SPEECH-BASED REAL-TIME SUBTITLING SERVICES
Andrew Lambourne*, Jill Hewitt, Caroline Lyon, Sandra Warren

* SysMedia Ltd. Riverdale House, 19-21 High Street, Wheathampstead, Herts., UK.
Department of Computer Science, University of Hertfordshire, Hatfield, Herts., UK
Andrew.Lambourne@sysmedia.com, J.A.Hewitt@herts.ac.uk

Abstract

Recent advances in technology have led to the availability of powerful speech recognizers at low cost and to the possibility of using speech interaction in a variety of new and exciting practical applications. The purpose of this research was to investigate and develop the use of speech recognition in live television subtitling. This paper describes how the "SpeakTitle" project met the challenges of real time speech recognition and live subtitling through the development of a customisable speaker interface and use of 'Topics' for specific subject domains. In the prototype system (described in Hewitt et. al. 2000 and Bateman et. al. 2001) output from the speech recognition system (the IBM ViaVoice ® engine) is passed in to a custom-built editor from where it can be corrected and passed on to an existing subtitling system. The system was developed to the extent that it was acceptable for the production of subtitles for live television broadcasts and it has been adopted by three subtitle production facilities in the UK.

The evolution of the product and the experiences of users in developing the system in a live subtitling environment are considered, and the system is analysed against industry standards. Ease-of-use and accuracy are also discussed and further research areas are identified.

Keywords: real-time subtitling, speech recognition, language models

Speech-based real-time subtitling services

1   Introduction

The value of subtitles in providing access for the hearing-impaired audience of television programmes has long been recognized and is reflected in legislation in the US and Europe (see for example references: UK legislation 1990, 1996, 2001; US legislation 1990, 1996, 2001). As broadcasters seek to fulfil the mandated increases in subtitling coverage, more and more subtitling is being created in real-time for "live" or "as live" television. This presents technical and editorial challenges, as well as the problem of finding suitably skilled subtitlers. As an alternative to different kinds of fast keyboard devices, the technique of "re-speaking" a subtitle commentary into a speech recognizer has been investigated. This paper describes a project to assess the feasibility of this method and to develop a practical system for speech-based live subtitling. Following a general overview of initial work in the live subtitling field, the paper goes on to describe the SpeakTitle project, in which speech recognition technology is used to produce real-time subtitles for live broadcasts in the UK (Hewitt et al., 2000 and Bateman et al. 2001).

1.1   *Background*

Technology to enable transmission of "closed captions" (subtitles which are only visible on the picture with the aid of special decoder circuitry) was developed independently in the US and the UK during the 1970s (reference: US standards 1991, 1992, 1999, 2000 and history, UK standards 1975, 1976; WGBH). In 1982, the National Captioning Institute in the US started producing real-time captions for live programmes using specially trained court reporters to input the text as phonetic codes on special stenographic keyboards (reference: National Captioning Institute). The codes were converted back into conventional text using transcription software with phonetic-to-English dictionaries (for general information, see reference: Stenograph).

The teletext standard adopted in the UK and Europe (reference: UK Standards, 1975, 1976) provided for closed captioning or "optional subtitling" by reserving one or more page numbers to carry subtitle services. The first regular UK subtitling service was launched on Independent Television in 1979, and regular live news subtitling started in 1987. Alternative approaches for real-time input implemented in the UK involved using normal QWERTY keyboards (up to around 90 wpm sustained) or Velotype syllabic chord keyboards, on which two or more keys may be pressed simultaneously (up to around 120 wpm sustained), to provide an edited version of the spoken word (for general information, see reference: Velotype). Specialist techniques such as the use of "shortforms" (automatically expanding predefined abbreviation

codes for the names of people and places and hard-to-type words) were developed. Multiplexing two slower operators in tandem, each transcribing alternate utterances, allowed high quality "edited" real-time subtitling to be produced.

The problems with this approach are that it is labour-intensive (two multiplexed operators are needed for best results using QWERTY or Velotype), there is a shortage of trained Velotype operators and training times are extended (one year or more), and it does not easily suit all programme types (for example**,** sports where there are many different player names). As an alternative, stenographic keyboards and associated transcription software were also adopted in the UK during the 1990s. However, the use of stenography is also problematic for the reason that operators are in short supply and require 2 or 3 years of training and experience to achieve the necessary speed and accuracy.

1.2   Possibilities for speech input

Since the early days of real-time subtitling, it had been foreseen that speech recognition could eventually provide a viable future text input modality. While affordable recognition technology was still in its infancy, Damper, Lambourne and Guy had in 1985 proposed using speech input as an adjunct to keyboard entry in television subtitling (Damper et al. 1985). The system used a series of simple restricted speech commands to control the position and style (principally colour) of live subtitles entered on a QWERTY keyboard, thus enabling the operator to focus maximum effort on text entry. Early trials demonstrated the difficulty of using speech and keyboard input simultaneously at the same workstation due to keyboard noise affecting recognition.

