

# Automatic speech recognition in media asset management

Johanna Björklund, Emil Lundh

Dept. Computing Science, CodeMill AB  
Umeå University, Riddaregatan 8,  
901 87 Umeå, 90336 Umeå  
johanna@cs.umu.se, emil.lundh@codemill.se

## Abstract

This contribution describes on-going development work on a media asset management (MAM) system, capable of automatically transcribing ingested video and audio. The MAM system invokes LIUM for segmentation, followed by the ASR toolkit CMU Sphinx for transcription. We list three practical aspects of ASR that we find particularly challenging, with the hope of attracting interest in these from the national language technology community.

## 1. Background

Media asset management (MAM) systems are specialized databases for media assets, e.g. video, audio, and image files. They are commonly used by broadcasters to support collaborative workflows and to automate tasks such as cataloging, transcoding, distributing, and retrieving media. To facilitate search, the assets are annotated with descriptive metadata at the time of ingestion into the system. Modern MAM systems are expected to handle large quantities of data. The BBC digital archive alone contains material dating back to 1890, and offers 1 million hours of playable material (Kiss, 2010). By 2018, global IP traffic is predicted to exceed 1.6 zettabytes per year, out of which 79 per cent will be video (CVNI, 2014).

Due to the increasing size of the data sets, there is a shift from manual to automated metadata annotation. For some combinations of asset and label, the transition is straightforward. For instance, the time and location a video was shot can typically be read from the camera, and there are efficient algorithms to detect scene changes and assess image quality. Then there is information that is harder to get at, and in this category, an accurate transcription is perhaps the most important. Not only is a transcription valuable in itself, as it allows free-text search, but it is extremely helpful in subtitling and other post processing steps. We may for instance want to summarize a political debate, tag segments with over-arching themes, or do named-entity extraction to generate appropriate second-screen content. For all these purposes, it helps to start from a high-quality text.

The authors are currently involved in the development of a commercial MAM system that, among other things, automatically transcribes speech in video. The approach consists of (i) a preprocessing step, in which the input audio content is segmented into sentences using the open-source tool LIUM Speaker Diarization (Meignier and Merlin, 2010), (ii) the automatic transcription with CMU Pocketsphinx 0.8 (Lee et al., 1990), and finally (iii), a manual post-processing step to correct mistakes and add descriptive tags. A screenshot of an early version of the post-processing tool is shown in Figure 1. The user edits the transcription in real time, and the result is stored as time-stamped metadata.

The proposed MAM system is not the first with ASR capabilities. Due to the complexity of the task, this functionality is almost always provided by a third party different from the MAM vendor. When a new video is entered into the MAM system, the audio track is uploaded to some cloud-based ASR system for transcription, BBC's audio-analysis platform Comma and Vocapia's SaaS service VoxSigma being two notable examples. This kind of distributed architecture is tractable because the audio track is typically small compared to the entire video file, so the network usage need not be overwhelming. However, many customers have privacy policies that prevent data from leaving their premises. On-site installations are therefore also relevant, and it is towards this latter category that we aim.

## 2. Challenging aspects

Our customers are typically interested in ASR either because they want to reduce the cost of producing subtitles for new material, or because they want to make banks of existing material searchable. In both cases, they expect precision to be near perfect before they are willing to make the investment, so quality is of the highest concern. In the following, we list some aspects that we find particularly challenging.

### 2.1 Corrective feedback

Even the best of ASR systems will sometimes make mistakes. This can either be due to missing entries in the dictionary, which can never be complete for human language, or due to the language and acoustic models. A somewhat amusing example of the first kind is when CMU Sphinx, with the default dictionary, thinks that Samuel L. Jackson is saying *how I would send an enema*, rather than *Hollywood cinema*. Mistakes like these are discovered and corrected during in the post-processing step, by manually rewriting the transcription. Once the ASR system has 'misheard' a word, the  $n$ -gram model tends to put it off track until the start of a new sentence. We would therefore like to see an ASR system that would start from a partial, user-provided, transcription, and take this as certain in all probability calculations. Such a system could, if it was fast enough, be invoked every time the user corrected an error, and would hopefully discover the subsequent errors by itself.

