Adapting the Filibuster text-to-speech system for Norwegian bokmål

Kåre Sjölander and Christina Tånnander
The Swedish Library of Talking Books and Braille (TPB)

Abstract

The Filibuster text-to-speech system is specifically designed and developed for the production of digital talking textbooks at university level for students with print impairments. Currently, the system has one Swedish voice, 'Folke', which has been used in production at the Swedish Library of Talking Books and Braille (TPB) since 2007. In August 2008 the development of a Norwegian voice (bokmål) started, financed by the Norwegian Library of Talking Books and Braille (NLB). This paper describes the requirements of a text-to-speech system used for the production of talking textbooks, as well as the developing process of the Norwegian voice, 'Brage'.

Introduction

The Swedish Library of Talking Books and Braille (TPB) is a governmental body that provides people with print impairments with Braille and talking books. Since 2007, the in-house text-to-speech (TTS) system Filibuster with its Swedish voice 'Folke' has been used in the production of digital talking books at TPB (Sjölander et al., 2008). About 50% of the Swedish university level textbooks is currently produced with synthetic speech, which is a faster and cheaper production method compared to the production of books with recorded human speech. An additional advantage is that the student gets access to the electronic text, which is synchronized with the audio. All tools and nearly all components in the TTS system components are developed at TPB.

In August 2008, the development of a Norwegian voice (bokmål) started, financed by the Norwegian Library of Talking Books and Braille (NLB). The Norwegian voice 'Brage' will primarily be used for the production of university level textbooks, but also for news text and the universities' own production of shorter study materials. The books will be produced as DAISY-books, the international standard for digital talking books, via the open-source DAISY Pipeline production system (DAISY Pipeline, 2009).

The Filibuster system is a unit selection TTS, where the utterances are automatically generated through selection and concatenation of segments from a large corpus of recorded sentences (Black and Taylor, 1997).

An important feature of the Filibuster TTS system is that the production team has total control of the system components. An unlimited number of new pronunciations can be added, as well as modifications and extensions of the text processing system and rebuilding of the speech database. To achieve this, the system must be open and transparent and free from black boxes.

The language specific components such as the pronunciation dictionaries, the speech database and the text processing system are NLB's property, while the language independent components are licensed as open-source.

Requirements for a narrative textbook text-to-speech system

The development of a TTS system for the production of university level textbooks calls for considerations that are not always required for a conventional TTS system.

The text corpus should preferably consist of text from the same area as the intended production purpose. Consequently, the corpus should contain a lot of non-fiction literature to cover various topics such as religion, medicine, biology, and law. From this corpus, high frequency terms and names are collected and added to the pronunciation dictionary.

The text corpus doubles as a base for construction of recording manuscripts, which in addition to general text should contain representative non-fiction text passages such as bibliographic and biblical references, formulas and URL's. A larger recording manuscript than what is conventionally used is required in order to cover phone sequences in foreign names, terms, passages in English and so on. In addition, the above-mentioned type of textbook specific passages necessitates complex and well-developed text processing.

The number of out-of-vocabulary (OOV) words is likely to be high, as new terms and names frequently appear in the textbooks, requiring sophisticated tools for automatic generation of pronunciations. The Filibuster system distinguishes between four word types; proper names, compounds and simplex words in the target language, and English words.

In order to reach the goal of making the textbooks available for studies, all text - plain Norwegian text and English text passages, OOV words and proper names - need to be intelligible, raising the demands for a distinct and pragmatic voice.

The development of the Norwegian voice

The development of the Norwegian voice can be divided into four stages: (1) adjustments and completion of the pronunciation dictionary and the text corpus, and the development of the recording manuscripts, (2) recordings of the Norwegian speaker, (3) segmentation and building the speech database, and (4) quality assurance.

Pronunciation dictionaries

The Norwegian HLT Resource Collection has been made available for research and commercial use by the Language Council for Norwegian (http://www.sprakbanken.uib.no/). The resources include a pronunciation dictionary for Norwegian bokmål with about 780,000 entries, which were used in the Filibuster Norwegian TTS. The pronunciations are transcribed in a somewhat revised SAMPA, and follow mainly the transcription conventions in Øverland (2000). Some changes to the pronunciations were done, mainly consistent adaptations to the Norwegian speaker's pronunciation and removal of inconsistencies, but a number of true errors were also corrected, and a few changes were made due to revisions of the transcription conventions.