Once speech recognition technology reached the point that an affordable system could deliver near real-time transcription of continuous speech from a trained speaker, it was worth seriously investigating its application to live subtitling. Production of acceptable subtitles by direct recognition from the TV soundtrack was judged to be unfeasible for a number of reasons: interference from background music or noise; likelihood of multiple simultaneous speakers; the need for highly accurate speaker-independent recognition in real time; and the need to control style and position. The technique of speech input by a trained "editing re-speaker" (hereafter referred to as the Speaker) was therefore chosen. The Speaker would be trained in the use of the speech recognition system, would train the recognition system to recognize his/her voice, and would develop vocabulary appropriate to the expected subject matter. The purpose of this project was therefore to investigate whether and how real-time speech input could be used to create acceptable subtitles for live TV broadcasts, and to create the necessary interface tools to facilitate it.

Speech-based real-time subtitling services

The research project was initiated in 1998 between a broadcast software development company Synapsys Ltd (now SysMedia Ltd) and the University of Hertfordshire, under the auspices of a DTI (Department of Trade and Industry) LINK scheme (reference: LINK, 1998). This was a government-sponsored scheme which partially funded broadcasting technology initiatives.

1.3 Project goals

This use of speech recognition technology particularly focuses on the problems of real-time recognition (i.e., the speech engine must deliver a transcript with minimal delay) and high accuracy. This can be supplemented by having a second operator to catch last-minute errors and rapidly correct them. It was hoped that the technique would enable operators to be selected and trained far more rapidly than for the fast keyboard technologies. If quality criteria were met, the use of speech recognition would provide an alternative to keyboard technologies for entering text in real-time for live subtitling purposes – and indeed for other transcription areas.

The tasks and goals were defined as follows:

- assess the suitability of speech input for various programme genres
- assess different speech engines and choose a suitable candidate
- devise a suitable software interface to the selected speech engine
- devise a suitable user interface for the speech subtitling system
- reach an acceptable quality threshold in the recognized text
- cope with changing vocabulary such as names of sports teams
- deliver subtitle text with an acceptably low throughput delay

During the course of the research and development different models for the production and correction of text were assessed. Initially it was proposed that two people would be needed – one listening, editing if necessary and re-speaking (the Speaker) one correcting any errors (the Corrector) – but by the end of the work the recognition accuracy for suitable programme material was judged to be high enough to dispense with the Corrector. Subtitles were typically presented in "scrolling mode" rather than the traditional "block mode" in order to minimise the delay between an utterance and the appearance of the words on-screen.

Speech-based real-time subtitling services

The results of the work are embodied in a new product called "SpeakTitle" which utilizes a commercial recognition engine to produce real-time teletext subtitles for live programmes. It needs to be operated by a trained Speaker, who in turn has trained the recognition engine to recognize his/her voice, and in a suitably quiet acoustic environment. In addition, performance is enhanced if the recognition engine is supplied with topic-specific vocabulary files, since different topics will not only use specific words but they may use combinations of words in different ways. Given these provisions, SpeakTitle is being used successfully to subtitle a variety of sporting events and other live programmes. It has enabled television companies to widen the pool of real-time subtitle production staff and thus increase the potential for subtitling an increased number of live broadcasts.

This paper describes how the project met the challenges of real-time speech recognition and live subtitling. It outlines how the project was adapted to achieve television guidelines on subtitle acceptability (see section 1.4) as well as the philosophy of the approach adopted. The paper also describes the evolution of the product and the experiences of users in developing the system in a live subtitling environment.

1.4   Design criteria for speech-based subtitling

In order to set up and operate a service delivering speech-based TV subtitles that will be useful to viewers, general criteria for successful real-time subtitling need to be met (reference: ITC 1999). Such criteria are not hard-and-fast, since it is clear that by definition the production of subtitles in real time cannot be perfect since text entry will always take a finite time, and there is little opportunity to correct errors.

The general quality and operational criteria were:

> 1) Accuracy : the recognition accuracy needs to be around 97-98%. Assuming an average of 14 words in a 2-line subtitle, this still equates to an error roughly once in every three or four such subtitles.
> 2) Throughput : where picture events and sound are tightly in synchrony (ignoring lip-sync: focusing on subject-matter) a throughput delay of more than about 5-6 seconds can be problematic for the viewer.
> 3) Style control : if multiple speakers are being subtitled, it is desirable to be able to control the colour of individual subtitles in order to reflect speaker identity.
> 4) Position control : if the centre of visual interest can vary its position on the screen, then it is desirable for subtitles to be moved to avoid obscuring this area.

5) Ease of use : to sustain an economic service, it is important to be able to select and train new staff to be subtitle Speakers relatively straightforwardly. Fast keyboard operators can take around 1 year (Velotype) or 2-3 years (Stenograph) to train and become productive; speech subtitlers would ideally be productive in a matter of 2-3 months.

6) Flexibility : to be able to respond to changing subject matter, it should be possible to add new specialist topics and new vocabulary with the minimum of overhead and complexity.

2   The SpeakTitle system

A range of speech recognition systems were investigated, with reference to the results of the 1997 National Institute of Standards and Technology (NIST) evaluations (Pallett et al., 1997). It was found that recognition results were most favourable for the IBM ViaVoice Executive system with recognition rates of 95-98% consistently recorded in trials based on trained Speakers reading text at 150 wpm, where there were no out-of-vocabulary words (reference: ViaVoice®).

