DEVELOPMENT OF PROSODIC MODELS FOR A SOUTHERN SWEDISH CLUSTERGEN VOICE FOR SPEECH SYNTHESIS

Johan Frid
Centre for Languages and Literature, Lund University

1 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 voice source and prosody, with the robustness of corpus-based methods. In this paper we describe our efforts at building such a voice, and, in particular, our recent efforts at improving the intonation model for this voice.

1.1 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, then generate the mean of a number of similarly sounding speech segments and finally resynthesizes speech using MLSA (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).

2 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.

2.1 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.

2.2 Recording

The sentences were recorded in a quite office with door closed, using a rather standard headset microphone connected to a laptop. 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.

2.3 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. Also note that Anumanchipalli, G., Prahallad, K. and Black A. (2008) use letters directly. This approach may be interesting for Swedish.

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 ɕ ɧ 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 pronounces 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.

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.

2.4 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.

3 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. Resyntesis is then done through MLSA.

4 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).

5 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.

6 Intonation modeling with QTA

Recent work has been focussed on using the Target Approximation model (Xu & Wang 2001) in its quantified version - QTA (Prom-on et al 2006a; 2006b and submitted) - for intonation. QTA is a model where the F0 curve is viewed as a syllable-based approximation towards a linear target. It may be used to parameterize the F0 curve (per syllable) in terms of:

- strength (how fast targets are approximated)
- slope of the target
- height of the target

In general terms, finding the parameters of the QTA model is done by defining a function with parameters that describes an F0 contour in terms of the model and then letting a mathematical algorithm finding the parameter values that makes the function produce the contour that is most similar to a naturally occuring F0 contour. Finding parameters is done with the Levenberg-Marqardt algorithm as implemented in the **fityk** curve fitting and data analysis program (http://www.unipress.waw.pl/fityk/). Figure 1 shows an example of the model at work. In Figure 1, panel (a) shows the original F0 contour, panel (b) shows the synthesized contour given the model and the parameters found by the curve fitting method, and panel (c) shows the corresponding targets, per syllable. We follow Prom-on et al (submitted) in that we impose a restriction on the height of the pitch target, which follows from findings that the syllable offset is where the surface F0 becomes closest to the target (Xu & Wang 2001). In the current version we actually force the height to be the same as the original F0 value, which may be an over-simplification.

