**Proceedings of the Institute of Acoustics**

# DATA-SUPPORTED CASE FOR THE EXTENDED COVERAGE OF REPAIRS IN THE RECOGNITION OF NATURAL SPEECH

**K Johnson, R J Collingham & R Garigliano**

**Laboratory for Natural Language Engineering**
**Department of Computer Science, University of Durham**

## 1. Introduction

The advancement of Automatic Speech Recognition (ASR) systems over the past few years, as seen from the increased complexity of the developing speech corpora, from the recognition of isolated words to recognising continuous, natural speech has led to a number of specific problems. In the move from isolated speech (word recognition) to continuous read speech the extra requirements of a speech recognition system included identifying word boundaries and sentence boundaries. The move from read speech to spontaneous speech has led to a further problem with the grammatical nature of the speech. Spontaneous speech is known to have an, what can be classed as, ungrammatical nature. Speech repairs, repetitions or false starts (repairs will be the term used to class all of these speech problems) are something that are common in spontaneous speech but do not appear regularly in other forms of speech data, such as isolated words or even read speech. [4] show that the extension of speech recognition systems from read speech to spontaneous speech can "significantly degrade recognition performance". It is known [1] [11] that repairs are a definite problem found in spontaneous natural speech and it is expected that repairs do not exactly follow the normal constructs of natural speech therefore any passage containing a repair would not be recognised without some form of repair processing.

The work described in this paper is part of a larger analysis looking into the effects of repairs on the performance of an ASR system being developed at the University of Durham. This work is of an exploratory nature which is looking into the identification of speech repairs and on to their solution. Work already presented in the fields of linguistics and speech research have identified possible solutions but those that are more than theoretical can only cover some of the repairs that appear in spontaneous speech. This paper looks at developing these methods by using two further matching procedures in an attempt to extend the coverage of repairs.

## 2. Current System.

The aim of the system being developed at the University of Durham [5] [6] (AURAID - A University speech Recognition AID) is to automatically recognise the spoken words of a lecturer as an aid to hearing impaired students undertaking a normal university course. The very nature of the problem being tackled causes a great number of design and implementation constraints. The system must be speaker independent to allow any lecturer to perform in any lecture theatre at any time thus not restricting the hard of hearing student in their selection of courses. The vocabulary of the system must be large to allow the speaker to say exactly what they mean, in a way in which they believe the true meaning will be conveyed to the students. Therefore the speaker must not be told how to say things or restricted in their word or structure/construct usage. These requirements give rise to many problems with spontaneous speech repairs being a major problem.

The system, at present, is only part of a final product. The front end is a simulated phoneme recogniser with a corruption rate of up to 25%. The corruptions are split between insertions, substitutions and deletions based on realistic phoneme recognition data. The system uses the first level of a two level dynamic programming algorithm to build a word lattice and a pyramid beam search to determine sentence hypotheses occurring in the lattice. Word frequency information and anti-grammar (AG) rules are then used to select the most appropriate hypothesis. The word frequency processing uses information from the Oxford Advanced Learners Dictionary [12] to categorise words as common, normal or rare. The AG rules [5] are a set of rules which indicate what cannot be said or more to the point how things are not normally said during spontaneous natural speech. It is a set of rules which decreases the likelihood of hypotheses being selected and therefore removes ill-formed hypotheses from a list of potential hypotheses.

## 3. Current Research.

There are a number of theories on how problems of spontaneous speech can be overcome and most rely on acoustic and prosodic knowledge. Speech, and in particular spontaneous speech, contains various pieces of information in its acoustic wave forms. What this actually means is as yet unknown and can only be speculated upon. But both the fields of linguistics [9] [2] and speech [1] [13] [14] have identified that the use of prosody would be beneficial in overcoming speech repairs. [9] identify the possible use of structure and semantic information in dealing with repairs while [7] identify the possibility of predicting the structure of the next input based on the previous input and history information. [15] calculate possible boundary locations based on the length of previous boundaries using the theory that "...consecutive phrases have roughly equal length". [1] identify that acoustics is important, and use it in the latter stages of their processing, but "...make no assumption about the existence of an explicit edit signal ... as a reliable edit signal has yet to be found" and therefore rely heavily on pattern matching techniques for identifying speech repairs.

Though there is a consensus in the requirement of some method for overcoming speech repairs only [1] make inroads into overcoming repairs.

