Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Neofilologii
Instytut Językoznawstwa

A comprehensive model of intonation for application in speech synthesis

Agnieszka Wagner

Rozprawa doktorska

Opiekun naukowy:
Prof. UAM Dr hab. Inż. Grażyna Demenko

Acknowledgements

The work presented in this thesis was carried out within the scope of the research grant no. 1 H01D 003 30 received from the Minister of Science and Higher Education.

Badania przedstawione w pracy zostały zrealizowane w ramach grantu promotorskiego nr 1 H01D 003 30 przyznanego przez Ministra Nauki i Szkolnictwa Wyższego.

Contents

Chapter 1. Introduction

In this chapter the research area of the thesis - *intonation modeling* is presented. At first, a brief explanation is given why intonation is important and why there is a need for intonation models in speech synthesis and other speech applications. Then the goal of the thesis is defined, namely *development of a comprehensive intonation model for application in speech synthesis*. In the next sections, the terms that appear in the title of the thesis are discussed. Various definitions of *intonation* which can be found in the literature are discussed and the sense in which *intonation* is used in the current work is explained in detail. The term *comprehensive* is defined in the context of intonation modeling and components and tasks of intonation models are specified. This is followed by an overview of speech synthesis systems, because depending on the synthesis type the output of an intonation model may be used in a slightly different way.

1.1. Preliminaries

In this section the role of intonation in speech is outlined and the terms used in the title of the thesis are explained.

1.1.1. Why intonation is important?

The area of research presented in this dissertation is *intonation* and the final goal is to provide a *comprehensive intonation model for application in speech synthesis systems*. The other application domain where the results presented in the thesis can be useful is speech recognition.

As regards speech synthesis, the reason for the development of intonation models is generation of a naturally sounding synthetic speech. Apart from that, an intonation model can be used in intonation research as a framework within which intonation is described and compared across languages.

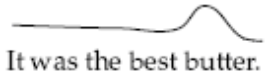
As regards the role of intonation in speech recognition the intonation content of utterances provides a lot of useful information for example, the phrasing information can be used to resolve ambiguous parses, the information on the location of pitch accents indicates that a syllable is stressed and can be used to determine the lexical content of utterances.

The short discussion presented below will explain to some extent why intonation has to be taken into account in both speech synthesis and recognition.

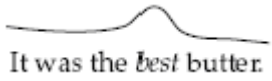
First of all, intonation constitutes an inherent, suprasegmental component of speech and conveys information concerning speaker's message which significantly affects its interpretation

by the listener. The examples of two utterances presented below (adopted from Clark 2003:3) illustrate some of the intonation functions.

(1)



(2)



The utterances have the same segmental structure and differ in the intonational features. The shift of the pitch accent from the penultimate syllable of the word "butter" to the word "best" is an example of the intonation function known as *focus*. The intention of the speaker was to highlight different part of the message and consequently, the two messages are interpreted differently by the listener. The utterance in (1) has a pitch contour typical for an unemphatic statement, whereas the accentuation in (2) produces an emphatic contour.

One of the reasons why intonation modeling is necessary in speech synthesis regards perception. Listeners are less sensitive to errors in the segmental structure of an utterance than to those which concern suprasegmental features and which may lead to misinterpretation of the message or completely blur it. What is more, even if the segmental content of the message is blurred, its interpretation on the basis of intonational features is still possible to some extent – even if it is hard to understand what the speaker is talking about it is possible to get to know:

a) whether he gives a suggestion or a request, or whether he asks about something or states some facts
b) his current attitude and emotional state e.g., whether he is angry or sad
c) some features of the speaker himself such as sex or age

The information in a) is an example of the *linguistic* message conveyed by intonation. *Paralinguistic* messages like those given in b) are regarded as modifications of the linguistic meanings e.g., "in an accent contour that rises to a peak and then falls to the bottom (...) the height of the peak can vary paralinguistically to convey greater emphasis without affecting the linguistic identity of the contour" (Ladd 1996:39). Consequently, it is assumed that linguistic and paralinguistic intonational meaning should be represented and analyzed separately - on the phonological and phonetic level respectively (see discussion in Ladd 1996:38ff). The issues related to intonational meaning are discussed in a greater detail in the next sections.

The third message conveyed by intonation presented in c) is regarded as *non-linguistic* or *extralinguistic* and refers to features of a speaker such as sex or age e.g., female and children voices are much higher than male voices.

This short presentation shows that intonation is involved in conveying various messages which can not be derived from the text or segmental structure of the utterance. For this reason intonation has to taken into account in speech applications.

1.1.2. Definition of intonation

The first term to be defined is *intonation*.

In a broad sense intonation includes *lexical features: stress, accent* and *tone*. In a narrow sense intonation is restricted to *non-lexical* (also: *postlexical* or *supralexical*) features "consisting of such phenomena as the overall form of pitch patterns, declination, boundary phenomena etc." (Hirst & Di Cristo 1998:4). In the second, narrow sense intonation conveys *sentence-level meanings* i.e., "meanings that apply to phrases or utterances as a whole, such as sentence type or speech act, or focus and information structure" (Ladd 1996:7). The sentence-level meanings are regarded as linguistic. Intonation defined in the way presented here is referred to as *intonation proper*. In this sense the term *intonation* is used also by other authors e.g. Gussenhoven and Cruttenden: "intonation refers to the structured variation in pitch which is not determined by lexical distinctions as in tone languages" (Gussenhoven 2006:1) and "intonation involves the occurrence of recurring pitch patterns, each of which is used with a set of relatively consistent meanings, either on single words or on groups of words of varying length" (Cruttenden 1997:7).

Intonation proper is represented and analyzed on the *abstract, cognitive, phonological level* in a *linguistically structured way*: "intonational features are organized in terms of categorically distinct entities (e.g. low tone or boundary rise) and relations (e.g. stronger than/weaker than). They exclude 'paralinguistic' features, in which continuously variable physical parameters (e.g. tempo and loudness) directly signal continuously variable states of the speaker (e.g. degree of involvements or arousal)" (Ladd 1996:8).

In (Hirst & Di Cristo 1998:7) one more definition of intonation is proposed. Intonation is regarded as an *intermediate system on the phonetic level*, "a construction by which prosodic primitives on the lexical level and the non-lexical level, however we choose to represent these formally, are related to acoustic prosodic parameters" (op.cit.:7). So, high-level phonological representations of intonation are mapped onto/derived from intermediate phonetic representations, and the latter are mapped onto/derived from prosodic parameters (f0, intensity, duration, spectral emphasis) on the acoustic, physical level of description.

Prosody is defined as a system which comprises lexical and non-lexical phenomena on the one hand, and prosodic parameters on the other.

In Figure 1 various interpretations of intonation discussed here are illustrated. On the cognitive phonological level lexical and non-lexical (intonation proper) prosodic systems are represented and analyzed. The physical acoustic level is the level of prosodic parameters. In between these two levels there is space for the phonetic level where intonation is described and analyzed in terms of continuous parameters which can be mapped onto phonological categories on the one hand and acoustic parameters on the other.

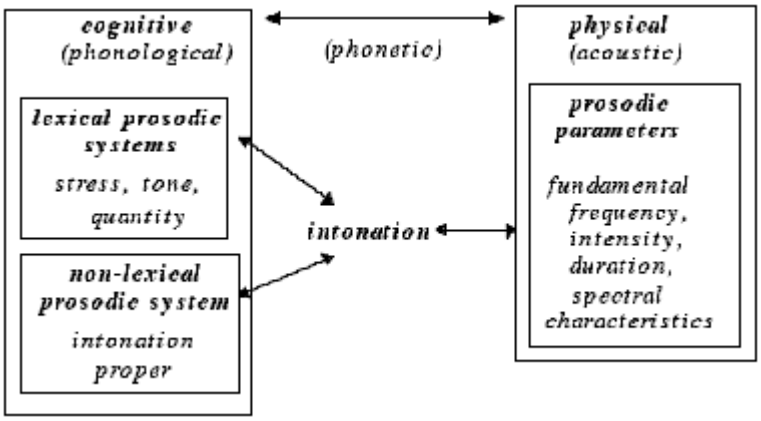


Figure 1 (adopted from Hirst & Di Cristo 1998:7): Illustration of the general prosodic characteristics of the languages.

As it can be concluded from the above discussion, the problem of defining intonation is related to the issue of *levels of representation and analysis for the description of intonation*.

Apart from the three levels already mentioned i.e., *physical (acoustic), phonetic* and *phonological (cognitive, linguistic*) fourth level of description of intonation is proposed, namely the *surface phonological level* (Hirst, Di Cristo & Espesser 2001:4).

The authors point out that a global theory of intonation should take all these levels into account and a constraint on the representations on various levels is formulated according to which "intermediate representations must be interpretable at both adjacent levels: the more abstract and the more concrete" (Hirst, Di Cristo & Espesser 2000:3). Below the most important features of various representations are discussed starting from the most concrete level and ending at the most abstract one.

The physical acoustic level is the level of *suprasegmental features*: "observable and measureable physical parameters" of fundamental frequency, duration, intensity and spectral tilt (Hirst & Di Cristo 1998:4). Among them "fundamental frequency is universally acknowledged to be the primary parameter" (op.cit:4). *Fundamental frequency* (*f0*) is described in terms of the rate of vocal folds vibrations and measured in Hz. In this thesis fundamental frequency/f0 will be used interchangeably with *pitch* which is its auditory correlate. *Intensity* is described in terms of strength of subglottal pressure; its acoustic correlate is *loudness*. *Duration* is the length of time necessary to set articulators in a proper position to produce a sound. *Spectral characteristics* also belong to prosodic parameters, their auditory correlate is *timbre*.

The *phonetic level* is the level of phonetic descriptions which use continuous parameters to describe intonation. It is known that apart from speaker-based influences, pitch is also affected by segmental and suprasegmental structure of an utterance. Individual speech segments produce minor perturbations in the f0 contour which are referred to as *microprosody*. Microprosody does not contribute to intonational meaning and therefore, from the point of view of intonation modeling, is unimportant. There exists a number of intonation stylization methods which aim at elimination of the microprosodic component of an f0 curve - what is left is the macroprosodic component reflecting the choice of intonation pattern for the utterance (Hirst & Espesser 1993). Phonetic representations are *intermediate* between and can be mapped both

onto physical and phonological representations. Examples of phonetic representations are Momel, Tilt, PaIntE and Fujisaki's model described in Chapter 3.

Above the phonetic level, there is the level of *surface phonological descriptions* (e.g. INTSINT, see sec. 3.5). On the surface phonological level a distinction is drawn between intonation forms and functions and only the former are described. The motivation behind this distinction is that in different languages the same intonation functions are encoded by different intonation forms. Unlike physical-acoustic and phonetic representations, the surface phonological ones use "distinctive discrete categories with which we can describe surface phenomena cross-linguistically" (Hirst, Di Cristo & Espesser 2000:3).

The underlying *phonological level* is the level of *functional, linguistic and cognitive descriptions*. They use distinctive discrete categories and encode information necessary for the syntactic and semantic interpretation of utterance's intonation (Hirst, Di Cristo & Espesser 2000:3). An example of underlying phonological representation is presented in (Pierrehumbert 1980).

In view of the discussion on possible interpretations of intonation and the levels of representation and analysis of intonation presented above it is proposed that in the current thesis the following terms are used in the senses explained below.

1. *Prosodic* or *suprasegmental parameters* refer to fundamental frequency, intensity, duration and spectral tilt/emphasis. Alternatively, the term *acoustic parameters* or *features* will be used.

2. Distinction is drawn between *lexical features* (stress, accent and tone determined in the lexicon) and *intonational features* (tunes, boundary phenomena). The latter are also referred to as *non-lexical, postlexical* or *sentence-level* features. The term *prosodic features* will be used to refer to both lexical and non-lexical features.

3. The term *prosody* is used in a general sense to refer to phonological (lexical and non-lexical) features, phonetic features and physical, acoustic parameters.

4. The term *suprasegmental structure* refers to segmental units which constitute the domain of suprasegmental analysis: syllables, accent groups/feet, intonational phrases.

5. For the needs of the research presented in this thesis and following the phonetic interpretation of intonation given in (Hirst & Di Cristo 1998) *intonation* is defined as an *interface* between *lexical and intonational features* on the one hand and *suprasegmental parameters* on the other. When used this sense, intonation is analyzed and represented on the phonetic level in terms of continuous parameters which can be mapped onto physical and phonological representations. At this level of description and analysis no distinction is drawn between linguistic and paralinguistic intonational meaning.

6. On the phonological, cognitive level intonation conveys *sentence-level meanings* which are *linguistic* and excludes *paralinguistic* messages. The properties of *paralanguage* can be described as follows (Ladd 1996:33ff):

a) "Paralinguistic messages deal primarily with basic aspects of interpersonal interaction - such as aggression, appeasement, solidarity, and condescension - and with the speaker's current emotional state - such as fear, surprise, anger, joy, boredom."
b) "Paralinguistic cues are global properties of the speech signal such as loudness, voice quality, and pitch range". This means that linguistic and paralinguistic meanings can be difficult to distinguish, because they are realized by the same acoustic parameters.
c) "Paralinguistic cues can be consistently interpreted even in the absence of the linguistic message".
d) "Paralinguistic cues should be regarded as modifications of the way in which phonological categories are realized". As a result of paralinguistic modification "instances of a given phonological category (may) sound as instances of some other".
e) "The central difference between paralinguistic and linguistic messages resides in the quantal or categorical structure of linguistic signaling and the scalar or gradient nature of paralanguage".

The consequences of the interpretation of the term intonation proposed here on assumptions underlying a comprehensive intonation model are discussed in sec. 1.1.4.

1.1.3. Intonation model

In general, a complete intonation model consists of three components (Hirst, Di Cristo & Espesser 2000, see Figure 2):
a) description/representation of intonation in terms of some theory
b) a method of mapping this description onto pitch targets
c) a method of deriving the description of intonation from a continuous f0 curve

As regards the first issue descriptions of intonation at different levels of analysis and representation are proposed: phonetic, surface phonological and phonological. Theoretical assumptions underlying these descriptions and their features are discussed in Chapter 3.

As regards the task of mapping the intonation description onto pitch targets, various machine learning methods can be applied such as neural networks (e.g. Mixdorff 2002a) or decision trees (e.g. Black & Hunt 1996, Dusterhoff & Black 1997). An alternative solution is to define so called mapping rules (e.g. Anderson, Pierrehumbert & Lieberman 1984, Jilka 1996, Jilka, Möhler & Dogil 1999).

Pitch targets can be defined in a number of ways depending on the level of intonation description. In phonological representations they are regarded as turning points in the contour which correspond to phonologically specified tones characterized by invariant scaling and temporal alignment (Ladd 1996:63ff). In surface phonological representations such as INTSINT (Hirst 1999,2000) target points are also regarded as turning points in f0 contours with the difference that no claim is made concerning their phonological status i.e., some targets correspond to tones, while others not. In the current thesis pitch targets are considered in the same way as proposed in (Black & Hunt 1996) as f0 values at specific position in the syllable structure which does not necessarily have to correspond to turning points and/or phonological

tones. This approach has been adopted in a number of other studies as well (e.g. Oliver & Clark 2005, Brinckmann 2006): its advantages are discussed in the sec. 7.1.

As regards the third component of an intonation model statistical modeling techniques such as decision trees or neural networks can be used to derive the description of intonation from a continuous f0 curve (e.g. Wightman & Ostendorf 1992, 1994, Kießling et al. 1996, Sridhar, Bangalore & Narayanan 2007, see sec. 6.1.2). This task involves detection and classification of intonational events - pitch accents and boundary tones on the basis of prosodic parameters derived from the speech signal and additionally (but not necessarily) information provided in utterance's segmentation/annotation.

An intonation model defined in this way will involve both coding of an f0 contour into some representation of intonation and predicting pitch targets from this representation. In contour generation the targets are smoothed and interpolated to give a continuous f0 curve. The diagram below illustrates these tasks of an intonation model.

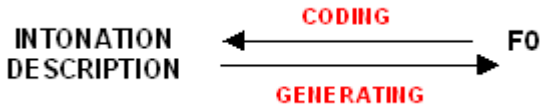


Figure 2 (after Hirst, Di Cristo & Espesser 2000:5): Intonation modeling as a two-way process.

It should be noted that apart from the two tasks of an intonation model i.e., coding and generation of f0 contours there is one more task which is related to the problem of intonation modeling, namely prosody prediction from text. Among other things, the task of such a module is to predict intonation description for a given input text from a number of factors including for example (Brinckmann 2006) word-level features (POS, word frequency), sentence-level features (sentence length and type), syntactic features, punctuation, syllable-level features (lexical stress, syllable length and structure), positional features (absolute and relative position of a syllable in the word/sentence).

Since the research presented in the current thesis is dedicated to intonation modeling as defined in (Hirst, Di Cristo & Espesser 2000:5) these issues are out of the scope of the current thesis.

As regards the tasks of an intonation model, depending on speech synthesis system type the output of an intonation model may be used in a slightly different way (cf. sec.1.1.5).

Some concatenative TTS systems generate speech from uniform units, usually diphones (e.g. Festival). These units have to be free from the influence of suprasegmental features, which means that among other things they should be produced with a neutral "flat" intonation.

In uniform TTS synthesis, once speech units that match the input text are found, speech can be generated. But the resulting speech will have no specific intonation pattern and consequently will be perceived as unnatural. Therefore it is necessary to take the following steps: a) predict the desired intonation for a given text, b) map the intonation description onto f0 targets from which an f0 contour can be generated, c) adjust the f0 of the speech units to the predicted f0 contour by means of signal processing techniques.

Unlike uniform speech synthesis, unit selection speech synthesis is based on a corpus consisting of speech units of various sizes (they may vary from half-phonemes to sentences or even paragraphs). The most important task in developing a unit selection corpus is to ensure that the speech material is representative. For this purpose text material is designed in such a way as to ensure good phonetic coverage and provide examples of the most important (from the point of view of synthesis) speech units (e.g. triphones, words) in all contexts which significantly affect their suprasegmental features. This means that e.g., the most frequent syllable structures are given in different position in a phrase, in different phrase types (interrogative, declarative, continuation etc.), with different pitch accent types associated with them. Apart from the requirement of the representativeness of the speech material (not to mention requirements concerning the recordings) another key issue is the consistent annotation of suprasegmental features. This task is known to be time consuming and laborious, which motivates the design of methods of automatic intonation labeling (examples are given in sec. 6.1). They achieve accuracy similar to consistency among human-labelers when transcribing intonation. These issues are addressed in Chapter 6 where methods of automatic detection and classification of pitch accents and boundary tones are presented.

Like in uniform speech synthesis, in unit selection TTS systems the first step towards intonation modeling is prediction of the desired intonation for the input text. The second step is also the same as in uniform systems and involves prediction of f0 targets from which contour can be generated for a given input utterance. This prediction is based on a number of syllable-, word-, phrase/sentence-level features including various prosodic features (stress, pitch accent type, phrase type: minor vs. major, type of the edge tone or intonational tune). In the next step, the database is searched for segmental units that match the best the predicted segmental, prosodic and f0 specification, and they are selected for synthesis. In unit selection TTS synthesis there is generally no need for speech signal processing and fine-tuning of the f0 of the target segmental string.

Recently, a method referred to as *prosody transplantation* has been proposed (van Santen et al. 2005) which consists in imposing of a pitch contour that matches the closest the target prosody onto unit sequences that match the target phone string by means of standard speech processing methods.

1.1.4. What does 'comprehensive' mean?

The term *comprehensive* as it will be used in this thesis refers to the structure and tasks of the intonation model, as well as to the levels of representation and analysis for description of intonation.

As regards the first issue it was already stated in the previous section that intonation modeling involves two tasks: coding and generation of f0 contours. The first task consists in detection and classification of intonational events such as pitch accents and boundary tones, thus it can be regarded as a recognition problem. The second task involves estimation of pitch targets from a number of features (e.g. prosodic, positional, sentence/phrase-level features). In contour generation these targets are smoothed and interpolated to give a continuous f0 curve (Black & Hunt 1996) which is applied to the waveform using e.g. PSOLA (Moulines &

Charpentier 1990). In the comprehensive intonation modeling methods are designed which are capable of performing both detection/classification of intonational events and prediction/generation of f0 contours.

As regards the issue of the levels of representation and analysis for description of intonation, a variety of approaches is presented in the literature (an overview is given in Chapter 3) and within the framework of different models descriptions of intonation at various levels are proposed. Here, only some of them are mentioned and their most important features are summarized in terms of the typology introduced by Hirst & Di Cristo (Hirst & Di Cristo 1998:7, see Figure 1).

