University of Leeds
SCHOOL OF COMPUTER STUDIES
RESEARCH REPORT SERIES
Report 95.20

Developing a Corpus -Based Grammar Model Within a Continuous Commercial Speech Recognition Package

by

Eric Atwell, Gavin Churcher & Clive Souter
Division of Artificial Intelligence

June 1995

Abstract

This paper is derived from experiments with a commercial 'off-the-shelf' continuous speech recognition system, applied to the apparently restricted domain of Air Traffic Control (ATC) for light aircraft. The system is required to transcribe key sub-phrases in a transmission by the ATC to a particular aircraft, the commercial speech recognition system providing the main recognition component. After the development of a corpus of transmissions, it was realised that key information is often interspersed with unconstrained English. Initial attempts focused on using a wildcard mechanism for the non-key sub-phrases. The mechanism, however, proved to be valuable only in simplistic grammars due to its overgenerative nature. The speech recognition system showed us that whilst useful mechanisms are provided, such as the wildcard mechanism, they tend to make over-simplistic assumptions about English grammar and dialogue structure.

1 Introduction

This paper is concerned with the initial findings of investigations into the use of a commercial 'off-the-shelf' continuous speech recognition system for the development of an Air Traffic Control (ATC) speech recogniser for light aircraft. The paper demonstrates the properties, desirable or otherwise, of the speech recognition system and of the intended application.

The application we wish to develop involves a speech recognition system which would recognise key elements of a transmission by the ATC to an aircraft. The recognition would take place in the ATC Tower, where an application would listen to the controller as he/she speaks. The information transcribed by the system would then be digitally transmitted to an aircraft and then be displayed to the pilot. Cockpit display of ATC transmissions would then ease the pressure on the pilot when relaying the information back to the ATC and when making appropriate changes to the aircraft's controls. For details of key information and examples of transmissions, see section 2.

Since, at the start of the investigation we did not know the true requirements of a speech recognition device, we chose the commercially available Speech Systems Incorporated Phonetic Engine 500 (SSI PE500) speech recognition development kit (SDK). The SSI PE500 aims to provide for continuous, speaker-independent speech recognition, with a 400,000-word vocabulary. The system provided two generic speaker models: American male and American female. It is not yet clear how the performance is impaired when using these models for a British speaking person. The development of such speaker independent models is extensive and must be carried out under contract by Speech Systems Inc. At the moment the prospect of a British set of speaker models is unlikely given the costs and resources involved.

For applications using specialised words, the vocabulary can be extended via a generalised phonetic transcription algorithm. In order to reduce the search problem and achieve reasonable recognition accuracy, the system requires the application developer to constrain possible input to the system by formulating a strict context-free grammar of allowable utterances. For greater efficiency, a regular grammar could be used instead.

The use of the SSI PE500 is not the only approach being considered. There are a number of other off-the-shelf speech recognition packages available, including Dragon and Philips. A more research orientated approach would be using systems such as Carnegie Mellon University's Sphinx II and Phoenix [Huang et al., 1993, Isaar and Ward 94], BBN's HARC [Bates et al. 92, Bobrow et al. 92] and Cambridge University's CU-HTK system [Woodland et al. 94a, Woodland et al. 94b].

2 Corpus Collection

In order to construct a grammar suitable for the SSI PE500, several sources were analysed. The primary source was the British Radiotelephony Manual [RTF CAP413]. This establishes a standard phraseology for the communication between ATC and pilots and explains the format and interaction structure for set tasks, such as changing to another radio frequency.

A corpus of ATC dialogue transcripts from Leeds-Bradford Airport (LBA) was also examined. It consists of a series of transmissions between the ATC and the numerous aircraft within the airspace over a reasonably short period. The corpus consists of approximately 430 sequential utterances, where an utterance is one transmission made by either the ATC or a pilot. The corpus was constructed by recording the transmissions from an Air band receiver on a particular frequency using a voice actuated cassette recorder. Using this technique, we recorded the equivalent of 45 minutes of speech, which is a greater amount in real time. From the recordings the transmissions were transcribed. Due to interference problems and the use of voice actuation there were some parts of the tape where the transmission was not clear.

The key information with which we are concerned falls into five main categories:

1. Instructions to the pilot to change his/her altitude. Information would be an altitude in terms of either a height in feet or a *flight level*, a system for abbreviating regularly used altitudes.

