BBC

R&D White Paper

WHP 065

July 2003

Speech Recognition in Assisted and Live Subtitling for Television

M.J. Evans

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 065

Speech Recognition in Assisted and Live Subtitling for Television

M.J. Evans

Abstract

For several years the BBC has been using speaker-independent speech recognition to assist television subtitling and dramatically reduce the time taken for a skilled subtitler to complete a programme. This paper describes recent developments in this technology. Our new system for robust generation of subtitles for live television is also introduced. This system uses speaker-specific speech recognition and novel networking software to allow subtitlers to work from home or other remote locations whilst permitting straightforward handovers between subtitlers or television services.

This was an invited paper published by the Technological Advancement Organization of Japan in the Proceedings of the TAO Workshop on Universal Design for Information, Communication and Broadcasting Technologies, Tokyo, 7 June 2003

Key words: speech recognition, television subtitling,

Speech Recognition in Assisted and Live Subtitling for Television

DR. MICHAEL J. EVANS

*Research and Development Department, British Broadcasting Corporation,
Kingswood Warren, Surrey KT20 6NP, United Kingdom*

For several years the BBC has been using speaker-independent speech recognition to assist television subtitling and dramatically reduce the time taken for a skilled subtitler to complete a programme. This paper describes recent developments in this technology. Our new system for robust generation of subtitles for live television is also introduced. This system uses speaker-specific speech recognition and novel networking software to allow subtitlers to work from home or other remote locations whilst permitting straightforward handovers between subtitlers or television services.

0 INTRODUCTION

The BBC is committed to subtitling all of its television programmes by 2008. This represents a huge amount of programming, as well as a huge range of types of programming; from coverage of live events, to drama programmes scripted and recorded several months in advance of broadcast. Subtitles for BBC programmes are produced by skilled subtitlers, who compose sensibly condensed, helpfully coloured and well formatted subtitles to accompany the programmes. Subtitles are delivered using teletext for analogue services and DVB for digital TV.

In the year 2001/02, the BBC broadcast approximately 44000 hours of TV, of which about two-thirds was subtitled. Within our multichannel services, the proportion ranged from 73.7% of programmes on the general interest, multiplatform services BBC1 and BBC2 to 26.6% of the digital-only service BBC News 24. The BBC Parliament service was completely unsubtitled during this period. Increasing these subtitling rates to 100% across the board places great demands on the BBC's subtitling capacity.

To support this increase in subtitling without compromising quality and, in particular, the BBC's commitment to use skilled subtitlers, BBC Research and Development has developed a range of software tools to allow subtitlers to work more effectively. These tools take a number of different forms, to meet the range of styles of programmes that they support. In each case a key technology is speech recognition. The BBC, however, does not use speech recognition to automatically generate a transcript of a programme's soundtrack, and use this as the basis of subtitles. The multitude of different voices in a typical programme's soundtrack would require the use of speaker-independent recognition and the transcription accuracy of such systems remains low and overheads in post-correction of errors high. Also, part of the skill of the subtitler is the condensation of programme dialogue, to moderate the time taken by viewers to read the subtitle.

The type of speech recognition system, and the manner in which it is applied in aiding subtitling are dependent upon the availability of particular resources; namely a transcript of the programme and a recording of it. The implications are summarised in the table below:

Recording Available?	Transcript Available?	Type of Speech Recognition
Yes	Yes	speaker-independent
Yes	No	speaker-specific followed by speaker-independent
No	No	speaker-specific

This paper describes three systems developed by BBC Research and Development to increase the effectiveness of skilled subtitlers. Each uses speech recognition:

- *Assisted Subtitling* uses a speaker independent recogniser together with a transcript and recording of a programme to produce an initial set of subtitles for the subtitler to amend as necessary.
- *Script Capture* uses a speaker-specific recogniser to allow pre-recorded

programmes that do not have a transcript to be used with the Assisted Subtitling system.

- *Live Subtitling* allows subtitlers to use speaker-specific recognition to provide real-time subtitles for live programmes and programmes edited close to transmission.

1 ASSISTED SUBTITLING

1.1 Overview

Assisted Subtitling (or AS) is a software system allowing an operator to produce an initial set of subtitles from a recording of a programme and a textual transcript. The initial subtitles formed by AS can then be modified by a skilled subtitler. Overall, this process is much less time-consuming than having the skilled subtitler produce the subtitles without any automation.

The AS system consists of a sequence of functional modules. These extract relevant information from the recording and transcript of the programme, identify the time at which each word in the programme is spoken, and assign distinct subtitle colours to speakers. The final module produces well-timed, well-formed subtitles for the programmes, corresponding to the stylistic preferences set up by the user. Figure 1 shows a modular overview of the AS system.

