COMPONENTS OF A QUANTITATIVE MODEL OF GERMAN INTONATION

Bernd Möbius
AT&T Bell Laboratories, Murray Hill, NJ, USA

ABSTRACT

In this paper a quantitative description of German intonation is presented. It will be demonstrated that intonation contours can be efficiently analyzed, and predicted, by interpreting the components and parameters of Fujisaki's model in terms of linguistic features and categories. It will also be argued that a superpositionally organized model is particularly suitable for a quantitatve description.

STRUCTURE OF INTONATION

Tone Sequences Or Layered Components?

Two major classes of intonation models have evolved in the course of the last two decades. There are, on the one hand, hierarchically organized models which interpret *F*0 contours as a complex pattern resulting from the superposition of several components. Their counterparts are usually seen in the models which claim that *F*0 contours are generated from a sequence of phonologically distinctive tones, or categorially different pitch accents, that are locally determined and do not interact.

Two quotations illustrate the competing points of view:

"[...] the pitch movements associated with accented syllables are themselves what make up sentence intonation [...] there is no layer or component of intonation separate from accent: intonation consists of a sequence of accents, or, to put it more generally, a sequence of tonal elements." ([9], p. 40)

"[...] Standard Danish intonational phenomena are structured in a hierarchically organized system, where components of smaller temporal scope are superposed on components of larger temporal domain [...] These components are simultaneous, parametric, non-categorical and highly interacting in their actual production." ([25], p. 2)

Ladd [9, 10] argues that although the tone sequence and the superpositional models diverge in formal and notational terms, they nevertheless may be more similar from a descriptive point of view than usually admitted. Although I agree with the argument ([24], p. 1041) that the two types of intonation models not only differ in formal respect but from a conceptual point of view as well, I don't think they are ultimately incompatible. As a matter of fact, in more recent publications (e.g., [11]) Ladd proposed a metrical approach that incorporates both linear and hierarchical elements.

The main difference between the 'pure' linear and overlaying models can be seen in how the relation between local movements and global trends in the intonation contour is defined, or, in other words, in the view of the relation between word accent and sentence intonation. The underlying problem is that word- and utterance-(or phrase-)prosodic aspects all express themselves by one and the same acoustic variable: the variation of fundamental frequency as a function of time. There is no way of deciding either by acoustic measurements or by perceptual criteria whether *F*0 movements are caused by accentuation or by intonation. A separation of these effects, however, can be done on a linguistic, i.e. more abstract, level of description. Here rules can be formulated that predict accent- or intonation-related patterns independent of, as well as in interaction with, each other.

Autosegmental theory allows for the independence of various levels of suprasegmental description and their respective effects on the intonation contour by an appropriate phonological representation. According to Edwards and Beckman [2], the most promising principle of intonation models ought to be seen in the capability to determine the effects of each individual level, and of their interactions. Although probably not intended by the authors, this is precisely the most important argument in favor of a hierarchical approach and of superpositional models of intonation. Thus, the conceptual gap between the different theories of intonation does not seem to be too wide to be bridged.

After presenting supporting data, I will continue this line of argument in the concluding section.

Motivation For A Superpositional Approach

Even among researchers representing different types of intonation models there is widespread agreement on the fact that the *F*0 contour of an utterance should be regarded as the complex result of effects exerted by a multitude of factors. Some of these factors are related to articulatory or segmental effects but others clearly have to be assigned to linguistic categories.

In contradiction to the explicit assumption in [20] that intonation is determined exclusively on a local level, there is ample evidence for non-local factors. In a study of utterances containing parentheses [8], the authors show that the intonation contour is interrupted by the parenthesis, and resumed right afterwards in a way the contour would have looked like in the 'same' utterance without parenthesis. Also, in [12] the authors explain how the first accent peak in an utterance is adjusted depending on the underlying syntactic constituent structure. Furthermore, there is some evidence that the speaker pre-plans the global aspects of the intonation contour, not only with respect to utterance-initial *F*0 values but to phrasing and inter-stress intervals as well [23].