At one end of the typology of intonation models there are *phonetic models*. On the phonetic level intonation is described in terms of continuous parameters such as amplitudes, durations, slopes of rising and falling pitch movements, level and position of f0 peaks and minima (e.g. Tilt, PaIntE, Fujisaki model) or scaling and temporal alignment of pitch targets (e.g. Momel). This description encodes the macroprosodic component of an f0 curve reflecting the choice of intonation pattern for the utterance (Hirst & Espesser 1993). The choice of specific parameters depends on what is considered to be the invariant and primary feature of the elements of intonational tunes (pitch accents, boundary tones) and what is regarded as their best acoustic correlate.

In *sequential phonetic* representations (Taylor 2000, Möhler 2001) only the "meaningful" portions of an f0 contour (i.e., those which correspond to pitch accents and boundary tones) are described (see sec.3.2). In *superpositional models* (Möbius 1993, Mixdorff 2002, van Santen et al. 2005) the parameterization is carried out separately for phrase and accent components of an f0 contour (see sec. 3.3). Phonetic representations can be mapped onto physical parameters on the one hand and phonological representations on the other. On the phonetic level of intonation description no distinction is made between linguistic and paralinguistic intonational meaning - they are distinguished on the phonological level.

At the other end of the typology there are *phonological models* (e.g. Pierrehumbert's model, original ToBI system and its adaptations to other languages, see sec. 3.6). They use distinctive discrete categories and "encode the information necessary for the syntactic and semantic interpretation of the prosody of an utterance" (Hirst, Di Cristo & Espesser 2000:3). Phonological descriptions account for both intonation form and function, and in the same way as the relation intonation form-function differs among languages, these descriptions are language-specific.

One of the key features of phonological models is that they tell linguistic meanings conveyed by intonation and paralanguage apart. Since both linguistic and paralinguistic features of utterances are realized by the same acoustic properties it is impossible to distinguish between them at lower levels of analysis (Ladd 1996:20ff). This feature of phonological models is considered to be their advantage, but on the other hand, they are often regarded as inadequate for application in speech technology systems. Among other things they are regarded as unable to represent the whole variety of naturally occurring intonational tunes (Taylor 2000) and introduce a quantization error into intonation modeling by reducing "the whole variety of f0 values available in the acoustics (...) to a mere binary opposition Low vs. High, and to some few additional, diacritic distinctions" (Möbius et al. 2000, Batliner et al. 2000). Another disadvantage of phonological representations mentioned in the literature is that they pretend to

be of categorical and discrete nature, but boundaries between categories are not strict and easily identifiable (Taylor 2000).

In between phonetic and phonological levels of representation and analysis of intonation there is the *surface phonological level* on which intonation is described in terms of discrete categories with which surface phenomena can be described cross-linguistically (Hirst, Di Cristo & Espesser 2000:3). On the surface phonological level a distinction is drawn between intonation functions and forms, and only the latter are described, because in different languages the same functions are encoded by different forms. Unlike phonological representations the representations on the surface phonological level make no attempt to identify the underlying intonational phonology and consequently, they describe linguistic as well as paralinguistic intonational meaning. Examples of surface phonological models are INTSINT (Hirst & Di Cristo 1998) and GToBI (Grice, Baumann & Benzmueller 2005).

INTSINT (International System of Intonation Transcription) provides a formal encoding of the macroprosodic curve in terms of tones: some of them correspond to phonological tones, whereas others do not. INTSINT description is restricted to representation of the form of intonation contours.

The major difference between the GToBI representation and INTSINT consists in that GToBI accounts for both intonation form and function. Like strictly phonological descriptions GToBI uses discrete categories to encode linguistic meaning conveyed by intonation. But still, it can be regarded as a surface phonological representation. As stated in (Grice, Baumann & Benzmueller 2005:10) GToBI is "not a strictly phonological description of German intonation" - no distinction is made between linguistic and paralinguistic functions of intonation, because "both types of function can be expressed by discrete means such as the choice of pitch accent and boundary tones" (op.cit:20). It is assumed that surface phonological representations like GToBI can be more useful for application to speech synthesis and recognition systems than strictly (underlying) phonological descriptions, because they are more perceptually-oriented. Besides, the example of the GToBI model shows that there is no need to account for intonation form and function at different levels of analysis as it is proposed in (Hirst & Di Cristo 1998) and consequently, to define more intermediate descriptions then it is really needed (cf. Batliner et al. 2000).

In view of the discussion on intonation representations at various levels presented here the following assumptions concerning the comprehensive approach to intonation description are defined.

1. The comprehensive approach provides a global description of intonation - it takes various levels of analysis and representation into account. In view of the drawbacks of strictly phonological models on the one hand and the advantages of surface phonological GToBI-like models on the other hand, it is proposed to exclude the underlying phonological level and to "restrict" analysis and representation of intonation to physical/acoustic, phonetic and surface phonological levels.

2. As already said in sec. 1.1.2 in this thesis the phonetic interpretation of intonation proposed in (Hirst & Di Cristo 1998:7) is adopted according to which intonation is considered as interface between lexical and intonational features on the one hand and suprasegmental parameters on the other. Intonation defined in this way is analyzed and represented on the phonetic level and

described in terms of continuous parameters which can be mapped onto lower-level (physical/acoustic) and higher-level (surface phonological) representations. Consequently, not only phonetic, but also more abstract, higher-level description of intonation has to be proposed, which fits well to the previous assumption that the comprehensive approach accounts for various levels of analysis and representation for intonation description.

3. The comprehensive approach to description of intonation takes both linguistic and paralinguistic intonational meaning into account and consequently, provides a framework within which both melodic and functional aspects of intonation are analyzed and represented. As already explained in sec. 1.1.2: "intonation conveys meanings that apply to phrases or utterances as a whole" (Ladd 1996:7). Linguistic functions of intonation include (among other things) signaling of different types of sentence modes or speech acts, focus, phrasing and information structure (cf. sec. 2.4). Unlike linguistic messages the paralinguistic ones "deal primarily with basic aspects of interpersonal interaction (...) and with speaker's current emotional state" (op.cit:33). On the one hand the linguistic and paralinguistic messages are realized by the same acoustic properties, which is why it is so difficult to tell them apart on the level of speech production. On the other hand "both types of function can be expressed by discrete means such as the choice of pitch accent and boundary tone" (Grice, Baumann & Benzmueller 2005:20). What's more, both linguistic and paralinguistic distinctions realized by suprasegmental features involve perceptually significant differences. Consequently, it seems reasonable to concentrate on melodic aspects of intonation in the first place and to identify the elements of the intonation system on the basis of the fact that they are perceptually distinct from other elements. This is one of the key theoretical assumptions of the IPO approach (t'Hart, Collier & Cohen 1990, see also discussion in Ladd 1996:18ff), where "meaning or function plays no role in the analysis" (Ladd 1996:19). Yet, intonation models for application to speech technology systems have to rely on a description which serves as a framework for representation and analysis of intonational meaning as well. Therefore, unlike in the IPO system, apart from a phonetic description of intonation (in terms of acoustic properties) which describes only melodic aspects of intonation, the comprehensive intonation model proposed in this thesis provides also a description on the surface phonological level where elements are distinguished with reference to both melodic and functional aspects of intonation.

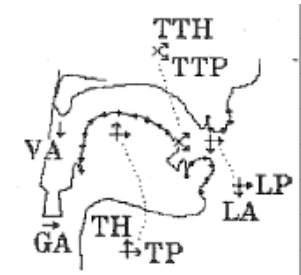
1.1.5. Speech synthesis

Speech synthesis concerns generation of speech from some kind of input, i.e. parametric representation of speech acoustics, or speech production or text. The primary division of speech synthesizers is into rule-based and data-driven. The former include articulatory and formant synthesizers, whereas the latter are based on concatenation of speech units of the same size (uniform units) or different size (non-uniform units). While the discussion on speech synthesis systems presented in this section gives a general overview of types and features of synthesizers, and problems that can be encountered extensive information on these topics can be found in (Sproat 1997, van Santen et al. 1997, Narayanan 2004).

1. Articulatory synthesis

Articulatory synthesis simulates speech production. This type of synthesis requires a dynamic model of vocal tract which makes it possible to simulate motions of the articulators during speech production. Another component is glottis model which generates the excitation signal (i.e. random noise or a quasi periodic sequence of pulses). The synthesis requires also a method of generating and monitoring air pressure and velocity.

In the source-filter model of speech production the excitation signal is considered as the source. The signal passes through the vocal tract (and nasal tract) which acts as a filter: depending on its shape there occur resonances or formants at certain frequencies. Their location determines the identity of a sound. The other filters are lips whose rounding also determines formant frequencies. The vocal tract is represented as a transfer function (because it transfers the excitation signal) and in terms of parameters as illustrated and described in Figure 3.



TTP:	TONGUE TIP POSITION
TTH:	TONGUE TIP HEIGHT
VA:	VELIC APERTURE
GA:	GLOTTAL APERTURE
LP:	LIP PROTRUSION
LA:	LIP APERTURE
TH:	TONGUE HEIGHT
TP:	TONGUE POSITION

Figure 3 (adopted from Krueger 1998): Vocal tract model.

During synthesis the parameters representing the vocal tract are assigned specific values in time to model motions of the articulators, the glottis is set in a proper position by defining its aperture and tension of the vocal cords, lung pressure and air velocity are also defined. It has to be determined whether the nasal tract functions as an additional resonance tube or not.

Articulatory synthesis is very useful tool for research on speech production and perception, but the quality of speech generated in this way is far from perfect. This is mainly due to computational and mathematical complexity of the underlying models and the insufficient knowledge concerning articulatory processes involved in the production of natural speech. Articulatory synthesis is available in *Praat* (Boersma & Weenink 2005) based on the models developed in (Boersma 1998). For more literature and examples of a dynamic articulatory synthesis see http://www.haskins.yale.edu/facilities/INFO/info.html

2. Formant synthesis

Formant synthesis is based on the source-filter-model and knowledge concerning speech acoustics. A formant synthesizer consists of two components - the generator of an excitation signal and formant filters that represent the resonances of the vocal tract. The former is considered as a source and the latter as a filter according to the source-filter model of speech production. Formant frequencies and bandwidths are modeled by means of a two-pole resonator. The number of formants varies from 3 to 6: with the latter number a high quality

speech can be obtained. Additional information concerning radiation characteristics of the mouth which influence formant frequencies is also necessary.

Usually from more than a dozen up to a couple of dozens of control parameters are used as an input to the waveform synthesizer. Some of the control parameters vary in time e.g. formant frequencies, amplitudes and bandwidths, fundamental frequency, amplitude of voicing, aspiration, friction and turbulence, spectral tilt of voicing etc., whereas other parameters remain constant e.g. sampling rate, number of formants in cascade vocal tract (Klatt 1980).

Formant synthesizers can have parallel cascade or combined parallel-cascade structure. The best results can be obtained with parallel-cascade synthesizers whose structure is illustrated in Figure 4.

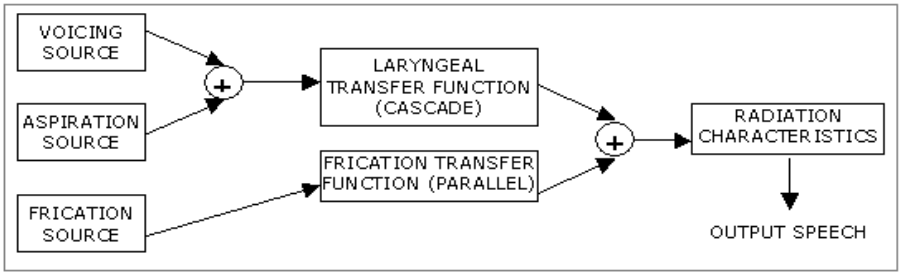


Figure 4 (adopted from (Allen et al. 1987): Structure of the MITTalk systems

Formant synthesis can be a very useful tool for research in the field of speech acoustics, phonetics and speech perception.

Speech generated by formant synthesizers is characterized by metallic sounding which can be eliminated only after hand-tuning of the control parameters, which makes format synthesis impractical if not impossible for application in a fully automatic speech synthesis system.

Examples of formant synthesizers are: KlattTalk, MITalk, DECtalk, Prose-2000, Delta and Infovox.

3. Concatenative synthesis

Concatenative synthesis refers to generation of speech from an input text. The speech results from concatenation of acoustic units stored in a speech corpus. The units are annotated at segmental and suprasegmental level. During synthesis the corpus is searched for units which match the closest the specification of the target utterance. This specification is determined in the analyses carried out in the NLP (Natural Language Processing) component of a TTS system. In general, NLP includes: text preprocessing module, morphosyntactic analyzer, phonetizer and prosody generator (Dutoit & Stylianou 2003).

The second component of a TTS system is called Digital Signal Processing module. Acoustic units are stored in a parametric and compressed form – techniques such as LPC (Linear Predictive Coding), TD-PSOLA (Time-Domain Pitch-Synchronous-Overlap-Add) or MBROLA (Multi-Band Resynthesis Overlap Add) are in charge of this task. They also perform concatenation of the units, smoothing out of the concatenation points and prosody manipulation.

Unlike in the rule-based speech synthesis which deals with speech production problems, these problems do not have to be addressed in concatenative synthesis. However, other problems arise that are related to the choice of the units (e.g. diphones, triphones, uniform units - units of the same length or non-uniform units - units of a varying length), the concatenation itself (how to find units which match the closest the target utterance specification), prosody modification (manipulation of intonation and duration in order to adapt the units to the target prosody) and compression of the acoustic units.

The basic idea behind concatenative speech synthesis is that speech can be generated from a limited inventory of acoustic units and that their concatenation should take coarticulation effects into account. Therefore systems based on phonemes and words are impractical – they do deal with coarticulation phenomena properly and as a result unnatural synthetic speech is obtained. Moreover, an inventory consisting of all words (or syllables) occurring in a given language is too numerous, requires too much memory and is computationally too costly. These problems do not arise if diphones or demisyllables are used as acoustic units. On the one hand they account for coarticulation, and on the other their inventories are not too numerous. However, in order to generate natural sounding synthetic speech, the modification of f0 and duration of the units is required.

Another way to achieve "natural" synthetic speech is to use an inventory of non-uniform units and to select the longest units that match the target utterance specifications the best. This method is known as unit selection speech synthesis (Hunt & Black 1996).

The selection of units is based on two costs – concatenation and target costs. The former specifies the amount of discontinuity at the concatenation point between two acoustic units, whereas the latter determines to what extent the unit matches the target utterance specification. This approach makes it possible to decrease the number of concatenation points. In general, it is assumed that the target utterance can be synthesized without any post-processing such as fine-tuning of f0 and duration.

One of the latest approaches to text-to-speech synthesis is Hidden Markov Model-based speech synthesis (e.g. Tokuda, Zen & Black 2002, Yamagashi et al.2007, Liang Qian & Soong 2007). This approach is much more flexible in adding new voices or speaking styles in comparison to classical unit selection synthesis which requires large amount of speech data. For example in (Yamagashi et al.2007) within the framework of a HMM-based synthesis a speaker-independent approach is proposed which makes use of average voice models. Firstly, an average voice is trained using speech from a number of different speakers and then, it is adapted to a target speaker. The authors report that a database of 100 utterances (about 6 min.) is sufficient to obtain a synthetic speech for a new target speaker, which shows robustness of this approach.

1.2. The goals of the thesis

In the previous sections the terms *intonation* and *comprehensive* were explained and the goal of intonation modeling and tasks of intonation models were specified. Apart from that an overview of speech synthesis systems was given. This descriptive and explanatory part aimed at presentation of the research area of the current thesis. In this section the hypotheses which will be tested in this dissertation are formulated.

The research presented in this work is focused on the *development of a comprehensive intonation model* for speech synthesis. The term *comprehensive* as it is used in this thesis refers to three aspects:

a) the levels of representation and analysis for description of intonation taken into account in the model
b) the structure of the model and the tasks it performs
c) speaking styles taken into account in the model

It was claimed that as regards the levels for description of intonation in the comprehensive intonation modeling the *phonetic* and *surface phonological levels* should be taken into account in order to make analysis of various aspects of intonation possible. And thus, melodic aspects are analyzed on the phonetic level in terms of continuous parameters. On the surface phonological level not only melodic, but also functional aspects of intonation are analyzed in terms of discrete distinctive categories.

As regards the structure and tasks performed by the model it was claimed that it should be *bi-directional* - on the one hand it should provide a framework for *detection and classification of intonational events* (i.e. coding of f0 contours, see Figure 2) and on the other hand it should offer solutions to the problem of *prediction and generation of f0 contours*.

Finally, the comprehensive intonation model should not be confined to a single speaking style - apart from news-reading style speech provided by most of the existing TTS systems, it should also make *generation of a naturally sounding emphatic speech* possible.

Bearing these considerations in mind the hypotheses which will be tested in this thesis are formulated: Some of them are related to the issue of the description of intonation or components and tasks of the model, while others refer to the speaking styles that are taken into account in intonation modeling. The first hypothesis is the following:

Hypothesis 1. *The intonation model developed in this thesis can be regarded as comprehensive if it provides a framework for description of intonation at various levels of analysis which is useful for coding and generation of f0 contours.*

The specification of the description of intonation can be regarded as the first task when developing an intonation model, because it is used by the other two components of the intonation model which deal with coding and generation of f0 contours. As mentioned above, the model should take into account two levels of analysis - phonetic and surface phonological. On the phonetic level intonational features are analyzed in terms of vectors of acoustic features

which describe the macroprosodic component of an f0 contour reflecting the choice of intonation pattern for the utterance (see sec. 1.1.2). The phonetic description is used in detection and classification of intonational events (see Chapter 6). Intonational events - pitch accents and boundary tones constitute elements of intonational tunes and are described on the surface phonological level in terms of distinctive categories. This higher-level and abstract description encodes not only melodic, but also functional aspects of intonation. It is not a strictly phonological description, because as explained in the discussion in sec. 1.1.4, it is assumed that phonological systems are less useful for speech applications than surface phonological descriptions such as for example GToBI (see sec. 3.6.5). The latter encode perceptually significant differences between tunes irrespective of whether they reflect distinct linguistic or paralinguistic intonational meaning. The surface phonological description proposed in this thesis (see Chapter 5) is based on the system used for prosody labeling in the Polish unit selection corpus (see Chapter 4). In order to prove Hypothesis 1 which says that the intonation model can be regarded as comprehensive it provides a framework for description of intonation at various levels of analysis which is useful for both coding and generation of f0 contours a number of fine-grained hypotheses will have to be confirmed first. They are formulated as follows:

Hypothesis 1a. *The phonetic description of intonation proposed in this thesis provides information which is significant to the detection and classification of the elements of intonational tunes - pitch accents and boundary tones.*

Hypothesis 1b. *The surface phonological description proposed in this thesis which reflects melodic and functional aspects of intonation provides information of a high significance to the estimation of pitch targets and thus, to the results of contour generation in speech synthesis.*

Apart from the description of pitch accents and boundary tones on the phonetic and surface phonological levels a finer description of prosodic structure will be proposed. It will be shown that pitch variation can be controlled by referring to a two level phrasing system in which phrases of a different structure and position in utterance are grouped together. The study presented in sec. 5.1 which aims at definition of such a description is motivated by the results described in (Clark 2003): they proved that incorporation of a more detailed information concerning prosodic structure affects positively the quality of f0 generation. For this reason the following hypothesis is defined:

Hypothesis 1c. *The description of prosodic structure proposed in this thesis provides an important information for the estimation of pitch targets and thus, affects the performance of the regression model and the overall quality of intonation modeling.*

The hypotheses presented so far are related to the first aspect of intonation modeling, namely the levels of representation and analysis for description of intonation.

The second aspect refers to the components and tasks of the model. In order to prove the comprehensive character of the intonation model developed in this thesis the following hypothesis has to be confirmed.

Hypothesis 2. *The intonation model proposed in this thesis can be regarded as comprehensive if it is bi-directional i.e., makes coding and generation of intonation contours possible and performs these tasks with a high accuracy.*

Coding of f0 contours deals with the design of methods capable of automatic detection and classification of intonational events. The former involve prediction of the location of accented syllables and phrase boundaries in the utterance and the latter deal with recognition of the type of pitch accents and boundary tones distinguished in the surface phonological description.

