Reducing Word Error Rates of Found Speech -
XPERT Tool for Transcription Analysis

S.E. Johnson

CUED/F-INFENG/TR 330

July 1998

Speech, Vision and Robotics Group
Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

E-mail: sej28@eng.cam.ac.uk

CONTENTS

1 INTRODUCTION

Automatic Speech Recognition (ASR) research is moving increasingly away from clean speech dictation systems, such as with single-speaker voice-dictation, to so-called "found" speech. This is when natural speech has been recorded, for example from a television broadcast, and an automatically-generated transcription is required. There are less stringent time constraints on such systems, and multi-pass strategies can be used, but the problem of recognition itself become much more difficult.

An illustration of a found-speech task is maintaining an archive of audio, for example, transmitted broadcast news. If an accurate transcription can be made of all the audio, then the required space to store the information content is reduced, information retrieval methods can produce efficient audio indexing and the archive becomes an audio library, where people can scan documents and find information they need without having to listen to the entire audio. The priority for a found-speech automatic speech recogniser in this case is therefore to produce as low a word error rate as possible.

Found speech raises many new problems which have not previously been tackled in single-speaker clean-speech dictation problems. Several different speakers will occur during the sound-track [1] and there is no artificial indication of when a speaker change occurs. Also, since no restriction is made on who is speaking, it is possible to have non-native speakers whose voice characteristics are very different from the recogniser model. Similarly, speaking styles may vary. Speech is no longer in the controlled form people use when they know they are talking to a machine. Some of the speech can be prepared, producing grammatical sentences with a clear voice pattern, but some can be spontaneous. The latter has a greater variability in speaking rate, a greater frequency of hesitations and false-starts of words and sentences and generally less grammaticality than is found in prepared speech.

Broadcast News transcription is also complicated by the presence of different audio conditions. A reporter in the field may be speaking over a telephone line, an announcer in a studio may be reading headlines over background music, there may be background noise, varying channel properties, degraded acoustics or any combination of at any time during the sound-track.

Finally, the audio stream is continuous. Methods of automatically segmenting the speech into homogeneous segments using sentence boundaries become necessary. These segments ideally should only contain one speaker and one acoustic condition but again this is not always reliable and another source of error is introduced.

Since found speech is a much more difficult to transcribe than clean-speech, and the number of sources of error is greatly increased, there is a need for a specialised analysis tool to help identify the occurrence of systematic errors and if possible why they are occurring. Once identified, ways of tackling the causes of these errors could perhaps be devised.

In order to make the identification of errors, their characteristics and correlations between them possible a multi-media browsing tool, X-Program for Evaluating Recogniser Transcriptions (`xpert`), has been designed. This allows the analyser to listen to the audio, view the waveform and read both the correct transcription and recogniser output simultaneously. It highlights the errors which occur, and allows the user to zoom in at several levels to analyse the errors in more detail.

This report outlines the design requirements of the browser and briefly describes the tasks used here in section 2, explains how to use `xpert` in section 3, shows some of the results from using the browser in section 4 and offers some conclusions in section 5.

[1] Indeed in extreme cases more than one person may be speaking at one time.

2 DESIGNING THE BROWSER

Analysing the transcription of audio is difficult because there are many sources of information, all of which must be used to obtain the best indication of what happened during the broadcast. Firstly there is the audio itself. This is recorded from the radio/tv directly and stored digitally on disk. A person then listens to the audio and types into the computer the reference (or manual) transcription. This is assumed to be exactly what was said and may contain additional information, such as the background noise conditions or the bandwidth of the speech. The sound-track is sampled, encoded by its salient features (MF-PLP coefficients) and run through the recogniser to give an automatic (or recogniser) transcription. This may recognise a word correctly (no error), wrongly (substitution error), miss a word out (deletion error) or add an extra word (insertion error). A computer program then takes the reference transcription and the automatically generated one and aligns them to produce a record of which type of error (if any) is associated with each word. Information can therefore be derived from the audio itself, the manual transcription, the recogniser transcription and the alignment of the transcriptions. This is illustrated in figure 1.

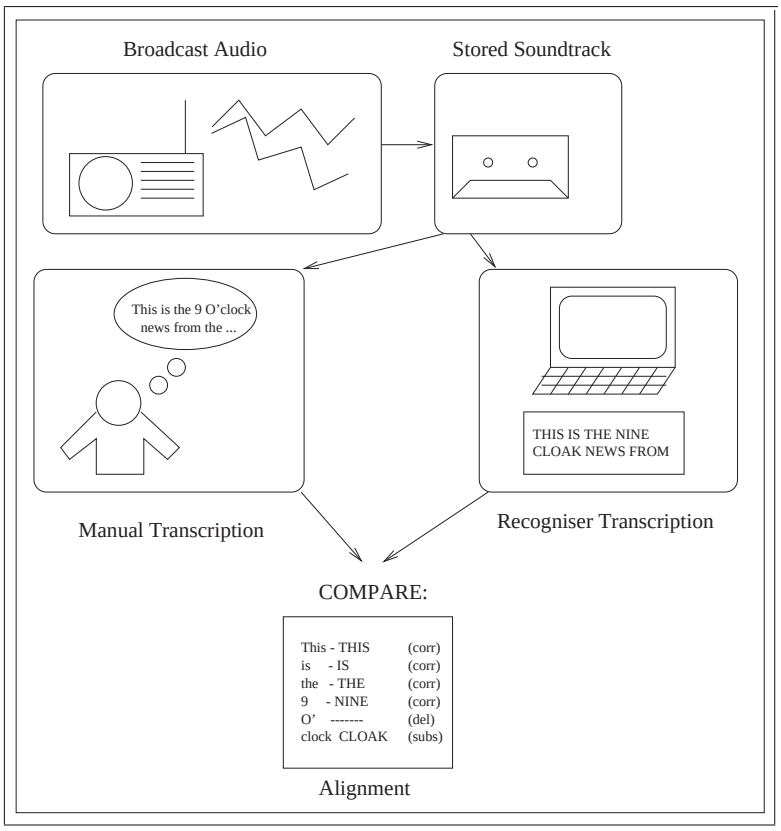

Figure 1: *Stages in Recognition Analysis*

`xpert` uses NIST-standard file formats (explained in appendix B). These are:

- SPH file - Contains the raw audio information.
- SRT file - Contains the word start time, end time and word itself of recogniser output.
- STM file - Contains the manual transcription which has been divided into segments each with labelled speaker and focus condition.
- SGML file - Contains the alignment of the transcriptions indicating the errors.

The recogniser transcription must of course be visible, to allow the user to see the final system output. Making the reference (manual) transcription readable simultaneously allows the user to compare the transcriptions and locate the errors. Manual location of all the errors by reading the transcriptions is a very slow, inaccurate and fairly boring process. To allow the user to home in on the errors, an automatic method of identifying the mismatches and displaying this information to the user is required. This can be done by utilising the information in the SGML file.

A system which did just this would be useful for locating the errors and may help indicate why some errors have occurred, for example those due to confusable acoustics. In order to help the user look for correlations between errors however, the divisions given in the STM file could be used to split up both transcriptions into segments of known speaker and condition. This allows the user to identify the speaker and type of speech for any given error. Further refinements can be gained by allowing the user to analyse the error statistics by segment, speaker or condition. The user should also be able to focus in on a given speaker/condition and see at a glance which segments were produced by them. The converse is also true, given a segment of the text, the user should be able to easily ascertain the speaker/condition that produced it.

The system described so far is helpful in comparing the two transcriptions, but some of the information contained within the audio will never appear in the transcriptions. This is information such as what is happening in the background, how the speaking rate varies, the voice quality, whether the speaker laughs/coughs etc. In order to maintain this information, the audio itself must also be available to the user.

The audio can be represented in two forms. The first is graphically presenting amplitude versus time. This allows the user to see the characteristics of the segments at a glance, picking out for example, periods of silence. Also by allowing the user to focus in on a part of the waveform, the periodicity and shape of the acoustic wave can be determined.

The most natural way of presenting audio is of course using sound. All the information about what occurred in the sound-track is presented straight to the user, and the finer points of the broadcast can be easily determined. Also, having an aural and visual input simultaneous presents the user with more analysis options.

The user should be able to "track" the show. This means playing the audio, moving a marker across the waveform, highlighting the words of the recogniser transcription , highlighting the segments of the manual transcription, indicating the occurrence of errors in the transcription and showing the speaker, segment and current focus condition in synchrony.

In summary then, the user should be presented with the sound, the waveform, the manual transcription, the recognition transcription and the errors simultaneously. They should be able to use any one source of information to focus in on any of the others, see the statistics of the errors at a glance, and be able to follow through all the sources of information easily and synchronously.

The examples of automatic transcriptions analysed using `xpert` in this report are taken from the 1997 Hub4 Broadcast News evaluation [2] and 1997 TREC6 Spoken Document Retrieval (SDR) Track [1] data using the CU-HTK recogniser [3]. They both contain seven difference audio "focus" conditions (see appendix A). The CU-HTK system obtained the best recognition performance on the 1997 Hub4 evaluation [2] with a word error rate of 16.2% and a simplified version of this system produced 28.6% WER on the SDR task.

