An Overview of Prosodic Modelling for Croatian Speech Synthesis

Lucia Načinović Prskalo, Sanda Martinčić-Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{lnacinovic, smarti}@inf.uniri.hr

Abstract: *In order to include prosody into the text to speech (TTS) systems prosody knowledge needs to be acquired, represented and incorporated. Two main features of prosody important for modelling prosody for TTS systems are duration and F0 contour. There are various approaches to modelling those features and they can be categorized into three main groups: rule based, statistical and minimalistic. Some of the best known approaches to duration acquiring are Klatt's model, classification and regression trees (CARTS) and neural networks and to F0 modelling TOBI, Fujisaki and Tilt. A procedure for automatic intonation event detection on Croatian texts based on the Tilt model was evaluated in terms of Root Mean Square Error values for generated F0 contours.*

Key Words*: prosody modelling, speech synthesis, TTS, duration models, F0 contour models, prosodic characteristics of Croatian*

1 Introduction

The main task of speech synthesis is the generation of voice signal understandable to listener from the text input. This implies that the synthesized speech should sound natural, and that it should own prosodic characteristics of natural human speech. Language conveys a wide range of information about the duration, intonation, emphasis, grouping words into phrases, voice quality, rhythm, etc., and these features are collectively referred to as - prosody. Prosody plays a great role in intelligibility, and especially in the naturalness of synthesized speech.

The ability of humans of using the prosody knowledge is naturally acquired but difficult to articulate. For synthesizing speech from a text by a machine this prosody knowledge needs to be acquired, represented and incorporated. Therefore prediction of the prosodic patterns directly from text is not an easy task [1]. However, for this purpose, there are different models or algorithms that attempt to predict the prosodic elements from text. These models vary from models based on a set of rules to data driven models, such as classification and regression trees (CARTS) [2] and Hidden Markov models [3]. Besides the mentioned models that tend to fall into one of the basic categories, there are models that use additional methodology (JEMA: Joint feature extraction and modelling) [4] or combine rule-based approach with data driven approach [5].

This paper discusses the component of prosodic analysis in TTS systems. A procedure

for automatic intonation event detection on Croatian texts is evaluated with Root Mean Square Error values for generated F0 contours using Tilt.

The paper is organized as follows. In the second section basic concepts of prosody in TTS systems are described. In the third section rule based, statistical and minimalistic approaches to prosody acquiring are outlined. Duration models are presented in the fourth section and F0 contour models in the fifth section. In the sixth section basic prosodic characteristics of Croatian and in the seventh related work for languages cognate to Croatian are outlined. A procedure for automatic intonation event detection for Croatian speech synthesis is presented in the eighth section. The paper concludes with our plans for future work for prosody modelling for Croatian TTS.

2 Prosody in TTS – basic concepts and definitions

Prosody is a complex combination of phonetic factors which has a task to express attitude, assumptions and draw attention as a parallel channel of communication in our daily speech [1].

Semantic content that is transmitted via voice or text message is also called denotation, and emotional aspects and the effects of intent that speaker wishes to convey are a part of the message that is called connotation. Prosody plays an important role in the transmission of denotation, and a major role in the transmission of connotations (speaker's attitude toward the message, toward the listener and toward the overall communication event) [1].

Prosody represents the acoustic properties of speech that transmit information, which is not conveyed by the word meaning such as emotions, discourse features, syntax [6].

Two most important prosodic features that affect the quality of the synthesized speech are considered to be duration and F0 contour.

Duration refers to the duration of all speech particles: paragraph, sentence, intonation unit, speech word, syllables and phonemes. However, for TTS the duration of phonetic segments rather than the duration of words and syllables is used [7] [8]. One of the reasons for this is that the pause (boundaries) between segments, which is one of the most important prosodic features, can be relatively easy to determine automatically. Research regarding the duration on the level of syllables and phonemes were mostly focused on the duration of syllables in read speech [9] [10]. It has been shown that the duration of vowels depends on many factors, some of which include the articulatory context (phonemes before and after vowels), accent (both word accent and sentence accent), position (the position of syllables in a word and speech unit).