To cover the need for English pronunciations, the English dictionary used by the Swedish voice, consisting of about 16,000 entries, was used. The pronunciations in this dictionary are ‘Swedish-style’ English. Accordingly, they were adapted into ‘Norwegian-style’ English pronunciations. 24 xenophones were implemented in the phoneme set, of which about 15 have a sufficiently number of representations in the speech database, and will be used by the TTS system. The remaining xenophones will be mapped into phonemes that are more frequent in the speech database.

In addition, some proper names from the Swedish pronunciation dictionary were adapted to Norwegian pronunciations, resulting in a proper name dictionary of about 50,000 entries.

Text corpus

The text corpus used for manuscript construction and word frequency statistics consists of about 10.8 million words from news and magazine text, university level textbooks of different topics, and Official Norwegian Reports (http://www.regjeringen.no/nb/dok/NOUer.html?id=1767). The text corpus has been cleaned and sentence chunked.

Recording manuscripts

The construction of the Norwegian recording manuscript was achieved by searching phonetically rich utterances iteratively. While diphones was used as the main search unit, searches also included high-frequency triphones and syllables.

As mentioned above, university level textbooks include a vast range of different domains and text types, and demands larger recording manuscripts than most TTS systems in order to cover the search units for different text types and languages. Biographical references, for example, can have a very complex construction, with authors of different nationalities, name initials of different formats, titles in other languages, page intervals and so on. To maintain a high performance of the TTS system for more complex text structures, the recording manuscript must contain a lot of these kinds of utterances.

To cover the need of English phone sequences, a separate English manuscript was recorded. The CMU ARCTIC database for speech synthesis with nearly 1,150 English utterances (Kominek and Black, 2003) was used for this purpose. In addition, the Norwegian manuscript contained many utterances with mixed Norwegian and English, as well as email addresses, acronyms, spelling, numerals, lists, announcements of DAISY specific structures such as page numbers, tables, parallel text and so on.

Recordings

The speech was recorded in NLB’s recording studio. An experienced male textbook speaker was recorded by a native supervisor. The recordings were carried out in 44.1 KHz with 24-bit resolution. Totally, 15,604 utterances were recorded.

Table 1. A comparison of the length of the recorded speech databases for different categories, Norwegian and Swedish

	Norwegian	Swedish
Total time	26:03:15	28:27:24
Total time (speech)	18:24:39	16:15:09
Segments	568 606	781 769
Phones	519 065	660 349
Words	118 104	132 806
Sentences	15 604	14 788

A comparison of the figures above shows that the Swedish speaker is about 45% faster than the Norwegian speaker (11.37 vs. 7.83 phones per second). This will result in very large file-sets for the Norwegian textbooks, which often consists of more than 400 pages, and a very slow speech rate of the synthetic speech. However, the speech rate can be adjusted in the student's DAISY-player or by the TTS-system itself. On the other hand, a slow speech rate comes with the benefit that it well articulated and clear speech can be attained in a more natural way compared to slowing down a voice with an inherently fast speech rate.

Segmentation

Unlike the Swedish voice, for which all recordings were automatically and manually segmented (Ericsson et al., 2007), all the Norwegian utterances were control listened, and the phonetic transcriptions were corrected before the automatic segmentation was done. In that way, only the pronunciation variants that actually occurred in the audio had to be taken into account by the speech recognition tool (Sjölander, 2003). Another difference from the Swedish voice is that plosives are treated as one continuous segment, instead of being split into obstruction and release.

Misplaced phone boundaries and incorrect phone assignments will possibly be corrected in the quality assurance project.

Unit selection and concatenation

The unit selection method used in the Filibuster system is based mainly on phone decision trees, which find candidates with desired properties, and strives to find as long phone sequences as possible to minimise the number of concatenation points. The optimal phone sequence is chosen using an optimisation technique, which looks at the phone's joining capability, as well as its spectral distance from the mean of all candidates. The best concatenation point between two sound clips is found by correlating their waveforms.

Text processing

The Swedish text processing system was used as a base for the Norwegian system. Although the two languages are similar in many ways, many modifications were needed.