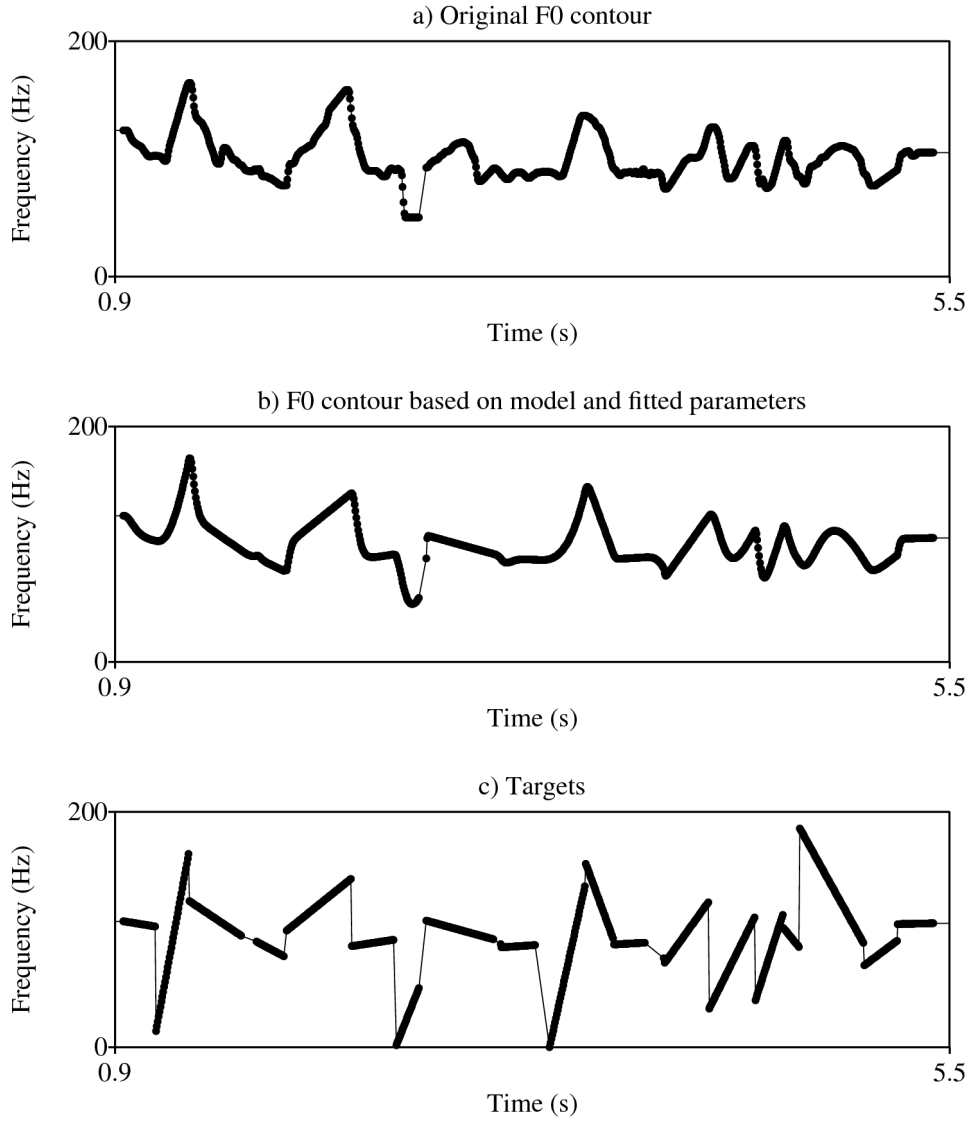

Figure 1. The figure shows a) the original pitch contour b) the pitch contor as reconstructed by the model and fitted parameters per syllable c) the targets per syllable.

6.1 An experiment

A small experiment was conducted to examine the ability of the model to capture one of the most relevant characteristics of intonation in Swedish: the word accents. We used the ~600 sentences described above This totalled a number of

8856 syllables. The F0 curves were parameterized (per syllable) by the QTA model and furthermore classified according to word accent and stress level into four categories:

- accent 1
- accent 2
- secondary stress
- unstressed

Results

We have only had time to perform a small, preliminary check of the results by taking the trimmed mean. The trimmed means were calculated by, after sorting the measurement values numerically by each parameter, removing the top and bottom 25% of the data. The results are presented in Table 1.

Table 1. Results for **slope** and **strength** parameters per category.

	(trimmed means)		
	slope	strength	n
unstressed	3	89	2855
sec. stress	-48	101	1225
accent 1	-74	70	3483
accent 2	-11	58	1293

Analysis

The major finding is that we indeed find a large difference between *accent 1* and *accent 2* for the **slope** parameter. This suggests that the QTA model is able to capture this distinction. Note also that the category *unstressed* has a **slope** value close to 0. This could be interpreted as rather flat pitch contour is preferred for this category.

7 Summary and outlook

We have presented a preliminary version of a Swedish clustergen voice for use with the festival speech synthesis system. We have also described initial work on using the QTA model. Our future goals include expanding the model with more categores (e.g. focus) and implement the model in the festival/clustergen context.

References

Anumanchipalli, G., Prahallad, K. and Black A. (2008) Significance of Early Tagged Contextual Graphemes in Grapheme Based Speech Synthesis and Recognition Systems, *ICASSP2008*, Las Vegas, NV.

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.

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.

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 frquency scale, 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.

Prom-on, S., Xu, Y. and Thipakorn, B. (2006a). Functional-oriented articulatory modeling of tones and intonations. In *Proceedings of Speech Prosody* 2006, Dresden, Germany. PS2-14_0089.

Prom-on, S., Xu, Y. and Thipakorn, B. (2006b). Quantitative Target Approximation model: Simulating underlying mechanisms of tones and intonations. In *Proceedings of The 31st International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France. pp. I-749-752.

Prom-on, S., Xu, Y. and Thipakorn, B. (submitted) Modeling tone and intonation in Mandarin and English as a process of target approximation.

Tokuda, K., Zen, H., and Black, A. (2002) An HMM-based speech synthesis system applied to English, *Proc. of 2002 IEEE SSW*, Sept. 2002.

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33: 319-337.