IMPROVING AUTOMATIC SPEECH TRANSCRIPTION FOR MULTIMEDIA CONTENT

Masood Masoodian
Department of Computer Science
The University of Waikato
Hamilton, New Zealand
m.masoodian@cs.waikato.ac.nz

Bill Rogers
Department of Computer Science
The University of Waikato
Hamilton, New Zealand
b.rogers@cs.waikato.ac.nz

Saturnino Luz
Department of Computer Science
Trinity College Dublin
Dublin, Ireland
luzs@cs.tcd.ie

ABSTRACT

Automatic Speech Recognition systems are increasingly being used in multimedia retrieval applications, where speech recognition is used to aid the creation of high-quality transcripts for data such as multimedia meeting recordings, lectures and presentations, video and audio libraries, and broadcast news. These transcripts are then used for indexing and retrieval of the multimedia content over the Internet. Automatic transcription in these domains is however a challenging task for a number of reasons, including unfavourable recording conditions, high frequency of out-of-vocabulary words and multiplicity of speakers and accents. Furthermore, with the increasing volume of multimedia data, particularly video and speech, being recorded, stored, and shared over the Internet, full manual correction of transcribed speech is impractical. Hybrid transcription systems are therefore needed to allow combining automatic transcriptions for the most part, and manual corrections to some extent, in generating accurate transcription of multimedia content. This paper presents a graphical system which limits human involvement to correcting some, but not all, transcription errors. These corrections can then be used to dynamically update the system vocabulary, thus helping the system to remove related transcription errors automatically.

KEYWORDS

Multimedia retrieval, multimedia indexing, speech recognition, computer-assisted transcription, transcript error correction, usability.

1. INTRODUCTION

Once confined to specialist domains such as law and language studies, the demand for accurate speech transcripts has increased dramatically in recent years. In addition to providing training data for natural language processing applications, speech transcription is becoming an essential component of information retrieval applications that target a proliferating variety of on-line multimedia resources, ranging from recordings of lectures (Hazen, 2006) to news broadcast (Wayne, 1998) to recorded meetings (Luz and Masoodian, 2005).

However, automatic speech recognition of such content presents several challenges. For instance these types of recordings often involve multiple speakers; it is usually not possible to train the system on all the speakers involved; their speech can contain words critical to good indexing (e.g. names of people and places) which usually aren't in the systems' vocabularies.

Under ideal circumstances automatic recognition rates achieved by commercial systems are typically in the 85% to 95% range, and can be further improved to 99% for a single speaker who speaks clearly, only uttering words that are in the system's vocabulary, and on whose voice the system has been explicitly trained. However, outside ideal conditions, in cases such as recorded online multimedia content, error rates can be as high as 60%.

Furthermore, with the increasing volume of multimedia data being recorded, stored and shared over the Internet, it is becoming crucial to develop mechanisms for automatic indexing and retrieval of multimedia content, in a manner similar to what has been achieved with textual information. It is also clearly impossible to carry out full manual correction of the transcribed speech on such content. There is therefore a need for tools to support high-quality transcription of long speech recordings.

Our research focuses on design and development of tools which provide visualisation of the automatic speech recognition process and output, so that users can dynamically intervene in the recognition process, and thus allow the system to propagate their input through the rest of the recognition process. The aim is to limit human involvement in automatic speech recognition to correcting some, but not all, transcription errors; for instance by allowing the user to select from a list of alternatives suggested by the speech recogniser. These corrections can then be used to dynamically update training models and vocabularies, and by doing so the system can remove the remaining related transcription errors automatically.