The original SpeakTitle system was designed for use by two operators, the Speaker and the Corrector. The Speaker listened to the live programme and spoke the subtitle text while the Corrector corrected the output before it went on air. It was subsequently found that an experienced Speaker could achieve recognition rates without correction that were acceptable for live broadcasts, and the systems currently in use do not utilize a Corrector.

2.1   System Overview

A development system was built which incorporated a video recorder so that television programmes could be repeated for experimental purposes. An overview of the development system is given in Figure 1. It is designed to be operated by a Speaker and a Corrector.

A television programme is played on the video recorder, simulating a live feed from the television cameras used in the operational system. The programme is transmitted to the Speaker via his workstation and it is also output on a TV monitor accessible by the Corrector. The Speaker listens to the television programme on a headset, and can watch the programme in a video image window on the Speaker workstation. The Speaker repeats what has been said, s/he may need to apply a degree of editing and précis in cases where the television programme includes rapid streams of speech which cannot not be converted to subtitles in an acceptable timeframe.

Speech-based real-time subtitling services

In order to make the subtitles acceptable to the viewers, they must contain punctuation, and the Speaker may either speak punctuation or utilize a touch screen monitor (the Speaker Interface, described in section 2.2) to generate the punctuation words which are incorporated into the audio stream prior to it being passed to the recognition engine. The output from the Speaker thus comprises a single stream incorporating the spoken subtitles augmented with punctuation commands. This output can be split, with one stream passed directly to the ViaVoice recognition engine in the Corrector's workstation, and one to the Corrector's headset – this allows the possibility of inserting a delay in the stream to the Corrector which can be adjusted so that the Corrector hears the output from the Speaker simultaneously with it showing on the computer monitor. The delay corresponds to the amount of time taken for the voice recognition engine to process the speech.

The Corrector watches the output on the computer monitor and corrects any errors before the subtitles are passed on to an existing subtitling system such as WinCAPS (reference: WinCAPS 2003). This uses a teletext inserter card to merge the subtitle data with the television video signal for decoding on an ordinary teletext TV set.

Figure 1: The Design of the Speaktitle Development System
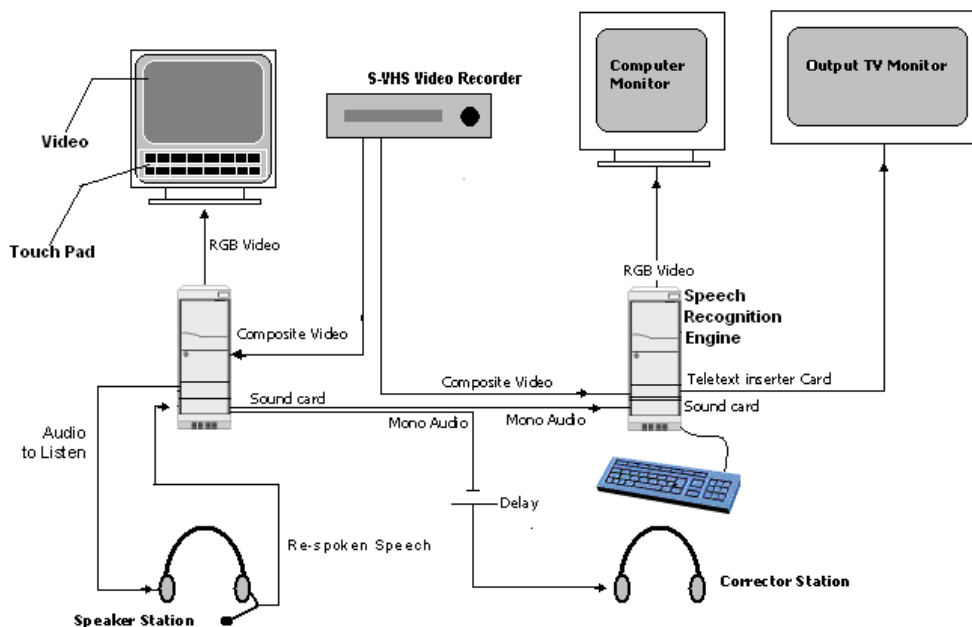
The Speaker Interface, described in detail in 2.2, gives the Speaker various options such as the use of buttons for punctuation and macros rather than speaking commands as is necessary with the basic speech recognition system. The Corrector Interface presents the

text in a scrolling window from which it can be edited in a short time frame before going out 'on air'. This interface incorporates software to amend the output according to the presenter's 'house-style', for example the use of upper or lower case characters for words such as "North-East" and the addition or removal of hyphens. It is also possible to filter out any offensive words that might otherwise escape the notice of the Corrector.

The Corrector Interface incorporates three scrolling modes – "elastic", where text that is being edited is not sent out until the edit is completed, "semi-elastic" which allows text to be delayed for only a limited amount of time, and "bulldozer" where text is sent out continuously, edited or not. A variation of this last mode, the "pass-through" sends text out directly without any editing, although house-style changes are still applied. This has been employed successfully in live subtitling of snooker with reported error rates of only 2-3%. It is recognized, however, that for faster paced programmes, and ones where a high degree of accuracy is required, such as parliamentary interviews, a Corrector might be required.