The fundamental frequency (F0 contour) is determined by many factors such as segmental factors (microintonation), patterns of stress, melody, rhythm, gender, attitude, and physical and emotional state of the speaker. Two main approaches to intonation acquiring are phonological models and phonetic (parameter based) models.

3 Basic approaches to prosody acquiring

Three main approaches to prosody acquiring have been distinguished so far: rule based approach, statistical approach and minimalistic approach.

3.1 Rule based

Rule based approach of implementing prosody into the synthesized speech uses written rules to predict prosodic characteristics from text. One of the best known rule based approaches for the duration modelling is Klatt's MTalk system [11]. For F0 contour modelling the best known rule based model is Pierrehumbert's system [12] in which the

contour is described as a series of target values which are connected together by transition rules. The target values are expressed as locations within the current pitch range. Which syllables within the phrase are assigned a target depends on the stress pattern. For example, in a declarative neutral intonation, all pitch accents are high (H), when the phrase is terminal, the phrase final tones are low-low (L-L) and if it is nonterminal, they are low-high (L-H).

3.2 Statistical

The statistical model is trained on labelled data. Hand-labelled prosodic features are used for parameter estimation. Parameters represent the probability of prosodic events in the context of different linguistic features. Model is used to predict the most likely prosodic labels on any input text.
One of the methods used in statistical approach are decision trees (CART - classification and regression trees) [13] [14] [15]. A list of possible features must be determined, and the system automatically selects features that have the greatest ability of prediction. Hidden Markov Models (HMMs) is another method that can be used to predict prosodic events. In [3] HMMs are used to predict phrase boundaries, and the model is trained on the information about the type of word and preceding anticipated border. This approach requires a large amount of data for model training.

3.3 Minimalistic approach

In minimalistic approach, large natural language corpuses are used to train prosodic models, and as a source for units needed in the concatenation synthesis. There are several instances of units (most often diphones) with different characteristics in different phonetic and prosodic environment. One of the first systems that used unit selection approach in speech synthesis was CHATR: a generic speech synthesis system [16].

4 Approaches to duration modelling

As mentioned before, one of the two most important prosodic features in speech synthesis is duration. There are different approaches of duration modelling and some of them are described in this chapter.

4.1 Klatt's duration model

This model was developed in the 70ies and 80ies of the 20th ct and is an integral part of a MITalk formant speech synthesizer [11]. It is composed of sequential rules that include phonetic environment features, accents, shortening and lengthening of syllables at certain positions etc. The basic assumption in Klatt's model is that each segment has its inherent duration; each rule increases or decreases the duration of the segment for a certain percentage, and the duration of each segment cannot be decreased beyond minimal length.

4.2. CARTs

Some of the features that can be included in the duration modelling with classification and regression trees are phoneme identity, identity of phoneme to the left, identity of phoneme to the right etc. There are different programs for CARTS training and one of them is for example Wagon procedure in the Festival Speech Synthesis Systems [17].

4.3. Neural networks

Neural networks can be used in duration modelling [18]. The model first predicts the duration of the syllable and then complements it with the phoneme duration. For each

syllable vector which consists of information about the number of the phonemes in the syllable, accent, part of speech tag etc. is calculated.

5 Approaches to F0 modelling

Phonological approaches to prosodic analysis of speech use a set of abstract phonological categories (tone, breaks etc.) to describe F0 contour and each category has its own linguistic function. An example of this approach is ToBI intonation model [19]. Parameter based approaches attempt to describe F0 contour using a set of continuous parameters. Such approaches are, for example, Tilt intonation model [20] and Fujisaki model [21].

5.1 ToBI

ToBI (Tones and Break Indices) [19] takes a linguistic or phonological approach specifying a small set of discrete labels which identify the intonational space of accents and tones. It is used for transcribing accents and phrasing (grouping of words). ToBI differs two pitch accents: H* or L* and four main boundary tones L-L%, L-H%, H-H%, H-L%. One pitch accent is associated to each accented word and one boundary tone is associated to the end of each prosodic phrase.

