Original scientific paper

Automatic Intonation Event Detection Using Tilt Model for Croatian Speech Synthesis

Lucia Načinović
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
lnacinovic@inf.uniri.hr

Miran Pobar
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
mpobar@inf.uniri.hr

Sanda Martinčić-Ipšić
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
smarti@inf.uniri.hr

Ivo Ipšić
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
ivoi@inf.uniri.hr

Summary

Text-to-speech systems convert text into speech. Synthesized speech without prosody sounds unnatural and monotonous. In order to sound natural, prosodic elements have to be implemented. The generation of prosodic elements directly from text is a rather demanding task. Our final goals are building a complete prosodic model for Croatian and implementing it into our TTS system. In this work, we present one of the steps in implementation of prosody into TTSs – detection of intonation events using Tilt intonation model. We propose a training procedure which is composed of several subtasks. First, we hand-labelled a set of utterances and within each of them, marked four types of prosodic events. Then we trained HMMs and used them to mark prosodic events on a larger set of utterances. We estimate parameters for each of the intonation event and generated f0 contours from the parameters. Finally, we evaluated the obtained f0 contours.

Key words: prosody in TTS, intonation model, Tilt

Introduction

Intonation modelling plays a great role in TTS systems. Synthesized speech without intonation component sounds unnatural and monotonous. Prediction of intonation patterns from text has been a difficult task due to their complex nature. There are, however, various prosodic models that predict prosodic elements from a text. They vary from rule based prescriptive models to data driven models such as CART decision trees (Dusterhoff et al., 1999), lazy learning approaches (Blin & Miclet, 2000) and unit selection based models (Meron, 2001).
Phonological approaches to prosodic analysis of speech use a set of abstract phonological categories (tone, breaks etc.) and each category has its own linguistic function. An example of this approach is ToBI intonation model (Silverman et al., 1992). Parameter based approaches attempt to describe f0 contour using a set of continuous parameters. Such approaches are, for example, Tilt intonation model (Taylor, 2000) and Fujisaki model (Fujisaki & Ohno, 2005).
Besides the mentioned models that tend to fall into one of the basic categories, there are models that use additional methodology (JEMA) (Rojc et al., 2005) or combine rule-based approach with data driven approach (Aylett et al., 2003).
Regarding Croatian, there is a list of rules about how accents on words in a sentence are combined (Mikelić Preradović, 2008). Using method "analysis by synthesis", basic intonation categories: "rise", "fall" and "flat" have been determined (Bakran et al., 2001). Our goal is to build a complete prosodic model for Croatian and implement it into our TTS system. In this paper we will present the way we automatically detected intonation events for Croatian using Tilt intonation model and statistical models – hidden Markov models (HMM). In accordance with the results of the research on the basic intonation categories for Croatian, we have chosen Tilt intonation model which also differentiates three main prosodic events – rise, fall and connection.
The paper is organized as follows: in the next chapter we give an overview of the Tilt intonation model. In the third chapter we explain the procedure of the automatic detection of prosodic events. We describe the speech database we used, the process of hand-labelling, f0 feature sets extraction, the procedure of HMMs training and the process of f0 generation. We conclude the paper with the main results we obtained.

Tilt model overview

Tilt (Taylor, 2000) is a phonetic model of intonation that represents intonation as a sequence of continuously parameterised events (pitch accents or boundary tones). These parameters are called tilt parameters, determined directly from the f0 contour.
Basic units of a Tilt model are intonation events – the linguistically relevant parts of the f0 contour (circled parts in picture 1). From such a representation, it is possible to encode the linguistically relevant information in an f0 contour, and then recreate the original f0 from this coding.

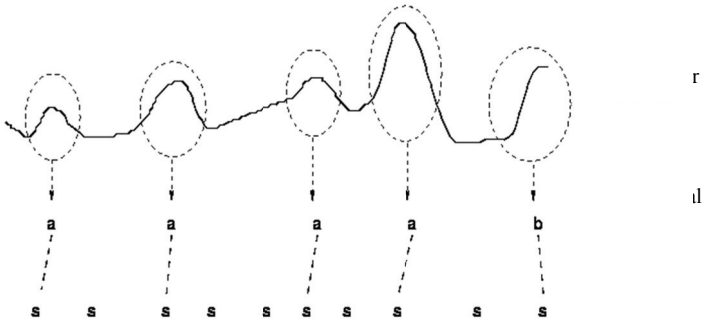

Figure 1: Intonation events in the Tilt model

Tilt model can be described with a simpler model – RFC model (R-rise, F-fall, C-connection). In the RFC model, each event is modelled by a rise part followed by a fall part. Each part has an amplitude and duration, and two parameters are used to give the time position of the event in the utterance and the f0 height of the event. (Taylor, 1995).

