MODELING SEGMENT INTONATION FOR SLOVENE TTS SYSTEM

Aleš Dobnikar

J. Stefan Institute
Jamova 39, 1000 Ljubljana, SLOVENIA
E-mail: Ales.Dobnikar@ijs.si

ABSTRACT

A scheme for modeling the F_0 contour for different types of intonation units for the Slovene language is presented. It is based on results of analyzing F_0 contours, using a quantitative model. Data from ten speakers was collected, resulting in a large corpora, mainly of declarative sentences. A way of generating the F_0 contour for given utterances was defined, using only the text of the utterance as input. Near-to-natural synthesized F_0 contour was obtained by rules which regard the F_0 contour as the sum of global and local components.

1. INTRODUCTION

The prosodic characteristics of natural speech, especially intonation, have at the same time very universal and very language specific features. Previous observations, studies and results obtained for Slovene intonation [1-6] have rarely been made in large data of acoustic domains. Such insufficient knowledge led to our improving and redefining the rules for determining fundamental frequency contours from phonologically relevant descriptions.

The generation of rules for various intonations consists of two main phases:

- segmentation of the analyzed speech into intonation units, and
- definition of prosodic rules allowing automatic derivation of F_0 from the text.

A few authors mention the use of the paragraph as intonation unit for speech synthesis [7, 8]. However, these authors conclude that this approach is significant but as yet fairly vague. The reason is the insufficient knowledge of intonation organization of the paragraph and the fact that no account is taken of the effect of the text level on the intonation structure [9]. In these studies, the onset or offset of the pitch range does not essentially depend on the location of the sentence or intonation unit in the paragraph. In the model presented here, the intonation unit is any connected signal between two pauses, greater that 40 msec. The duration of the intonation unit is defined as the duration of its segmental string and the pause after the segment.

The scheme for modeling F_0 contours proposed on this paper is based on results of analyzing F_0 contours, using the INTSINT system (**IN**ternational **T**ranscription **S**ystem for **INT**onation), proposed by D. Hirst [9, 10], which incorporates some ideas from TOBI (**TO**ne and **B**reak **I**ndex) transcription [11]. The analysis algorithm uses a spline fitting approach that reduces F_0 to a number of target points. The F_0 contour is built up by interpolation between these points. The target points can then be automatically coded into INTSINT symbols, but the orthographic transcription of the intonation units or boundaries must be manually introduced and aligned with the target points.

The system provides two sorts of tonal symbols:

- relative tones, which make reference only to the immediately preceding pitch-target, and
- absolute tones, which are assumed to refer to the speaker's overall pitch range over the current intonation unit.

In this study, the F_0 contour was built only from linguistic information. For generating an adequate F_0 shape, we need to know the relationship between linguistic units and the structures of an utterance with appropriate F_0 contour. This approach maximally reduces the amount of input prosodic information by applying a set of rules directly to the text. The so-called quantitative model of analysis and interpretation of the F_0 contour was proposed by many authors [12, 13, 14], with a differing number and complexity of the functions which try to simulate natural F_0 contours. In this paper, a global approach to modeling the F_0 contour is defined, mainly based on the so-called superpositional approach [14], which regards an F_0 contour as consisting of two different types of components:

- **global components** related to the whole intonation unit, and
- **local components** related to accented syllables or syntactic boundaries.

Global components rise in the beginning of the intonation unit and slightly decrease towards the end. This gives the baseline for the F_0 contour for the whole intonation unit. The local components

present local (rise, fall, rise and fall) movements of the shape at accented syllables or syntactic boundaries. Syntactic boundaries with local ascent often indicate the final F_0 shape at various types of intonation units. The generated F_0 contour is then represented as the sum of both components.

2. SPEECH MATERIAL

In order to generate rules for our synthesis scheme, data was collected by analysing the readings of ten speakers. All of them are native Slovene speakers, five males and five females. Eight of them (male and female equally) are professional speakers on national radio. The largest part of the speech material consists of declarative sentences, in short stories, monologues, containing sentences of various complexities and types, news, weather reports and commercial announcements. This speech data-base largely contains lexical emphasis and aims to be maximally intelligible and convincing. Other parts of the corpora are interrogative sentences with yes/no and wh-questions and imperative sentences. The first part of the corpora contains 500 declarative sentences, uttered by eight speakers, and the second part 100 questions and 30 imperative clauses uttered by 2 speakers.

