Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization

Johan Frid

Department of Linguistics and Phonetics
Lund University, Sweden
johan.frid@ling.lu.se

Abstract

This paper describes an initial attempt at the construction of a data-driven model of Swedish intonation. The study is mainly concerned with model-building and prediction of the intonation patterns of accented words in a corpus of read news in Swedish. Extraction of pitch information is achieved by performing a stylization of the pitch contours. The information is used to build a model for the prediction of pitch patterns using linguistic features such as accent type and position of stress. The model is tested against unseen data from the same corpus. The evaluation is done by numerical comparisons. The RMSE between predicted and original contours for the different categories ranges between 3.7 and 31.4 Hz. The results are quite promising for future studies.

1. Introduction

The existing model of F_0 prediction in our department's systems and tools for speech synthesis [1], [2] is rule-based. It can be described as a 'ToBI-style' intonation model in that it uses tonal turning points, represented by Ls and Hs, which are mapped to time and frequency values that are connected by straight lines in order to produce a pitch contour. The model has been fairly successful at producing a neutral intonation of Standard Swedish [3] and has also been applied to different dialects of Swedish [4]. There have also been attempts, summarized in [5], to incorporate discourse and dialogue features into the model.

Given that recent attempts [6], [7] at data-driven methods have been rather succesful within the area of speech synthesis, and that such approaches, to our knowledge, havn't been pursued previously for Swedish, we decided to investigate this technique.

In this study we concentrate on the words with an actual realization of word accent. We have not yet tried to predict which words in a phrase that get accents, or why some words are deaccented. Neither have we accounted for phrasal phenomena like focussing or post-focal downstepping. This study is therefore somewhat tentative, rather a test of a possible methodology for future studies than a full account of Swedish prosody.

2. Speech data

The speech data for this study was taken from a corpus consisting of read news (from the hourly Swedish news program 'Ekot'). The speech in the corpus is read at a clear, rather formal style, and the complexity of utterances can be rather high. The mean length of a read sentence is 6.8 s, and the average number of words/sentence is 15.76. The corpus consists of several speakers, both male and female, which all speak a variety of 'Standard' Swedish. The voice quality of the speakers is very good and poses few problems for pitch extraction. The corpus currently consists of 300 sentences and from this we extracted about 2300 content words (nouns, verbs and adjectives) which had a word accent.

3. Linguistic Analysis

A linguistic analysis was performed in order to classify the words in different categories. Each word's lexical accent was first determined by looking it up in a word accent lexicon. This lookup also gave information about the number of syllables and the position of the mainly stressed syllable.

The accent information was then checked manually by listening to the words and determining whether the designated labels were correct or not by using the pitch information of each word. Visual inspection of the F_0 was used in some doubtful cases, where the intuitions of the author had to settle the issue.

The accent types used were:

- Accent 1 (acute)
- Accent 2 (grave)
- Compound accent

Note that compounds almost always have Accent 2 (except in some dialects in southern Sweden), but a distinction is made here since compounds always have a secondary stressed syllable, which the Accent 2 words may or may not have.

In addition to the accent category, the position type of the mainly stressed syllable was also included as a feature used for prediction. The position types were:

- Initial
- Medial
- Final
- Single (for monosyllabic words)

There are many other features that influence the pitch properties of words (cf. [6]), in particular the degree of prosodic prominence of a word, but also at levels above and below the word, such as position in the phrase from the left and right edges, pre- or post-focal position, openness and heaviness of syllables and foot structure. These have not been used in the present study, but this will be the subject of future investigations.

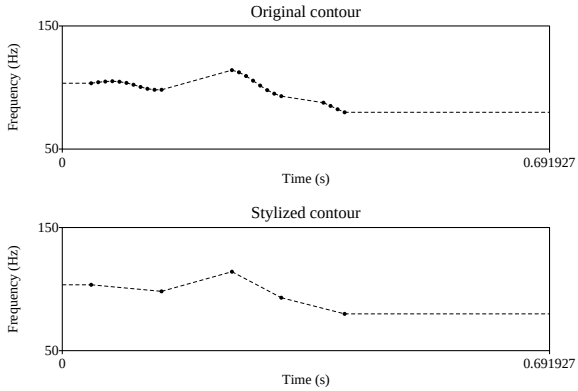

Figure 1: *Original and stylized pitch contours. The dots show the pitch points, and the dashes how they are connected.*

4. Acoustic Analysis

The goal of the acoustic analysis was to extract pitch information for the selected words in order to build a model for prediction of F_0 contours from the linguistic features. Pitch information was extracted from the words by first obtaining F_0 contours, then smoothing the contour in order to remove voice quality conditioned minor perturbations that have little intonational content, and then performing a stylization of the pitch contour.