An XML-based format called AML (Audio Markup Language) has been developed for Assisted Subtitling. AML files consist of sections which comprise, on completion of processing, a character list with assigned colours, a timed version of the script, a shot change list and the initial subtitles themselves.

1.2 Script Recognition and Analysis

Assisted Subtitling requires a transcription of the dialogue in a programme. Often, this transcription exists in the form of a verbatim post-production script. In general, the script consists, not only of dialogue, but also stage directions, descriptions of music cues and timing information. Extraction of the relevant information from the script is also complicated by the fact that scripts are typed up using a huge variety of different formatting styles.

Script Recognition extracts speaker names, spoken words and scene boundaries from post production script documents produced in Microsoft Word, Rich Text Format (RTF) or plain text. Any additional text in the document is ignored. (Text can also be originated from existing subtitles, typically from an acquired programme; in which case the AS system is used to ‘reversion’ the subtitles.) The system learns new script formats by means of a high-level graphical user interface which allows an operator to manually specify the nature (dialogue, stage direction etc.) of blocks of text from a small section of an example script. After manually marking-up 5-20% of a script, the system has usually learnt enough about the formatting conventions to recognise and interpret the contents of any script document sharing the same format, with a very high degree of reliability. Scripts with recognisable formats can then be directly converted into an unambiguous AML representation by the AS system.

1.3 Alignment using Speaker-Independent Speech Recognition

Alignment is the process of determining timing information for each word spoken in the programme. A programme’s soundtrack is processed by a speech recognition system. The recogniser does not need to

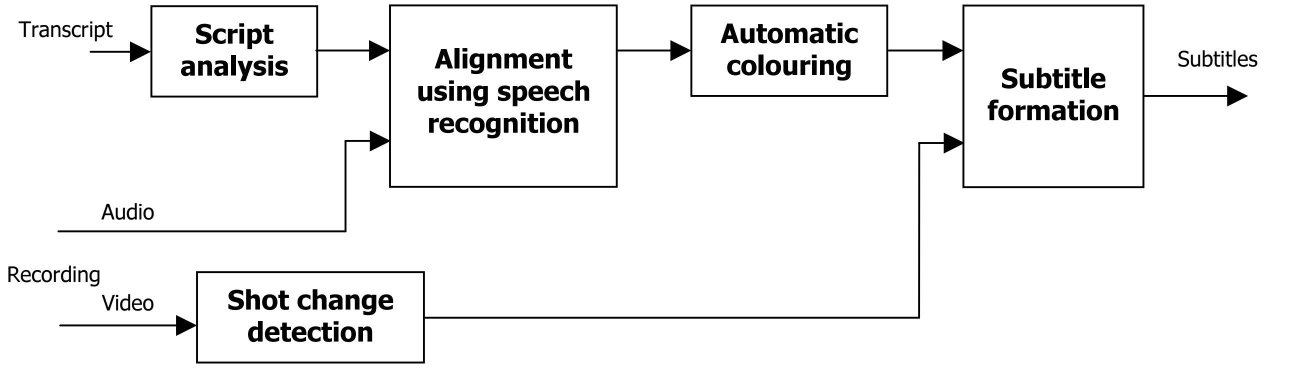


Figure 1: Modular structure of Assisted Subtitling

produce a transcription of the programme, since speaker and dialogue information has already been extracted from the script document. Instead, we simply require the recogniser to provide an accurate estimate of the time at which each word is spoken. This process is considerably more reliable than trying to produce a transcript using a speaker-independent recogniser.

Assisted Subtitling's alignment module uses the Aurix speech recognition software from 20/20 Speech Ltd [1]. Its accuracy in determining the timing of words spoken is normally well within the acceptable range for a subtitle. Alignment is the most time-consuming stage of the AS process. On a high-end PC, alignment is generally about 4-5 times faster than real-time; i.e. a 60 minute programme takes about 12-15 minutes to align.

1.4 Automatic Speaker Colouring

Subtitles on the BBC, and in Europe generally, make use of different colours to help viewers distinguish between the words spoken by different people. AS has a colouring module which automatically generates an optimum assignment of colours to speakers; minimising the incidence of interacting speakers sharing the same colour. The system supports different styles of subtitle colouring; including manual or exclusive assignment of some colours, different numbers and preference orders of colour, restrictions on the colours of interacting speakers, and whether a speaker must retain the same colour for the duration of the programme. For example, the general BBC subtitling style states that

- White, yellow, cyan and green are the available colours - with green the least preferred;
- Only white can be assigned to more than one speaker in a given scene;
- The last speaker in a scene cannot have the same non-white colour as the first speaker in the next scene;
- Colour assignments persist for the duration of the programme;
- In documentaries, yellow is (usually) assigned exclusively to the Narrator.