A number of tools exist which support human transcribers in segmenting and labelling speech signals. However, this type of manual transcription is a time-consuming process which would render applications such as indexing of recorded lectures or broadcast news impractical. Fully automated transcription, on the other hand, typically fails to yield the level of accuracy required by many applications. An approach to speeding up production of accurate, time-aligned speech transcripts has been proposed (Hazen, 2006) which employs a two-stage strategy: first an "approximate" (imperfect, nonaligned) transcript is manually generated, and then automatic speech recognition is employed to align this manually generated transcript to the speech signal, often correcting transcription errors in the process. The approach we propose in this paper takes the opposite route. The first step consists of using an automatic speech recogniser to generate a (typically imperfect but time-aligned) transcription of a long speech recording. From that point on fine tuning is performed through iterations of human-driven error correction steps supported by an interactive system. This approach therefore shifts the focus from automatic recognition to user interface support for error correction.

Error correction is an issue which has been extensively studied by speech recognition and human-factors researchers over the past decades. As with other user interface interaction modalities, the nature of the errors made by users while interacting through the speech modality is inherently dependent on the nature of the application in which speech is used as well as the application's context of use. Thus, error correction strategies for dialogue systems (Hone and Baber, 1999), for instance, differ greatly from the strategies employed in dictation systems (Suhm et al., 2001). The former generally focus on support for increased accuracy and effective repair through exploitation of contextual (pragmatic) constraints (Hone and Baber, 1999) while the latter focus on interface design strategies involving error recognition via the visual modality and repair through speech interaction ("respeaking") (McNair and Waibel, 1994; Ainsworth and Pratt, 1992) and, more recently, through combinations of modalities (Suhm et al., 2001; Halverson et al., 1999; Liu and Soong, 2006).

Interface design strategies for transcription of long speech recordings is an area that has received much less attention than either dialogue or dictation applications. Yet, few of the error prevention and repair techniques employed in dialogue systems and dictation applications can be effectively adapted for use in computer-assisted transcription. The techniques which could be used include, for instance, alternative lists (Ainsworth and Pratt, 1992) and multimodal interaction (Suhm et al., 2001). Although these techniques have been investigated (mainly) in connection with dictation applications, little research exists on using such techniques for transcription of large audio files.

In what follows we present a prototype system to support rapid creation of time-aligned transcripts of speech recordings by human transcribers. This prototype is a basic platform on which various interaction design techniques for computer assisted transcription of long recordings can be investigated.

2. DYNAMIC TRANSCRIPT CORRECTION

In order to investigate techniques for graphical visualisation of automatically generated transcripts and interactive error correction, we have developed a prototype "transcript editor" called TRAEDUP. The current version of the system uses the Microsoft™ Speech Recognition system's interface.

The graphical user interface of the prototype is based on a text editor metaphor augmented with components for visualisation and interactive manipulation of the input speech signal. After "loading" the transcript onto the editor screen, through the recogniser, the system allows its users to correct transcription errors manually. This initial feedback is then used to dynamically update the speech recognition system's vocabulary and speaker model, thus reducing the likelihood of later occurrences of the same transcription errors during the recognition process. TRAEDUP assumes the following scenario of use:

1. The user opens an audio file containing the recorded speech to be transcribed.
2. The system recognises speech from the entire file, producing a transcript (which is likely to contain errors).
3. This imperfect transcript is presented to the user, allowing the user to begin manual correction of errors. After a time, part of the transcript will be (substantially) correct.
4. Some of these corrections will typically be words that are missing from the system's vocabulary (e.g. names of places and people etc). These words are then dynamically added to the vocabulary.
5. Corrected sections of the transcript, which include corrections of mistakes other than out-of-vocabulary words, are used to update training of the speech recognition system's speaker model.
6. The system then moves on to consider the remaining parts of the transcript not yet corrected. These are re-recognised and steps 3-5 of the process are repeated.

It is clear that as the process is repeated through the recorded speech from the beginning to the end, the number of transcription errors due to missing vocabulary words will gradually be reduced at each iteration; thus reducing the amount of manual intervention on the part of the user.

