Interactive Visualisation Techniques for Dynamic Speech Transcription, Correction and Training

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

Masood Masoodian  
Dept. of Computer Science  
The University of Waikato  
New Zealand  
masood@cs.waikato.ac.nz

Bill Rogers  
Dept. of Computer Science  
The University of Waikato  
New Zealand  
coms0108@cs.waikato.ac.nz

ABSTRACT
As performance gains in automatic speech recognition systems plateau, improvements to existing applications of speech recognition technology seem more likely to come from better user interface design than from further progress in core recognition components. Among all applications of speech recognition, the usability of systems for transcription of spontaneous speech is particularly sensitive to high word error rates. This paper presents a series of approaches to improving the usability of such applications. We propose new mechanisms for error correction, use of contextual information, and use of 3D visualisation techniques to improve user interaction with a recogniser and maximise the impact of user feedback. These proposals are illustrated through several prototypes which target tasks such as: on-line transcript editing, dynamic transcript editing, and real-time visualisation of recognition paths. An evaluation of our dynamic transcript editing system demonstrates the gains that can be made by adding the corrected words to the recogniser’s dictionary and then propagating the user’s corrections.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems; H.5.2 [User Interfaces]: Natural Language

General Terms
Human Factors

Keywords
Automatic Speech Transcription, Error-correction, Speech Recogniser Training, Semi-automatic Speech Transcription

1. INTRODUCTION
Transcription is one of the major applications of automatic speech recognition, as evidenced by the existence of workshops and evaluation tasks specialised on this topic [21, 10].

Regardless of the initial accuracy levels achieved through automatic speech recognition, when 100% accuracy is required, manual correction of transcripts by human operators becomes necessary [10, 2]. Depending on the quality of the automatically generated transcripts the task of making corrections by a human transcriber can be rather time consuming [19, 11]. For instance when the original transcripts contain something like 50% errors, the manual correction task is almost as time consuming as transcribing the entire speech from scratch; requiring several iterations of listening to the speech audio and making corrections.

The task of generating transcripts from live speech in real-time is even more complicated. In such cases automatic speech recognisers are even less accurate than when working on recorded speech where transcription speed is not an issue. Real-time speech transcription is therefore generally carried out by highly trained human transcribers, which is costly and therefore of limited in application.

These demonstrative examples of problems associated with automatic as well as manual transcription of speech identify a clear need for systems that combine the best aspects of automatic speech transcription (i.e. speed) with that of manual transcription (i.e. accuracy) [15]. Such semi-automatic systems, however, need to minimise the human intervention...
while maximising the speed and accuracy of the generated transcripts.

In this paper we review a range of solutions to the problem of balancing the need for speed and accuracy in semi-automatic speech transcription through the development of interactive visual techniques which aim to assist human transcribers intervening in the process of speech recognition and transcription. Several interactive prototype systems are described which demonstrate the use of visual techniques in a range of tasks including correcting transcripts of recorded speech, dynamic correction of speech transcripts, and dynamic training of speech recognition systems.

2. HUMAN INTERVENTION IN SPEECH RECOGNITION

There are many cases, such as in dictation, where it is crucial to generate completely accurate transcripts of speech. In such cases transcription of the recordings can be generated manually by human operators. However, as mentioned above, this process can be slow and expensive. Similarly, automatic speech transcription systems, with their inherent inaccuracies, are less than ideal for this type of task.

Clearly a better solution is to use a semi-automatic approach where an automatic speech recognition system can be used to generate transcripts (with possible errors), which can then be manually corrected by a human operator. The success of this mode of operation depends largely on providing the human operator with visual tools that would allow easy identification of the errors made by the automatic speech recogniser, and then allow error correction with minimal effort on the part of the operator. Error correction mechanisms have been studied in the human-factors literature mainly in connection with dictation and dialogue systems [20, 11, 1, 9]. In general, these error correction facilities can be divided into two groups:

1. interactive visual facilities for locating and correcting errors, without involving the speech recogniser,
2. facilities for automatic propagation of manual corrections made by the operator, through iterative involvement of the speech recogniser.