For a typical programme, automatically colouring to this specification takes a fraction of a second, whereas colouring manually can often be very time-consuming, and still fail to yield an optimal colour scheme.

1.5 Shot Change Detection and Subtitle Formation

Skilled subtitlers take careful note of shot changes when producing subtitles, to avoid distracting the viewers. Generally, subtitles should not be displayed during a shot change, as viewers tend to look away from the text at such instants. Similarly, subtitles should neither start nor end within a second of a shot change, as this is distracting. Starting or ending subtitles in synchrony with shot changes is preferable.

A shot change detector is included in Assisted Subtitling; analysing the video of the programme and producing a list of the times of each shot change. These times, together with the timed words determined by the alignment module and the colours assigned by the colouring module feed the final module; subtitle generation.

The words from the script are concatenated to form subtitles, but a large number of user-configurable subtitle preferences are also applied, so that the subtitles have the best possible overall appearance. The task is to optimise start and end times of each subtitle, together with the grouping of words between adjacent subtitles, and between lines in the current subtitle. Line and subtitle breaks can interrupt the grammar and slow reading, but subtitle timing cannot drift very far from the timing of the dialogue, subtitles must be displayed long enough to be read and, as already discussed, care must be taken around shot changes. The subtitle generation module in AS manages this complicated balance between timing, synchronisation and appearance to produce as good a set of subtitles as possible. The user-configurable preferences specify the relative importance of particular attributes in the finished subtitles. These subtitles are output in the form of a standard EBU format subtitle file. Figure 2 shows subtitles produced by the AS system.

1.6 Impact

The subtitles produced automatically by the AS system will not necessarily be ready for broadcast. Normally, the quality of the subtitles can be significantly improved by having a skilled subtitler using his/her experience in abbreviating the text and adjusting the timing and formatting of the subtitles. However, the AS system considerably reduces the time taken to subtitle a programme. Instead of between 12 and 16 hours subtitling time per hour of

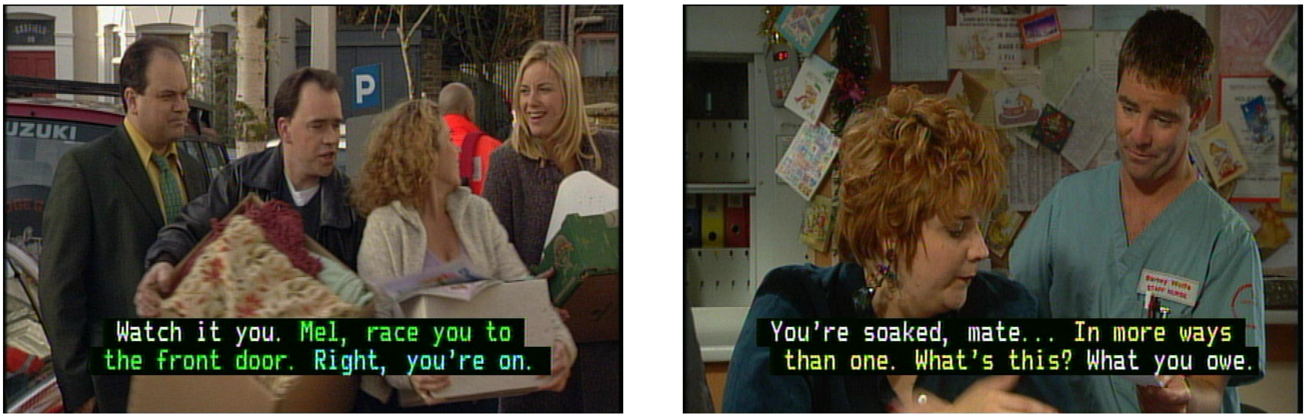

Watch it you. Mel, race you to
the front door. Right, you're on.
You're soaked, mate... In more ways
than one. What's this? What you owe.

Figure 2: Examples of subtitles formed using Assisted Subtitling

programme, the skilled subtitler might now be required for as little as 8 hours. Indeed, for some programmes, such as narrated documentaries, with very simple timing and speaker patterns, the time saving can be even more dramatic. Subtitlers, therefore, can increase the number of programmes they can work on in a week. Within the BBC, Assisted Subtitling has been in continuous use for three years and programmes processed using the system are broadcast daily.