## 4. Data Analysis.

A number of lectures performed at the University of Durham were recorded and transcribed to include all of the disfluencies of speech including the "hums" and "ahs" as well as pauses, which were categorised into long and short pauses. This forms the basis for the data used in all of the analyses carried out at at the University of Durham [8] [10]. The data used in the analysis presented in this paper was taken from a single lecture on 'Software Engineering' given as the first introductory lecture for second year students. The lecture contained 4903 words and 382 sentences or part sentences with an average of 12.84 words per-sentence. 119

(31.15%) of the sentences contained repairs while 37.79% of sentences of more than nine words long contained repairs. This is similar to the 34% given by [11] but much greater than the 10% given by [1]. Of the 119 sentences containing repairs 101 contained a single repair, 17 contained two repairs and 1 sentence contained three repairs. The repairs themselves do not all follow the same pattern and an analysis of these repairs showed that there were five repair types with one type being split into four different sub-groups.

- **Phonetic.** This is where a sound is repeated. That sound could be a part word or whole word.
- **Grammatical.** This is where a word is inserted, deleted or corrected to change the grammar of the sentence but without changing the meaning.
- **Semantic.** This is where a word is inserted, deleted or corrected to change the actual meaning of what is being said.
- **Sentence Abortion.** This is where a clear sentence has been aborted and another started without the original sentence being completed (i.e it is only a part sentence). Sentence abortions can be one of four types.
  - (a) Un-required information. The details of the aborted sentence are not required within the dialogue and can therefore be ignored.
  - (b) With required information. The information from the part sentence is required to be able to understand latter (probably the following) sentences.
  - (c) Major repair. This is where it is possible, with major modifications to add (join) the part sentence to the following sentence.
  - (d) Completed sentence. This is where it is possible to add (join) the part sentence to a latter (not the following) sentence. This could be seen as sentence Divergence.
- **Dialogue repair.** This is where the meaning of what is said is incorrect but the actual meaning is obvious and no repair by the speaker is made. Rather the repair is made by the listener. This could be an incorrect word, such as "your" rather than "you've", or the insertion of an un-required word in the middle of a sentence or more typically double negatives which are continually corrected by human listeners.

Repair types 1, 2 and 3 can be classed as normal repairs and types 4 and 5 are further repair types which can cause problems when processing spontaneous speech. It is possible for one repair to be classed as a number of repair types such as a type 3 repair could easily be a type 4 repair, if it was at the beginning of a sentence. In this case type 3 took priority as the repair would be classed as a type that is normally seen as a repair. The priorities used in classifying the repairs were, types 1, 2 and 3 first, as they are the normal types of repairs, followed by type 4 and finally type 5. Using this arrangement there were 53 (38.4%) type 1 repairs, 25 (18.12%) type 2 repairs, 26 (18.84%) type three repairs, 25 (18.12%) type 4 repairs and 9 (6.52%) type five repairs.

## 5. Current Coverage.

Due to the effect of repairs in spontaneous speech a system using no repair processing would in effect lose 31.15% of all possible sentences before any processing is carried out. If the system was 100% successful with all other sentences then the output would still only have 68.85% of the sentences identified correctly and this does not take into account any role on effects of the repair information or the use of semantic and pragmatic processing which would use previously recognised passages. It is more likely that the performance would decrease with the repair information confusing the ASR system. Recognition rates of this level are unacceptable.

Using the pattern matching techniques of [1] 66% of those repairs identified in the data would be covered, based on a 100% success rate. This would mean that 10% of those sentences in the passage will not be identified, reducing the total possible correct sentences to 90%. This is again an unacceptable starting point for any speech recognition system even before the effect of the unrecognised repair information is taken into account.

Something must be done to try to incorporate this 10% into the recognition process and try to decrease the effect of repair information on the performance of an ASR system.

## 6. Extended Coverage.

In an attempt to extend the coverage repairs beyond those covered by the simple pattern matching, an analysis was carried out of the repairs found in the data. Those repairs that were covered by the simple pattern matching were removed from the list and the remaining repairs where split into two separate groups. The first group (10% of the total repairs) where those repairs that contained part words, such as:

> The hou —— red house looks good in the sun.

while the second group (24% of the total repairs) contained no significant pattern, such as:

> The green —— red house looks good in the sun.

These two sets of repairs required two different techniques for dealing with them. In both cases it was necessary to look lower than word matching for a potential solution. For the first group the use of phoneme pattern matching was investigated in an attempt to overcome part words which cause great problems in ASR systems. For the second group structure pattern matching was investigated as as a possible solution.