5.2 Fujisaki

Fujisaki model [21] describes F0 contour as a superposition of two contributions: a phrase component and an accent component. The phrase component models the baseline component and the accent component models micro prosodic variations. F0 contour is generated as a result of the superposition of the outputs of two second order linear filters with a base frequency value. The second order linear filters generate the phrase and accent components. The base frequency is the minimum frequency value of the speaker.

5.3 Tilt

Tilt [20] is a phonetic model of intonation that represents intonation as a sequence of continuously parameterized events (pitch accents or boundary tones). These parameters are called tilt parameters, determined directly from F0 contour. Basic units of a Tilt model are intonation events – the linguistically relevant parts of the F0 contour.
Parameters important for events detection are rise amplitude (Hz), rise duration (seconds), fall amplitude (Hz), fall duration (seconds), position (seconds) and 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:

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

6 Prosodic characteristics of the standard Croatian language

The core of the most European languages makes the accented syllable in a stressed word of the intonation unit, while in Croatian the core is comprised of the accented syllable and syllable behind the accented syllable because of the differentiation of the ascending and descending stress.

In Croatian, there are six different intonation cores: descending (\), ascending (/), descending-ascending (\/), descending-ascending-descending or reversed (\ / \), ascending and descending or complex (/ + \) and flat (-). Their distribution is not related to the grammatical syntactic types [22].
The most common intonation beginning in Croatian is descending, after which any type of intonation core can follow. The intonation ending is always descending or low and flat except after a flat core, when it is high and flat. If the end of intonation core is low, intonation ending extends into a flat, low tone.
Syllables in the standard Croatian can be accented or unaccented, long or short and high or low (tone). In one spoken word, only one accented syllable is allowed in Croatian. The most common accented syllable is the first syllable of the word (in about 66% of the words in the text), then the second (in about 23% of the words), the third (6.7%) and the fourth (1.6%) [22].
Only one syllable in a spoken word is accented and all others are unaccented. Before the accented syllable all syllables are of high tone and short, and after of low tone and short or long.
Long accented syllables are 50% longer then long unaccented syllables and short accented are 30% longer then short unaccented [22].
Prosodic structure is an aspect of a prosody which refers to the fact that some words group together and some have a break or natural pause between them. At the boundaries between prosodic phrases we often hear a change in the rhythm of the speech or a pause. Prosodic unit smaller than prosodic phrase and greater than phonological word is called clitic group or "spoken word". It consists of a word and proclitic or enclitic. A clitic is a morpheme that is grammatically independent, but phonologically dependent on another word (e.g. /uškoli/). In Croatian low tone accent can only be found on the first syllable of a word and when there is a proclitic in front of a word the accent moves from the first syllable in a word to the proclitic. If a word had three or more syllables, the accent stays on the first syllable of a word [22].

7 Related work for languages cognate to Croatian

The Slovenian language is by prosodic characteristics similar to Croatian. Several studies regarding prosody implementation into TTS have been conducted for Slovene. Šef and Gams [23] developed a prosody generating system for TTS. They used the approach of duration modelling at two levels: intrinsic (type of voice, the voice environment, record type, syllable emphasis, etc.) and extrinsic (speed of pronunciation, position of the words within phrases and the number of syllables in a word). In F0 modelling, they differ two main phases: text segmentation on intonation units and definition of F0 contours for specific intonation units. Šef [24] also explored the automatic accentuation of words for Slovene words. First, it was determined whether each vowel is stressed or unstressed, and then the accents were corrected using decision trees, and taking into account the number of accented vowels and word length. Marinčić et al. [25] analyzed the automatic accentuation in the Slovene language, and compared the human and machine capacity of accent allocation. Gros [26] recorded a long continuous speech database and studied the influence of speech rate on the duration of syllables and phonemes. She presented models of intonation for the Slovenian language, based on the intrinsic level (word level) and extrinsic level (level higher than word level).
The Czech language can to some extent be compared to the Croatian language in its prosodic characteristics. Romportl and Kala [27] described the statistical F0 modelling, intensity and duration of the Czech language. Tihelka et al. [28] describe a speech

synthesis system for Czech language which includes prosodic characteristic module based on the unit selection approach.
Tihelka and Matoušek [29] also incorporated phonetic transcription and prosodic rules to convert an input text to its phonetic form and to estimate its suprasegmental features in ARTIC system for Slovak. Kondelova et al. [30] proposed statistical approach for prosody contour modelling based on sentence classification for the Slovak language.
Sečujski [31] has developed dictionary of accents for Serbian designed for the Serbian speech synthesis.