It is known that manual transcription of prosody is laborious and sometimes discrepancies between labelers occur (this topic is discussed in sec. 6.1.3). On the contrary, automatic methods can provide a fast and more consistent prosody labeling, which is especially useful if large speech corpora for various applications such as speech synthesis or recognition are to be annotated (see sec. 6.1.2).

In speech synthesis the similarity between the natural and generated intonation contours is greater and the perceived naturalness of synthetic speech increases if a more detailed information concerning prosody is taken into account in contour prediction (Möhler 2001, Syrdal et al. 1998). The information provided by automatic prosody labeling methods can be useful for speech recognition systems as well. The information concerning prosodic structure (location of boundaries and strength of phrase break) may help to resolve ambiguous parses. The type of nuclear contour (nuclear pitch accent and subsequent boundary tone) indicates sentence modality – this information can be used directly to indicate what kind of response is expected from a dialogue system. The knowledge that an accent is associated with a particular syllable indicates that the syllable is stressed. In fixed stress languages such as Polish stress has demarcative property: it signals the boundary of a bigger prosodic constituent, namely prosodic word. Thus, the location of a word boundary can be deduced from the information on stress. In languages which make lexical distinctions based on stress position this information may be helpful to resolve word ambiguity e.g. /'dizain/ vs. /di'zain/. The examples given here show that the information concerning intonational features and prosodic structure is very important and methods which are capable of deriving it automatically can be very useful, but only if they perform with a high accuracy. Otherwise, the error introduced by automatic prosody labeling may have no particular effect on the quality of f0 generation as there may be no perceptual difference between the output of systems which exploit prosodic features and those that do not (e.g. Brinckmann 2006).

In order to prove the hypothesis on the comprehensiveness of bi-directional intonation models methods of automatic detection and classification of prosodic constituents will be designed and solutions to the problem of pitch target estimation for the purpose of f0 contour generation will be proposed. The performance of these methods will be evaluated in terms of their accuracy. As regards the problem of coding of f0 contours two further hypotheses are formulated, namely:

Hypothesis 2a. *Automatic detection and classification of intonational events can yield results comparable the inter-labeler consistency in manual transcription of prosody.*

Hypothesis 2b. *A high accuracy in the automatic detection and classification of intonational events can be achieved even if only a small vector of acoustic parameters and information extracted from utterance's transcription/segmentation are used as input features to the model.*

In order to prove these two hypotheses the performance of the classification models designed in the current study will be compared with the results reported in the literature (an overview is given in sec. 6.1).

As regards the problem of f0 contour generation the following hypothesis is defined:

Hypothesis 2c. *An approach to f0 generation in which f0 contours result from interpolation between pitch targets of a predefined position in the syllable structure (at the onset start, in the middle of the nucleus and at the end of the coda) whose values are estimated by means of a regression model provides a high quality speech characterized by natural intonation.*

In order to test this hypothesis, the performance of the regression models designed in the current study will be evaluated in the objective and subjective manner. Objective evaluation will show to which extent the generated f0 contours are similar to the original contours by calculating the correlation between them – a high correlation coefficient reflects high similarity. Moreover, the performance of the designed models should be at least comparable to the performance of models described in other studies (an overview is given in sec. 7.1).

Contours which are similar should also be perceptually close to each other, but this can be confirmed only in a perception test. The results of the test indicating naturalness of the generated intonation will prove the usefulness of the approach proposed in this thesis. It is assumed that if the three requirements are met, namely 1) the correlation between original and generated f0 contours is high; 2) the correlation is comparable to that reported in other studies and 3) the generated intonation "sounds" natural the Hypothesis 2c will be confirmed.

The last hypothesis refers to speaking styles that are taken into account in the modeling. As mentioned before, the comprehensive approach to intonation modeling should not be confined to a single speaking style, but on the contrary it should provide methods and solutions which can successfully be applied to intonation generation of unemphatic as well as expressive speech. On the basis of this assumption the following hypothesis is formulated:

Hypothesis 3. *The intonation model can be regarded as comprehensive if it provides a framework for generation of a high-quality, naturally sounding intonation in expressive speech.*

In order to test this hypothesis a model for pitch contour generation in expressive speech will be proposed. The hypothesis will be proven if the model generates intonation contours which are similar to the original contours and if the performance of this model is comparable to the results reported in the literature. Apart from that, the generated intonation should be characterized by naturalness, which will be investigated in a perception study.

As the speech corpus used to design the model for expressive speech contains material from three different speakers f0 normalization similar to that proposed in (Clark 2003) and adopted in (Oliver & Clark 2005) will be used. The targets predicted by the regression model will be z-score normalized f0 values with respect to mean and standard deviation determined for a given speaker and for a given phrase type (distinguished in the analyses presented in sec. 5.1). In contour generation the values will be re-scaled, smoothed and interpolated through to produce a continuous f0 contour.

The hypotheses presented above will be tested in Chapter 6 and Chapter 7 and confirmation of the hypotheses can be regarded as a proof of the *comprehensive* character of the intonation model developed in this thesis.

Figure 5 illustrates the structure and tasks of the comprehensive intonation model developed in the scope of this thesis.

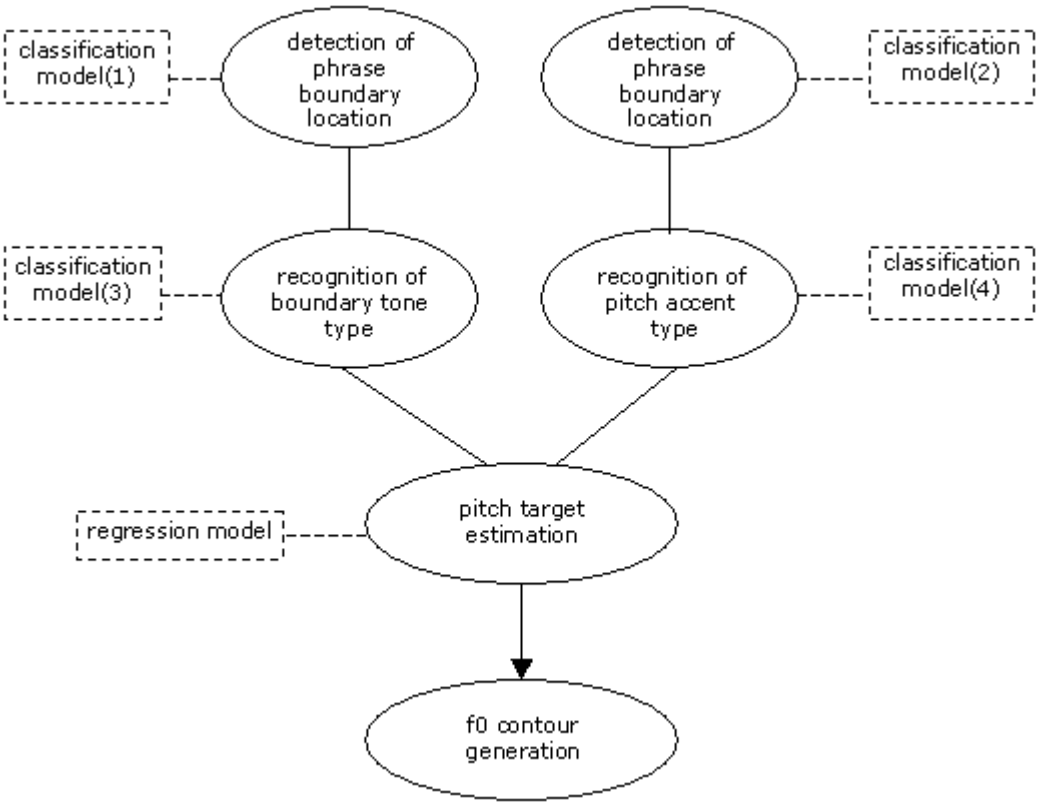


Figure 5: Outline of the comprehensive intonation model proposed in the current thesis. Models developed in the thesis are given in boxes, whereas specific tasks performed by the models are given in circles.

In the first place, the location of pitch accents and phrase boundaries is determined at the word-level from f0 and duration features which constitute part of the description of intonation on the phonetic level. Two classification models are designed to perform this task and they rely also on the information on word/syllable/phoneme boundary and lexical stress, but the latter information is used solely in detection of accented syllables.

The second task consists in the recognition of pitch accent and boundary tone types. The classification is performed at the word level - only the syllables recognized as accented or pre-boundary are taken into account. Like the models designed for the detection of the location of accents and phrase boundaries they use acoustic features which constitute phonetic description of intonation. As a result, the surface phonological description of utterance's intonation is obtained.

In the next step, on the basis of the information provided in the surface-phonological description of intonation, description of prosodic structure and a number of features derived from utterance's annotation/segmentation the symbolic representation of the structure of the utterance is obtained. It is used as the input to the regression model which estimates pitch targets. In contour generation which is the last step in intonation modeling interpolation is carried out between smoothed f0 targets to produce a continuous f0 curve.

1.3. Outline

In this chapter an introduction to the research area of this thesis was given. In the first place, some of the basic functions of intonation were mentioned and the role of intonation in speech was sketched. Secondly, the senses in which the term intonation appears in the literature and in which it is used in this thesis were explained. Next, the reasons behind the development of intonation models and their tasks were discussed, which was followed by an overview of the types of speech synthesis systems. In the end, the hypotheses which will be tested in the analyses presented in this thesis were formulated and research goals were defined.

In Chapter 2 melodic and functional aspects of intonation are discussed. In the first place, lexical level (also referred to as syllable level) is taken into account and consequently, terms such as tone, tonal accent, stress, pitch accent and nuclear accent are explained. Next intonation forms and functions on the sentence/phrase level are presented including focus, phrasing, prosodic structure, etc. In the end the discourse/dialogue level is taken into account. The chapter ends with a short discussion on intonational meaning.

In Chapter 3 overview of the literature on intonation modeling is given with special emphasis on description of intonation at various levels of analysis and representation. The chapter starts with discussion on the typologies according to which intonation models are classified. In the next sections features of phonetic sequential and superpositional models, perceptual models and phonological models are discussed in detail. This is followed by presentation of the state-of-the-art knowledge on Polish intonation and studies whose results can be used in various speech applications. The chapter ends with a discussion on the usefulness of different approaches to intonation modeling and concluding remarks are presented.

Chapter 4 describes the speech material and methods used in the analyses presented in this thesis. In the first place, the structure and features of two speech corpora i.e., subset of the Polish unit selection corpus and especially designed and recorded expressive speech corpus are discussed. Then the segmental and prosodic annotation of the corpora are presented, which is followed by description of the collection of data for analyses and model building. In the last section of the chapter an overview of the basic assumptions underlying statistical analyses and modeling methods used in the thesis and their application is given.

In Chapter 5 the comprehensive description of intonation is proposed. The chapter is divided into three parts. In the first one the analyses towards definition of the model of prosodic structure are presented. It is expected that the proposed description will provide an important information for the estimation of pitch targets and thus, will have an impact on the performance of the regression model and the overall quality of intonation modeling. (Hypothesis 1c).

In the second part of Chapter 5, the surface phonological description of intonation is proposed. On this level intonational tunes are considered as sequences of discrete, distinctive categories of pitch accents and boundary tones. An inventory of pitch accent types is defined and their distribution and structural roles in tunes are discussed. This is followed by definition of between various boundary tone types. The resulting description encodes not only melodic, but also functional aspects of intonation. Its usefulness for the estimation of pitch targets used for contour generation in speech synthesis (Hypothesis 1b) will be tested in chapter 6. The chapter ends with definition of the description of pitch accents and boundary tones in on the phonetic level. The usefulness of this description (Hypothesis 1a) is tested in the next chapter.

Chapter 6 is dedicated to the problem of coding of f0 contours. On the basis of the description of intonation on the phonetic and surface phonological levels models performing automatic labeling of intonational events and detection of accentual prominence and phrase breaks are designed. The chapter starts with discussion on the acoustic cues to accentual prominence and phrase boundary presented in the literature. Next, an overview of the existing solutions to the problems of detection and classification of intonational events is given and the results of studies on inter-transcriber consistency in manual labeling intonational events are reported. The information presented in this theoretical part of the chapter is necessary for the evaluation of the performance of the designed models and confirmation of Hypothesis 2a and Hypothesis 2b.

In Chapter 7 the third component of the intonation model is presented which deals with f0 estimation and generation in speech synthesis. Two hypotheses are tested in this chapter: Hypothesis 2c according to which an approach to f0 generation in which f0 contours result from interpolation between pitch targets of a predefined position in the syllable structure (at the onset start, in the middle of the nucleus and at the end of the coda) whose values are estimated by means of a regression model provides a high quality speech characterized by natural intonation.

The other hypothesis (Hypothesis 3) refers to the speech styles taken into account in the intonation modeling and says that the comprehensive model can be regarded as universal if it provides a framework for generation of a high-quality, naturally sounding intonation of expressive speech.

The chapter starts with preliminary remarks and reasons for adopting the specific sequence-based approach to contour generation are explained. Next, the selection of features for estimation of pitch targets is described and regression model for unemphatic speech is designed: the model is trained and tested on the speech material from the unit selection corpus. The training, testing and performance of the other regression model designed for expressive speech is reported in the next section. The chapter ends with presentation of the results of subjective evaluation (perception study) of the quality of the intonation generated with the proposed model.

The last chapter gives a summary of the main findings of the research carried out in the scope of this thesis. Most importantly, it presents the results of testing the hypotheses formulated in the beginning.

The thesis ends with an outlook on future work.

Chapter 2. Intonation functions and forms

In this chapter linguistic functions of intonation are discussed as well as forms in which they are encoded. This overview has two goals. First of all, it shows the variety of ways in which intonation contributes to communication between speakers. Secondly, it signals some essential issues related to intonation and investigated in this thesis.

For the purpose of the overview given in this chapter a broad definition of intonation is adopted (see sec. 1.1.2) and consequently both lexical (syllable-level) and non-lexical features (sentence/phrase-level) are discussed.

Following (Botinis, Granstroem & Möbius 2001) functions and forms of intonation are presented with respect to three levels including lexical, phrase/sentence and discourse/dialogue level. At the end of the chapter intonational meaning is briefly discussed.

2.1. Lexical level

At the lexical or syllable level[1], the function of intonation is to make stress, tone and accent distinctions.

2.1.1. Tone and tonal accent

Tonal languages e.g. Chinese, Thai, Vietnamese make use of contrastive tones which have distinctive functions in the lexicon and morphology. Tones are realized by pitch and intensity variation and can be static (pitch height remains constant over the length of a syllable), or dynamic (there is a pitch movement on the syllable). For example, Chinese has four distinctive tones: one static, H (high-level) and three dynamic: R (mid-rising), L (low-dipping) and F (high-falling). Apart from tone distinctions Chinese, Vietnamese and Thai "possess also a distinction between stressed and unstressed syllables which is lexically distinctive in Chinese but not in Vietnamese or Thai" (Hirst & Di Cristo 1998). Tonal patterns of polysyllabic words are determined in the lexicon.

Tonal accent languages e.g. Japanese and Swedish make use of a lexically specified accent realized by tonal prominence. Like in tonal languages intonation does not affect relative pitch height/shape of an accent/tone which is specified in the lexicon.

In Japanese both accent position and accent presence/absence cause lexical contrasts e.g. `kaki - oyster, ka`ki - fence, kaki – persimmon, where ` indicates accent on the following syllable (Hirst & Di Cristo 1998:10).

[1] It is assummed that the term lexical level is equivalent to syllable level, because syllable constitutes the domain of lexical distinctions.

The other tonal accent language, Swedish, makes use of two distinctive word accents: acute and grave. They both are associated with stressed syllables but only the acute accent can occur on a word-final syllable.

Figure 6 illustrates possible patterns for disyllabic words in Japanese and Swedish.

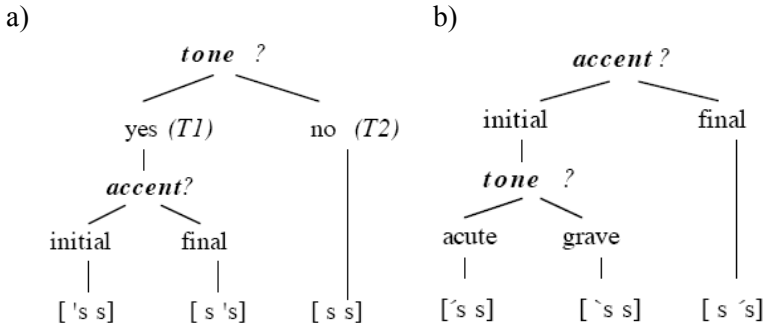


Figure 6 (adopted from Hirst & Di Cristo 1998:11): Distinctions in tonal accent/pitch accent languages: a) Japanese, b) Swedish

As it can be seen in
Figure 6 the difference between Japanese (a) and Swedish (b) is that while the distinction between accent presence/absence is primary in Japanese it is secondary in Swedish; in Swedish accent position is primary - in Japanese it is the opposite.

2.1.2. Stress and pitch accent

In some works dealing with intonation the terms accent or pitch accent and stress are used interchangeably. Yet, in this thesis an assumption is made that pitch accent and stress are not the same thing and consequently a distinction is drawn between them.

Stress is defined as a lexical property of individual syllables and is associated with acoustic salience or prominence. From the production point of view stress can be realized by properties such as intensity, duration, spectral tilt and force of articulation. Stress can have the functions described below.

a) it may distinguish lexical meaning e.g. **`defer** vs. **di`ffer**, or Greek **`nomos** (‘law’) vs. **no`mos** (‘county’) (Botinis, Granstroem & Möbius 2001)
b) it may distinguish grammatical classes e.g. **`content** (noun) vs. **con`tent** (verb)
c) grouping - each stressed syllable forms a stress group or foot with adjacent/preceding unstressed syllables. Stressed syllable is regarded as the head of the foot. Depending on whether the foot begins or ends with the stressed syllable the foot is left-headed (e.g. in Dutch, English, German) or right-headed (e.g. in French, Italian, Romanian). Stress patterns reflect prominence relations between syllables and rhythmic structures of utterances (van der Hulst 1999) e.g. binary rhythm (iambic/trochaic feet), ternary rhythm (dactyl – left headed, anapest – right headed).

d) demarcative - position of the stressed syllable serves as a cue of the presence of the word boundary (Botinis, Granstroem & Möbius 2001). This function is observed in fixed-stress languages e.g. Polish (penultimate syllable), Czech and Finnish (left initial syllable), Turkish and French (right final syllable).
e) culminative property of stress consists in signaling the presence of a prosodic domain e.g. word-level stress is an indicator of the presence of a prosodic word, and sentence-level stress signals presence of a phrase (van der Hulst 1999)

A pitch accent is considered as „a local feature of a pitch contour – usually but not invariably a pitch change and often involving a local maximum or minimum" (Ladd 1996:46). Pitch accents are associated with metrically strong, stressed syllables and serve as cues to syllable's prominence. However, in some languages pitch accents may be associated with unstressed syllables (e.g. Italian, Ladd 1996:129) and additionally they may have no prominence-cueing function (e.g. French, Indonesian, Ladd 1996:59).

The definitions of stress and pitch accent provide us with two reasons for drawing a distinction between them. The first one is production: while duration and spectral tilt are the main acoustic correlates of stress, pitch movements and overall intensity are the main correlates of pitch accents (Tamburini 2005). The second reason is the level of analysis and description of stress and pitch accents: stress belongs to lexical prosodic system, whereas pitch accents – to non-lexical prosodic system (Hirst, Di Cristo & Espesser 2000).

2.1.3. Nuclear pitch accent

Nucleus or *nuclear pitch accent* (also: nuclear tone, primary accent, primary stress) is regarded as the most prominent and salient accent in the intonation group (Cruttenden 1997). In phonological systems (e.g. Pierrehumbert 1980) nucleus is considered as a combination of a phrase-final pitch accent (i.e., accent associated with the last stressed syllable in the phrase) and the subsequent boundary tone. No structural distinction is made between nuclear and prenuclear pitch accents apart from their position in a phrase i.e., nucleus is simply the final pitch accent and prenuclear accents are associated with non-phrase-final stressed syllables. Yet, it is argued (see: Ladd 1996:211) that there is need for such distinction, because even though prenuclear and nuclear accents are not phonetically distinct they have different structural roles in tunes.

First of all, while nuclear pitch accents are obligatory components[2] of intonational phrases, prenuclear accents are not. In monosyllabic utterances the tune is realized by a single pitch accent, nucleus (followed by the boundary tone) which determines the intonational meaning that this tune conveys. In polysyllabic utterances the nucleus is associated with the most prominent stressed syllable which usually occurs in the last content word in an utterance.