2 SCRIPT CAPTURE FOR ASSISTED SUBTITLING

2.1 Overview

Assisted Subtitling is dependent upon an accurate transcript of the programme's dialogue. However, verbatim post-production scripts are not always available for single programme. A fast and reliable method for capturing a transcript of a programme is required. As discussed, applying speaker-independent speech recognition to the programme's soundtrack will not produce a sufficiently accurate transcript. However, we can employ a speech recogniser which has been trained specifically to recognise the words spoken by an operator. The operator 're-speaks' the programme's dialogue, flagging changes of speaker and scene as he/she goes. The result is an accurate transcript of the programme which, by virtue of being respoken alongside playback of the programme, is already partially synchronised with the programme's soundtrack

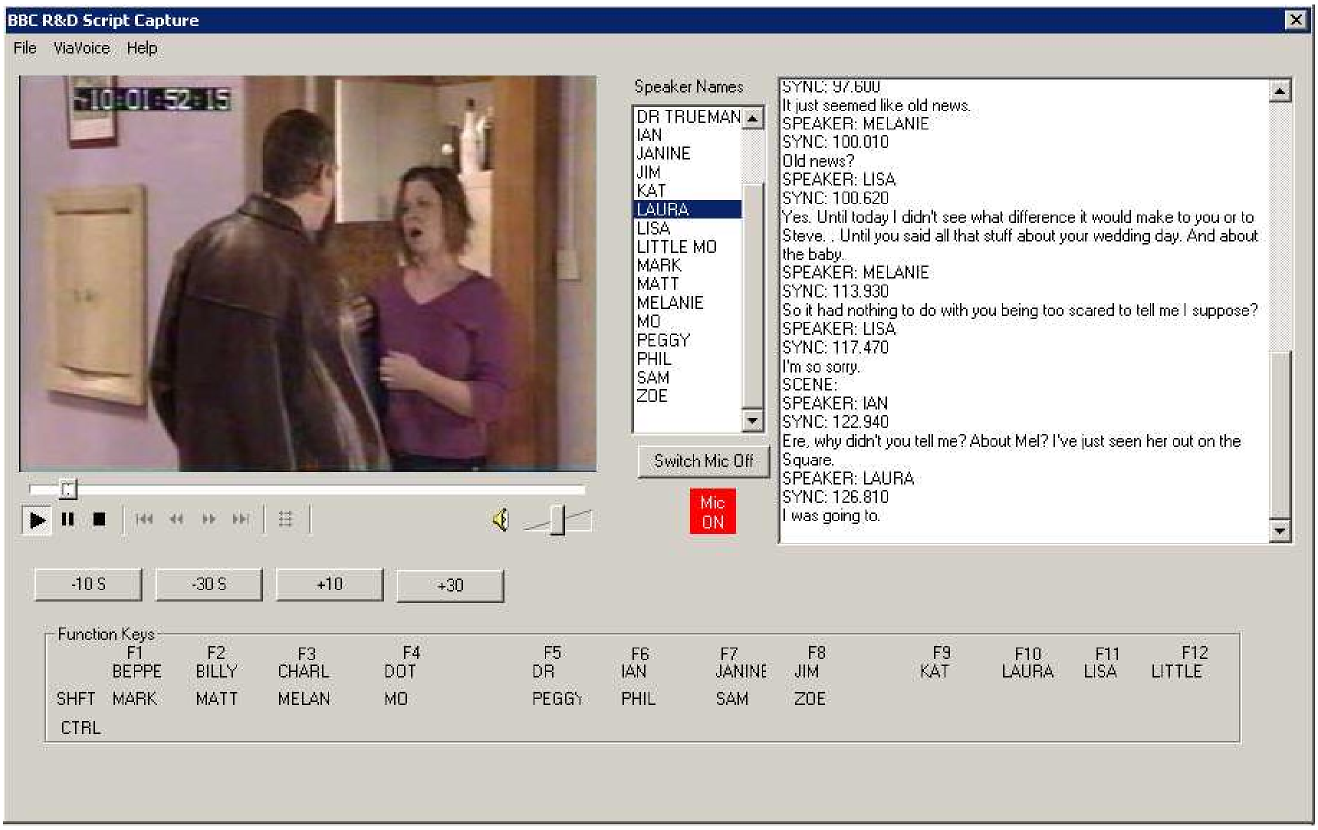
2.2 Script Capture using Speaker-Specific Speech Recognition

Script Capture currently uses ViaVoice from IBM [2], although a number of other suitable recognisers are available. ViaVoice is a speaker-specific voice recognition package in which a user profile is built up during a training period lasting approximately 2 hours, and optimised during subsequent use. Experienced ViaVoice users can achieve transcription accuracy in excess of 95%.

