# Synthesising intonational varieties of Swedish

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

*[1]Dept. of Linguistics & Phonetics, Centre for Languages & Literature, Lund University*
*[2]Dept. of Speech, Music & Hearing, School of Computer Science & Communication, KTH*

### *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 modeling intonation using the* SWING *(*SW*edish* IN*tonation* G*enerator) 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.*

## Background

Our object of study in the research project SIMULEKT (Simulating Intonational Varieties of Swedish) (Bruce et al., 2007) is the prosodic variation characteristic of different regions of the Swedish-speaking area. The SIMULEKT project, supported by the Swedish Research Council 2007-2010, is a collaboration between Linguistics, Lund University, and Speech, Music, Hearing, KTH, Stockholm. The primary goal of the project is to gain more precise knowledge about the major intonational varieties of Swedish. A concomitant goal is to develop the Swedish prosody model theoretically and experimentally. In addition to the regular description of intonational patterns from studying $F_0$contours, the use of speech synthesis in different forms is a major feature of our research project. Our starting-point is that prosody and specifically intonation is a fundamental constituent of the different, native accents characterising the distinct regional varieties of Swedish. We believe that studying the pitch patterns of different varieties of a language like Swedish contrastively will sharpen our analysis and description of their intonation. The main regional varieties or dialect groups of Swedish are South, Göta, Svea, Dala, Gotland, North, and Finland Swedish. They are also considered to be the major intonational varieties of Swedish. Figure 1 shows a map of these regions, corresponding to our present dialect classification scheme.

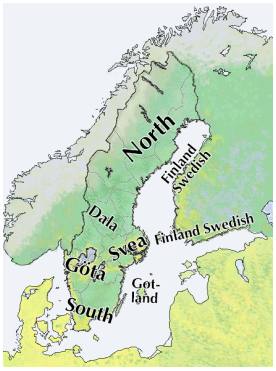


*Figure 1: Approximate geographical distribution of the seven main regional varieties of Swedish.*

## The Swedish prosody model

The main parameters of the Swedish prosody model (Bruce and Gårding, 1978; Bruce and Granström, 1993; Bruce, 2007) are for word prosody 1) word accent timing, i.e. timing characteristics of pitch gestures of word accents (accent1/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. Thus the specific timing of pitch patterns of word accents in relation to stressed syllables (both primary and secondary), i.e. in simplex as well as in compound words, is considered to be a potential distinctive feature for the classification of different intonational varieties. Major emphasis in the SIMULEKT project

has been given to postlexical prosody, specifically phrase and utterance intonation, assuming intonational prominence levels and tonal concatenation patterns to be particularly revealing for the impression and identification of different native accents of Swedish. There are regional varieties of Swedish, which can be shown to be exploiting two levels of intonational prominence (focal vs. non-focal accentuation), specifically the Central Swedish regional varieties Svea and Göta. In these varieties there will be a fairly regular alternation between focal and non-focal accentuation for successive accented words of a phrase. Other varieties have more equal weighting between such words of a phrase and regularly exploit only one level of intonational prominence as a characteristic feature, namely the geographically more peripheral varieties South, Gotland, North, and also Finland Swedish as well as Dala. We have also been emphasising patterns of tonal concatenation, both as coherence patterns between prominent words – high/low plateau, downslope/upslope – and as boundary signalling. These patterns would also appear to contribute to our differentiating between distinct varieties of Swedish.

# Methodology

## Speech databases

Our main sources for analysis here are the three Swedish speech databases SweDia 2000, SpeechDat and NST (Nordisk Språkteknologi 'Nordic Language Technology').

The SweDia 2000 database comprises a word list, an elicited prosody material, and spontaneous monologues from 12 speakers (younger and elderly men and women) each from more than 100 different places in Sweden and Swedish-speaking parts of Finland, selected for dialectal speech. From this database a minor elicited prosody material and primarily the extensive spontaneous speech material are relevant for our project work.

SpeechDat (Elenius, 1999) 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 (Elert, 1994) that is a more fine-grained classification with 18 regions in Sweden. This database contained two particularly interesting read sentences to our project: *Mobiltelefonen är nittiotalets stora fluga, både bland företagare och privatpersoner.* 'The mobile phone is the big hit of the nineties, both among business people and private persons' and *Flyget, tåget och bilbranschen tävlar om lönsamhet och folkets gunst* 'Airlines, train companies and the automobile industry are competing for profitability and people's appreciation'.