Secondly, nucleus signals focus and is involved in signaling different grammatical, attitudinal and discoursal meanings (Cruttenden 1997:96). For example the rising-falling English nuclear accent „involves a sense of finality, completeness, definiteness and separateness, particularly used with declaratives" (Cruttenden 1997:100). The special status of nuclear pitch accents is also confirmed by the fact that modifications in the prenuclear melody

[2] However, there are languages like Dutch in which intonation phrases may have no nuclear pitch accent (Gussenhoven 2005).

do not change the message conveyed by the nuclear melody. Consider the following example (from Ladd 1996:208).

Relative neutral intonation:

a) **H*** **H*L-L%**
I read it to Julia.

b) **H*** **H*** **H*L-L%**
I wanted to read it to Julia.

(1)

Surprise-redundancy contour:

a) **L*** **H*L-L%**
I read it to Julia.

b) **L*** **L*** **H*L-L%**
I wanted to read it to Julia.

(2)

In the above examples, the nuclear melody remains the same, but the prenuclear one changes both with respect to the number and type of prenuclear accents. Changes in the number of prenuclear accents in phrases a) and b) do not affect the meaning conveyed by the tune. On the contrary, the substitution of H* prenuclear pitch accents by L* pitch accents in (2) adds a nuance of surprise to the neutral declarative intonation contours in (1). Thus, while the nuclear melody signals the same modality (i.e. declarative sentence) in all the examples, some paralinguistic modification concerning attitudinal meaning is introduced by the prenuclear melody.

The position of nucleus may shift from the last stressed syllable to a stressed syllable of some earlier position in the phrase, in which case the nucleus functions as a *focus* or *focal accent.*

2.2. Phrase/sentence level

Focalization is the first intonation function at the *phrase/sentence* level to be discussed in this section. Other functions at this level include *phrasing,* structuring information in the utterance into *topic* and *comment,* and conveying different *modes* and *expressivity.* Attention will also be paid to issues related to *prosodic structure* and *strict layer hypothesis*.

Phrase and sentence are also the domain of intonation forms – *pitch range, register*, *key and* recurring pitch patterns which include *intonational tunes* conveying various intonational messages (linguistic/paralinguistic) and *downward trends* in the pitch contours such as *declination* or *downstep*.

2.2.1. Focus

The term *focus* or *focal accent* refers to relations of relative prominence in a prominence pattern of an utterance. Apart from pitch prominence, like other pitch accents focus can additionally be realized by intensity and duration. In some languages (e.g. Hungarian) it can be manifested by changes in the syntactic structure of an utterance as well.

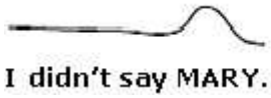
In the *narrow* sense, the term focus refers to "a part of utterance which is highlighted by pitch prominence" (Hirst & Di Cristo 1998:31), whereas *broad focus* makes all parts of an utterance equally prominent. The distinction between broad (1) and narrow focus (2) is illustrated below (author's example).

(1)

(2)

(1) is a neutral statement with a nuclear pitch accent on the last content word. By shifting the nucleus from PAINT to WHITE, the speaker highlights the most important information in an utterance and *emphasizes* it. *Emphasis* is the first function of focus. It gives rise to the division of the information structure in an utterance into *new* and *old* (or *given*): „old information will (…) fall outside the scope of focus, and new information will generally constitute the scope of focus" (Cruttenden 1997:89). The part of the utterance which refers to the old information is referred to as *presupposition*.

Focus can also be used to express *contrast*, in which case the focal part of an utterance may include an old information. This function of focus is depicted in (3) below (author's example), where the name MARY is contrasted with MANNY.



(3)

2.2.2. Phrasing

Intonation organizes an utterance into a hierarchical prosodic structure. An *intonational phrase* (also: intonation unit, prosodic sentence) constitutes the second highest-level unit (after utterance) in this structure. Intonational phrases are considered as sequences of *intonational events*, i.e. pitch accents and boundary tones. They consist of *feet* (also referred to as accent groups) and syllables; The former are regarded as prosodic units intermediate between stressed syllables (the smallest units in the hierarchical prosodic structures) and intonational phrases.

Intonational phrases have to include one obligatory (nuclear) pitch accent and are characterized by semantic and syntactic coherence. Moreover, they constitute the domain of recurring *intonation patterns* which convey various messages (e.g. sentence mode, expressivity) and therefore can be considered as *units of information* (Hirst & di Cristo 1998:30). Although the correspondence between syntactic and semantic units on the one hand, and intonational phrases on the other is not straightforward (e.g. Botinis, Granstroem & Möbius 2001:269), it is generally agreed that to some extent intonational phrases correspond to clauses (Hirst & di Cristo 1998:36).

The experiment presented in (Harris, Umeda & Bourne 1981) showed that while listeners agree on the position of intonational phrase boundaries to a great extent (in 83%), their decision is motivated by different factors. And indeed, phrase boundaries can be signaled by a variety of acoustic cues: pause (e.g. Lea 1979, Horne, Strangert & Heldner 2003, Bulyko &

Ostendorf 2001), final syllable lengthening, (Berkovits 1994, Cruttenden 1997, Horne, Strangert & Heldner 2003, Yoon, Cole & Hasegawa-Johnson 2007), terminal pitch movement (Yoon, Kim & Chavarria), presence of nucleus, change in tempo or pitch resetting (Hirst & Di Cristo 1998). As these are continuous parameters it is expected that some boundaries are perceived as stronger and others as weaker. This issue was investigated by De Pijper and Sanderman (De Pijper & Sanderman 1993, 1994) who found three factors influencing the perception of boundary strength: P - *pause length* (6 classes), M - *nuclear melody type* (7 classes) and R - *pitch reset* (3 classes: none, upstep, downstep). One of the consequences of this phenomenon is the search for units that could be delimited by boundaries of different strength. A very good example is the introduction of the *intermediate phrase* - a structural unit intermediate between foot and intonational phrase (Beckman & Pierrehumbert 1986).

Studies on intonational phrase boundaries in Polish have proven the significance of the acoustic factors listed above for perception of phrase boundaries. Extensive acoustic analyses presented in (Demenko 1999) have shown that lengthening of vowels in phrase-final and penultimate syllables serves as a cue for perception of a boundary. There exist also a strong correlation between the occurrence of pauses, pitch movements and declination of the f0 parameter to minimum f0 value at the end of a phrase and perception of phrase boundaries (Demenko 2000).

In (Steffen-Batóg & Katulska 1984) it was shown that the occurrence of nuclear accent is not always regarded as a cue for a phrase boundary: in a perception test listeners found more nuclear accents than phrase boundaries.

2.2.3. Prosodic structure

Intonational tunes are regarded as strings of pitch accents and edge tones (Pierrehumbert 1980, Taylor 2000), but apart from that tunes have *hierarchical constituent structure* (Ladd 1996:204). The most important part is the *nucleus* – pitch accent associated with the most prominent stressed syllable that belongs usually to the last content word of an utterance. Nucleus constitutes an obligatory and the most important element of any prosodic structure. The idea of nucleus has been incorporated into various theories of prosodic structure beginning with (Palmer 1922) where tunes consist of three components: *head* (prenuclear part of the tune), *nucleus* and *tail* (including postnuclear accents). In later works (see: Ladd 1996:209) head is defined as a sequence of syllables starting at the major stressed syllable, whereas preceding part of a tune is considered as a *prehead*. In the IPO model (t'Hart, Collier & Cohen 1990) a distinction is drawn between three components of different structural roles in the tune, i.e. *prefix* (which corresponds to head), *root* (→ nucleus) and *suffix* (→ tail). At a nuclear position a number of specific pitch movements can occur - they determine the shape of the postnuclear contour. The prenuclear part is independent from the nuclear pitch accent, but some combinations of prenuclear and nuclear accents are more common than others. In (Ladd 1996:208) it is said that all prenuclear accents in a tune must be of the same type (i.e., must belong to the same tonal category). As shown in the examples (1) and (2) in sec. 2.1.3 modifications in the prenuclear part of the tune (i.e., different number of accents resulting from association with different texts, different types of prenuclear pitch accents) do not affect the

basic linguistic meaning conveyed by nucleus (e.g. statement or interrogation), but adds some intonational nuance to that meaning (Ladd 1996:208).

In the Pierrehumbert system (Pierrehumbert 1980) the idea of a structural difference between prenuclear and nuclear accents was rejected and nucleus was considered solely as the last pitch accent that occurs in an intonational phrase. The nuclear melody is determined by the pattern associated with nuclear pitch accent and phrase accent and\or boundary tone. There is also no place for the tail in the intonational tune structure, because pitch movements that follow the nucleus are attributed to the phrase accent/boundary tone. Pierrehumbert's model makes also other assumptions concerning constituent structure of tunes, namely that "there is a hierarchy of prosodic domain types such, that in a prosodic tree any domain at a given level of the hierarchy consists exclusively of domains at the next lower level of the hierarchy" (after Ladd 1996:238). This approach is referred to as *strict layer hypothesis.* It assumes that prosodic domains are *non-recursive*, i.e. in a prosodic tree no constituent can dominate a constituent at the same level (e.g. an intonational phrase can not include another intonational phrase). However, consider the example below (from Ladd 1996:242):

(A) Warren is a stronger campaigner, and (B) Ryan has more popular policies, but
(C) Allen has a lot more money.

It can be said that the cohesion is greater between phrases (A) and (B) than (B) and (C), because there is a conjunction between the former and opposition between the latter. Therefore, it can be expected that the prosodic boundary is stronger between (B) and (C) than between (A) and (B). There exist a number of acoustic cues (e.g. duration, range of the pitch movement, see sec. 2.2.2 and 6.1.1) which signal different *boundary strength* and cohesion between intonational phrases. As proved in a number of studies (e.g. Horne, Strangert & Heldner 1995, Carlson & Swerts 2003) these differences are perceptually significant. The strict layer hypothesis as defined above does not give an opportunity to reflect this phenomenon, and the three intonational phrases in (1) would be represented as equal nodes of some higher level prosodic domain, which is not really the case (see Figure 7a). The prosodic structure of the utterance in (1) can be represented by the tree in (b) below.

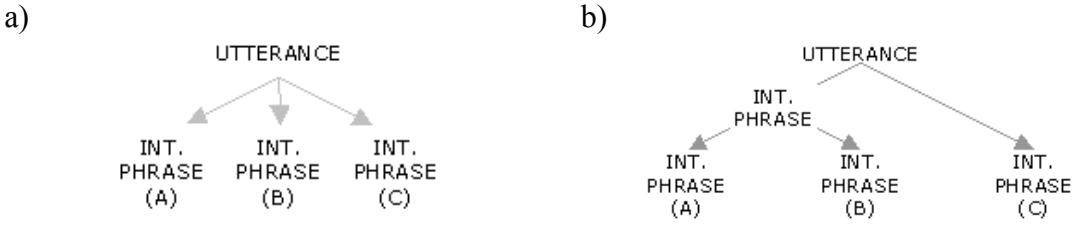


Figure 7 (after Ladd 1996:243): Examples of hierarchical a) and compound prosodic structure b).

To solve this problem a constituent smaller than intonational phrase was introduced - an *intermediate phrase* (Beckman & Pierrehumbert 1986) which is marked by a weaker boundary. In (O'Connor & Arnold 1973) single and double bars were used to signal weaker and stronger prosodic boundaries respectively. However, it seems that there is no reason for prosodic structure to be restricted to two levels of phrasing, since phrase boundaries are signaled by

continuous acoustic parameters which can take many values and consequently, signal boundaries of a different strength. These issues are addressed in (Ladd 1996) where a new element of prosodic structure is introduced, i.e. *compound prosodic domain*. It is defined it as follows (Ladd 1996:244): “a CPD is a prosodic domain of a given type X whose immediate constituents are themselves of type X”. This approach allows representing a different boundary strength (and cohesion between prosodic constituents) without the need of generating more prosodic domain types than it is really needed.

Further discussion on prosodic structure and phrasing Polish, and analyses of their effect on pitch variation are presented in sec. 5.1.

2.2.4. Topic and comment

One of the functions of intonation at the phrase/sentence level is "to provide a way of breaking up an utterance into information units" (Hirst & Di Cristo 1998:30): *topic* and *comment* (also referred to as *theme* and *rheme*). This binary division has its origin in gestalt psychology, where in visual perception of a stimulus a distinction is drawn between *figure* and *ground*.

Topic is considered as the subject of the utterance i.e., what the utterance is about, whereas *comment* is considered as the part of an utterance which says something about the subject. The concept of topic and comment is closely related to that of *focus* and *presupposition* (discussed in sec.2.2.1), because all of them refer to information structure, and very often they are treated as equivalents.

There is no general consensus on the nature of correspondence between intonation and information structure (after Hirst & Di Cristo 1998:30), examples are listed below:

a) in (Halliday 1970) an assumption was made that nucleus of a given intonation unit marks the end of a new information; an opposite claim can be found in (Schmerling 1976)
b) nucleus does not necessarily has to mark boundaries of an intonation unit, because the latter may contain more than one nucleus (Brown 1983)
c) "given information can be just as accented as new information" (Noteboom & Kruyt 1987)
d) different items can be characterized by different degree of giveness, thus the opposition given-new is scalar rather than binary

Topic and comment are realized by specific pitch patterns: some authors (see Hirst & Di Cristo 1998:30) assume that there is a specific accent type which occurs in the rheme, others associate higher accents with comments and lower accents with topics. The choice of a specific patterns for the topic and comment depends on the mode of utterance and position of the theme in the utterance.

2.2.5. Intonational tunes

Tunes (also referred to as *intonation patterns*) are properties of intonational phrases (Ladd 1996) and can be regarded as recurring, distinctive pitch patterns. On the phonological level tunes are strings of abstract tones: prenuclear, nuclear and postnuclear; The latter may "surface either as accents or edge tones depending on the metrical structure of the segmental material to which they are associated "(Ladd 1996:219).

Many intonational tunes are universal: they constitute elements of intonation systems of languages of different origin and very often have similar functions. Examples are given in (Hirst & Di Cristo 1998):

a) most languages use falling intonation for *statements* and *WH-questions*
b) *unemphatic yes/no questions* and *question tags* are characterized by rising intonation and especially by high phrase-final pitch in about 70% of a sample of 250 languages (Bolinger 1978 after Hirst & Di Cristo 1998:25). Among the remaining 30% languages there are Danish, Finnish, Hungarian, Bulgarian, Russian, Western Arabic, Brazilian Portuguese.
c) *repeat questions*, i.e. a questions which echo the previous question or call for repetition of what was said have rising nuclear intonation
d) *unfinished utterances* or *continuation phrases* are characterized by rising nuclear melody (or at least by no drop in phrase-final pitch, Ladd 1996:114). In some languages the rise can have bigger range than it has in questions (e.g. Greek, British English).
e) *stylized contours* are characterized by intonation which is more similar to that of singing than speech. *Calling contour* is the most common example. This is a simple intonational phrase which consists of a nuclear accent realized by fall in pitch from the top of speaker's range to a level of a downstepped high tone or to the middle of speaker's range.
f) other universal intonational tunes include for example (Hirst & Di Cristo 1998:36): *enumerations* (Spanish, French, German), *greetings* (German), *warnings* (Thai)

The examples of intonational tunes listed above show that intonation contributes to the *expression of modes*, because the use of a specific intonation pattern signals a specific mode e.g. *interrogative* or *declarative*. There are also other correspondences between intonation and modes/expressivity e.g., high pitch may indicate *surprise*, high register may be used to signal *obviousness,* whereas delayed peak may express *hinting* (Hirst & Di Cristo 1998:27). In some languages (e.g. Brazilian Portuguese) distinct intonation patterns express different types of utterances e.g. *commands, requests, suggestions* or *threats*.

2.2.6. Downward trends

The term *downward trends* refers to *declination, downstep* and *downdrift* considered as global f0 trends involving lowering of the pitch range, and *final lowering* – a local characteristic of f0 contours consisting in a rapid fall in pitch associated with the phrase final pitch accent. The three trends are described in the following way (Hirst & Di Cristo 1998:21).

a) *declination* is defined as "a strictly phonetic characteristic of utterances consisting of a continuous lowering from the beginning to the end of the intonation unit"
b) *downstep* consists in iterative lowering of f0 peaks of successive pitch accents such, that "the value of each accent peak in a downstep series is a constant proportion (…) of the previous peak" (Ladd 1996:77)
c) *downdrift* can be given the same definition as *downstep* with the difference that there are intervening low tones between pitch accents

There exists experimental evidence that declination, downstep and downdrift are features of laboratory speech rather than spontaneous speech (Hirst & Di Cristo 1998, Botinis, Granstroem & Möbius 2001). They affect prenuclear parts of utterances (mostly declarative) or whole intonational phrases and the reset of the lowered pitch range takes place before the nuclear part of the intonational tune or at phrase boundary.

From the production point of view, the downwards trends are related to the decrease of subglottal pressure (t'Hart, Collier & Cohen 1990). However, there is no agreement whether they are controlled by the speakers or not: for example, Pierrehumbert (Pierrehumbert 1980) considers downstep as an effect triggered automatically by specific tone sequences, whereas Ladd (Ladd 1996) defines it as a speaker's choice and points out that downstep conveys specific meanings. In the IPO system (t'Hart, Collier & Cohen 1990) declination is regarded as an automatic process.

2.2.7. Pitch range

The realization of intonational/pitch features is affected by many factors: "pitch differs from speaker to speaker (e.g. male vs. female speech), from occasion to occasion (e.g. bored vs. angry speech), and even from one part of utterance to another (e.g. declination and other similar effects)" (Ladd 1996:252). Therefore, in the analysis of intonation some point of reference has to be determined. *Pitch range* can be used as such a reference point. In general, there is no agreement how pitch range should be measured: whether it is the difference between the maximum and minimum f0 of a given speaker's voice (Clark 2003), a difference in ST between 90th and 10th percentile or +/-2 standard deviations around the mean (Mennen, Schaeffler & Doherty 2007). There are also proposals suggesting that pitch range should be described in terms of clearly defined linguistic targets: "sentence initial peaks, accent peaks, post-accent valleys and sentence final lows" (Patterson & Ladd 1999:1169).

Descriptions which provide characterization of pitch relative to speaker's voice are called *normalizing.* (e.g. Jilka, Möhler & Dogil 1999, Möhler 2001). There are also *initializing descriptions* in terms of which pitch target level is expressed relative to the level of the preceding target. The initializing description was partly[3] adopted in the INTSINT intonation transcription system (e.g. Campione, Hirst & Véronis 2000, Hirst, Di Cristo & Espesser 2000, Hirst 2001). In general, results of various experiments discussed in (Ladd 1996:252ff.) suggest that normalizing descriptions of intonation are more useful for intonation modeling then

[3] In INTSINT both normalising and initialising approaches are applied: the scaling of tonal targets is determined both with respect to speaker's overall range and with respect to the level of the preceding target.

initializing descriptions: they are capable of abstracting away from differences caused by extrinsic factors (different speakers, prosodic structure) to a greater extent and in a more consistent way than initializing descriptions do.

As a matter of fact, pitch range is not a single variable, because pitch variation involves either modification of the *overall level* or *span,* or both (Ladd 1996:260). Changes in the overall level affect all tonal targets in the same way i.e., if the overall level is raised then the level of all tonal targets is raised too. But changes in span (i.e. expansion or compression of the range of the frequencies used by the speaker) have impact on scaling of high and low tones, but the position of M tones is unaffected. These effects are depicted in Figure 8.

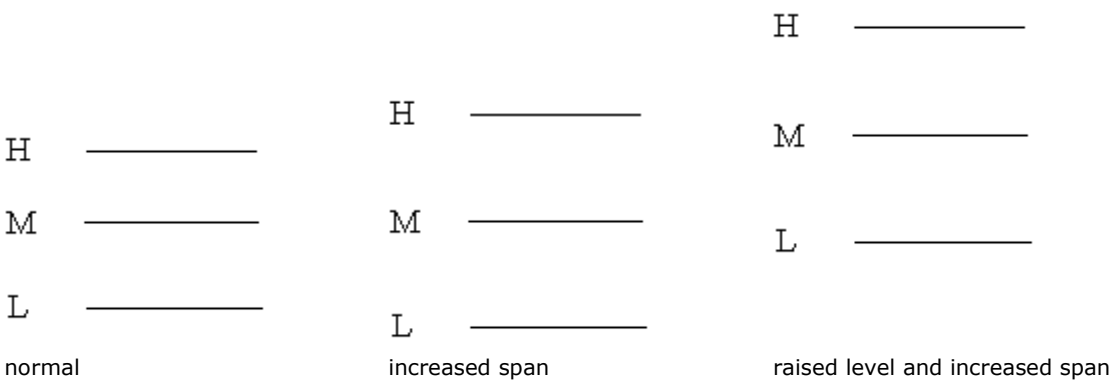


Figure 8 (after Ladd 1996:262): Illustration of the distinction between level and span.