8 Automatic intonation event detection for Croatian speech synthesis

A procedure for automatic intonation event detection on Croatian texts based on the Tilt model was proposed in [32]. In order to detect intonation events automatically, we chose a representative set of utterances and marked four main prosodic events (pitch accents, boundaries, connections and silences) within each utterance. Then we trained HMMs to mark events automatically on a larger set of utterances. To extract F0 features from the training set of utterances we used RAPT algorithm [33] as implemented in Voicebox Matlab toolbox. 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. Finally, we obtained three different F0 feature sets: raw output from the RAPT algorithm smoothed and interpolated. We parameterized the detected events with tilt parameters and generated F0 contours out of those parameters. In order to evaluate the obtained F0 contours, we compared three different F0 contours based on three models for automatic event detection, trained on raw, smoothed and interpolated F0 features to the original contour. The F0 contour synthesized using hand-labelled events was also compared with the original F0. The usual measure for F0 contour evaluation is the root mean square error (RMSE) between the original and generated F0 contour. The obtained results are shown in Table 1 and graphical comparison is shown in Fig. 1.

Table 1: Root Mean Square Error values for generated F0 contours

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

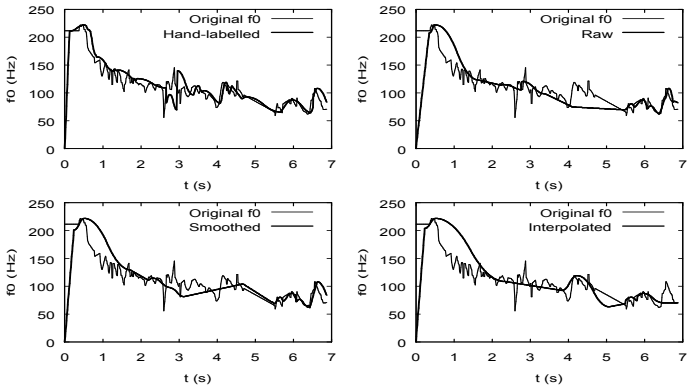

Figure 1: Comparison of the generated F0 contours with the original F0

9 Conclusion and future work

A procedure for automatic intonation event detection on Croatian texts based on the Tilt model was evaluated in terms of Root Mean Square Error values for generated F0 contours. Three different F0 feature sets: raw output from the RAPT algorithm; smoothed and interpolated were compared.

The results that we obtained are preliminary and we expect to get better results after we train the model on a larger set of sentences. All F0 contours obtained from automatically detected events have similar RMSE values, and perform comparably to the hand-labelled case which encourages us to use this method in the future work. We plan to build CARTS for Tilt parameter prediction from text. We also plan to build duration model for Croatian and to automatically accent the Croatian words with CARTS. Then we will incorporate the obtained duration and F0 models into Croatian TTS system and evaluate the generated speech.

10 References

[1] Huang X.; Acero A.; Hon H. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. New Jersey: Prentice Hall, 2001.

[2] Dusterhoff K. E.; Black A.; Taylor P. Using Decision Trees within the Tilt Intonation Model to Predict f0 Contours. Eurospeech, pp. 1627-1630, 1999.

[3] Taylor P.; Black A. Assigning Phrase Breaks from Part-of-Speech Sequences. Computer Speech and Language, pp. 99-117, 1998.

[4] Rojc M; Aguero P. D.; Bonafonte A; Kacic Z. Training the tilt intonation model using the JEMA methodology. Interspeech, pp. 3273-3276, 2005.

[5] Aylett M. Merging Data Driven and Rule Based Prosodic Models for Unit Selection TTS. Pitsburgh, 2004.

[6] Fordyce C.S. Prosody Prediction for Speech Synthesis Using Transormational Rule-Based Learning. 1998.

[7] Santen van J. Segmental Duration and Speech Timing. Computing Prosody, pp. 225-250, 1997.

