

Development of a southern Swedish clustergen voice for speech synthesis

Johan Frid

Centre for Languages and Literature, Lund University

Abstract

This paper describes the development of a speech synthesis voice with a southern Swedish accent. The voice is built for the Festival speech synthesis system using the tools in the festvox suite. The voice type is clustergen, which is a statistical-parametrical synthesis method where parametrical models for phonemes, duration and pitch all are built from a labeled speech database.

Introduction

In recent years, much of the progress within the field of speech synthesis have come within the concatenative paradigm. Corpus-based methods with speech material collected from several thousands of utterances have been dominating the field. This method has reached a high level of naturalness and is widely used in commercial systems today. These systems have a drawback though; they are limited in voice flexibility. Therefore, a recent development is to use corpus methods within parametric synthesis as well. Several techniques have emerged under the name of *statistical-parametrical synthesis methods*. The goal of these methods is to combine the flexibility of parametric synthesis, thus allowing variation in the voice source and the prosody, with the robustness of corpus-based methods.

Statistical-parametrical synthesis methods

The method we will use in this work is called clustergen. The clustergen synthesis method was developed by Alan Black (Black 2006, Black, Zen & Tokuda 2007). The basic idea is to represent speech as MFCCs (Mel Frequency Cepstral Coefficients), then generate the mean of a number of similarly sounding speech segments and finally resynthesizes speech using MLSA (Mel Log Spectrum Approximation, Imai 1983).

Another branch within statistical-parametrical synthesis methods is HMM-based synthesis, e.g. HTS by Tokuda, Zen and Black (2002). An HMM-based system was developed for

Swedish by Lundgren (2005).

Developing a clustergen voice

In this section, we describe the different steps involved in building a clustergen voice. This involves corpus collection and preparation, recording the prompts (sentences) in the corpus, autolabelling the corpus, and finally the actual voice building process.

Corpus development

The clustergen voice building process needs a database with good phonetic coverage. It is also favorable if the sentences to be read does not contain too many uncommon words and are otherwise easy to read. A procedure for finding suitable and phonetically balanced sentences is described in Kominek and Black (2003). The key idea is to, rather than to make up sentence after sentence and in the end hope that you get it right, start with lots of text material and have an automatic procedure look for the right things among your sentences.

The first thing to do is to collect a sufficiently large body of text. We selected the Swedish wikipedia encyclopedia, a version dating from 2007-07-25. This consists of about 600 MB of xml formatted data. After some processing, involving removing tags, captions, headers and more, about 600000 sentences remained. This was then reduced down to around 600 sentences using a festvox script that applies the following criteria:

- each sentence should consist of 4-10 words
- each word should be among the 5000 most frequent
- avoid all pictographic characters (only letters, periods and comma were allowed)
- maximize phonetic coverage by including as many different two-letter sequences as possible

Here are some example sentences:

- Aristoteles ansåg att människor av naturen är politiska varelser.
- Dessa fynd gjordes i Afrika, Asien, Europa och Nordamerika
- Karl Gerhard föddes som Karl Emil Georg Johnson
- Resten av sträckan till Sankt Petersburg är vanlig landsväg
- Säsongen blev mycket framgångsrik och laget vann Stanley Cup
- Efter fem månader stod tyskarna utanför Moskva.

Recording

The sentences were recorded in a quiet office with door closed, using a rather standard headset microphone connected to a laptop. By using a headset the distance from the microphone to the mouth, and hence the recording level, was kept constant. All the sentences were recorded in one session with a very short break about every 50 sentences. For optimal pitch analysis EEG recordings would be desirable. Additional reduction of noise levels would have been achieved in an anechoic chamber. However, the resulting sound quality was found to be sufficient, at least for the research purposes targeted here.

Automatic labeling

The database must be phonemically labeled. This can be done fully automatically if you have a pronunciation lexicon for the words in your sentences.

Defining the phoneset

In order to develop a lexicon, we first need to develop a phone inventory or phone set. Southern Swedish differs from standard Swedish in that retroflexes rarely occur. Otherwise, the phone set included all regularly occurring phonemes in southern Swedish with the addition of a few xenophones (Eklund and Lindström 2001). Here is a summary:

- nine long and nine short vowels
- schwa. This is sometimes used in final unstressed syllables
- consonants: [p t k b d g m n ŋ f s ɛ ʃ h v j l]
- front and back r. Southern Swedish normally has a back r, but some words were foreign place names, which often is pronounced with a front r
- w, also since a few words have english

origins

Lexicon development

Earlier speech synthesis work at the department has used the CTH lexicon (Hedelin, Jonsson and Lindblad 1987), but for the current project we decided not to use it for the following reasons:

- it is not targeted for southern Swedish
- it has a restricted license

Instead, work was started to develop an in-house lexicon from scratch. The 600 sentences contained about 1600 different unique words so the task was not overwhelming. Here are some example entries from the pronunciation lexicon:

- (första (f oe4 r s t a))
- (föddes (f oe3 d e2 s))
- (får (f ao+ r))
- (finland (f i4 n l a n d))
- (där (d ae r))
- (delar (d e+3 l a2 r))
- (båda (b ao+3 d a2))
- (bland (b l a n d))

We use a simple phonetic alphabet where only ASCII characters are allowed in pronunciation entries. This is because its easier to enter these characters and keeps things simple for the computer.

