

Efficiency of Speech Alignment for Semi-Automated Subtitling in Dutch

Patrick Wambacq and Kris Demuynck

ESAT/PSI-Speech, Katholieke Universiteit Leuven, Belgium,
{wambacq,krisdm}@esat.kuleuven.be

Abstract. This paper describes the use of speech alignment to aid in the process of subtitling Dutch TV programs. The recognizer aligns the audio stream with an existing transcript. The goal is therefore not to transcribe but to generate the correct timing of every word. The system performs subtasks such as audio segmentation, transcript preprocessing, alignment and subtitle compression. The result is not perfect but good enough to gain efficiency when used by a professional subtitler as a starting point to refine and finalize the subtitles. In our tests, considerable time savings of 47 to 53% on average are obtained, such that the generation of subtitles for a 1 hour program, is lowered from between 4 and 7 hours to between 2.5 and 4 hours. This is all the more important in the context of an increased pressure from user groups on governments and broadcasters to reach 100% subtitled TV programs.

1 Introduction

Organizations of the deaf and hard of hearing are since long pushing broadcasters and governments to increase the amount of subtitled television programs. This asks for large investments in new technologies and personnel. One of the technologies that can come to aid is speech recognition. In principle, speech recognition can be used in several ways to produce subtitles:

- Generate the transcript and use this as the basis for subtitles. Depending on the type of TV program (e.g. documentary vs. discussions on politics) this works rather well or not at all. Although possible for some tasks, on average word error rate (WER) is too high and due to the numerous required manual corrections in the subsequent post-editing step, no time is saved in that case.
- A well trained speaker re-speaks the audio and an automatic speech recognition (ASR) system adapted to his voice produces a good quality transcript that is already more or less time-synchronized to the soundtrack. For best quality it should however be manually checked (this time requiring much less time given the quality of the source transcript), and perfectly time-aligned to the soundtrack.
- When transcripts are available (either from a re-speaking step as described above or from the program production) they can be aligned to the audio using a speech recognizer. The result is manually checked in a post-editing step but time savings are considerable.

Several efforts to use speech recognizers for subtitle generation have been reported, e.g. [1] transcribes broadcast news and [2] uses both the re-speaking approach and transcription. However the context is too different to be able to compare results. In a project called NEON (“Nederlandstalige Ondertiteling”, Dutch for “Dutch Subtitling”)¹ technology providers and broadcasters have teamed to evaluate the third approach. The results of this evaluation are presented here.

The organization of the paper is as follows: first the segmentation, transcript preprocessing and alignment subtasks are described. Then the complete system is presented. The results of the evaluation of the system are discussed and finally conclusions are given.

2 Segmentation of the Audio Stream

This stream-based subsystem detects long intervals of non-speech that can be discarded in further processing. It produces the following segmentations with low delay and computational effort:

- speech/non-speech: feed the aligner with speech only segments to lower its error rate and processing time;
- male/female: allow to select an appropriate male or female acoustic model;
- speaker clustering: used for speaker adaptation through speaker specific vocal tract length normalization (VTLN) and spectral mean normalization.

The speech/non-speech segmentation uses gaussian mixture models (GMM’s) to distinguish several categories of audio events (speech, music, speech+music, speech+other). The speaker segmentation and clustering is based on a Bayesian Information Criterion (BIC). The details of the system can be found in [3].

3 Preprocessing of the Transcripts

Next to the audio stream, the second input to the aligner is the available transcript which also contains any or all of the following metadata: speaker identities, timing information, stage directions, description of music cues, etc. This meta information is not fed to the aligner, but is kept aside since it can improve the quality of the generated subtitles. In a merging step it augments the alignment results with extra information. Therefore the transcripts are split into the true transcript (which is tokenized) and the metadata.