The main data used in our HMM approach are from the Norwegian Språkbanken. This large speech synthesis database from a professional speaker of standard Swedish was recorded as part of the NST synthesis development. About 5000 read sentences are included in the corpus, adding up to about 11 hours of speech.

## Analysing intonation with SWING

An important part of our work concerns analysis and modeling of Swedish intonation by resynthesis. The SWING (SWedish INtonation Generator) tool was developed for this task. It comprises several parts joined by the speech analysis software Praat (Boersma and Weenink, 2010), 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.

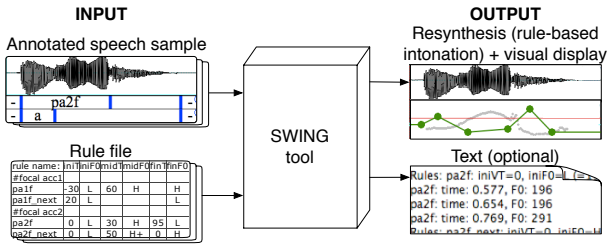


*Figure 2: Schematic overview of the* SWING *tool.*

### 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: *tävlar om lönsamhet och folkets gunst* 'are competing for profitability and people's appreciation', while Table 1 shows the prosodic labels that are handled by the current version of the tool.

### 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

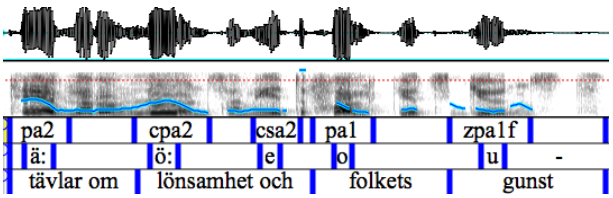


*Figure 3: Example of an annotated input speech sample.*

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

| *Label* | *Description* |
|---|---|
| pa1 | primary stressed (pre-focal) accent 1 |
| pa1f | focal focal accent 1 |
| pa1po | primary stressed (post-focal) accent 1 |
| zpa1f | finally stressed focal accent 1 |
| pa2 | primary stressed (non-focal) accent 2 |
| pa2f | focal focal accent 2 |
| cpa | primary stressed (ps) compound accent |
| csa | secondary stressed (ss) compound accent |
| cpaf | focal ps compound accent |
| csaf | focal ss compound accent |
| cpapo | post-focal ps compound accent |
| csapo | post-focal ss compound accent |

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 *i* (initial), *m* (medial), and *f* (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 e.g. 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 Gothenburg (Göta) Swedish, adjusted in accordance with the Swedish prosody model and Segerup (2005). Several rules change values when applied to phrase-final position (see *(pf)* in Table 2), and some 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 '*_n*' (n stands for next segment, the domain of the extra gesture) attached to its rule name. Examples of such rules are *pa1f*, *zpa1f*, *pa2f* and *csaf* in Table 2.

*Table 2: Example rule file for Gothenburg (Göta) Swedish with timing (T) and $F_0$ (F0) values for initial (i), mid (m) and final (f) points. Values in italics within parentheses are phrase-final (pf) rules. For rule name abbreviations, see Table 2.*

| *Rule name* | iT/*(pf)* | iF0/*(pf)* | mT/*(pf)* | mF0/*(pf)* | fT/*(pf)* | fF0/*(pf)* |
|---|---|---|---|---|---|---|
| global | | L | | | | L |
| pa1 | -40 | L | | H | 70 | L |
| pa1f | -40 | L | | H | 70 | L |
| pa1f_n | 20 | L | | */(60)* | */(H+)* | |
| pa1po | | | | H+ | 70 | L |
| zpa1f | -50 | L | -20 | H | 30 | L |
| zpa1f_n | -20 | H+ | | | | |
| pa2 | -50/*(0)* | L/*(H+)* | */(100)* | H+/*(L)* | */(180)* | L/*(H)* |
| pa2f | -40 | L | 20/*(0)* | H+/*(H)* | */(70)* | L |
| pa2f_n | 40 | L | | | 70/*(90)* | H+/*(H)* |
| cpa | -20 | L | 20 | H | | L |
| csa | | | | | | L |
| cpaf | -20 | L | 20 | H+ | | L |
| csaf | | | | | | L |
| csaf_n | */(20)* | */(L)* | | | */(80)* | */(H+)* |
| cpapo | | | 20 | H+ | | L |
| csapo | | | | | | L |

## Procedure

Analysis with SWING is fairly straightforward. 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.