In the example above, the pronunciation entries are not syllabified but this is done automatically later. Prosodic information about vowel length, stress position and word accent is included. The + indicates a long vowel, otherwise all vowels are short by default. The numbers mean:

- 4: main stress, accent 1
- 3: main stress, accent 2
- 2: secondary stress

In monosyllabic words, prosody information is redundant as these are always stressed on the final syllable and have word accent 1.

Note that in southern Swedish, compound words may retain accent 1 if the first element of the compound is an accent 1 word (Bruce 1974, Frid 2003). This is different from other varieties of Swedish where compound words almost invariably has accent 2. Since the lexicon is targeted towards southern Swedish, compound words with accent 1 are given a pronunciation

entry with accent 1. This is easy to modify for other varieties of Swedish, where the main stress accent 1 indicator '4' could be changed into main stress accent 2 indicator '3'.

The parentheses structure seen in the example is the normal format used for festival lexicons.

Doing the labeling

The actual labeling is done through forced alignment. For each sentence, the pronunciation of each word is looked up. This results in a phoneme string. The phoneme strings are then aligned with the utterances using the EHMM procedure in the festvox package.

Building

The voice is constructed by building decision-tree based models from data. Each phoneme is divided into three states in order to handle coarticulation effects. For each state, trees are built for prediction of:

- MFCC
- F0
- duration

In the trees, features such as phonetic context, syllable structure and word position are used. The whole building process is automated and done with tools provided in the festvox package.

Resynthesis

The resynthesis process works as follows: The NLP component produces a phoneme string, where each phoneme again is divided into three phone states. Duration is produced by the duration tree. As we now have temporal information, we can step through the utterance at an interval of, e.g. 5 ms and at every n:th millisecond predict F0 and MFCC parameters from the phone state that is 'active' at the current time frame. Resynthesis is then done through MLSA.

Results

Included in the festvox tools is a script to produce some numerical measurements based on comparisons of synthesized utterances with real utterances. Here are the results:

- all mean 1.78 std 55.09
- F0 mean 8.76 std 261.55

- noF0 mean 0.3 std 0.79
- MCD mean 7.62 std 5.77

The numbers give the mean difference for all features in the parameter vector, for F0 alone, for all but F0, and MCD (mel cepstral distortion).

Notes on building a voice in a new language

The following steps are needed to build a voice in a new language:

- develop a corpus (> 500 prompt sentences)
- record the prompts
- develop a phoneset and a phonetic lexicon for the words in the corpus
- decide on a prosodic model. The default model only differs between stressed and unstressed syllables, but for Swedish we need to handle word accents.

The rest of the process is done through tools in the festvox package. The corpus processing can take some time, especially if you have a large material. Recording can be done in less than a day. The lexicon development can be tedious, but something like 500 words a day is possible. The rest of the voice building is more or less automatic. The labeling takes lots of time, the voice building a little less. However, once the voice is built it can be used instantaneously; it is as fast as any festival voice.

Improvements

The procedure described above is enough to give you a working voice. However, the quality of the voice can be improved in several ways. One easy way is to use more sentences. This will increase the material used in the model building. Another thing is to recheck the labels. The automatic labeling process works rather well; however, there is room for some improvement. We have found instances where the lexicon contains a full pronunciation form but the prompt recording contains a reduced pronunciation. Since the labeling works by forced alignment this may introduce errors. It would also be interesting to explore more elaborate prosodic modeling, for instance feet structure. Currently there is no support for this in festival.

References

- Black, A. (2006), CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling, Interspeech 2006 - ICSLP, Pittsburgh, PA.
- Black, A., Zen, H., and Tokuda, K., (2007) Statistical Parametric Synthesis, ICASSP 2007, Hawaii.
- Bruce, G., (1974) Tonaccentregler för sammansatta ord i några sydsvenska stadsmål. In Platzac, C., editor, *Svenskans beskrivning*, number 8, pages 62-75.
- Eklund, R., and Lindström, A., (2001) Xenophones: An Investigation of Phone Set Expansion in Swedish and Implications for Speech Recognition and Speech Synthesis. *Speech Communication* 35, vols. 1-2, pp. 81-102.
- Frid, J., (2003) *Lexical and Acoustic Modelling of Swedish Prosody*, Department of Linguistics and Phonetics, Lund University.
- Hedelin, P., Jonsson, A., and Lindblad, P., (1987) *Svenskt uttalslexikon: 3 ed.* Tech Report, Chalmers University of Technology.
- Imai, S., (1983) Cepstral analysis/synthesis on the Mel frequencyscale," in ICASSP-83, Boston, MA, 1983, pp. 93-96.
- Kominek, J. and Black, A. (2003) CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute, Tech Report CMU-LTI-03-177
- Lundgren, A. (2005) HMM-baserad talsyntes. Master's Thesis.
- Tokuda, K., Zen, H., and Black, A. (2002) An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sept. 2002.