These considerations obviously favor models that directly represent both global and local properties of intonation. These models also provide a way of extracting prosodic features related to the syntactic structure of the utterance and to sentence mode. Generally speaking, the analytical separation of all the potential factors considerably helps decide under which conditions and to what extent the concrete shape of a given *F*0 contour is determined by linguistic factors (including lexical tone), non-linguistic factors, such as, e.g., intrinsic and coarticulatory *F*0 variations, and speaker-dependent factors.

Superpositionally organized models lend themselves to such a quantitative approach: Contours generated by such a model result from an additive superposition of components that are in principle orthogonal to, or independent of, each other. The components in turn can be related to certain linguistic or non-linguistic categories. Thus, the factors contributing to the variability of *F*0 contours can be investigated separately. In addition, the temporal course pertinent to each individual component can be computed independently. A production-oriented model providing components for accentuation on the one hand and sentence or phrase intonation on the other hand and generating the pertinent patterns by means of parametric commands appears to be particularly promising.

The only approach exploiting the principle of superposition in a strictly mathematical sense is the model proposed by Fujisaki and co-workers (e.g., [5, 3, 4]). This particular model has several advantages. Since it satisfies the principle of superposition, the respective effect of a given factor can be determined for a predefined temporal segment or for a given linguistically or prosodically defined unit, such as a phrase or a stress group. For every desired point in time in the course of an utterance, the resulting *F*0 value can be computed. The values of the model parameters (see following section) are constant at least within one stress group. This data reduction can be an important aspect for certain applications like speech synthesis. The smooth contour resulting from the superposition of the model's components is appropriate for the approximation of naturally produced *F*0 contours.

Generally speaking, adequate models are expected to provide both predictive and explanatory elements [1]. In terms of prediction, models have to be as precise and quantitative as possible, ideally being mathematically formulated. A model provides explanations if it is capable of analyzing a complex system in such a way that both the effects of individual components and their combined results become apparent. Fujisaki's model meets both requirements; and all effects can be described uniquely by their causes.

The model does not, however, explain by itself why a given component behaves the way it does. The particular approach and the application presented in this paper aim at providing these explanations, especially by applying a linguistic interpretation of the model's components.

Another explanatory approach can be seen in the potential physiological founda-

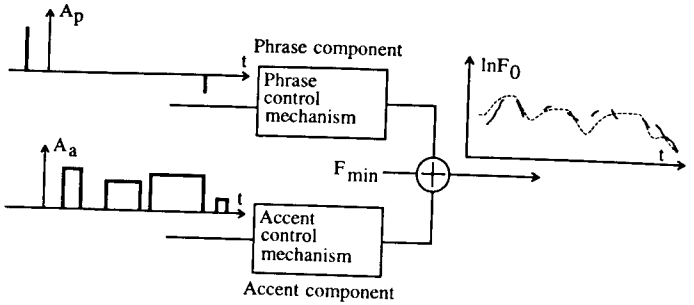

Figure 1. Block diagram of Fujisaki's quantitative model that additively superposes a basic Fo value (F_{min}), a phrase component, and an accent component on a logarithmic scale (lnFo). The control mechanisms of the components respond to impulse (phrase component) and rectangular commands (accent component), respectively (A_P = amplitude of phrase commands; A_a = amplitude of accent commands; t = time).

tion, in terms of laryngeal structures and interactions of laryngeal muscles, as discussed by Fujisaki [3]. His model is the only one I am aware of that explicitly includes a quantitative simulation of the *Fo* production and control mechanisms inherent in a human speaker; the approach is based on work by Öhman and Lindqvist [18].

The model represents each partial glottal mechanism of fundamental frequency control by a separate component. Although it does not include a component that models intrinsic or coarticulatory *Fo* variations, such a mechanism could easily be added in case it is considered essential for natural-sounding synthesis.