The type of facilities provided by semi-automatic systems are also dependant on the granularity or the level at which human interventions (i.e. corrections) are made possible. This in turn is dependant on the application area of transcription, which can either be real-time transcription (e.g. word by word or sentence by sentence) or time-delayed transcription (e.g. paragraph by paragraph, or an entire recorded session).

Interactive visual facilities of type 1, which only allow locating and correcting errors on automatically generated transcripts, without further involvement of the speech recogniser, only need to work at the entire recorded session level.

Figure 1: TRAED speech transcript editor.
The scenario of work in this case is to use the recogniser to transcribe the entire session and then allow the user to make corrections in any order; although sequential mode of making correction from the beginning to end is more likely to be adopted.

Error correction facilities of type 2, on the other hand, which allow propagation of manual corrections through further iterative involvement of the speech recogniser, need to work at the word, sentence, or paragraph level. The scenario of work in these situations is to use the recogniser to start transcribing speech input (either from live or recorded speech input), and then depending on the level of granularity provide means of intervention (e.g. making manual corrections or selecting from a list of suggestions) appropriately. Then, once some user interventions have been made the system can use those single or multiple interventions to guide the recogniser through the remainder of the transcription process.

Our research has focused on designing interactive visualisation techniques which provide these types of facilities to the users. We have developed a range of prototype systems incorporating these techniques, with the aim of making the task of semi-automatic transcription as efficient as possible in different modes of operation.

### 3. LOCATING AND CORRECTING ERRORS

The first prototype system we developed was specifically designed to allow users to make corrections to transcripts generated using Microsoft’s Speech Application Programming Interface (SAPI). The user interface of this system, called TRAED (TRAnscript EDITor), is shown in Figure 1.

This system, which is more fully described elsewhere, displays the audio speech signal in waveform along with its transcription line by line. Boxes around each recognised word serve to delimit the associated speech signal. The waveform signal for each word is displayed at a fixed timescale, and their associated words are centred under the waveforms. The chosen scale is such that the waveform is usually wider than the word, and in cases where the text is wider, the left and right borders of the waveform display area are widened, leaving the wave itself correctly scaled.

This simple technique for associating the transcribed text with their time-aligned speech signals allows the users to easily identify mis-recognised words, and if in doubt play back the associated speech audio, without having to search through the audio recording using the more common, and time consuming, fast-forward and rewind method.

Another obvious advantage of the time-alignment of transcribed text and audio signal is that the areas of silence, between word gaps, etc, which have no transcription can easily be seen; as unrecognised audio segments are represented by single space characters (see the spaces between some of the recognised words in Figure 2).

TRAED also provides a simple mechanism for dealing with one of the most common type of errors made by speech recognisers; which is not being able to identify word boundaries correctly. This problem results in a long word being broken down into several smaller words, or alternatively (though less commonly) in smaller words being joined up into a long mis-recognised word. Sometimes speech recognisers also mis-align boundaries of correctly recognised words, in which case being able to align the boundaries manually becomes necessary. This is specially true of tasks such as creation of speech corpora, in which getting the alignment between transcript and audio waveform (or spectrogram) right is essential. In TRAED adjusting word boundaries can be carried out using simple mouse operations (e.g. clicking and dragging).

Although TRAED allows easy correction of transcript errors, it does not provide any clues as to why the speech recogniser used to generate the transcripts has made those errors. Modern automatic speech recognition systems generally use Hidden Markov Models to compare statistical probabilities associated with different words in their generated “word lattice” to select a word from a set of likely possibilities. However, after a word has been selected (which may actually be incorrect) the other possibilities (which may include the correct word) are simply ignored.