4 Speech Alignment

The alignment step takes at its input both the soundtrack and the tokenized transcript and generates a time aligned output, i.e. every word receives exact

¹ The NEON project is carried out within the STEVIN program which is funded by the Dutch and Flemish Governments (<http://www.stevin-tst.org>).

begin and end times, needed to show the subtitle at the right moment. To this end the SPRAAK system [4] is used in recognition mode (and not in alignment mode) with a restricted finite state grammar (FSG) as explained below.

In the **preprocessing stage** simple speaker adaptation is performed through speaker specific spectral mean normalization and VTLN.

The **language model** (LM) consists of a FSG that is built from the input transcript. First, every sentence in the transcript is numbered and assigned to a time window through a linear interpolation rule that takes into account the lengths of the sentence, of the transcript and of the soundtrack. For every segment that is labeled as speech, a set of candidate sentences is constructed. This set contains the sentence that is closest in time to the segment, and a number of previous and following sentences. The set is then used to construct a small FSG that serves as the language model for the alignment. This means that for every speech segment, a dedicated LM is constructed. The set is expanded or contracted and time shifted as the aligner steps through the sequence of segments following some heuristics. This keeps a small number of possibilities to choose from by the aligner while at the same time allows to cope with deviations between transcript and audio. Two types of deviations can occur.

Firstly, there can be extra sentences in the transcript: this occurs typically when the direction decides last minute changes to what is aired (usually to shorten an item). This type of deviation is taken care of by the above heuristic as long as the skips are not too large. In the presence of very large skips, the aligner derails and the user has to restart it from the point where it goes wrong by adjusting the set of candidate sentences. This was observed now and then but given the fast processing time this does not pose many problems (and only the remaining part needs to be aligned again).

Secondly there may be non transcribed audio: obviously whatever the aligner tries to map onto it, it will not be correct. The aligner gets back on track if the set of candidate sentences still contains the correct sentence for the subsequent piece of transcribed audio (i.e. when the non transcribed audio is not too long). In the future we will add a fallback to full recognition mode with a general language model to try to remedy this.

The generated **lexicon** contains all words of the transcript. It is created by an updated version of the system described in [6]. The core lexicon is Fonilex [7] which provides multiple phonetic transcriptions for 170k common Dutch words. Based on a simple classification (initial capital, all capitals, etc.), the remaining words are sent to one or more of the following modules:

- An inflection, derivation and compounding module which finds possible decompositions and merges the phonetic transcriptions of the composing parts using the appropriate assimilation rules.
- A module that handles letter/digits words such as acronyms.
- A grapheme to phoneme (g2p) converter. The g2p system was trained on the Fonilex lexicon and hence produces pronunciations in line with that of standard Dutch words. When handling proper nouns, some rules are first applied to convert the archaic spelling conventions commonly used in Dutch

names to a more modern form. Nevertheless, the automatic transcription of proper nouns (especially from foreign origin) remains problematic.

Fonilex also provides rules to generate the alternative pronunciation variants of a word. An extended version of this rule set was used to generate all likely pronunciation variants which resulted in a median of 3.8 pronunciations per word or 1.13 variants per phone in the canonical transcriptions of the words.

The **acoustic model** (AM) is taken from a Flemish Broadcast News transcription task [5]. In summary, this is a triphone HMM with state emissions modeled by GMMs with globally tied gaussians (4k states and a pool of 50k gaussians), based on MIDA features (mutual information based discriminant analysis) and using 49 three-state cross-word triphones and one single state triphone (short schwa). Since the transcript is known, the search space is very restricted and hence the AM is not very critical to the performance.

5 Overview of the Complete System

A block diagram of the complete system is shown in Fig. 1. The different subtasks

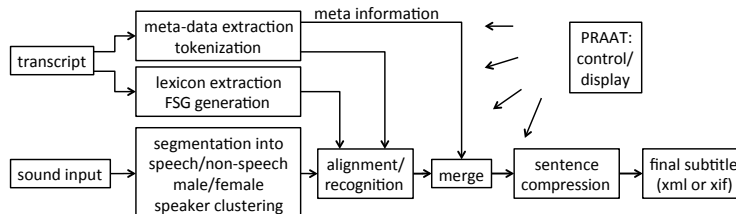


Fig. 1. Block diagram of the system.

