BBC

R&D White Paper

WHP 015

September 2001

Applications of speech recognition technology in broadcasting

D. G. Kirby

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper 015

Applications of speech recognition technology in broadcasting

D. G. Kirby

Abstract

This paper describes work which is exploring how speech recognition can be used in programme production. It examines the performance of large vocabulary, speaker independent, speech recognition in a broadcast context and illustrates this with results showing the performance likely to be achieved under these conditions. These show how the variety of speech and audio qualities to be expected in broadcast programmes makes this a challenging task but nevertheless, performance adequate for a range of applications can be achieved.

Practical applications for the technology are described. Some of these are already demonstrating significant benefits, whilst others suggest completely new ways of working will be possible as this technology improves further.

It is concluded that speech recognition is now able to deliver worthwhile results if applications are considered carefully and that, over time, its introduction will make a more widespread impact.

Given at the International Broadcasting Convention (IBC 2000), Amsterdam, 7-12 September. Conference Publication.

Key words: speech recognition, programme production, archive retrieval,
teleprompter, subtitling.

APPLICATIONS OF SPEECH RECOGNITION TECHNOLOGY IN BROADCASTING

D. G. Kirby

BBC R&D, U.K.

ABSTRACT

This paper describes work which is exploring how speech recognition can be used in programme production. It examines the performance of large vocabulary, speaker independent, speech recognition in a broadcast context and illustrates this with results showing the performance likely to be achieved under these conditions. These show how the variety of speech and audio qualities to be expected in broadcast programmes makes this a challenging task but nevertheless, performance adequate for a range of applications can be achieved.

Practical applications for the technology are described. Some of these are already demonstrating significant benefits, whilst others suggest completely new ways of working will be possible as this technology improves further.

It is concluded that speech recognition is now able to deliver worthwhile results if applications are considered carefully and that, over time, its introduction will make a more widespread impact

INTRODUCTION

The performance of speech recognition technology has improved to the point where it is useful to consider how it could be applied to broadcasting.

Although dictation software using speech recognition is now widely available, using speech recognition on the audio tracks of broadcast programmes is a much greater challenge. Despite this pushing the technology to (and frequently beyond) its limits, new ways of working in programme production, archiving and many other areas become possible.

This paper describes some of the ideas that have been developed at BBC R&D to explore what speech recognition can offer and how broadcast production operations could benefit from its use.

CATEGORIES OF SPEECH RECOGNISER

There are several key factors in speech recognition that impose conflicting demands on performance and hence its suitability in different applications. Those factors of particular importance in our context are size of vocabulary, being independent of the speaker, and tolerance to differing acoustic conditions and delivery styles, e.g. scripted or spontaneous speech.

The most widely available speech recognition packages are the speaker-dependent systems, aimed at dictation applications. These are trained to the voice of the user and are used in reasonably well-controlled acoustic conditions and with a carefully positioned headset microphone. Hence, although they have a large vocabulary and can achieve an accuracy of 95% or so, their application in broadcasting is limited to those situations where only a few users are required and configuration files can be changed between each user.

In contrast, speech recognition used for telephone-based services, e.g. banking and automated switchboards, work with many different voices but they have a vocabulary limited to keywords or a range of names. Hence they are not suitable in a broadcasting context that we are addressing here.

In our applications we are interested in using speech recognition on the content of broadcast programmes to provide improved production methods. In this situation, we have little or no control over any aspect of the audio content passed to the speech recogniser and hence we are working in an extremely demanding situation. Ideally, we require a speech recogniser that is capable of providing good performance with a large vocabulary, a wide range of speakers, and acoustic conditions varying from broadcast studios to noisy venues on location. This range of demands cannot, at present, be fulfilled by any speech recogniser. Hence it is important to understand the performance that can be achieved in practice and develop those applications which are not unduly demanding of the technology.

TRAINING SPEECH RECOGNISERS

All speech recognisers need to run in a training mode prior to use in order to achieve optimum performance for the quality and type of dialogue intended. For a dictation-based product, this is not particularly onerous, requiring the user to work through an initial training session lasting about 30 minutes. The software then continues to adapt itself as the user corrects errors during use. In contrast, the training of a large vocabulary, speaker-independent recogniser is a more complex process. It requires many hours of accurately transcribed recordings and a very large quantity of text documents, representative of the usage of that language, in order to model word usage and hence the likelihood of particular word sequences occurring.

