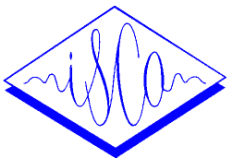

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

REAL-TIME SPEECH-GENERATED SUBTITLES: PROBLEMS AND SOLUTIONS

J.Hewitt, A.Bateman, A.Lambourne*, A.Ariyaeeinia, P.Sivakumaran

University of Hertfordshire, College Lane, Hatfield, Herts. UK. AL10 9AB
*Synapsys Ltd., Riverdale House, 19-21 High Street, Wheathampstead, Herts. UK. AL4 8BB

ABSTRACT

This paper refers to work carried out in the Subspeak project [1] in which we are investigating the use of speech recognition in live television subtitling. Research to date has shown that with current speech recognition technology it is not possible to achieve a satisfactory level of accuracy in the direct transcription of broadcast material. To circumvent this problem in our system the broadcast speech data is respoken by a native English speaker in a quiet environment. Recognition rates of up to 98% can be achieved by a trained speaker where there are no out of vocabulary words. However, using conventional keyboard input, subtitlers can currently achieve near to 100%, with typically only minor errors of spelling or punctuation. The challenge is therefore to provide a speech-based subtitling system which mirrors the conventional systems in accuracy and speed, but which requires far less time to train subtitlers to use. Subtitles must typically be output at between 150 and 180 words per minute and the delay between the broadcast speech and the appearance of the subtitle must be at most 8 seconds. In the prototype system, output from the speech recognition system is passed in to a custom-built editor from where it can be corrected and passed on to an existing subtitling system.

1. BACKGROUND

The subtitling of television programmes for the hearing impaired is a mandatory requirement placed on Independent Television Broadcasters in the UK by the 1990 Broadcasting Act. The later 1996 Act licensing Digital Terrestrial Broadcasting contains similar obligations. The amount of subtitling under this legislation is set to rise year by year, and will increase dramatically as more channels become available. Because subtitling is labour-intensive and there is a shortage of skilled operators, effort is being invested in seeking improved and more efficient means of producing subtitles – especially for live programmes. Currently this is the area where skills shortage is the most acute, and specialisation is high. Techniques are needed which maintain quality but increase the potential base of production staff.

2. THE SUBSPEAK PROJECT

Subspeak, which began in October 1998, is a 3-year LINK project under the Broadcast Technology Initiative, jointly funded by the DTI and EPSRC. The main partners are the University of Hertfordshire and Synapsy Ltd., a company which specialises in broadcast subtitling and digital information services.

The aim of the project is to investigate how the techniques of speech recognition and speaker recognition can be used to increase the efficiency of subtitling and to enable subtitles to be produced for a wider range of programme material, particularly live programmes. In the first half of the project the emphasis has been on an assessment of the currently available tools and on the development of an interface to enable an operator to quickly correct spoken subtitles prior to them going ‘ on air’ . A test-bed system has been developed to facilitate various experiments related to the specific problems encountered with speech recognition and text editing in the real-time subtitling environment.

3. SPEECH RECOGNITION TOOLS

The direct transcription of speech data found in the real world including spoken material in broadcast television programmes poses a number of challenges to large vocabulary continuous speech recognition systems [2-6]. This is due to the fact that the speech data is non-homogeneous and may contain data types which have not been used in the training process. A typical television programme may contain speech data with a variety of speaking styles (e.g. read, spontaneous), the talkers may be non-native as well as English speakers and the data may be affected by variation in communication channel conditions and by background noise. To tackle these problems a number of methods based on segment processing [4-6] have been investigated, but the results of the National Institute of Standards and Technology (NIST) investigations in 1997 clearly indicated that although some improvements can be achieved, error rates are still high and range from 30% to over 60% [2-7]. It is apparent that for the forseeable future direct transcription of broadcast spoken material is not practical, therefore an interim measure is needed. The approach taken by this project is to employ a trained respeaker.

Before any practical prototype system could be developed it was necessary to determine which of the available speech recognition systems would be most appropriate to the task of live televsion subtitling. A range of speech recognition systems were investigated with reference to the results of the 1997 NIST evaluations and two were chosen for further trials, the Cambridge Entropic System and the IBM ViaVoice Executive. Initial recognition results were more favourable for the IBM system and, although the Entropic system was seen as perhaps providing a greater degree of flexibility in the long term, it was decided to proceed with initial investigations using the ViaVoice system and its SDK (software developers toolkit). In trials based on a trained speaker reading text at 150 wpm, where there were no out of vocabulary words, recognition rates of 98% were consistently recorded.