### 6.1 Phoneme Matching

This is a basic pattern matching exercise which spans the whole length of a string of phonemes from left to right. For each phoneme (main phoneme) the rest of the string is checked for possible matches (a match can be equality or two phonemes being in the same articulation class, as described in [3]). When a match is found (secondary phoneme) the phoneme after the main phoneme is matched with the phoneme after the secondary phoneme. If this is a match then the next phonemes are checked. This continues until the match is unsuccessful. A score is calculated (depending on match types, e.g. 2 for equality and 1 for articulation class) for the matched strings and then the process continues directly after the first main phoneme that matched. This process is carried out for every phoneme in the string. A list containing the score, the starting position of the main matched string, the main matched string, the starting position of the secondary matched string and the secondary matched string for every matched pair will be produced.

```
Sen. = The ques the first question to answer is what is software engineering
Pho. = D@  kwEs D@  f3st kwEstS@n t@ Ans@r  Iz wQt  Iz sQftwe@r EndzInI@rIN
Cor. = D@  kwEs @   f3st kwEstSN  @  As@r   Iz wQp  Iz softwe@r EndzInI@rIN

1    =     ----           ----
2    =      ---            ---
3    =                                  ------                          ---
4    =                                   -----                           --
5    =                                  ------                ----
6    =                                                       -----      ----
7    =                                                        ----       ---
8    =       ----                     ----
9    =                                       --------xx------
```

Figure 1: Example of basic phoneme pattern matching

```
Sen. = The ques the first question to answer is what is software engineering
Pho. = D@  kwEs D@  f3st kwEstS@n t@ Ans@r  Iz wQt  Iz sQftwe@r EndzInI@rIN
Cor. = D@  kwEs @   f3st kwEstSN  @  As@r   Iz wQp  Iz softwe@r EndzInI@rIN

1    =     ----           ----
2    =                                  ------                          ---
3    =                                  ------                ----
4    =                                                       -----      ----
5    =       ----                     ----
6    =                                       ------  ------
```

Figure 2: Example of extended phoneme pattern matching

Figure 1 shows the sentence (Sen.) original phoneme string (Pho.), the corrupted phoneme string (Cor.) and the position of the different matching strings. A number of problems can be identified. The first is with pair 9 where the end of the main matching string overlaps the start of the secondary string. The second problem is where each string from one pair is encased within the strings of another pair. This can be seen with pairs 1 and 2 as well as pairs 3 and 4. A more advanced experiment was developed which removes these errors by stopping a search when the end phoneme of the main string is the start phoneme of the secondary string. A change was also required to remove those pairs that are part of another pair. This system would produce those pairs shown in Figure 2.

These six pairs can be used with a hypothesis list from any speech recognition system to identify a repair containing part words (It can also be helpful for pattern matching of whole words). It can be used to identify where extra care needs to taken in trying to overcome recognition problems. Using the above example it can be seen that pairs 2, 3, 4, 5 and 6 would not be relevant as each of those pairs contain a string that covers two words, though it may be possible to condense the strings and use only part of the matched phoneme strings (i.e. remove the first matched phonem from each string in a pair). What to do with the matched strings has not been investigated, but this example, which is typical of many of the repairs examined, shows that phonemc pattern matching has potential in helping overcome repairs.

## 6.2 Structure Matching

Structural pattern matching looks at the structure of the repairs. Speech follows some form of grammar [8] and therefore we use some form of grammar (anti-grammar in AURAID) in ASR systems to help us identify what is likely to have been said. The problem is that repairs break this grammar. But we all understand what each other says with these disfluencies embedded into the speech. This points to the fact that these disfluencies also have some common nature to them. People have proposed acoustic and prosodic cues as the common element in speech repairs. Here the actual structure of the repairs is examined and in a way a form of grammar for these ungrammatical constructs is produced.

In order to make this investigation more substantial the whole set of repairs were taken into consideration. Each sentence in the initial lecture was broken in all combinations of possible structures from the beginning to the end of the sentence. By structure we mean using the word type (e.g. VERB, NOUN, etc.) rather than the word itself. The sentence 'I like red Cars.' has the structure 'PRON VERB ADJ NOUN' and would therefore create 10 different structure (PRON VERB ADJ NOUN, PRON VERB ADJ, PRON VERB, PRON, VERB ADJ NOUN, VERB ADJ, VERB, ADJ NOUN, ADJ, NOUN). From the test lecture 27268 different structure types where created of which one is "VERB", which appeared 1049 times, and another is "CONJ PRON VERB", which appeared 108 times.

