Filibuster – a new Swedish text-to-speech system

Christina Ericsson, Jesper Klein, Kåre Sjölander, and Lars Sönnebo
Talboks- och Punktskriftsbiblioteket

Abstract

A Swedish text-to-speech system has been developed at the Swedish Library of Talking Books and Braille (TPB). The system, named Filibuster, is open and extensible and makes it possible to generate synthetic speech with a high degree of control.

The Filibuster system is used in the production of talking books in TPBs service for print handicapped students at university level. Through the use of text-to-speech, students can receive their talking books much faster. Also, each book costs less to produce. The system was deployed in February 2007 and during this year the plan is to produce a total of 200 titles.

The system has been designed specifically for creating talking book versions of university textbooks. It has a large lexicon, covering some 573,000 words and names. Filibuster includes a comprehensive text pre-processor to write out non-word entities, such as numbers, characters and expressions. So far, one male voice, Folke, has been created, but more are planned.

Introduction

The Swedish Library of Talking Books and Braille provides people with print handicaps with access to printed text. This also includes a service for print disabled students at university level. Typically, this means that digital talking books are created through recording of human narration. This enables students to listen to text books using either software or hardware digital talking book players. In order to provide students with course material in the form of talking books more quickly, text-to-speech (TTS) technology seemed like a natural solution. This approach had already been investigated in a project were talking book versions of printed magazines had been produced and evaluated, (Klein, 2004). The adoption of synthetic speech would also bring the added advantage that more books could be produced within the given budget constraints.

Most text-to-speech systems today are corpus-based (Black et. Al, 1997 and Beutnagel et al., 1999). Synthesized utterances are automatically generated through selection and concatenation of segments from a large collection (corpus) of recorded sentences, this process is often referred to as *unit selection*. The segments can vary in length from a single phone to several words. The manuscript for the sentences to be recorded is chosen carefully to cover as much normal articulatory and prosodic variation as possible, while at the same time keeping the total size of the recorded speech within the speaker time and processing resources available.

It was decided to develop a new Swedish text-to-speech system using a corpus-based approach. The idea was to use text from a collection of university textbooks in the recordings. Similarly, a lexicon with relevant words from the same text corpus would be created with the goal of achieving a system specially targeted at this domain. Development started in 2005 and the first completed version of the system, named Filibuster, was deployed in February 2007. The first voice recorded was male and was given the name Folke. Up until the end of April, 74 talking books have been produced. By the end of the year the plan is to have produced a total of 200 talking books.

Pronunciation lexicon

The pronunciation lexicon is based on *Svenska språknämndens uttalsordbok* (Garlén, 2003). An electronic version of this book was licensed and converted for text-to-speech purposes. Part-of-speech information and inflected forms where added automatically. This resulted in about 484,000 lexical entries. Another 52,000 words occurring 3 or more times in a text corpus created from a set of 400 books available at TPB were transcribed and added. In addition, about 23,000 names were included as well as 14,000 English words. The total number of lexical entries thus came to 573,000. The text corpus was also used to compute word frequencies needed in the system to disambiguate decompositions for unknown compound words.

Speech corpus recording

TPB has thousands of talking books available. Since a typical book consists of many hours of read speech it seemed like an interesting possibility to use an existing talking book as the speech corpus in creating a speech synthesis voice. This approach was ruled out, after some initial testing, for sound and speech quality reasons. Talking books are not produced to meet the quality level needed for this task. Since there were no other available speech corpora for Swedish a new one needed to be designed and recorded.

Voice selection

Initially, it was decided to use a male voice, since there were several promising candidates known at the beginning of the project. The project group jointly listened to samples from about 30 different professional speakers. The main requirements were clear articulation and stable prosody. Also, the speaker had to be able to read for several hours per recording day without any audible fatigue effects on the voice. An initial recording session took place and the results were used to test the speaker's viability by creating a minimal speech synthesizer for a limited domain.

Manuscript preparation

The text corpus, mentioned above, was also used for preparing the recording manuscript. An automatic procedure was used to extract sentences, resulting in a wide distribution of different syllable and phone combinations in varying prosodic contexts. (Kominek & Black 2003). Sentences had a length of between 5 and 15 words. A number of phrases and words commonly occuring in talking books were also included. In total, 15000 sentences were chosen and grouped together in lots of 200, which seemed like an appropriate amount for a recording session with the chosen speaker.

Studio set-up

A recording booth was acquired and assembled in a quiet room in the TPB offices. The recordings were carried out in 44.1KHz, with 24-bit resolution, directly to hard-disk. A Neumann TLM-103 microphone was used in conjunction with a TL Audio preamplifier and a PreSonus FIREBOX for analog to digital conversion. Recording was done sentence by sentence using custom software which also presented the text on a screen in the recording booth. Each sentence was checked both by listening and by observing its waveform during recording. Dubious cases were checked again before proceeding to the next sentences, and re-recorded if necessary.

Speech corpus segmentation

The recordings were segmented automatically using the text manuscript as input (Sjölander, 2003). Aligned word and phone transcription files were generated. Subsequently all files were manually checked and corrected by a single person over a period of two months. Special software was developed for this work, where the process was entirely keyboard driven and the number of keystrokes minimized. A number of special annotations were used in the manual transcriptions. For example, glottalized boundaries and reduced phones were marked. Two problematic sentences, out of a total of 14790 recorded, were discarded in this stage. The size of the recorded corpus after segmentation is shown in Table 1.

Table 1. Length of the recorded speech corpus for different categories.