The system imposes no direct control over the sequence in which the user works through the document, but in order to be able to distinguish corrected from uncorrected sections of transcript and to make the process more efficient and minimally disruptive to the user, we assume that the system will initially recognise and present the entire speech audio file and then wait for the user to correct transcription errors in a segment of the transcripts (e.g. a few sentences). Once the user moves on to correcting errors in the next segment, the system will update the vocabulary and use the corrected transcript for its training. A process then starts concurrently with the user's editing activity which consists of re-recognising and presenting transcripts of speech segments below the point at which the user is making corrections. We also assume that the processes of user-initiated error correction and system-driven propagation of corrections will be sequential, starting from the beginning and continuing through to the end of the recorded speech.

Figure 1 shows the user interface of TRAEDUP. In each line TRAEDUP shows the waveform representation of the input speech audio above the transcription of the associated speech. There are several errors in the transcript shown in Figure 1, the most obvious one being the word "Waikato", which despite being a well-known province name in New Zealand, is missing from the speech recogniser's American English vocabulary. The word "Waikato" has been mis-recognised twice, once in the first line as "pointcast's a" and then in the last line as "why can't a".

In this particular instance, TRAEDUP would allow a user to correct the first mis-recognised occurrence of "Waikato", by first joining the words "pointcast's" and "a" into a single word (Figure 2), and then associating "Waikato" with the new single word. The system can then automatically check its vocabulary to see if the word "Waikato" is present. If the word is not present it will be added to the vocabulary. Depending on the granularity with which the system performs re-recognition, further instances of added words that occur later in the document are likely to be correctly recognised, though instances very close to the initial correction will not be subject to re-recognition and will still need manual correction. In the case of the word "Waikato", TRAEDUP would find the next occurrence of it (i.e. "why can't a") and dynamically correct the previously mis-recognised interval (as shown in Figure 3).