A QUANTITATIVE MODEL OF INTONATION

Since Fujisaki's model has been described by the original authors on many occasions, I will restrict myself to only presenting the most important properties of the model. I will focus instead on motivating the linguistic interpretation of the components as it emerged from applying the model to the analysis of German intonation.

The model additively superposes a basic *Fo* value (F_{min}), a phrase component, and an accent component on a logarithmic scale (Figure 1). The control mechanisms of the two components are realized as critically damped second-order systems responding to impulse functions in the case of the phrase component, and rectangular functions in the case of the accent component. These functions are generated by two different sets of parameters: the timing and amplitudes of the phrase commands as well as the damping factors of the phrase control mechanism on the one hand, and the amplitudes and the timing of the onsets and offsets of the accent commands as well as the damping factors of the accent control mechanism on the other hand. All these parameter values are constant for a defined time interval: the parameters of the phrase component within one prosodic phrase, the parameters of the accent component within one accent group, and the basic value F_{min} within the whole utterance.

The *Fo* contour of a given utterance can be decomposed into the components of the model by applying an analysis-by-synthesis procedure. This is achieved by successively optimizing the parameter values, eventually leading to a close approximation of the original *Fo* curve. Thus, the model provides a parametric representation of intonation contours.

Linguistic Interpretation

As I have argued in more detail elsewhere [14, 17], the quantitative description of intonation can be more efficient if modeling a given *Fo* contour and extracting the pertinent parameters is subjected to the constraints given by a linguistic and prosodic interpretation in the first place and by the criterion of optimal approximation in a mathematical sense only in the second place.

Here are the key elements of my interpretation of the model:

- The phrase component of the model represents the global slope and the slow variations of the *Fo* contour in the utterance. Obviously, the phrase component is very suitable to describe F0 declination since the phrase contour reaches its maximum rather early and descends monotonically along the major part of the utterance. Therefore, the contour that results from adding the basic value F_{min} to the phrase component serves as a baseline of the intonation contour, the magnitude of the phrase command amplitude being a direct measure for *Fo* declination in the utterance.

- Besides the obligatory utterance-initial phrase command, additional phrase commands are only provided at major syntactic boundaries, e.g., between main and subordinate clauses, thereby resetting the declination line. The procedure of inserting phrase commands wherever the criterion of optimal approximation seems to demand it [5] is rejected.

- The conspicuous final lowering of *Fo* which is regularly observed in declarative utterances and often in wh-questions is modeled by a negative phrase command. Likewise, we provide positive utterance-final phrase commands for other sentence modes, such as yes/no and echo questions. Thus, the phrase component of the model can be related to the linguistic category *sentence mode*, via the shape of the phrase contour and the underlying commands and parameter values. There are both global (the overall slope) and local (final rise or lowering) cues that contribute to differentiating between sentence modes.

- Local *Fo* movements that are associated with accented syllables are represented by the accent component and superposed onto the global contour. Closely following Thorsen's definition of *stress groups* [22] I apply an accent group concept, an accent group being defined as a prosodic unit that consists of an accented syllable optionally followed by any number of unaccented syllables. Accent groups are independent of word boundaries but sensitive to major syntactic boundaries, as will be shown below.

The concept of accent groups fits in the hierarchical structure of the model. While the linguistic category *sentence mode* is reflected in the phrase component, the linguistic feature *word accent* is manifested in the locations and shapes of accent commands. Consequently, the *Fo* course of a given accent group should be modeled by the contour generated by exactly one accent command. Thus, the parameter configurations of the accent component can be interpreted as correlates of the linguistic feature word accent.

ANALYSIS OF GERMAN INTONATION

Estimation Of Parameter Values

In principle, the parameter values that approximate the *Fo* contour of a given utterance can be determined automatically or by hand. Nevertheless, only an automatic procedure guarantees that the optimal values are extracted in an objective and reproducible way. Preliminary experiments showed that there are considerable intra- and interindividual divergencies when an interactive, i.e., partly manual method is used. Therefore, the parameter values of the model are determined by means of a computer program [19] that automatically approximates measured *Fo* contours by successively optimizing the parameters within the framework of the linguistic interpretation of the model. Input information is the *Fo* data for the given utterance and the locations of accent group boundaries.