To deal with the situation in which one Speaker is re-speaking the words of more than one television speaker, such as in an interview situation, the possibility of colour coding the text output needed to be explored as a way to identify the different speakers to the viewer.
As only one human operator is responsible for both listening to and then repeating material to be transcribed into the speech recognizer the input to the speech recognizer does not convey the information required to colour-code the subtitles automatically. Automatic colour-coding using the original speech would require two recognition systems (original speech and Speaker) to operate independently. They would need different and varying operational speeds, different inputs in terms of speakers' voices, and, in some instances, in terms of the text content. As a result, it would not be possible to synchronise the output of the speech recognizer with that of the colour coding (speaker discrimination) system.

Two techniques have been developed to deal with this. The first method uses special speech "macros" which produce commands that can be interpreted by the subtitling system. The second uses buttons on the Speaker interface.

A further development which has been designed to improve the accuracy of the system is the use of 'Topics' for specific domains (described in 2.3). Here, accuracy has been improved by integrating specialised language models into the system. As reported above, acceptable recognition results for the IBM ViaVoice system were achieved with trained Speakers where there were no out-of-vocabulary words. The SpeakTitle system aimed to address the treatment of out-of-vocabulary words using specialist language models or "Topics".

Speech-based real-time subtitling services

In the speech recognizer, the acoustic processor produces a set of ranked candidate words from the acoustic signal. The language model then provides information on the probability of a given word or phrase occurring in the context, and these two sources of information are combined to give the output words. In specific domains, such as a particular profession, specialized vocabularies are used. In domains such as a sports commentary, the vocabulary may be largely unchanged from general speech, but certain word patterns are likely to occur, for example "on the black" is a typical phrase heard in snooker commentary but seldom anywhere else.

2.2 The Speaker Interface

A prototype Speaker Interface has been developed which comprises several components (see Figure 2).

Figure 2 – The Speaker Interface
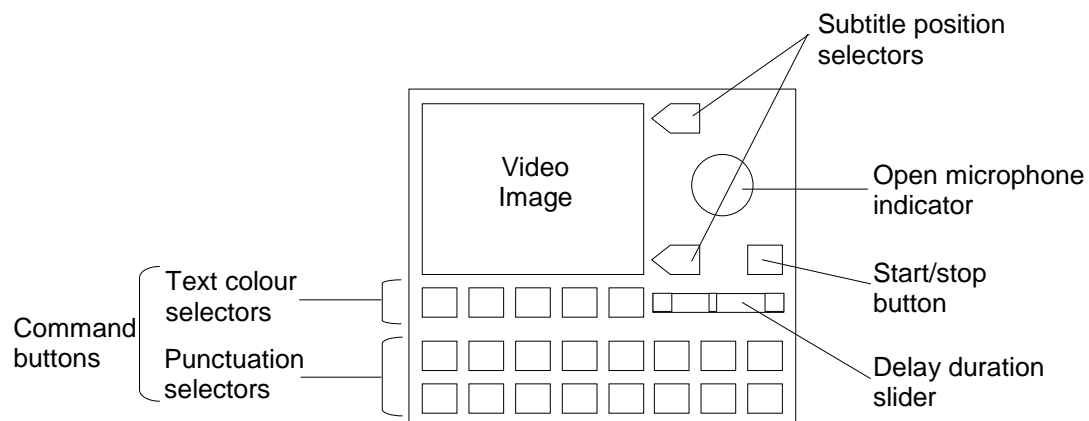
The Video Image window shows the live television broadcast (or, when used in test or development mode, a simulation from a video recording). The subtitle position selectors allow the Speaker to position the broadcast subtitle at the top or bottom of the picture so that it avoids any captions that are already present on-screen. The Open Microphone Indicator informs the Speaker when the system is ready to accept speech. The Start/Stop button activates and deactivates the system. The Delay Duration Slider allows the Speaker to pre-set a delay in the audio signal that is sent to the Corrector. This is a key feature, since it permits better synchronisation between what the Corrector hears (i.e. a delayed version of the dictation), and the corresponding appearance of the recognised text output from ViaVoice which they are to review and correct. With the development system this delay was typically tuned to between one and two seconds. The Command Buttons allow the Speaker to insert

punctuation or macros into the audio signal that is sent to ViaVoice. These buttons can be customised by the user and typically include common punctuation and macros to change the colour of subtitles.

The Delay Mechanism and Command Inserter were originally implemented as two separate pieces of hardware. The delay was provided by a broadcast profanity delay – an expensive piece of equipment designed to be used primarily to avoid swearing going out on the air in live situations. The SpeakTitle system used it in a mode in which it provided a fixed duration delay. The commands were generated by a laptop computer running a simple program that played back a pre-recorded audio file from disk whenever a pre-determined key was pressed.

Both the Delay Mechanism and the Command Inserter functions are now implemented in software on a single desktop computer. Figure 3 shows the multiple buffer mechanism that is employed. This consists of a capture buffer that records a mono audio signal, a Corrector playback buffer and a Speaker playback buffer. The Corrector playback buffer has its output panned to the left channel of the computer's stereo audio output and the Speaker playback buffer has its output panned to the right channel. This allows the audio outputs of these buffers to be routed to different destinations. An array of command sample buffers is created to hold the pre-recorded punctuation and macro audio command files.

