The UGent Institutional Repository is the electronic archiving and dissemination platform for all UGent research publications. Ghent University has implemented a mandate stipulating that all academic publications of UGent researchers should be deposited and archived in this repository. Except for items where current copyright restrictions apply, these papers are available in Open Access.

This item is the archived peer-reviewed author-version of:

STON Efficient Subtitling in Dutch Using State-of-the-Art Tools

Lyan Verwimp, Brecht Desplanques, Kris Demuynck, Joris Pelemans, Marieke Lycke, and Patrick Wambacq


To refer to or to cite this work, please use the citation to the published version:

STON: Efficient Subtitling in Dutch Using State-of-the-Art Tools

Lyan Verwimp\textsuperscript{1}, Brecht Desplanques\textsuperscript{2}, Kris Demuynck\textsuperscript{2},
Joris Pelemans\textsuperscript{1}, Marieke Lycke\textsuperscript{3}, Patrick Wambacq\textsuperscript{1}

\textsuperscript{1}ESAT, KU Leuven, Belgium,  
\textsuperscript{2}ELIS, Ghent University - iMinds, Belgium,  
\textsuperscript{3}VRT, Belgium,  

firstname.lastname@esat.kuleuven.be,elis.ugent.be,vrt.be

Abstract

We present a modular video subtitling platform that integrates speech/non-speech segmentation, speaker diarisation, language identification, Dutch speech recognition with state-of-the-art acoustic models and language models optimised for efficient subtitling, appropriate pre- and post-processing of the data and alignment of the final result with the video fragment. Moreover, the system is able to learn from subtitles that are newly created. The platform is developed for the Flemish national broadcaster VRT in the context of the project STON, and enables the easy upload of a new fragment and inspection of both the timings and results of each step in the subtitling process.

Index Terms: automated subtitling, speaker diarisation, acoustic modelling, language modelling

1. Introduction

In Flanders, the broadcast networks are obliged to provide subtitles to help the deaf and hard of hearing, and this for the majority of their programmes. The production of subtitles is very labour intensive, but with support from speech and language technology (SLT) the efficiency can be improved substantially. In the project STON (Spraak- en Taaltechnologisch Ondertitelen in het Nederlands – Dutch subtitling using speech and language technology) initiated by the Flemish national broadcaster (VRT), we have developed an integrated and modular platform that combines several SLT tools in a flexible and user-friendly subtitling application. An important design requirement is that the system is able to automatically learn from already produced subtitles such that its performance will increase throughout its lifetime. Given the VRT’s very high quality standards, the system was not designed to replace the human subtitler entirely, but to provide a substantial efficiency gain, requiring only limited human interventions on the automatically produced result. The main use cases for the project are high volume productions with high to medium quality speech such as documentaries, news and current affairs. To cope with this scenario we use a recogniser that primarily follows the system is able to automatically learn from already produced subtitles such that its performance will increase throughout its lifetime.

The ASR output is a word stream augmented with markers (ASR timestamps derived from the ASR output).

Index Terms: automated subtitling, speaker diarisation, acoustic modelling, language modelling

2. Text pre- and postprocessing

Text preprocessing is required for the LM training material and for screenplays. The text normalisation is based on the one described in [1], but several changes were made: capital correction was improved, short ‘garbage’ words (that easily lead to confusion in recognition) are removed and fillers such as uh are treated in a class-based way such that semantically and/or acoustically similar fillers are collapsed in the LM.

The ASR output is a word stream augmented with markers for plausible sentence breaks, fillers and silences. A postprocessing step converts this to a format that is more suitable for subtitles by making the sentence-final punctuation explicit, capitalising sentence-initial words, writing long numbers in digits, converting time indication and percentages into a standardised form, and compounding words [1] whenever appropriate.

3. Audio segmentation

The AAS incorporates three main components: speech/non-speech segmentation, speaker diarisation and language identification. The speech/non-speech system detects long non-speech intervals (>1s) that can be discarded in the further processing of the audio stream. Non-speech intervals can contain music and strong background sounds such as applause and street noise, so we rely on a model-based approach [2] to detect these segments.