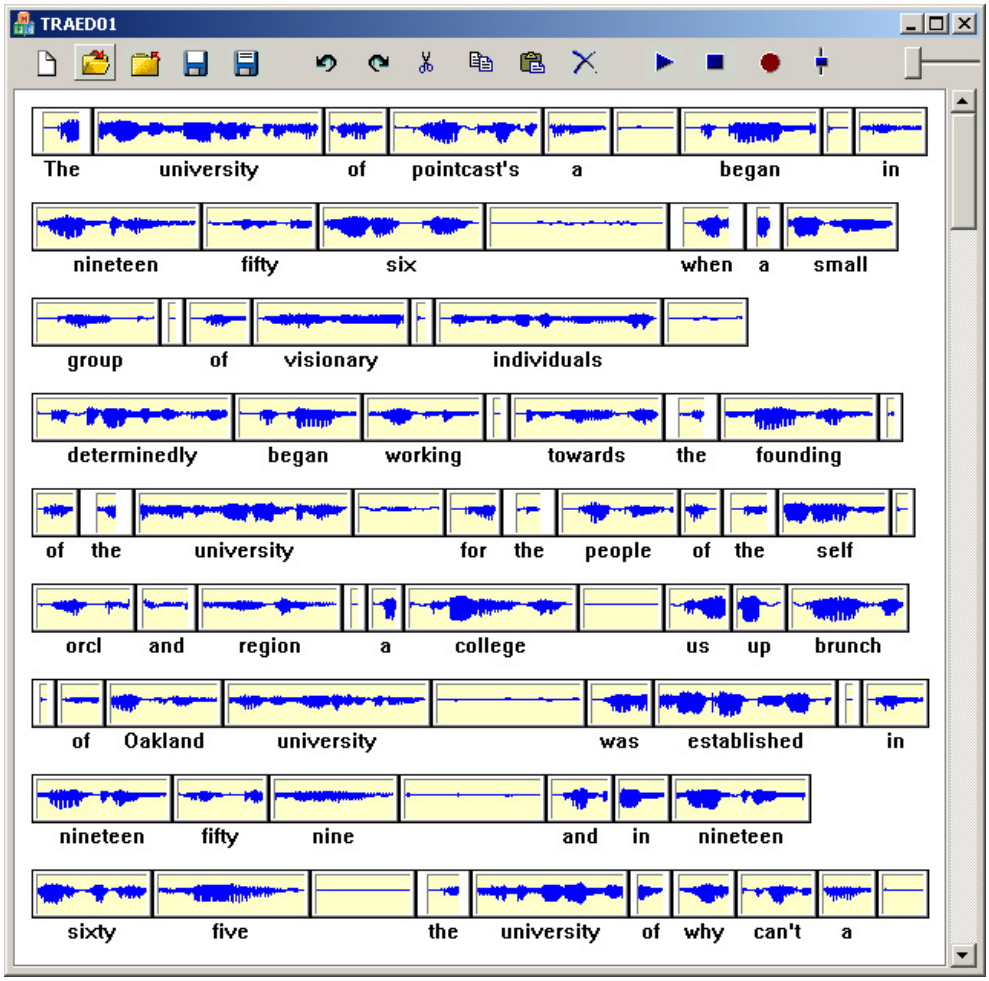

Figure 1: TRAEDUP showing imperfect transcript of a recorded speech audio file

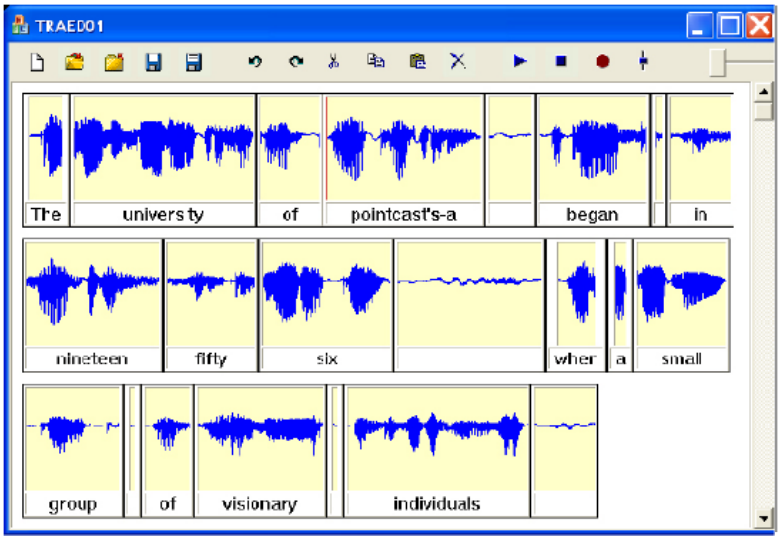

Figure 2: Changing the word boundaries in TRAEDUP

However, out-of-vocabulary words are not the only reasons for transcription errors, and in fact Figure 1 shows several other mistakes made by the speech recognition system. For instance, further to mistakes related to the missing vocabulary words "Waikato" and "Auckland", the system mis-transcribes "a university" and "the south" in line 5, and "a sub branch" in line 6 (Figure 1). Manual correction of these types of mistakes is also used by TRAEDUP for training the speech recognition engine, with the aim of improving the system's accuracy over the remainder of the document in subsequent recognition cycles.

It is unreasonable to assume that users will always be able to fully correct an entire transcript. TRAEDUP allows them to spend as much time making corrections as they can in a given session, and then stop, knowing that the efforts they have made has contributed to improving the quality of the parts of the transcript not checked, as well as any future transcripts that might be made for the same speaker. In addition, TRAEDUP saves its transcripts along with any modifications the user might have made to the recorded speech audio so that work can continue on a transcript document at a later stage. These capabilities target specifically our goals of facilitating dynamic update of the recognition system's vocabulary, improve its accuracy through dynamic training, and to provide software support for efficient manual correction of imperfect computer-generated transcripts.

