

A System for Dynamic 3D Visualisation of Speech Recognition Paths

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

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

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

Bo Zhang
Dept. of Engineering
The University of Waikato
Hamilton, New Zealand
bz48@waikato.ac.nz

ABSTRACT

This paper presents an interactive visualisation system that assists users of semi-automatic speech transcription systems to assess alternative recognition results in real time and provide feedback to the speech recognition back-end in an intuitive manner. This prototype uses the OpenGL libraries to implement an animated 3D visual representation of alternative recognition results generated by the Sphinx automatic speech recognition system. It is expected that displaying alternatives dynamically will facilitate early detection of recognition errors and encourage user interaction, which in turn can be used to improve future recognition performance.

Categories and Subject Descriptors

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, Interactive visualisation, Animated interfaces, Error correction

1. BACKGROUND

Automatic speech recognition (ASR) technology has progressed remarkably in the last decades, evolving from small-vocabulary research prototypes into commercial systems,

*This work was supported by a Science Foundation Ireland *Research Frontiers* grant.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AVI'08 28–30 May, Naples, Italy

Copyright 2008 ACM 1-978-60558-141-5...\$5.00.

with applications that range from domain-specific dialogue systems to unconstrained dictation. However, despite being clearly workable in a variety of applications, ASR remains an imperfect technology for which error correction mechanisms need to be carefully designed [1, 5].

The issue of error correction has been extensively studied in the area of spoken language dialogue systems where recognition rates and user acceptance of an imperfect input modality can be improved through clever interaction design and exploitation of domain constraints [2]. In applications such as dictation systems, for which domain-specific constraints will not readily come to the rescue of system designers, user-specific factors can sometimes be brought to bear. Dictation systems usually incorporate on-line training functionality which allows the system to adapt to the user's voice, thereby improving recognition rates above those attained by the baseline system.

When domain and user constraints fail, however, error correction will typically need to be done through an input modality other than speech [5]. This is often the case of speech transcription applications, where domains are unconstrained and speakers vary greatly in voice and accent. Error-correction in such applications has been dealt with by presenting the user with a linear transcript and allowing words to be highlighted, deleted, inserted or modified directly. In this scenario, the role of the ASR module ends once the initial, imperfect transcript has been produced. The user-corrected transcript then becomes the final product of the transcription process. More recently, applications have been proposed which extend the role of the recogniser by allowing it to effect global changes to the transcript as a result of local user feedback [3, 4]. However, these applications still employ a linear text document metaphor to mediate error correction and user feedback.

We have developed a prototype, called DYTRAED (DY-namic TRAnscript Editor), which uses the word lattice produced by the recogniser in order to propagate local error correction through the transcript, but also employs an animated 3D representation of the sentence being transcribed, showing alternative recognition paths as they unfold.

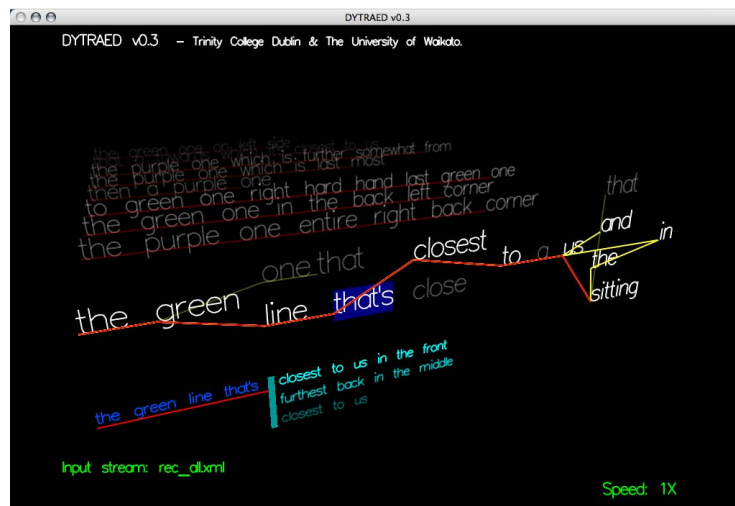


Figure 1: The user interface of DYTRAED

2. THE SYSTEM

The user interface of DYTRAED is shown in Figure 1. The user initially selects an audio source (live or recorded speech) and starts the transcription process. As recognition alternatives are generated by the ASR engine, the words are displayed as edges of a directed graph laid out on the middle-foreground of the application window. The partial recognition alternatives assigned the greatest scores by the ASR decoder are highlighted and connected through red-coloured edges. Lower-score hypotheses are dimmed, and alternatives undergoing active search (the rightmost words on the graph) are highlighted and connected to the choice point through yellow-coloured edges.

The user can pause the animation, increase or decrease its speed, and interact with the transcription graph. If the user clicks on a word (node) the animation stops and completion alternatives based on the buffered word lattice are presented. The user can then either select an alternative, thus accepting the entire sentence and bypassing the remainder of the visualisation for the current utterance, or simply ignore all options and continue to visualise the recogniser’s preferred paths. This form of user input is illustrated in Figure 1, where the words next to the vertical bar near the bottom of the screen are possible sentence completions ordered by the likelihood assigned to them by the ASR engine.

As the recognition process ends for a given sentence (either through user selection or due to the system reaching the end of the search on the lattice) sentences move to the background. Old sentences are slowly pushed towards the horizon by newly finished sentences until they disappear completely from view. This form of presentation serves to maintain a degree of context of the transcription task visually available to the user without hindering the necessary focus on the current sentence.

DYTRAED uses Sphinx-4¹ as its speech recognition backend, and OpenGL for 3D rendering and animation. Sphinx encodes hypotheses actively under consideration by the recog-

¹<http://cmusphinx.sf.net/>

niser (i.e. recognition paths that have not been pruned out) as a “word lattice” object containing acoustic scores (from the Hidden Markov Model) and language scores (in the present case, from a 3-gram language model). Our system displays “snapshots” of such structures and provides feedback to the search process through user interaction by, for instance, forcing the recogniser to select a lower-score path that would normally have been pruned out.

3. CONCLUSION AND FUTURE WORK

Informal evaluation suggests that our prototype can potentially increase user performance in ASR-assisted transcription tasks as well as making the experience more enjoyable. User feedback at the moment is restricted to the hypothesis already under consideration. We are currently working on mechanisms to incorporate new hypotheses (e.g. alternative segmentation, out-of-vocabulary words) to the system, along the lines of what has been done in [4].

4. REFERENCES

- [1] W. A. Ainsworth and S. R. Pratt. Feedback strategies for error correction in speech recognition systems. *International Journal of Man-Machine Studies*, 36(6):833–842, June 1992.
- [2] K. S. Hone and C. Baber. Modelling the effects of constraint upon speech-based human-computer interaction. *Interface Journal of Human-Computer Studies*, 50(1):85–107, 1999.
- [3] P. Liu and F. K. Soong. Word graph based speech recognition error correction by handwriting input. In *Procs. of the 8th Intl. Conference on Multimodal Interfaces*, pages 339–346. ACM Press, 2006.
- [4] M. Masoodian, B. Rogers, and S. Luz. Improving automatic speech transcription for multimedia content. In P. Isaías and M. B. Nunes, editors, *Proceedings of WWW/Internet '07*, pages 145–152, Vila Real, 2007.
- [5] B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM Trans. Comput.-Hum. Interact.*, 8(1):60–98, 2001.