(blocks in the figure) are controlled by a PRAAT script. The automatic sentence compression block is not discussed here because it has nothing to do with the speech alignment. Readers who want to know more about it are referred to [8]. We have chosen for the PRAAT software [9] to control the process because it provides a simple way to present the audio signal and several tiers are available that can be tailored to many tasks. Although not optimal, this was certainly a quick and good way to demonstrate the power of the system and to let the broadcasters evaluate its potential. In a real subtitling application, a specialized GUI should be developed or the subtitle aligner should be integrated in an existing software environment for subtitling. In our case, different tiers indicate the result from several steps: speech/non-speech detection (tier 2 in Fig. 2), word alignment (tier 1), subtitle alignment (tier 3). Tier 4 shows the set of candidate sentences that the aligner can choose from for the current segment, as discussed in Sect. 4. The user can change the contents of the tiers if required (to correct a

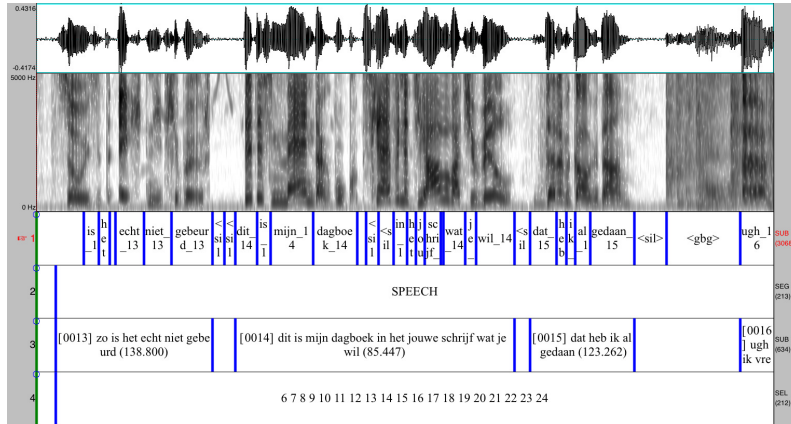


Fig. 2. PRAAT screen shot with the tiers indicating results of different steps.

wrong segmentation, to put the aligner back on track after alignment errors, to eliminate non-transcribed foreign speech, etc.).

The system’s speed is shown in the table below (average figures for 20 minutes of broadcast audio, measured on a 3 GHz Intel core2duo machine (using only one core however) with 2 GBytes of memory, running WinXP SP3).

Soundtrack extraction from video	28 sec
Transcript preprocessing	1 sec
Segmentation and clustering	34 sec
Lexicon generation	4 sec
Alignment	434 sec
Subtitle compression	≈ 1 sec/sentence
Total	≈ 550 sec

On average, the system is a little more than two times faster than real-time. This does not include any human intervention. Time gains that include human post-editing are discussed in Sect. 6.

6 System Evaluation

A prototype of the system was evaluated by broadcasters in Flanders (VRT) and the Netherlands (NPO) on a large variety of programs: documentaries, soaps, animation, human interest, programs for children, action series, church service, etc. with a varying degree of intermixed foreign speech parts and voice-overs that were classified as speech by the segmentation step.

6.1 Qualitative Evaluation

The general impression of the users was very positive. At the beginning, the learning curve (especially the use of the PRAAT controls) was a bit steep, but

after a while the application ran very smooth and proved to be very robust. The software saves lots of time compared to the manual process and the quality of the result was much better than expected.

Sentence compression was rarely used (although of good quality): it was either not required or solved differently, because either the alignment was done with already condensed transcripts (obtained through e.g. re-speaking) or the human editor who conducted the post-editing took care of it manually.