Category	Length
Total time:	28:27:24
Total time (speech):	16:15:09
Segments:	781769
Phones:	660349
Words:	132806
Sentences:	14788

System implementation

The implementation of the TTS system relies on previous work described in (Gustafson and Sjölander 2004), which borrows many ideas from the Festival speech synthesis system (Black et al. 1998).

The segmented corpus is processed and indexed in order to create a phone unit inventory that can be easily searched. This process is fully automatic. Phones are grouped together according to spectral similarity using decision tree induction (Black et al., 2000a).

The synthesized speech is generated from text in several consecutive and distinct stages.

Text pre-processing

The text is transformed into a string of words that can be found in the lexicon. Digits and number are written out. Other characters are transformed depending on context, the character "-", for example, can be read as "minus", "to" or nothing depending on how it is used in a sentence. Transcriptions are generated for acronyms and more complex expressions such as web-addresses. Pauses and phrase boundaries are also inserted in order to increase intelligibility. Unknown words can be tagged as "probable name" or "probable English word" in order for the system to adapt its pronunciation rules as necessary. Also, many homographs are disambiguated at this stage.

Word pronunciation

For every word its transcription, morpheme and syllable boundaries, and part-of-speech is looked up in the pronunciation dictionary. Unknown words are handled by removing inflectional suffixes before lexicon lookup. Alternatively, the system will try to decompound them into two, or more, shorter words or stems. Ambiguous cases are handled by choosing the most frequent component words. For example, for the word *rostrött*, the decomposition *rost-rött* will be chosen over the less likely alternatives *ros-trött* and *ro-strött*. As a last resort, pronunciation rules, in the form of statistically trained decision trees, will be applied. There are separate rule sets for Swedish words, Swedish names, and English words. The text pre-processor will have provided the tags necessary to select the appropriate set.

Target generation

All the information gathered in the previous stage is collected in a data structure describing the whole target sentence. This facilitates handling of the utterance as a hierarchical word-syllable-phone structure for the processing needed in the next stage. Also, this makes it easy to handle assimilation between words.

Unit candidate generation

The target sentence structure is used to find suitable candidates for each phone in the utterance. The main method is to search the phone decision trees. There is one decision tree per phone is used and these have been constructed automatically. At each branch in the tree, questions on phonetic context are used to find a leaf with phone candidates with desired properties. Several techniques are used to increase the probability of getting good matches from the corpus, for example, special searches are made to find matching longer phone sequences in the corpus. (Black et al., 2000b) Several constraints are also applied to further control with phone units that will be considered for further processing. The number of resulting candidates can vary between a few and several hundred for each phone in an utterance.

Unit selection

Given the set of possible candidates for each phone in the target utterance, the optimal phone sequence is chosen using an optimization technique. The spectrum at the end points of two units determines how well the can be joined and this factor is used in the process. Similarly, the spectral distance for a phone from the mean of all candidates for a given phone is used. During this selection process weights are applied in order to balance where segments break should be more likely to happen: between phones, syllables, or whole words.

Unit concatenation

The sound clips representing each phone in the optimal sequence of phones is finally retrieved and concatenated. The best fit, or concatenation point, between two clips is found by correlating their waveforms. For the special case where two or more phones are adjacent in the corpus concatenation is trivial, the whole segment can be used as is.

Conclusions and future work

A new Swedish text-to-speech system, Filibuster, specially designed for producing talking books has been developed. The system is corpus-based and the manuscript used for recording the voice, as well as the lexicon, have been created based on texts from university textbooks, which also is the domain of texts that Filibuster should be able to generate synthetic speech for. Some sound examples are available at the project's web site, www.filibuster.se.

The Filibuster system has been improved continuously since its deployment in February 2007. Improved text pre-processing capabilities and new words have been added to the lexicon. Filibuster's algorithms have been tuned and improved, both regarding speech quality and processing requirements. Several other improvements are planned, for example, improved part-of-speech tagging, domain specific text-processing, and prominence prediction. Another foreseen development is the creation of a female voice.

During 2007, TPB plans to produce 200 books using Filibuster. In doing this a larger text corpus from relevant literature will be created, that can be used to produce an improved recording manuscript. One weakness of the current system is the lack of short utterances in the corpus as well as special words like acronyms and non-Swedish words. This will also be addressed in order to improve coverage of phone units.

Another development is the recent start of a project that aims at making government reports available as digital talking books. In this project the Filibuster system will be extended with new words and concepts, as well as pre-processing capabilities, needed to handle text occurring in these types of reports.

References

Beutnagel M, Conkie A., Schroeter J, Stylianou Y, and Syrdal A (1999). The AT&T Next-Gen TTS System. *137th Acoustical Society of America meeting*, Berlin.

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

Black A, Taylor P, and Caley R (1998). The Festival speech synthesis system, http://festvox.org/festival.

Black A, and Lenzo K (2000). Building Voices in the Festival Speech Synthesis System, http://festvox.org/bsv.

Black A, and Lenzo K (2000). Limited Domain Synthesis. *Proceedings of ICSLP2000*, Beijing, China,.

Garlén C (2003). Svenska språknämndens uttalsordbok - 67 000 ord i svenskan och deras uttal. Sweden, Norstedts akademiska förlag.

Gustafson J, and Sjölander K (2004). Voice creation for conversational fairy-tale Characters. Proceedings of the

Klein J (2004). http://www.daisy.tpb.se/kurt/

Kominek J, and Black A (2003). CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177, Carnegie Mellon University.

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