Thus, when dealing with different speaker in intonation analysis and modeling a choice has to be made between initializing vs. normalizing description. Another issue concerns the frequency scale on which pitch will be represented (Hermes & van Gestel 1991). Usually the logarithmic scale is adopted which, unlike the linear Hertz scale, makes it possible to compare intonation of female and male speakers. Average pitch range of female speakers is between 180 and 400Hz (but in singing the top of the range may be shifted up to 1000Hz and even higher), whereas pitch range of male speakers is on average between 60-200Hz. A rise in pitch from 100 Hz to 150 Hz in a male voice has the span of 50 Hz and the rise from 180 Hz to 270 Hz in a female voice has the span of 90 Hz (after t'Hart, Collier & Cohen 1990:24). But in a fact, these two rises are perceived as having the same range, which can be accounted for after logarithmic transformation of the f0 values (then both rises equal 7.02 ST).

In (Cruttenden 1997) instead of span and overall level, the terms *key* and *register* are used to describe variation in pitch range. *Key* is defined as varying "the width of the pitch range over whole intonation groups" (Cruttenden 1997:129). *Register* is considered as an "overall (upward and downward) shifting of the whole pitch range within which the speaker is speaking" (Cruttenden 1997:55). The function of register is to signal the emotional state of a speaker e.g. tension, stress or anger. High register may signal defense or social politeness (Cruttenden 1997:130).

2.3. Discourse/dialogue level

The last functions of intonation to be discussed regard the *discourse/dialogue level*. Intonation is involved in structuring discourse and dialogue into information units such as *topics* and *sub-topics*. For this purpose *key* can be used: it signals cohesion between intonation groups (Cruttenden 1997:54) and indicates "the beginning and end of a topic: high key indicates the beginning of a new topic and low key indicates the end of a topic" (Cruttenden 1997:129). Similar remark is made in (Botinis, Granstroem & Möbius 2001:278): higher tonal patterns signal start of the topic and lower tonal patterns signal topic end.

Another level of discourse and dialogue organization concerns division of the utterance into *turn-units* which reflect speakers' contribution to the conversation development. Here, units such as *turn-keeping, turn-taking* and *turn-leaving* are marked by intonation. Continuation rises, i.e. tunes with a final rise, are applied by the speakers to signal turn-keeping; falling pitch at the phrase boundary is used to signal turn-leaving (Batliner et al. 2001:279).

2.4. Intonational meaning

Intonation conveys universal and language specific meanings which derive from two different components of a language: phonetics and phonology/intonational morphology (Gussenhoven 2002). On the phonetic level intonation conveys *paralinguistic meanings* which concern aspects of interpersonal interaction e.g. aggression, appeasement, solidarity, condescension and features of the speaker e.g. anger, joy, boredom (Ladd 1996:33). On the phonological level intonation is involved in conveying *linguistic meanings*. They concern features of an utterance/message and include: *focus*, *modes* and (to some extent) *expressivity*, *topic* and *comment*, structuring discourse/dialogue into *turn-units.* Linguistic and paralinguistic meanings are realized by the same acoustic properties: loudness, voice quality and pitch range. What distinguishes paralinguistic from linguistic meanings is that while the former have scalar or gradient nature the latter are characterized by quantal and categorical structure (Ladd 1996:36). Paralinguistic cues can be regarded as "modifications of the way in which phonological categories are realized" (Ladd 1996:35). As a result of paralinguistic modification "an instance of a given phonological category may sound like an instance of some other", but even then the basic linguistic identity of a contour remains unaffected. For example, in a statement contour that rises to a peak and then falls to the bottom of the range, higher peak may be used to signal emphasis (from Ladd 1996:39).

Extralinguistic or *non-linguistic messages* deal with speaker's age, sex, social status etc.

The universal intonational meanings derive from three biologically determined codes (Gussenhoven 2002). The codes reflect various aspects of speech and f0 production and can be interpreted as different relations between intonation form and function. They include frequency, effort and production codes.

1. *Frequency code* reflects "correlation between larynx size and rate of vocal folds vibration" (Chen, Rietveld & Gussenhoven 2000:91). Children larynxes are smaller and produce speech of a higher fundamental frequency. This may explain why "high pitch sounds

vulnerable and submissive" in contrast to low pitch of male adult speech which is perceived as "protective and dominant" (Gussenhoven 2002:48). Frequency code is used to make sentence mode distinctions: high final pitch signals questioning and low pitch signals declarative utterance, because it sounds assertive.

2. *Effort code* reflects relation between the amount of energy put by the speaker in speech (and f0) production process on the one hand and pitch variation on the other: "greater effort corresponds with wider [pitch] excursions, and as a result, a higher peak will sound more emphatic than a lower peak" (Gussenhoven 2002:48). The function of the effort code is to signal emphasis, surprise and obligingness.

3. *Production code* reflects how "energy is parceled out in chunks that coincide with the exhalation phases of the breathing process" (Chen, Rietveld & Gussenhoven 2000:91). This code is applied to signal beginnings (→high pitch) and ends of phrases (→low pitch), and to distinguish between continuation phrases (→high final pitch) and statements (→low final pitch) (Gussenhoven 2002:54).

The universal intonational meanings may undergo grammaticalization, in which case they become part of the phonology of a given language system. This explains why mapping between intonation forms and functions differs among languages and why "languages (…) possess form-meaning relation in their grammars, which go against the universal, biological codes" (Gussenhoven 2002:48). For example, once signaling of questioning by high final pitch became part of the grammar (in 70% out of 250 analyzed languages, Hirst & Di Cristo 1998:25), intonation systems of some languages evolved to signal questioning by falling pitch (e.g. Hungarian).

To sum up, universal intonational meanings have their origin in phonetics and the three biological codes. Language-specific intonation meanings result from grammaticalization and are embedded in the phonology of the language.

Chapter 3. Overview of intonation modeling – literature

This chapter gives an overview of the literature on intonation modeling. According to the typologies introduced in the next section some of the well known intonation models will be presented. From among the existing approaches only those were selected which have actually been applied in speech synthesis and present the theoretical assumptions underlying the models. It will be seen that different models share features and each model can be classified to more than one type. The discussion in this chapter will also provide an explanation for the choice of a surface phonological representation for the comprehensive intonation model proposed in this thesis (see also the section 1.1.4).

3.1. Typology of intonation models

Intonation models can be grouped with respect to features listed below.

1. The level of analysis and representation for description of intonation. On the basis of this factor a distinction is drawn between *phonetic* and *phonological* models. In between the phonetics and phonology an intermediate level of *surface-phonology* is proposed (Hirst, Di Cristo & Espesser 2000).

2. The way in which tunes are modeled gives rise to the distinction between *superpositional* and *sequential* models. In the former tunes result from superposition of two components of a different temporal scope i.e., phrase and accent components. The latter regard tunes as sequences of discrete elements (pitch accents, boundary tones, connections) which are associated with the elements of the segmental string.

3. The direction in which analysis of intonation is carried out: a distinction is drawn between *generative* and *analytical* models. In general, the generative approach involves a top down process in which f0 contours are produced from higher-level information. On the contrary, the analytical approach involves a bottom up process infers higher-level information from the f0 contour.

4. The way in which coding of f0 contours into description of intonation is carried out gives rise to the distinction between *data-driven* and *rule-based* models. The former use various machine learning techniques (neural networks, regression trees) to generate f0 contours from the symbolic input. In the latter approach f0 contours are generated from rules defined by experts.

In the following sections some basic theoretical assumptions underlying the models are presented.

3.1.1. Phonetic versus phonological

Phonetic models are regarded as quantitative. In phonetic models intonational features are described in terms of vectors of acoustic features or continuous parameters (e.g. duration, amplitude, slope, f0peak position) which interact with one another. In f0 contour generation the values of the parameters are estimated from symbolic input by a regression model (e.g. Black & Hunt 1996, Dusterhoff & Black 1997, Mixdorff 2002a). Depending on whether the model is sequential or superpositional the f0 contour of an utterance results from interpolation between the estimated pitch targets (e.g. Momel, PaIntE, Tilt) or superposition of the components of different temporal scopes (e.g. Fujisaki model and its adaptations to different languages).

Phonological models are qualitative and sequential. In phonological models intonational tunes are considered as sequences of distinctive discrete tonal categories. As a result of detailed acoustic analyses an inventory of tonal categories and intonational grammar are defined which provide framework for transcription of intonation. As opposed to phonetic models which account for melodic aspects of intonation in the first place, phonological models represent the analytical approach: in the first place they account for functional aspects of intonation which are related to higher-level linguistic information.

In f0 contour generation the alignment and scaling of tonal targets is determined from rules devised by a human expert (Anderson, Pierrehumbert & Liebermann 1984, Jilka 1996, Jilka, Möhler & Dogil 1999).

3.1.2. Superpositional versus sequential

In superpositional model intonational tunes are described in terms of two non-categorical components of different temporal scopes. One of them models the global shape of the contour over the length of an intonational phrase, whereas the other component models local pitch changes associated with accented syllables/accent groups and pre-boundary syllables. The global and local pitch variation is described in terms of acoustic parameters such as duration and amplitude of the commands. In generation of f0 contours the component of the shorter temporal scope (accent component) is layered onto the component of a longer scope (phrase component).

In sequential models f0 contours are considered as strings of intonational events (pitch accents, boundary tones) associated with the elements of the segmental string (accented syllables, pre-boundary syllables). In the Tilt model (Taylor 2000) there is an additional element - connection which models the sections of the contour between the events. Depending on the model type - phonetic vs. phonological, pitch accents and boundary tones are described in terms of tonal categories (e.g. ToBI) or in terms of vectors of acoustic features (e.g. Tilt or PaIntE). The f0 contours are produced by interpolation between the intonational events.

3.1.3. Other types

Other types of intonation models include *data-driven* and *rule-based models*, and generative and *analytical models*. They will not be discussed here in detail, because these distinctions are secondary and all the models are primarily classified according to the typologies presented in the previous sections.

The difference between *data-driven* (e.g. Black & Hunt 1996, Dusterhoff & Black 1997, Reichel 2007) and *rule-based* models (e.g. Anderson, Pierrehumbert & Liebermann 1984, Jilka 1996, Jilka, Möhler & Dogil 1999) concerns the way of a) *coding* of f0 contours into intonation description at some level of representation and analysis and b) *generating* the contours from symbolic description of an utterance. These two processes can be carried out on the basis of rules devised by an expert (rule-based approach) or derived automatically by means of some machine learning techniques like decision trees or neural networks (data-driven approach). In general, data-driven models are considered as more universal, because they can be easily adapted to modeling of intonation of different languages and speech styles.

Generative models (e.g. Pierrehumbert 1980) are based on phonological descriptions of intonation determined with respect to functional aspects of intonation in the first place and melodic aspects in the second. Building of a generative model involves "specifying accent levels and phrasal boundaries, transforming these into an abstract description of the duration and f0 contours which is then by application of phonetic rules converted into the actual f0 contour" (Mixdorff 2002a:27). On the contrary, analytical models (e.g. Tilt, PaIntE) derive the intonation description from purely phonetic information, i.e. observed pitch contour.

3.2. Phonetic sequential models

In this section two of the most influential phonetic sequential intonation models applied in speech synthesis systems are presented, namely *Tilt* and *PaIntE* models.

3.2.1. Tilt intonation model

In the *Tilt* model of intonation (Taylor 1998, 2000) tunes are regarded as sequences of discrete *intonational events*: *pitch accents* and *boundary tones* associated with elements of the segmental string: accented syllables and phrase final syllables. Stretches of the contour between the events are described as *connections*. These constituents provide a framework for transcription of intonation on the phonological level and are defined in (Taylor 2000:1698):

a) *pitch accents* are labeled with **a** and defined as "f0 excursions associated with syllables which are used by the speaker to give some degree of emphasis to a particular word or syllable"
b) *boundary tones* are labeled with **b** and considered as "rising f0 excursions which occur at the edges of intonational phrases and as well as giving the hearer a cue as to the end of the phrase, can also signal effects such as continuation and questioning"

c) *connections* are denoted by **c** and as mentioned above are sections of contour between the events
d) **ab** label is used to describe situations when "a pitch accent and boundary tone occur so close to one another that only a single pitch movement is observed"

If there is need this basic label set can be extended, for example instead of a single label describing a phrase boundary a distinction can be drawn between rising and falling boundaries and appropriate labels can be used to mark them. The resulting description of intonation is compact (there is a limited set of labels) and functional (i.e., it accounts for the functional aspects of intonation). The intonational events i.e. pitch accents, boundary tones (and combined intonational events ab) are realized by distinctive pitch movements (bi-directional rising-falling in case of pitch accents and uni-directional in case of boundary phenomena) and described in terms of continuous parameters derived automatically from the f0 curve (Taylor 2000):

a) *amplitude* is the sum of magnitudes of the amplitudes of the falling and rising f0 curves: "the rise amplitude is the difference in f0 between the f0 value at the peak and at the start of an event, and the fall amplitude is the difference in f0 between the f0 value at the peak and at the end of an event" (op.cit.: 1699). Normalized with respect to pitch range, the amplitude parameter reflects the perceived prominence of the event.
b) *duration* is the sum of durations of the falling and rising f0 curves: "the rise duration is the distance in time from the start of the event to the peak and the fall duration is the distance from the peak to the end" (op.cit.: 1699)
c) *tilt* parameter encodes the overall shape of the intonational event (see Figure 10). It is calculated from the amplitudes and durations of the rising and falling f0 curves (op.cit.: 1706)

$$\text{tilt} = \frac{|A_{\text{rise}}| - |A_{\text{fall}}|}{2(|A_{\text{rise}}| + |A_{\text{fall}}|)} + \frac{D_{\text{rise}} - D_{\text{fall}}}{2(D_{\text{rise}} + D_{\text{fall}})}$$

Equation 1: Equation describing the tilt parameter.

d) for each intonational event the start-f0 parameter is defined. It contains f0 value at the beginning of the event and determines its position in time either with respect to the start of the utterance or nuclei of the syllable with which the event is associated
e) syllabic position parameter determines the alignment of the intonational event with respect to the accented syllable or accented vowel. Together with the tilt parameter it encodes categorical differences between pitch accents.

Figure 9 and Figure 10 illustrate the interaction between Tilt and syllabic position parameter and show correspondences between Tilt and ToBI representations. On the right, shapes of intonational events of a different Tilt value are given.

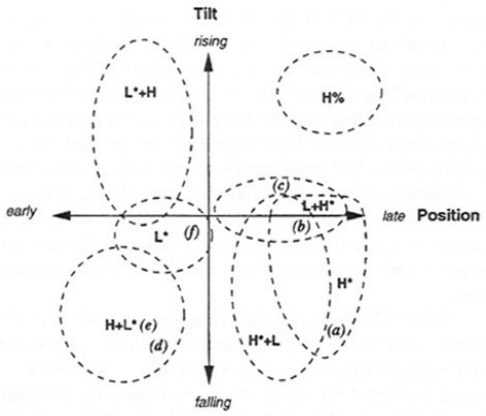


Figure 9 (adopted from Taylor 2000:1711): The interaction of Tilt and syllabic position parameters

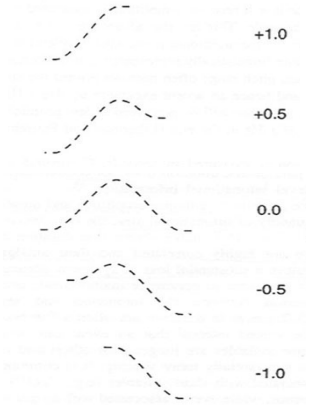


Figure 10 (adopted from Taylor 2000:1706): Different shapes of f0 curves described by Tilt parameter

Tilt model provides a method of coding of a continuous f0 curve into higher-level categorical representation (in terms of intonational events etc.) as well as a method of converting this representation into a continuous f0 curve. The model is data-driven and has to be trained on a speech database labeled with the intonational events, connections and silences. The first step to derive Tilt representation from an f0 curve involves detection of the intonational tune constituents. The intonational event detector uses a Hidden Markov Model (HMM) recognizer for the segmentation of an f0 curve into its constituents on the basis of acoustic information only. The next step consists in RFC parameterization (Taylor 1993, 1995): the intonational events are described in terms of amplitudes and durations calculated separately for the falling and rising f0 curves of the event. From this parametric representation the Tilt representation is derived.

In the speech synthesis scenario, the first task is prediction of the intonational structure from text, which is performed most often by means of classification trees. The next step consists in estimation of the Tilt parameters from a symbolic input which is performed by regression models (e.g. Dusterhoff & Black 1997). Each of the Tilt parameters is predicted for each of the constituents of the tune by a separate model. The estimated parameters are converted into RFC representation (Taylor 1992). Finally, the f0 contour is reconstructed from the RFC parameters.

Both analysis and synthesis of intonation contours in the framework of Tilt theory is based on parameterization of intonation contours which is carried out during stylization of the contours. In the Tilt and RFC models the stylization is carried out only on the fragments of the contour which are associated with intonational events (connections are modeled by straight lines).

Monomial functions are used for this purpose; they are described by the equation below (Taylor 1992:71) where γ is the coefficient of curvature:

$$y = 1 - 2^{\gamma-1}.x^{\gamma} \quad 0 < x < 0.5$$
$$y = 2^{\gamma-1}.(1-x)^{\gamma} \quad 0.5 < x < 1.0$$

Equation 2: Equations describing the Tilt model functions.

Before the approximation function is fitted to fragments of an f0 contour delimited by the intonational event detector, they are smoothed with a 15-point median filter in order to remove erroneous f0 values. Preprocessing involves also a linear interpolation through unvoiced regions.

Monomial functions proved to give very accurate approximations of f0 curves of different curvatures and corners (i.e. regions near the turning points in an f0 contour where the start and end of the approximation function are located). The coefficient of curvature was set to two. After approximation the f0 curve of an event is parameterized in terms of durations and amplitudes of its falling and rising components.

An example of application of the approximation method used in the Tilt model is the *PitchLine* method described in detail in the sec. 3.7.4.

3.2.2. The PaIntE model

PaIntE model (Möhler 1998, 1999) shares many features with the Tilt model. First of all, intonation contours are considered as strings of intonational events: pitch accents and boundary tone associated with accented and pre-boundary syllables, and described by a set of continuous parameters.

In the PaIntE model the parameterization of the intonational events is carried out by a 4th order polynomial function described by the equation below (Möhler 2001:1).

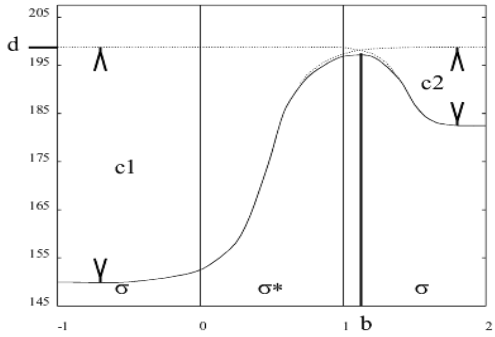
$$f(x) = d - \frac{c_1}{1+\exp(-a_1(b-x)+\gamma)} - \frac{c_2}{1+\exp(-a_2(x-b)+\gamma)}$$

Equation 3: Equation describing the PaIntE model functions.

The parameterization proceeds as follows: “First the f0 contours are segmented according to their syllabic annotation. The approximation takes place within a 3 syllable window around the syllable carrying the accent. The f0 contour within this window is normalized with respect to the syllable lengths” (Möhler & Conkie 1998:2). Within the window a top line of the pitch range is determined so that the two curves share the same upper limit. Figure 11 illustrates a three syllable analysis window around the accented syllable and the

parameters which describe intonational events. On the right the six PaIntE parameters are described.