Based on the principle of superposition, determination of the phrase command parameters and the basic value F_{min}, which is the first step in the algorithm, can be separated from the subsequent determination of the accent command parameters. The contour resulting from F_{min} and the phrase parameters is approximated to the measured *Fo* curve. Once the parameters of the phrase component are optimized, the resulting differential signal is interpreted by the accent component of the model.

The accent component is made up of partial contours that are in turn generated by accent commands. Each accent group is modeled by the contour resulting from exactly one accent command. Algorithmically speaking, the individual accent groups are processed from left to right in a non-iterative way.

Figure 2 illustrates the close approximation of a measured *Fo* contour.

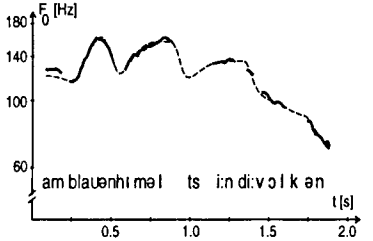

Figure 2. Close approximation (dashed line) of the Fo contour of the declarative utterance "Am blauen Himmel ziehen die Wolken" (male voice).

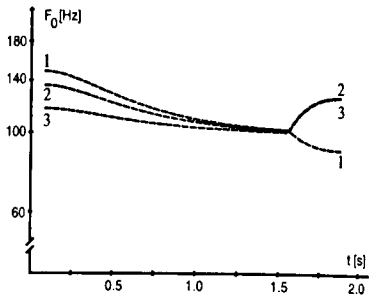

Figure 3. Typical phrase contours for the interrogative modes wh-question (1), yes/no question (2), and echo question(3), for Fmin=100 Hz.

Results

The speech materials cover declarative sentences with one or two major syntactic clauses, the latter realized as two prosodic phrases, and three types of interrogatives, namely echo, yes/no, and wh-questions. Speech data for six male and three female speakers were collected under 'laboratory' conditions.

The potential sources of variation of the parameter values were explored by means of statistical procedures, taking into account both linguistic and speaker-dependent factors. The results have been presented at full length elsewhere [14], so only the major trends and findings will be presented here.

Damping factors. The damping factors of the phrase and accent components are treated as constants. My experiments confirm the claim that the approximation of *Fo* contours is not impaired by this assumption [3]. For the phrase component, a standard value of 3.1/s is both appropriate for the purpose of approximation and reasonable as far as the physiological foundation of the model is concerned. A constant value of 16/s corresponding to the arithmetic mean for all speakers and all accent groups is suitable for the damping factor of the accent component.

Basic value *Fmin*. For all speakers, the dispersion of the basic value *Fmin* is relatively small, yielding 50% of the observed values within the range of about 3.0 Hz around the arithmetic mean for the respective speaker. This finding suggests that it is reasonable to keep *Fmin* constant for a given speaker. Typical values are 75-80 Hz for male speakers and 145-150 Hz for female speakers.

Phrase command timing and amplitude. Since the phrase component serves as a baseline to the intonation contour with the peak of each phrase contour coinciding with the beginning of the utterance, or the prosodic phrase, the exact timing of a phrase command directly depends upon the value of the damping factor (3.1/s). Therefore, the first phrase command is set at 323 ms before the onset of the utterance. This agrees with findings from studies on *Fo* production and control which reveal prephonatory activities of the laryngeal muscles [7].

Phrase command amplitudes are largely speaker, or rather speaker type, dependent. Sentence mode is the most important linguistic factor; it is globally signaled by the contour of the phrase component. While phrase contours of wh-questions are very similar to those of declaratives, yes/no-questions and the syntactically unmarked echo questions show a much less steep declination (see also [22] for Danish). Typical phrase contours for these three interrogative modes are shown in Figure 3. No consistent dependency of phrase command amplitude upon utterance duration or speech tempo was observed.