Such training was carried out for the THISL project, Robinson et al (1), (described later) which uses the Abbot speech recogniser developed at Cambridge University (UK). In this case, 45 hours of BBC radio and TV news was transcribed as accurately as possible for the acoustic training of the recogniser. For the language modelling, two-years' of scripts from News and other programmes were gathered and, together with the entire text content of the BBC News web-site, resulted in documents amounting to 30 million words being available. Further texts from US sources were used to supplement this selection.

The gathering, transcribing and checking of this amount of material is a time-consuming task. Starting with the basic programme transcript and turning it into an accurate and fully marked-up version for the training, can take between four and ten times the duration of the programme. Radio programmes proved to be easier to work with than television programmes, probably because there are fewer interruptions to the speakers and less intrusive background noise.

PERFORMANCE

The wide range of conditions and speakers in typical broadcast programmes make it difficult to give a clear measure of the performance of a speech recogniser. Additionally, there are many factors that affect the accuracy of speech recognition and hence a single, unqualified performance figure can conceal many underlying differences. For example, running a speech recogniser slower than real-time may improve its accuracy but will significantly limit its usefulness: having to wait ten minutes for a one-minute recording to be processed is unlikely to be acceptable. Published results therefore need to be interpreted with care when considering suitability for a particular application.

Despite these caveats, the Word Error Rate (WER) figure is normally used to give an indication of recognition accuracy. This is the number of words in error expressed as a percentage of the entire document. This figure will include inserted, deleted and substituted words although these figures are sometimes quoted separately.

Referring again to the THISL project (1), after training had been completed, tests using a separate group of six news programmes, three radio and three television were carried out. These programmes were from a few months after the training data had been gathered and so had not been 'seen' by the speech recogniser during its training.

For the six news programmes together, the average WER achieved was 26.8% when running at close to real-time speed and with a 65,000-word vocabulary. Perhaps not surprisingly, TV news has a worse recognition rate at about 33% than radio news which averages about 20%.

Although a single figure of merit indicates the general trend, a more revealing indication of actual performance is shown in Figure 1, which shows the variation in word error rate throughout one of the 30-minute radio news programmes.

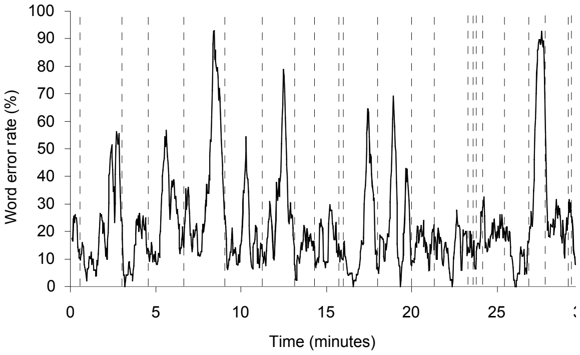


Figure 1. The word error rate from a speech recogniser throughout a 30-minute radio news programme.

Two features are noticeable from this graph: for the majority of the programme the error rate is in the 10% to 25% region but there are peaks, some of which are above 90% WER. Looking at the content of the programme in more detail, reveals an interesting pattern. The broken vertical lines mark the boundaries between each news story and within each story there tends to be a region of good performance followed by worse performance. The regions of good performance generally correspond to a scripted contribution read in a studio and probably by the news presenter. The peaks in the 50% region tend to be unscripted and more casual speech, whilst those reaching 90% are further degraded generally by the poorer quality of the speech or the audio link.

Following these assessments, excerpts from studio interviews, narration from wildlife programmes and examples of other broadcaster's English language material, gathered by the BBC's Monitoring Service, were used as further test items. In all these cases, the performance of the speech recogniser was markedly worse that that achieved with the TV and radio news test items. The factors causing this degradation have yet to be determined, however these informal results suggest that speech recognisers may require 'tuning' in order to achieve optimum performance for different genres of programmes.

Although the results presented here are for one particular configuration of a speech recogniser, results published from annual evaluations in the US by the National Institute for Standards and Technology (NIST) (www.itl.nist.gov/iaui/894.01/) generally indicate the same trends. These wider evaluations give a valuable indication of how current research is advancing and the forthcoming improvements that might be expected.

Overall the published results show that, for broadcast audio, speech recognition performance is likely to be highly variable, at times being very good whilst at other times being poor. To some extent, the performance follows the intelligibility of the audio track itself, although this is by no means always the case.

APPLICATIONS IN BROADCASTING

The foregoing results show that it is important to be realistic when it comes to matching the performance of the technology to applications. Applications must be tolerant of recognition errors; the inevitable errors must either have little impact on overall performance or provision must be made to correct them, probably by hand. This section describes some of the applications that are being considered in the field of broadcasting.