3 GUIDE TO USING `xpert`

This section describes the browser in detail. For a quick reference user-guide see appendix C.

There are two ways to run `xpert`. For analysis, if you have generated the SRT, STM, SPH and SGML file, then you can run

```
xpert $stmfile $srtfile $sphfile $sgmlfile
```

and this will open `xpert` running on the files concerned. Note it does not matter which order you specify the filenames in as long as you give all four filenames.

The alternative is to just run `xpert`
This will load in the default story. This is especially useful if giving demonstrations using the system. Note it is also possible to switch between stories and load in new files once the program has started.

The text for the manual transcription (STM file) is divided up into homogeneous regions (same speaker, same focus condition). These regions vary in length from a single word to maybe an entire paragraph. The recogniser transcription (SRT file) has time markings for each word. Each word is read in in turn, and if the midpoint of the word lies within the current region of the manual transcription the word is added to the current line, otherwise a line break is inserted in the recogniser transcription. This produces a rough alignment of the transcriptions.

`xpert` has several drop down menus. To obtain the menu either press alt+key, where key is the letter underlined on the menu, or press the left mouse button on the menu title. The relevant options will then be displayed as a drop down menu on the screen. Left-click on the option you want and the program will take the appropriate action. The menus are:

File	Enables the user to quit the program, or change files.
Highlight	Allows relevant parts of the transcriptions to be highlighted.
Play	Plays the relevant part of the audio.
View	Alters the view of the waveform.
Track	Plays the audio, follows through the waveform, and highlights the relevant parts of the transcriptions simultaneously.
Inspect	Displays various details about the speech/transcription on the screen.

Text-focusing procedures allow the user to home in on a certain part of the transcription. This is achieved by pressing the right mouse button (given word only), middle mouse button (given word with 4 words of context) or left mouse button (entire line) on the transcription.

Additional information is permanently displayed on the screen. This informs the user of the current show, story, speaker, focus condition and line number, to facilitate the identification of the exact error locations.

Finally, three buttons are provided as short cuts, namely, :

- `Quit` - enables the user to leave the program
- `Start` - begins the option "Track Entire File".
  This allows the user to go through the entire file and pick out regions for further investigation.
- `Stop` - stops all tracking and playing of the audio which is currently being carried out.

Each of these features are described in more detail in the following sections and the interface is illustrated in figure 2.

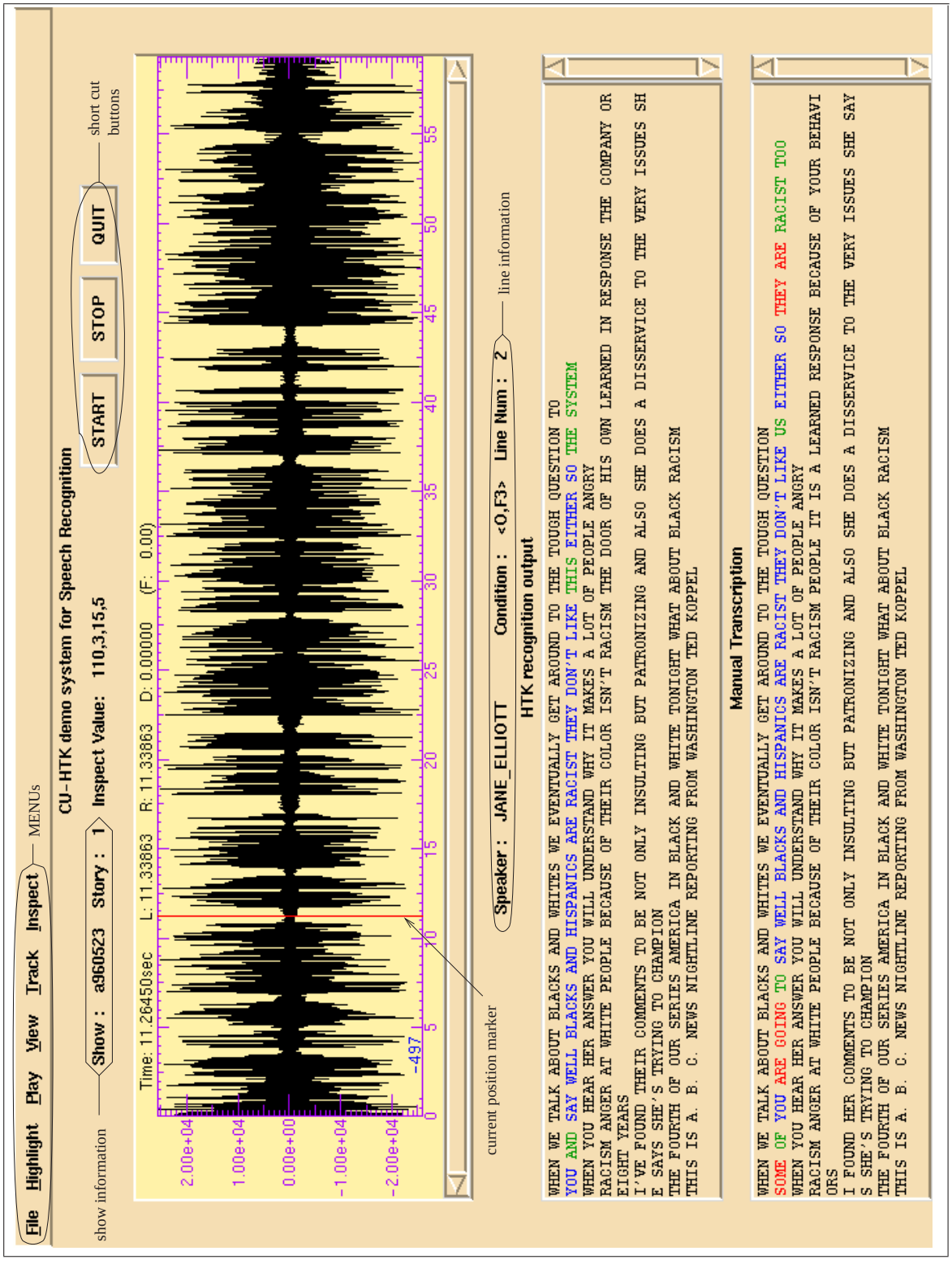

Figure 2: *Main Screen*

3.1 Changing Files

The file menu currently has the following options:

- New File
- eval97
- a960523
- g960515
- Quit

Selecting the new file option causes a new window to be produced, which enables you to type in the name of the files you wish to load. (see figure 3). To get between the fields press enter or tab. Note, all four files must be specified for the loading to take place and if you specify a file which does not exist a message appears informing you of this and allows you to correct the filename.

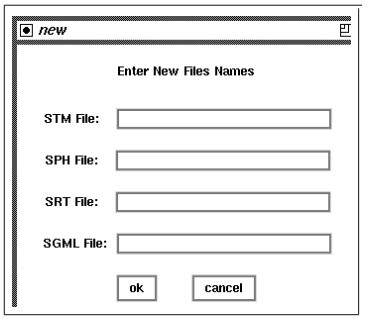

Figure 3: *Specifying a New File within* xpert

An alternative to specifying a new file is to view one of the ones already held in the system. These currently consist of 2 evaluation files and 4 SDR stories. The files take a few seconds to load, and a message is displayed asking the user to wait. The process of the changing files with the file menu is shown in figure 4.

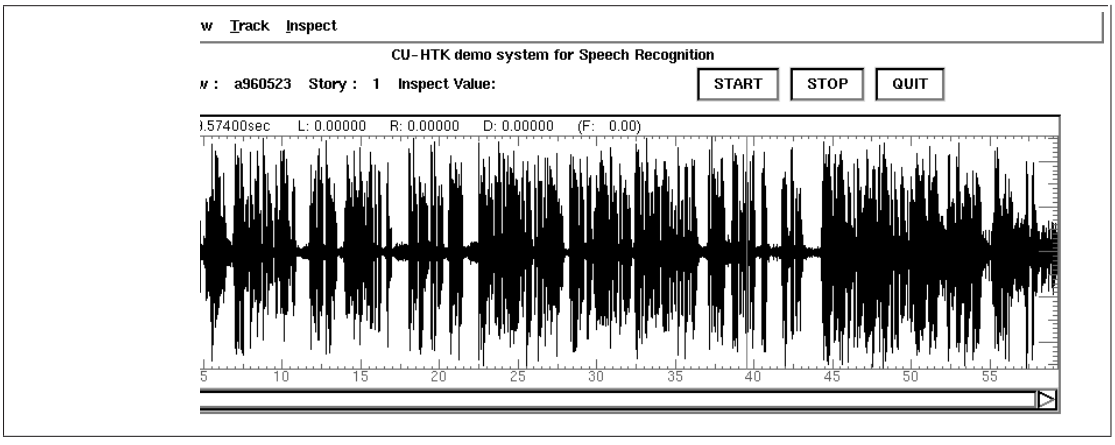

Figure 4: *Changing Story within* xpert

The eval97 option holds two five-minute excerpts from the 1997 Hub4 evaluation as an example of evaluation data. The a960523 and g960515 options each hold 2 stories from the 1997 TREC6 SDR task. Each story has different characteristics and the combination of these stories allows many of the problems of the SDR recogniser to be seen.