Figure 3 – The Delay Mechanism and Command Inserter
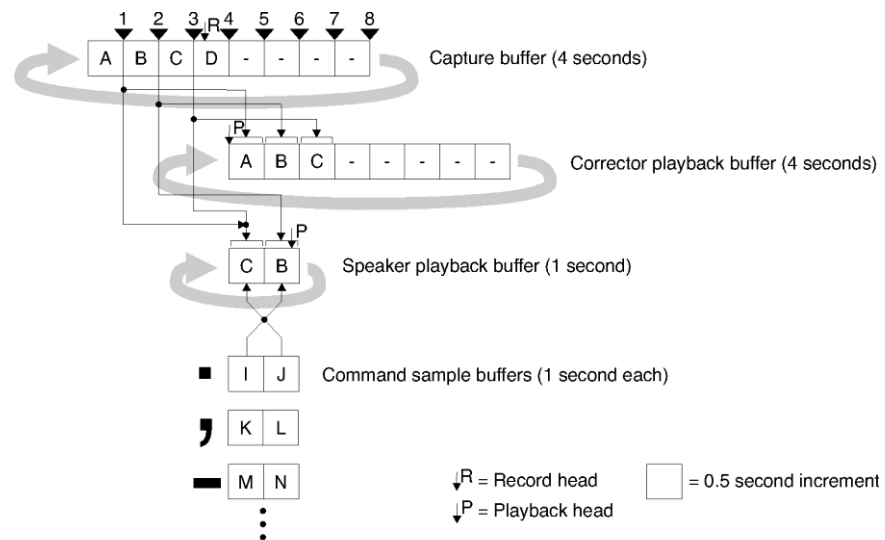
When the system is active, the Speaker's speech is recorded in the capture buffer. The capture buffer is divided into segments. The size of the segment can be varied, but the minimum size is limited by the speed of the host computer. Tests with a 733 MHz Pentium III

show that four increments per second is the most that can be achieved without loss of continuity. As soon as a segment has been recorded, it is immediately copied into the Corrector playback buffer and the Speaker playback buffer.

If the Speaker playback buffer is not already playing, then playback is started. Playback of the Corrector playback buffer only commences when the buffer is full, hence the length of this buffer determines the duration of delay to the signal that the Corrector hears. All the buffers wrap around. Once they are full, the insertion of data is started again at the beginning, and when the record and playback "heads" reach the end of the buffer, they are immediately returned to the start. When the Speaker presses a command button, the system is notified. The next two copy operations for the Speaker playback buffer are then sourced from the relevant sample buffer, not the capture buffer. In this way, the command output is passed to ViaVoice, but the Corrector does not hear it.

2.3   *"Topics" - Specialized Language Models*

The use of specialized language models can improve the accuracy of the speech recognizer. The main language model is an integral part of any speech recognizer. It takes the candidate words from the acoustic processor and ranks possible output in order of probability. Thus "a tax on petrol" would be more likely than "attacks on petrol". In order to do this two components have to be processed. First there is the set of single words that comprise the vocabulary of the language model (LM). The second component of the LM is the set of trigrams, three adjacent words, that define contexts in which words are likely to occur. These two components constitute the training data for the LM. Now, the main language model can be augmented by a specialized LM, also called a Topic, developed for a particular subject field, and then integrated with the main LM. Via Voice has the facility to develop such Topics, and they have been successfully used in the subtitling of live TV sports programmes – snooker, golf, tennis, football (soccer), athletics. The Topic can be switched in or out as required.

In customizing a language model by integrating a Topic there are two steps. First, single words have to be added to the vocabulary. Topics developed for particular professional use, such as legal or medical domains, would have appropriate technical terms In sporting domains the names of players are specific to each sport and need to be added to the vocabulary. This may be done by the user who can update the main vocabulary right up to the last minute. Some names can be accepted through spoken input, but it may be necessary to use a phonetic transcription facility, described below, particularly for foreign names. However, if the addition is made to the main vocabulary it will then be a permanent word there. If

names are added to the Topic, which is prepared in advance, they can be used when relevant only, as the Topic can be switched in and out.

The second component of the Topic is the domain specific set of trigrams. In sporting domains, the specialized vocabulary is often quite limited, apart from names of players. Few words occur that are not in the base vocabulary of the speech recognition system. However, there are characteristic combinations of words that seldom occur in general text or speech. For instance in football commentary trigrams like "a free kick", "hit the post", "yellow card for" occur frequently, yet they did not arise in 2.5 million words of TV chat show speech. In tennis we need to avoid solecisms such as number to court. The words themselves in these examples are not peculiar to the football or tennis topic, but their combinations are.
In speech recognition the "sparse data problem" is a key issue (Gibbon et al.,1997; Ney et al., 1997). This is the observed phenomenon that a small number of words occur frequently, but most words occur rarely - a zipfian distribution. This phenomenon is more pronounced for bigrams and trigrams. For instance in 39 million words from the Wall Street Journal, 77% of the trigrams have only occurred once. If a new passage in this limited domain is considered, most of the trigrams will not have occurred before (Gibbon et al., p.258). By modelling the domain characteristics, we begin to address this problem.