Transcribing

There is considerable interest in using speech recognition to transcribe recorded material both to assist in post-production work and also to provide a record of the content of completed programmes.

However, it is clear from the results shown above, that, with the current level of performance, this is too demanding an application for the technology. Although manual correction of the text from the speech recogniser can be carried out, this may frequently take longer than transcribing the entire document by hand in the first place. As the speech recogniser produces phrases that, although incorrect, are plausible within themselves, identifying and correcting regions of the text that are in error can take significantly longer than might at first be expected.

If the broadcast material is of a restricted range and of consistent quality then transcribing may be possible, particularly if the speech recogniser can be re-trained as more recordings are processed. However, with typical broadcast material it seems that this application cannot yet be addressed to any useful extent.

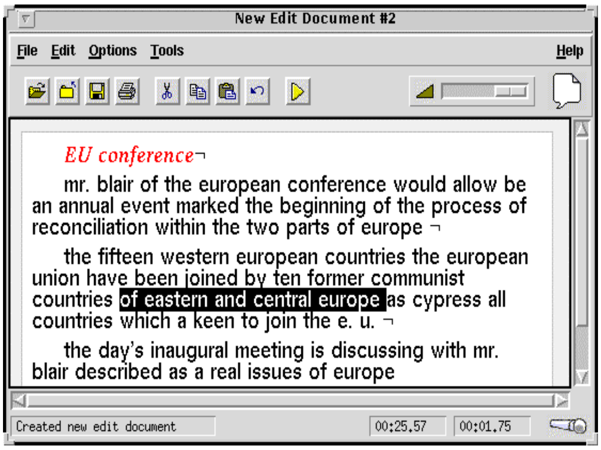
Editing

As part of the speech recognition process, the start and end of each word are identified and these timings can be used to locate each word within the original recording. This provides the basis for an alternative method of editing audio: a word processor can be used to edit the text produced by the speech recogniser then, as sections of this text are deleted or moved, so the corresponding edits are performed on the original audio recording. In this way the audio recording can be quickly edited for content and running order without having to listen to the recording in its entirety.

It is not intended that this approach would eliminate the need for the careful and precise audio editing currently performed to ensure that edits are inaudible. Variations in intonation and the like, that will not be obvious from the text, will mean that this will not be possible. However, as a way of gathering together selected sections of recordings and getting them into the appropriate order, it can offer a much faster way of working.

Figure 2 shows an experimental word processor that offers this facility. The text displayed is from a speech recogniser: sections can be highlighted and cut and pasted elsewhere. At any time the audio can be replayed to hear all or part of the editing. The running time for the entire excerpt or a selection from it is displayed on the lower edge of the window.

To facilitate further editing, an edit decision list (EDL) can be exported so that work can continue from that point but using a conventional audio or video editing system to complete the task. Figure 3 shows an editing display using such an EDL from the speech recogniser. The text from the speech recogniser is displayed along the audio waveform with marking between each word. These features



EU conference¬

mr. blair of the european conference would allow be an annual event marked the beginning of the process of reconciliation within the two parts of europe ¬

the fifteen western european countries the european union have been joined by ten former communist countries of eastern and central europe as cypress all countries which a keen to join the e. u. ¬

the day's inaugural meeting is discussing with mr. blair described as a real issues of europe

Created new edit document 00:25.57 00:01.75

Figure 2. An audio editor which uses speech recognition coupled with a word processor, to allow cut and paste editing of an audio recording.

allow rapid navigation though the recording and enable features such as 'Find Word' to be provided.

Whilst this application does not demand the same level of accuracy from the speech recogniser as does transcribing, too many errors in the text will make it difficult to use. The current level of performance being achieved suggests that it can be effective for some types of programme but improved accuracy for a wider range of programmes is needed before is could be adopted more generally. Despite this limitation at present, this approach undoubtedly holds much promise for future editing systems, both audio and video.

Archive Retrieval

Using speech recognition to index broadcast programmes is a topic that is being widely researched by the speech community. The essence of this application is that broadcast programmes are passed through a speech recogniser and the resulting text indexed, to provide a means of searching the content of those programmes.

To investigate the feasibility of this approach, the BBC participated in a recent EC-supported project, THISL (1), that developed a practical system for indexing and subsequently searching BBC News broadcasts in this way. THISL brought together speech recognition and information retrieval techniques and with a web-based user interface, provides an experimental archive retrieval service on the BBC's Intranet. A fully automatic, off-air recording and processing arrangement was set up early in the project. Each day, this records an average of 5 hours of BBC television and radio broadcasts including all main news bulletins and these are available from the archive shortly after broadcast. Although set up as an experimental facility for the project, the THISL archive now amounts to about 2000 hours of programmes, going back to January 1998 and has itself, become a valuable asset.

