Simulating Intonation in Regional Varieties of Swedish

Susanne Schötz[1], Jonas Beskow[2], Gösta Bruce[1], Björn Granström[2], Joakim Gustafson[2]

[1]Dept. of Linguistics & Phonetics, Centre for Languages & Literature, Lund University, Sweden
[2]Dept. of Speech, Music & Hearing, School of Computer Science & Communication, KTH, Sweden
{gosta.bruce, susanne.schotz}@ling.lu.se, {bjorn, jocke}@speech.kth.se, beskow@kth.se

Abstract

Within the research project SIMULEKT (Simulating Intonational Varieties of Swedish), our recent work includes two approaches to simulating intonation in regional varieties of Swedish. The first involves a method for modelling intonation using the SWING (SWedish INtonation Generator) tool, where annotated speech samples are resynthesised with rule-based intonation and audio-visually analysed with regards to the major intonational varieties of Swedish. The second approach concerns a method for simulating dialects with HMM synthesis, where speech is generated from emphasis-tagged text. We consider both approaches important in our aim to test and further develop the Swedish prosody model, as well as to convincingly simulate Swedish regional varieties using speech synthesis.

Index Terms: Swedish dialects, prosody, speech synthesis

1. Introduction

Our object of study in the research project SIMULEKT (Simulating Intonational Varieties of Swedish) [1] is the prosodic variation characteristic of different regions of the Swedish-speaking area. Figure 1 shows a map of these regions, corresponding to our present dialect classification scheme. In our work, various forms of speech synthesis and the Swedish prosody model [2, 3, 4] play prominent roles. Our main sources for analysis here are the two Swedish speech databases SpeechDat [6] and the NST corpus (see further section 3.1). SpeechDat contains speech recorded over the telephone from 5000 speakers, registered by age, gender, current location and self-labeled dialect type, according to Elert's suggested Swedish dialect groups [7] that is a more fine-grained classification with 18 regions in Sweden. This material was used in our first approach together with our SWING tool. The HMM material used in our second approach was selected from the NST database for training of speech recognition, which covers some major regional varieties of Swedish (see also section 3.1). The large speech synthesis database from a professional speaker of standard Swedish also used in the second approach was recorded as part of the NST (Nordisk Språkteknologi 'Nordic Language Technology') synthesis development.

1.1. The Swedish prosody model

The main parameters for the Swedish prosody model [2, 3, 4] are for word prosody 1) word accent timing, i.e. timing characteristics of pitch gestures of word accents (accent 1/accent 2) relative to a stressed syllable, and 2) pitch patterns of compounds, and for utterance prosody 3) intonational prominence levels (focal/non-focal accentuation), and 4) patterns of concatenation between pitch gestures of prominent words.

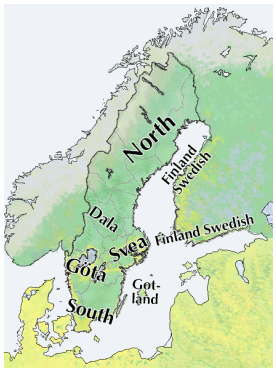

Figure 1: *Approximate geographical distribution of the seven main regional varieties of Swedish.*

1.2. Outline of the paper

This paper exemplifies two recent approaches involving simulation of Swedish regional varieties: one analysis tool for testing and further developing our prosody model using rule-based intonation resynthesis, and one HMM synthesis approach for simulating dialects, where speech is generated from emphasis-tagged text.

2. The SWING intonation analysis tool

SWING (SWedish INtonation Generator) is a tool for analysis and modelling of Swedish intonation by resynthesis, developed within our project. It comprises several parts joined by the speech analysis software Praat [8], which also serves as graphical interface. Using an input annotated speech sample and an input rule file, SWING generates and plays PSOLA resynthesis – with rule-based and speaker-normalised intonation – of the input speech sample. Additional features include visual display of the output on the screen, and options for printing various kinds of information to the Praat console (Info window), e.g. rule names and values, the time and F_0 of generated pitch points etc. Figure 2 shows a schematic overview of the tool.

2.1. Input speech material

The input speech sample to be used with the tool is manually annotated. Stressed syllables are labelled prosodically and the corresponding vowels are transcribed orthographically. Figure 3 displays an example utterance with prosodic annotation: *De' på kvällarna som vi sänder* 'It's in the evenings that we are transmitting', while Table 1 shows the prosodic labels that are handled by the current version of the tool.