## Testing the Swedish prosody model

SWING has been used in our work with testing and developing the Swedish prosody model for simplex and compound words as well as phrasing. 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 between the $F_0$ contour of the original version and the rule-based one generated by the tool. Figure 4 shows example SWING output of three phrases for the two intonational varieties Göta and South Swedish. As can bee seen there is a close match between the original pitch of the input speech samples and the simulated pitch contour in all phrases.

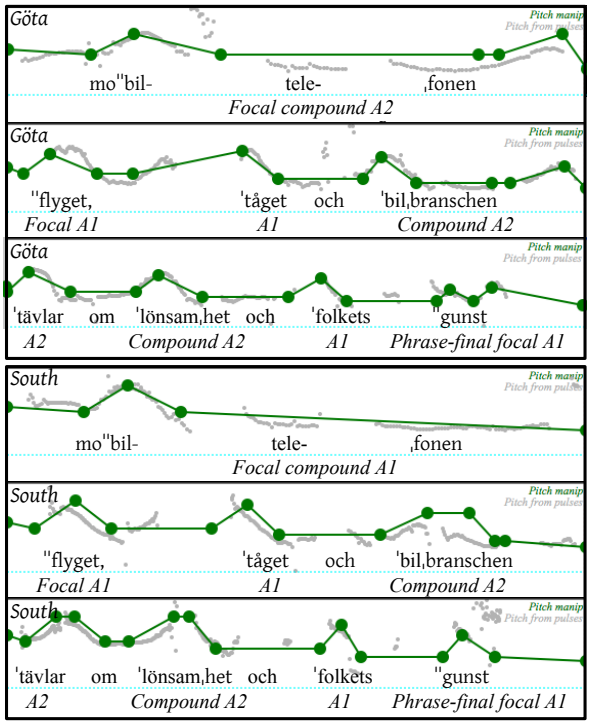


*Figure 4: Simulation examples in* SWING. *Three phrases in Göta and South Swedish (simulation: circles connected by solid line; original pitch: light-grey line; A1: accent 1; A2: accent 2).*

## HMM synthesis

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 (Taylor, 2009). 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 (Tokuda et al., 2000). 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 (Lundgren, 2005).

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 (Watts et al., 2008). In this study we used 20-30 minutes of dialectal speech for experiments on speaker adaption of the initially trained HMM synthesis voice. The data we used in this study are from the Norwegian Språkbanken (see Section Speech Databases).

### Data description

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 (Sjölander and Heldner, 2004) 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 (Megyesi, 2002). 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.

### 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 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

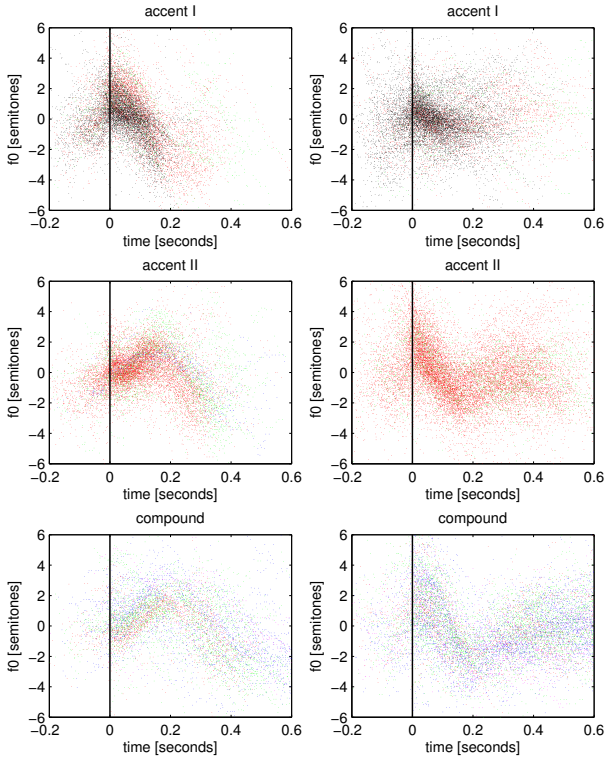


*Figure 5: Pitch clouds for* South *(left) and* Svea *(right) dialects.*

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 is often a significant dialectal feature in Swedish (Bruce et al., 2007). At the syllable level we have added explicit information about lexical accent type (accent 1, accent 2 or compound accent).