4.1. Stylization

The stylization works by selecting tonal turning points in the contour. The points are selected so that when reconnecting the points with straight lines, there may not, at any given point along the contour, be a difference in pitch between the reconstructed contour and the original contour that is larger than a set value, in this case one (1) semitone. This results in a series of time/frequency pairs, which describe the contour of a pitch pattern accurately, but with a smaller number of points than the full contour.

The stylization process is illustrated in Fig. 1, which shows two pitch contours—original and stylized—of the word 'Helsingfors' spoken by one of the speakers. Note the much smaller number of pitch points in the stylized contour.

In one sense, this method follows the IPO [8] paradigm of intonation stylization in that it tries to approximate F_0 contours using "close-copy stylizations". This method is, however, fully automatic, and does not use the interactive step.

Note that the points are not anchored to any particular point in the word, neither a temporally aligned one, such as start, middle or end of the word, nor a linguistically motivated one, e.g. the vowel onset of the stressed syllable. This might of course introduce some unnecessary variation due to the durations of segments being different from word to word and that some words begin with voiced segments and others with unvoiced, but we are mainly interested in general tendencies here and we think that the amount of such variation will be roughly the same in each linguistic category.

Another issue is that there is no guarantee that the number of points selected will be the same from word to word. The stylization simply selects the lowest number of points it needs in order to produce a line within the given limit. This means that the number of points used in modelling the original contours potentially may vary with the complexity of each pitch contour. This acoustically grounded variation in the number of features to be predicted is a bit problematic unless the number itself is possible to predict from the other linguistic features. Without claiming that we have used the best solution to this problem, we will return to this issue in the model-building described in the next section.

Other stylization models include e.g. MOMEL [9], Tilt [10] and the d'Alessandro-Mertens model [11]. The current model has not yet been compared to these models.

4.2. Pitch extraction

Both the F_0 analysis, the smoothing, and the contour stylization was performed using the functions available in the PRAAT [12] program. Some words (about 20) were spoken with a very harsh or whispered voice quality, and they were discarded from the study as it was impossible to calculate F_0 in these words. This left 2311 words for the remaining study.

In order to test the stylization procedure's ability to accurately model F_0 contours of real speech, new contours were reconstructed from the stylized data and compared with the originals. The RMSE (*root mean squared error*) between voiced frames of the original and reconstructed contours was 2.68 Hz (0.296 semitones) for the 2311 words.

4.3. Normalization

Since the corpus contained speech from both male and female speakers, we normalized the actual pitch values by dividing the values of the pitch points by a value calculated by dividing the mean F_0 of each word by 100.

The words were also normalized in time by dividing the times of the pitch points by the duration of each word. The values thus obtained correspond to how many percent of the word's total duration into the word a pitch point is.

5. Building models

In the previous section we described the extraction of pitch information. Model-building from this data is not completely straighforward since the pitch feature vectors contain different number of elements. If one word has been stylized with two points and another in the same category with three points, how can we integrate these parameters in one model? The strategy selected here is: if two different pitch patterns have been stylized with a different number of points in order to keep the stylized contour within the allowed range from the orginal contour, we deem that both pitch patterns are worth using. We thus subcategorize the contours within a word category according to the number of pitch points used in stylizing each contour. Other types of more linguistically motivated splitting of the data, such as syllable weight, number of syllables in the word or the word's position in the phrase may also prove to be valuable, but this requires a finer analysis than what has been performed at the moment and hence has to be saved for a future study.