The Script Capture consists of an MPEG player for playback of the programme, as well as an editing window for the compilation of the script. The operator controls the playback of the programme, using shuttle controls on the MPEG player. He/she 're-speaks' the programme's dialogue whilst the programme is running. Alongside this, the operator can also use the keyboard and mouse to flag scene breaks, select speaker names and insert synchronisation points. An illustration of Script Capture in use is shown in Figure 3.

A Script Capture session generally involves the operator specifying the MPEG file containing the programme and a text file containing the names of its characters. The operator re-speaks the dialogue, selecting the appropriate speaker from the list as he/she progresses. (Controls are available on the MPEG player to shuttle the recording backwards and forward, to identify characters and their entrances and exits, and to correct errors.) The transcript of the programmes is gradually built up in the editing window. Whenever the operator indicates that the character who is speaking has changed, a sync marker is inserted into the transcript. Therefore, the transcript contains the approximate start time of each piece of dialogue. This will eventually aid the exact speech alignment of the script to the soundtrack.

The software ensures that the transcript is formatted in a formal and consistent manner. In this way, when the capture of a script is completed, the transcript is used to form an AML file for the programme without the need to use the probabilistic



```
SYNC: 97.600
It just seemed like old news.
SPEAKER: MELANIE
SYNC: 100.010
Old news?
SPEAKER: LISA
SYNC: 100.620
Yes. Until today I didn't see what difference it would make to you or to
Steve. . Until you said all that stuff about your wedding day. And about
the baby.
SPEAKER: MELANIE
SYNC: 113.930
So it had nothing to do with you being too scared to tell me I suppose?
SPEAKER: LISA
SYNC: 117.470
I'm so sorry.
SCENE:
SPEAKER: IAN
SYNC: 122.940
Ere, why didn't you tell me? About Mel? I've just seen her out on the
Square.
SPEAKER: LAURA
SYNC: 126.810
I was going to.
```

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
	BEPPE	BILLY	CHARL	DOT	DR	IAN	JANINE	JIM	KAT	LAURA	LISA	LITTLE
SHFT	MARK	MATT	MELAN	MO	PEGGY	PHIL	SAM	ZOE				
CTRL												

Figure 3: Script Capture application in use

script analysis techniques generally used by the Assisted Subtitling system. An AML file produced by Script Capture consists primarily of character list and script sections, including sync points. This file, along with the MPEG file containing the programme's recording, can then be imported into the Assisted Subtitling application. Alignment (using Aurix's speaker-independent recogniser), speaker colouring, shot change detection and subtitle generation are then carried out as usual. The result is an initial, easily correctable set of subtitles obtained in a quick and reliable manner for a programme for which no post-production script was available.

3 LIVE SUBTITLING

3.1 Background

The commitment to subtitle 100% of TV output encompasses a large amount of programming for which no pre-recording or transcript exists; i.e. live programmes or programmes edited very close to the time of transmission. Such programmes include news reports, sport, parliamentary coverage and special events (such as the Queen Mother's funeral) and are subtitled live. Live subtitling requires a very fast and accurate text input mechanism, to allow subtitles to be created without excessive delay. Specialised keyboard input and Computer-Aided Transcription (CAT) are often used, and recent years have seen the effective use of speech recognition to create the text.

As with Script Capture, live subtitling by speech requires a speaker specific recogniser, trained to the individual voice of the subtitler. The subtitler 're-speaks' the live programme's dialogue, condensing and rephrasing the text as required. The BBC has been using such a live subtitling technique for three years and found it to be extremely effective. However, the costs associated with scaling up the amount of live subtitling the BBC needs to perform to meet our commitment is prohibitive, in terms of equipment costs.

Live subtitling through re-speaking has proven effective for BBC programmes. However, to meet its needs, the BBC requires a flexible, simple and low-cost live subtitling system, in which the subtitler's spoken (or CAT) input of text is formed into subtitles with the minimum of delay. For live programmes, optimal editing and correction of the subtitles would generally lead to an excessive lag between dialogue and subtitle. We want the subtitler's spoken input to go directly to air.

An additional consideration is the multichannel nature of the BBC's television output. There are up to 14 BBC TV services on air at any moment. Conceivably, any or all of these may require live subtitling at some point. As many services and live programmes include regional opt-outs - for example a regional news update within a national news programme - some subtitlers are generally required at the regional broadcast centres to cover what might often be a very short broadcast. For flexibility, the BBC requires its subtitle production to be networked in such a way so that geography is not a significant factor. Subtitlers can provide subtitles from any regional centre, or indeed from any location with the appropriate network connection. This allows the subtitlers to work from home, when necessary.