Quit allows the program to be exited.

3.2 Highlighting the Text

The highlight menu has the following options:

- Speaker
- Condition
- Line Number
- Word Match
- Between Marks
- Errors
- Unhighlight

Highlighting by speaker is shown in figure 5. The list of speakers is generated automatically from the manual transcription. All the words associated with the chosen speaker are then highlighted (correct words in blue, insertions/deletions in red and substitutions in green). The display also indicates the identity of the currently chosen speaker.

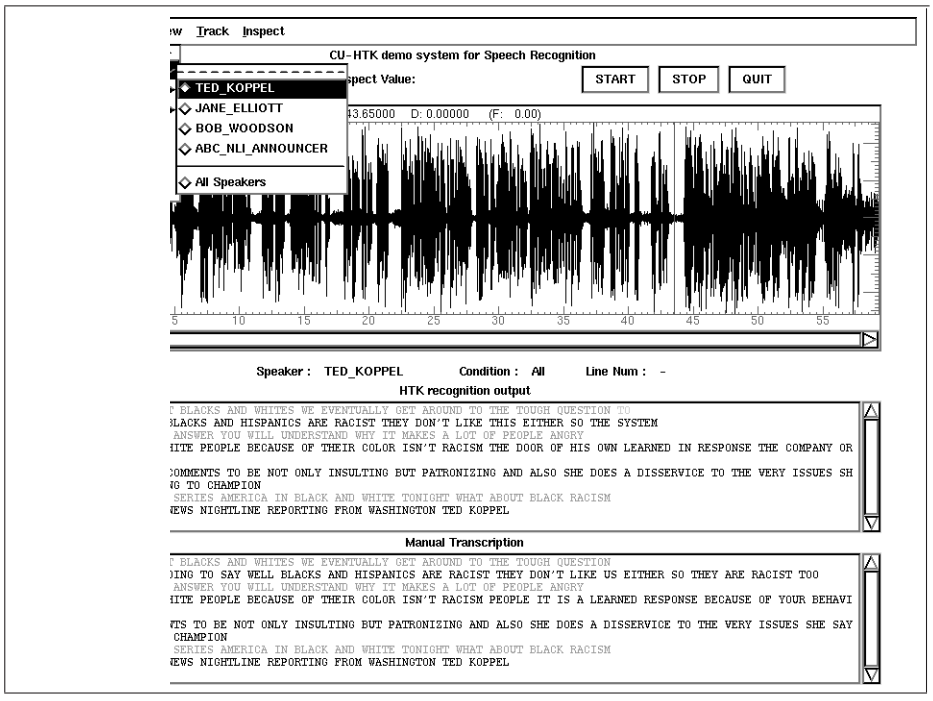

Figure 5: *Highlighting by Speaker*

Highlighting by condition is identical to that by speaker, except the focus condition is used to determine which words to highlight in the transcription.

The program also allows you to highlight any given line by its line number. This is useful if you want to focus in on a particular line.

Highlighting a word-match is included for work on the SDR files. In information retrieval the stems of words are very important, so for example you could highlight the occurrence of "insult" to find "insulting", "insulted" and "insults" as well as "insult". This enables the SDR analyser to quickly see if a given query word appears in the transcriptions.

Highlight→Between Marks allows the user to analyse the transcription corresponding to a marked piece of waveform. The waveform can be marked directly by clicking the left mouse button on the start of the desired region of the waveform and the middle mouse button at the end of it. (Alternatively push the left mouse button down at the start of the region and release it at the end). This marking of the waveform is shown by the change of colours in the waveform display.

An alternative method of marking the waveform is to use the text-focusing facilities described in section 3.7

One of the most useful features of `xpert` is to allow the user to see the errors made by the recogniser. Substitution errors are highlighted in green, whilst insertion and deletion errors are shown in red. Whilst all highlight options show the error status of the words involved, the highlight→ errors option allows the user to see at a glance all the mistakes in a given transcription. This is illustrated in figure 6.

The `Unhighlight` option simply removes all the current highlighting.

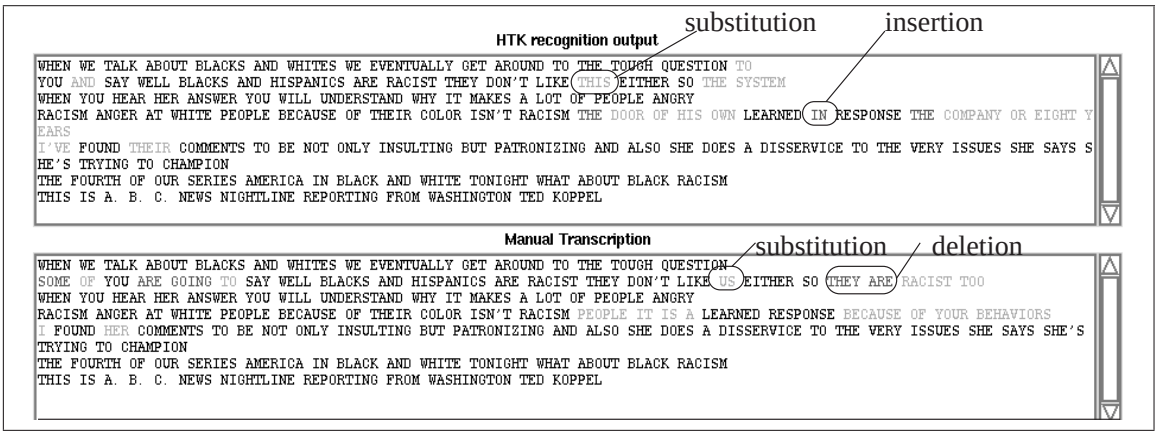

Figure 6: *Highlighting Errors*

3.3 Playing The Audio

The play menu has the following options:

- Play Entire File
- Play Between Marks
- Line Number
- Stop

The play options do not highlight the transcriptions or alter the view of the waveform, they simply play the audio. If the region being played is currently within the display of the waveform, then the appropriate region is highlighted and the cursor moves across the waveform display as the speech is played.

"Play Between Marks" can be used to play part of the waveform which has been marked in the way described in section 3.2. An example of this is illustrated in figure 7

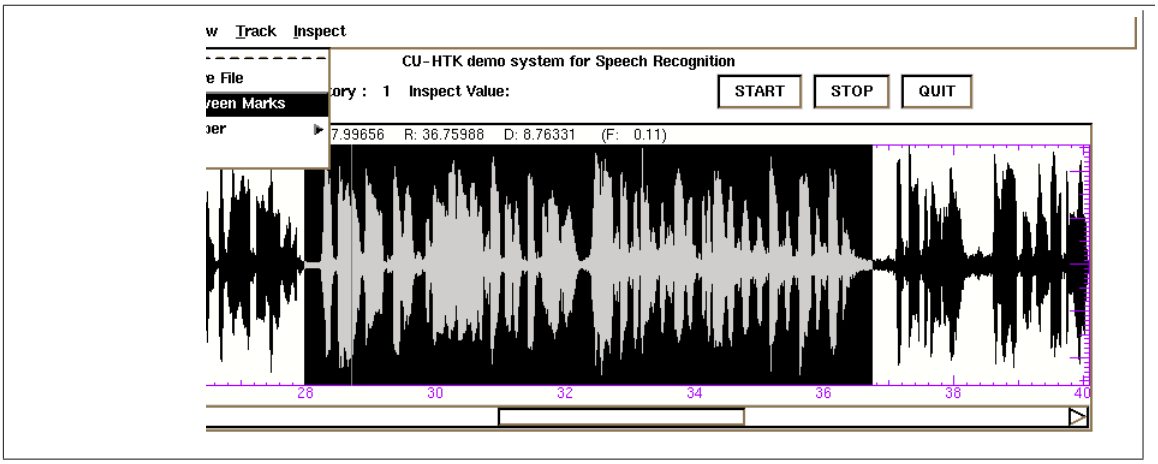

Figure 7: *Playing Between Marks*

Audio can also be played by line-number. If the user is unsure which line number they want to play, they can click on a word in the transcription and read the line-number information field, or used the highlight → line-number option.

The "stop" option can be used to stop all audio (and tracking if applicable) and is identical to the "STOP" button in the the right of the screen.

3.4 Viewing The Waveform

The view menu has the following options:

- Zoom In
- Zoom Out
- Zoom Full Out
- View Between Marks
- Page Forward
- Page Back
- Line Number
- Reset

View commands simply change the view of the waveform displayed on the screen. They do not play the audio or highlight the transcriptions.

The zoom commands allow the user to analyses part of the waveform in more detail. The effect of zooming into the waveform is illustrated in figure 8.

The marks for "View Between Marks" are set in the same way as for highlighting or playing between marks, and this option allows more detail to be seen when playing between marks, as the view can zoom in to just the relevant part of the waveform.

Page forward and back allows the user to move through the waveform using the same magnification throughout. The scrollbar can also be used to move through the waveform. Pushing the left mouse button down on the position indicator in the scrollbar and dragging to the correct place before releasing moves the waveform view continuously. An alternative to this is to click the right mouse button on the location you want to move to.