4. EXPERIMENTAL TEST-BED SYSTEM

A test-bed has been set up to enable us to carry out experimental work and training in the two aspects of the system – respeaking subtitles and correcting the output of the speech recogniser.

This test-bed allows several different modes of operation. The person speaking the subtitles (the Speaker) can choose to hear speech from different sources – from the sound track of a videotape, a specially recorded audio CD (for practice at correction and reproducable test conditions), or from a live television broadcast. Similarly the person correcting the subtitles (the Corrector) can either listen to the Speaker, to a soundtrack recorded on a CD, or s/he can play a text file directly into the editor at a specified rate with no soundtrack. The Punctuation Generator is a laptop computer which can play back sound samples of the Speaker saying particular punctuation phrases such as "Full Stop", "Comma", "Dash", "Exclamation Mark".

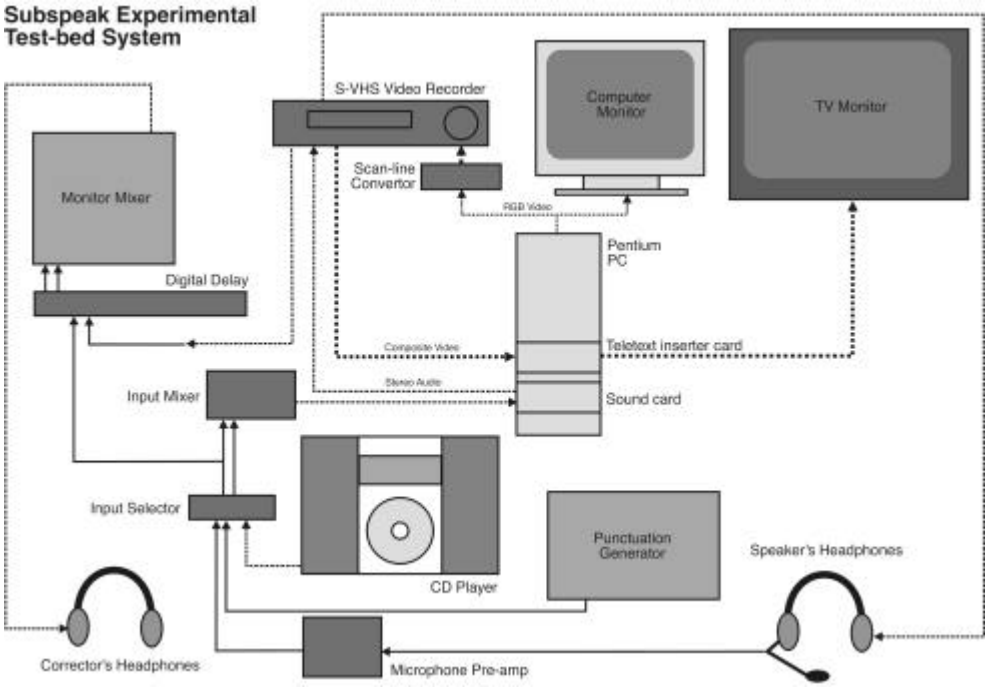

Figure 1: Diagram of the experimental test-bed system.

5. INTERFACE DESIGN

The process of creating live subtitles by speech can be broken down into a number of parallel tasks. These are:

> *watch program* || *listen to dialogue* || *process dialogue* || *speak subtitle* || *read output from recogniser* || *correct output* || *send subtitle*

The task 'process dialogue' may involve a degree of précis of the dialogue along with decisions as where to add punctuation marks. Although it may be possible for all these tasks to be carried out by a single operator, initial studies showed that for most programmes, where subtitles must be output at between 150 and 180 wpm, it would be more feasible if two operators were employed, one to do the speaking and one to do the correcting. The tasks would then be as follows:

Operator one – the Speaker:

> *watch program* || *listen to dialogue* || *process dialogue* || *speak subtitle*

Operator two – the Corrector:

> *listen to output from operator one* || *read output from recogniser* || *correct output* || *send subtitle*

The editing system has been designed on this basis and a number of variations of the interface are currently being tested.

5.1 The Speaker Interface

The Speaker listens to the live television programme on a headset and repeats what has been said. In order to ensure maximum efficiency of the recognition engine s/he must add punctuation to her dictation. Two methods are being tested, firstly where the Speaker says the punctuation as part of the dictation and secondly where the Speaker presses a key on the Punctuation Generator which correspond to pre-recorded punctuation commands. The first method directly mirrors current dictation systems, the advantage of the second is that it gives the Speaker time in which to catch breath between sentences; a possible disadvantage is that the Speaker can 'speak over' the punctuation thus confusing the recogniser. Further work is being done to allow the Speaker to indicate changes in speaker on the programme (as these are colour coded in the final subtitle) and to indicate an unknown word to the Corrector – typically this could be a name that s/he didn' t know.

