Speech Recognition and Scholarly Research: Usability and Sustainability

Roeland Ordelman and Arjan van Hessen

Netherlands Institute for Sound and Vision and University of Twente, The Netherlands

Objectives

In spite of significant efforts and progress, automatic speech recognition (ASR) is not yet a tool that scholars can easily deploy in their research. In our project we focus explicitly on those aspects that are crucial for scholars with respect to actually deploying the technology:

- Usability of the technology: working with an ASR engine and using its output for research
- Sustainability: the longer-term availability of the technology for scholars, with state-of-the-art performance, maintained, updated, and accessible.

Introduction

In the past decades, we have been working with humanities scholars, especially Oral Historians, on topics related to ASR in a variety of projects:

- **CHoral** (2006-2011)
- **Verteld Verleden** (2010-2012)
- **Oral History Today** (2013-2014)
- **CLARIAH** (2015-2018)
- **CLARIN** (2016-2018)

These projects led to a better understanding and collaboration between humanities scholars and ICT-researchers and developers. Moreover, the projects provided a better insight into the variety of uses of digital collections, how a specific tool such as ASR could play a role here, and finally also, how ASR should be provided as a tool in a research infrastructure. In co-development with scholars, the Dutch CLARIAH project is now implementing all the parts of the ASR workflow in an iterative manner, including user testing and strategies for longer-term sustainability.

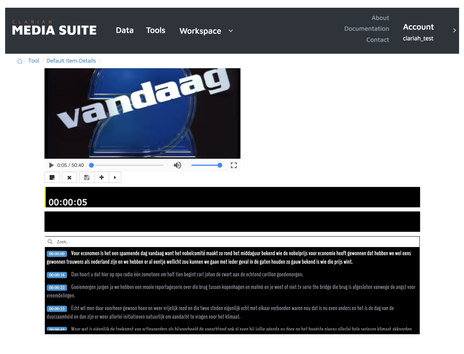

Figure 1: Within-document browsing

Requirements

Key requirements of scholars with respect to usability of ASR in scholarly research are:

1. Quality for annotation
   - Approximately correct transcriptions
   - Facilitate error correction
   - Fit into the annotation workflow
2. Quality for discovery
   - Increase the chances for discovery
   - Support recall (e.g., OOV solving)
3. ASR as a service
   - Easy accessible (low tech)
   - Adaptable (e.g., vocabulary, models)
   - Secure (privacy & IPR)
4. Large scale ASR
   - Access to large collections for bulk processing
   - Dedicated machinery and management (computer clusters)
   - Keep track of provenance
5. Using the output of ASR
   - Indexing (standard search)
   - Word cloud generation
   - Named entity extraction
   - Topic modeling

CLARIAH Media Suite / CLARIN OH portal

KALDI_NL is running as a sustainable service in the **CLARIAH** infrastructure at one of the CLARIAH Centers. Until now, we have been processing $350K$ hours of audiovisual content via a High Performance Computing data infrastructure, searchable via `mediasuite.clariah.nl`. KALD_NL is also available via the **CLARIN** Oral History Portal accessible via `www.phonetik.uni-muenchen.de/apps/oh-portal`.

ASR Engine

We are using the open-source KALDI [1] speech recognition toolkit that supports deep neural nets, together with the LIUM speech diarization toolkit. Dutch models have been developed at University of Twente using the Spoken Dutch Corpus and a large corpus of text data from a variety of sources. The resulting open-source "Kaldi_NL" instance (`www.opensource-spraakherkenning.nl`), has a lexicon of around $250K$ words and with NNET3-TDNN-LSTM models. Its performance is state-of-the-art with around 10% WER (< 0.5xRT) tested on the NBest BN-NL benchmark set [2].

Implementation

ASR is implemented in the CLARIAH infrastructure given the requirements of scholars discussed above. KALDI_NL is running as a service in a local cloud at the Netherlands Institute for Sound and Vision (NISV), one of the CLARIAH Centers, for individual scholars and medium size collections, accessible for scholars via the closed environment with federated authentication of the CLARIAH Media Suite. For bulk processing we installed a KALDI_NL instance on the High Performance Computing data infrastructure for science and industry, SURFSara. For sustainability and support on acoustic models, language models and updates we work together with speech groups at University of Twente and Radboud University Nijmegen.

Results

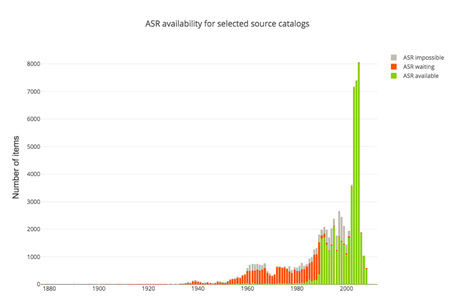

Figure 2: Processing status ASR at NISV

Until now, we have processed $350K$ hours of audiovisual content via the High Performance Computing cluster, approximately one third of the catalogue of the Netherlands Institute for Sound and Vision. Figure 1 shows the number of processed items (green) and items in the cue (red) for the NISV catalogue. We have indexed the transcripts and made them searchable via the Media Suite, keeping the time information to allow also for within-document-search. See also Figure 1, a screen-shot of the document viewer of the Media Suite.

Conclusion

We made significant progress in making ASR available for scholarly research by focusing on the lessons learned in previous projects and taking the requirements of scholars into account in the development process itself, but also in developing *governance* strategies for sustainability. In spite of the progress, there is still work lying ahead, both on the technical level and governance level. On the technical level, one of the most important topics is to enable (dynamic) adaptations in vocabulary, for example for scholars working with collections in specific domains or time periods. On the governance level, we need a more elaborate business case for maintaining ASR instances on computer clusters, defining the role of the various stakeholders (computer science research, heritage institutions, humanities clusters, possibly also commercial providers).

References

[1] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely.
The kaldi speech recognition toolkit.
In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011.
IEEE Catalog No.: CFP11SRW-USB.

[2] Judith Kessens and David A van Leeuwen.
N-best: The northern-and southern-dutch benchmark evaluation of speech recognition technology.
In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

Acknowledgements

This work is funded by CLARIAH.

Contact Information

- clariah.nl
- rordelman@beeldengeluid.nl
- a.j.vanhessen@utwente.nl