3. ANALYSIS OF SLOVENE UTTERANCES

An intonation unit is defined as any unit of speech between two pauses, longer that 40msec. This length represents the low-limited value for distinguishing between different units of speech. The classical points of pauses in the speech occur:

- at prefaces, new paragraphs and new topics of readings,
- at the end of clauses,
- at places of prosodic phrases inside clauses
- at places of rhythmical division of some clauses or prosodic phrases, and
- at places of increased attention to some word or part of the text.

Depending on orthographic delimiters, four phrase boundaries were introduced:

1. boundaries without orthographic delimiters

   1.A at prefaces, between paragraphs, ...

   1.B at rhythmical divisions in the clause - before the Slovene grammatical words *in*, *ter* (and), *pa* (but), ...

2. boundaries with the delimiters '.' '...' '?' '!'
3. boundaries with the delimiters ',' ';' ':' '-' '(...)' '"..."'

The durations for various types of pauses are shown in figure 1. Taking into account the fact that speakers show vast variations of speaking style, for the average value a median was taken, because the mean value greatly depends on extreme values, often added for different reasons (physical and emotional states of the speaker, style, attitude,...).

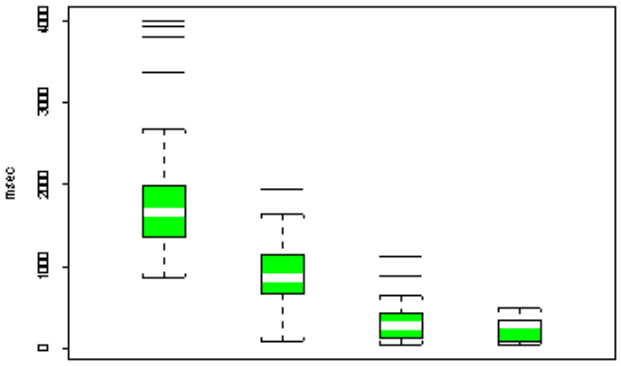

Figure 1: Average pause durations for the groups of orthographic delimiters: 1.A, 2., 3. and 1.B.

In the analysis of F_0 contour, the following parameters were studied:

- durations of intonation units
- onset frequency of intonation units
- offset frequency of intonation units
- frequency and number of syllables of the main accent
- frequency and number of syllables of secondary accents

A few studies of Slovene macroprosody mention that most of the main accents stay at the beginning (first three syllables) of the sentence [2, 6, 15]. Those studies agree with our results in table 1, showing the percentage of intonation units, where accents appear at the beginning (first three syllables) of the sentence, the percentage of intonation units with secondary accents and the percentage of accents at the end (last three syllables) of the intonation units. An accent was assumed to be any rise in frequency which differs more than 10 % in Hz from its vicinity.

Position of accents in the intonation unit	Percentage
At first three syllables	82
At last three syllables	52
At other places	61

Table 1: Percentage of intonation units with accents at different places in the unit.

4. THE SYNTHESIS SCHEME

After determination of word types, accent positions and (microprosodic) durations in the sentence, the proposed synthesis scheme consists in two main stages:

- text is divided into intonation units with punctuation marks and syntactic analysis
- main (and secondary) accents are determined by rule, considering the type of words and the type of prosodic phrase (declarative statement, interrogative statement, yes/no question, non-terminal,...).

The duration of pauses is determined from the type of intonation unit, as described in section 3. The duration of pauses is in the range between the first and the third quartile, as shown in the boxplots of figure 1 with the gray areas. Values closer to the median, denoted in the boxplots with the white stripe, have a greater probability. This stochastic variance in the range of pause durations prevents the synthetic, discrete nature of pauses in synthetic speech. The F_0 contour is defined with the function, composed from global (phrase) and local (accent) components, mentioned above. Many functions were tested (linear, power, transfer, decay, exponential) for the best approximation of the natural F_0 contour. In the system presented, an exponential function for the phrase component $P_c(t)$ [14, 16, 17] was adopted and a cosinusoidal function for accents and final boundary contours $A_c(t)$. The F_0 contour is thus defined by the following equation:

$$F_0(t) = P_c(t) + A_c(t)$$

where $P_c(t)$ and $A_c(t)$ are defined as:

$$P_c(t) = F_a \, e^{A_p \alpha \, t} e^{-\alpha t}$$

$$A_c(t) = A_a \, (1 + \cos \frac{T_a - t}{d})$$

where the expression $(Ta-t)$ must be in the range $(-\pi,\pi)$, otherwise $A_c(t) = 0$.