For this application the speech recogniser runs in real-time on a high performance PC and has a 65,000-word vocabulary. The resulting word error rate is about 26%. However the overall effectiveness, measured in terms of how relevant the retrieved stories are to the user's query, is surprisingly immune to errors in speech recognition. The results suggest that with a WER of better than 30% or so, it is the performance of the accompanying information retrieval system that dominates overall system performance. This is therefore a very suitable task considering the current level of speech recognition performance that can be achieved.

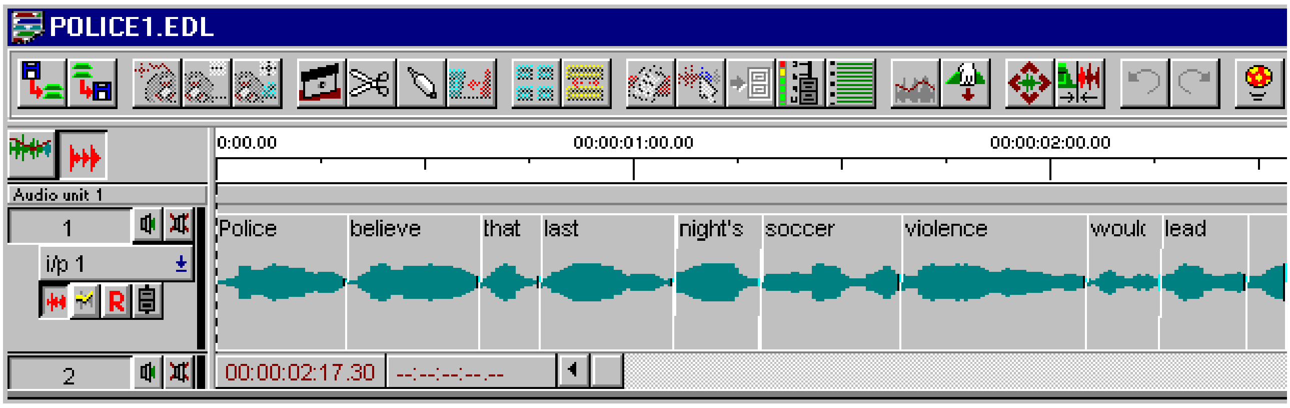


Figure 3. A conventional audio desktop editor display enhanced by the text of the spoken words which are produced automatically by a speech recogniser and displayed along the audio waveform.

Formal evaluations of spoken document retrieval systems of this type are carried out each year as part of the annual Text Retrieval Conference (TREC) organised by NIST. A summary of results for THISL and other systems, under a variety of conditions, are available from the their web-site at http://trec.nist.gov.

Users access THISL via a web-page in a similar way to a search engine. The user enters a search phrase and can optionally restrict the search to a range of dates or programmes. THISL returns a list of stories ranked in order of relevance and the text for the currently selected story. Although the text will contain errors from the speech recogniser it is very useful in determining if the story contains the details sought. A browse quality version of the audio is also available allowing users to hear the selected excerpt immediately and hence confirm its contents.

From a user's perspective the advantages of the THISL system are that it gives immediate access to the archive. Users are generally finding that the search is very effective: the top ranking stories returned by the search are relevant to their query. Furthermore, the display of the text from the speech recogniser and being able to replay the audio immediately, means that the results can be reviewed quickly and, if necessary, the search terms refined.

From an operational point of view, the complete automation of the service, from initial recording, through to the processing, indexing and web-server stages leads to low running costs and hence a wider range of broadcast output can be indexed and made available in this way. Whilst the quality of such an automatically generated index cannot match that produced by experienced cataloguers, the quantity of material that can be captured and indexed in this entirely automatic way makes the system very attractive.

The experience gained so far by the BBC's Information and Archives Department in evaluating THISL suggests that this approach offers very significant potential for future archive operations.

Teleprompter

The requirement for accuracy from the speech recogniser can be relaxed in applications where there is prior knowledge of what may be spoken. One such example of this is for a teleprompter, where the presenter is reading from scripted text shown in a partially reflecting mirror across the camera lens. As only about 8 to 10 words are shown at once, the text is scrolled upwards in step with the presenter's delivery either by an operator or by the presenter using a foot-operated control.