## Transcript

All changes saved (14:04)

---

0      this administration also puts for a false choice between the liberties we cherish the mystery we provide

i'll provide our intelligence and want horsemen agencies with tools they need to track him take out the terrorists without undermining our car

the fusion alfred that means no more illegal wire tapping of americans that is no more national security letters to spy

alan citizens were not suspected of crime

no more tracking citizens who do nothing more than protest in misguided war no more ignoring the law when i'm inconvenient


34,44      that is not we are not what is necessary to defeat the terrorists

the wiser court works the separation of powers works are constitution works we will against

example for the world that the law does not subject to the winds of sovereign rulers in the justice is not arbitrary

this administration acts like filing civil liberties is the way to enhance our security it is not there are no short cuts to protect

in america



### Metadata

Video Title:	Obama Speech 2007
Transcript Language:	English
Use for indexing:	<input type="checkbox"/>
Progress:	Not started ▾

Figure 1: The transcription tool for a MAM system under development.

## 2.2 Structured language models

Another way to improve accuracy would be to use a more structured language model than  $n$ -grams, one that also took into account the grammatical structure of what was being said. The advantage would be that the system could recover from an error sooner, perhaps at the beginning of the next phrase or clause. It is clear that the language model in question would have to be extremely robust to cope with noise and uncertainty in the input. For this reason, formalisms such as probabilistic phrase-structure grammars are probably not appropriate. There are however more recent alternatives that may be worth investigating, for example probabilistic lexicalized tree-insertion grammars (Schabes and Waters, 1995; Hwa, 1998).

We believe that grammatically correct sentences, with appropriate punctuation and capitalization, would improve the user experience, even if the total number of correct words is unchanged. It would also be interesting to know if quality could be improved by first finding the verbs of a sentence, and then growing clauses around them. About a seventh of the approximately 170 000 words in current use in the second edition of the Oxford English Dictionary are verbs, so one could consider using CMU Sphinx to mark these out, and then use Kaldi speech recognition toolkit, which is supposedly good for large-vocabulary ASR (Povey et al., 2011), to fill in the blanks between.

## 2.3 Automatic quality assessment

The final challenge that we want to mention here is that of automatically detecting poor transcription. There may not be time to manually check the entire transcription, so it would be nice if the system notified the user when the likelihood for a segment is particularly low. The CMU Sphinx

does assign a score to its transcriptions, but it appears that this score is largely determined by the length of the transcribed sentence. In addition, the next-best sentences offered are rarely an improvement. Automatic assessment is a difficult problem, but if the same unlikely word sequence appears several times in a transcription, then there is reason to believe that there is a word missing in the dictionary. Compare for example the transcribed *want horsemen agencies* with the actual and more likely *law enforcement agencies* in a speech by Barack Obama from 2007 (see Fig. 1).

## Acknowledgements

We are thankful to Joel Långström, Christina Svensson, and Urban Söderberg at CodeMill AB for their work on the implementation, and to Raja Kuppuswamy at Last-mile Consulting for his helpful advice. We also acknowledge the support of the EU FP7 project MICO.

## References

- CVNI. 2014. The zettabyte era - trends and analysis. Technical report, Cisco Systems, Inc.
- R. Hwa. 1998. An empirical evaluation of probabilistic lexicalized tree insertion grammars. In *Proc. ACL 1998*, pages 557–563, Stroudsburg, PA, USA. Association for Computational Linguistics.
- J. Kiss. 2010. In the BBC archive. In *Tech Weekly*. Guardian News & Media Ltd, August.
- K.-F. Lee, H.-W. Hon, and R. Reddy. 1990. An overview of the Sphinx speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(1):35 – 45, January.
- S. Meignier and T. Merlin. 2010. LIUM spkdiarization:

- An open source toolkit for diarization. In *Proc. CMU SPUD Workshop*, Dallas, Texas.
- D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.
- Y. Schabes and R. C. Waters. 1995. Tree insertion grammar: Cubic-time, parsable formalism that lexicalizes context-free grammar without changing the trees produced. *Computational Linguistics*, 21(4):479–513.