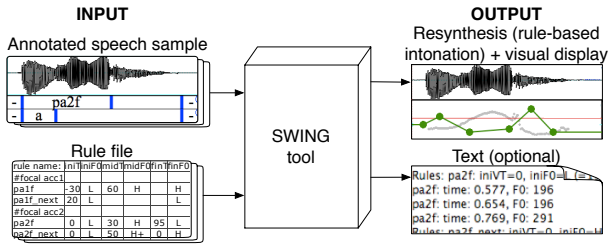

Figure 2: *Schematic overview of the* SWING *tool.*

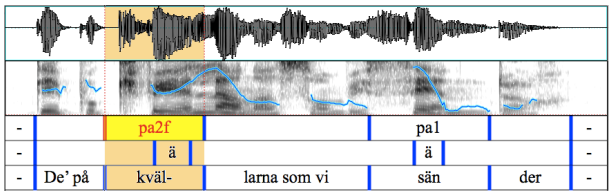

Figure 3: *Example of an annotated input speech sample.*

Table 1: *Prosodic labels used for annotation of speech samples to be analysed by* SWING.

Label	*Description*
pa1	primary stressed (non-focal) accent 1
pa2	primary stressed (non-focal) accent 2
pa1f	focal focal accent 1
pa2f	focal focal accent 2
cpa1	primary stressed accent 1 in compounds
cpa2	primary stressed accent 2 in compounds
csa1	secondary stressed accent 1 in compounds
csa2	secondary stressed accent 2 in compounds

2.2. Rules

The Swedish prosody model is implemented as a set of rule files – one for each regional variety in the model – with timing and F_0 values for critical points in the rules. These files are text files with a number of columns; the first contains the rule names, and the following comprise three pairs of values, corresponding to the timing and F_0 of the critical pitch points of the rules. The three points are called *ini* (initial), *mid* (medial), and *fin* (final). Each point contains values for timing (*T*) and F_0 (*F0*). Timing is expressed as a percentage into the stressed syllable, starting from the onset of the stressed vowel, which is the default. Three values are used for F_0: L (low), H (high) and H+ (extra high, used in focal accents). The pitch points are optional; they can be left out if they are not needed by a rule. New rules can easily be added and existing ones adjusted by editing the rule file. Table 2 shows an example of the rules for South Swedish. Several rules contain a second part, which is used for the pitch contour of the following (unstressed) interval (segment) in the annotated input speech sample. This extra part has '*_next*' attached to its rule name. Examples of such rules are *pa1f*, *pa2f* and *cpa2* in Table 2.

2.3. Procedure

Analysis is fairly straightforward with SWING. The user selects one input speech sample and one rule file to be used with the tool, and which (if any) text (rules, pitch points, debugging information) to print to the Praat console. A Praat script generates resynthesis of the input speech sample with a rule-based output pitch contour based on 1) the pitch range of the input speech sample, used for speaker normalisation, 2) the annotation, used to find the time and pitch gestures to be generated, and 3) the rule file, containing the values of the critical pitch points. The Praat graphical user interface provides immediate audio-visual feedback of how well the rules work, and also allows for easy additional manipulation of pitch points with the Praat built-in *Manipulation* feature.

Table 2: *Example rule file for South Swedish with timing (T) and F_0 (F0) values for initial (ini), mid (mid) and final (fin) points.*

Rule name	*iniT*	*iniF0*	*midT*	*midF0*	*finT*	*finF0*
global (phrase)		L				L
concatenation		L				L
pa1f (focal accent 1)	-30	L	40	H+		
pa1f_next (extra gesture)		L				
pa2f (focal accent 2)		L	40	L	80	H+
pa2f_next (extra gesture)	10	H+	50	L		L
pa1 (non-focal accent 1)	-30	L	40	H		L
pa2 (non-focal accent 2)		L				H
pa2_next (extra gesture)	30	H			80	L
cpa1 (compound accent 1)	-30	L	20	H+	80	L
cpa2 (compound accent 2)		L				H
cpa2_next (extra gesture)	30	H			80	L

2.4. Testing the Swedish prosody model

SWING is now being used in our work with testing and developing the Swedish prosody model for compound words. Testing is done by selecting an input speech sample and a rule file of the same intonational variety. If the model works adequately, there should be a close match be-tween the F_0 contour of the original version and the rule-based one generated by the tool. Figure 4 shows compound intonation for the three dialect regions Gotland, Svea (Central standard Swedish), and South Swedish.

