Latvian Speech-To-Text Transcription Service

Askars Salimbajevs 1, 2, Jevgenijs Strigins 1, 2

1 Tilde, 75a Vienibas gatve, LV-1004, Riga, Latvia
2 University of Latvia, 19 Raina Blvd., LV-1586, Riga, Latvia

askars.salimbajevs@tilde.lv, jevgenijs.strigins@tilde.lv

Abstract

In this demonstration paper, we introduce the first publicly available Speech-To-Text transcription service for the Latvian language. We present its main features, the details of automatic speech recognition (ASR) system used in this service, software architecture, and an evaluation of recognition quality. The service will provide regular people with the opportunity to transcribe their own audio files for various purposes, such as lectures, meetings, etc. Also, the users will be given an opportunity to give their evaluation and feedback about the quality and usability of this service, which will be used by developers to make changes in the ASR in order to improve it.

Index Terms: speech recognition, large vocabulary, language model, web service

1. Introduction

Although there exist many services and software that transcribe speech to text for bigger languages, such as English, German, etc., there is no such service for the Latvian language. One explanation for this might be that Latvian is spoken by only about 1.5 million people, which becomes a reason why big companies, who actively research and make ASR applications available for the public, do not develop ASR systems for Latvian speech. It should also be noted that Latvian is an under-resourced language, and there is not much language and acoustic corpora available for Latvian. Another reason could be the fact that Latvian is a very different language, e.g., in comparison to English, so the existing acoustic and language models cannot just be retrained on Latvian data and provide good performance. The differences between Latvian and other "big" languages include:

- A very rich and complex nominal and verbal morphology
- The possibility of one word having dozens of surface forms
- No strict rules in accordance to which separate words can follow each other in a sentence

Because of these reasons, not only is there no publicly available service for transcribing Latvian speech, but there is practically no research done on this topic, as there are only a few published results on speech recognition for Latvian [1], [2], [3].

In spite of these difficulties, there is quite strong motivation to develop a good performing LVASR system for Latvian, as it would allow people to use it in various areas, such as business, education, security, etc., and help to develop them in this way.

The main purpose of our service is to enable regular users to automatically transcribe their speech recordings in Latvian, which was impossible before. The service accepts most of the popular audio formats and provides the transcription in a plain text or a subtitle format (i.e., text with time marks). This service will also allow to test the Latvian LVASR system in a real life scenario, evaluate how it performs, and gather user experience stories. To obtain the required information, users will be prompted to leave an evaluation and feedback about the quality of transcriptions and their experience. Last but not least, this service will give developers very useful data about what people actually tend to transcribe, which will help to develop more optimized speech recognition systems.

2. Latvian Speech Recognition System

The main part of our transcription service is a Latvian Large Vocabulary Automatic Speech Recognition system, which is based on an open-source Kaldi toolkit [4].

The acoustic model (AM) is trained on a 100 hour-long Latvian Speech Recognition Corpus [5]. The HMM-DNN modeling approach with iVector adaptation [6] is used together with PLP features and LDA transformation. The phoneme repository consists of 37 grapheme-based phonemes, 1 unified filler/silence model, and 1 model for fragmented words and other garbage.

We use n-gram language models (LM) that are trained on a 104M sentence and 1.5B word text corpus, which was collected by crawling Latvian web news portals. A vocabulary of 950K units is used, which are selected by their frequency in the training corpus.

The decoding process is two-pass, a 2-gram heavily pruned model is used during the first-pass, and a full not-pruned 3-gram model is used for rescoring lattices and obtaining the final decoding result.

3. Transcription Service

The transcription service is based on the Alumae Full-duplex Speech-to-text System for Estonian [7]. It consists of a web frontend, a single master server, and multiple workers, which can be used independently. For example, workers and master servers can be hosted on different machines. This makes it possible to easily scale up the system just by adding more servers and hardware. It is also possible to use the transcription service without the web frontend, i.e., create a desktop or mobile application that communicates directly with the master server. An overview of the service is presented in Figure 1.

The web frontend consists of two HTML/JS web pages. The first page enables the user to select a file, enter an email address, and submit this data to the transcription service. The other page is used for viewing and downloading transcriptions. It also contains a feedback form for evaluation of user experience.
The master server is responsible for receiving files from users, converting them to a uniform format, and maintaining a job queue (one job represents one file). The master server is also responsible for sending notification e-mails about the transcription process, e.g., to notify the user that his file is successfully processed, and the transcription is ready.

![Transcription service overview](image)

Workers connect to the master server, receive jobs, and perform the actual transcription. Each worker processes only 1 job at a time. Each job consists of several stages:
1. Segmentation and speaker diarization
2. PLP feature and iVector extraction
3. Decoding with 2-gram model
4. Lattice rescoring with 3-gram model
5. Converting end result to SRT format

The transcriptions are then sent back to the master server where they can be accessed by users through the special URL that is sent in a notification e-mail.

4. Evaluation

Before publishing the Latvian Speech-To-Text transcription service, we have performed several evaluations of speech recognition quality.

We have used two test sets: (1) about 1h of freely available lecture recordings found on the Internet and (2) a very small (23 min.) corpus of Latvian speech, which was obtained by recording various people reading articles from Web news portals.

The results of evaluation are summarized in Table 1. While very far from perfect, we believe that these results show that the transcription service has the potential for a real-word practical application.

### Table 1. Evaluation of Word error rate

<table>
<thead>
<tr>
<th>Test set</th>
<th>OOV</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectures</td>
<td>3%</td>
<td>20.71%</td>
</tr>
<tr>
<td>News</td>
<td>6%</td>
<td>19.63%</td>
</tr>
</tbody>
</table>

While word error rates (WER) of 19-21% can be considered as pretty high, not all errors are equal. Our analysis of errors shows that only 47% of misrecognized words in these test sets make utterances difficult or impossible to understand. E.g., 42% of errors are in word endings, and in most cases it is easy for a human reader to recover from such errors.

5. Conclusions

In this demonstration paper, we presented our Latvian Speech-To-Text transcription service, which is the first speech recognition service publicly available for the Latvian language. Transcription quality evaluation shows that it has potential for real-word use. The architecture of the service enables scalability, support of many different audio formats, and the ability to use different user interfaces. The transcription service is free and available at: http://tilde.lv/asr-demo.

6. Acknowledgements

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Centre” of EU Structural funds, contract nr. L-KC-11-0003 signed between the ICT Competence Centre (www.itkc.lv) and the Investment and Development Agency of Latvia, research No. 2.4 “Speech recognition technologies”.

7. References


