INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

System for speech transcription and post-editing in Microsoft Word

Askars Salimbajevs[1,2], *Indra Ikauniece*[1]

[1]Tilde, Vienibas gatve 75A, Riga, Latvia
[2]University of Latvia, Raina bulv. 19, Riga, Latvia
askars.salimbajevs@tilde.lv, indra.ikauniece@tilde.lv

Abstract

In this demonstration paper, we introduce a transcription service that can be used for transcription of different meetings, sessions etc. The service performs speaker diarization, automatic speech recognition, punctuation restoration and produces human-readable transcripts as special Microsoft Word documents that have audio and word alignments embedded. Thereby, a widely-used word processor is transformed into a transcription post-editing tool. Currently, Latvian and Lithuanian languages are supported, but other languages can be easily added.

Index Terms: speech recognition, transcription tool, Microsoft Word.

1. Introduction

Many everyday life scenarios require speech transcription, e.g. parliament meetings, court hearings, interviews, business meetings etc. Typically, these transcriptions are done manually, which is time-consuming and tedious work. Automatic speech recognition (ASR) can be very useful for these cases because it can reduce the work of a human to only correcting recognition mistakes, adding punctuation and formatting. The ASR function can be complemented with automatic speaker diarization and punctuation restoration to further reduce the work.

Although there exist many alternatives, such as specific audio transcription editors, Microsoft Word is a de-facto standard tool for preparing many types of documents. It's a familiar tool for millions of people. Moreover, in many scenarios, the final transcription document will be a Microsoft Word document. So, it is only convenient and natural if the whole transcription process can be done in one tool.

We present a transcription system that performs speaker diarization, speech recognition and punctuation restoration and then produces a special macro-enabled Word document with embedded audio that provides a convenient and familiar editing environment. Currently, only Latvian and Lithuanian languages are supported, however we also plan to add the English language for the demonstration.

2. Transcription Web-Service

2.1. Overview

The transcription service is based on the Latvian Speech-To-Text transcription service [1], which in turn is a fork of Alumae Full-duplex Speech-to-text System [2].

It consists of a single master server and multiple workers, which can be used independently. For example, workers can be hosted on different machines in different data centres. This makes it possible to easily scale up the system by just adding more servers and hardware. The service can be used from desktop, mobile or web applications.

The master server is responsible for providing the API, receiving files from users, converting them to a uniform format and distributing jobs to workers. The master server is also responsible for converting an ASR transcript into a special Word document with embedded audio. The API will accept a wide variety of multimedia file formats and codecs.

Workers connect to the master server, receive jobs and perform the actual transcription. Each worker processes only one job at a time. Each job consists of several stages:

- Segmentation and speaker diarization
- Speech recognition and lattice rescoring
- Converting end result to SRT format
- Punctuation restoration

Speaker diarization is performed by the LIUM SpkDiarization tool [3]. Punctuation restoration is performed by the bidirectional LSTM neural network[4]. The method is language-independent, and support for new languages can be added easily by training a new model on a monolingual text corpus. While both diarization and punctuation restoration are not perfect and can produce a lot of errors, the readability of the document is greatly improved, allowing users to edit transcription more efficiently.

2.2. Speech recognition

Our web-service provides Lithuanian and Latvian speech recognition, which is based on an open-source Kaldi toolkit [5]. There are no specific language dependencies, so it is possible to add Kaldi-based systems for any other language.

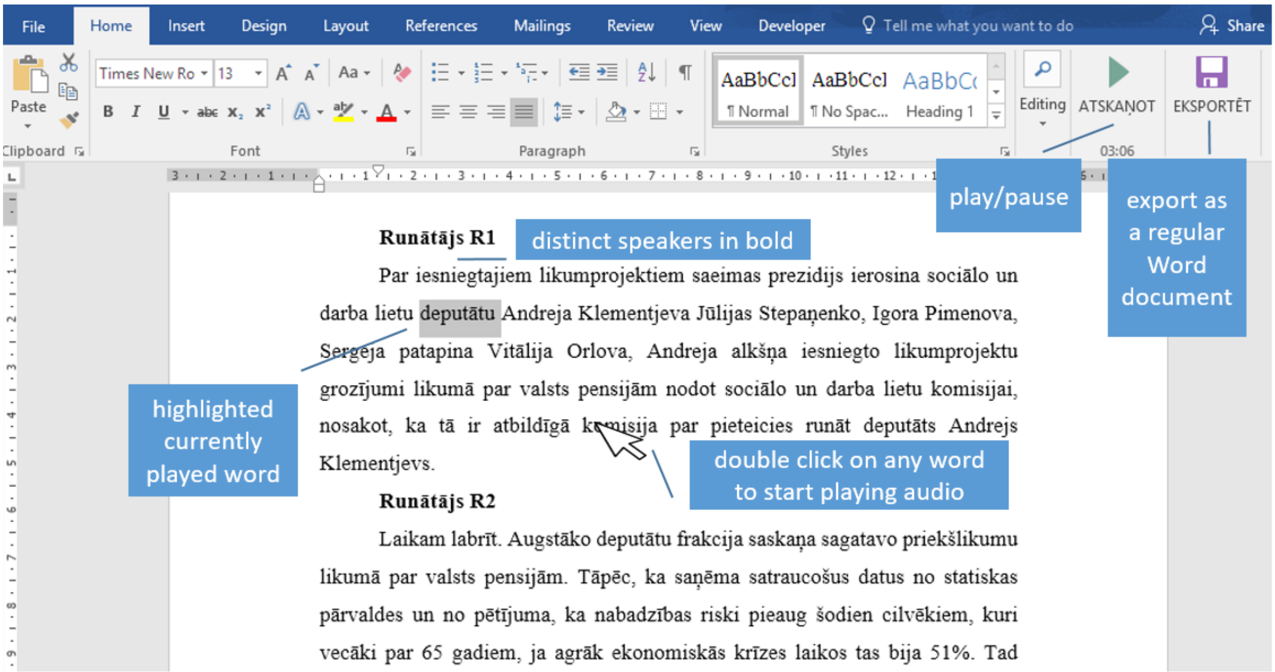
The recipes for both Latvian and Lithuanian systems are very similar:

- HMM-DNN acoustic model with grapheme-based pronunciation model and iVectors for speaker adaptation.
- N-gram language model with vocabulary of about 800,000 word forms.

The Lithuanian ASR is trained on a 248h speech corpus, which contains 195h of automatically annotated Lithuanian Parliament (Seimas) recordings from 2015-2016 and speech recordings from the LIEPA project[1]. The Latvian ASR is trained on a 286h speech corpus, which contains a 100h Latvian Speech Recognition Corpus (LSRC) [6] and 186h of automatically annotated Latvian Parliament (Saeima) recordings from 2011-2014.

The Word error rate (WER) on the general domain test set for the Latvian ASR is 17% and 24% for the Lithuanian ASR. A much lower WER can be achieved if the system is adapted for some specific domain, for example, the Latvian ASR with language model adapted for parliament session transcription achieves 7% WER on a corresponding in-domain test set.

[1] LIEPA (Services Controlled by Lithuanian Voice). https://www.xn–ratija-ckb.lt/liepa



Runātājs R1

Par iesniegtajiem likumprojektiem saeimas prezidijs ierosina sociālo un darba lietu deputātu Andreja Klementjeva Jūlijas Stepaņenko, Igora Pimenova, Sergeja patapina Vitālija Orlova, Andreja alkšņa iesniegto likumprojektu grozījumi likumā par valsts pensijām nodot sociālo un darba lietu komisijai, nosakot, ka tā ir atbildīgā komisija par pieteicies runāt deputāts Andrejs Klementjevs.

Runātājs R2

Laikam labrīt. Augstāko deputātu frakcija saskaņa sagatavo priekšlikumu likumā par valsts pensijām. Tāpēc, ka saņēma satraucošus datus no statiskas pārvaldes un no pētījuma, ka nabadzības riski pieaug šodien cilvēkiem, kuri vecāki par 65 gadiem, ja agrāk ekonomiskās krīzes laikos tas bija 51%. Tad

play/pause

export as a regular Word document

distinct speakers in bold

highlighted currently played word

double click on any word to start playing audio

Figure 1: *Editing transcriptions in Microsoft Word.*

2.3. User application

The user application is a single web page (but it can be anything: mobile app, CRON script that scans some directory etc.). The page enables the user to select a file, enter an email address and submit this data to the transcription service. Then the user can either close the page and wait for e-mail notification or stay and wait until the file is processed. Typically, a one hour file is processed in 30 minutes. After processing, the page displays a link to the Word document that contains the transcription. The same link is also sent by e-mail.

3. Transcription editing

The transcription is saved as a special macro-enabled Word document that allows to play the embedded audio from any place in the document and highlights the words as they are being played (see Figure 1).

The document's text is divided into speakers (with speaker titles in bold) and speech (in formatted paragraphs with punctuation and formatting).

There are two custom buttons on the Word ribbon: the ”play/pause” button plays and pauses the embedded audio, and the ”export” button saves the document as a regular Word document without audio and macros. It is also possible to start playing the audio from any word by double clicking on it. There are also keyboard shortcuts for pausing and resuming the playback.

While the audio is being played, each current word is highlighted. If the user sees or hears a mistake, he can pause the playback by using the ”pause” button or by clicking on the first incorrect word. He can then edit the erroneous segment as he would do in a regular Word document, and it will not break the audio and word alignment. The playback then can be resumed.

Thus, the speech transcription post-editing can be done efficiently by using only Microsoft Word, which is a familiar tool for many professional and non-professional users.

4. Acknowledgements

The research leading to these results has received funding from the research project ”Competence Centre of Information and Communication Technologies” of EU Structural funds, contract No. 1.2.1.1/16/A/007 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.3 “Speech recognition and synthesis for professional application”.

5. References

[1] A. Salimbajevs and J. Strigins, “Latvian speech-to-text transcription service,” in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 722–723.

[2] T. Alumäe, “Full-duplex speech-to-text system for Estonian,” Kaunas, Lihtuania, 2014.

[3] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, “An open-source state-of-the-art toolbox for broadcast news diarization,” in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'2013)*, Aug. 2013.

[4] A. Salimbajevs, “Bidirectional LSTM for automatic punctuation restoration,” in *Human Language Technologies - The Baltic Perspective - Proceedings of the Seventh International Conference Baltic HLT 2016, Riga, Latvia, October 6-7, 2016*, 2016, pp. 59–65.

[5] 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, “The Kaldi Speech Recognition Toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[6] M. Pinnis, I. Auzina, and K. Goba, “Designing the Latvian Speech Recognition Corpus,” in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*, 2014, pp. 1547–1553.