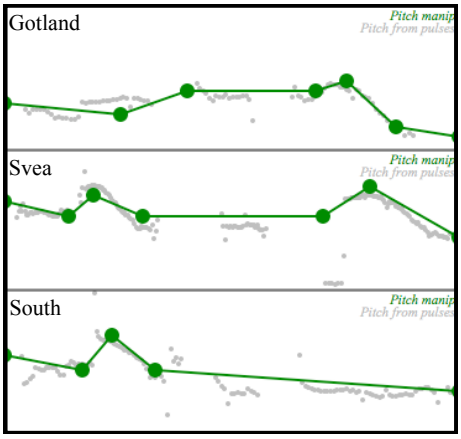

Figure 4: *Simulation of compound words in* SWING. *Praat Manipulation displays of the compound word* mo'biltele,fonen *'the mobile phone' of the three dialect regions Gotland, Svea (Central standard Swedish), and South Swedish (simulation: circles connected by solid line; original pitch: light-grey line).*

As can bee seen in Figure 4, there is a close match between the original pitch of the input speech samples and the simulated pitch contour in all three dialectal regions.

3. Synthesis Experiments

During the last decade, most speech synthesisers have been based on prerecorded pieces of speech resulting in improved quality, but with lack of control in modifying prosodic patterns [9]. The research focus has been directed towards how to optimally search and combine speech units of different lengths. A synthesis approach that has gained interest in recent years is HMM based synthesis [10]. In this solution the generation of speech is based on a parametric representation, while the grapheme-to-phoneme conversion still relies on a large pronunciation dictionary. This approach has been successfully applied to a large number of languages, including Swedish [11].

HMM synthesis is an entirely data-driven approach to speech synthesis. As such it gains all its knowledge about segmental, intonational and durational variation in speech from training on an annotated speech corpus. Given that the appropriate features are annotated and made available to the training process, it is possible to synthesise speech with high quality at both segmental and prosodic levels. Another important feature of HMM synthesis that makes it an interesting choice in studying dialectal variation, is that it is possible to adapt a voice trained on a large data set (2-10 hours of speech) to a new speaker with only 15-30 minutes of transcribed speech [12]. In this study we used 20-30 minutes of dialectal speech for experiments on speaker adaption of the initially trained HMM synthesis voice.

3.1. Data description

The data we used in this study are from the Norwegian Språkbanken. This large speech synthesis database from a professional speaker of standard Swedish was recorded as part of the NST (Nordisk Språkteknologi 'Nordic Language Technology') synthesis development. It was recorded in stereo, with the voice signal in one channel, and signal from a laryngograph in the second channel. About 5000 read sentences are included in the corpus, adding up to about 11 hours of speech. The manuscripts for the recordings were based on the NST corpus, and the selection was done to make them phonetically balanced and to ensure diphone coverage. Though not prosodically balanced, the manuscripts still contain different types of sentences that ensure prosodic variation, e.g. statements, wh-questions, yes/no questions and enumerations. The 11 hour speech database was aligned on the phonetic and word levels using our Nalign software [13] with the NST dictionary as pronunciation dictionary. This comprises more than 900.000 phonetically transcribed items with syllable boundaries marked. In addition, the text was tagged for part-of-speech using a TNT tagger trained on the SUC corpus [14]. From the NST database for training of speech recognition we selected a small number of unprofessional speakers from the following Swedish dialectal areas: North, Dala, Göta, Gotland and South (see Figure 1). The data samples were considerably smaller than the speech synthesis database; they ranged from 22 to 60 minutes, compared to the 11 hours by the professional speaker.

3.2. HMM Contextual Features

The typical HMM synthesis model can be decomposed into a number of distinct layers. At the acoustic level, a parametric source-filter model (MLSA-vocoder) is responsible for signal generation. Context dependent HMMs, containing probability distributions for the parameters and their 1st and 2nd order derivatives, are used for generation of control parameter trajectories. In order to select context dependent HMMs, a decision tree that uses input from a large feature set to cluster the HMM models was applied.

In this study, we used the standard model for acoustic and HMM level processing, and we focussed on on adapting the feature set for the decision tree for the task of modeling dialectal variation. The feature set typically used in HMM synthesis includes features on segment, syllable, word, phrase and utterance level. Segment level features include immediate context and position in syllable; syllable features include stress and position in word and phrase; word features include emphasis, part-of-speech tag (content or function word), number of syllables, position in phrase etc., phrase features include phrase length in terms of syllables and words; utterance level includes length in syllables, words and phrases. For our present experiments, we have also added a speaker level to the feature set, since we train a voice on multiple speakers. The only feature in this category at present is dialect group, which is one of North, Dala, Svea, Göta, Gotland and South. In addition to this, we have chosen to add to the word level a morphological feature stating whether or not the word is a compound, since compound stress pattern often is a significant dialectal feature in Swedish [1]. At the syllable level we have added explicit information about lexical accent type (accent 1, accent 2 or compound accent).