The symbols in these equations denote:

F_a: the asymptotic value of F_0

A_p: global F_0 maximum

α: parameter for F_0 shape control

T_a: time of accent

A_a: accent magnitude

d: local accent shape duration

The parameters F_a, A_p, α and A_a change during the synthesis process according to the analysis results of the F_0 contour. The parameter d models the microprosodic duration of accented syllables.

Figure 2 illustrates the results obtained. The sentence for comparison is uttered by a female speaker. The parameters for the synthesized F_0 are the same for the whole sentence. The panels display (top to bottom) the speech wave, the beginnings of orthographic notation for every intonation unit, the original F_0 contour modeled with the INTSINT system, indicated by squares, and the synthesized F_0 contour, generated with the equations presented, indicated by circles.

5. CONCLUSION

The paper describes an attempt to model the F_0 contour for Slovene intonation units by rules, generated through analysis of a large set of utterances. Analysis revealed that every speaker has a peculiar speaking style and that a great amount of individual differences in the patterns of phrasing and accentuation is present. One of the ways to define general intonation parameters could be to take the average values of large sets of utterances. The results of synthesized F_0 contour, based on average parameters, confirmed that the presented model could roughly, but realistically simulate the natural F_0 contour. In any case, the presented model, using just the given raw orthographic text, makes much better approximations to the natural F_0 contour than models which regard the F_0 contour as a linear function between different values at F_0 points. With additional information in the given text (especially levels and durations of local accents), the similarity of natural and synthesized F_0 contours was essentially improved. The analyzed speech corpus was limited, so that all aspects of the original speech could not be covered. Work towards the implementation of different speech rates and speaking styles still requires further prosodic and linguistic analysis and is currently in progress.

8. REFERENCES

1. Toporišič, J., *Slovenska stavčna intonacija*, V. seminar slovenskega jezika, literature in kulture, Faculty of Philosophy, University of Ljubljana, 1969.
2. Toporišič, J., *Slovenska slovnica*, Založba obzorja, Maribor, 1984.
3. Toporišič, J. et al., *Slovenski pravopis 1 - pravila*, Slovenska akademija znanosti in umetnosti, DZS, Ljubljana, 1994.
4. Šuštaršič, R., *Kontrastivna analiza angleške in slovenske stavčne intonacije*, Ph.D. thesis, Faculty of Arts, University of Ljubljana, 1994.
5. Rakar, A., *Modul makroprozodike za oblikovanje stavčne intonacije v okviru sinteze slovenskega govora*, B.Sc. thesis, Faculty of Electrical and Computer Engineering, University of Ljubljana, 1995.