TRAED also only displays the final transcript (correct or incorrect) returned by its recogniser. The human transcriber using TRAED does not get to see the set of possible words from the word lattice, and can only manually correct the mistakes made by the recogniser by typing in the corrections, and when necessary change the word boundaries.

Since Microsoft’s SAPI does not provide access to its generated word lattice, we developed another prototype system, called TRAENX, to provide the words selection mechanism to the users. TRAENX uses the open-source Sphinx speech recognition engine, which allows access to intermediate word lattices as well as greater control over the recognition process.

TRAENX also provides a simple mechanism for dealing with one of the most common type of errors made by speech recognisers; which is not being able to identify word boundaries correctly. This problem results in a long word being broken down into several smaller words, or alternatively (though less commonly) in smaller words being joined up into a long mis-recognised word. Sometimes speech recognisers also mis-align boundaries of correctly recognised words, in which case being able to align the boundaries manually becomes necessary. This is specially true of tasks such as creation of speech corpora, in which getting the alignment between transcript and audio waveform (or spectrogram) right is essential. In TRAED adjusting word boundaries can be carried out using simple mouse operations (e.g. clicking and dragging).

Figure 2 shows the user interface of TRAENX. Although the visual aspect of TRAENX is similar to that of TRAED, and despite the fact that it provides similar functionality to TRAED, there is one crucial difference between the two prototypes. Unlike TRAED, TRAENX clearly shows which words in the transcription were selected from a set of possible words by the speech recogniser. These words are shown as buttons in the transcript, and the user can click on them to see what the other possibilities were. If a different option is the correct word, the user can select it instead of the incorrect (automatically chosen) alternative. This is illustrated in Figure 3.

---

1[^1]: [http://research.microsoft.com/research/srg/sapi.aspx](http://research.microsoft.com/research/srg/sapi.aspx)

2[^2]: [http://cmusphinx.sf.net/](http://cmusphinx.sf.net/)

Figure 3: TRAENX speech transcript editor.

Figure 4: Word selections in TRAENX.

4. PROPAGATING USER CORRECTIONS

Despite the fact that both TRAED and TRAENX provide useful transcript error correction facilities to human operators, their functionality is limited to the initial transcription results returned by their backend speech recogniser, without providing any functionality for dynamic interaction between the operator and the speech recogniser. Although the form of “post-transcription correction” mode of operation implemented by TRAED and TRAENX might be sufficient in some use cases, further assistance can be provided to the users through dynamic interaction with the speech recogniser.

In many speech transcription cases (e.g. transcription of broadcast news) there are often phrases or nouns that are repeated throughout a particular segment of speech being transcribed. If, for example, these repeated nouns are missing from the recogniser’s vocabulary (as often is the case with proper nouns), or if the system has not been trained on that particular speaker (also very common in cases such as broadcast news), then it is likely that the speech recogniser will transcribe those nouns or phrases incorrectly repeatedly throughout the transcription. In such cases, TRAED or TRAENX will assist the user with the task of finding and correcting those transcription errors, but neither of these systems has any mechanism for learning from the initial corrections that the user makes so that other repeated cases can be fixed automatically.

Furthermore, since it is often likely that the speech recogniser would have mis-recognised the same noun differently (though be it incorrectly) at different points within the segment, an automatic find and replace functionality will have very little value in most cases.

Therefore, more advanced techniques are needed to allow the system to learn from the corrections that the user makes, and propagate the results of the user’s interventions as much as possible by guiding the backend speech recogniser dynamically throughout the entire transcription process in an iterative manner. These advanced dynamic correction techniques can be divided into three groups:

1. making word additions to the speech recogniser’s vocabulary,
2. training the speech recogniser,
3. guiding the selection of alternatives from the word lattice.

We have developed two prototype systems to demonstrate how each of these techniques could be used to provide for dynamic user intervention in semi-automatic transcription tasks.