As expected, the PRAAT interface was deemed not optimal, since it was chosen for rapid prototyping, only demonstrating the potential of speech alignment.

6.2 Quantitative Evaluation

A quantitative evaluation was also pursued. Since this is an alignment task, word error rate is not the right metric. The Levenshtein distance between the aligned subtitles and some ground truth could be calculated but this does not consider timing. Also there is no real ground truth since there is not such a thing as a single perfect subtitle: every human subtitler produces slightly different results. Moreover, a Levenshtein distance does not indicate how much time saving the approach would deliver (although we can suspect that there is some correlation between both). Another measure that can be calculated is a histogram of the deviations between the correct timing and generated timing of the subtitles. Subtitlers tell us that segment boundaries should lie within 200 msec from the exact times. A previous experiment on English subtitles in another project, demonstrated that more than 97% of the subtitles met this criterion ([10]). This calculation was not undertaken for the current project but given the improvements in acoustic modelling and alignment implemented in NEON, similar or better results are expected. Also this measure does not tell anything about time savings. Fig. 3 shows a plot of timing errors (for a test on an english program in another project, [10]), clearly indicating that these are not frequent.

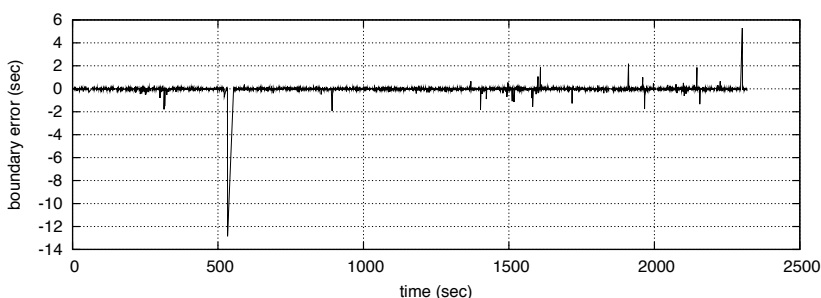


Fig. 3. Alignment timing errors.

The only useful evaluation metric in our view is the time saved. This was measured by professional subtitlers on a mix of TV programs, with and without

the aid of our system (called manual and NEON further on). Manual subtitle generation as well as post-editing of the automatically generated subtitles are performed with existing commercial software (Swift by Softel for VRT and WinCAPS by SysMedia for NPO) whose functionality is not addressed here. The same program was never subtitled by the same person using both approaches to avoid a possible influence of the first result on the second one.

VRT conducted a test where the same person was given different programs from the same series (e.g. two consecutive episodes) for the manual and NEON tests. After producing the subtitle, a final check was performed (running along with the TV program, so its duration equals the audio length). VRT confirmed after the evaluation that this extra step is not really required. The results are shown in the table below, indicating a 20% efficiency gain. The "NEON time" shows the time that the subtitler needs to control and use the NEON aligner with PRAAT; it does not correspond to the required CPU time. The column "manual time" shows the time required to produce the initial subtitle starting from the NEON result (or from scratch). When the final check that was not deemed necessary is removed, the time gain amounts to 53% for VRT.

VRT (total audio length of the test: 725 min); times below in mins					
# of programs	NEON time	manual time	final check	total	RT factor
15 (manual)	0	3130	520	3650	7.02
7 (NEON)	365	585	205	1155	5.63
efficiency gain = $100 \cdot (7.02 - 5.63) / 7.02 = 20\%$					
efficiency gain without final check = 53%					

NPO conducted a test with 9 programs. The manual and NEON tests were performed by different persons. The results in the table below show a 47% efficiency gain. Here the column "total time NEON" includes both the time needed to control the automatic alignment and the manual refining time.

NPO (total audio length of the test: 149 min); times below in mins		
# of programs	total time manual	total time NEON
9	1390	740
efficiency gain = $100 \cdot (1390 - 740) / 1390 = 47\%$		

The difference in gain between the two broadcasters is attributed to differences in experience that the subtitlers have gained with the application, to different procedures when subtitling and to differences in the TV programs.