An option is also included to allow you to zoom in to any given line to see it in more detail.

Reset just sets the view to the default shown on loading. This starts at the beginning of the file, and given a relatively small file, will show the whole file on the display.

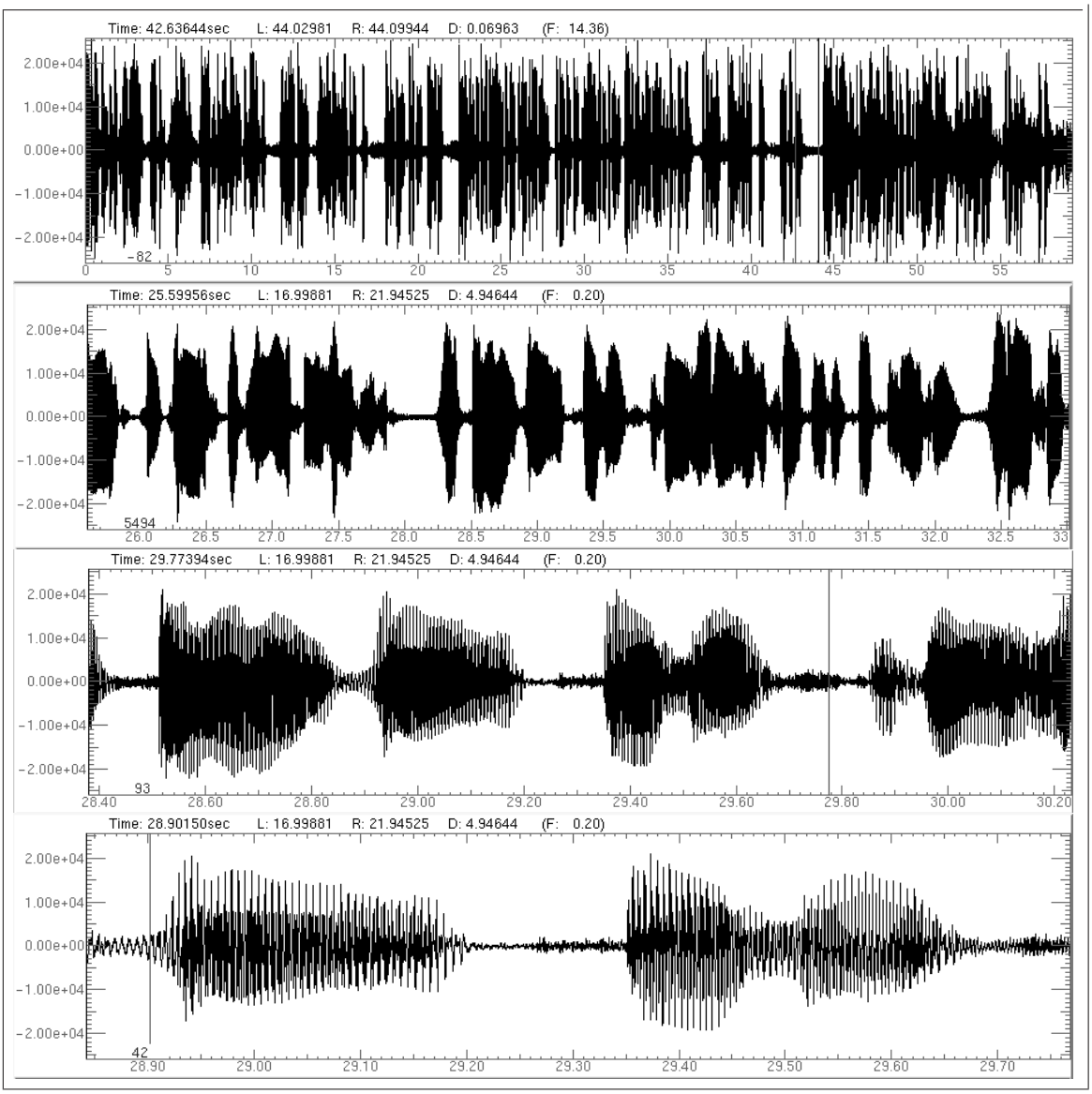

Figure 8: *Zooming into the Waveform*

3.5 Tracking The Performance

The tracking options allows synchronised viewing, highlighting by word, and audio playback to enable the user to access all the sources of information at once. The menu consists of:

- Entire File
- Line Number
- Speaker
- Condition

To track a line, the waveform display is set to show just that line and the corresponding line on the manual transcription is highlighted. The audio for that line is then played synchronously with the cursor moving over the waveform, and the words in the recognition output being highlighted. A correctly recognised word is highlighted in blue, an insertion error in red and a substitution error in green. Any correct words are unhighlighted when the audio reaches the next word, whereas errors are kept highlighted until the end of the line. This enables the user to track the errors more easily.

An example of tracking a file is given in figure 9

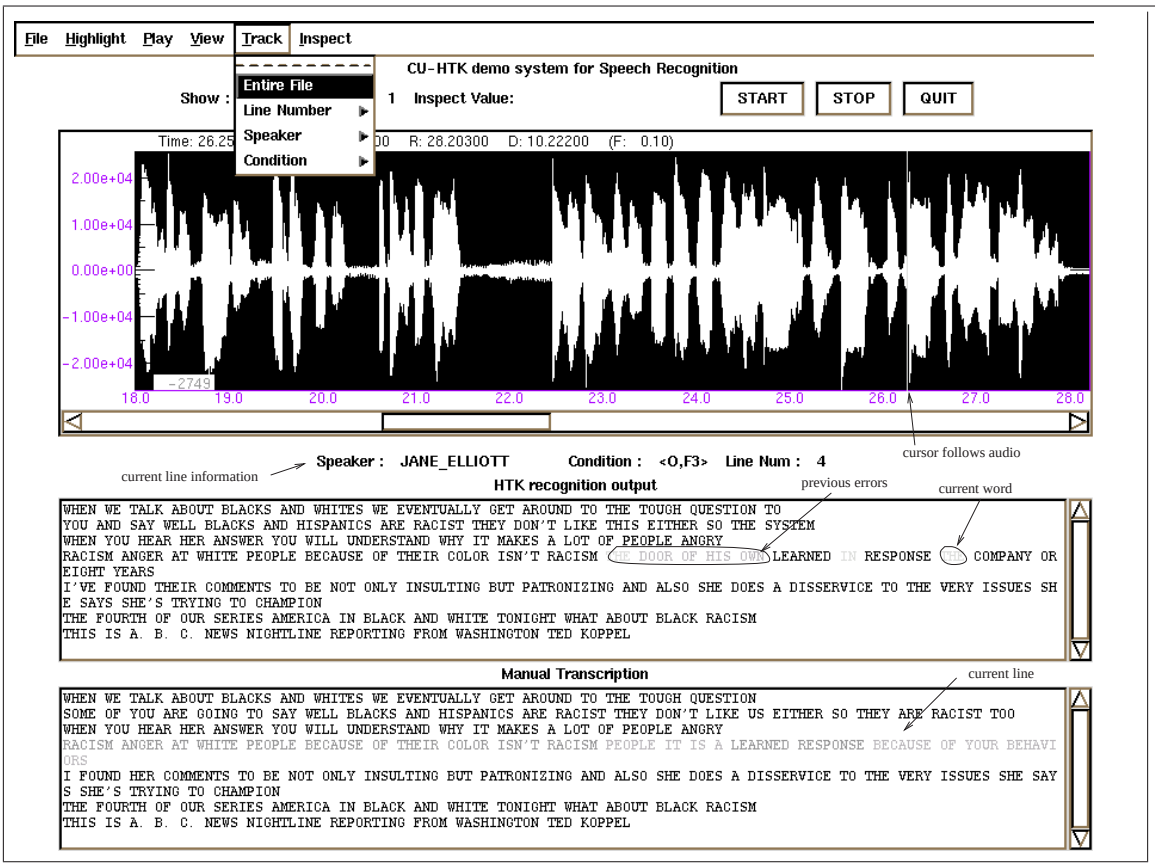

Figure 9: *Tracking the Waveform*

Tracking the entire file just goes through the lines in turn, whereas tracking a given speaker tracks the lines relevant to the chosen speaker and displays the speakers name on the display. Condition tracking is analogous to speaker tracking except it uses the focus condition as identification.

Pressing the "START" button is identical to tracking the entire file. If the tracking is no longer required at any point in time, simply press the "STOP" button.

3.6 Inspecting Parameters

The inspect menu consists of:

- Number of Lines
- HTK Number of Words
- Manual Number of Words
- Error Stats (Correct, Insertion, Substitution, Deletion)
- Beginning of File
- End of File
- Length of File
- Errors by Speaker
- Errors by Condition
- Errors in Line
- Relative Start of Line
- Relative End of Line
- Length of Line
- Speaker of Line
- Condition of Line
- View of Wave (Start:End)

Once the option has been selected, the parameter value appears on the screen after the "inspect value" label. An example showing the error stats is shown in figure 10.

The beginning and end of file values are useful when the files being analysed are part of a larger file group. For example, looking at one story in a show, or 5 minutes of the evaluation data. It allows the information about the ordering of the files to be retained in an additional form to the story number.