In developing Topics for sports, the relevant trigrams for training were obtained from data supplied by the broadcasters. Transcribed commentaries that had previously been broadcast were taken and processed into sets of trigrams. As much data as possible was collected, with a target of a million words, though we usually had to make do with less. This is an arbitrary figure, based on empirical investigations, but in general larger amounts of training data are desirable so that more trigrams can be captured. The training data is then cleaned, if necessary. Cleaning can include excising extraneous comments and removing glaring errors that are not wanted as models. The IBM Via Voice Topic Factory tool is used to process the training data, and produce the Topic which can be integrated with the main language model.

2.3.1   Example from commentary on snooker

To illustrate these issues, we report on an example from live TV commentary on snooker. The data used is from stenographers' broadcast output, produced in real time, and therefore occasionally noisy. The figures given are rounded to avoid spurious precision. A small corpus of snooker commentary, 59K words, has been compared to a base corpus of 2.5 million words from TV chat shows. A control experiment with another 59K corpus from new chat shows is also used. We see how many single words, word pairs and word trigrams are unique to the snooker corpus and compare these figures to those from the control corpus (Table 1).

Speech-based real-time subtitling services

If we exclude names, numbers and errors, there are nineteen words in the snooker corpus that did not occur in the base corpus, e.g. terms like "missable" and "pottable". Compared to the control corpus, there are few words that are in the snooker corpus but not in the base corpus. However, there are a comparable number of new trigrams. Phrases like "behind the red" or "the opening pot" occur frequently. There are characteristic constructions, typical of the domain, that can be exploited. See Table 2, but note "words" include names, numbers and errors.

Table 1 - Statistics of the corpora (including as words names, numbers and errors)

	Base corpus	Snooker	Control
Number of words	2,592K	59K	59K
Distinct words	53K	4K	7K
Distinct bigrams	725K	27K	34K
Distinct trigrams	1737K	48K	51K

Table 2 - Numbers (rounded) of all words, bigrams and trigrams not in base corpus (including as words names, numbers and errors)

	Snooker	Control
Words, frequency $\geq$ 2, not in base corpus	90	270
Bigrams, frequency $\geq$ 2, not in base corpus	1200	1300
Trigrams, frequency $\geq$ 2, not in base corpus	2000	2300

The training data are presented to Topic Factory, and any words that need phonetic representations are identified. The developer may be able to get a correct phonetic representation accepted by speaking the word. Otherwise, the phonetic representation has to be typed in, using a mapping in which phonetic symbols are mapped onto ASCII letter

combinations. After this is done, Topic Factory will process the prepared training data to produce the customized LM.

In the past, language models have typically been evaluated by assessing the perplexity of the model on test data. However, recent work indicates that direct word error rates may be a more useful metric (Clarkson et al., 1998), and that is the measure utilized here, using the CRER (Composite Record of Errors in Recognition) analyzer described below. Preliminary experiments indicate improvements of 1-3% when adding a sporting Topic.

Other work in this field, such as the development of story topics for stories in similar domains, has used single word similarities as a basis for clustering texts (Seymour et al., 1997). The snooker example illustrates how the use of word combinations such as trigrams is more powerful.

3   The System in Use

The CRER assessment tool was developed to enable more consistent measurement of accuracy, clearer representation of results and the ability to make a more detailed investigation. This enabled an analysis of the Speakers' performances under different conditions. This analysis and the experience of Speakers working in a live environment have in turn contributed to improvements in the SpeakTitle system.

3.1   CRER (Composite Record of Errors in Recognition) Tool

The CRER software is an analytical tool which compares scripts of actual spoken input and speech recognized output. It provides measurements of overall accuracy (defined as number of words minus substitutions, deletions and insertions) and correctness (defined as number of words minus substitutions and deletions), as well as a range of detailed error statistics for substitutions, deletions and insertions of words. An example of the difference between accuracy and correctness can be seen in the following illustration. If the word "away" is recognized as "the way" this gives one substitution and one insertion error. This would count as 2 errors on the accuracy metric, one on the correctness metric. Different types of errors are represented with different colour codes for ease of analysis, and all results are shown as percentages. This detailed and consistent analysis of live subtitling results has provided a base line measurement from which to test improvements made to the system with such developments as topics and the Speaker interface.

3.2   Accuracy

Preliminary tests were carried out to assess the initial relative accuracy of the basic ViaVoice speech recognition system (i.e. without the use of topics or the SpeakTitle Speaker Interface developments) with a range of untrained speakers and a range of input modes (the speaker either reading text or hearing speech). One objective of these tests was to assess whether certain speakers were better suited to the front-end of the live subtitling environment where abilities such as hearing/reading, comprehension and speaking would be utilized. With the Speaker utilizing this variety of different abilities in order complete the tests, it was felt that certain characteristics might be highlighted as producing greater accuracy. Another objective of these tests was to assess the basic accuracy level of a minimally trained speech recognition system with inexperienced Speakers, in order to identify the range of improvement needed to be achieved with the SpeakTitle system. These tests were carried out at the start of the project, in 1998, using an earlier version of ViaVoice than was subsequently employed.

