

A Computer Assisted Speech Transcription System

Alejandro Revuelta-Martínez, Luis Rodríguez, Ismael García-Varea

Computer Systems Department
University of Castilla-La Mancha
Albacete, Spain

{Alejandro.Revuelta,Luis.RRuiz,Ismael.Garcia}@uclm.es

Abstract

Current automatic speech transcription systems can achieve a high accuracy although they still make mistakes. In some scenarios, high quality transcriptions are needed and, therefore, fully automatic systems are not suitable for them. These high accuracy tasks require a human transcriber. However, we consider that automatic techniques could improve the transcriber's efficiency. With this idea we present an interactive speech recognition system integrated with a word processor in order to assist users when transcribing speech. This system automatically recognizes speech while allowing the user to interactively modify the transcription.

1 Introduction

Speech has been the main mean of communication for thousands of years and, hence, is the most natural human interaction mode. For this reason, Automatic Speech Recognition (ASR) has been one of the major research interests within the Natural Language Processing (NLP) community.

Although current speech recognition approaches (which are based on statistical learning theory (Jelinek, 1998)) are speaker independent and achieve high accuracy, ASR systems are not perfect and transcription errors rise drastically when considering large vocabularies, dealing with noise environments or spontaneous speech. In those tasks (as for example, automatic transcription of parliaments proceedings) where perfect recognition results are required, ASR can not be fully reliable so far and, a human transcriber has to check and supervise the automatically generated transcriptions.

In the last years, cooperative systems, where a human user and an automatic system work together, have gained growing attention. Here we present a system that interactively assists a human transcriber when using an ASR software. The proposed tool is fully embedded into a widely used and open source word processor and it relies on an ASR system that is proposing suggestions to the user in the form of practical transcriptions for the input speech. The user is allowed to introduce corrections at any moment of the discourse and, each time an amendment is performed, the system will take it into account in order to propose a new transcription (always preserving the decision made by the user, as can be seen in Fig. 1). The rationale behind this idea is to reduce the human user's effort and increase efficiency.

Our proposal's main contribution is that it carries out an interactive ASR process, continually proposing new transcriptions that take into account user amendments to increase their usefulness. To our knowledge, no current transcription package provides such an interactive process.

2 Theoretical Background

Computer Assisted Speech Recognition (CAST) can be addressed by extending the statistical approach to ASR. Specifically, we have an input signal to be transcribed \mathbf{x} and the user feedback in the form of a fully correct transcription prefix \mathbf{p} (an example of a CAST session is shown in Fig. 1). From this, the recognition system has to search for the optimal completion (suffix) $\hat{\mathbf{s}}$ as:

$$\begin{aligned}\hat{\mathbf{s}} &= \arg \max_{\mathbf{s}} \Pr(\mathbf{s} \mid \mathbf{x}, \mathbf{p}) \\ &= \arg \max_{\mathbf{s}} \Pr(\mathbf{x} \mid \mathbf{p}, \mathbf{s}) \cdot \Pr(\mathbf{s} \mid \mathbf{p}) \quad (1)\end{aligned}$$

where, as in traditional ASR, we have an acoustic model $\Pr(x \mid \mathbf{p}, s)$ and a language model $\Pr(s \mid \mathbf{p})$. The main difference is that, here, part of the correct transcription is available (prefix) and we can use this information to improve the suffix recognition. This can be achieved by properly adapting the language model to account for the user validated prefix as it is detailed in (Rodríguez et al., 2007; Toselli et al., 2011).

As was commented before, the main goal of this approach is to improve the efficiency of the transcription process by saving user keystrokes. Off-line experiments have shown that this approach can save about 30% of typing effort when compared to the traditional approach of off-line post-editing results from an ASR system.

3 Prototype Description

A fully functional prototype, which implements the CAST techniques described in section 2, has been developed. The main goal is to provide a completely usable tool. To this end, we have implemented a tool that easily allows for organizing and accessing different transcription projects. Besides, the prototype has been embedded into a widely used office suite. This way, the transcribed document can be properly formatted since all the features provided by a word processor are available during the transcription process.

3.1 Implementation Issues

The system has been implemented following a modular architecture consisting of several components:

- *User interface*. Manages the graphical features of the prototype user interface.
- *Project management*: Allows the user to define and deal with transcription projects. These projects are stored in XML files containing parameters such as input files to be transcribed, output documents, etc.
- *System controller*. Manages communication among all the components.
- *OpenOffice integration*: This subsystem provides an appropriate integration between the CAST tool and the OpenOffice¹ software suite. The transcriber has, therefore, full access to a word processor functionality.

¹www.openoffice.org

- *Speech manager*: Implements audio playback and synchronization with the ASR outcomes.
- *CAST engine*: Provides the interactive ASR suggestion mechanism.

