SPOKENDATA.COM

Igor Szöke and Josef Žižka

Brno University of Technology[1] and ReplayWell, s. r. o., Brno, Czech Republic
info@spokendata.com

1. INTRODUCTION

SpokenData.com is an online transcription service for any audiovisual data specified by the user. In brief, SpokenData works in four steps:

1. Signing in and adding a media file referencing a YouTube URL or by upload from user's local drive.
2. Selection of the recognizer (voice activity detector, Czech, English)
3. When the automatic transcription is ready, the user is notified by email.
4. The user can edit the transcription by him/herself or buy a professional transcription.

SpokenData is run by BUT Speech@FIT spin-off ReplayWell in tight cooperation with the research group.

The core idea of the service lies in providing users an automatic speech recognizer which automatically adapts on their data – both acoustic and language. As the user is provided with GUI to correct ASR errors, this information should be used by ASR to learn.

2. UNDER THE HOOD

The block scheme of SpokenData data flow is shown in Figure 1.

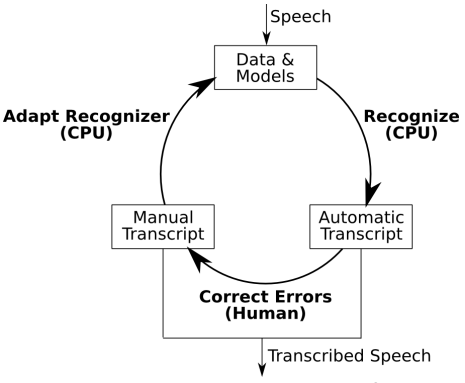

Figure 1: SpokenData data flow.

2.1. Voice activity detector

The voice activity detector is the core part of the system. Currently, we are using robust VAD trained on multilingual data and various channels (telephone, lecture and distant microphone). The detector is based on Kaldi toolkit and neural network classifier with two output classes (speech/non-speech) and it is a simplified version of BUT VAD developed for the RATS project [1].

2.2. Diarization

The output of VAD is processed by diarization to attribute segments to speakers. This leads to better accuracy of the speech recognizer. The estimated speaker information helps to unsupervised adaptation of the acoustic model. The principle used in spoken data lies in using an HMM instead of a simple mixture model when modeling generation of segments (or even frames) from speakers. HMM limits the probability of switching between speakers when changing frames, which makes it possible to use the model on frame-by-frame basis without any need to iterate between 1) clustering speech segments and 2) re-segmentation [2]. VAD is offered as a separate service and is a part of speech recognizers. Diarization is implemented and is now being tested. It will be deployed in the SpokenData.com service soon.

2.3. Speech recognizer

The recognition engine is based on BUT's production system. The system contains PLPs, neural network-based stacked bottleneck features, RDT, discriminative, and speaker adaptive training. The system is a simplified version of BUT BABEL ASR [3].

Currently, English (16 kHz) and Czech (8 kHz) are supported with works on Czech (16 kHz including distant microphones) and English (16kHz including distant microphones) in progress. We also plan to add Russian, and Levantine Arabic (in cooperation with Phonexia).

[1] *This work was partly supported by Technology Agency of the Czech Republic grant No. TA01011328.*

2.4. Adaptation

The system needs to be adapted to the target domain and acoustic channel. There are several levels of adaptations we are planning to use:

1. On-the-fly adaptation of acoustic model of the recognizer. Here, the information provided by diarization is used.
2. Off-line adaptation of acoustic and language model. The language model can be easily adapted in supervised manner in case user corrects the transcript. Acoustic model can also be easily adapted in unsupervised manner [4].
3. Off-line semi-supervised learning of language and acoustic models.

The first and the second level is already implemented and deployed. However, the second one is invoked manually. The third step is one of our future plans.

2.5. Cluster computation

The system allows for distribution of ASR tasks in a computing cluster based on users' demand. It allows parallelization on the level of submitted jobs and on the level of the ASR core. Presently, there are two separate queues. The first one is fast for high priority jobs (paying customers) and second one for standard jobs (free users). The whole processing runs on standard U1 server with Linux OS now and can be scaled up to the BUT computational server (up to 2000 cores) or commercial virtual servers such as Amazon EC2, or Microsoft Azure.

3. OBTAINING PERFECT TRANSCRIPTION

While ASR accuracy never reaches 100%, SpokenData aims at providing the user with perfect transcription. The service allows this by:

1. correction of the transcriptions by the user.
2. paying hand-transcription.

For both scenarios, we have developed a web-based editor that synchronizes transcription with audio and video and allows user operations known from other systems (for example ICSI Transcriber):

- correcting text
- moving segment's time boundaries
- merge captions
- insert captions

in an efficient way using keyboard shortcuts.

We believe, that this system is quite interesting itself allowing for distributed processing of audiovisual data in corpus transcription, subtitling, and other tasks.

4. USER INTERFACE AND SOFTWARE ARCHITECTURE

The way the system interacts with the user is tailored for naïve users and the most intuitive possible. It was developed in tight cooperation with transcriptionists. The UI for SpokenData features:

- keyboard shortcuts for editing the transcript without touching the mouse
- personalized multimedia library with categories and tagging

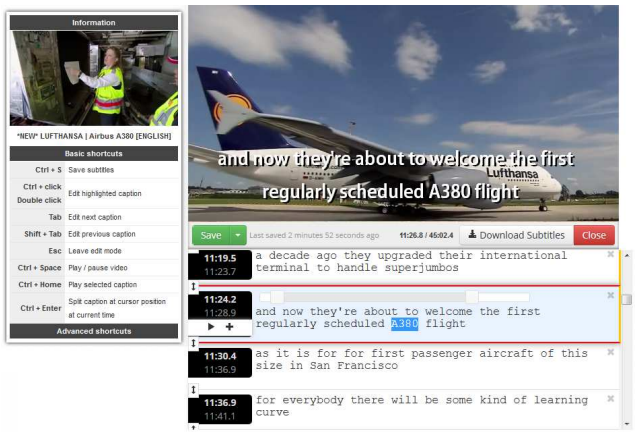

Figure 2: UI of the caption editor.

The SpokenData website is built with PHP, HTML5, CSS3 and JQuery and was tested in all major browsers on all major platforms. The backend is connected to our computing cluster. The processing can be quite time consuming depending on the settings and audio length, therefore the user is notified by email when all data is ready.

5. REFERENCES

[1] Ng, T., Veselý, K., Matějka P. et al., Developing a Speech Activity Detection System for the DARPA RATS Program, in Proceedings of InterSpeech 2012, September 2012

[2] Kenny, P. et al., Diarization of Telephone Conversations using Factor Analysis IEEE Journal of Selected Topics in Signal Processing, December 2010

[3] Karafiat, M. et al., BUT BABEL System for Spontaneous Cantonese, in Proceedings of InterSpeech 2013, August 2013

[4] Grezl, F. et al., Semi-supervised Bootstrapping Approach for Neural Network Feature Extraction Training, to appear in Proceeding of ASRU 2013