Accent command amplitude. The values of the accent command amplitudes split the speakers into two groups. The location of the accent group in the utterance turned out to be the most important linguistic factor. Utterance-final accent commands show significantly smaller amplitudes than accent commands in any other position in the utterance. Other important factors are the part-of-speech

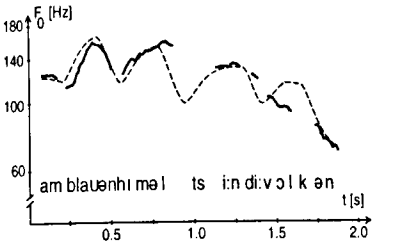

Figure 4. Rule-generated Fo contour (dashed line) compared to the original Fo contour of the declarative utterance "Am blauen Himmel ziehen die Wolken." (male voice); cf. Figure 2.

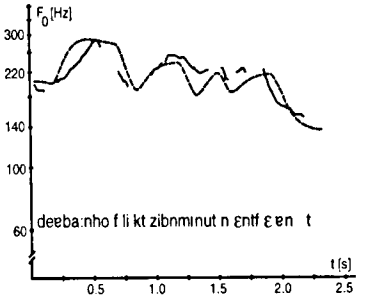

Figure 5. Rule-generated Fo contour (dashed line) compared to the original Fo contour of the declarative utterance "Der Bahnhof liegt sieben Minuten entfernt." (female voice).

class for the word carrying the accent, with nouns requiring higher amplitudes than other classes, and the presence of a phrase boundary. Amplitudes of accent commands preceding a phrase boundary tend to be about 25% higher than in other positions.

Accent command duration. Duration of an accent command can be reliably predicted from the duration of the respective accent group. There is a high correlation ($r = 0.84$) between these two variables, i.e., more than 70% of the variance observed in durations of accent commands can be explained by accent group duration. An effect of phrase-final lengthening is observed for several speakers.

Accent command position. The most important factor controlling the relative temporal position of an accent command within a given accent group is the location of the accent group in the utterance. While in non-final positions the temporal distance between the beginning of the accent group and the command onset is about 10% of the accent group duration, it tends toward zero in utterance-final accent groups.

Fo Synthesis By Rule

Parameters are adjusted by rules based on the analysis described above. The rules capture speaker dependent as well as linguistic features, such as sentence mode, sentence accent, phrase boundary signals, or word accent, and generate an artificial intonation contour for a given target utterance. The input information needed is the location of accented syllables in the utterance, the durations of accent groups, and, although less important, part-of-speech information for the words that carry accents.

Since the rules are based on the results of statistical analyses, the parameter values they provide are averages, producing contours that were not actually observed for any real speaker. On the other hand, they were shown to capture speaker-dependent features; they produce intonation patterns that to a fair degree correspond to what the modeled speaker could have produced. Thus, one should expect a mixture of frequently very good predictions with occasionally rather poor ones, the latter being due to either insufficient data or inadequate predictive power for a particular context.

Illustrations of *Fo* contours generated by rule are given in Figures 4, 5, and 6.

The adequacy of the rules was tested in a series of perceptual experiments whose results are presented elsewhere [15, 16]. The rules have been implemented in the German concatenative speech synthesis system HADIFIX developed at IKP Bonn [21].

