Speech Recognition Web Services for Dutch

Joris Pelemans[1], **Kris Demuynck**[2], **Hugo Van hamme**[1], **Patrick Wambacq**[1]
[1]Dept. ESAT, Katholieke Universiteit Leuven, Belgium
[2]DSSP, ELIS, Ghent University, Belgium
{joris.pelemans,hugo.vanhamme,patrick.wambacq}@esat.kuleuven.be
kris.demuynck@elis.ugent.be

Abstract

In this paper we present 3 applications in the domain of Automatic Speech Recognition for Dutch, all of which are developed using our in-house speech recognition toolkit SPRAAK. The speech-to-text transcriber is a large vocabulary continuous speech recognizer, optimized for Southern Dutch. It is capable to select components and adjust parameters on the fly, based on the observed conditions in the audio and was recently extended with the capability of adding new words to the lexicon. The grapheme-to-phoneme converter generates possible pronunciations for Dutch words, based on lexicon lookup and linguistic rules. The speech-text alignment system takes audio and text as input and constructs a time aligned output where every word receives exact begin and end times. All three of the applications (and others) are freely available, after registration, as a web application on `http://www.spraak.org/webservice/` and in addition, can be accessed as a web service in automated tools.

Keywords: speech recognition, web services, Dutch

1. Introduction

Automatic Speech Recognition (ASR) has turned out to be a more daunting task than was originally perceived. Despite several decades of research, it is still no match for Human Speech Recognition (HSR) and it is likely that this won't change in the near future. However, this doesn't mean that ASR is entirely useless, on the contrary. It has already proven its value in a lot of applications and continues to do so today. Children are using automatic tutors to improve their reading skills; doctors are gaining time and money using dictating software; disabled people are able to control the computer with their voice instead of the keyboard, ...

Also in the world of Human and Social Sciences (HSS) there exists a great need to be able to search and access spoken data in a more efficient way. A lot of material is stored as audio only: recorded interviews, speeches, news broadcasts, ... Easy access to this material by historians, political scientists or linguists requires a complete or partial transcription. This can then be used to find certain keywords or speakers, or can be analyzed using standard text mining applications. Some of the material has already been transcribed manually but is not aligned with the speech which makes it cumbersome to process. In all these cases ASR technology can help either by aligning the incomplete transcriptions with the audio or by providing automatic transcriptions which can be adopted as such or can be utilized as a baseline transcription to speed up the manual annotation process.

Using the current state-of-the-art technology however often requires expert knowledge: speech recognition systems typically require a specific audio input format and have multiple parameters and models which must be chosen in such a way as to assure optimal performance, choices which depend heavily on the different acoustic and linguistic conditions of the audio. This choice is far from trivial and in most cases the interpretation of the different parameters requires an ASR background. In addition, the end user first has to be able to run the application successfully. If installation is required, it might be platform-dependent or simply laborious to do. Moreover, it is not guaranteed that the user's computer is powerful enough to run the software. All of this is a burden on the end user who does not want to be concerned with such technical details and would prefer to simply input an audio file and receive a transcription.

The Common Language Resources and Technology Infrastructure (CLARIN) [1] initiative is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable for the whole European HSS community. CLARIN offers scholars the tools to allow computer-aided language processing, addressing the multiple roles language plays in the HSS. The CLARIN project TTNWW (Language and Speech Tools for Dutch as Webservice in a Workflow) has the goal to develop and provide Speech & Language Technology (SLT) tools for Dutch to HSS researchers with little or no technical background in the field of ASR. These tools are meant for them to answer their current research questions better and provide them with the possibility to formulate new types of research questions.

In a previous publication (Pelemans et al., 2012), we aimed to put a first step towards achieving the TTNWW goals by presenting our Dutch large vocabulary continuous speech-to-text transcriber (Demuynck et al., 2009) to the user in an environment that meets the CLARIN usability criteria. The software was embedded into a website which we developed using the Computational Linguistics Application Mediator (CLAM) (van Gompel, 2012). It can be accessed with a web browser on `http://www.spraak.org/webservice/` or as a web service in automated tools and as such guarantees fast and easy access. The need for expert knowledge was reduced by providing input conversion as a separate web service and tuning system parameters based on user input.

[1]http://www.clarin.eu

In this paper we discuss the transcriber's new functionality of adding words to the lexicon and present 2 additional applications in the domain of Automatic Speech Recognition (ASR) for Dutch: a grapheme-to-phoneme (g2p) converter and a speech-text alignment system. The focus of this paper is on the actual applications and their free availability, not on the development of the web services. For more information on the web service architecture and on usability, we refer the reader to (Pelemans et al., 2012).
The rest of the paper is organized as follows. Section 2. describes our Dutch speech-to-text transcriber and its recent extension. In Section 3. and 4., we present our grapheme-to-phoneme conversion and text-speech alignment systems, respectively. We end with a conclusion.