This architecture is oriented to be flexible and portable so that different scenarios, word processor software or ASR engines can be adopted without requiring big changes in the current implementation. Although this initial prototype works as a standalone application the followed design should allow for a future “in the cloud” tool, where the CAST engine is located in a server and the user can employ a mobile device to carry out the transcription process.


With the purpose of providing a real-time system response, CAST is actually performed over a set of word lattices. A lattice, representing a huge set of hypotheses for the current utterance, is initially used to parse the user validated prefix and then to search for the best completion (suggestion).

3.2 System Interface and Usage

The prototype has been designed to be intuitive for professional speech transcribers and general users; we expect most users to quickly get used to the system without any previous experience or external assistance.

The prototype features and operation mode are described in the following items:

- The initial screen (Fig. 2) guides the user on how to address a transcription project. Here, the transcriber can select one of the three main tasks that have to be performed to obtain the final result.
- In the project management screen (Fig. 3), the user can interact with the current projects or create a new one. A project is a set of input audio files to be transcribed along with the partial transcription achieved and some other related parameters.
- Once the current project has been selected, a transcription session is started (Fig. 4). During this session, the application looks like a standard OpenOffice word processor incorporating CAST features. Specifically, the user can perform the following operations:



| | | |
|---------------|------------|---|
| | utterance | |
| ITER-0 | prefix | () |
| ITER-1 | suffix | (<i>Nine extra soul are planned half beam discovered these years</i>) |
| | validated | (Nine) |
| | correction | (extrasolar) |
| ITER-2 | prefix | (Nine extrasolar) |
| | suffix | (<i>planets have been discovered these years</i>) |
| | validated | (planets have been discovered) |
| FINAL | correction | (this) |
| | prefix | (Nine extrasolar planets have been discovered this) |
| | suffix | (<i>year</i>) |
| | validated | (#) |
| | prefix | (Nine <u>extrasolar</u> planets have been discovered <u>this</u> year) |

Figure 1: Example of a CAST session. In each iteration, the system suggests a suffix based on the input utterance and the previous prefix. After this, the user can validate part of the suggestion and type a correction to generate a new prefix that can be used in the next iteration. This process is iterated until the full utterance is transcribed.

The user can move between audio segments by pressing the “fast forward” and “rewind” buttons. Once the a segment to be transcribed has been chosen, the “play” button starts the audio replay and transcription. The system produces the text in synchrony with the audio so that the user can check in “real time” the proposed transcription. As soon as a mistake is produced, the transcriber can use the “pause” button to interrupt the process. Then, the error is corrected and by pressing “play” again the process is continued. At this point, the CAST engine will use the user amendment to improve the rest of the transcription.

- When all the segments have been transcribed, the final task in the initial screen allows the user to open the OpenOffice’s PDF export dialog to generate the final document.

A video, showing the prototype operation mode, can be found on the following website: www.youtube.com/watch?v=vc6bQCtYVR4.

4 Conclusions and Future Work

In this paper we have presented a CAST system which has been fully implemented and integrated into the OpenOffice word processing software. The implemented techniques have been tested of-line and the prototype has been presented to a reduced number of real users.

Preliminary results suggest that the system

could be useful for transcribers when high quality transcriptions are needed. It is expected to save effort, increase efficiency and allow inexperienced users to take advantage of ASR systems all along the transcription process. However, these results should be corroborated by performing a formal usability evaluation.

Currently, we are in the process of carrying out a formal usability evaluation with real users that has been designed following the ISO/IEC 9126-4 (2004) standard according to the efficiency, effectiveness and satisfaction characteristics.

As future work, it will be interesting to consider concurrent collaborative work at both, project and transcription levels. Other promising line is to consider a multimodal user interface in order to allow users to control the playback and transcription features using their own speech. This has been explored in the literature (Rodríguez et al., 2010) and would allow the system to be used in devices with constrained interfaces such as mobile phones or tablet PCs.

Acknowledgments

Work supported by the EC (ERDF/ESF) and the Spanish government under the MIPRCV “Consolider Ingenio 2010” program (CSD2007-00018), and the Spanish *Junta de Comunidades de Castilla-La Mancha* regional government under projects PBI08-0210-7127 and PPII11-0309-6935.

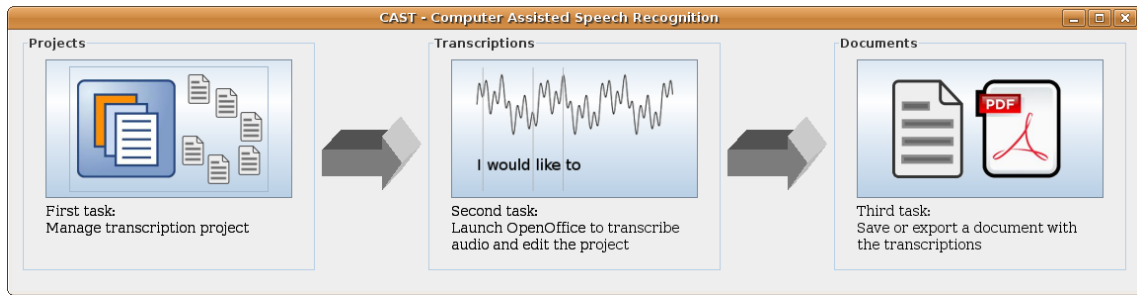


Figure 2: Main window prototype. The three stages of a transcription project are shown.