### Data-driven exploration of intonation

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 5 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.

# Synthesis of our approaches

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.

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. We have integrated SWING into the HMM synthesis framework, so that it may be driven by the same input features as the HMM synthesis, and the durations generated by the HMMs trained on a speaker of the target dialect. This makes it possible to replace or adapt the HMM-generated $F_0$ track by the one generated by SWING before rendering the waveform. As an example of how the new hybrid SWING/HMM synthesiser works, Figure 6 shows the $F_0$ tracks generated by the two systems for South Swedish SWING.

The new hybrid HMM/SWING synthesiser will allow more careful investigation of the SWING rules, since large sets of perceptual stimuli can be automatically generated under controlled conditions.

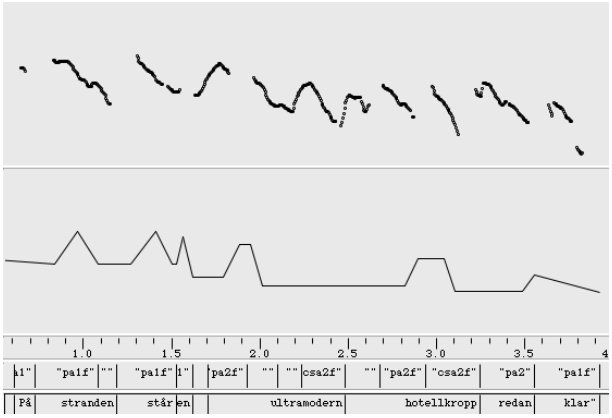


*Figure 6: Log $F_0$ curves generated by HMM (top) and* SWING *(bottom) for the* South *dialect.*

# Discussion and future work

We are planning to run some perceptual testing to see whether listeners will be able to reliably identify a speaker as belonging to the different regional varieties of Swedish depending on the specific pitch shapes of target utterances. We will also perform listening tests comparing different methods of merging the data-driven and rule-generated $F_0$ tracks in the hybrid HMM/SWING synthesiser.

# Acknowledgement

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

# References

Boersma P and Weenink D (2010). Praat: doing phonetics by computer (version 4.6.17) [computer program]. Webpage http://www.praat.org/, visited 14-March-10.

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

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

Bruce G and Granström B (1993). Prosodic modelling in Swedish speech synthesis. a prosodic typology for Swedish dialects. *Speech Communication*, 63–73.

Bruce G, Granström B and Schötz S (2007). Simulating intonational varieties of Swedish. In *Proc. of ICPhS XVI, Saarbrücken, Germany*.

Elenius K (1999). Two Swedish SpeechDat databases - some experiences and results. In *Proc. of Eurospeech 99*, vol. 4, 2243–2246.

Elert C C (1994). Indelning och gränser inom området för den nu talade svenskan - en aktuell dialektografi. In L Edlund, ed., *Kulturgränser - myt eller verklighet.*, vol. 1, 215–228. Umeå, Sweden: Diabas.

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

Megyesi B (2002). *Data-Driven Syntactic Analysis - Methods and Applications for Swedish*. Ph.D. thesis, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.

Segerup M (2005). The interaction of word accent and quantity in Gothenburg Swedish. In *Proc. of the XVIIIth Swedish Phonetics Conference, Fonetik 2005*. Department of Linguistics, Göteborg University.

Sjölander K and Heldner M (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proc. of The XVIIth Swedish Phonetics Conference, Fonetik 2004*, 116–119. Stockholm University.

Taylor P (2009). *Text-To-Speech Synthesis*. Cambridge University Press.

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

Watts O, Yamagishi J, Berkling K and King S (2008). Hmm-based synthesis of child speech. In *Proc. of The 1st Workshop on Child, Computer and Interaction*.