Unlike in the Tilt model where an approach is developed to automatically derive the position and type of intonational events, for the purpose of PaIntE modeling the information on syllable boundaries and position of pitch accents and boundary tones is required. The parameterization is carried out only for accented and boundary syllables and every parameterized intonation event are described as a sum of a rising and falling curves. The f0 contour results from linear interpolation between successive curves.



a1, a2 – steepness of the rising and falling sigmoid respectively

b – the alignment of the function (the unity is the syllable length)

c1, c2 – the amplitudes of the rising and falling sigmoid respectively which reflect the perceived prominence of an event

d – the function's peak

Figure 11 (adopted from Möhler 1999:2): The PaIntE model function and parameterization of the contour.

Vector Quantized PaIntE (Möhler & Conkie 1998) is an extended version of the PaIntE model and results from application of vector quantization onto the six PaIntE parameters. This approach is motivated by two facts. First of all "intonation theory suggests that the intonation can be described by a number of distinct shapes, which in turn can be related to semantic meanings" (Möhler & Conkie 1998:2) and secondly, reduction of the number of parameters and predicting them together instead of each one separately should result in a more effective f0 contour generation.

In order to normalize across speakers every PaIntE vector is determined by the highest *d* and the lowest *b* parameter found over the length of a given phrase and mapped onto the interval from 0 to 1.

To carry out VQ analysis of intonation, first a codebook of basic intonation patterns has to be designed. Codebooks may consist of a different number of entries – codewords, which constitute basic units of the model. Every codeword reflects a specific f0 movement associated with the syllable bearing a pitch accent or boundary tone. It seems that this approach makes it possible to create a linguistically relevant representation of intonation, because codewords can be related to specific intonational functions such phrasing, emphasis, modality. Yet, the determination of the number of codewords is based on statistical analyses results and is not verified in any linguistic or perceptual analyses. Although some perception tests were carried out (Möhler 1999, Möhler & Conkie 1998, Syrdal et al. 1998) they aimed at evaluation of the intonation modeling quality, not at investigating the intonation form-function relation.

Figure 12 presents the intonational events in a codebook including 8 entries.

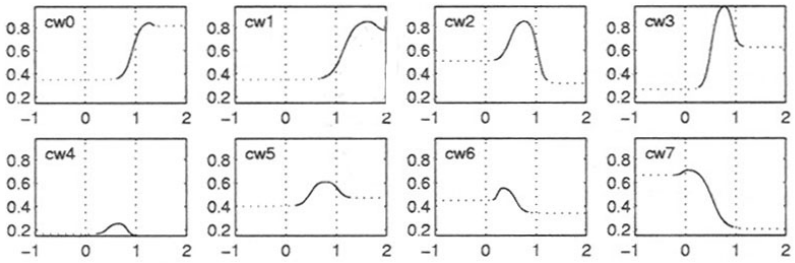


Figure 12 (adopted from Möhler 1999:3): The types of events represented in a codebook including 8 entries.

CART trees are trained to predict codewords from the symbolic input describing features of the utterance. Perception tests on the quality of generated intonation showed that better results can be obtained when some higher-level information is provided such as pitch accent type or features of accented and phrase boundary syllables (e.g. position within a phrase, total duration, onset and rhyme durations). The f0 contour of an utterance results from linear interpolation between predicted codewords. Its accuracy depends not only on the type of input information, but also on the codebook size. The best results in terms of the quality of the synthesized intonation are reported for codebooks of 8 and 16 codewords (Möhler & Conkie 1998, Syrdal et al. 1998).

The last improvements of the PaIntE model reported in (Möhler 2001) consisted in introducing a more linguistically motivated description of peak (d) position: it is mapped on one of the intervals corresponding to the onset, sonorant nucleus and coda of the accented syllable or the previous/next syllable.

3.3. Phonetic superpositional models

In this section assumptions underlying superpositional intonation models are discussed. The presentation is confined to those models which have been applied in speech synthesis systems.

3.3.1. The Fujisaki model

The Fujisaki model (Fujisaki & Hirose 1982, Fujisaki 1983) is a fully formal and superpositional phonetic model which be applied to analysis and synthesis of intonation of different languages. It provides physiological interpretation relating f0 movements with the activity of intrinsic larynx muscles. The model's representation consists of two components and continuous parameters which interact with one another and describe intonational phenomena in terms of duration and amplitude.

a) Phrase component consists of phrase commands in form of impulses and phrase control mechanism.
b) Accent component consists of accent commands in form of stepwise functions and accent control mechanism.

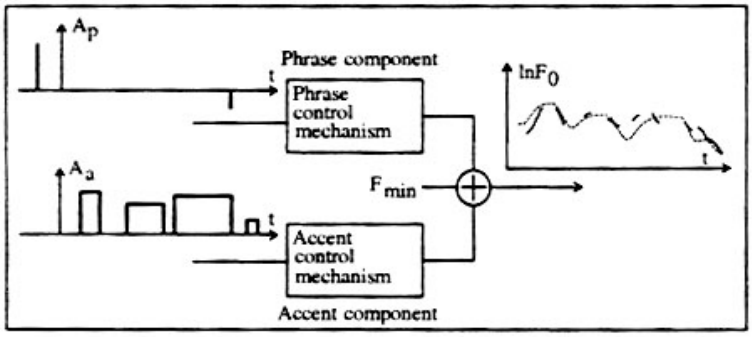


Figure 13 (adopted from Batliner et al. 2001:283 after Fujisaki 1998): Block diagram of the Fujisaki model.

The first component models global shape of an f0 contour and corresponds to declination (which is resetted at the beginning of a new phrase). The second component models pitch accents. Phrase and accents commands are processed in corresponding control mechanisms and added to Fmin value (a baseline determined for a given f0 contour): its Fmin value depends on the speaker, sentence mode and utterance length. The f0 contour of an utterance results from superposition of the phrase and accent components which makes the model superpositional.

The phrase component uses parameters such as amplitudes and timing of phrase commands and damping factors of the phrase control mechanism. The values of the parameters are determined for a specific phrase and have constant value over its duration. The accent component uses parameters such as amplitudes and timing of the onsets and offsets of accent commands, and damping factors of the accent control mechanism. They are determined for a specific accent group and have constant values over its duration.

Numerous synthesis experiments for different languages were carried out to assess the generated intonation contours. Generally good results were obtained but also some problems occurred. For English (Taylor 1992, 2000) the model's representation results to be too constrained and there are some intonation patterns that are difficult to synthesize. It is mainly due to the phrase component, which dictates the global shape of the f0 contour over the length of a phrase. It has fixed gradients and restricts variation in the f0 movements on accented syllables. The treatment of declination as a fixed component of the model has been often criticized (e.g. Hirst & Di Cristo 1998, Batliner et al. 2001), because declination is observed mainly in laboratory speech.

In Mixdorff application of the Fujisaki model for German (Mixdorff 2001, 2002a) special care had to be taken in the determination of timing rules for the accent commands with respect to accented syllable (or nucleus) onset. It was observed that for specific syllable structures the generated accents were too weak which induced perception of an accent shift to

the following syllable (Mixdorff & Fujisaki 1998). Another problem concerned proper detection and modeling of high boundary tones (Mixdorff & Fujisaki 2000).

In case of the application of the Fujisaki model to Polish intonation (Demenko 1999) the problem consisted in determining parameters of the accent commands: only two out of nine nuclear accents were properly modeled. The other problem concerned timing of f0 peaks and minima relative to the accented vowel onset. The general conclusion was the same as that presented in (Taylor 2000), namely that the model's representation is too constrained and the model lacks flexibility. Both authors report that the modeling of intonation using a sequential approach yielded better results.

3.3.2. Application of the Fujisaki model to German

The first application of the Fujisaki model to German to be discussed here comes from (Möbius et al. 1993, Möbius 1995). It follows the Fujisaki model in its composition: the model consists of two independent components layered on the Fmin value determined and constant for a given speaker. But unlike the Fujisaki model the Möbius model accounts for various sentence modes and defines rules to model them effectively.

Phrase component "serves as a baseline to the intonation contour" (Möbius 1995:5). The f0 contour of an utterance is shaped with reference to the Fmin value and with respect to phrasing, and difference between the f0 values on successive accented syllables. The global shape of an f0 contour over the length of a phrase corresponds to declination line and reflects sentence mode.

The domain of the accent component is accent group. It is defined as a sequence of syllables which begins with a syllable bearing an accent and ends before the next accented syllable.

The model accounts for various types of tunes: declarative, interrogative: wh-question, yes/no question and echo question. As it can be seen in the Figure 14 the tunes are divided into two parts and the final part which carries the information about sentence mode is treated as a separate phrase component. Therefore the values of the phrase and accent command parameters differ with sentence mode e.g. the amplitude of the final phrase accent command has a fixed value of –0.1 for statements and wh-questions (as to model the final lowering of the f0 contour); in case of yes/no and echo questions it has a fixed value of 0.2.

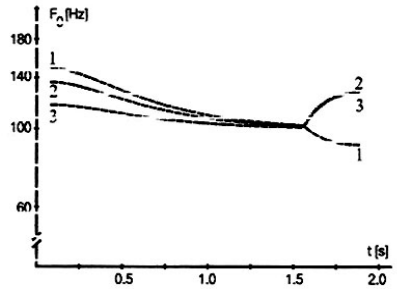


Figure 14 (adopted from Möbius 1995:5): Typical phrase contours of interrogative wh-question (1), yes/no question (2), and echo question (3)

The number of phrases in an utterance and syntactic relations between them are other factors determining the amplitudes and timing of the phrase command e.g. in a statement consisting of two phrases the amplitudes of both phrases will be relatively high if the main clause occurs in the first phrase; if it occurs in the second phrase the amplitude of both phrases will be relatively low. Also the Fmin (baseline) value is relative to the number of phrases in an utterance: in a one-phrase statement it is set at 75Hz for a male speaker and 145Hz for a female speakers; in a two-phrase statement the value increases about 15% for both speakers.

As far as amplitude and timing of accent commands are concerned the following dependencies are observed:

a) the accent command duration is determined by the duration of the corresponding accent group
b) accent groups of shorter durations have higher amplitudes than longer accent groups
c) the relation between duration of the accent group and its position within a phrase is not straightforward as "an effect of phrase-final lengthening is observed for several speakers" (Möbius 1995:6)
d) position of the accent group in the utterance and POS of the accented word determine the amplitude: "nouns require higher amplitudes than other classes" (Möbius 1995:6)

Another application of the Fujisaki model was presented in (Mixdorff 1998, 2001, 2002a).

The basic assumptions behind the model are listed below.

1. The model should be bi-directional (Mixdorff 2002b) i.e., it should provide a description of intonation on the phonological level which reflects the functional aspects of intonation and can be derived from higher-level representation (syntactic, semantic, contextual features) on the one hand and from a continuous f0 curve on the other. For the latter purpose the model should provide a phonetic description of intonation contours which accounts for melodic aspects of intonation in the first place and from which the abstract phonological description can be derived or mapped onto.

2. Linguistic and non-linguistic functions of intonation should be modeled separately (Mixdorff 1998:52; Mixdorff 2002b): the linguistic functions should be represented in terms of discrete phonological categories and the paralinguistic messages should be modeled by varying the acoustic parameters that realize the intonational events.

3. The linguistic functions of intonation can be modeled in terms of intonemes and boundary tones. On the phonological level tunes are regarded as sequences of intonemes and boundary tones. A single intoneme may spread over a longer part of the contour but it is realized by a discrete tone switch associated with an accented syllable.

Table 1 presents intoneme classes used in the model.

INFORMATION INTONEME (↓)	associated with declarative-final pitch accents, realized by a falling tone switch, conveys a message
CONTACT INTONEME (↑)	associated with question-final pitch accents, realized by a rising tone switch, establishes contact
NON-TERMINAL INTONEME (↑)	associated with non-final pitch accents, realized by a rising tone switch, signals non-finality
INFORMATION INTONEME (E↑)	variant of the information intoneme, signals emphasis

Table 1 (adopted from Mixdorff 2001:136): Intoneme classes, the arrows show the tone switch direction.

A boundary tone is defined as a section of an f0 contour including a phonologically distinctive tone at a phrase boundary. Intonemes and boundary tones can be considered as intonation description on the phonological level.

Tone switches and boundary tones are modeled respectively by accent and phrase components of the Fujisaki model. The accent component uses amplitude (which is a correlate of perceived prominence) and duration of the accent command and the phrase component uses amplitude and duration of the phrase command. A number of solutions were proposed to the problem of specification of the number of accent commands and values of the acoustic parameters (Mixdorff & Fujisaki 1997, Mixdorff & Fujisaki 1998, 2000). The Fujisaki-based parameterization provides a representation on the phonetic level.

As suprasegmental features of speech (intonation, rhythm) are strongly correlated with one another the model integrates also the module of syllable duration prediction (for details see: Mixdorff 2001, 2002a).

3.3.3. The Bell Labs intonation model

The Bell Labs intonation model (van Santen & Möbius 1997, van Santen et al. 1998, Möbius & van Santen 2000) presents superpositional approach to intonation modeling: intonational tunes consist of pitch curves, accent curves and perturbation curves layered on each other. The three components model intonational phenomena observable on the level of an intonational phrase, accent group and phoneme respectively. The model has been implemented into a multilingual TTS system developed in the AT&T labs for English, French, German, Italian, Spanish, Romanian, Russian and Japanese.

Accent curves consist of pitch peaks and pitch movements associated with accent groups i.e. sections of f0 contours which stretch from one accented syllable to the next one. Accent groups are divided into 1) onset and 2) rhyme of the accented syllable and 3) the reminder (unaccented syllables). Accent curves are generated by the linear alignment model responsible for the proper alignment of accent curves with accent groups. The first step consists in determination of f0 peak position using information related to segmental structure and "sub-duartions" of the accent group (i.e., durations of the specific components of the accent group) is

taken into account. For example, in sonorant final monosyllabic accent groups like the word *pin*, peaks occur later than in obstruent final monosyllabic accent groups e.g. *pit* (van Santen & Möbius 1997:322). Peak location is calculated from the equation below (van Santen et al. 1998).

$$T_{peak}(a) = \sum_j \alpha_{\mathbf{S},j} \times D_j(a) + \mu_{\mathbf{S}}.$$

Equation 4: The equation describing peak position.

As only as the peak has been located, a number of anchor points is determined so as to obtain good approximation of the accent curve. Their location is predicted relative to the pitch peak position and then they are aligned with the accent group. A complete accent curve results from linear interpolation between successive anchor points.

Phrase curves in the Bell Labs intonation model do not have any fixed gradients and therefore the model has more degree of freedom in comparison to the classical Fujisaki model and can generate a wide range of intonation contours. Phrase curves result from a non-linear interpolation between: "1) the start of the phrase, 2) the start of the last accent group in the phrase, 3) the end of the phrase" (van Santen & Möbius 1997:323). The phrase curve shape is determined by sentence mode and its position in the paragraph. Unlike the phrase component of the Fujisaki model, in the Bell Labs model the f0 course over the length of a phrase does not correspond to the declination line which is modeled separately.

The model accounts for microprosody and includes a special component i.e. perturbation curves to model microprosodic phenomena. These are short-term curves associated with those parts of the modeled f0 contour where segmental effects occur e.g. initial parts of sonorant following a transition from an obstruent (van Santen et al. 1998:296).

The f0 contour of an utterance results from "generalized addition of various classes of component curves" (van Santen & Möbius 1997:324) according to the equation below where: ***C*** corresponds to a set of curve classes {accent, phrase and perturbation curves}, ***c*** represents a particular curve class, ***k*** stands for an individual curve and ⊗ is an operator.

$$F_0(t) = \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t).$$

Equation 5: Equation describing the superposition of various classes of f0 curves.

The superpositional approach presented here is still being developed: the latest results are reported in (van Santen et al. 2005, Mishra, van Santen & Klabbers 2006).

3.4. Perceptual models

The basic assumption underlying perceptual models is that an f0 contour consists of a number of pitch movements which are irrelevant for the perception of intonation and consequently, an f0 contour can be reduced to relevant pitch movements in a stylization

procedure. It is also assumed that while interpreting messages conveyed by intonation human listeners do not have access to the acoustic form of the contour, but do it perceptually.

3.4.1. The IPO model of intonation

The basic assumption of the IPO model (t'Hart, Collier & Cohen 1990) is that an f0 contour of an utterance is a sequence of pitch movements. The listener is sensitive only to these pitch movements which are intended by the speaker, and are referred to as perceptually relevant pitch movements. An f0 contour can be reduced to these movements in a stylization procedure. Stylization is carried out by a human experimenter. It consists in replacing the original f0 contour with a minimum number of straight lines. As a result a representation is obtained in terms of perceptually relevant pitch movements in the logarithmic f0 domain. The resulting f0 contour, i.e. a close copy contour, is perceptually identical to the original contour.

The next step is the standardization procedure. It consists in determination of the common features of the close copy contours and representing them by a single pitch movement. As a result an inventory of discrete perceptually relevant pitch movements is established; this categorization is based on the melodic aspects of intonation and acoustic properties such as: direction, range and slope of the pitch movement and its alignment with respect to the onset of the vowel (pitch movements are not aligned with syllables and the model requires only that vowel onset position is provided). Each entry of the inventory is a distinctive category described by vectors of perceptually relevant features and can be interpreted on the phonetic level e.g. 1 (i.e. fast, early rise) stands for a rise of 50 semitones in 120 ms, the peak occurs 50 ms after the accented vowel. The number of categories in an inventory is restricted due to limitations of human perception of pitch.

As regards prosodic structure, the model assumes that an intonational phrase consists of three elements: prefix, root and suffix among which only the root is obligatory.

The IPO model is rule-based: it requires definition of an intonation grammar. The grammar determines which combinations of pitch movement categories into longer-range contours are possible and which categories are associated with specific elements of the intonational phrase structure.

The IPO approach matches features of various models Like phonological models, the it describes intonation in terms of a number of discrete categories i.e., perceptually relevant pitch movements. Like phonological models, the model is rule-based and sequential: f0 contour of an utterance is regarded as a sequence of discrete categories.

The IPO model is phonetic and superpositional in that it takes into account global (declination) and local factors (pitch movements associated with accented and phrase-final syllables) in shaping f0 contours. Like phonetic models it does not account for functional aspects of intonation: the description of intonation is in terms of categories distinguished on the basis of acoustic properties and melodic contrasts.

3.4.2. *Prosogram* and tonal perception model

Prosogram is a method of semi-automatic transcription of prosody based on the model of tonal perception (Mertens & d'Alessandro 1995, d'Alessandro, Mertens & Beaugendre 1994).

Intonation modeling in the framework of the tonal perception model involves the steps illustrated in Figure 15. The model consists of three basic modules.

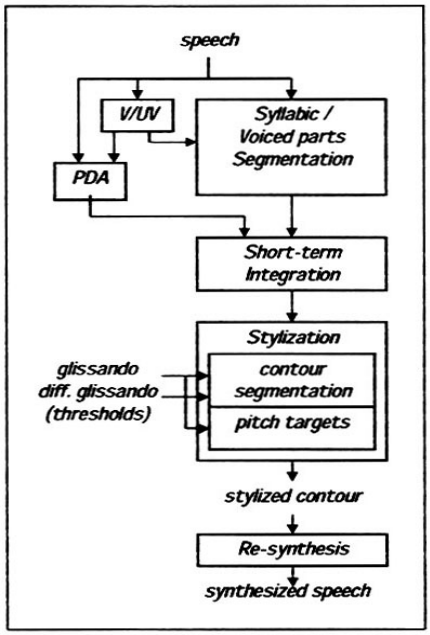


Figure 15 (adopted from Mertens & Alessandro 1995): Structure of the tonal perception model

1. *Syllabic/voiced part segmentation*: the speech signal is segmented into syllables and then syllables are segmented either relative to the spectral changes and syllabic nucleus, or relative to amplitude changes and loudness peak of a syllable.

2. *Short-term integration*: WTA (window time average) function models static tonal events. It generates smooth f0 contour on the basis of changes in the spectrum and amplitude of the speech signal.

3. *Stylization procedure* starts with segmentation of the f0 contour over the length of a syllable into tonal segments. Then G (glissando) and DG (differential glissando) parameters are applied to tonal segments and provide information which parts of the smoothed f0 contour are perceptually relevant.

The last step consists in assigning pitch target values to perceptually relevant parts of the contour. The stylized f0 contour results from interpolation between successive pitch targets.

The main assumption underlying *Prosogram* is that syllabic nuclei constitute basic units for pitch perception (Mertens 2004).