The structure of each repair was then calculated and the frequencies of the appearances of the repair structures in the actual lecture were examined.

- 76.79% of the repair structures were unique (i.e. only appeared as repairs, of these 66% appeared only once, as the repair).
- 82.65% of the repair structures that appeared in the lecture were actual repairs.
- 17.35% of those structures that are repair structures that appear in the lecture are not actual repairs.

Therefore using the repair structures as a grammar for repairs 17.35% of those sentences investigated in more detail would have been false alarms. This is quite high but the use of extra knowledge on pattern matching should help overcome this. This technique does not rely on the matching of exact words therefore it has the potential to cover the remaining 24% of repairs not covered by simple pattern matching or phoneme matching.

### 6.3 Coverage

These techniques, though only shown here as exploratory research, do show potential in helping overcome the remaining repairs not covered by a simple (word) pattern matching technique. By combining these two methods with those identified in [1] a more substantial coverage of repairs will be achieved. This coverage may not reach 100% as certain repair types will not be covered (e.g. repairs with part words of which the full word is not used in the actual passage) by even these techniques.

## 7. Conclusion

It is shown that phoneme pattern matching can help to overcome part words which appear frequently in speech repairs and that it can also be helpful, in combination with word pattern matching, in identifying and overcoming other repairs. The use of a grammar for the ungrammatical aspects of speech is also shown to be of potential help. Though not conclusive these techniques could go some way towards enhancing the pattern matching techniques used in repair identification.

It is not being claimed that the techniques presented here would fully cover everything required for dealing with repairs in spontaneous speech recognition. What is presented here is a possible way of extending the coverage of repairs beyond that covered by word matching. The use of acoustic information, as identified in [1], is very important in dealing with repairs and identifying between repairs and correct speech. Prosody would also be very beneficial but as yet is not available and therefore can not be incorporated into any possible practical procedure.

## REFERENCES

[1] J. BEAR, J. DOWDING, and E. SHRIBERG. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63, June 1992. Delaware, USA.

[2] E. R. BLACKMER and J. L. MITTON. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194, 1991.

[3] S. R. BROWNING, R. K. MOORE, K. M. PONTING, and M. J. RUSSELL. A phonetically motivated analysis of the performance of the ARM continuous speech recognition

system. In *Proceedings of the Institute of Acoustics Speech and Hearing Conference*, November 1990. Windermere.

[4] J. BUTZBERGER, H. MURVEIT, E. SHRIBERG, and P. PRICE. Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 339–343, 1992.

[5] R. J. COLLINGHAM. *An Automatic Speech Recognition System for use by Deaf Students in Lectures.* PhD thesis, University of Durham, 1994.

[6] R. J. COLLINGHAM and R. GARIGLIANO. Using anti-grammar and semantic categories for the recognition of spontaneous speech. In *Proceedings of Eurospeech, the 3rd European Conference on Speech Communication and Technology*. ESCA, September 1993. Berlin.

[7] P. K. FINK and A. W. BIERMANN. The correction of ill-formed input using history-based expectation with applications to speech understanding. *Computational Linguistics*, 12(1):13–36, 1986.

[8] R. GARIGLIANO, K. JOHNSON, and R. J. COLLINGHAM. A data-supported case for a spontaneous speech grammar. In *Proceedings of Eurospeech, the 3rd European Conference on Speech Communication and Technology*. ESCA, September 1993. Berlin.

[9] P. HOWELL and K. YOUNG. The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology*, 43A(3):733–758, 1991.

[10] K. JOHNSON, R. GARIGLIANO, and R. J. COLLINGHAM. Data-based control of the search space generated by multiple knowledge bases for speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, September 1994. Yokohama, Japan.

[11] W. J. M. LEVELT and A. CUTLER. Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–217, 1983.

[12] R. MITTON. *A Description of a Computer-Usable Dictionary File Based on the Oxford Advanced Learner's Dictionary of Current English*, June 1992.

[13] D. O'SHAUGHNESSY. Analysis and automatic recognition of false starts in spontaneous speech. In *Proceedings of the IEEE International Acoustics Speech and Signal Processing*, pages 724–727, 1993. Minneapolis.

[14] C. ROWELS and X. HUANG. Prosodic aids to syntactic and semantic analysis of spoken english. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 112–119, June 1992. Delaware, USA.

[15] M. Q. WANG and J. HIRSCHBERG. Predicting intonational boundaries automatically from text: The ATIS domain. In *Proceedings of the DARPA workshop on Speech and Natural Language*, pages 378–383, February 1991.