The tokenisation (at sentence, multi-word and word level) is largely the same for Swedish and Norwegian. One of the exceptions is the sentence division at ordinals, where the standard Norwegian annotation is to mark the digit with a period as in '17. mai', which is an annotation that is not used in Swedish.

The Swedish part-of-speech tagging is done by a hybrid tagger, a statistical tagger that uses the POS trigrams of the Swedish SUC2.0 corpus (Källgren et al., 2006), and a rule-based complement which handles critical part-of-speech disambiguation. It should be mentioned that the aim of the inclusion of a part-of-speech tagger is not to achieve perfectly tagged sentences; its main purpose is to disambiguate homographs. Although the Swedish and Norwegian morphology and syntax differ, the Swedish tagger and the SUC trigram statistics should be used also for the Norwegian system, even though it seems like homographs in Norwegian bokmål need more attention than in Swedish. As an example, the relatively frequent Norwegian homographs where one form is a noun and the other a verb or a past participle, for example 'laget', in which the supine verb form (or past participle) is pronounced with the 't' and with accent II [''lɑ:.gət], while the noun form is pronounced without the 't' and with accent I ['lɑ:.gə]. As it stands, it seems as the system can handle these cases to satisfaction. OOV words are assigned their part-of-speech according to language specific statistics of suffixes of different lengths, and from contextual rules. No phrase parsing is done for Norwegian, but there is a future option to predict phrase boundaries from the part-of-speech tagged sentence.

Regarding text identification, that is classifying text chunks as numerals or ordinals, years, intervals, acronyms, abbreviations, email addresses or URLs, formulas, biographical, biblical or law references, name initials and so on, the modifications mainly involved translation of for instance units and numeral lists, new lists of abbreviations and formats for ordinals, date expressions and suchlike. Similar modifications were carried out for text expansions of the above-mentioned classifications.

A language detector that distinguishes the target language from English was also included in the Norwegian system. This module looks up the words in all dictionaries and suggests language tag (Norwegian or English) for each word depending on unambiguous language types of surrounding words.

OOV words are automatically predicted to be proper names, simplex or compound Norwegian words or English words. Some of the pronunciations of these words are generated by rules, but the main part of the pronunciations is generated with CART trees, one for each word type.

The output from the text processor is sent to the TTS engine in SSML format.

Quality assurance

The quality assurance phase consists of two parts, the developers' own testing to catch general errors, and a listening test period where native speakers report errors in segmentation, pronunciation and text analysis to the developing team. They are also able to correct minor errors by adding or changing transcriptions or editing simpler text processing rules. Some ten textbooks will be produced for this purpose, as well as test documents with utterances of high complexity.

Current status

The second phase of the quality assurance phase with native speakers will start in May 2009. The system is scheduled to be put in production of Norwegian textbooks by the autumn term of 2009. Currently, the results are promising. The voice appears clear and highly intelligible, also in the generation of more complex utterances such as code switching between Norwegian and English.

References

Black A. and Taylor P. (1997). Automatically clustering similar units for unit selection in speech synthesis. Proceedings of Eurospeech 97, Rhodes, Greece.

DAISY Pipeline (2009). http://www.daisy.org/projekcts/pipeline.

Ericsson C., Klein J., Sjölander K. and Sönnebo L. (2007). Filibuster – a new Swedish text-to-speech system. Proceedings of Fonetik, TMH-QPSR 50(1), 33-36, Stockholm.

Kominek J. and Black A. (2003). CMU ARCTIC database for speech synthesis. Language Technologies Institute. Carnegie Mellon University, Pittsburgh PA . Technical Report CMU-LTI-03-177. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

Källgren G., Gustafson-Capkova S. and Hartman B. (2006). Stockholm Umeå Corpus 2.0 (SUC2.0). Department of Linguistics, Stockholm University, Stockholm.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. Proceedings of Fonetik 2003, 93-96, Stockholm.

Sjölander K., Sönnebo L. and Tånnander C. (2008). Recent advancements in the Filibuster text-to-speech system. SLTC 2008.

Øverland H. (2000). Transcription Conventions for Norwegian. Technical Report. Nordisk Språkteknologi AS