Media Monitoring System for Latvian Radio and TV Broadcasts

Artūrs Znotiņš[1], *Kaspars Polis*[2], *Roberts Darģis*[1]

[1] Institute of Mathematics and Computer Science, University of Latvia
[2] LETA, Latvia

arturs.znotins@lumii.lv, kaspars.polis@leta.lv, roberts.dargis@lumii.lv

Abstract

Media monitoring allows to capture media exposure of people, organizations and other important topics. This paper presents a media monitoring system for Latvian radio and television broadcasts. This system uses an automatic speech recognition (ASR) module to convert audio and video files to text and to extract keywords of interest. The system has been developed in close cooperation with Latvian information agency LETA.

Index Terms: media monitoring, multimedia, keyword spotting, automatic speech recognition

1. Introduction

Media monitoring is an important tool to identify mentions of organizations, brands and people in media content. It can be used to get actual information about events relevant to the organization, brand reputation, competitors, etc. Manual audio and video monitoring is highly labor intensive task and it is limited due to the large amount of data which is produced every day.

In this paper we describe current efforts in creating Latvian radio and television broadcast monitoring system that is possible due to recent advancements in Latvian automatic speech recognition. [1,2] The system uses a trained large vocabulary speech recognition (LVCSR) system. This system transforms audio into text output and word lattices that are used to find keywords of interest. A schematic overview of the proposed system is shown in figure 1. The output of the system is a transcribed text with marked keywords and precise word positions in an audio file as shown in figure 2.

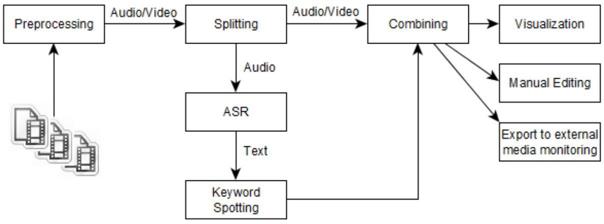
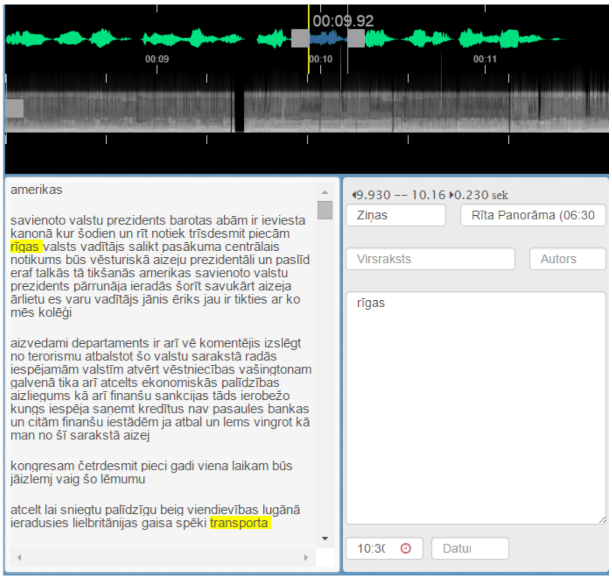


Figure 1: *A schematic diagram of audio/video monitoring system.*



amerikas

savienoto valstu prezidents barotas abām ir ieviesta kanonā kur šodien un rīt notiek trīsdesmit piecām rīgas valsts vadītājs salikt pasākuma centrālais notikums būs vēsturiskā aizeju prezidentāli un paslīd eraf talkās tā tikšanās amerikas savienoto valstu prezidents pārrunāja ieradās šorīt savukārt aizeja ārlietu es varu vadītājs jānis ēriks jau ir tikties ar ko mēs kolēģi

aizvedami departaments ir arī vē komentējis izslēgt no terorismu atbalstot šo valstu sarakstā radās iespējamām valstīm atvērt vēstniecības vašingtonam galvenā tika arī atcelts ekonomiskās palīdzības aizliegums kā arī finanšu sankcijas tāds ierobežo kungs iespēja saņemt kredītus nav pasaules bankas un citām finanšu iestādēm ja atbal un lems vingrot kā man no šī sarakstā aizej

kongresam četrdesmit pieci gadi viena laikam būs jāizlemj vaig šo lēmumu

atcelt lai sniegtu palīdzīgu beig viendievības lugānā ieradusies lielbritānijas gaisa spēki transporta

◂9.930 -- 10.16 ▸0.230 sek

Ziņas

Rīta Panorāma (06:30

rīgas

10:3(

Figure 2: *The user interface of the system, containing audio waveform, spectrogram, ASR output and highlighted keywords.*

2. Preprocessing and Audio Splitting

In beginning the system converts all audio and video contents to MSWAV format that is used for ASR.