3.2 The BBC Live Subtitling System

The BBC's new live subtitling system is called *KLive* and allows a distributed team of skilled subtitlers to provide live subtitles for any of our television services from anywhere that they can be connected to the system's IP (Internet Protocol) network. PC-based subtitling workstation software can be used at regional broadcast centres or in the subtitlers' homes. The software allows subtitlers to hand over responsibility for services between each other. For example, Subtitler A - at BBC Glasgow - can relinquish responsibility for BBC1 to Subtitler B, just joining the system from home. Subtitler A then takes a break before requesting handover of responsibility for BBC Parliament from Subtitler C, working at BBC Bristol.

Figure 4 shows a simple block diagram for the KLive system.

3.3 The Real-Time Subtitling Workstation

Figure 5 shows the KLive workstation application in use. When subtitling live, the subtitler watches the programme (usually the live on-air broadcast) and uses IBM ViaVoice (or CAT) to enter the subtitle text corresponding to the dialogue they hear. The workstation also allows a very small amount of subtitle formatting, including positioning and colour selection. However, essentially the text from ViaVoice or the CAT keyboard is divided directly into subtitles without further processing and delay.

The workstation application runs on a PC connected to the IP network. The output of the workstation is a sequence of subtitles which are sent over the IP network to the subtitle transmission gateway.

3.4 KLive Name Server and Gateway

Subtitles are inserted into the on-air signal via a subtitle inserter. In the live subtitling system, each inserter is coupled with live subtitling gateway. The gateway is a piece of software which receives the IP data containing the live subtitle stream from the workstation and extracts the subtitles. These are passed to the inserter. Each television service with live subtitles requires a gateway-inserter pair. It is also prudent to include some redundancy. Normally there will be main and reserve broadcast chains, each with its own subtitle insertion point. The system is