TRAEDUP is based on an earlier prototype system called TRAED (Masoodian et al., 2006), and therefore inherits and extends many of the useful speech audio editing and transcript correction functionality of that system. The most important editing mode for transcript correction is the tape-recorder play/stop functionality (shown in top right-hand side toolbox of Figure 3). This has been carefully integrated with the text editing keyboard commands to allow the user to smoothly move between the listen/check and transcript correction processes. In many applications it is also helpful to be able to "clean up" the transcript and the speech audio by removing noise, speech dysfluencies (such as false starts, repetitions, filled pauses, hesitation, etc), and sometimes even whole sentences. Commercial human transcription services often correct and remove speech errors (Hazen, 2006). TRAEDUP supports this type of task by offering the option, inherited from TRAED, of editing the audio at the same time as the transcript.

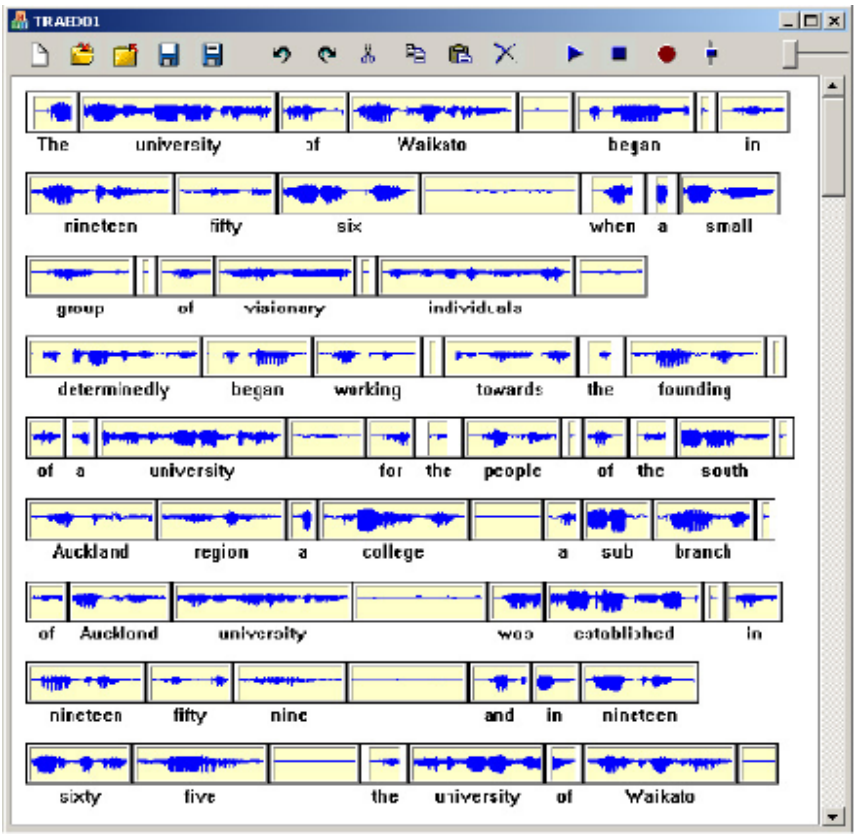

Figure 3: TRAEDUP showing dynamically corrected transcript of a recorded speech audio file