Tonal segment is considered to be the basic intonation unit and is defined as a simple rising or falling pitch movement (i.e. glissando), or level pitch. Tonal segments are associated with syllabic nuclei. *Prosogram* requires phonetic annotation of speech signal at the input (various types are allowed and it influences the stylization) so that the determination of vocalic nuclei location is easier. Sequences of tonal segments constitute compound pitch contours e.g. rise-fall or rise-fall-rise. The intonation stylization in *Prosogram* simulates pitch perception by humans by applying perceptual thresholds to f0 contours:

a) for the perception of a variation in pitch (i.e. glissando) a minimal f0 change has to occur: it is expressed by glissando threshold (G) parameter (set to 0.32 for vowels)
b) for the perception of a change in a pitch movement slope a minimal difference in the slope must occur: it is expressed by the differential glissando threshold (DG) parameter (set to 20 ST/s)

If the data is not provided at the input of the program, *Prosogram* extracts f0 of the speech signal, calculates intensity and creates a segmentation of the signal based on loudness. It

also produces a pitch tier with the values of the stylized f0 contour which makes resynthesis and perceptual evaluation of the stylization possible.

Figure 16 presents an example of *Prosogram* stylization of a Polish phrase "nie możemy też stwierdzić" (Engl. *we can not say*). As regards simulation of tonal perception by humans *Prosogram* is reported to produce very accurate stylizations (Mertens 2004) which can also be seen in the figure below.

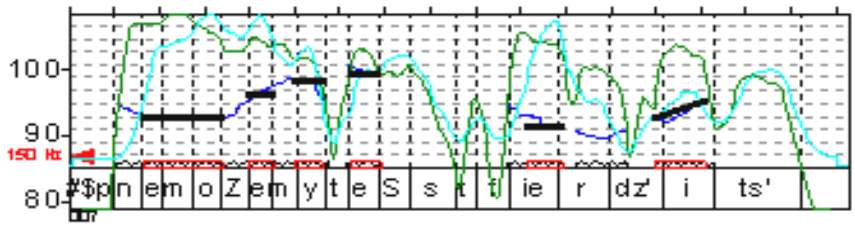


Figure 16 (author's example): *Prosogram* stylization of a Polish phrase: the thick solid lines mark boundaries of tonal segments. Vocalic nuclei boundaries are marked at the bottom of the stylization panel and above segment boundaries and SAMPA transcription of the phrase.

3.5. From phonetic to surface phonological description

As already explained in the sec. 1.1.2 intonation forms can be described at different levels of representation and analysis. Starting with the most abstract level and ending at the most concrete one these levels are: phonological, surface phonological, phonetic and physical (acoustic-physiological). On the phonological level intonation is described in terms of discrete distinctive categories, whereas phonetic descriptions use continuous parameters. The surface phonological level is regarded as intermediate between the phonological and phonetic levels. It is defined as "a level of distinctive discrete categories with which we can describe surface phenomena cross-linguistically" (Hirst, Di Cristo & Espesser 2000:3).

INTSINT intonation transcription system presented below is an example of a surface phonological description of intonation. The main assumption underlying it is that the form and function of intonation should be represented separately (Hirst 2004). The system uses an inventory of tones which describe the intonation contour on the surface phonological level and are interpreted on the phonetic level as pitch targets associated with syllables. The system is not strictly phonological for two main reasons. Firstly, it does not account for functional aspects of intonation and secondly, not all of the tones identified in the contour correspond to phonological tones. Their position in time and frequency domain is determined during stylization by means of the Momel algorithm.

Momel stylization (Hirst & Espesser 1993) involves four steps. The first one is f0 preprocessing which consists in setting to 0 of all f0 values which are higher than a ratio of 5% from the closest neighboring f0 values. This reduces the number of erroneous f0 values and at the same time the number of false f0 target values (i.e. target candidates) for application of the approximation function which is a quadratic spline function. Afterwards, the estimation of

target candidates takes place within an analysis window of 300ms. It involves neutralization of f0 values that might be faulty (i.e. if they fall outside the range determined by hz_{min}=50Hz and hz_{max}=500Hz, they are treated as missing), application of a modal quadratic regression to all non-neutralized f0 values and further neutralization if the distance between the original f0 value and the value determined in regression is more than 5%. The regression and neutralization are repeated until no more f0 values to be neutralized remain. Then, for each instant x the target point <t,h> is calculated which is the extreme of the corresponding parabola (see Hirst & Espesser 1993:79). If h falls outside the range (hz_{min}, hz_{max}) the target <*t,h*> is neutralized. As a result, for each f0 value in the contour one target point is determined. The final two stages are *partitioning* and *reduction* of target candidates which consist in division of the f0 contour into rising and falling components and finding (within an analysis window of 200ms) single targets which correspond to the inflection points between the components where the new function has its start.

Figure 17 contains *Praat* manipulation window showing the stylization results of a phrase from the Polish unit selection corpus: "w kilka tygodni później, film znalazł się na pierwszym miejscu najbardziej kasowych filmów wczechczasów" (A couple of weeks later the film has appeared on the top of the list of the best films ever). At the top of the figure the waveform is depicted, the darker regions mark the voiced regions in the speech signal. Below the waveform panel the original and stylized pitch contours are depicted: the former is marked with a dotted line and the latter - with a solid line. The points visible on the stylized contour mark the positions of target points found with the Momel algorithm. The stylization was carried out using the *Praat* implementation of Momel (Auran 2004).

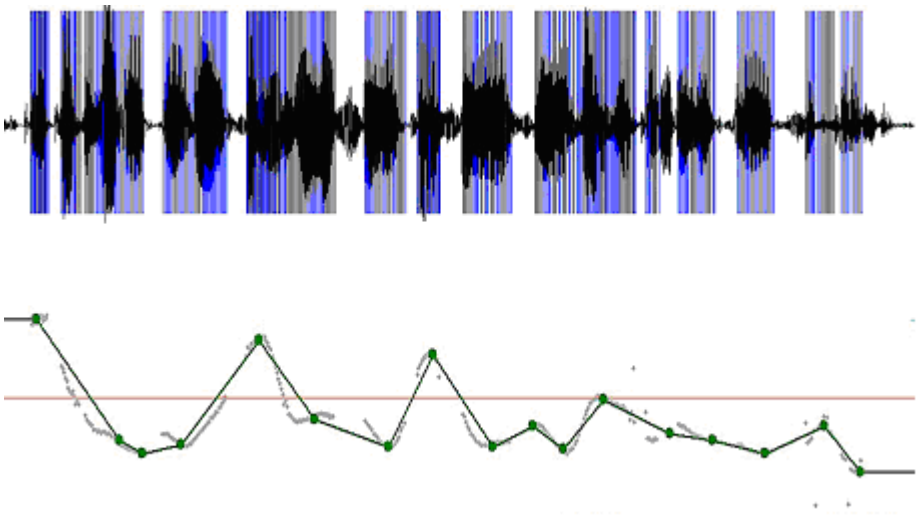


Figure 17 (author's example): Momel stylization of an f0 contour.
The original contour is marked with the dotted line, the stylized contour - with the solid line.

An example of application of the Momel algorithm to Polish is given in (Oliver 2005). Momel stylization was used as a preprocessing stage in intonation modeling in the Festival system (Oliver & Clark 2005).

As mentioned before in this section, the representation of pitch contours resulting from Momel stylization is used to derive a higher-level representation which is in terms of discrete distinctive tones - INTSINT.

Table 2 presents the tonal inventory used for transcription of the melodic aspects of intonation in the INTSINT system.

		Positive	***Neutral***	***Negative***
ABSOLUTE		T ⇑	M ⇒	B ⇓
RELATIVE	*Non-Iterative*	H ↑	S →	L ↓
	Iterative	U <	•	D >

Table 2 (adopted from Hirst, Di Cristo & Espesser 2000:12): Orthographic and iconic symbols for the INTSINT transcription system.

The inventory of tone labels consists of absolute and relative tones. In the latter group distinction is drawn between iterative and non-iterative tones:

a) the scaling of absolute tones: T (top), M (mid), B (bottom) is determined with respect to speaker's overall pitch range (the normalizing approach, see sec. 2.2.7)
b) the scaling of relative and non-iterative tones: H (higher), S (same), L (lower) is determined with respect to the position of the preceding tone on the frequency scale (the initializing approach, see sec. 2.2.7)
c) relative and iterative tones: U (upstepped), D (downstepped) differ from the non-iterative tones in the range of the f0 interval which is bigger in case of the non-iterative tones. Their scaling is determined also with respect to the level of the preceding tonal target.

INTSINT uses a set of rules to control mapping of tones onto tonal categories which constitute the description of intonation on the phonological level. These rules may be treated as a coding algorithm for automatic transcription of intonation (for details see: Hirst, Di Cristo & Espesser 2000:12). INTSINT includes also additional symbols to mark intonation unit boundaries, which makes it possible to control pitch range e.g. [and] stands for the beginning and end of an intonation unit respectively. There is also possibility of marking the overall range and register of the intonation unit by placing symbols in round brackets at the beginning of an intonation unit e.g. (⇑), (>). Marking of boundaries and pitch range of an intonation unit facilitates mapping of pitch targets onto tone labels.

Figure 18 contains *Praat* manipulation window showing the results of the Momel stylization and coding of the pitch targets into INTSINT description. The phrase was: "w kilka tygodni później, film znalazł się na pierwszym miejscu najbardziej kasowych filmów wczechczasów" (A couple of weeks later the film has appeared on the top of the list of the best films ever). At the top of the figure the waveform is depicted, below it the pitch contour is shown. The top transcription panel contains values of pitch targets determined with the Momel algorithm, the panel below shows coding of the Momel description into INTSINT description and the bottom panel contains values of f0 targets estimated from rules on the basis of the INTSINT description of the contour. The estimation of pitch targets and tones was performed using the *Praat* implementation of Momel and INTSINT (Auran 2004).

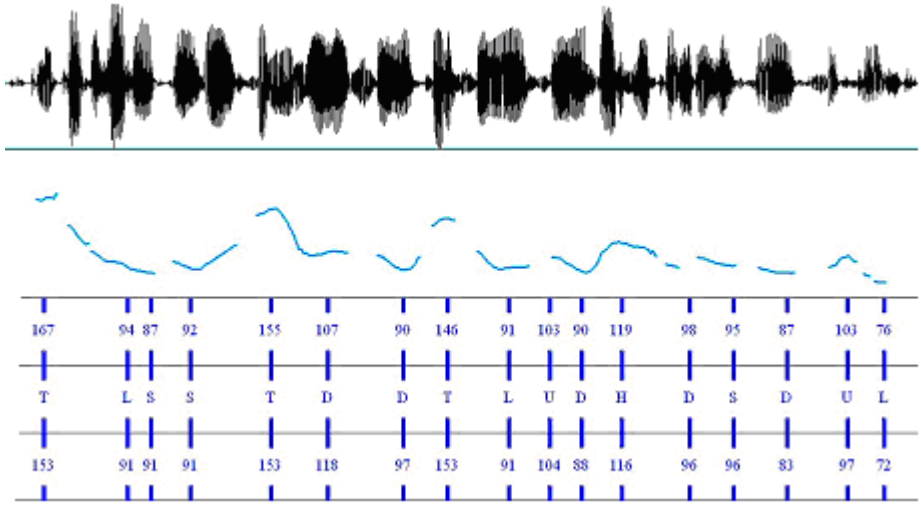


Figure 18 (author's example): INTSINT transcription of a Polish phrase.
In the top panel: waveform, below - the pitch contour. The three transcription panels contain (from the top to the bottom): f0 values of pitch targets determined by Momel algorithm, INTSINT coding of the targets, f0 targets generated from rules on the basis of the INTSINT transcription.

3.6. Phonological models

This section gives an overview of phonological models of intonation which follow the approach to intonation analysis and representation defined in the framework of autosegmental-metrical theory. Therefore, in the first place the fundamental concepts of the autosegmental-metrical theory are presented. Then, features of phonological models which have been applied in speech technology are discussed.

3.6.1. Autosegmental-metrical (AM) theory

The *autosegmental-metrical (AM) theory* presents a phonological approach to analysis and representation of intonation and evolved with the work of Lieberman (Lieberman 1975), Bruce (Bruce 1977) and Pierrehumbert (Pierrehumbert 1980). The four basic assumptions underlying the AM theory are (Ladd 1996:42): linearity of tonal structure, distinction between pitch accent and stress, analysis of pitch accents in terms of level tones and local sources for global trends. These assumptions are discussed in detail below.

1. Linearity of tonal structure.

Tunes are regarded as sequences of distinctive discrete events associated with elements of the segmental string. These events are pitch accents and edge tones and they are associated with stressed and pre-boundary syllables respectively. Stretches of the contour between the events are called transitions. Transitions do not contribute to conveying intonational meaning.

2. Distinction between pitch accent and stress.

In the AM theory stress is regarded as "acoustic salience: (…) a complex of properties that can be related to greater force of articulation, including increased intensity and duration, and shallower spectral tilt" (Ladd 1996:58). Pitch accents are regarded as phonological elements associated with metrically strong stressed syllables. The AM approach suggests that pitch accents are prominence-cueing (Ladd 1996:52) rather than prominence-lending i.e., a pitch accent manifest presence of a prominent, metrically strong stressed syllable rather than "constitute the prominent syllables prominence" (Ladd 1996:59).

3. Analysis of pitch accents in terms of level tones.

Pitch accents and edge tones are represented by single abstract phonological level tones: H and L and their combination. Pitch accents are defined as "a local feature of a pitch contour – usually but not invariably a pitch change, and often involving a local maximum or minimum" (Ladd 1996:45). This definition suggests that on the one hand pitch accents are realized by *pitch movements* or *configurations* and on the other hand – as sequences of *tonal targets* or *levels*. In the survey of intonation the *configuration based* approach has a long tradition (e.g. Crystal 1969, Halliday 1970, O'Connor & Arnold 1973, Cruttenden 1997, t'Hart, Collier & Cohen 1990, Kohler 1987, 2005). However, there is a rich experimental evidence supporting the *level based* approach.

First of all, it was shown in a number of studies that tones are aligned with specific segmental landmarks such as a particular syllable or segment boundary (e.g. Silverman & Pierrehumbert 1990, Prieto, van Santen & Hirschberg 1995, Arvaniti, Ladd & Mennen 1998, Ladd et al. 1999, Atterer & Ladd 2004). Secondly, tones exhibit independent scaling in the frequency domain too (e.g. Lieberman & Pierrehumbert 1984, Ladd 1998, Ladd, Mennen & Schepmann 2000).

Another experimental result supporting the level based approach is that what plays a role in the perception of a pitch accent is not size of a pitch movement but reaching a specific f0 target (e.g. Gussenhoven & Rietveld 1988). Experiments on the phonetic realization of Greek prenuclear LH pitch accents showed that both L and H tones "are aligned with reference to the accented syllable, and there is no evidence that one of them aligns with reference to the other" (Ladd et al.1998:11). If size and duration were the invariant and primary features of pitch accents speakers would control them instead of the alignment of f0 maxima (H) and minima (L) with respect to segmental anchors.

It is claimed (Ladd 1996) that with the description based on a single contrast between H and L tones all distinct intonation patterns can be modeled efficiently. On the phonological level only linguistic intonational meaning is accounted for, in order to be able to describe all perceptually distinct tunes (also those which convey paralinguistic messages) one needs to specify mapping rules which determine the scaling and alignment of tones on the f0 and time scale. In this way it is possible to tell linguistic and paralinguistic intonational meaning apart and account for them on the phonological and phonetic levels respectively.

4. Phonological interpretation of global f0 trends.

Global f0 trends include on the one hand lowering of a pitch range which results in an overall downward trend of a pitch contour and on the other - rising of a pitch range which results in an overall upward trend (upstep). The former effect can be modeled either globally – as *declination*, as it is done in the IPO model (t'Hart, Collier & Cohen 1990) and superpositional models (Fujisaki 1983, Fujisaki & Hirose 1982, Mixdorff 2000, 2001, 2002a, Möbius et al. 1993, Möbius 1995, van Santen & Möbius 1997), or locally as *downstep* – an effect that occurs at specific points in the utterance (Ladd 1996:75). The downstep interpretation was adopted in phonological descriptions of intonation (e.g. Pierrehumbert 1980, Grice & Savino 1995, Arvaniti & Baltazani 2000, Grice, Baumann & Benzmueller 2005), and there are two main arguments supporting this approach.

First of all, there is an experimental evidence (see Ladd 1996:76ff.) showing that speakers control the relationship between the levels of tonal targets: "the value of each accent peak in a downstep series is a constant proportion (…) of the previous peak" and thus "target values can be generated from left to right, with only a small 'window' looking back to a previous value" (op.cit.77). It means that global trends of the pitch contour can be modeled locally.

Apart from that, various authors reported on the problems encountered in modeling intonation with the superpositional models. In (Demenko 1999) difficulties were encountered in modeling yes/no question contours in Polish with the Fujisaki model due the fixed gradient of the phrase component. The phrase component of the Fujisaki model appeared also to be too constrained to model an f0 contour which rises slowly and steadily after the first pitch accent (Taylor 2000). This kind of difficulty could be overcome "by stacking several phrase components on top of each other in short intervals" (op.cit.:1711). Such problems do not arise if global f0 trends are interpreted as local effects.

The fundamental concepts of the AM theory discussed here were adopted in a number of phonological models of intonation presented in the following subsections.

3.6.2. Pierrehumbert (1980)

The model of American English intonation presented in (Pierrehumbert 1980) was developed in the framework of the autosegmental-metrical theory and significantly contributed to popularization of the phonological approach to intonation description and analysis.

All intonational tunes naturally occurring in American English are represented in the model as sequences of tonal categories depicted in Table 3. Tonal categories must always occur in the order indicated by arrows. The categories include: edge tones (initial boundary - optional and final boundary - obligatory), pitch accents (single H and L tones, and their combinations) and phrase accents (single H and L tones). * indicates the alignment of the tone with the accented syllable, % indicates alignment with the phrase boundary, whereas - is used to mark the alignment of the tone with word edge. + is used to mark linking of tones in bitonal pitch accents.

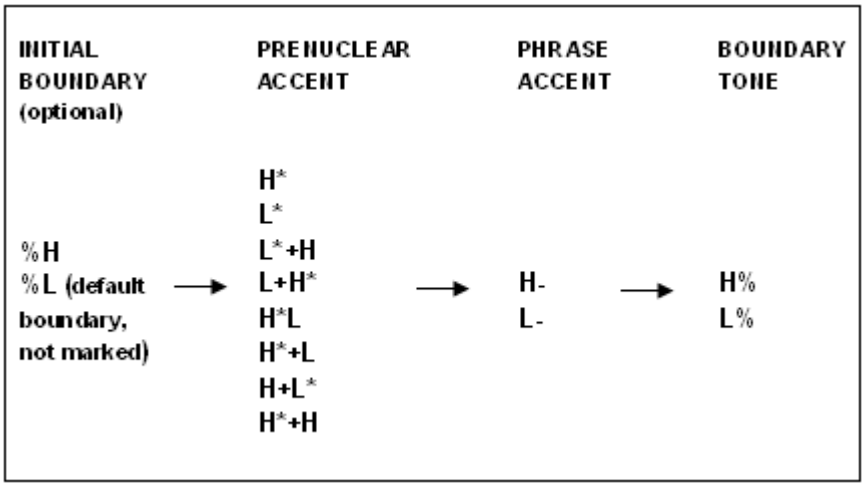


Table 3: (on the basis of Pierrehumbert 1980): Inventory of tonal categories in the Pierrehumbert model.

Below basic assumptions underlying the model are presented.

1. Linearity of tonal structure.

Intonational tunes are regarded as sequences of tonal categories - pitch accents and edge tones (phrase accents and boundary tones). Pitch accents can be either monotonal (a single H or L tone), or bitonal (a combination of H and L tones). No distinction is drawn between prenuclear and nuclear pitch accents; nucleus is simply considered as the last pitch accent in the phrase thus, it is assumed that accents do not have any particular structural roles in the tunes. Edge tones are always monotonal. Stretches of unaccented syllables between pitch accents are called transitions. Transitions do not contribute to conveying intonational meaning and are modeled by straight lines between pitch targets realizing tonal categories.