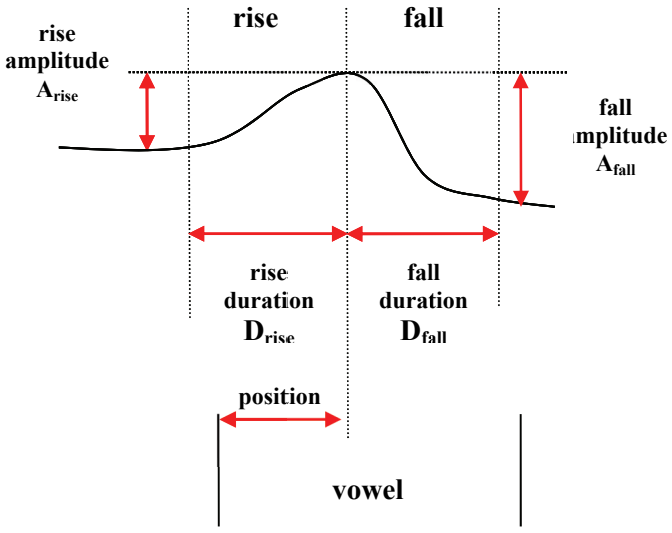

Figure 2: RFC parameters in the Tilt model

The RFC parameters for an utterance are:

- rise amplitude (Hz),
- rise duration (seconds),
- fall amplitude (Hz),
- fall duration (seconds),
- position (seconds),
- f0 height (Hz).

Those parameters can be transformed into Tilt parameters:

- Tilt-amplitude (Hz): the sum of the magnitudes of the rise and fall amplitudes:

$$tilt_{amp} = \frac{|\mathrm{A_{rise}}| - |\mathrm{A_{fall}}|}{|\mathrm{A_{rise}}| + |\mathrm{A_{fall}}|}$$

- Tilt-duration (seconds): the sum of the rise and fall durations:

$$tilt_{dur} = \frac{|\mathrm{D_{rise}}| - |\mathrm{D_{fall}}|}{|\mathrm{D_{rise}}| + |\mathrm{D_{fall}}|}$$

- Tilt: a dimensionless number which expresses the overall shape of the event, independent of its amplitude or duration:

$$tilt = \frac{|\mathrm{A_{rise}}| - |\mathrm{A_{fall}}|}{2|\mathrm{A_{rise}}| + |\mathrm{A_{fall}}|} + \frac{|\mathrm{D_{rise}}| - |\mathrm{D_{fall}}|}{2|\mathrm{D_{rise}}| + |\mathrm{D_{fall}}|}$$

Tilt is calculated from the relative sizes of the rise and fall components in the event. A value of +1 indicates the event is purely a rise, -1 indicates it is purely a fall. Any value between says that the event has both a rise and fall component, with a value of 0 indicating they are the same size.

Intonation event detection

In order to detect intonation events and label the whole database, an automatic HMM based procedure which we described in this chapter was used. The procedure uses four HMMs to predict the four intonation events from the f0 features. To train the parameters of the HMMs, a set of hand labelled utterances was used.

Speech database

Speech database that we used in our research consists of 1 hour and 54 minutes of speech from a collection of fairy tales spoken by one speaker. We hand-la-

belled hundred utterances out of which we selected a subset of 25 utterances for testing and 75 for training the model.

Hand labelling

From the speech database, we chose a hundred of utterances which were labelled by hand to produce intonation transcriptions. We located pitch accents, boundaries, connections and silences within each utterance, in accordance with the intonation event model (Figure 3.)

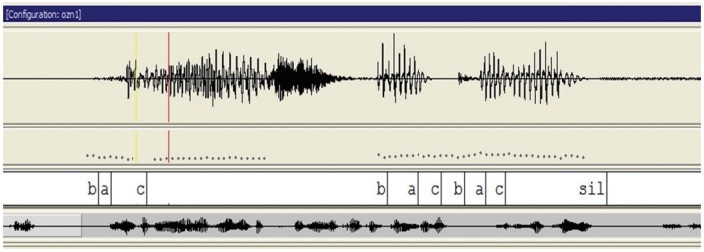

Figure 3: Intonation event transcription of an utterance

Labels that we used for labelling events are: *sil* for unvoiced parts, *a* for pitch accents, *b* for all rising boundaries and *c* for all falling boundaries.
The process of labelling was performed using the WaveSurfer tool (Sjölander & Beskow, 2000) . Labelled files were exported and saved as xlabel files and were then used in the automatic detection of intonation events.