[8] Santen van J. Assignment of Segmental Duration in Text-to-Spech Synthesis. Computer Speech and Language, pp. 95-128, 1994.

[9] Kato H; Tsuzaki M; Sagisaka Y. Acceptability for Temporal Modification of Single Vowel Segments in Isolated Words. J. Acoust. Soc. Am., pp. 540-549, 1998.

[10] Stergar J.; Erdem C. Adapting Prosody in a Text-to-Speech System. Products and Services; from R&D to Final Solutions, 2010.

[11] Allen J.; Hunnicut S.; Klatt D. Text-to-Speech: The MITalk System. Cambridge: Cambridge University Press, 1987.

[12] Pierrehumbert J.B. Synthesizing intonation. J. Acoust. Soc. Am., pp. 985-995, 1981.

[13] Hirschberg J. Pitch Accent in Context: Predicting Intonational Prominence from Text. Artificial Intelligence, vol. 3, pp. 305-340, 1995.

[14] Ross K.; Ostendorf M. Prediction of abstract prosodic labels for speech synthesis. Computer Speech and Language, vol. 10, pp. 155-185, 1996.

[15] Ostendorf M.; Veileux N. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. Computational Linguistics, vol. 20, pp. 27-54, 1994.

[16] Taylor P.; Black A. CHATR: a generic speech synthesis system. In COLING '94, pp. 983-986, 1994.

[17] The Festival Speech Synthesis System. [Online]. http://www.cstr.ed.ac.uk/projects/festival/

[18] Campbell W. N. Syllable-based segmental durations. Talking Machines: Theories, Models, and Designs, pp. 43-60, 1992.

[19] Silverman K.M. et al., TOBI: A Standard Scheme for Labeling Prosody. Banff, 1992.

[20] Taylor P. Analysis and Synthesis of Intonation using the Tilt Model. Journal of the Acoustical Society of America, pp. 1697-1714, 2000.

[21] Fujisaki H.; Ohno S. Analysis and Modeling of Fundamental Frequency Contours of English Utterances. In Speech Communication, vol. 47, 2005, pp. 59-70.

[22] Babić S. et al. Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika. Globus, Nakladni zavod, Zagreb, 1991.

[23] Šef T; Gams M. SPEAKER (GOVOREC): A Complete Slovenian Text-to Speech System. INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, pp. 277-287, 2003.

[24] Šef T. Automatic Accentuation of Words for Slovenian TTS System. In Proceedings of the 5th WSEAS International Conference on Signal Processing, pp. 155-160, 2006.

[25] Marinčič D.; Tušar T.; Gams M.; Šef T. Analysis of Automatic Stress Assignment in Slovene. Informatica, pp. 35-50, 2009.

[26] Gros J. Samodejno tvorjenje govora iz besedil. Doktorska disertacija, 1997.

[27] Romportl J.; Kala J. Prosody Modelling in Czech Text-to-Speech Synthesis. In Proceedings of the 6th ISCA Workshop on Speech Synthesis, pp. 200-205, Bonn, 2007.

[28] Tihelka D.; Kala J; Matousek J. Enhancements of viterbi search for fast unit selection synthesis. Interspeech, pp. 174-177, 2010.

[29] Tihelka D.; Matoušek J.; Romportl J. Current state of Czech text-to-speech system ARTIC. Berlin, Heidelberg, 2006.

[30] Kondelova A.; Toth J.; Rozinaj G. Statistical Approach for Prosody Contour Modeling Based on Sentence Classification. Elektrorevue, pp. 40-44, 2013.

[31] Sečujski M. Akcenatski rečnik srpskog jezika namenjen sintezi govora na osnovu teksta. DOGS2002, 2002.

[32] Načinović L.; Pobar M.; Martinčić-Ipšić S.; Ipšić I. Automatic Intonation Event Detection Using Tilt Model for Croatian Speech Synthesis. In Information Sciences and e-Society, Zagreb, 2011, pp. 383-391.

[33] Talkin D. A robust algorithm for pitch tracking (RAPT). Speech coding and synthesis, pp. 495-518, 1995.