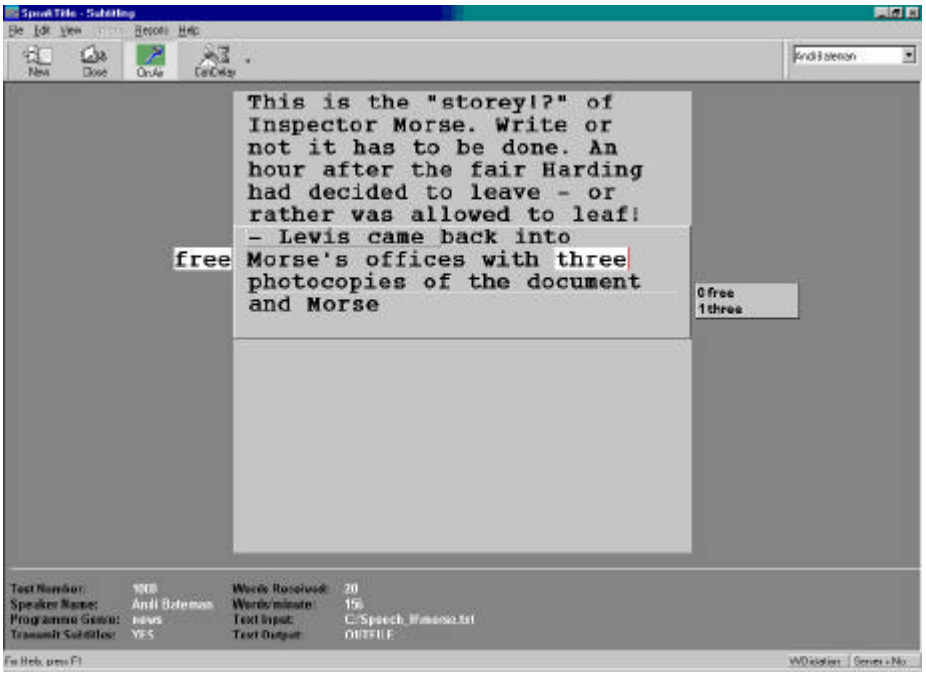

Figure 2: Speaktitle prototype interface.

5.2 The Corrector Interface

The interface of the prototype editing application, known as "Speaktitle" is shown in Figure 2. The text from the speech recogniser appears in a scrolling window, it must be edited within a short timeframe before being sent out as a subtitle. Various scrolling modes are being investigated, these include 'elastic' where text that is being edited is not sent out until the edit is completed, 'semi-elastic' where text can be delayed for only a limited amount of time before it is automatically sent out as a subtitle and 'bulldozer' where text is sent out continuously, edited or not. The process of editing is under close scrutiny, it must allow the selection of a word or phrase and corrections in word or character mode in addition to the possibility of replacements from a list of alternatives provided by the recogniser. The nature of speech recognisers is that they do not usually mis-spell words but they sometimes provide the wrong word and can give spurious endings (adding an 's' or leaving it off are fairly typical), the errors also tend to 'bunch up' rather than being evenly spread; hence the interface must be specially designed rather than duplicating that of a standard word processor, but it must retain sufficient similarity of operation to allow the user to transfer their word processing skills. Additionally, different input devices are being investigated. Experiments are being carried out with a touch screen, a standard keyboard and mouse and a trackerball.

6. EDITING TRIALS

A CD was made of the Speaker re-speaking soundtracks from a number of videotaped live television broadcasts. These include the pre-recorded spoken punctuation which was inserted during the recording session using the Punctuation Generator. The speaking rates of the tracks on this CD vary from about 150 to 196 wpm. Any out-of-vocabulary words were added to the speech recogniser's vocabulary prior to the trials.

A track was selected and passed through the two-second delay to the headphones of the operator. The Corrector used the Speaktitle interface to correct as much of the text as s/he could.

At the end of the session two recognition rates were calculated:

- the raw input to Speaktitle compared with the original text
- the corrected output from Speaktitle compared with the original text.