2. Pressure settings for observed pressure (QFE) and the altimeter/subscale (QNH). Pressure settings are measured in millibars.

3. Secondary Surveillance Radar settings for aircraft transponders, indicated by a *squawk* values. These settings allow the ATC to identify an aircraft by radar.

4. Instructions to the pilot to change to another radio frequency.

5. Instructions to the pilot to change his/her heading, a setting measured in magnetic degrees.

Although the manual defines a precise and rigid format for individual utterances and the structure of interaction, the LBA corpus shows that pilots and controllers sporadically intersperse the dialogue with a variety of unrestricted English, such as greetings, interjections and other phrases. So, whilst an utterance may contain key information, it is surrounded not only by other information in which we have no interest, but also relatively free English sub-phrases. Consider the following example utterances taken from the corpus. They show key information on its own, combined with other key information, with non-key information, and interspersed with unrestricted, free English.

1. Utterance contains one piece of key information (ignoring callsign)
   "GFR, squawk zero four one five"
   Meaning: pilot of aircraft registration G-??FR, change transponder code to 0415

2. Utterance contains two pieces of key information combined (ignoring callsign)
   "UK 655, turn right five heading one three five, descend altitude one five hundred feet"
   Meaning: pilot of flight UK 655, change heading to 135 and descend to height 1500 feet

3. Utterance contains both key information and non-key information
   "GFY, turn left heading one three five, radar vectors ILS approach runway three two, information delta, QNH one zero one five, QFE if required is nine nine one millibars"
   Meaning: there are five pieces of information contained in the utterance, only three of which fall into the above categories of key information. The pilot is requested to change heading to 135, and is informed that the pressure settings for QNH and QFE are 1015 and 991 millibars, respectively.

4. Utterance contains key information but is interspersed with unconstrained English

   "701, thanks, I believe you're just passing north west of Leeds by one three miles, I've nothing further for you, you may want to call Linton on one two nine decimal one five, and they're located in the vale of York"

   Meaning: there are three pieces of information, in which we are only interested in one, the instruction to the pilot to change radio frequency.

3 Initial Attempts At Grammar Development

As is often the case with 'sub-languages', the language in our domain is not as clearly constrained as one might assume; and it is difficult to deal with the unwanted 'noise-phrases' effectively within the context-free formalism required by the speech recognition package.

An initial attempt of developing a grammar involved the use of keyword spotting within an utterance. The Speech Development Kit claimed that it was possible to use keywords within an utterance, and to effectively ignore the noisy sub-phrases within the utterance. It did this by using three mechanisms, the first of which divided words in the lexicon into two categories. One category contained the keywords which the system was interested in, the other contained all of the other words which were not required. In effect, a 'wildcard' category was generated which could absorb unwanted words.

The second mechanism allowed the wildcard category to be used iteriaritively, in effect making multiple copies of itself. The third mechanism incorporated optionality, enabling the wildcard category to be used zero or more times.

The use of just one wildcard category containing all of the undesired words is not versatile enough for a large grammar. The approach of using several wildcard categories enables the grammar to have greater expressibility, and constrains each category to a smaller selection of throw-away words. The initial grammar used wildcards between important sub-phrases. The wildcard categories were generated by examining key words in context (KWIC) concordances within the corpus.

Consider the example grammar rule below which makes use of keywords and wildcards. It shows a simplified version of the initial top level rule which includes the wildcards. Here optionality is represented by the ordinary brackets, "(" and ")", iteration is represented by braces with asterisks, "{*" and "*}", and disjunction by the symbol "|". Letters in uppercase indicate a non-terminal tag.

```
U ->    (CALLSIGN)
        {  ({* BH *})  HEADING  ({* AH *})  |  ({* BA *})  ALTITUDE  ({* AA *})  }
```

The example shows that an utterance can be made up from an optional callsign and either an instruction to change heading, or an instruction to change altitude. The heading instruction can be preceded by a number of words defined in the wildcard category BH, and succeeded by a number of words defined in the wildcard category AH. Similarly with the altitude instruction.

The initial grammar, of which the above rule is only a small part, consisted of a total of 37 tags, and 17 rules; 6 of the tags present used the wildcard mechanism.