The user interface of the first of these two prototypes is shown in Figure 5. This system which is called TRAEDUP...
Figure 5: TRAEDUP speech transcript editor.

is based on TRAED and has been more fully described elsewhere [13]. Although TRAEDUP is visually similar to TRAED, their modes of operation are rather different. TRA-ED utilises its backend speech recogniser to produce a complete transcription of its input speech data, which is then shown to the user so that manual corrections can be made. TRAEDUP also utilises Microsoft SAPI as its backend speech recogniser. However, unlike TRAED, its mode of operation is iterative and can be summarised as follows:

1. the recogniser is provided with the speech input to be transcribed,
2. the entire input is transcribed (probably with errors) and presented to the user,
3. the user starts making manual correction from the beginning, producing transcripts that will be (substantially) correct,
4. corrections to words that are missing from the system’s vocabulary are dynamically added to the vocabulary,
5. corrections of mistakes other than out-of-vocabulary words are used for training the speech recogniser’s speaker model,
6. the system re-transcribes the speech related to the parts that are not yet corrected by the user,
7. steps 3-6 are repeated until no more corrections are needed.

The advanced transcription correction facilities of TRAEDUP come under the first two categories mentioned above (i.e. additions to the recogniser’s vocabulary and training). However, once again, since TRAEDUP is based on TRAED and utilises a proprietary API, we were unable to access the recogniser’s internal working components during transcription, including the word lattice it generates. Therefore, to demonstrate the third category of more advanced dynamic correction techniques we developed another prototype using the previously mentioned Sphinx-4 open-source speech recognition engine.

Figure 6 shows the interface of this prototype called DYTREAO (DYnamic TRANscript EDitor). DYTREAO is radically different from all the other systems described here, and is the first system to actually use the word lattice produced by its backend speech recogniser in real-time in order to propagate the results of user interventions (e.g. selections and corrections) through the transcript.

Also, unlike the previous systems, which have a linearly textual representation of transcripts along with their associated speech signal in waveform, DYTREAO employs a directed graph-based 3D visualisation technique to depict the recognition alternatives as generated by the speech recogniser in real-time.

As the technical details of DYTREAO and its implementation are described elsewhere [12], it is sufficient here to mention how it is operated by a human transcriber.
When DYTRAED receives its speech audio signal, either from a live speaker or a recorded file, it starts transcribing its input using the speech recogniser engine, which in turn generates the word lattice as part of its recognition process. DYTRAED receives these recognition alternatives from its recogniser and displays them in the form of a directed graph to the user. The partial recognition alternatives with the greatest probability scores are highlighted and connected with red lines, the lower-score alternatives are dimmed, and the alternatives undergoing active search are highlighted and connected with yellow lines.

The user is able to: pause, speed up, or slow down the animation; click on a word to select and accept the entire sentence being transcribed; or simply ignore and continue to visualise the recogniser’s preferred transcription being generated. When the transcription of a given sentence is completed (either by the operator or the system) it moves to the background, and then gradually is pushed towards the horizon by sentences completed afterwards, until it disappears completely from the screen.

5. EVALUATION

Although the tools and techniques described in this paper seem to be valuable in providing mechanisms for assisting users with the tasks of semi-automatic speech transcription, it is clearly important to evaluate them to make sure that they are indeed as efficient as possible in their usability.

We have started with the process of conducting formal evaluations of these tools, both in terms of their usability, as well as their efficiency, accuracy, etc. A brief testing of TRAEDUP system showed promising results, and allowed us to design specific tests to evaluate the effects of dynamic word addition and recogniser training.

So far we have carried out a further test to see if dynamic word additions make a sufficient difference to improving the accuracy of TRAEDUP’s back-end speech recogniser. For this test we focused on the fact that one of the main difficulties speech recognisers often face is recognising proper nouns, which tend to occur regularly in cases such as broadcast news. The problem with proper nouns is that they are either missing from the recognisers’ vocabulary, or are pronounced differently by people.