The error breakdown by line, condition and speaker allows any problems due to poor modelling of any given class to be seen at a glance.

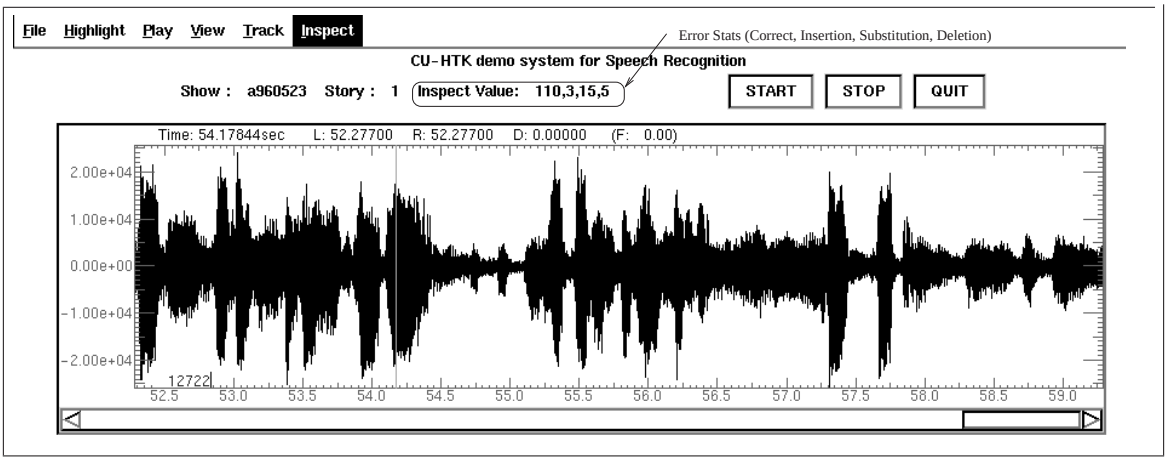

Figure 10: *Inspecting the Transcription Errors*

3.7 Text Focusing

Text focusing allows the user to home in on a particular part of the broadcast using the text. This is especially useful because errors highlighted in the text can quickly be identified in the waveform, and the corresponding transcription and audio analysed.

By right-clicking on a word in the recogniser transcription, that word is highlighted, marked on the waveform and zoomed-in on. **Play→Between Marks** can then be used to play the audio for that word. The corresponding line of the reference transcription is also highlighted. This allows intra-word difficulties, such as the introduction of background music during a word, to be analysed

Middle-clicking on a word in the recogniser output focuses on the word plus two words of context on either side, giving 5 words in total. This allows problems which occur at word boundaries or those due to incorrect word sequences to be analysed. This is illustrated in figure 11.

Left-clicking on a word in either the manual or the recogniser transcription highlights and marks the corresponding line. This is most useful when analysing errors of a given speaker and/or focus condition.

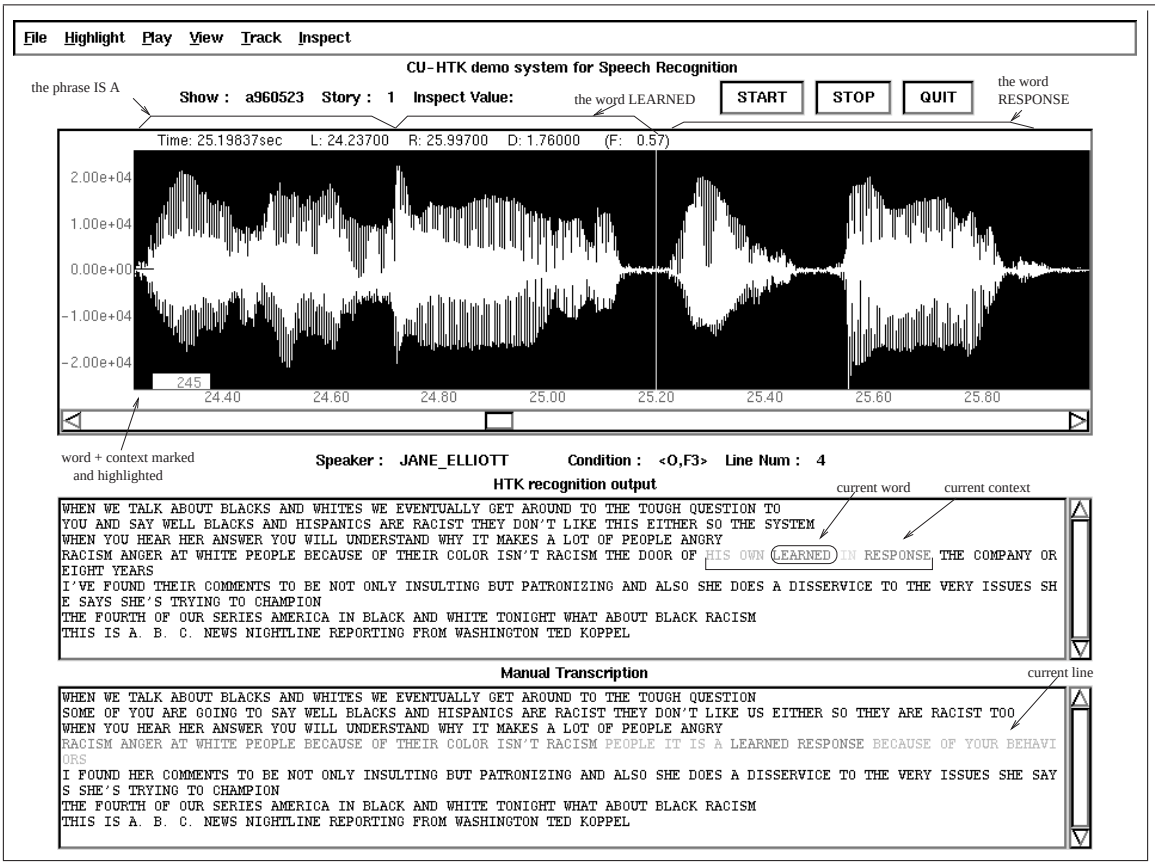

Figure 11: *Text-Focusing on a Word with Context*

4 ANALYSIS WITH xpert

xpert has been used to analyse several transcriptions and locate the causes of the errors within them. This has been particularly successful with the SDR transcriptions, due to their higher error rates. The conclusions from using xpert to analyse some SDR stories are given below.

One possible indication of the cause of errors can be to look at the length of the error sequence. Similar acoustic/syntactic words which are simply misrecognised cause only 1 mistake. Often this does not prevent the reader from extracting the meaning of the transcription. If a whole phrase is recognised incorrectly however, it is very difficult for the user to recover the information (or even topic) in the original audio. This is especially problematic for retrieval and topic identification tasks. For this reason the table of error-length versus frequency of occurrence is presented for all the evaluated stories.

Story : a960523.1
Stats : 96 correct, 4 insertion, 15 substitution, 18 deletion.
Error Rate: 28%
Average Length of Error: 3.17 words
Mode Length of Error: 1 word

	Corr	Ins	Sub	Del	WER
TED_KOPPEL	45	0	1	0	2%
JANE_ELLIOT	25	4	13	6	52%
BOB_WOODSON	26	0	1	0	3%
ABC_NLI_ANNOUNCER	0	0	0	12	100%
F3	96	4	15	18	28%

Breakdown of Errors by Speaker and Condition

Num Words	1	2	5	9
Frequency	3	1	1	1

Analysis of Error by Word Length

Problem: Music/Speech Detection
Causes: 12 deletion errors
The story ends with some music followed by speech with the music still playing. The pre-recognition segmenter has correctly recognised the portion of pure music as music and thrown it away, but has also incorrectly recognised the following portion of music and speech as pure music and also discarded it. Reinstating this segment and running the recogniser on it produces all 12 words correctly. Thus, these 12 errors can be prevented by using a more accurate music/speech classifier.

Problem: Female Modelling
Causes: 13 substitution, 6 deletion, 4 insertion errors
In the story there are 2 errors in 73 male-spoken words (excluding the music problem mentioned above), but 23 errors in 44 female-spoken words. This suggests work on improving models for female speakers may help reduce error rate.

Problem: Word Substitution
Causes: 2 substitution errors
Twice the word “HER” has been recognised as “THEIR”. This is the only mistake in the male-spoken words. The words are similar both acoustically and syntactically, but the correct one could be chosen from the situation context, namely that there is only one person being referred to. This however is not a major error.

Story: g960515.3
Stats: 223 correct, 4 insertion, 38 substitution, 8 deletion.
Error Rate: 18%
Average Length of Error: 1.91
Mode Length of Error: 1

	Corr	Ins	Sub	Del	WER
LINDEN_SOLES	140	1	6	3	6%
BOB_DOLE	50	0	19	3	30%
AL_GORE	33	3	13	2	37%
F0	123	1	6	3	7%
F2	50	0	19	3	30%*
FX	33	3	13	2	37%*
F4	17	0	0	0	0%

Num Words	1	2	3	4	5
Frequency	12	4	3	2	1

Problem: Onset of New Condition/Speaker
May Cause: 7 substitution, 2 insertion.
The first 4 words of the onset of BOB_DOLE(F2) and 5 words of AL_GORE(FX) are incorrect. This suggests there may be problems at speaker boundaries. Could possibly be helped by clustering similar speakers together (if a larger amount of data is being used).