The choice of words in the wildcard category is critical to the performance of the system. Mistakes are made by either 'misses' or 'false alarms'. A 'miss' is where the wildcard category is mistaken for a keyword in the utterance. A 'false alarm' is where a keyword is found in the place of a non-keyword, and should have been ignored. Errors such as these occur because of the phonetic similarity of the words within either category. The recommended method of improving accuracy is to edit either category and, in the case of misses, the word which is phonetically similar to a keyword should be

removed from the wildcard category. The method assumes that the wildcard category itself is sufficiently diverse to overcome minor changes of a few words.

However, using a mechanism such as the wildcard presents two problems. The first is the overgenerative nature of the category which could result in a grammar which is insufficiently constrained to yield acceptable recognition accuracy. The final system is expected to have a high accuracy level, although the inaccuracies of the speech recognition unit could be compensated for by an external natural language unit incorporating contextual knowledge (see section 5). The SDK has in-built capabilities for testing the speech recognition unit's accuracy with a particular grammar by collecting speech samples from different people and situations. Accuracy levels for grammars, which contain different uses of the wildcard mechanism, can then be assessed. We intend to use this facility to determine the improvement in recognition accuracy as grammar size and complexity increases.

The second problem as applicable to this grammar is the sheer diversity of the keywords and the overlap between what is in some cases to be ignored in the wildcard category, and what should be considered as a keyword. For example, the keyword 'descend' could trigger off the recognition of an instruction to a pilot to change his/her altitude. However, the word 'descend' could also occur as a wildcard word since it also exists in both non-key information and the unrestricted English used by the ATC. The following example shows two utterances made in the corpus by the ATC, both of which contain the word "descend", the first of which occurs in a key information phrase, the second in a non-key information phrase.

> "zero three six four, *descend* at altitude two thousand, q n h one zero one five."
>
> "roger, eight zero one, what altitude are you looking to *descend* to?"

The mechanism is quite unrealistic when using a large grammar, and can lead to syntactic ambiguity, thus compounding the inaccuracy of the recogniser. The noise-phrases are relatively rare compared to the legal phrases at any point in an utterance. The wildcard mechanism currently gives equal weight to all alternatives at a choice point. An ideal way forward would be a modification of the package to allow grammar rules and lexical entries to be augmented with corpus-derived relative probabilities. The rules could then be used in probabilistic parsing to favour legal and more common recognition candidates. Low probability noise-phrases would then only be selected when there is a clear acoustic match.

4 The SSI PE500 Speech Development Kit

The SSI PE500 is effectively a 'black box', where an utterance is presented to the system and decoded according to a pre-compiled grammar. The system returns an N-best list of decoded utterances, ordered by an overall score which reflects how well the words match the actual utterance. Unfortunately, it is not possible to modify the grammar constraint formalism to directly include the probabilities obtained from a corpus.

The software documentation suggests that users are free to use the package to generate N-best recognition candidate utterances, and then pass these on to a linguistic constraint post-processor. The post-processor can then apply any language model suitable to the list. This may not be as effectual as it initially sounds since it is not *optimal* to use corpus-derived probabilities to *derive* the N-best utterance transcripts, only to re-order the N-best candidates left by the non-probabilistic context-free grammar.

The system allows for another variation on the standard context-free grammar constraint model: the recogniser can hold several variant grammars, known as 'contexts', which can be applied to the same utterance. An application can then switch between the different contexts as the overall dialogue structure passes through different phases. Unfortunately, once more it is not easy to make effective use of this facility. It only makes sense to store and switch between alternative grammars if a dialogue can be confidently partitioned into distinct dialogue segments, each with a notably different grammar.

Our LBA Corpus has been edited to facilitate this analysis and has been manually phrase-tagged with around 50 semantic/functional labels. For a complete listing of labels and examples see Appendix 1. The creation of discourse and semantic functional phrase tags is intended to enhance the existing context-free grammar in order that it might be partitioned to take advantage of the PE500's aforementioned 'context' facility. The utterances have been grouped into dialogues between the ATC and a particular pilot. The controller may be interacting with several pilots in parallel, in which case each pilot-controller 'thread' constitutes a separate dialogue. This training set should provide evidence of habitual repeated patterns or structures within dialogues, if they exist. For example, consider the two interactions between the pilot of aircraft G-AJCT and the ATC, below. The ATC's utterance ("A:") has been tagged in terms of semantic/functional labels. The number in brackets preceding the utterance is the transmission index.