For this application speech recognition applied to the presenter's speech can be used to track the corresponding point in the script and hence calculate the average speech rate. This position and speed data is then used to control the scrolling of the text displayed to the presenter as required.

An experimental system to demonstrate the feasibility of this idea was developed at Cambridge University (UK) and shows that reliable tracking of the speech can be achieved under most circumstances (Pickford, 2).

A further development of this technique is to produce trigger signals from the text-tracking module as certain keywords in the script are reached. For example, when a person's name is reached in the script, a trigger could be produced to cut to a corresponding picture of that person. Similarly, pre-recorded items could be played-in at the instant that their corresponding keywords in the script are spoken.

This could provide semi-automatic operation of the studio controlled from the presenter's voice. Whilst not necessarily appropriate in a main studio, for smaller studios, where fewer staff are available, this automation may be an effective way of reducing their workload.

Subtitling

Live subtitling is perhaps the most demanding of applications for speech recognition: not only is high accuracy required but processing time must also be kept short. In the BBC, live subtitling is carried out by stenographers who, using a dedicated 'steno' keyboard, achieve an accuracy exceeding 97%, i.e. at worst, a 3% WER, and with a maximum delay of one second. As the earlier results show, speech recognition accuracy will fall well short of this level and is not currently a practical alternative on its own. If speech recognition is used then a method of correcting the text manually will be required. As Figure 1 shows, at times the amount of correction necessary will be significant.

There may be some scope for using some of the underlying techniques in speech recognition to help minimise keying errors in the present system. As already mentioned, speech recognition uses a language model to give the likelihood of words in a particular phrase. Many of the errors in live subtitling arise from slight mis-keying of sounds which result characters being produced that match the speech phonetically but for which no word can be found in the dictionary, e.g. "are you shaur?"

rather than "are you sure?". The decoding stage within a speech recogniser should be able to decode this correctly as its language model and dictionary would ensure that any erroneous words are replaced by the most likely word in that situation.

An alternative approach to live subtitling is to use a speaker-dependent speech recogniser, such as one of the commercially available dictation packages, and have the user re-voice the programme dialogue as it is broadcast. This approach has some potential but processing delays inherent in the speech recognition will mean that subtitles will lag behind the speech by a more significant amount than at present.

FURTHER CONSIDERATIONS

In experimenting with speech recognition several limitations have become apparent that should to be considered in any practical system.

Vocabulary

Over time, language usage will change and new words will come into the vocabulary. In News, for example, new names, either of people or places, are coming into use almost daily. A convenient way of adding these to the speech recogniser's vocabulary is essential otherwise it will gradually become less effective when used with topical material. Furthermore, gathering examples of such new language usage for re-training is impractical. Words need to be added to the vocabulary quickly although further training, to improve performance might be possible, as more examples of usage become available.

Punctuation and Capitalisation

The output from a speech recogniser is likely to be continuous text in upper case without punctuation or paragraph breaks. In some applications, such as archive retrieval, this is not a significant problem, as the user does not need to refer to the text in detail. However, when the user needs to work with the text, e.g. as a transcription, it needs to be broken up in some way to improve readability and avoid confusion. Ideally, sentences need to be broken by punctuation and paragraphs breaks inserted when appropriate. This is clearly a complex problem that the research has yet to explore.

A simpler first step that would help with readability would be to start a new paragraph of text when a change of speaker is detected. This is a more realistic option considering the current state of the technology and for many situations, would suffice.

Appropriate capitalisation of words is also important and the speech recogniser's word list should correctly capitalise names. However, rules for capitalisation in a general context are complex and would require a subsequent analysis of the text to determine its structure and language usage.

CONCLUSIONS

Although broadcast programmes are challenging for speech recognition, the technology has advanced to the level where it can already be applied very effectively in some applications.

As the accuracy improves further, several other applications will become feasible that will, in themselves, offer enhanced or completely new ways of working in broadcast production.

REFERENCES

1. Robinson, T., Abberley, D., Kirby, D. and Renals, S. 1999. Recognition, indexing and retrieval of British broadcast news with the THISL system. Proceedings of Eurospeech 1999. pp. 1067 to 1070.

2. Pickford, D. 1998. Automated Control of a teleprompter. MPhil thesis. University of Cambridge, August 1998.

ACKNOWLEDGEMENTS

The THISL project was an ESPRIT Long Term Research Project supported by the EC (EP23495). The project partners and their various and significant contributions to that work and the results presented here, are detailed on the project's web-site http://www.dcs.shef.ac.uk/spandh/projects/thisl

The author would like to thank the BBC for permission to publish this paper.