CONCLUSION

In conclusion, I resume the controversial discussion of tone-sequential and superpositional intonation models, taking the prosodic marking of phrase boundaries as a starting point. The results presented here indicate that major syntactic boundaries invoke a resetting of declination, or the *Fo* baseline, which is realized in my quantitative model by inserting a phrase command. Additionally, signaling of the

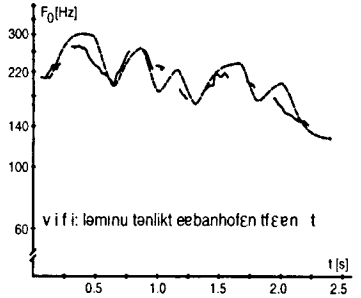

Figure 6. Rule-generated F0 contour (dashed line) compared to the original F0 contour of the wh-question "Wieviele Minuten liegt der Bahnhof entfernt?" (female voice).

boundary is enhanced by parameters of the accent component that are sensitive to major syntactic boundaries. The hierarchical structure of the model enhances elaborating the respective effects of the factors involved, even though both strategies of signaling phrase boundaries make use of the same phonetic variable.

It is important to note, though, that the notion of *hierarchy* is not necessarily an appropriate criterion for differentiating tone sequence and superpositional models, especially since its meaning is ambiguous. Both types of concepts contain hierarchical elements in the sense that utterances consist of prosodic phrases which in turn consist of accent groups or pitch accents; and even the most influential tone sequence model [20] provides a non-local element, i.e. declination. There is another meaning of *hierarchy*: making choices in various components of the prosodic system of a given language, higher levels having priority over, and setting constraints for, lower levels (cf. [25]). In the superpositional model presented here, however, there is no preponderance of one component over another.

Furthermore, since all the models discussed here (explicitly or implicitly) assume a mechanism of pre-planning in speech production, the difference between them should rather be seen in terms of how they represent this mechanism. Tone sequence models provide a higher *F0* onset in longer utterances but the relations between the individual pitch accents are not affected. According to [6], utterance length determines the slope of declination, short utterances having a steeper baseline, but not the utterance-initial *F0* value.

The formulae given by [13] in their version of the linear tone sequence approach are based on the analysis and approximation of intonation contours. Meaning is only assigned to the relations between pitch accents which are in turn defined by the feature of downstepping. However, it seems to be more appropriate to also assign meaning to the arguments in the formula, i.e., to the variables and constants. Genuinely superpositional models meet this requirement: The output behavior of the model as a response to the sum of several input signals can be predicted from the responses to each of the individual input signals.

Arguing in favor of a hierarchical organization of prosodic systems does not imply a rejection of phonological approaches. On the contrary, the integration of a superpositionally organized intonation model with an underlying phonological representation of the prosodic system of a given language is ultimately desirable. The phonological foundation of the quantitative model for German presented here remains a desideratum.

ACKNOWLEDGMENTS

The experiments presented in this paper were done at IKP, Univ. of Bonn, supported by grants from the German Research Council (DFG) and the German Federal Ministry of Research and Technology (BMFT). The author wishes to thank Grazyna Demenko, Wolfgang Hess, Julia Hirschberg, Matthias Pätzold, Thomas Portele, and Jan van Santen, for support and valuable discussions.

REFERENCES

[1] Cooper, F.S. (1983): "Some reflections on speech research", In P.F. MacNeilage (ed.), *The production of speech*, New York: Springer, pp. 275-290.

[2] Edwards, J., Beckman, M.E. (1988): "Articulatory timing and the prosodic interpretation of syllable duration", *Phonetica*, vol. 45, pp. 156-174.

[3] Fujisaki, H. (1983): "Dynamic characteristics of voice fundamental frequency in speech and singing", In P.F. MacNeilage (ed.), *The production of speech*, Berlin: Springer, pp. 39-55.

[4] Fujisaki, H. (1988): "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", In O. Fujimura (ed.), *Vocal physiology: voice production, mechanisms and functions*, New York: Raven, pp. 347-355.

[5] Fujisaki, H., Hirose, K., Ohta, K. (1979): "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese", *Annual Bulletin of the Research Institute for Logopedics and Phoniatrics (Tokyo)*, vol. 13, pp. 163-172.

[6] Grønnum, N. (1990): "Prosodic parameters in a variety of Danish standard languages, with a view towards Swedish and German", *Phonetica*, vol. 47, pp. 182-214.