F0 feature sets extraction

To extract f0 features from the training set of utterances we used RAPT algorithm (Talkin, 1995) as implemented in Voicebox Matlab toolbox. The f0 was sampled at 10 ms. The obtained f0 contours contained some noise which we smoothed with a three point median filter. We set the f0 value to 0 Hz to represent the unvoiced segments where f0 cannot be determined and in another attempt we used linear interpolation to determine the missing values. We got three different f0 feature sets: raw output from the RAPT algorithm, smoothed and interpolated. In addition to the f0 contour, we used dynamic features (delta f0 and delta-delta f0) for training HMMs of intonation events.

Automatic event recognition

We trained HMMs for detection of accents, boundaries, connections and silences. A five-state HMM was used for each event type. Models were trained with Baum-Welch algorithm on f0 features and hand-labelled event positions.
For all three f0 feature variants mentioned in the chapter above, a different HMM set was trained.

To limit valid event sequences, a grammar with permitted combination of events was defined. Viterbi algorithm was used to detect the events.
For testing the automatic event detection, the utterances are divided into two sets which were identical to the sets that we used in the process of training. We applied models trained on different f0 features from the training set to the set of utterances with different f0 features from the test set. The performances for all type of events and for each event separately are shown in Table 1.

Table 1: Performance for different feature sets

Feature set	Correctness
f0 raw	45.62
f0 smoothed	53.75
f0 interpolated	45.77

The correctness was computed using the Levenshtein distance between the automatically generated and hand-labelled event labels. We got the best results with models trained on median filter smoothed f0 feature set and applied to feature set obtained in the same way.
The interpolation of missing f0 values did not improve the event detection, as distinguishing between voiced and unvoiced speech may give important clues for event locations, and by interpolation this information was lost.

Tilt analysis
When the events are detected, the location of the start, peak and end position of each event has to be determined. The analysis performs only on those parts of f0 contours which were detected as the events. Each of those parts is smoothed by median smoothing algorithm and unvoiced regions are interpolated. Each event has to be described as a rise or fall shape within the f0 contour so tilt parameters have to be assigned to each of them. Algorithm used in tilt analysis minimizes the difference between the original contour and the fitted shape. We get a model represented by tilt parameters which were explained in chapter 2.

Tilt synthesis
From tilt/RFC parameters, we can generate f0 contours using tilt synthesis and given equations:

$f0(t) = A_{abs} + A - 2.A.(t/D)^2 \qquad 0 < t < D/2$
$f0(t) = A_{abs} + 2.A.(1 - t/D)^2 \qquad D/2 < t < D$

where A is rise or fall amplitude, D is rise or fall duration and A_{abs} is the absolute f0 value at the start of the rise or fall, which is given by the end value of the previous event of connection.
Places on the f0 contour between the events are filled using the method of interpolation.

Results

The usual measure for evaluating generated f0 contour is the root mean square error (RMSE) between the original contour and the obtained generated f0 contour. We compared the performance of three models for automatic event detection, trained on raw, smoothed and interpolated f0 features. The models produced event labels for the test data set (f0 features extracted from 25 utterances, with interpolation for unvoiced segments). Tilt analysis was performed using these labels and f0 features, yielding tilt parameters from which the f0 contours were synthesized. The resulting f0 contours were compared with the original (interpolated) f0 contour. In the same way the f0 contour was synthesized using hand-labelled events and compared with the original f0. The results of comparison are shown in Table 2.

Table 2. Mean RMSE values for generated f0 contours.

Event label model	RMSE (Hz)
raw	**25.16**
smoothed	26.69
interpolated	25.57
hand-labelled	**23.11**

We got the best results using the model trained on raw (unprocessed) f0 features. The obtained results are satisfactory but further improvements might be achieved.

Figure 4 shows an example of the generated contours, each compared with the original f0.