4. Exemplifying work in progress

The SWING tool requires information about phoneme alignment, pitch range, syllable stress and accents. These features are all automatically generated in the HMM synthesis process, which makes it possible to use SWING rules to generate pitch contours automatically from an emphasis-tagged text, which in turn can be used to replace or supplement the HMM-generated pitch curves prior to sound synthesis.

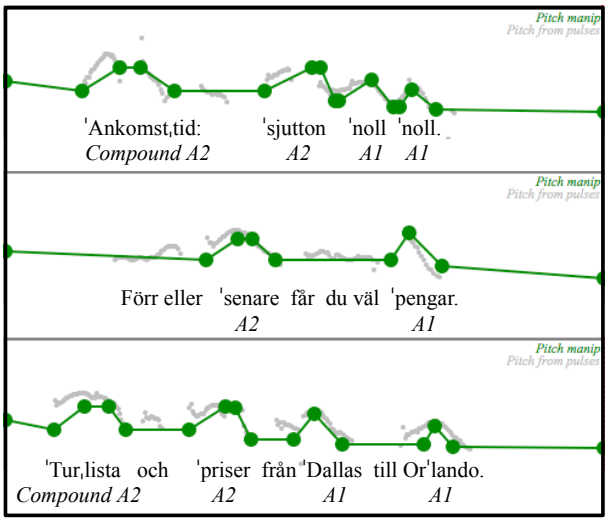

Figure 5: SWING *rules for South Swedish applied to the HMM material. Praat Manipulation displays of the three phrases* Ankomsttid: sjutton noll noll. *'Time of arrival: seventeen zero zero.',* Förr eller senare får du väl pengar. *'You should get money sooner or later.' and* Turlista och priser från Dallas till Orlando *'Timetable and prices from Dallas to Orlando'. (simulation: circles connected by solid line; original pitch: light-grey line; A1: accent 1; A2: accent 2).*

Current work in our project concerns using the rules obtained with SWING to generate intonation for the seven main regional varieties of Swedish together with the HMM synthesiser. As an example of how the new hybrid SWING/HMM synthesiser works, Figure 5 shows how South Swedish SWING rules can be applied to the material from the HMM synthesis. There is a rather close match between the original F_0-contour and the synthesis. One exception is phrase-initial intonation, which has not been regulated yet.

The content-rich feature files generated from large annotated speech corpora, used in HMM synthesis training, also allow for statisical and explorative investigation of prosodic characteristics of different speakers and dialects. Figure 6 visualises F_0-patterns for two speakers with different dialects as *pitch clouds*. We selected approximately 1000 content words, ranging from 1 to 5 syllables, with primary stress on the first syllable, from a large set of read utterances. F_0-curves were extracted, mean-normalised and temporally aligned according to vowel onset in the stressed syllable (marked with a vertical line in the figure). For each dialect, separate clouds were generated for three accent types: accent 1, accent 2 and compounds. The figure clearly shows the dialect difference in accent 2 and compounds, with two peaks in the *Svea* case and a single peak for *South*. For *South* it is clear that the temporal alignment of the peak is later in accent 2 than in accent 1. An additional dimension in the figure is syllable length, which is represented by color. Monosyllabic words are black, 2-5 syllable words are red, green, blue and magenta respectively. Not unexpectedly, there is an overrepresentation of monosyllabic accent 1 words, since we selected only those with stress on the first syllable. Accent 2 words are primarily disyllabic, while a majority of the longer words are compounds. This type of analysis gives insight into features that influence prosodic realisations, which is valuable both in HMM synthesis and for fine-tuning the SWING rules.

5. Discussion and future work

Although SWING still needs work, we already find it useful in our project work of analysing speech material as well as testing our model. Our preliminary simulation of compound word intonation for Gotland, Svea (Central standard Swedish) and South Swedish with the tool is also encouraging. The new hybrid SWING/HMM synthesiser will allow more careful investigation of the SWING rules, since large sets of perceptual stimuli can be automatically generated under controlled conditions.