Fifteen untrained Speakers were asked to read out loud strings of words where the input source was either text input or audio input. The rate of speaking for the audio input tests was gradually increased thus permitting less time between the speaker hearing and repeating the words. Not only did the findings show, as might be expected, that the read text input provided the greatest accuracy overall (92%) but that highest accuracy was found with audio input when the speakers had more time between hearing and saying the words (average of 83%). However, the findings highlighted not so much speaker variation or speaker suitability for online subtitling, as was initially anticipated, but the importance of training. The recognition accuracy needs to be around 97-98% for a live-subtitling environment, and the earlier quoted figures of 95-98% being found using ViaVoice were reliant on both the level of ViaVoice training and the avoidance of out-of-vocabulary words.

The development of the Speaker Interface together with the use of Topics for the SpeakTitle system (described above in 2.2 and 2.3), has not only improved accuracy, and therefore the viability of automated subtitling of live broadcasts, but has also provided the users of the system with a more manageable and easy-to-use interface.

The system was adopted for use by a television company before the end of the project and a number of dedicated speakers were engaged to make it operational for particular types of live programme, particularly sport. Evidence that the system is now acceptable comes from the commercial decisions made to use it to replace the traditional methods of subtitling. It was not possible to divert commercial operations away from their prime functions to conduct objective

tests except on a very limited scale, but early reports on the accuracy of the systems in real-time live use with dedicated and well-trained operators have already given figures in the range of 98%.

3.3   Experience of Speakers

There were certain findings gained from the user perspective and the experience of running the system in a live subtitling environment that have contributed to further development of the SpeakTitle system to improve ease of use and accuracy. For example, speaker training at the speed of expected normal delivery and without undue hesitation was found to give improved recognition rates. Setting up the microphone before a live session and configuring the SpeakTitle for throughput (i.e. setting transmission rate and maximum lines of text visible on-screen for editing) before output to transmission buffer have both improved accuracy.

It has been found that no special microphone or sound card equipment is required to reach satisfactory recognition levels. Most modern computers have an adequate sound card that is compatible with the SoundBlaster sound card standard, and the Andrea NC-61 microphone shipped with ViaVoice gives as good recognition as more expensive microphones in normal situations. Whilst ViaVoice will take background noise into account when performing speech recognition, a consistent background level, or ideally a silent acoustic environment, gives better results. A specialised Proximity Microphone is more suitable in environments where there is variable local noise, such as other speakers.

Throughput has been further improved with the fine-tuning of the speech system for best throughput balance in a number of ways: For example, setting an appropriate words per minute (WPM) rate for delivery of subtitle text to WinCAPS and setting the maximum number of lines in the Edit box so that the text does not build up have both been effective. The identification at the outset of which subtitle presentation style best meets the needs of the transmission can provide benefits to throughput. The subtitles may be displayed in block-mode where the whole subtitle is displayed at once or scrolling mode where the subtitle is displayed word by word. On a practical level, and from the speakers' perspective, having clear rules on error correction greatly increases throughput. For example knowing to correct only text that seriously detracts from the meaning intended in the spoken text, rather than all errors, can speed up the editing process.

Other SpeakTitle features enable the fine-tuning of the system for specific treatment of homophones, shortforms and the text processing required for specific in-house styles pre-

defined by the user. A greater or lesser degree of editing can also be controlled by the system.

From the staff selection and training perspective a "good" Speaker has been defined as one that gives a consistent delivery with fully pronounced words. Slight accents seem not to lower recognition accuracy, and tests completed with a range of Speakers from different backgrounds (see 3.2 above) found little variation in accuracy due to age, sex, ethnic origin, education level, vocal tone, volume, experience of dictation, speed or even first language other than English.

4   Conclusions

The main achievement of the project has been the development of an effective working system for live-subtitling. The SpeakTitle project has developed from an initially basic system which utilized a standard speech recognition package together with standard speaker and correction modules to one which can be finely tuned to the user requirements in terms of 'topics' and speaker interface requirements. This has provided a useable working system which is now used live on-air for three major UK broadcasters, giving sufficiently accurate results and providing hearing impaired viewers with access to more live television.

The development of CRER as a tool for assessing the performance of continuous speech recognizers has been another valuable outcome from the project. It provides a consistent measure for assessing continuous speech recognizers. Work is now under way to utilize this tool more extensively to assess the on-air results of more and more 'live subtitling' which is being undertaken by the two major users of SpeakTitle. The results from this exercise will direct future work on further improving the accuracy levels towards human perception levels.

A number of future applications of and spin-offs from the 'live subtitling' technology developed for the SpeakTitle project are being explored. The presentation on screen of the text of lectures, meetings and telephone conversations are other areas for further evaluation, and tests are already underway to assess the suitability for lecture situations. The obvious advantages to the hearing-impaired of converting from audio input to text output might also be applicable to viewers who do not have English as their first language and who may find text easier to follow than spoken words. The lessons learnt during the process of achieving real-time subtitling for English and the processes developed as part of this project could be applied in the future to languages other than English, work towards this end has been done in Japan (reference: NHK). One further application of the SpeakTitle technology could be in the field of translation where a bilingual Speaker listening to output spoken in one language could provide text output in another. These areas are being further explored based upon the results

Speech-based real-time subtitling services

of the CRER analysis and experiences of running subtitling in a live setting for major British broadcasters since summer 2002.

