Lund University, Dept. of Linguistics
Working Papers 49 (2001), 42–45

Prediction of intonation patterns of accented words in a corpus of read Swedish news

Johan Frid
Dept. of Linguistics and Phonetics, Lund University
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 (Filipsson & Bruce 1997, Frid 1999) 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 (Bruce & Granström 1993), has been applied to different dialects of Swedish (Bruce & Gårding 1978). Efforts at incorporating discourse and dialogue features into the model are summarized in (Bruce et al. 2000).

Given that recent attempts (Black & Hunt 1996, Dusterhoff 2000) at data-driven methods have been rather successful within the area of speech synthesis, and that such approaches, to our knowledge, haven't been pursued previously for Swedish, we decided to investigate this technique. The study is 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') of several speakers, both male and female, which all speak a variety of 'Standard' Swedish. The corpus currently consists of 300 sentences and about 2300 content words (nouns, verbs and adjectives) with word accents.

3 Linguistic Analysis

Each word's lexical accent was first determined by looking it up in a word accent lexicon. The accent information was 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) and **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. The position type of the mainly stressed syllable was also analyzed. The position types were: **Initial**, **Medial**, **Final** and **Single** (for monosyllabic words)

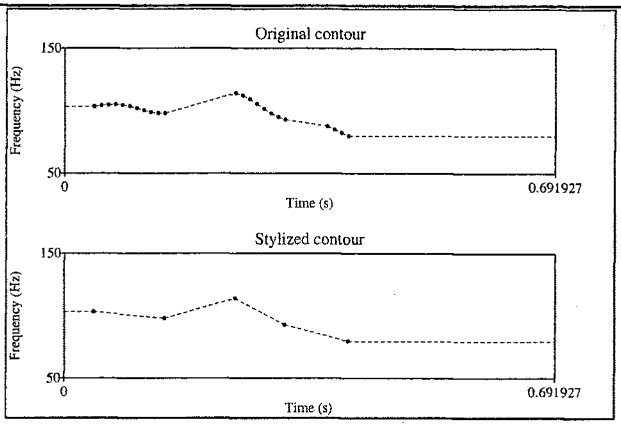

Figure 1. Original and stylized pitch contours. The dots show the pitch points, and the dashes how they are connected.

Many other features influence the pitch properties of words (cf. Black & Hunt 1996), 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.

4 Acoustic Analysis

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.

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. For further discussion on the stylization procedure, see (Frid 2001).

The F_0 analysis, the smoothing, and the contour stylization was performed using the functions available in the PRAAT (Boersma & Weenink 1992-2001) program. Some words

Table 1. RMSEs (in Hz and semitones) between reconstructed and original F_0 contours.

	Hz	*Semitones*
mean	10.6	1.19
median	7.6	0.96

(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.

Since the corpus contains 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.

5 Building and evaluating models

In the previous section we described the extraction of pitch information. Model building from this data is not completely straightforward 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 original 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.

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 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. For each group, every fifth word was placed in the test set. 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).

In order to evaluate the model, intonation contours for the words in the test set were reconstructed. For each word, the model was 'denormalized' using the word's actual length and mean pitch. The reconstructed contours were compared with the originals by calculating RMSEs.

6 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. For a more detailed presentation of the results, see (Frid 2001).

The results show a tendency that the lower the stylization order, the lower the RMSE. Medial and Single syllable types are generally lower than Initial. 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.

7 Discussion

Many aspects of this study is sub-optimal and work in progress, particularly the linguistic analysis needs much more refinement. 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.

Note that the numerical measure should only be taken as an indication of how successful 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 laborious and not always easily interpretable and numerical measures can give some indication of whether a model is worth of further testing or not.

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.

8 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.

References

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

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

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.

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

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

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

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.

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

Frid, J., "Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization", submitted to Eurospeech 2001.