2. Speech-to-text Transcriber

2.1. Recognition system

The first installment of the transcriber web service (Pelemans et al., 2012) was built around and has the same performance as the recognizer that was made for the N-Best project (Demuynck et al., 2009) and follows the block diagram of Figure 1. This speaker independent large vocabulary continuous speech recognizer was developed using the SPRAAK (Demuynck et al., 2008; Demuynck, 2001) toolkit and has the capability to select components and adjust parameter settings on the fly, based on the observed conditions in the audio. In the current setup, two conditions are distinguished yielding different acoustic models (AMs) and parameters: studio quality broadband speech and narrowband telephone speech. Audio upload is limited to 2 hours.
For the **acoustic models**, 49 three-state acoustic units (46 phones, silence, garbage and speaker noise) and one single-state phone (short schwa) are modeled using our default tied gaussian approach, i.e. the density function for each of the 4k cross-word context-dependent tied states is modeled as a mixture of an arbitrary subset of Gaussians drawn from a global pool of 50k Gaussians. The mixtures use on average 180 Gaussians to model a 36 dimensional observation vector of features which are obtained by means of a mutual information based discriminant linear transform (MIDA) on vocal-tract length normalized (VTLN) and mean-normalized MEL-scale spectral features and their first and second order time derivatives (Demuynck, 2001).
Using a vocabulary of 400k words, 5-gram word **language models** (LMs) are trained using modified Kneser-Ney discounting on 4 main text components: 12 Southern Dutch newspapers, 10 Northern Dutch newspapers and transcriptions of broadcast news and conversational telephone speech. The 4 LMs are interpolated linearly and perplexity minimization is done to find the optimal interpolation weights.
Lexicon creation is handled by the g2p converter described in Section 3.. Dutch has a decent amount of (regional) pronunciation variation which is addressed by using phonological rules to generate the likely pronunciation variants. This results in a median of 3.8 pronunciations per word or 1.13 variants per phone in the canonical word transcriptions.
Since Dutch compounds are always written as a single word, the word recognition results are optionally post-processed for **compounding**. Two subsequent words are replaced by their compound if the following criteria are met: 1) the words are longer than 3 letters, 2) the words are not very rare, 3) the unigram count of the compound is higher than the bigram count of the individual words. This approach essentially extends the 400k lexicon to a 6M lexicon.
The main **parameters** of the system concern hypothesis pruning and combining language model and acoustic model. To combine the model scores, we employ our standard way of handling this problem (Demuynck, 2001), by having a LM scaling factor and a word startup cost. Beam search pruning is applied to control the amount of hypotheses in the search space (Steinbiss et al., 1994): a threshold indicates how much the score of a hypothesis can drop below the score of the most likely hypothesis; if most hypotheses have a similar score, a beam width parameter is applied to indicate how many hypotheses can be retained, keeping only the best ones.
Table 1 shows the results on the NBest evaluation and development data, comparing different configurations for the real-time factor. The post-processing method was used and separate results are given for the wideband and telephone segments of development data. The substantially larger WER on the evaluation data is due to a larger portion of the data being spontaneous speech or accented speech. For more results, we refer to (Demuynck et al., 2009).

dev, wideband		**dev, telephone**		**eval**	
xRT	WER	xRT	WER	xRT	WER
0.9	6.18%	2.3	29.5%	1.5	21.8%
2.7	5.64%	8.0	28.0%	9.4	20.7%
9.0	5.23%	45.0	25.8%	37.6	20.3%

Table 1: WERs for NBest development and evaluation data with different configurations for real-time factor

2.2. Adding words

Since its first publication (Pelemans et al., 2012), the transcriber has been extended with the capability of adding new words. This allows the users to take full advantage of any background information concerning the task at hand. For a recognizer to successfully adopt a new word, it needs to have information about its pronunciation and its linguistic behavior. A phonemic transcription of the word can be provided manually or it can be acquired by running our g2p service, described in section 3., after which the user can choose to correct the transcription. The uploaded words will then be added to the pronunciation lexicon.
For a new word to be added to the LM, we need to know how it behaves in the context of other words. Although it is possible to ask the user to provide training material for each new word, we wanted to prevent the upload of large texts and subsequent retraining of the LM. Moreover, collecting valuable training material is often a cumbersome task with which we do not want to burden the user. Instead we opted for synonym mapping: the user provides the new word together with a word that behaves in an almost identical way, both syntactically and semantically. The language model

