompted to repeat the sentence.

The above results validate the initial concept [1] and prove that speech recognition is a feasible solution for replacing the classic voice radio communication system and also improve the input method for the LINK2000+ system.

The independence between the control interface and the message transmission system (given by the language model) allows further improvement of the system, which will make the object of future research. One direction for improvement is the further increase of the voice control interface by means of implementing of a mother tongue language model. This approach is facilitated by the fact that the LINK2000+ system contains a standard set of messages with a standard (and strict phraseology) which can be easily implemented with a language model. Also, given the fact that messages are identified by the message type identifier and the message parameters, it is easy to create a language model and a speech recognition system that can return these parameters independent of the input language, thus enabling the user to speak in his mother tongue and eliminating the language issues completely.

Furthermore, for a more "natural" feel, a speech synthesis system can be implemented for the delivery of the messages, allowing a seamless integration of the LINK2000+ system with the current voice radio communication system.

Of course, the system accuracy remains an open issue, and a real-life implementation requires an error-handling procedure. Such a procedure could be based on a Figure-Of-Merit (FOM) message validation procedure, in which messages for which recognition accuracy is below a specified threshold are automatically rejected and the user is required to re-input the message or use an alternative input method (such as free text messages or FMS keyboard).

ACKNOWLEDGEMENT

The work has been funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labor, Family and Social Protection through the Financial Agreement POSDRU/6/1.5/S/19.

REFERENCES

[1] Claudiu-Mihai Geacăr, Reducing pilot / ATC communication errors using voice recognition, *Proceedings of ICAS 2010*, 2010.
[2] B. H. Juang; L. R. Rabiner, Hidden Markov Models for Speech Recognition, *Technometrics,* vol **33**, No. 3, Aug., pp. 251-272, 1991.
[3] *** The HTK Book, User Guide, 2009.
[4] *** ICAO Doc 9432.
[5] *** ATC Data Link Operational Guidance for LINK 2000+ Services.
[6].*** Carnegie Mellon University Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
[7] L. E. Baum, An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes, *Inequalities,* Vol. **3**, pp. 1-8, 1972.
[8] G. D. Forney, The Viterbi Algorithm, *Proceedings of the IEEE,* **61**, pp. 268-278, 1973.