[7] Jafari, M., Wong, K.-H., Behbehani, K., Kondraske, G.V. (1989): "Performance characterization of human pitch control system: an acoustic approach", *Journal of the Acoustical Society of America*, vol. 85, pp. 1322-1328.

[8] Kutik, E.J., Cooper, W.E., Boyce, S. (1983): "Declination of fundamental frequency in speaker's production of parenthetical and main clauses", *Journal of the Acoustical Society of America*, vol. 73, pp. 1731-1738.

[9] Ladd, D.R. (1983): "Peak features and overall slope", In A. Cutler, D.R. Ladd (eds.), *Prosody: models and measurements*, Berlin: Springer, pp. 39-52.

[10] Ladd, D.R. (1983): "Phonological features of intonational peaks", *Language*, vol. 59, pp. 721-759.

[11] Ladd, D.R. (1993): "In defense of a metrical theory of intonational downstep", In H. van der Hulst, K. Snider (eds.), *The phonology of tone. The representation of tonal register*, Berlin: Mouton de Gruyter, pp. 109-132.

[12] Ladd, D.R., Johnson, C. (1987): "'Metrical' factors in the scaling of sentence-initial accent peaks", *Phonetica*, vol. 44, pp. 238-245.

[13] Liberman, M.Y., Pierrehumbert, J. (1984): "Intonational invariance under changes in pitch range and length", In M. Aronoff, R. Ochrle (eds.), *Language sound structure*, Cambridge: MIT Press, pp. 157-233.

[14] Möbius, B. (1993): *Ein quantitatives Modell der deutschen Intonation—Analyse und Synthese von Grundfrequenzverläufen*, Tübingen: Niemeyer.

[15] Möbius, B. (1993): "Perceptual evaluation of rule-generated intonation contours for German interrogatives", *Working Papers (Dept. of Linguistics and Phonetics, Univ. Lund) (=Proceedings of the ESCA Workshop on Prosody, Lund, 27-29 Sept. 1993)*, vol. 41, pp. 216-219.

[16] Möbius, B., Pätzold, M. (1992): "F0 synthesis based on a quantitative model of German intonation", *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, vol. 1, pp. 361-364.

[17] Möbius, B., Pätzold, M., Hess, W. (1993): "Analysis and synthesis of German *F0* contours by means of Fujisaki's model", *Speech Communication*, vol. 13, pp. 53-61.

[18] Öhman, S.E.G., Lindqvist, J. (1966): "Analysis-by-synthesis of prosodic pitch contours", *Royal Inst. of Technology (Stockholm), STL-QPSR*, vol. 4 (1965), pp. 1-6.

[19] Pätzold, M. (1991): Nachbildung von Intonationskonturen mit dem Modell von Fujisaki - Implementierung des Algorithmus und erste Experimente mit ein- und zweiphrasigen Aussagesätzen (Ms., Univ. Bonn).

[20] Pierrehumbert, J. (1980): The phonology and phonetics of English intonation (Diss., MIT, Cambridge).

[21] Portele, T., Steffan, B., Preuss, R., Sendlmeier, W.F., Hess, W. (1992): "HADIFIX - a speech synthesis system for German", *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, vol. 2, pp. 1227-1230.

[22] Thorsen, N. (1979): "Lexical stress, emphasis for contrast, and sentence intonation in advanced standard Copenhagen Danish", *Annual Report of the Institute of Phonetics (Univ. Copenhagen), ARIPUC*, vol. 13, pp. 59-85.

[23] Thorsen, N.G. (1985): "Intonation and text in Standard Danish", *Journal of the Acoustical Society of America*, vol. 77, pp. 1205-1216.

[24] Thorsen, N.G. (1986): "Sentence intonation in textual context - supplementary data", *Journal of the Acoustical Society of America*, vol. 80, pp. 1041-1047.

[25] Thorsen, N.G. (1988): "Standard Danish intonation", *Annual Report of the Institute of Phonetics (U Copenhagen), ARIPUC*, vol. 22, pp. 1-23.