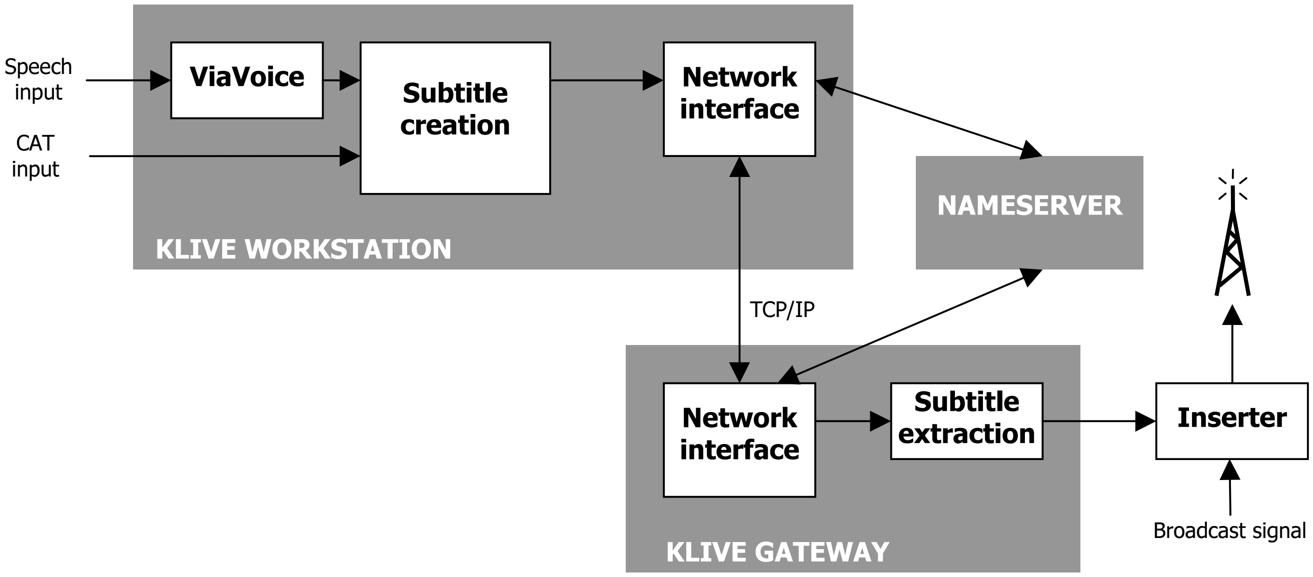
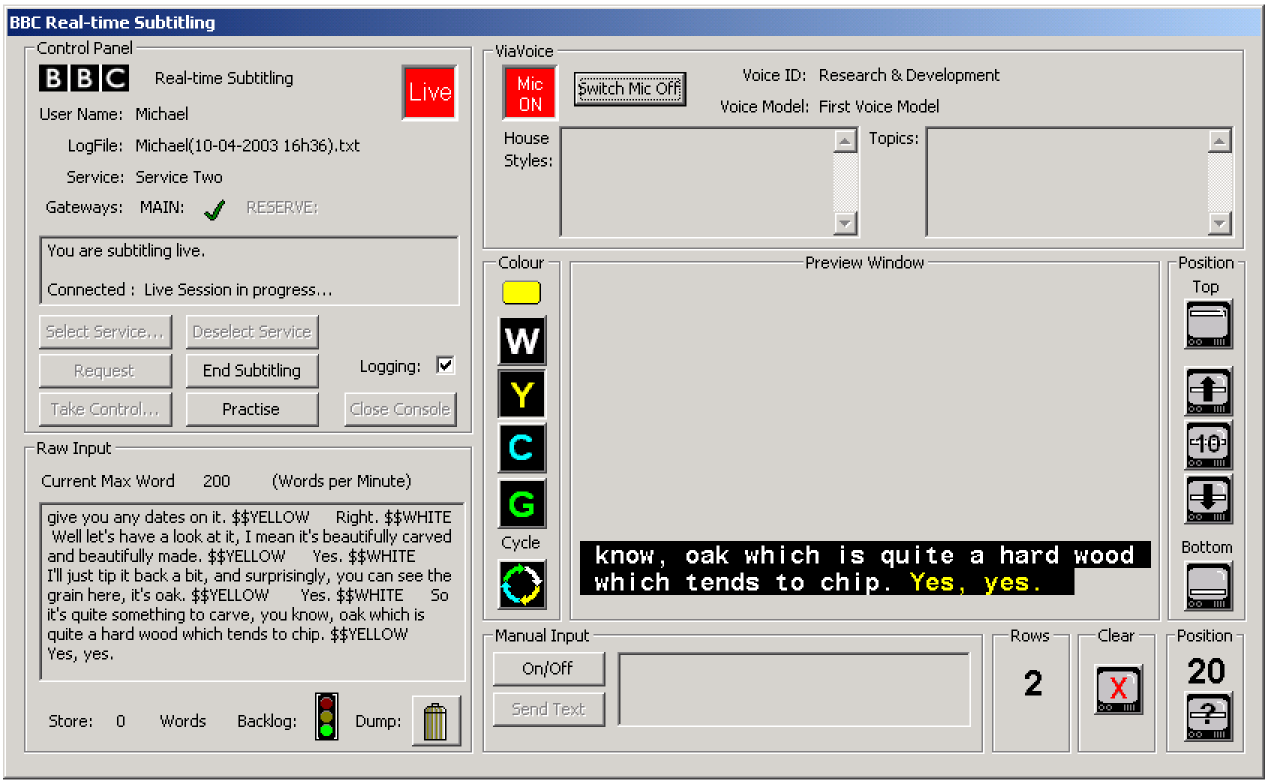


Figure 4: Simplified block diagram of KLive real-time subtitling system



BBC Real-time Subtitling

User Name: Michael

LogFile: Michael(10-04-2003 16h36).txt

Service: Service Two

Gateways: MAIN: ✓ RESERVE:

You are subtitling live.

Connected : Live Session in progress...

Logging: ☑

Current Max Word 200 (Words per Minute)

```
give you any dates on it. $$YELLOW     Right. $$WHITE
 Well let's have a look at it, I mean it's beautifully carved
and beautifully made. $$YELLOW     Yes. $$WHITE
I'll just tip it back a bit, and surprisingly, you can see the
grain here, it's oak. $$YELLOW     Yes. $$WHITE     So
it's quite something to carve, you know, oak which is
quite a hard wood which tends to chip. $$YELLOW
Yes, yes.
```

Store: 0 Words

Voice ID: Research & Development

Voice Model: First Voice Model

know, oak which is quite a hard wood
which tends to chip. Yes, yes.

Rows 2

Position 20

Figure 5: KLive live subtitling workstation

able to switch rapidly to the reserve in the event of a failure in the main chain. Similarly, live subtitling has a main and a reserve gateway for each service. There are additional gateways on the network, which can be brought into service with a particular inserter in case of a gateway failure.

Gateways can be added, removed and reassigned between television services without having to inform each live subtitling workstation of the IP addresses of their host PCs. The IP addresses of each gateway are registered in an IP nameserver. A workstation simply looks up main and reserve gateways for a selected television service by name using the nameserver. The IP addresses retrieved from the nameserver are used to route subtitles from the workstation to the appropriate gateway, and ensure that the subtitler using the workstation is providing live, on-air subtitles for the correct television service.