6. Acknowledgements

This work is supported by a grant from the Swedish Research Council.

7. References

[1] Bruce, G., Granström, B., Schötz, S., 2007. Simulating Intonational Varieties of Swedish. *Proc. of ICPhS XVI*, Saarbrücken, Germany.

[2] Bruce, G.; and Gårding, E., 1978. A prosodic typology for Swedish dialects. In *Nordic Prosody*, E. Gårding; G. Bruce; R. Bannert (eds.). Lund: Department of Linguistics, 219-228.

[3] Bruce, G. and Granström, B., 1993. Prosodic modelling in Swedish speech synthesis. A prosodic typology for Swedish dialects. *Speech Communication* 13, 63-73.

[4] Bruce, G., 2007. Components of a prosodic typology of Swedish intonation. In *Tones and Tunes*, Volume 1, T. Riad; C. Gussenhoven (eds.). Berlin: Mouton de Gruyter, 113-146.

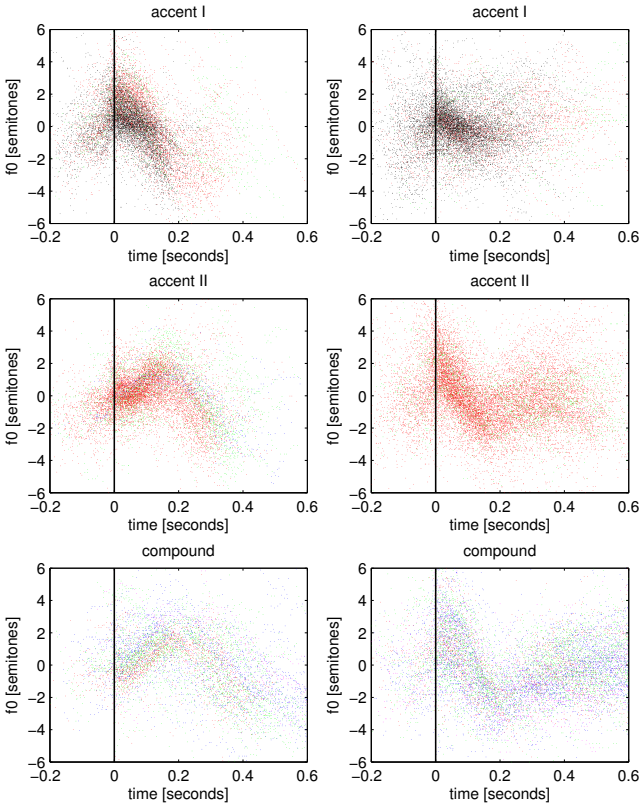

Figure 6: Pitch clouds for *South* (left) and *Svea* (right) dialects.

[5] Engstrand, O., Bannert, R., Bruce, G., Elert, C-C. and Eriksson, A., 1997. Phonetics and phonology of Swedish dialects around the year 2000: a research plan. *Papers from FONETIK 98, PHONUM* 4. Umeå: Department of Philosophy and Linguistics, 97-100.

[6] Elenius, K., 1999. Two Swedish SpeechDat databases - some experiences and results. *Proc. of Eurospeech 99*, 2243-2246.

[7] Elert, C.-C., 1994. Indelning och gränser inom området för den nu talade svenskan - en aktuell dialektografi. In *Kulturgränser - myt eller verklighet.*, Edlund, L.E. (Ed.). Umeå, Sweden: Diabas, 215-228.

[8] Boersma, P. and Weenink, D., 2009. *Praat: doing phonetics by computer (version 4.6.17)* [computer program]. http://www.praat.org/, visited 12-Oct-09.

[9] Taylor, P. (2009). Text-To-Speech Synthesis. Cambridge University Press.

[10] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. *Proc. of CASSP 2000*, 1315-1318.

[11] Lundgren, A. (2005). HMM-baserad talsyntes. Master's thesis, KTH, TMH, CTT.

[12] Watts, O., Yamagishi, J., Berkling, K., & King, S. (2008). HMM-Based Synthesis of Child Speech. *Proc. of The 1st Workshop on Child, Computer and Interaction*.

[13] Sjölander, K., & Heldner, M. (2004). Word level precision of the NALIGN automatic segmentation algorithm. *Proc. of The XVIIth Swedish Phonetics Conference, Fonetik 2004*, Stockholm University, pp. 116-119.

[14] Megyesi, B. (2002). Data-Driven Syntactic Analysis - Methods and Applications for Swedish. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.