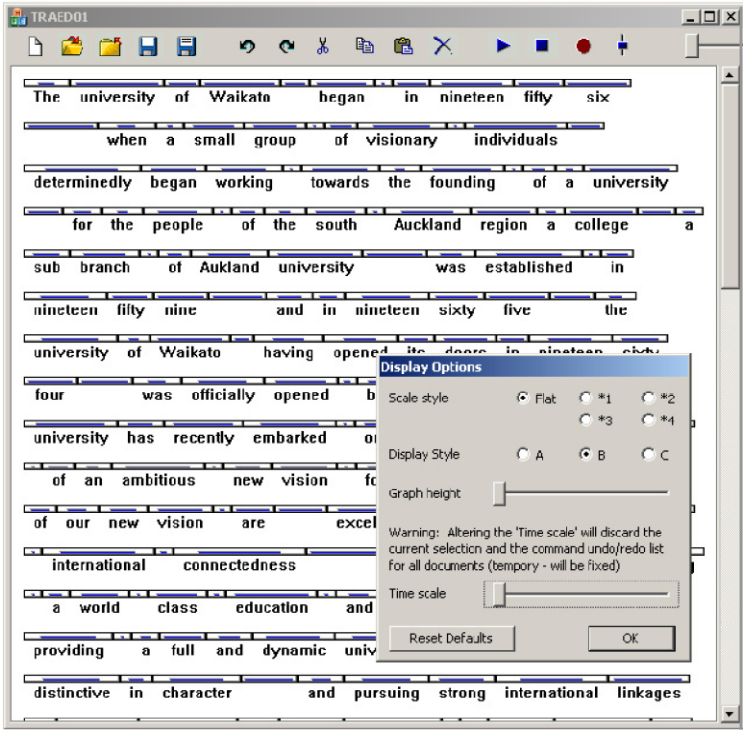

Figure 4: Display settings of TRAEDUP

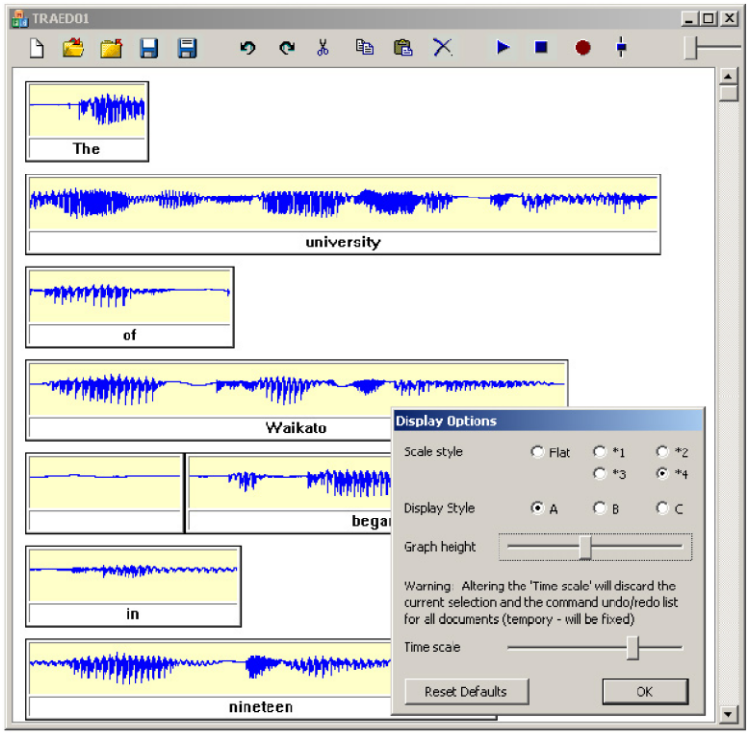

Figure 5: Zoomed view of showing word boundaries

TRAEDUP also has several display setting options, shown in Figure 4, which users can set depending on the type of transcription activity they are performing. In situations where it is necessary, for instance, to see the exact time alignment of speech and transcription word boundaries, the user can change the zooming level (i.e. time scale and waveform amplitude), and add speech and text boundary markers (see Figure 5).

3. EVALUATION

We are yet to conduct a comprehensive evaluation of the system in terms of its overall usability and effectiveness in improving the recognition rates for a typical transcription task. We have, however, performed a simple test using a shortened version of an arbitrarily chosen segment of text read by a New Zealand male English speaker. The recorded speech was transcribed using the Speech Recognition system, in two segments, generating 30 errors for each segment of about 400 words. Of these 30 errors, 7 were out-of-vocabulary words (e.g. Waikato, Maori, etc.), and 23 were other transcription errors.