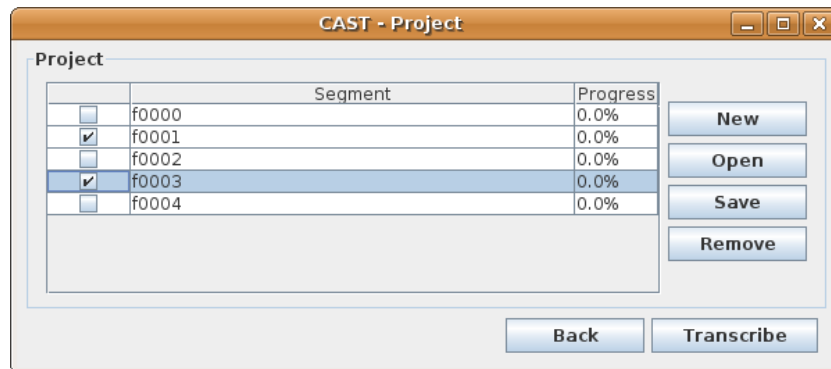


Figure 3: Screenshot of the project management window showing a loaded project. A project consists of several audio segments, each of them is stored in a file so that the user can easily add or remove files when needed. In this screen the user can choose the current working segments.

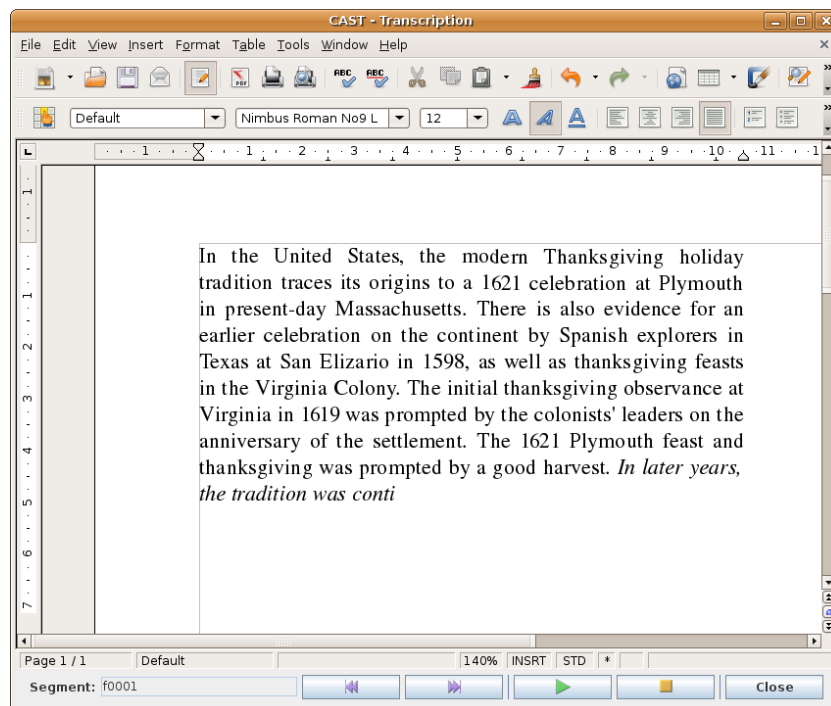


Figure 4: Screenshot of a transcription session. This shows the process of transcribing one audio segment. In this figure, all the text but the last incomplete sentence has already been transcribed and validated. The last partial sentence, shown in italics, is being produced by the ASR system while the transcriber listen to the audio. As soon as an error is detected the user momentarily interrupts the process to correct the mistake. Then, the system will continue transcribing the audio according to the new user feedback (prefix).

References

- ISO/IEC 9126-4. 2004. Software engineering — Product quality — Part 4: Quality in use metrics.
- F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, USA.
- Luis Rodríguez, Francisco Casacuberta, and Enrique Vidal. 2007. Computer assisted transcription of speech. In *Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I, IbPRIA '07*, pages 241–248, Berlin, Heidelberg. Springer-Verlag.
- Luis Rodríguez, Ismael García-Varea, and Enrique Vidal. 2010. Multi-modal computer assisted speech transcription. In *Proceedings of the 12th International Conference on Multimodal Interfaces and the 7th International Workshop on Machine Learning for Multimodal Interaction, ICMI-MLMI*.
- A.H. Toselli, E. Vidal, and F. Casacuberta. 2011. *Multimodal Interactive Pattern Recognition and Applications*. Springer.