In an initial study eight trials were carried out with one operator, the results are shown in table 1:

Track No.	Speaking Speed (wpm)	Raw Recognition Rate (%)	Corrected Recognition Rate (%)	Percent Improve-ment (%)
1	196	92.24	94.41	2.17
1	196	94.72	96.58	1.86
1	196	94.43	96.90	2.47
2	172	90.10	92.08	1.98
3	181	92.65	95.53	2.88
3	181	93.18	96.59	3.51
3	181	92.76	95.74	2.98
3	181	92.86	95.95	3.04

Mean % improvement over the 8 trials = 2.61%

*Recognition rate = (N-D-I-S)*100/N where: N= no. of words spoken, D= no. of omitted words, I = no. of inserted words, S= no. of substituted words.*

Table 1: Results for preliminary editing trials.

It is worth noting that both the Speaker and the Corrector were fairly inexperienced and the editing interface was not optimal. Interestingly the raw recognition rate for *recorded* speech varied each time a track was played. There is a significant difference between the recognition rates for these spoken subtitles compared with that for when they were read. (Typically rates of 97% – 98% were achieved).

6.1 Ongoing investigations

A further set of respoken tracks were recorded using a trained subtitler. These are currently being used in further trials with an improved version of Speaktitle to see the extent to which the Correctors can improve their performance over time. The sessions are captured on video to enable us to study the interface in depth and to identify any usability problems. In addition to the two recognition rates calculated above we are also assessing the *effective error rate* of the corrected output – that is, ignoring any errors which would not have been significant had they been put out as subtitles, for example the substitution of "a" for "the", errors of capitalisation and hyphenation.

A major issue under consideration is how to deal with out of vocabulary words since these can seriously affect the performance of the speech recogniser causing large delays and unrecoverable errors in transcription. It is extremely difficult to correct a completely mis-recognised sentence in the allowable time window. We are investigating three strategies to alleviate this problem:

- Building context specific vocabularies to be used for particular types of television programme, for example sport, chat show or news.
- Updating the speech model of Via Voice ' on the fly'
- Providing short cuts in the editor to previously typed corrections.

7. SPEAKING TRIALS

Eventually we will need a way of assessing a person's effectiveness as a potential subtitle speaker. The following trial was carried out on four subjects:

- The subject was enrolled into ViaVoice by reading the first 100 training sentences. (Each subject was set up as a separate enrolment within a single user name, in this way they could all share the same vocabulary).
- The subject read the transcript of Track 1 of the CD into Speakpad, corrected the errors and then read in the same text again.
- S/he then read the transcript of Track 3 into the system
- The subject attempted to respeak Tracks 1 and 3 played to them via headphones
- At each stage all recognition rates were recorded. These are shown in table 2 (the first reading for subject 3 was overwritten and subject 4 had time for only two tasks)

Subject 1 was familiar with the text as she had carried out editing trials although she had not used ViaVoice before.

Subject 4 was a trained subtitler with some experience of ViaVoice. It was apparent during the trial that the tasks were too difficult for novice subjects 2 and 3 and it put them under some stress.

Sub-ject	Track 1 First reading	Track 1 Second reading	Track 1 Respoken	Track 3 Read	Track 3 Respoken
1	91.61	92.55	72.67	87.86	64.86
2	49.69	80.12	29.19	67.48	23.61
3		88.2	32.3	69.49	36.3
4	91.61		89.84		

Table 2: Recognition rates for speaking trials.

The experiment has been re-designed around a series of tracks graded by speed of delivery. This allows subjects to familiarise themselves with the process of re-speaking, starting with a low speed of dictation and gaps between sentences to allow the respeaker to complete their rendition of the sentence before they hear the next one. The tracks presented to the subject become progressively closer to the actual speed and presentation of live broadcast material. To avoid stressing the subject, at any stage the experimenter can terminate the experiment if s/he feels that the subject is unlikely to be able to complete the next level of the task.

8. CONCLUSIONS

The work to date has demonstrated the viability of a system to generate subtitles from the spoken word. In the next phase of the work we intend to develop the Speaktitle interface further and to use the system in live broadcast situations. We are also formulating a methodology for the selection of appropriate people to act as Speakers.

9. REFERENCES

1. Link: GR/M15958
2. Pallet D.S. et al., "1997 Broadcast News Benchmark Test Results: English and Non-English", Proc. DARPA Speech Recognition Workshop, 1997.
3. Bakis, R. et al., "Transcription of BN Shows with the IBM LVCSR System", Proc. DARPA Speech Recognition Workshop, 1997.
4. Chen, S. et al., "IBM's LVCSR System for Transcription of Broadcast News Used int the 1997 HUB4 English Evaluation", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
5. Woodland P.C. et al., "The 1997 HTK Broadcast News Transcription System", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
6. Wegmann S. and Scattone, F., "Dragon Systems' 1997 Broadcast News Transcription System", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
7. Liggett W. et al., "Insights From the Broadcast News Benchmark Tests", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998