*Problem: Telephone Models
May Cause: 19 substitution, 3 deletion.
The error rate is much higher for the section labelled telephone bandwidth (F2) (and FX) than the standard (F0). Before recognition all the segments are automatically classified as narrowband or wideband. By subsequent manual inspection of the relevant segments it was found that every segment in this story was classified as wideband, despite the manual labelling of many words as F2 and FX. It is possible that the manual labelling may be unreliable, but it would be interesting to compare the recognition of the regions concerned using the telephone models.

Problem: Spontaneous Speech
May Cause: 3 insertion, 13 substitution, 2 deletion.
The FX section is spontaneous speech with slight background noise. The speaker constantly varies his speaking rate, adds extra pauses, and includes hesitations and repeated words. Possible improvements might come from building a language model which can cope with such ungrammaticalities.

Problem: Relative Weight of the Language Model
May Cause: 17 substitution and 2 insertion errors.
The breakdown of error lengths is given above. It is possible that the language model is biasing the recogniser against the correct sequence of words following an incorrect initial word, producing the 3-5 word error sequences. Less relative weighting of the language model against the acoustic model, faster recovery of out-of-vocabulary (OOV) words or incorrect words in the language model may reduce this problem, although this is not as significant in this show as in some of the others.

Problem: Abbreviations
Causes: 1 deletion, 2 substitution, 1 insertion.
Different interpretations of the same thing, such as “HE IS” instead of “HE’S” and “REELECTION” instead of “RE ELECTION” cause a few errors. This does not cause a problem for the reader, but a standard form could be advantageous for retrieval problems. Note though this is not generally a problem if filtered transcripts are used.

Story: g960515.8
Stats: 207 correct, 26 insertion, 122 substitution, 32 deletion
Error Rate: 49%
Average Length of Error: 3.36 words
Mode Length of Error: 2 words

	Corr	Ins	Sub	Del	WER
ROBERT_VITO	166	19	56	8	36%
RON_HOFSTETTER	27	4	48	22	76%
BUFFALO_TIGER	14	3	18	2	67%
FX	188	24	115	32	50% *
F4	19	2	7	1	37%

Problem: Story Boundary
Cause : 1 Deletion Error
The first word of the story was correctly recognised, but occurred across the story boundary and was therefore deleted from the transcription. This should be changed so that the word is assigned to the story in which its midpoint occurs, and the start/end time of the word adjusted accordingly. This would eliminate this error.

Problem: Background Noise and Uncertain Speaker Rate. [2]
The performance on the "location report" in this program is consistently bad throughout. There is a lot of background noise, such as water, helicopters and alligators. The speech is spontaneous and suffers from a large variation in speaker rate, with un-natural gaps in the middle of phrases. This seems to throw the language model and results in more insertions than seen in the previous two stories.

Problem: Weight of the Language Model
The length of the errors in the HTK transcription (i.e. substitution and insertion) given in words are:

Num Words	1	2	3	4	5	8	10	11	15
Frequency	6	15	10	7	2	1	1	1	1

giving an average length of each mistake of 3.36 words. In some cases the transcription does not get anywhere near the correct answer, e.g.

HTK: IT'S JUST LIKE A LOT OF THINGS OUT POLLS *SOME* PRACTICAL MATTER WHETHER YOU THINK *THAT CAN* BREATHE *UNDERWATER AND* THERE ARE VERY WELL LET ME GO ON *AND SO* ARE *THEY* CAN'T JUDGE *A* GOOD THING.

REFERENCE: WATER MOCCASINS UH POSE *SOME* THREAT IN THAT THEY'RE ONE OF THE FEW SNAKES *THAT CAN* BITE *UNDER WATER AND* UH THEY WILL BITE REPEATEDLY UH *AND SO* UH *THEY* CAN INJECT QUITE *A* BIT OF VENOM INTO YOU.

Maybe in this case the acoustic models should have had more weighting than the language model. This might cause more errors elsewhere, but may reduce the chance of a long series of errors.

[2] All this show, including the FX portions were classified as wideband

Story: k960603.8
Stats: 257 correct, 31 insertion, 78 substitution, 5 deletion
Error Rate: 33%
Average Length of Error: 2.42 words
Mode Length of Error: 1 word

	Corr	Ins	Sub	Del	WER
DAVID_BRANCACCIO	148	6	15	1	13%
FRANK_VOGEL	109	25	63	4	52%
F0	94	3	5	0	8%
F1	54	3	10	1	21%
F2	109	25	63	4	52%

Num Words	1	2	3	4	6	7	8	10
Frequency	22	9	6	2	3	1	1	1

Problem: Pronunciations variants missing from the dictionary:
Causes: 5 substitution, 1 insertion
The following entries appear in the dictionary:
KENYA k eh n y ax
AH aa
A ax
However, in the audio they are pronounced:
KENYA k iy n y ax
AH ax

Problem: Spontaneous Speech (F1)
The error rate increases from 8% to 21% when switching from planned to spontaneous speech. This may be due in part to a breakdown in grammaticality of the speech confusing the language model and also the varying speaker rate and intermittent silences and "filler" words. Work on designing special acoustic and language models for spontaneous speech may reduce this problem.

Problem: Telephone Models (F2)
In this show all the data manually labelled as F2 was automatically correctly classed as telephone bandwidth, and was recognised with the narrowband models. The error rate went from 13% to 52% when changing to the telephone bandwidth. This suggests better methods of modelling narrowband speech could dramatically improve the system performance.

HUB4: Evaluation 97, first 5 minutes
Stats: 759 correct, 19 insertion, 76 substitution, 48 deletion
Error Rate: 16%
Average Length of Error: 1.61 words
Mode Length of Error: 1 word

	Corr	Ins	Sub	Del	WER
DAVID_BRANCACCIO	484	2	40	32	13%
CHRISTOPHER_LOWE	142	9	12	9	18%
FRANK_CONTRERAS	116	6	17	3	19%
DAVID_SHIELDS	17	2	7	4	46%
F0	101	0	8	1	8%
F3	32	0	1	2	8%
F4	321	1	25	28	14%
FX	305	18	42	17	21%

Num Words	1	2	3	4	5
Frequency	37	11	4	4	1

Problem: Tense of word
In some cases the correct stem of the word has been recognised, but the tense is wrong, leading to a substitution error. This occurs for example with:

KEEP	→	KEEPS
WANT	→	WANTS
SINCE IT CLOSED	→	TO CLOSE
BLAZE IS	→	BLAZES
ACCIDENTS	→	ACCIDENT
MOTORS'	→	MOTORS
THE MAN	→	MEN
INDUSTRIAL AVERAGE	→	INDUSTRIALS

Since the majority of these are very similar acoustically, maybe a grammatical constraint is needed in the language modelling to choose the correct form of the word.

Problem: Standardisation of Transcriptions
Several errors are due to different transcriptions of the same thing. For example "MINIVAN" not "MINI VAN" or "HE'S" not "HE IS". This problem would be substantially reduced if standard filtering was carried out on the files used by `xpert`.

Problem: Background Noise (FX)
The line spoken by David Shields has some background clicking on the audio. Coupled with a variable speaker rate, this accounts for the higher error rate of this line.

5 CONCLUSIONS

`xpert` has been shown to allow the user to quickly identify the location of all the errors in a transcription. These errors can easily be examined in further detail by listening to the audio and viewing the waveform. The user can focus in on the error from the waveform, the manual transcription, the recogniser transcription, and the properties of the segment (such as speaker/condition) with all the sources of information being linked together.

The tracking feature allows the user to examine all the information both aurally and visually simultaneously and helps understand the context of the errors.

The presenting of error statistics by line, speaker, condition and story allows the user to easily spot trends due to poor modelling and therefore suggest areas for further work.

`xpert` successfully allows the user to locate the errors in a transcription, and although it has been shown in section 4 to help identify systematic errors and their causes, future development of this system should be aimed at incorporating more automatic analysis of the errors. This could include automatic evaluation of the length of error sequences and the proportion of errors occurring at segment changes. The latter is especially important if speaker adaptation is being used.

Further extensions include allowing the user to examine the segments used in the recognition to determine whether a wideband, narrowband, male, female or gender-independent model set was used; and including the pronunciation dictionary to identify missed pronunciations or OOV words.

6 ACKNOWLEDGEMENTS

The `twsh` libraries which `xpert` requires are provided by Entropic Inc. (© 1998). Thanks also go to Paco Gimenez-Galanes and Patrick Gosling for help in writing the code and Steve Young and Phil Woodland for patiently trying it out whilst it was being developed.

This work was supported by EPSRC grant GR/L49611

A FOCUS CONDITIONS FOR BROADCAST NEWS

F0	baseline broadcast speech (clean, planned)
F1	spontaneous broadcast speech (clean)
F2	low fidelity speech (wideband/narrowband)
F3	speech in the presence of background music
F4	speech under degraded acoustical conditions
F5	non-native speakers (clean, planned)
FX	all other speech (e.g. spontaneous non-native)

B FILES REQUIRED FOR PROCESSING

The Spoken Document Retrieval (SDR) System uses several different types of files. Each of these are described briefly in the following sections.