(166) P: leeds approach good morning golf alpha juliet charlie tango is passing 1400 feet on the heading of 240

(167) A: [CALLSIGN charlie tango CALLSIGN] [GREET leeds good morning GREET] [INFO_ID you are identified INFO_ID] [MAN_HEAD continue heading two four zero MAN_HEAD]

(209) A: [CALLSIGN charlie tango CALLSIGN] [INFO_RADAR radar service terminates INFO_RADAR] [ALT_SQUAWK squawk seven thousand ALT_SQUAWK] [ALT_FREQ and continue now with east midlands approach frequency one one nine decimal six five ALT_FREQ] [BYE bye bye BYE]

(210) P: east midlands 119 decimal 65. bye for now

The functional labelling of utterances in the corpus will hopefully shed light on whether this technique will be useful.

5 The Use of Contextual Knowledge

The use of a natural language component to constrain the output of the system could increase the system's recognition performance. In this domain, there is also a wide range of contextual knowledge which could be incorporated into the system, either by means of a database containing information applicable to the local area around the ATC, or by controlling the speech recognition unit itself. The contextual knowledge which could be applicable includes the following:

1. Current callsigns being used in airspace.
2. Current transponder settings (squawks) being used by aircraft.
3. Current pressure settings of the local area, etc.
4. Regional geographical landmarks.
5. Transponder code ranges used at LBA.
6. Radio frequencies used at or around LBA.
7. Runway identifiers used at LBA.

The first three items contain information which exists for differing periods of time. For example, the callsigns currently being used exist only for the duration that the pilot is in LBA airspace. The remainder of the information is local to LBA, itself.

As an example of how this information may be used, consider the transponder or 'squawk' codes which range in value from 0400 to 0420, in octal and that only one aircraft in LBA airspace can have a particular code. This information can assist the choice of the correct code.

6 Conclusion

In conclusion, the lesson from "English Corpus Linguistics" is that a sophisticated continuous speech recognition system can be let down by over-simplistic assumptions about English grammar and dialogue structure. It is interesting to note that SSI have decided to incorporate probabilistic language models at the grammar formalism level, and are currently marrying their technology with that of another company for this effect. While acoustic pattern-matching has made great advances to the stage where sophisticated continuous speech recognition packages are available 'off the shelf', there is still a need for further research into higher-level linguistic models of grammar and dialogue structure for practical enterprises such as ours.

ACKNOWLEDGEMENT:

We thank Tony Denson and Visionair International Ltd. for financial support for this research, including provision of the Speech Systems Incorporated Phonetic Engine 500 speech recognition application development kit.

Bibliography

[Bates et al. 92] M Bates, R Bobrow, P Fung, R Ingria, F Kubala, J Makhoul, L Nguyen, R Schwarz, D Stallard. “Design and Performance of HARC, the BBN Spoken Language Understanding System”, in Proceedings of the International Conference on Spoken Language Processing (ICSLP), pages 241-244, October 1992, Alberta, Canada.

[Bobrow et al. 92] R Bobrow, R Ingria, D Stallard. “Syntactic/Semantic Coupling in the BBN Delphi System”, in Proceedings of the DARPA Speech and Natural Language Workshop, pages 311-315, Morgan Kaufmann, February 1992.

[Huang et al. 93] X Huang, F Alleva, H Hon, M Hwang, K Lee, R Rosenfeld. “The SPHINX-II Speech Recognition System: An Overview”, Computer Speech and Language, 2: 137-148, 1993.

[Isaar and Ward 94] S Isaar and W Ward. “Flexible Parsing: CMU’s Approach to Spoken Language Understanding”, in Proceedings of the ARPA Spoken Language Systems Technology Workshop, March 1994.

[RTF CAP413] Radiotelephony Manual (CAP 413), Civil Aviation Authority, London, 1992.

[Woodland et al. 94a] P Woodland, J Odell, V Valtchev, S Young. “Large vocabulary Continuous Speech Recognition Using HTK”, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1994, Adelaide.

[Woodland et al., 1994b] P Woodland, J Odell, V Valtchev, S Young. “The HTK Large Vocabulary Recognition System: An Overview”, in Proceedings of the ARPA Spoken Language Systems Technology Workshop, March 1994.