Six audio files were used in this experiment. They were from the CMU ARCTIC database and downloaded from the Voxforge corpus collection, which includes the speech audio recordings and their correct transcriptions. The speech files we chose were by two US female speakers, two US male speakers, a Canadian male speaker, and a Scottish male speaker. Each speech file contained 1132 sentences, consisting of around 10000 words.

We chose five proper nouns (Philip, Saxon, MacDougall, Gregson, Jeanne) which together occurred 76 times in each of the recordings (0.8% of the content). We used one instance of each of these nouns from one of the US female speaker’s recording to train the system. In the case of Gregson and Jeanne this also added them to the recogniser’s vocabulary.

Table 1 shows the results of the word additions and training test. The results clearly indicate the benefits of adding words and training on even a single case of each word.
fact the results are quite impressive considering the fact that addition and training of the recogniser on 5 words improves their recognition across different speakers on average from 37% to 89%. In fact if we exclude the word “Philip” in this sample case the average accuracy of recognition for the other 4 words increases from 9% to nearly 89%.

Further investigation of the errors after word additions and training in our test shows that we could achieve 100% accuracy by adding another instance of each noun which includes the possessive “’s”, for example “Gregson’s”, “Jeanne’s”, etc.

### 6. OTHER TECHNIQUES

The techniques described in this paper have focused on error correction through manual intervention. However, in certain scenarios it is possible to provide more assistance to operators of semi-automatic speech transcription tasks through techniques that automatically intervene in the speech recognition process by analysing the context of speech through other modalities. Such techniques can sometimes be used to improve the accuracy of speech transcription to a level that is suitable for applications where 100% accuracy is not critically required, and therefore remove the need for human intervention altogether. That is typically the case of applications that use transcripts to aid user navigation and search of time-based data such as multimedia recordings of meetings. “Meeting browsers” [4], for instance, can benefit from transcribed speech even if the transcription is not perfect. On the other hand, records of textual data produced at the meetings can be used to guide speech recognition results.

The MeetingMiner [3], for instance, is a system that uses time-stamped textual records to improve speech recognition results. It does so by constraining recognition alternatives, re-scoring word lattices by boosting the scores of words that occur on text segments in the “temporal neighbourhood” of the sentence being transcribed and adding out-of-vocabulary items to the recogniser’s vocabulary through text analysis. A detailed evaluation of an information retrieval task supported by speech recognition alone versus the same task supported by combined speech recognition, temporal information and text analysis has shown that increases of up to 43% in precision can be obtained even with baseline (speech-recognition only) transcripts that have a mean error rate as high as 60.2% [3].

### 7. CONCLUSIONS

In this paper we have described a range of interactive visualisation techniques developed to assist human operators with the process of correcting transcription errors in semi-automatic transcription applications. These techniques have been implemented in several prototype systems which demonstrate their effectiveness in assisting the users.

Due to the limitations of the proprietary automatic speech recogniser we chose initially we were unable to implement all the techniques we have described in this paper in a single prototype system. We have now moved to using the open-source Sphinx speech recognition engine which has given us more capabilities. Our aim is to port our initial prototypes and combine them all into a single system to give the users the option of using the techniques that best suit their needs.

It is possible that indeed in some cases users may utilise a combination of facilities sequentially to minimise their efforts in creating correct transcriptions. For instance a user may wish to rely on the techniques provided by DYTRAED to transcribe live speech, which may result in missing some transcription errors due to the real-time limitations of the task. Once the live speech session is over, the user may then utilise facilities provided by TRAEDUP and TRAENX to correct any errors missed during the live transcription session. Of course while the user is carrying out these transcription correction activities the system would be utilising the user’s input to train its speaker models as well as updating its vocabulary collection.

### 8. ACKNOWLEDGEMENTS

The work of Saturnino Luz is partially supported by Science Foundation Ireland through a CSET (CNGL) grant.
9. REFERENCES