TRAEDUP was used to correct the transcription errors in the first segment, by adding the out of vocabulary words to the system, and fixing the others manually. The corrected transcripts were then used to trigger the system's training cycle. We then used the system to transcribe the recorded speech for the second segment. Although the transcription errors for the added out of vocabulary words were almost completely fixed, the training process did not seem to produce a significant improvement.

Although these preliminary tests are clearly insufficient, they indicate that further work needs to be carried out to take better advantage of the information provided by the user's corrective actions. This could take the form of improvements to the process of incremental training of the speech recognition system used by TRAEDUP, but could also explore simpler strategies. An approach that involves re-scoring of the recognition lattice is currently being tested.

4. CONCLUSIONS

This paper described TRAEDUP, a dynamic speech transcription correction system. Most of our past effort in the development of this prototype has focused on providing the type of functionality needed to allow users to easily edit the imperfect transcriptions produced by an automatic speech recognition system. Techniques implemented in this prototype aimed to provide better functionality for natural integration of text and speech signal visualisation, including play/pause/resume actions, changing word boundaries, etc.

Our more recent work, as described in this paper, shows the possible improvements which can be made to the task of transcription correction, by automating the dynamic addition of out of vocabulary words to the speech recognition system's dictionary, and using segments of the corrected speech transcripts for training the recogniser's speaker model. These two methods of improvement are particularly valuable for automatic transcription of recorded multimedia data, where training of the recognition system on speakers is not feasible, and many out-of-vocabulary words such as names of places and people appear regularly in the recorded speech.

We are currently working on a new implementation using CMU's open-source speech recognition engine, Sphinx. This new implementation will hopefully allow us greater flexibility in implementing error correction strategies based on incremental training of the recogniser and re-scoring of the recognition lattice.

ACKNOWLEDGMENTS

The work of Saturnino Luz was supported by the Science Foundation Ireland under Grant No. SFI RFP 06/RFP/CMS054.

REFERENCES

Ainsworth, W. A. and Pratt, S. R., 1992. Feedback Strategies for Error Correction in Speech Recognition Systems. *International Journal of Man-Machine Studies,* Vol. 36, No. 6, pp. 833–842.

Halverson, C. A. et al., 1999. The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems. *Proceedings of INTERACT'99*, pp. 133–140.

Hazen, T., 2006. Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings. *Procedings of Interspeech'06*, Pittsburgh, Pennsylvania, pp. 1606–1609.

Hone, K. S. and Baber, C., 1999. Modelling the Effects of Constraint upon Speech-based Human-computer Interaction. *Interface Journal of Human-Computer Studies*, Vol. 50, No. 1, 85–107.

Liu, P. and Soong, F. K., 2006. Word Graph Based Speech Recognition Error Correction by Handwriting Input. *Proceedings of the 8th international conference on Multimodal interfaces,* New York, USA, pp. 339–346.

Luz, S. and Masoodian, M., 2005. A Model for Meeting Content Storage and Retrieval. *Proceedings of the 11th International Conference on Multi-Media Modeling,* Melbourne, Australia, pp. 392–398.

Masoodian, M. et al., 2006. TRAED: Speech Audio Editing Using Imperfect Transcripts. *Proceedings of the 12th International Conference on Multi-Media Modeling,* Beijing, China, pp. 454–259.

McNair, A. and Waibel, A., 1994. Improving Recognizer Acceptance Through Robust, Natural Speech Repair. *Proceedings of the 3rd International Conference on Spoken Language Processing,* Japan, pp. 1299–1303.

Suhm, B. et al., 2001. Multimodal Error Correction for Speech User Interfaces. *ACM Transactions on Computer-Human Interaction*, Vol. 8, No. 1, pp. 60–98.

Wayne, C. L., 1998. Topic Detection and Tracking (TDT): Overview and Perspectives. *Proceedings of the Broadcast News Transcription and Understanding Workshop,* Virginia, USA.