The recognition is done on a show basis, but information about a given story can be extracted using the tool `extract_story.pl`

Once the relevant files have been produced for a given story, `xpert` can be run on this story, independent of its length, number of speakers, conditions etc.

B.1 SPH Files

The Sphere (SPH) files are provided by NIST on a show basis. They can be viewed in `xwaves` or chopped into smaller (segment-based) chunks using `w_edit`. This can be done automatically within the script `extract_story.pl` for a given story with the output being placed in the file `$homedir/x_$story_$num.sph` .

B.2 NDX Files

The index (NDX) file is provided by NIST for the spoken document retrieval task. It contains the start and end times of each "section" along with a classification into the type

```
[ Story | Filler | Local_News | Commercial | Sports_Report ]
```

and an ID number of the form "`$show.$number`".
Each "Story" also has a "Topic" field which is nominally undetermined.

An example of an NDX file for the a960523 show on the TREC6 eval task is

```
<Episode Filename=a960523.sph Program="ABC_Nightline" Scribe="Jonathan_Cole"
 Date="960523:2330" Version=2 Version_Date=970130>
<Section S_time=399.063 E_time=458.350 Type=Filler ID="a960523.1" >
<Section S_time=458.350 E_time=565.692 Type=Story Topic="undeterm" ID="a960
523.2" >
<Section S_time=565.692 E_time=968.063 Type=Story Topic="undeterm" ID="a960
523.3" >
<Section S_time=968.063 E_time=975.381 Type=Filler ID="a960523.4" >
<Section S_time=1131.854 E_time=1299.901 Type=Story Topic="undeterm" ID="a9
60523.6" >
<Section S_time=1299.901 E_time=1497.129 Type=Story Topic="undeterm" ID="a9
60523.7" >
<Section S_time=1497.129 E_time=1525.549 Type=Filler ID="a960523.8" >
<Section S_time=1691.365 E_time=1903.000 Type=Filler ID="a960523.10" >
</Episode>
```

B.3 UEM Files

The Unpartitioned Evaluation Map (UEM) File is generated from the NDX file using the `ndx2uem.pl` tool. This contains the showname, channel, start time and end time for each section in the corresponding ndx file.

An example of a UEM file for the a960523 show is given below.

```
;;
;; File:      a960523.uem
;; Date:      Mon Jan 12 12:22:07 1998
;; User:      sej28
;;
;; Field 1: File ID
;; Field 2: Channel      ;; always 1
;; Field 3: Speaker ID   ;; left out
;; Field 4: Start Time
;; Field 5: End Time
;;
a960523 1 399.063 458.350
a960523 1 458.350 565.692
a960523 1 565.692 968.063
a960523 1 968.063 975.381
a960523 1 1131.854 1299.901
a960523 1 1691.365 1903.000
;; end of file
```

B.4 MLF Files

The output of the HTK automatic speech recognition system is a Master Label File (MLF). This contains the name of each of the parts which the segmenter generated and the recogniser subsequently transcribed. This is followed by the sentence marker `<s>` and the recognised words for that part. Each word also provides the start and end time relative to the start of the part given in 100nS units (frames) and an associated confidence score. The end of a part is signalled by the end of sentence marker `</s>` and a full-stop on its own.

An example of part of an MLF file for the show a960523 is given below.

```
#!MLF!#
"*/a960523_FWM0000_0039906_0041656_FW.rec"
0 200000 <s> -117.364258
200000 1600000 WHEN -1078.953369
1600000 2800000 WE -881.799561
2800000 5600000 TALK -2138.101074
5600000 9100000 ABOUT -2374.892578
9100000 13600000 BLACKS -3179.288574
<and so on until>
158200000 166200000 PEOPLE -4960.742188
166200000 169900000 ANGRY -2599.453125
169900000 175100000 </s> -2795.421875
.
"*/a960523_FWM0001_0041656_0044339_FW.rec"
0 5100000 <s> -2825.743652
5100000 12100000 RACISM -5316.588379
12100000 19200000 ANGER -4958.212891
<etc>
```

B.5 SRT Files

The Spoken Recognition Transcription (SRT) file is generated by combining the MLF files for telephone and wideband speech from the HTK output with the story information contained within the NDX file. The MLFs are combined using the `mlfs2mlf` tool, and the conversion to SRT format is done using `mlfndx2srt`. [3]

Any word which falls across a story boundary is assigned to the story in which its midpoint occurs, and the start/end time is adjusted accordingly.

The SRT file for the MLF given above is:

```
<Episode Filename=a960523.sph Program="ABC_Nightline" Scribe="Jonathan_Cole"
 Date="960523:2330" Version=2 Version_Date=970130>
<Section S_time=399.063 E_time=458.350 Type=Filler ID="a960523.1" >
<Word S_time=399.080 E_time=399.220> WHEN </Word>
<Word S_time=399.220 E_time=399.340> WE </Word>
<Word S_time=399.340 E_time=399.620> TALK </Word>
<Word S_time=399.620 E_time=399.970> ABOUT </Word>
...
<Word S_time=456.290 E_time=456.630> TED </Word>
<Word S_time=456.630 E_time=456.990> KOPPEL </Word>
</Section>
<Section S_time=458.350 E_time=565.692 Type=Story Topic="undeterm" ID="a9605
23.2" >
<Word S_time=458.370 E_time=458.590> AND </Word>
<Word S_time=458.590 E_time=458.860> THEN </Word>
...
<Word S_time=1902.250 E_time=1902.320> THE </Word>
<Word S_time=1902.320 E_time=1902.890> PROGRAM </Word>
</Section>
</Episode>
```

Again the tool `get_story.pl` extracts the information of a given story from this SRT file. This just involves picking out the relevant words and episode/section heading and storing them in the file: `$homedir/version$version/x_$show_$story.srt`

B.6 STM Files

The Segment Time Marked (STM) file contains the manual transcription for each show, as supplied by NIST. It is marked on a segment basis (not a word basis like the SRT file) and gives the associated show, channel, speaker, start time, end time, condition and transcription of each "line" of speech. (A "line" here corresponds to a piece of homogeneous audio, namely that in which the speaker and audio condition stay constant).

The tool `extract_story.pl` again enables the user to extract the relevant lines for a given story and place the result in the file `$homedir/x_$show_$story.stm`.

Care must be taken at the boundaries, however, as problems arise when the stm boundaries do not coincide with the story boundaries.

The beginning of the STM file for the show a960523 is given below.

```
;; STM for File /home/solvea/hub4/ldc/hub4_trans/aABC_NLI/a960523.txt,
Show ABC_Nightline, Episode 960523:2330, Version 2 - 970130
```

[3] N.B. an SRT file can also be produced from a CTM file using `ctmndx2srt.pl`

```
;;
;; more comments about the format/condition categories
;;
a960523 1 Ted_Koppel 399.063 404.377 <O,F3> WHEN WE TALK ABOUT BLACKS
 AND WHITES WE EVENTUALLY GET AROUND TO THE TOUGH QUESTION
a960523 1 Jane_Elliott 404.377 410.766 <O,F3> SOME OF YOU ARE GOING T
O SAY WELL BLACKS AND HISPANICS ARE RACIST THEY DON'T LIKE US EITHER
SO THEY ARE RACIST TOO
a960523 1 Ted_Koppel 410.766 417.044 <O,F3> WHEN YOU HEAR HER ANSWER
YOU WILL UNDERSTAND WHY IT MAKES A LOT OF PEOPLE ANGRY
```

B.7 CTM Files

The Conversation Time Marked (CTM) file is needed in order to generate the subsequent SGML file. It can be created using the `srt2ctm.pl` tool from an SRT file, or on a show basis directly from the MLF using `mlf2ctm.pl`.

It contains the story, channel, start-time, duration and word information. The CTM file for the beginning of the show a960523 is given below:

```
a960523  1  399.08 0.14  WHEN
a960523  1  399.22 0.12  WE
a960523  1  399.34 0.28  TALK
a960523  1  399.62 0.35  ABOUT
a960523  1  399.97 0.45  BLACKS
a960523  1  400.42 0.34  AND
a960523  1  400.76 0.47  WHITES
a960523  1  401.23 0.19  WE
a960523  1  401.42 0.49  EVENTUALLY
a960523  1  401.91 0.14  GET
a960523  1  402.05 0.85  AROUND
a960523  1  402.90 0.12  TO
```

B.8 SGML Files

The SGML file contains the information about the differences between the manually transcribed STM file and the automatically transcribed CTM file. Generally filtering is applied before the comparison is made, but this is *not* the case with this analysis system.