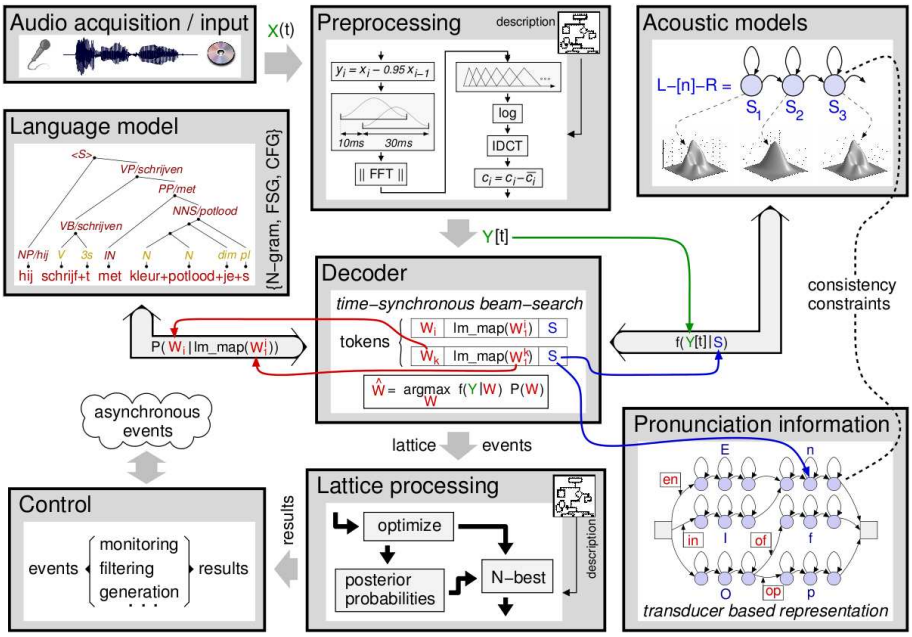

Figure 1: *Schematic diagram of SPRAAK*

will then use (a fraction of) the LM scores of the synonym to predict the new word. This is a simple way to cover many new words, as the words can be looked up in a thesaurus or can be specified by the user and it does not require any retraining of the LM.

This method of adding words has not yet been evaluated, but we are running experiments at the time of writing. Obviously the outcome of this technique is very dependent on the task at hand – the audio quality, speaker accent and spontaneity – and the uploaded words together with their transcription and synonym.

3. Grapheme-to-phoneme Conversion

Grapheme-to-phoneme conversion is the process of converting the orthographic form of a word into a phonemic transcription and is a necessary component in both speech recognition and synthesis. The g2p web service consists of an updated version of the system described in (Demuynck et al., 2002). To obtain plausible pronunciations for words, the following techniques and resources were used:

Lexicon lookup: Fonilex (Mertens and Vercammen, 1998) provides multiple phonemic transcriptions for all frequent standard Dutch words. For the foreign words we draw on Comlex (English), Celex (German) and Brulex (French). If a foreign word is part of more than one of these lexica, the different pronunciations are put in parallel since the orthography does not specify which foreign language is used. The same holds for capitalized words (e.g. 'Hamburg' which may either be pronounced in a Dutch, German or English fashion). Furthermore, specific lexica were made for the most frequent proper nouns (5892 entries), interjections, frequently used dialect words and items not covered in one of the other lexica (982 entries).

Compounding, derivation and inflection: As Dutch is a morphologically productive language, lexica by themselves cannot cover all possible word forms. The pronunciation of non-trivial compounds and derivations is found by decomposing the word into its basic constituents, concatenating their pronunciations and applying a set of assimilation rules. In our approach, all decompositions possible based on pure orthographic constraints are pursued, i.e. no morphotactic constraints are imposed, which leads to some degree of overgeneration (e.g. 'rijstroken' → 'rij' + 'stroken' / 'rijst' + 'roken'). This overgeneration rarely resulted in new pronunciation variants and even showed to be useful for handling Dutch proper nouns and mispronunciations.

Abbreviations and digits: Abbreviations are phonemically transcribed as the concatenation of the constituent letter word transcriptions. In case the abbreviation – converted to lowercase – maps to an existing word, the corresponding word pronunciation is added as well. Frequently occurring exceptions (e.g. NATO) are added to one of the specific lexica. The pronunciation of numbers inside the abbreviations is solved with a rule-based system.

Grapheme-to-phoneme system: A grapheme-to-phoneme system was developed as a fall-back. The g2p system is based on the Induction Decision Tree (ID3) mechanism (Pagel et al., 1998) and trained on the Fonilex database. More information on the configuration of the g2p system is given in (Demuynck et al., 2002).