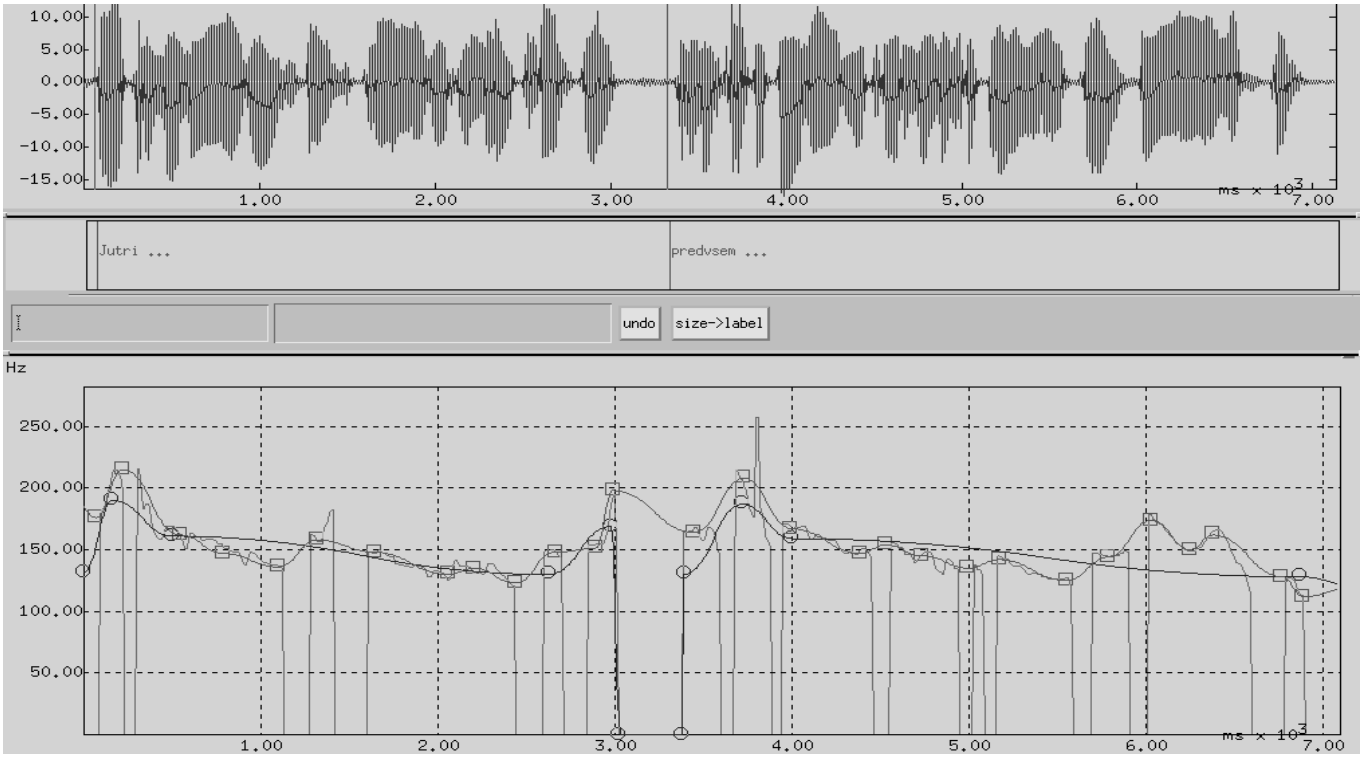

Figure 2: Example result of F_0 contour modeling for the Slovene sentence "Jutri bo jasno s spremenljivo oblačnostjo, predvsem popoldne in zvečer bodo še krajevne plohe in nevihte." Engl.: "Tomorrow will be clear, periodically cloudy, especially in the afternoon and in the evening local showers and storms could still come".

6. Vitez, P. and Aubergé, V. "Intonation Gesture of Slovene: First Indications", Proc. EUROSPEECH 95, Vol. 3, pp. 2073-2075, Madrid, 1995.

7. Sluijter, A.M.C., Terken, J.M.B. "The development and perceptive evaluation of a model for paragraph intonation in Dutch", Proc. ICSLP 92, Banff, Alberta, Canada, Vol. 1, pp. 353-356, 1992.

8. Terken, J.M.B., Collier, R. "Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch", Proc. EUROSPEECH 89, Edinbourgh, Vol. 1, pp. 357-359, 1989.

9. Nicolas, P. and Hirst, D.J. "Symbolic Coding of Higher-Level Characteristics of Fundamental Frequency Curves", Proc. EUROSPEECH 95, Vol. 3, pp. 2065-2068, Madrid, 1995.

10. Hirst, D. and Espesser, R. "Automatic Modelling of Fundamental Frequency", *MULTEXT*, LRE PROJECT 62-050, Task 2.6 Prosody Tools, Deliverable 2.6.1, Version B, Centre National de la Recherche Scientifique, 1995.

11. Silverman, K. et al. "TOBI: A standard for labeling English prosody", Proc. ICSLP 92, Banff, pp. 867-870, 1992

12. Moore, C.A. et al. "Quantitative description and differentiation of fundamental frequency contours", Computer Speech and Language, Vol. 8, Num. 4, pp. 385-404, 1994.

13. Taylor, P. "The rise/fall/connection model of intonation", Speech Communication, Vol. 15, Num. 1-2, pp. 169-186, 1994.

14. Fujisaki, H. and Ohno, S. "Analysis and modeling of fundamental frequency contours of English utterances", Proc. EUROSPEECH 95, Madrid, Vol. 2, pp. 985-988, 1995.

15. Aubergé, V. and Bailly, G. "Generation of intonation: a global approach", Proc. EUROSPEECH 95, Vol. 3, pp. 2065-2068, Madrid, 1995.

16. Mixdorf, H. and Fujisaki, H. "A scheme for a model-based synthesis by rule of F_0 contours of German utterances", Proc. EUROSPEECH 95, Vol. 3, pp. 1823-1826, Madrid, 1995.

17. Hirai, T., Higuchi, N. and Sagisaka, Y. "Automatic detection of major phrase boundaries using statistical properties of superpositional F_0 control model parameters", Proc. EUROSPEECH 95, Vol. 2, pp. 1341-1344, Madrid, 1995.