The broadcasters made several suggestions to further increase the time savings, mainly concerning the user interface (NPO estimates a further potential 45% speedup from a limited set of changes to the GUI). They also found some errors in the PRAAT scripts that control the alignment, that led to some extra alignment errors. Fixing these errors therefore would increase the time savings.

7 Conclusions

In this paper we have reviewed the use of speech alignment in the generation of subtitles for TV programs. Although the aligned subtitles are not always

correct and human intervention in a post-editing step is still required, the time savings are considerable (47 to 53% on average in our tests). The limitation of our approach is that a transcript is needed. This can however be produced by the re-speaking approach by a trained speaker, after which the generated transcription can be used as any other script in our approach. This method would also save time since re-speaking happens in real-time and the generated transcription would be of high quality requiring only minor post-editing on the subtitles that are based on it.

We regard the obtained time savings as minimum values. By fixing some errors, optimizing the alignment, providing automatic language detection and providing a user interface targeted at semi-automatic subtitling, the gains will increase further. We will also add fallback to full recognition mode in a next version. Our experiments described here are only the first steps towards a successful ASR aided subtitling system for Dutch.

References

1. Hugo Meinedo, Márcio Viveiros, João Paulo da Silva Neto, "Evaluation of a Live Broadcast News Subtitling System for Portuguese", Proc. Interspeech2008, pp. 508-511, Brisbane, Australia, September 2008.
2. Shinichi Homma, Akio Kobayashi, Takahiro Oku, Shoei Sato, Toru Imai and Tohru Takagi, "New Real-Time Closed-Captioning System for Japanese Broadcast News Programs", Lecture Notes in Computer Science, 2008, Volume 5105/2008, 651-654.
3. An Vandecatseye and Jean-Pierre Martens, "A Fast, Accurate and Stream-Based Speaker Segmentation and Clustering Algorithm", Proceedings of the 8th European Conference on Speech Communication and Technology, Eurospeech 2003, Vol. 2, pp. 941-944, Geneva, Switzerland, September 2003.
4. Kris Demuynck, Jan Roelens, Dirk Van Compernelle and Patrick Wambacq, "SPRAAK: An Open Source Speech Recognition and Automatic Annotation Kit", Proc. Interspeech2008, p. 495, Brisbane, Australia, September 2008.
5. Kris Demuynck, Antti Puurula, Dirk Van Compernelle and Patrick Wambacq, "The ESAT 2008 System for N-Best Dutch Speech Recognition Benchmark", Proc. IEEE ASRU Workshop, pp. 339-343, Merano, Italy, December 2009.
6. Kris Demuynck, Tom Laureys, Patrick Wambacq and Dirk Van Compernelle, Automatic phonemic labeling and segmentation of spoken Dutch, Proc. LREC-2004, pp. 61-64, Lisbon, Portugal, May 2004.
7. P. Mertens and F. Vercammen, "FONILEX manual", K.U.Leuven – CCL Technical report, 1998, <http://bach.arts.kuleuven.be/fonilex>.
8. Walter Daelemans, Anja Höthker and Erik Tjong Kim Sang, "Automatic sentence simplification for subtitling in Dutch and English", Proc. LREC-2004, pp. 1045-1048, Lisbon, Portugal, May 2004.
9. Paul Boersma and David Weenink, "Praat: doing phonetics by computer" (Version 5.1.05) [Computer program], retrieved May 1, 2009, from <http://www.praat.org/>.
10. Patrick Wambacq, Peter Vanroose, Xiaobin Yang, Jacques Duchateau and Dong Hoon Van Uytsel, "Speech Recognition for Subtitling Purposes", Proc. 5th Intl. Conf. Languages & The Media, p. 46, Berlin, Germany, November 2004. (Abstract).