Speaker diarisation deals with the “who-spoke-when?” problem. The objective is to assign a speaker label to every speech segment. A segmentation stage splits the audio stream into homogeneous segments, and a subsequent clustering stage groups the generated segments into clusters representing single speakers. We use an iVector-based method for both stages [3].

The speaker diarisation allows us to add informative colour codes to the generated subtitles and to profit from speaker adapted models during speech recognition. In addition, the speaker change points coincide with sentence boundaries and thus deliver useful information to the LM.

Many TV programmes in Flanders comprise multiple languages. Dubbing is a rare practice and most foreign speech segments are subtitled. To cope with this scenario we use a

This research is funded by IWT-INNOVATIEF AANBETEDEN and VRT in the STON project. The authors wish to thank the project partners for their contributions to the STON subtitling architecture.
Initially, we envisaged language model adaptation where smaller in-domain LMs, generated automatically based on previous subtitles, are interpolated with a larger background LM. However, tests have shown that this does not result in significant improvements. We attribute this to the fact that the large background LM is trained on 1.2B words from newspapers and magazines [1], which is a good match for our use cases (documentaries and news). Currently, we use a 4-gram LM with interpolated modified Kneser-Ney smoothing and a vocabulary of the 100k most frequent words for scripted programmes or 400k for non-scripted programmes. Given that speed is important and the available resources are limited, we refrained from using a larger 5-gram (no significant improvements) or neural network based LMs (no efficient implementation yet).

The ASR system was built with the SPRAAK toolkit [5] and started from the system developed in [1]. The existing GMM-based baseline system employs 3873 tied states to model the 49 three-state cross-word triphones (46 phones, silence, garage and speaker noise) and 1 single-state triphone (short schwa). A new DNN-based AM was created as well, using the exact same training material. Since SPRAAK currently lacks GPU support for training DNNs, Kaldi was used for training the AM, and the resulting DNN was converted to SPRAAK’s internal format. The ASR system was built with the SPRAAK toolkit [5] and includes all steps described in section 5. The system has a 2GB memory footprint, which is mainly determined by the LM.

The results show a relative improvement between 15% and 24% for the new system. Moreover, the fact that DNN scores are more discriminative helps in keeping the decoding time more in check when handling very difficult data such as Soap and Docu I which contain passages with dialect speech, loud background noise and/or music, concurrent speech and other compounding factors. For easy material with a script (Docu V+S) the overhead of the lexicon creation and post-processing (3 min) is now negligible. This will be solved in future updates of the system. The results on the soap taught us that the very challenging conditions encountered in such programs (dialect, loud background music/noise . . . ) have a detrimental impact on the accuracy of the current SLT tools. In Docu V+S, the remaining errors are due to errors in the G2P, text pre- and post-processing (capitals, interpretation of quotes, subtitling of a long pause in the script, overzealous compounder), interpretation differences (script versus annotator), and colloquial language. The errors in Docu I+S are also due to unscripted parts, very noisy speech passages that were labelled as noise by the AAS, and spontaneous speech phenomena (broken-off words, repetitions, dialect speech . . . ).

A short demonstration of the system can be found here: http://www.esat.kuleuven.be/psi/spraak/demo/STON.

7. References


### Table 1: ASR results for 1 documentary with voice-over only (Docu V) and 1 documentary with interviews (I) with (+) or without using the screenplay, and 1 episode of a daily soap.

<table>
<thead>
<tr>
<th>Progr.</th>
<th>Dur. (m:s)</th>
<th>WER (%)</th>
<th>Timing (m:s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>baseline</td>
<td>new</td>
<td>baseline</td>
</tr>
<tr>
<td>Docu V</td>
<td>49:50</td>
<td>9.10</td>
<td>7.76</td>
</tr>
<tr>
<td>Docu V+S</td>
<td>1.03</td>
<td>0.99</td>
<td>6:12</td>
</tr>
<tr>
<td>Docu I+S</td>
<td>17.06</td>
<td>13.00</td>
<td>33:57</td>
</tr>
<tr>
<td>Soap</td>
<td>30:17</td>
<td>75.18</td>
<td>61.79</td>
</tr>
</tbody>
</table>

Timing (m:s)