Following this discussion, all the data was classified according to:

- Order of stylization (the number of pitch points used to stylize the pitch contour)
- Position type of the main stressed syllable
- Accent

All words with the same order of stylization and the same position and accent types were placed in the same group. For

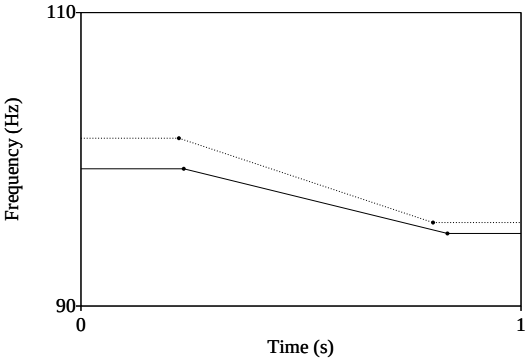

Figure 2: *Models with stylization order = 2, stress type = initial.*

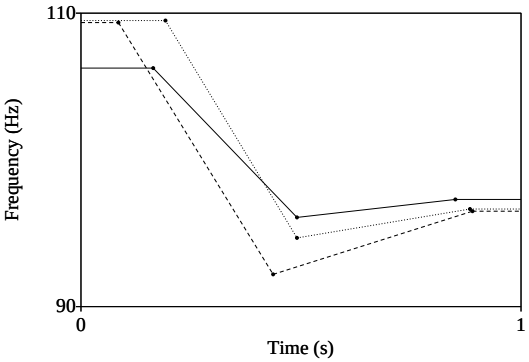

Figure 3: *Models with stylization order = 3, stress type = initial.*

each group, the mean time and frequency values for each pitch point was calculated. In order to get some reliability of the models, only groups with more than 30 occurrences were used. In each group, every fifth word was left out and placed in the test set for that group. In the end, 1974 word were used and the split between training and testing data was: training set = 1591 words, test set 383 words (roughly an 80%–20% split).

Illustration of the models are shown in Figs. 2 and 3. The figure shows the models for words stylized with two and three pitch points and initial stress. The plain line is the model for Accent 1, the dotted line the model for Accent 2 and the thick dashed line the model for compounds. They are shown in the normalized format, meaning that the pitch scale is around 100 Hz and the time scale from 0 to 1. The main difference seems to be that the Accent 2 and Compound models generally start from a higher level and fall to a lower level than the Accent 1 model does. This also gives them a steeper shape.

6. Evaluation

In order to evaluate the model, intonation contours for the words in the test set were reconstructed from the model. For each word, the model was 'denormalized' using the word's actual length and mean pitch. The reconstructed contours were then compared with the originals by calculating the numerical difference between reconstructed and original. Note that the numerical measure should only be taken as an indication of how succesful a model *might* be. The perceptual impression is always the ultimate test of any model of intonation. No perception test has been carried out so far. Perception tests are quite labourious and not always easily interpretable and numerical measures can give some indication of whether a model is worth of further testing or not.

7. Results

Table 1 shows the mean and median RMSE in all the groups included in the study. The overall measure should be taken with some care, since only the groups with 30 or more words are included, and this measure does thus not account for the other cases.

Table 1: *RMS Errors (in Hz and semitones) between reconstructed and original F_0 contours.*

	Hz	Semitones
mean	10.6	1.19
median	7.6	0.96

Table 2 shows the results for the different groups. Each error rate is a mean of all the words in that group.

Table 2: *RMS Errors (in Hz and semitones) between reconstructed and original F_0 contours for different groups.*

order	Syll. Type	Acc. Type	RMSE (Hz)	RMSE (st)
2	initial	Acc1	5.4	0.65
2	single	Acc1	3.7	0.46
2	initial	Acc2	4.1	0.48
3	initial	Acc1	6.7	0.80
3	mid	Acc1	7.9	0.98
3	single	Acc1	10.0	0.96
3	initial	Acc2	7.8	0.96
3	initial	Comp	6.1	0.79
4	final	Acc1	31.4	3.00
4	initial	Acc1	10.0	1.04
4	mid	Acc1	8.3	0.93
4	single	Acc1	15.0	1.67
4	initial	Acc2	10.4	1.25
4	initial	Comp	9.4	1.00
5	initial	Acc1	12.2	1.33
5	mid	Acc1	13.8	1.46
5	initial	Acc2	18.2	1.82
5	initial	Comp	13.0	1.55
6	mid	Acc1	9.9	1.19
6	initial	Acc2	16.1	2.14
6	initial	Comp	20.4	2.27
7	initial	Comp	14.5	1.65
8	initial	Comp	15.4	1.92
9	initial	Comp	21.3	2.24