3.5 Live Subtitling in Use

Up to two workstations can be connected to a KLive gateway simultaneously. This allows easy handovers between subtitlers working on a particular television service. The flexibility of the IP network allows those subtitles to be geographically separated from each other, and from the gateway.

As an example, consider two subtitlers; John, working from home in Bristol, and Anne, at her workstation at BBC Glasgow. John is due to begin providing live subtitles for a programme to be broadcast on the BBC1 service. The gateway and inserter for BBC1 are in London. John starts his KLive workstation and connects to the BBC1 gateway. When the programme begins he requests control of the gateway and begins subtitling live. His workstation at home is delivering live subtitles to air for BBC1.

Some time later, Anne is due to take over from John and subtitle the second half of the programme. She launches her KLive workstation application and connects to the remaining free slot on the BBC1 live subtitling gateway. The gateway tells Anne's workstation that John is connected and subtitling live. Anne requests control of the gateway, which automatically notifies John than Anne is ready to begin subtitling. (In an emergency, Anne could force a handover from John's workstation.) At a suitable handover point, John grants control of the gateway to Anne. Anne is now providing live subtitles for BBC1 from Glasgow. John is free to disconnect his workstation from the gateway and close down. When the programme finishes and live subtitling is no longer required, Anne relinquishes control and deselects the BBC1 gateway.

3.6 Impact

KLive live subtitling is being launched throughout the BBC in April 2003. At launch, the system will comprise around 30 gateways and 50-100 workstations (like the one shown in Figure 6). These will be distributed throughout the UK at the BBC's broadcast centres and at subtitlers' homes.

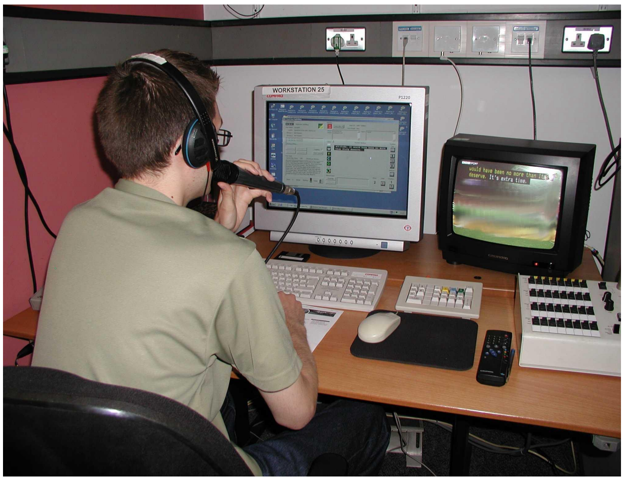


Figure 6: KLive in use by a subtitler

The system radically increases the capacity and flexibility of live subtitling of BBC television services. In the first few months alone, it is envisaged that the system will be used to subtitle events including the Wimbledon Tennis Championship and the BBC Parliament service, as well as a very large amount of regional news and other live programming.

4 CONCLUSION

This paper has described three systems developed by BBC Research and Development which use speech recognition to improve the effectiveness of skilled subtitlers. Applying speech recognition to a programme's soundtrack in order to yield a direct transcription is not practical where a wide variety of programme types need to be handled. However, our systems do make effective use of speech recognition as a subtitling tool: Assisted Subtitling uses speaker-independent recognition to perform forced alignment between a post-production transcript and a soundtrack, yielding subtitle timing information for each word. Script Capture supplements this with a re-speak system - using recognition trained to the re-speaker's voice - to produce transcripts for programmes for which none exists. Finally, the KLive live subtitling system also uses ViaVoice, allowing subtitlers to re-speak and condense the dialogue of a live broadcast, for real-time direct-to-air subtitles.

Applied speech recognition is a key technology in allowing each of these systems to increase the effectiveness of skilled subtitlers. However, as we have seen throughout this paper, speech recognition is only a relatively small part of each system. Other innovative techniques are also used in the subtitling tools. In particular, Assisted Subtitling includes a powerful script recognition and analysis engine, as well as generation of optimal subtitles conforming to a flexible set of preference; and KLive makes use of IP networking to provide a flexible and expandable live subtitling system. With Script Capture available to provide transcripts of pre-recorded programmes without post-production scripts, all types of BBC TV programmes are now able to benefit from speech recognition-based subtitling tools.

REFERENCES

[1] 'Aurix st - the script and audio alignment tool from 20/20 speech', *http://www.nxtsound.com/20-20/images/Aurix_st.pdf*

[2] 'IBMViaVoice', *http://www.software.ibm.com/speech/*