References

Bateman, A, Hewitt, J, Lambourne, A, (2001) Subtitles from Simultaneous Transdiction: Multi-modal Interfaces for Generating and Correcting Real-time Subtitles", HCII2001, New Orleans.

Clarkson, P. and Robinson, T.(1998). The Applicability of Adaptive Language Modelling for the Broadcast News Task. *Proceedings of ICSLP. pages 1699-1702,* Sydney, Australia.

Damper, R. I., Lambourne, A. D., and Guy, D. P. (1985) Speech input as an adjunct to keyboard entry in television subtitling . In Shackel, B., Eds. Proceedings Human-Computer Interaction---INTERACT'84, pages 203-208.

Gibbon, D., Moore, R. and Winski, R. (eds) (*1997).* Handbook of Standards and Resources for Spoken Language Systems, Chapter 7, *Mouton de Gruyter.*

Hewitt, J, Bateman, A, Lambourne, A, Ariyaeeinia, A, Sivakumaran, P. (2000)
Real-Time Speech Generated Subtitles: Problems and Solutions. 6th International Conference on Spoken Language Processing ICSLP 2000, Vol III.

ITC guidance on standards for subtitling (amended February 1999):
    http://www.itc.org.uk/itc_publications/codes_guidance/standards_for_subtitling/index.asp

LINK: (1998)
    The Use Of Speech Recognition In Live TV Subtitling, LINK Project No. GR/M15958/01, 1/10/1998 – 30/9/2001

    Overview of LINK Project:
    http://homepages.feis.herts.ac.uk/~nehaniv/idmf/abstracts/hewitt.doc

National Captioning Institute: http://www.ncicap.org/acapintro.asp

Ney, H., Martin, S. and Wessel, F. (1997). Statistical Language Modelling Using Leaving One Out. In Young, S. and Bloothoft, G. (eds) Corpus Based Methods in Language and Speech Processing*: Kluwer Academic.*

NHK 2002 – http://www.nhk.or.jp/strl/open2002/en/tenji/id03/03.html

Pallet D.S. et al. (1997) Broadcast News Benchmark Test Results: English and Non-English, Proc. DARPA Speech Recognition Workshop 1997.

Seymour K and Rosenfeld, R. (1997) Using Story Topics for Language Model Adaptation *Proceedings of Eurospeech97*

Speech-based real-time subtitling services

Sivakumaran, P, Fortuna, J, Ariyaeeinia, A M, (2001) On the use of the Bayesian Information Criterion in Multiple Speaker Detection, *Proceedings of Eurospeech2001*.

Sivakumaran, P, Ariyaeeinia, A, Fortuna, J (2002) An Effective Unsupervised Scheme for Multiple Speaker Detection, *ICSLP2002, Topic 16*, Denver, Colorado.

Stenograph: http://www.stenograph.com

UK legislation:
     Broadcasting Act 1990 (c. 42) section 35, HM Stationery Office UK
     Broadcasting Act 1996 (c. 42) section 20(3)(a), HM Stationery Office UK
     Statutory Instrument 2000 no 2378 : Broadcast (subtitling) order 2001, HM Stationery Office UK

UK standards:
      Unified Standard April 1974, BBC Engineering Sheet 4008(5), Oct. 1975
    Joint IBA/BBC/BREMA Publication : Broadcast Teletext Specification, September 1976

US legislation
   Television Decoder Circuitry Act of 1990, US Congress
   Telecommunications Act of 1996, US Congress
   Federal Communications Commission Rule 79 – Closed Captioning of Video Programming, updated 2001

US standards and history:
FCC Report and Order FCC 91-119 1991
    FCC Memorandum, Opinion and Order FCC 92-157 1992

    EIA / CEA-608-B : Recommended Practice for Line 21 Data Service, 31 Oct 2000, see http://www.ce.org/standards/standard_details.asp?id=270,
    EIA-708-B for Digital Television Closed Captioning, 29 Dec 1999, see http://www.ce.org/standards/standard_details.asp?id=249, Electronic Industries Association, Engineering Department, 20001 Pennsylvania Avenue, N.W., Washington, D.C. 20006

    http://www.robson.org/capfaq/caption-charset.html).

    http://ncam.wgbh.org/resources/icr/line21hist.html
    http://main.wgbh.org/wgbh/pages/mag/services/captioning/

Velotype:
    http://www.velotype.com/

    http://www/velotype.nl/

ViaVoice®: http://www.ibm.com/software/speech

WinCAPS: (2003) SysMedia Ltd at
http://www.sysmedia.com/subtitling/pdfs/wincaps_multimedia.pdf

List of Figures and Tables

Speech-based real-time subtitling services

Table 1 - Statistics of the corpora

Table 2 - Numbers (rounded) of words, bigrams and trigrams not in base corpus