The SGML file is generated by running:

```
$sclite -F -l 80 -r $stmfile stm -h $ctmfile ctm -o sgml
```

This results in a file `$homedir/version$version/x_$show_$num.ctm.sgml` of the form:

```
<SYSTEM title="/home/texas1/sej28/demo/version1/x_a960523_1.ctm"
ref_fname="/home/texas1/sej28/demo/x_a960523_1.stm" hyp_fname="/
home/texas1/sej28/demo/version1/x_a960523_1.ctm" creation_date="
Fri Mar 13 17:42:22 1998" format="2.2" frag_corr="TRUE" opt_del=
"FALSE">
<LABEL id="O" title="Overall" desc="Overall">
</LABEL>
<LABEL id="F0" title="Baseline//Broadcast//Speech" desc="">
</LABEL>
...
<LABEL id="FX" title="All other speech" desc="">
```

```
</LABEL>
<CATEGORY id="0" title="" desc="">
</CATEGORY>
<CATEGORY id="1" title="1996 Hub4 Focus Conditions" desc="">
</CATEGORY>
<SPEAKER id="ted_koppel">
<PATH id="(ted_koppel-000)" word_cnt="15" labels="<o,f3>" file=
"a960523" channel="1" sequence="0" R_T1="399.063" R_T2="404.377
" word_aux="h_t1+t2">
C,"when","when",399.080+399.220:C,"we","we",399.220+399.340:C,"
talk","talk",399.340+399.620:C,"about","about",399.620+399.970:
C,"blacks","blacks",399.970+400.420:C,"and","and",400.420+400.7
60:C,"whites","whites",400.760+401.230:C,"we","we",401.230+401.
420:C,"eventually","eventually",401.420+401.910:C,"get","get",4
01.910+402.050:C,"around","around",402.050+402.900:C,"to","to",
402.900+403.020:C,"the","the",403.020+403.150:C,"tough","tough"
,403.150+403.410:C,"question","question",403.410+404.160
</PATH>
<PATH id="(ted_koppel-001)" word_cnt="16" labels="<o,f3>" file=
"a960523" channel="1" sequence="2" R_T1="410.766" R_T2="417.044
" word_aux="h_t1+t2">
C,"when","when",410.780+410.970:C,"you","you",410.970+411.130:C
,"hear","hear",411.130+411.800:C,"her","her",411.800+412.090:C,
"answer","answer",412.090+412.950:C,"you","you",412.950+413.120
:C,"will","will",413.120+413.300:C,"understand","understand",41
3.300+413.880:C,"why","why",413.880+414.080:C,"it","it",414.080
+414.200:C,"makes","makes",414.200+414.420:C,"a","a",414.420+41
4.490:C,"lot","lot",414.490+414.760:C,"of","of",414.760+414.880
:C,"people","people",414.880+415.680:C,"angry","angry",415.680+
416.050
</PATH>
...
</SPEAKER>
<SPEAKER id="jane_elliott">
<PATH id="(jane_elliott-000)" word_cnt="23" labels="<o,f3>" file=
"a960523" channel="1" sequence="1" R_T1="404.377" R_T2="410.766"
word_aux="h_t1+t2">
D,"some",,:S,"of","to",404.160+404.720:C,"you","you",404.720+404.9
20:D,"are",,:D,"going",,:S,"to","and",404.920+405.170:C,"say","say
",405.170+405.950:C,"well","well",405.950+406.350:C,"blacks","blac
ks",406.350+406.720:C,"and","and",406.720+406.790:C,"hispanics","h
ispanics",406.790+407.330:C,"are","are",407.330+407.400:C,"racist"
,"racist",407.400+407.790:C,"they","they",407.790+407.890:C,"don't
","don't",407.890+408.080:C,"like","like",408.080+408.300:S,"us","
this",408.300+408.570:C,"either","either",408.570+408.880:C,"so","
so",408.880+409.270:D,"they",,:D,"are",,:S,"racist","the",409.270+
409.430:S,"too","system",409.430+410.780
</PATH>
etc
```

Note, the comparison is given in speaker order, *NOT* chronological order. Each pair is given a value C(orrect), I(nsertion), S(ubstitution) or D(eletion) along with the start and end time of the hypothesised word.

B.9 Evaluation 97 - Files

The evaluation data exists in a slightly different form to the SDR data. Firstly artificial boundaries must be decided upon to divide the data into more reasonably sized chunks (for example less than 5 minutes long). Once these boundaries have been determined, the SPH file `x_eval97_$number` can be created using `w_edit` as before. Assuming the output is given as a CTM file, this can be chopped up into the relevant "stories" and converted into an SRT format using the tool `ctmndx2srt.pl`. The final SRT file is then created by adding an artificial section/episode marking corresponding to the start and end time of the "story".

The STM file can be generated by extracting the relevant lines from the eval STM file and subsequently artificially filtered slightly by removing all instances of the "word" `(%HESITATION)`. The SGML file can then be created using `sclite` with the CTM and STM files as before. In this way the analysis tool can be used to look at small sections of the evaluation data in turn.

`xpert` has added functionality to cope with the phenomena of INTER_SEGMENT_GAP where no speech is transcribed, and EXCLUDED_REGION where speech is transcribed, but excluded from the evaluation scoring. This occurs for example when there are known faults in the manual transcriptions. These phenomena only occur in the evaluation files.

C QUICK USER GUIDE TO `xpert`

Introduction
X Program for Evaluating Recognition Transcriptions (`xpert`), is a tcl/tk tool to allow the user to analyse audio, waveform and transcription of a broadcast news show simultaneously. The errors can be easily seen and focused on to enable a better understanding of how the recognition could be improved.

It is run on an SGI and requires 4 files, namely

SPH file	the audio information
SRT file	the recogniser transcription
STM file	the manual transcription
SGML file	holds the information about the errors

The following gives a quick reference guide of the main features of xpert

Files

The files necessary can be specified in the command line (in any order)
e.g. `xpert <sph-file> <stm-file> <srt-file> <sgml-file>`
Alternatively the **File**→**New** option allows them to be specified once the program is running.
For demo purposes 4 SDR'96 stories and the first 10 minutes of eval'97 are included in the program. They are also available from the **File** menu.
The final option **File**→**Quit** exits the program.

Highlighting Transcriptions

Highlighting a portion of the transcription colours the correct words in blue, insertions (in the recogniser transcription) and deletions (in the manual transcription) in red and substitution errors in green. The following options exist for highlighting

by **speaker**	highlight everything said by a given speaker
by **condition**	highlight everything with a given audio condition
by **line number**	highlight a given line of the transcriptions
by **word match**	highlight any word containing the desired string
between marks*	highlight text corresponding to marked waveform
errors	highlight all the errors in the transcription
unhighlight	remove all highlighting

* N.B. marks can be set by pressing the left mouse button on the waveform and then dragging the mouse to the other mark and releasing. Alternatively the middle mouse button can be used to set the right marker.

Playing Audio

Playing the audio only is achieved via the **Play** menu, which offers:
Play Entire File
Play Between Marks
Play by **Line Number**
Stop playing

Viewing the Waveform

The **View** menu allows parts of the waveform to be analysed in more detail. The options are:

Zoom In	**Zoom Out**
Zoom Full Out	**View Between Marks**
Page Forward	**Page Backward**
by **Line Number**	**Reset**

Tracking

Tracking allows all of the known information to be presented to the user simultaneously. The audio is played, the cursor follows through the waveform, shifting pages at the appropriate time, the line of manual transcription (with its errors) is highlighted and each word in the recogniser transcription is highlighted at the exact moment it was recognised from the audio.

The **Track** options are:
Entire File (N.B. the "START" button also does this option)
by **Line Number**
by **Speaker**
by **Condition**
The "STOP" button stops the tracking.

Inspect

The **Inspect** Menu allows various properties of the files to be seen by the user. They are displayed under the title, next to the "Inspect Value" field. The Inspect options are:

Number of Lines	**HTK Number of Words**	**Manual Number of Words**
Error Stats (C,I,S,D)	that is correct, insertion, substitution and deletion	
Beginning of File	**End of File**	**Length of File**
Error by Speaker	**Error by Condition**	**Error by Line**
Relative Start of Line	**Relative End of Line**	**Length of Line**
Speaker of Line	**Condition of Line**	**View of Wave (start:end)**

Text Focusing

By right-clicking on a word in the recogniser transcription, that word is highlighted, marked on the waveform and zoomed-in on. **Play→Between Marks** can then be used to play the audio for that word.

Middle-clicking on a word in the recogniser output focuses on the word plus two words of context on either side, giving 5 words in total.

Left-clicking on a word in either the manual or the recogniser transcription highlights and marks the corresponding line.

Other Information

Information fields giving the Show, Story, Speaker, Condition and Line Number of the piece of transcription being analysed are constantly updated and displayed on the screen.

Additionally there are 3 buttons:

Quit	Exit the Program
Start	Track entire file
Stop	Stop all audio and tracking

REFERENCES

[1] E. Voorhees, D. Harman
Overview of the Sixth Text REtrieval Conference (TREC-6).
Proc. Sixth Text REtrieval Conference (TREC6) pp. 1-24

[2] D.S. Pallett, J.G. Fiscus, A. Martin, M.A. Przybocki
1997 Broadcast News Benchmark Test Results: English and Non-English.
Proc 1998 DARPA Broadcast News Transcription and Understanding Workshop, Virginia pp. 5-11

[3] P.C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, E. W. D. Whittaker, S. J. Young
The 1997 HTK Broadcast News Transcription System.
Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Virginia pp. 41-48