Recent advancements in the Filibuster text-to-speech system

Kåre Sjölander
Talboks- och punktskriftsbiblioteket
Enskede, Sweden
kares@tpb.se

Lars Sönnebo
Talboks- och punktskriftsbiblioteket
Enskede, Sweden
lars.sonnebo@tpb.se

Christina Tånnander
Talboks- och punktskriftsbiblioteket
Enskede, Sweden
christina.tannander@tpb.se

Abstract

The Filibuster text-to-speech system was designed and developed specifically to create synthetic speech for use in digital talking books. The aim was to create speech of sufficient pronunciation quality to allow readers to focus on the content of the speech and not on how the words were spoken. The talking books are accessible versions of university textbooks produced for students with print disabilities. Since its introduction in 2007, Filibuster has been used in the production of about 600 digital talking books. Recently, the system has been enhanced in several ways in order to improve the quality of the synthesized speech. One important development was the inclusion of a tagger and a parser in the text preprocessor, allowing for several types of word disambiguation, sentence detection and pause assignment. Speech intelligibility has been increased through the addition of additional phonetically motivated constraints in the speech generation process.

1 Introduction

Filibuster is a text-to-speech system for creating synthetic speech for use in the automatic production of digital talking books. The system been developed at the Swedish Library of Talking Books and Braille (TPB). TPB is a government body that, in collaboration with local libraries, provides access to printed materials for people with print impairment. Among the services provided is the production of accessible versions of textbooks for university students with print disabilities. Using text-to-speech this service could be improved through faster production times and also by including the electronic text with the audio. Also, more books could now be produced within the available budget. Since its introduction in 2007, Filibuster has been used in the production of about 600 digital talking books.

Most commercial text-to-speech systems today are based **unit selection** (Black and Taylor, 1997). Synthesized utterances are generated automatically through careful selection and concatenation of segments from a large collection (corpus) of recorded sentences, thus the name **unit selection**. The segments (units) can vary in length from a single phone to several words. The manuscript for the sentences to be recorded is designed 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.

2 Filibuster system enhancements

The Filibuster system uses the unit selection approach to text-to-speech synthesis. The system has been continuously improved and updated, since its initial deployment, in order to provide increased speech quality, for a wider domain of book topics. Recent improvements include handling of new domains such as religion and law, both with respect to extending the lexicon and creating domain specific rules in the text preprocessor. The Filibuster lexicon is based on conventions from the official pronunciation lexicon *Svenska språknämndens uttalsordbok* (Garlén, 2003). Another improvement was the inclusion of a tagger and a parser (Megyesi, 2002) to help in disambiguation, sentence detection, and pause

generation. The handling of foreign words and names was enhanced through language specific lexica, with adapted pronunciations which attempt to make the words sound somewhat more like Swedish language. For unknown words, which have been disambiguated as **name** or **foreign**, pronunciations are assigned using specific letter-to-sound rules. The quality of the synthesized speech has also been increased thanks to better segment selection mechanisms. Criteria such as lexical stress, syllable structure, and phonetic context are applied to select phones that will fit their position in a given target utterance and which can be joined without any perceptible discontinuity. The Filibuster system was also redesigned in order to be more easily adaptable for new languages in the future.

3 Conclusions

The Filibuster system has seen several important enhancements over the last year. These have in turn improved the quality of the digital talking books that TPB produces. The overall quality of the synthetic speech is high, but there are many areas that need improvement. In particularly, uncommon names and foreign words, which often exhibit unusual phone combinations, pose a problem for current unit selection methods. Also, in the area of prosody very little has been implemented so far.

References

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

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

Megyesi B. 2002. *Shallow parsing with PoS taggers and linguistic features.* Journal of Machine Learning Research, 2:639-668.