The performance of the g2p was evaluated by a 10-fold cross-validation experiment on Fonilex (Mertens and Vercammen, 1998) and on all components of the CGN corpus (Oostdijk, 2000) (Corpus Gesproken Nederlands/Corpus Spoken Dutch) by counting the number of insertions, deletions and substitutions with respect to a hand-checked reference transcription. This yielded an average phone error rate of 6.0% and 3.14% respectively. For more results we refer to (Demuynck et al., 2002) and (Demuynck et al., 2004).

4. Speech-text Alignment

The alignment service is an updated version of the system described in (Wambacq and Demuynck, 2011). It takes at its input both the audio (limited to 2 hours) and the tokenized transcript and generates a time aligned output, i.e. every word receives exact begin and end times. To this end the SPRAAK system (Demuynck et al., 2008) is used in recognition mode (and not in alignment mode) with a restricted finite state grammar (FSG) as explained below. A block diagram of the system is shown in Figure 2.

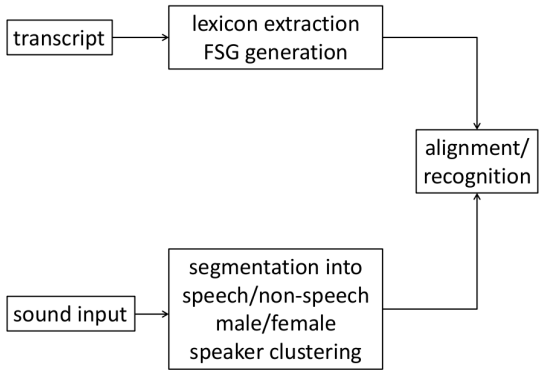

Figure 2: Block diagram of the speech alignment system

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

The **language model** 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 audio. 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 LM 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.

The **acoustic model** is taken from the recognizer described in Section 2. and the **lexicon**, containing all words of the transcript, is created by the g2p described in Section 3..

The alignment system was evaluated on a large variety of television 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. The performance was measured in the context of a semi-automated subtitling task, as the time gained by professional subtitlers on a mix of TV programs, with and without the aid of our system. In all cases the time gain was around 50%. More information can be found in (Wambacq and Demuynck, 2011).

5. Conclusion

We presented our 3 newest applications in the domain of Automatic Speech Recognition for Dutch, all of which are developed using our in-house speech recognition toolkit SPRAAK. The speech-to-text transcriber, grapheme-to-phoneme converter and speech-text alignment system are freely available as a web application on `http://www.spraak.org/webservice/` and can be accessed as a web service.

6. Acknowledgements

The authors would like to thank the CLARIN initiative and more specifically the TTNWW project for supporting their work which enabled sharing these tools with the rest of the world.

7. References

Demuynck, K., Laureys, T., and Gillis, S. (2002). Automatic generation of phonetic transcriptions for large speech corpora. In *Proc. ICSLP*, volume I, pages 333–336.

Demuynck, K., Laureys, T., Wambacq, P., and Compernolle, D. V. (2004). Automatic phonemic labeling and segmentation of spoken dutch. In *LREC*.

Demuynck, K., Roelens, J., Compernolle, D. V., and Wambacq, P. (2008). SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit. In *INTERSPEECH*, page 495.

Demuynck, K., Puurula, A., Van Compernolle, D., and Wambacq, P. (2009). The ESAT 2008 system for N-Best Dutch speech recognition benchmark. In *Proc. ASRU*, pages 339–343.

Demuynck, K. (2001). *Extracting, Modelling and Combining Information in Speech Recognition*. Ph.D. thesis, K.U.Leuven ESAT.

Mertens, P. and Vercammen, F. (1998). FONILEX manual. Technical report, K.U.Leuven – CCL. http://bach.arts.kuleuven.ac.be/fonilex/.

Oostdijk, N. (2000). The Spoken Dutch Corpus. *The ELRA Newsletter*, 5(2):4–8. http://lands.let.ru.nl/cgn/.

Pagel, V., Lenzo, K., and Black, A. (1998). Letter to sound rules for accented lexicon compression. In *Proc. ICSLP*, volume I, pages 252–255, Sydney, Australia.

Pelemans, J., Demuynck, K., and Wambacq, P. (2012). Dutch automatic speech recognition on the web: Towards a general purpose system. In *INTERSPEECH*.

Steinbiss, V., Tran, B.-H., and Ney, H. (1994). Improvements in beam search. In *Proc. ICSLP*, pages 2143–2146.

van Gompel, M., (2012). *CLAM: Computational Linguistics Application Mediator. Documentation. ILK Technical Report 12-02*. http://ilk.uvt.nl/downloads/pub/papers/ilk.1202.pdf.

Wambacq, P. and Demuynck, K. (2011). Efficiency of speech alignment for semi-automated subtitling in Dutch. In *TSD*, pages 123–130.