The results show a clear tendency that the lower the order of stylization, the lower the RMSE. Medial and Single stressed syllable types are generally lower than Initial. Only one case of Final is included in the study and this gives the largest error, 31.4 Hz. Acc1 and Acc2 groups generally have lower RMSEs than the Comp groups, but this is probably correlated with the order of stylization. The Comp group is more common in the

groups with a higher number of pitch points. For stylization orders 2 and 3, the RMSE is below 10 Hz and 1 st in all the groups.

8. Discussion

Many aspects of this study is sub-optimal and work in progress. Focussed words are not separated from unfocussed, prosodic context, i.e., position in phrase, pre- or postfocal position, information about presence of other and adjacent accents, syllabic information, is not used and the lexical word is not the most relevant domain for prosody.

Still, we interpret the results as indicative that the stylization method used in the study is able to model intonation patterns accurately. Particularly the words with the lower order of stylization have very low RMSEs.

Problems remaining to be solved, apart from using more linguistic features to categorize the groups, is how the different orders of stylization should be predicted. In a speech synthesis system this has to be guessed from the text processing, and it is at this stage not clear how this should be done.

In a way, the patterns resulting from calculating the means is quite similar to vector quantization (VQ) in that they represent a number of distinct pitch patterns. Using VQ in intonation modelling is a promising approach as shown in [14] and [15].

9. Conclusions

This study has shown that a model that uses stylization of pitch contours for accented words in Swedish and the linguistic features accent type and stress position type is able to predict pitch contours that numerically are quite similar to natural ones.

10. References

[1] Filipsson, M. and Bruce, G., "LUKAS - a preliminary report on a new Swedish speech synthesis", Working Papers 46:45–56, Dept. of Linguistics, Lund University, 1997.

[2] Frid, J., "An environment for testing prosodic and phonetic transcriptions", Proceedings of ICPhS 99, San Francisco, vol 3:2319–2322, 1999.

[3] Bruce, G. and Granström, B., "Prosodic modelling in Swedish", Speech Communication 13(1–2):63–73, 1993.

[4] Bruce, G. and Gårding, E., "A Prosodic Typology for Swedish Dialects", Nordic Prosody, Dept. of Linguistics, Lund University, 1978.

[5] Bruce, G., Filipsson, M., Frid, J., Granström, B., Gustafson, K., Horne, M. and House, D., "Modelling of Swedish Text and Discourse Intonation in a Speech Synthesis Framework", *in*: Botinis, A. (ed.) "Intonation. Analysis, Modelling and Technology", Kluwer Academic Publishers, Dordrecht, 291–320, 2000.

[6] Black, A. and Hunt, A., "Generating FO contours from ToBI labels using linear regression", Proceedings of ICSLP 96, Philadelphia, vol 3:1385–1388, 1996.

[7] Dusterhoff, K., "Synthesizing Fundamental Frequency Using Models Automatically Trained from Data", Phd Thesis, University of Edinburgh, 2000.

[8] 't Hart, J., Collier, R., and Cohen, A., "A perceptual study of intonation", Cambridge University Press, Cambridge, 1990.

[9] Hirst, D. and Espesser, R., "Automatic modelling of fundamental frequency using a quadratic spline function.", Travaux de l'Institut de Phontique d'Aix 15:71-85, 1993.

[10] Taylor, P., "Analysis and Synthesis of Intonation using the Tilt Model", Journal of the Acoustical Society of America 107(3):1697-1714, 2000.

[11] d' Alessandro, C. and Mertens, P., "Automatic pitch contour stylization using a model of tonal perception", Computer Speech and Language 9(3):257-288, 1995.

[12] Boersma, P. and Weenink, D. "PRAAT: doing phonetics by computer" Website: http://www.praat.org, 1992-2001.

[14] Möhler, G. and Conkie, A. "Parametric modeling of intonation using vector quantization", Proceedings of 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998.

[15] Syrdal, A., Möhler, G., Dusterhoff, K., Conkie, A. and Black, A., "Three methods of intonation modeling", Proceedings of 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998.