We use C4.5 decision tree classifier to segment audio into speech and non-speech fragments (music, loud noises, advertisements). More fine-grained classes can be further used to provide meaningful information to the operator and to adapt ASR for different signal types.

In order to process long audio files in a reasonable time, the system divides audio into segments of up to 1 minute splitting on pauses of significant duration (0.3 seconds) and taking into account the classified audio speech and non-speech segments. These fragments are then distributed among multiple instances of ASR system.

3. Automatic Speech Recognition

ASR system consists of two main parts: an acoustic model (AM) and a language model (LM). Used ASR achieves 51% Word Error Rate (WER).

Acoustic model for ASR was trained on 100 hours of audio data from Latvian Speech Recognition Corpus (LSRC) that was designed to represent the major speech characteristics of Latvian population [1]. The AM was trained using CMU Sphinx toolkit [3]. The developed AM is a context-depended continuous triphone Hidden Markov Model (HMM) with 4,000 tied states each described by 8 Gaussian mixture components. It contains 57 phoneme models, a silence model and 6 different noise models.

Trigram language model was interpolated from LMs trained on the LSRC transcriptions and a large corpus of LETA newswire containing Latvian interviews and news. The LM was trained using SRILM toolkit [4]. We have experimented with factored LMs utilizing information about word forms, word lemmas, tags and classes, however it did not improve the performance of the system significantly.

The ASR system uses an open vocabulary containing almost 600,000 word-forms that were extracted from a large set of high quality Latvian news articles. This vocabulary is extended using keyword vocabulary that is continuously enriched with new keywords. Phonetic transcriptions were generated by a rule based system that uses approximately 250 expert defined rules and 1300 exceptions.

Obtained transcriptions are post processed using sentence border classification and truecasing.

Output of the ASR system consists of words together with time intervals when they were spoken.

4. Keyword Spotting

The keyword spotting module searches for keywords in ASR text output and word lattices that allows to tune precision and recall of the system. The system achieves 81% F1 score (90% precision and 73% recall).

Currently keyword list contains approximately 3,500 unique expressions (also multiword) and approximately 16,000 inflected expressions.

New keywords can be added to the keyword vocabulary that extends the main ASR vocabulary. This is done using user interface that automatically generates pronunciations and allows to add more aliases as well as pronunciation variants.

Currently out-of-vocabulary (OOV) word modeling is the main problem for the keyword spotting task. Experiments showed that class based and factored LMs did not help significantly with OOV words that have not seen during training of LM.

5. Integrated System

The presented system is an integrated audio/video monitoring system. It automatically downloads audio and video files from servers that record radio and television broadcasts. Then the system preprocesses and splits these files for ASR on multiple server instances. Recognized audio fragments are merged from divided fragments into a single resulting audio. In the end the system exports processed files to a separate existing media monitoring system that allows to monitor media based on preferences of clients, keyword lists, date and time. This system also provides user interface to manually edit transcribed text, edit metadata (title, date, time, source, broadcast name).

6. Discussion

Although the performance of the speech recognition system is not high, it provides satisfactory results for the keyword spotting task.

The system is already deployed however it is continuously improved. Current efforts includes:

- using proxy keywords for OOV word modeling;
- adapt ASR for telephone speech and noisy speech.

Currently the proposed system works only for Latvian. By providing trained statistical models for Sphinx4 and SRILM and keyword lists, the presented media monitoring system could be used also for other languages.

7. Acknowledgments

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Center” (http://www.itkc.lv/) of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.10 “Exploring the potential of automated speech recognition for Latvian media monitoring”.

8. References

[1] M. Pinnis, I. Auziņa and K. Goba, “Designing the Latvian Speech Recognition Corpus,” in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC’14)*, Reykjavik, Iceland, 2014.

[2] R. Darģis and A. Znotiņš, “Baseline for Keyword Spotting in Latvian Broadcast Speech,” in *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014*, vol. 268, pp. 75–82, September 2014.

[3] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. A. Siegler, R. M. Stern and E. Thayer, “The 1997 CMU Sphinx-3 English Broadcast News Transcription System,” in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[4] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit,” in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, 2002.