APPENDIX 1:

List of Tags, number of occurances and example of most frequent

Tag: AFFIRM Total number of occurances: 40
roger

Tag: ALT_CLIMB Total number of occurances: 3
climb flight level nine zero

Tag: ALT_DESC Total number of occurances: 37
descend altitude four thousand feet

Tag: ALT_FREQ Total number of occurances: 22
contact the tower one two zero decimal three

Tag: ALT_HEAD Total number of occurances: 33
right heading two nine zero

Tag: ALT_SQUAWK Total number of occurances: 11
squawk seven thousand

Tag: BREAK Total number of occurances: 2
break

Tag: BYE Total number of occurances: 18
bye bye

Tag: CALLSIGN Total number of occurances: 190
u k six five five

Tag: CORRECTION Total number of occurances: 1
correction

Tag: GREET Total number of occurances: 27
leeds good morning

Tag: INFO_? Total number of occurances: 2
close the localiser from the right

Tag: INFO_APPR Total number of occurances: 24
further descent with the i l s

Tag: INFO_CLEAR Total number of occurances: 9
but you’re clear to the leeds zone boundary only altitude three thousand five hundred feet

Tag: INFO_CURRENT Total number of occurances: 15
information charlie current

Tag: INFO_END Total number of occurances: 4
i’ve nothing further for you

Tag: INFO_ID Total number of occurances: 11
you are identified

Tag: INFO_ILS Total number of occurances: 2

descent with the i l s

Tag: INFO_LAND Total number of occurances: 15
approximately two seven track miles to touchdown for a five mile final three two

Tag: INFO_LOC Total number of occurances: 1
and they're located in the vale of york

Tag: INFO_POS Total number of occurances: 10
i believe you're just passing north west of leeds by one three miles

Tag: INFO_QFE Total number of occurances: 20
q f e nine nine one millibars

Tag: INFO_QNH Total number of occurances: 26
leeds q n h one zero one five

Tag: INFO_RADAR Total number of occurances: 13
radar vectors i l s approach runway three two

Tag: INFO_RW Total number of occurances: 2
runway one four is available

Tag: INFO_STANDBY Total number of occurances: 1
i'll come back with the landing runway to you shortly

Tag: INFO_STATUS Total number of occurances: 7
entering the leeds zone shortly

Tag: INFO_TRAFFIC Total number of occurances: 14
be advised the circuit is active left hand runway three two

Tag: MAN_DISTTotal number of occurances: 1
continue until about a four or five mile further

Tag: MAN_HEAD Total number of occurances: 4
and continue on that heading on that heading

Tag: MAN_HEIGHT Total number of occurances: 4
maintain

Tag: MAN_SPEED Total number of occurances: 1
keep your speed up

Tag: OPTIONAL_ASSERT Total number of occurances: 4
if you wish

Tag: OPTIONAL_DESC Total number of occurances: 1
descend at your discretion

Tag: OPTIONAL_HOVER Total number of occurances: 1
if it's okay with leeming coordinate with them

Tag: REQ_ALTITUDE Total number of occurances: 4
verify your level

Tag: REQ_CHECK Total number of occurances: 1
it might be worth checking your giro against your compass

Tag: REQ_CONFIRM Total number of occurances: 19
confirm established

Tag: REQ_DEST Total number of occurances: 1
what is your destination

Tag: REQ_DETAILS Total number of occurances: 5
pass your details

Tag: REQ_HEAD Total number of occurances: 5
report your heading

Tag: REQ_LAND Total number of occurances: 1
what reg final would you like

Tag: REQ_RANGE Total number of occurances: 2
what range final would you like

Tag: REQ_REPORT Total number of occurances: 24
report established

Tag: REQ_SPEED Total number of occurances: 2
report your speed

Tag: REQ_SQUAWK Total number of occurances: 3
just confirm your squawk at the moment

Tag: RESTRICT_CLEAR Total number of occurances: 3
no further altitude restriction

Tag: RESTRICT_HEAD Total number of occurances: 2
proceed to the southern airfield boundary only

Tag: RESTRICT_HEIGHT Total number of occurances: 7
and transit the leeds zone vfr not above altitude two thousand five hundred feet

Tag: RESTRICT_POS Total number of occurances: 5
to close the localiser from the right

Tag: STANDBY Total number of occurances: 6
i'll keep you advised