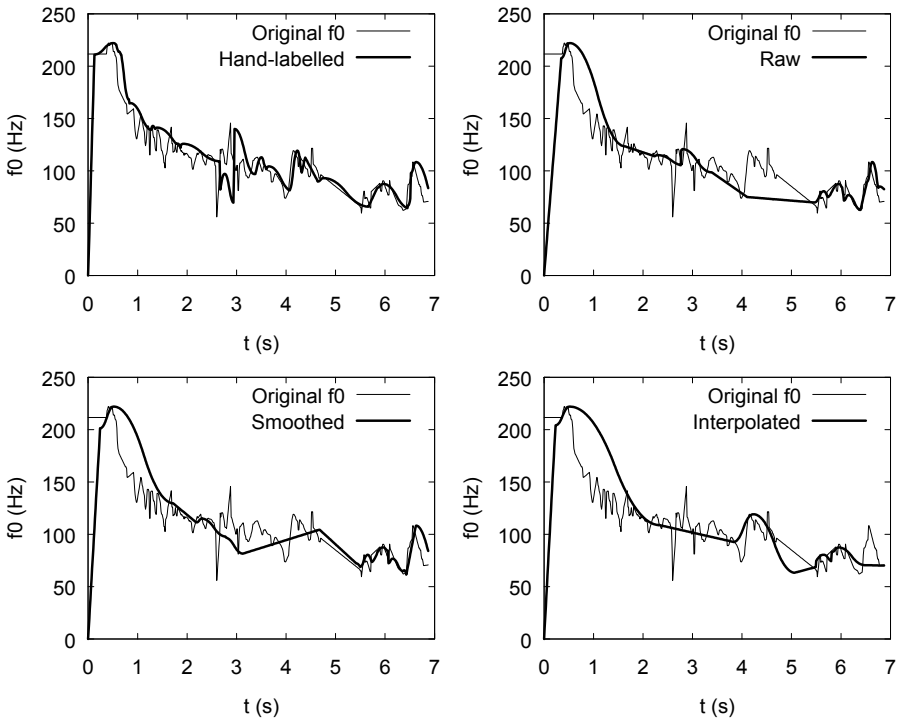

Figure 4: Comparison of the generated f0 contours with the original f0

Discussion

All f0 contours obtained from automatically detected events have similar RMSE values, and perform comparably to the hand-labelled case. More hand labelled training data may not be sufficient to improve the RMSE, but also improving the hand-labelling procedure could contribute to better RMSE.

We plan to improve the quality of hand-labelled event boundaries using an automated procedure. A search for the optimal position of the boundary could be done by trying several positions in the vicinity of labelled boundary and noting the change in observed RMSE. The boundary is fixed after a predefined number of iterations.

Further step towards automatic f0 generation from text will be CART (classification and regression trees) building. Based on the questions in tree nodes regarding chosen linguistic features extracted from text, trees will predict tilt parameters.

Conclusion

Implementation of prosodic elements into text-to-speech systems represents a demanding task. As a part of our final goal to implement prosody into our TTS, in this paper we proposed a procedure for automatic event detection. We chose a representative set of utterances and marked four main prosodic events within each utterance. We trained HMMs to mark events automatically on a larger set of utterances. We parameterized the detected events with tilt parameters and generated f0 contours out of those parameters. We evaluated the obtained f0 contours. Future work will include building a model for prediction of tilt parameters from text.

References

Aylett, M.P.; Fackrell, J.; Rutten, P. My voice your prosody: Sharing a speaker specific prosody model across speakers in unitselection tts. // *Eurospeech*. 2003; 321-324.

Bakran, J.; Erdeljac, V.; Lazić, N. Modeliranje temeljnih intonacijskih oblika. // *Govor*. vol. 18 (2001); 105-111.

Blin, L.; Miclet, L. Generating synthetic speech prosody with lazy learning in tree structures. // *CoNLL-2000 and LLL-2000*. 2000; 87–90.

Dusterhoff, K.E.; Black, A.W.; Taylor P. Using decision trees within the tilt intonation model to predict f0 contours. // *Eurospeech*. 1999; 1627–1630.

Fujisaki, H.; Ohno. Analysis and modeling of fundamental frequency contours of English utterances. //*Speech Communication*. 2005; 47:59-70.

Meron, J. Prosodic unit selection using an imitation speech database. // *4th ISCA Workshop on Speech Synthesis*. 2001; 53-57.

Mikelić Preradović N. Pristupi izradi strojnog tezaurusa za hrvatski jezik. Ph.D. thesis, University of Zagreb. 2008.

Rojc, M.; Agüero, P.D.; Bonafonte, A.; Kacic, Z. Training the Tilt intonation model using the JEMA methodology. // *Eurospeech*. 2005; 3273-3276.

Silverman et al. ToBI: A standard for labeling English prosody. // *ICSLP92*. 1995; 2:867-870.

Sjölander, K.; Beskow, J. Wavesurfer – An open source speech tool. // *Interspeech.* 2000; 464-467.

Talkin, D. A robust algorithm for pitch tracking (RAPT). // *Spech coding and synthesis*. 1995; 495-518.

Taylor, P. Analysis and synthesis of intonation using the tilt model. // *The Journal of the Acoustical Society of America.*2000; 1697-1714.

Taylor, P. 1995. The rise/fall/connection model of intonation. // *Speech Communication*. vol. 15 (1995); 169–186.