2. Scaling of tones, downstep and upstep.

L and H tones are identified as targets or turning points in the f0 contour – an f0 minimum and an f0 peak. In general, L and H are scaled at the bottom and top of speaker's range respectively. Yet, as tones never occur in isolation their scaling in reality is not that straightforward. When tones are combined as it is in bitonal pitch accents and sequences of a phrase accent + boundary tone, or occur in sequences of tonal categories in intonation contours, their scaling is affected by effects such as *downstep* (local lowering of pitch range) and *upstep* (local rising of pitch range). Both effects are regarded as phenomena of phonetic realization. But instead of being treated independently from the tonal structure (e.g. by labeling the affected tone with a special diacritic), they are considered to be triggered by specific sequences of tones. And thus each low tone of a bitonal pitch accent lowers the level of the subsequent pitch accent containing high tone (→downstep) and each high tone of a bitonal pitch accent raises the level of the following low tone (→upstep). High phrase accents trigger upstep on the subsequent boundary tones. In a sequence of downstepped/upstepped pitch accents, each subsequent tone is proportionally lower/higher than the preceding one.

For example, the L tone of the H*+L pitch accent is never realized as a low target, actually it does not represent a tonal target at all and has no phonological status. It functions only as a *downstep trigger* on the subsequent high tone. Therefore H*+L never represents a fall

from a peak or other high level to a local minimum and instead, the common falling contour is represented by a monotonal H* accent in the prenuclear position, and by sequence of a H* pitch accent followed by L-(L%) edge tone(s) in the nuclear position.

Similarly, the L tone in H+L* is never a local minimum, but rather a downstepped high tone i.e., it is scaled lower than the preceding high tone, but still high in speaker's range, which is the consequence of the upstepping effect of the preceding high tone.

The notation H-H% might suggest sustention of the H tone of the phrase accent until the phrase end. But as a matter of fact, the H- phrase accent triggers upstep on the subsequent H% boundary tone and the sequence H-H% is interpreted as a rise from a high pitch level to even higher pitch at the phrase end.

Much in the same way, the L% boundary tone does not necessarily have to signal a low tone at the intonational phrase end. When preceded by high phrase accent which triggers upstep on the following tonal target, L is scaled at the same level (e.g. in a *calling contour*).

3. Distinction between association and alignment.

Association refers to the relation between constituents of prosodic structure (syllables, phrases) and intonational features (pitch accents, edge tones). Pitch accents are associated with stressed syllables, phrase accents are associated with postnuclear syllables in between the last pitch accent and boundary tone. Boundary tones are associated with the end of intonational phrases. The fact that a given pitch accent is associated with a particular syllable does not necessarily mean that the whole pitch accent must be realized on that syllable. It is rather so, that pitch accents are realized partly on the accented syllable and partly on the previous or next unstressed syllable. In the notation of pitch accents the * symbol is used to indicate which tone occurs on the accented syllable. This property is called *alignment*.

4. Phrasing levels.

In (Pierrehumbert 1980) only one level of phrasing was proposed: *intonational phrase*. Its initial boundary can be optionally marked with a %H tone to indicate high pitch at the phrase onset. Final boundaries are obligatory. They are always preceded by phrase accents. In the modified version of the model (Beckman & Pierrehumbert 1986) a new prosodic constituent intermediate between intonational phrase and prosodic word was introduced - an *intermediate phrase*. Each intermediate phrase has to belong to some intonational phrase and has to include at least one pitch accent. Its final boundaries are marked with phrase tones. The introduction of the new prosodic constituent justified to some extent the use of phrase accents whose phonetic interpretation and phonological independence in certain cases was questionable (Ladd 1996:92).

3.6.3. American English ToBI

The model proposed in (Pierrehumbert 1980) became a basis for the development of ToBI system – *Tones and Break Indices* (Silvermann et al. 1992, Beckman & Hirschberg 1994, Beckman & Ayers 1997). Initially, ToBI was defined for the purpose of labeling of American English prosody in speech corpora. Since then it has been adapted to many other languages including Southern British RP and Australian varieties of English, German, Greek, Japanese and Korean.

The notion *tones* refers to description of intonational tunes in terms of pitch accents, phrase accents and boundary tones, whereas the notion indices refers to description of prosodic structure and accounts for grouping of smaller prosodic constituents into bigger ones. Basic assumptions underlying the transcription of prosody in the ToBI framework are given below.

1. Tones

ToBI follows most of the assumptions of the original model by Pierrehumbert. A significant difference between the two systems concerns *downstep*. In ToBI downstepped pitch accents receive the diacritic !. The definition of downstep remains the same but flagging of pitch accents affected by downstep makes the labeling more transparent. H*+L is no longer used as a downstep trigger – it was merged with the H* pitch accent which represents a fall from the peak on the accented syllable. H+*L pitch accent was replaced by H+!H*, which is a more transparent notation and reflects the fact, that the starred tone is not a local minimum but a high tone scaled lower in comparison to the high leading tone of the accent.

2. Break indices

Break indices reflect the strength of the boundary of a given prosodic constituent. They represent "a rating for the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance" (Beckman & Ayers 1997:31). Thus, each word boundary marker is accompanied by a proper break index.

- 0 describes the boundary between words which form *clitic groups*
- 1 marks the boundary between prosodic words
- 2 is used to describe the boundary between two words which has some properties of a phrase boundary but does not constitute a phrase boundary
- 3 indicates intermediate phrase boundary
- 4 is used for labeling of intonational phrase boundaries

Figure 19 illustrates ToBI transcription of two different tunes: a) continuation and b) statement. The top panel shows spectrogram, below it the pitch contour is depicted. The two bottom panels include ToBI annotation and word segmentation. The examples come from ToBI training materials available at: ftp://ftp.ling.ohio-state.edu/pub/TOBI/ame_wav_files/

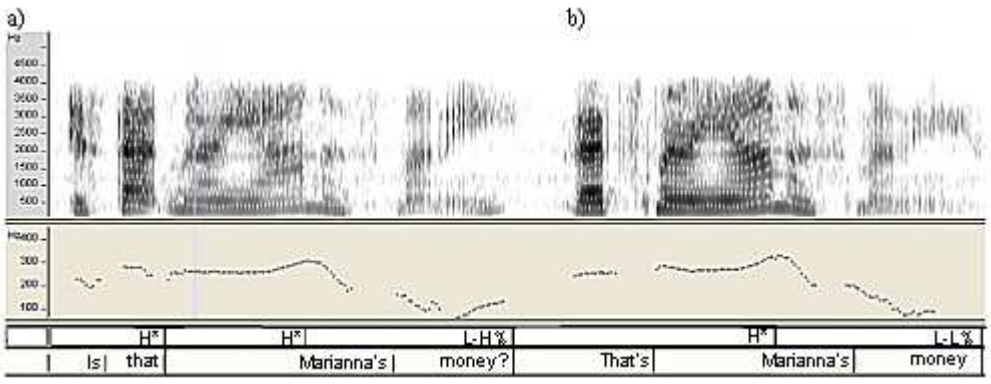


Figure 19: Examples of prosody transcription in the ToBI system.

The ToBI system was used for prosody transcription in a number of speech corpora designed for various speech applications e.g. the Boston Radio News Corpus (Ostendorf, Price & Shattuck-Hufnagel 1995), the Colorado University corpus (Pellom, Ward & Pradhan 2000) or Boston Direction Corpus (Nakatani, Hirschberg & Grosz 1995).

The information provided by ToBI transcription can be effectively used for estimation of pitch targets for the purpose of intonation generation in speech synthesis (e.g. Black & Hunt 1996, Syrdal et al. 1998). However, for the purpose of generation of f0 contours on the basis of ToBI phonological description rules have to be developed which specify scaling of tonal targets and their alignment with respect to segmental landmarks. An example of a rule-based system for generation of f0 contours from ToBI labels is given in (Jilka 1996, Jilka, Möhler & Dogil 1999). Table 4 presents mapping rules which determine the scaling of the L leading tone of a bitonal L+H* pitch accent.

position in pitch range	20% (not as low as a starred low tone)
position in voiced part of syllable	0.2 s before H* (reference point) ; 90% of voiced region to left of reference point if voiceless at that point; 20% of voiced region to right of reference point if no voicing before that point not farther to the left than target belonging to preceding tone label

Table 4: (adopted from Jilka, Möhler & Dogil 1999): Rules for scaling of L tone in a L+H* accent.

3.6.4. AM models of German intonation

The models of German intonation presented in this section are based on the assumptions of the AM theory. Yet, in some respects the models differ from each other. In this section, some of the most significant differences are discussed.

In some descriptions (Grice, Baumann, Benzmueller 2005) pitch accents can be both left- and right-headed, which means that pitch accents can account for the pitch preceding or following the accented syllable. In other models (Uhmann 1991, Féry 1993, Grabe 1998, Mayer 1995) pitch accents are always left-headed, i.e. there are no pitch accents consisting of a sequence of a leading tone + target tone (e.g. L*+H, late peak). This is the consequence of the definition of pitch accent domain. Following the traditional British school (Halliday 1967, O'Connor & Arnold 1973) the latter models assume that foot constitutes the domain of pitch accent. Foot is defined as a sequence of syllables starting on the accented syllable and extending over the subsequent unaccented syllables up to, but excluding the next accent. Thus, if there is a distinctive pitch variation on the pre-accented syllable(s) it is represented as a trailing tone of the previous pitch accent. It was proven that in German the preceding context is important for

the interpretation of intonational meaning. In (Kohler 1987, 2005) a categorical contrast was observed between pitch accents of a different peak position relative to the accented syllable. If the f0 maximum occurs on the previous unstressed syllable (early peak) the tune conveys the meaning of giveness or established fact; if the peak occurs around the middle of the accented syllable (medial peak) it signals a new fact. Whenever the peak occurs towards the end of the accented syllable or on the following unstressed syllable (so called late peak) it gives rise to the perception of an emphasis on the new fact. These findings suggest that the domain of pitch accents should be extended to the pre-accented syllable as it is in GToBI and the original English ToBI model, otherwise some important contrasts will be unaccounted for in the description of intonation. But of course, in some languages there may be no contrast among early, medial and late peaks and thus no need to account for leading tones (e.g. Gussenhoven 2005, 2006).

The inventories of tonal categories may differ among the models with respect to the number and type of elements. In (Uhmann 1991) a distinction is drawn between prenuclear pitch accents which can be either monotonal or bitonal and always bitonal nuclear pitch accents. The GToBI system (Grice, Baumann, Benzmueller 2005) accounts for both monotonal and bitonal pitch accents but does not make the distinction prenuclear/nuclear. The description proposed in (Féry, 1993) and The Stuttgart System (Mayer 1995) apart from nuclear pitch consisting of 2 or 3 tones use also monotonal prenuclear pitch accents which may occur as a result of *partial* or *total linking*.

The differences in the phonological organization of tones between the models are related to the issues of *tone spreading* and *linking*, and pitch range modification, i.e. *upstep* and *downstep*.

As regards *tone spreading* in (Féry, 1993) tone level at an intonational phrase boundary is specified by the level of the trailing tone of the preceding nuclear accent. The trailing tone *spreads* over the unaccented syllables between the nuclear pitch accent and phrase boundary (hence the term tonal spreading). Other models e.g. GToBI describe pitch level at phrase boundaries by means of phrase accents and boundary tones which constitute tonal categories independent from pitch accents. Phrase accents are associated either with a phrase final syllable or with a postnuclear stressed syllable and spread until the boundary tone.

As regards *tone linking* in (Uhmann 1991) monotonal prenuclear accents are considered as a result of complete linking which deletes the trailing tone of the bitonal prenuclear pitch accent. In the model by Féry and in The Stuttgart System it is assumed that prenuclear pitch accents can split off their trail tone which is then either associated with the syllable preceding the next accented syllable (->partial linking) or even completely omitted (->total linking). Figure 20 illustrates pitch accents which underwent complete (a) and partial linking (b). They can be referred to as allotones of the underlying H*L prenuclear pitch accent, because linking does not affect the interpretation of the message conveyed by intonation. In the top panel in the figures the waveform is displayed, below it the orthographic and ToBI transcription of the utterances is given; the dotted line below the text is the pitch contour. Vertical lines indicate word boundaries.

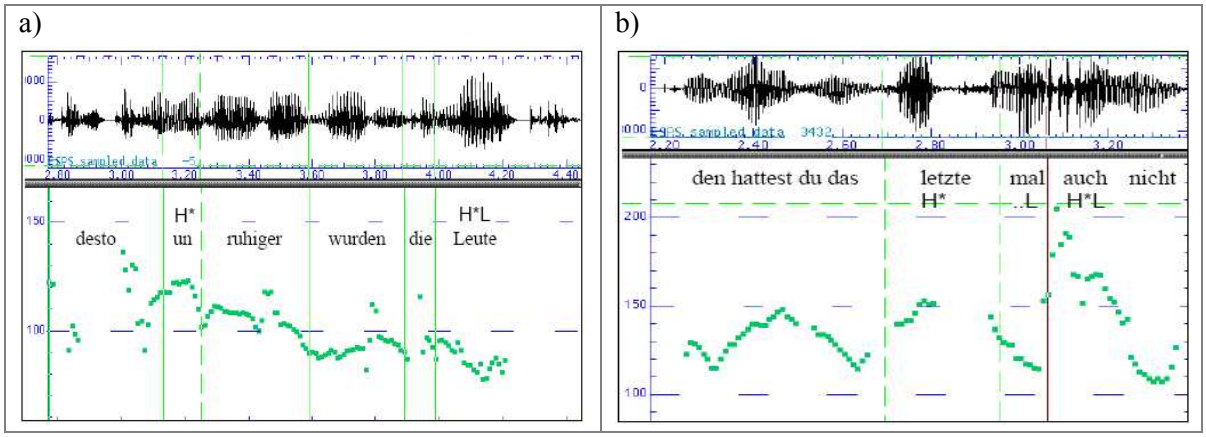


Figure 20 (adopted from Mayer 1995): Examples of complete (a) and partial linking (b) of the low trailing tone of the underlying H*L prenuclear accent

As regards *downstep/upstep* in the GToBI system downstep can affect each H tone (both in a pitch accent and phrase accent) and is marked with a ! diacritic. A sequence of downstepped tones ends at a phrase boundary or before the nuclear pitch accent, which in both cases involves pitch range reset. The GToBI system makes also use of a ^ symbol to indicate a local rise in the pitch range. The upstep effect may occur for example in the emphatic speech. The Stuttgart System also flags downstepped H tones with the ! symbol, but only pitch accents can be affected by downstep.

The German AM models disused here define either one or two levels of phrasing. In (Uhmann 1991, Grabe 1998) there is only one level – an intonational phrase, whereas in (Féry 1993), GToBI and The Stuttgart System a smaller prosodic constituent was introduced, namely an intermediate phrase. Each intonational phrase has to include at least one intermediate phrase, and each intermediate phrase has to include at least one pitch accent.

As a result of differences in the number of phrasing levels between the models there are differences in the usage of edge tones. The GToBI system provides phrase accents and boundary tones to mark weaker (intermediate phrase) and stronger (intonational phrase) boundaries respectively. A phrase accent may be associated either with a phrase final syllable or with a postnuclear stressed syllable. In the GToBI model all final boundaries (i.e., intermediate and intonational phrase boundaries) are obligatory with one exception: if there is no significant variation in pitch between the phrase accent and the subsequent boundary tone, the former does not have to be marked and its presence is signaled only by a hyphen. GToBI and The Stuttgart System provide an optional initial boundary tone to signal a high intonational phrase onset.

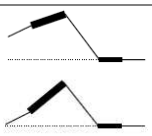
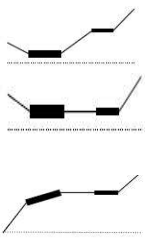
In (Uhmann 1991) the inventory of boundary tones consists of optional initial and obligatory final boundaries. In (Féry 1993) only final boundaries of intonational phrases are labeled and this is optional; the boundaries of intermediate phrases have no tonal specification. The Stuttgart System (based on Féry's model) marks intermediate phrase boundaries only with a hyphen, and only if there is no subsequent boundary tone. Final boundary tones are obligatory but apart from high and low a boundary tone there is a default boundary *%* to mark intonational phrase boundaries which have the same level as the trailing tone of the preceding pitch accent. In (Grabe 1998) they are labeled with a *0%* symbol.

3.6.5. GToBI

From among the German intonation models developed within the framework of autosegmental-metrical theory the GToBI model deserves a comment, because it is an example of a surface phonological description.

GToBI (*German Tones and Break Indices*) was developed mainly for the purpose of labeling of prosodic structure and intonational tunes of German. The main goal was to provide a description that is "able to capture distinctions drawn in the traditional auditory-based literature on German intonation (…) as well as in later autosegmental-metrical studies" (Baumann, Grice & Benzmueller 2000:1). As a result some GToBI patterns instead of describing strictly linguistic contrasts reflect paralinguistic modification, therefore "part of the description is redundant, in that some of these patterns are derivative or make reference to gradient features rather than categorical ones" (Grice, Baumann & Benzmueller 2005:30). For this reason, GToBI should be regarded as a surface phonological rather than strictly phonological description of intonation. Unlike in the original ToBI, in GToBI the scaling of H and L tones is more transparent and straightforward as a result of a different treatment of downstep and upstep. In ToBI both these effects are triggered by specific tone sequences, whereas GToBI uses diacritics ! and ^ to mark downstepped and upstepped tones respectively.

A summary of the discussion on German AM intonation models is given in Figure 21. It depicts the types of nuclear contours distinguished by the models in the current and previous sections, provides tonal transcription and schematic visualizations of different tunes: extra heavy lines indicate accented syllables, heavy lines indicate postnuclear syllables and the dotted line below it stands for the speaker's baseline. Additionally, contexts in which the tunes typically occur are given. It can be seen that the same nuclear contour has different tonal representation depending on the model.

TYPE	UHMANN	FERY/STUTT. SYSTEM	GTOBI	CONTOUR	MEANING
FALL	H*+L L%	H*+L	H* L-% L+H* L-%		neutral statement / neutral wh-question contrastive assertion
LATE PEAK	L*+H L%	L*+H+L	L*+H L-%		assertion
RISE	L*+H H%	L*+H	L*(+H) H-^H% L* L-H% (L+)H*H-^H%		neutral yes/no or echo question indignation/answering phone follow-up question
LEVEL		L*+H	(L+)H* H-(%)		incompleteness / ritual expression

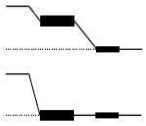
FALL RISE	H*+L H%	H*+L H%	(L+)H* L-H%		polite offer
EARLY PEAK		H+H*+L	H+!H* L-% H+L* L-%		established fact soothing or polite request
STYLIZED CONTOUR		H*+M	(L+)H* !H-%		calling contour

Figure 21 (adopted from Baumann, Grice & Benzmueller 2000:3): Comparison of the common nuclear contours in German intonation models.

3.6.6. Autosegmental models of Dutch

Gussenhoven's autosegmental model of Dutch (Gussenhoven 1983, 1984) had a significant influence on the theory of German intonation, examples are (Féry 1993) and (Grabe 1998) discussed in sec. 3.6.4. On the basis of Gussenhoven model a system for transcription of Dutch intonation *ToDI* was developed (Gussenhoven 2005). Both ToDI and the underlying phonological model are based on the assumptions of the AM theory, but in many respect they different from other AM-based systems, especially English ToBI and Pierrehumbert model.

The most important difference between Gussenhoven's model and ToDI on the one hand and other AM models on the other hand is that the former are not strictly phonological systems, because they were established on the basis of perceptual distinctions solely (drawn in the IPO system) with no reference to the functional aspects of intonation. In the rest of this section the assumptions underlying the ToDI system are described.

1. First of all, it is assumed that *feet are always left-headed* and thus, there are no pitch accents consisting of a leading tone followed by a target tone (e.g. L+H*). This is the consequence of the definition of foot as the domain of the pitch accent: a foot starts at the accented syllable, stretches over subsequent unaccented syllables and ends at the onset of the next accented syllable. The other reason and probably more important why bitonal accents are always starting with the target tone is that in Dutch there is no linguistic contrast between pitch accents of a different peak position.

2. Some *intonational phrases may contain no pitch accent.* This usually regards utterance final phrases which "often express some reformulation of a previous IP, or contain the reporting clause after a direct quotation" (Gussenhoven 2005:10). The pitch level at the beginning of an accentless IP is specified by the level of the trailing tone of the nuclear accent of the previous IP and the pitch level at the end of an accentless IP – by other tones occurring after the trailing tone of the nuclear accent of the previous IP.

3. There is only one level of phrasing - an intonational phrase. Its initial boundaries (%L, %HL and %H) are obligatory, whereas final boundaries (L%, H% and %) are optional. In contrast with the ToBI system where a sequence of a phrase accent and a boundary tone such as L-H% marks mid-level pitch at the phrase boundary, in ToDI the L and H boundary tones identify

individual tonal targets and are scaled in the bottom quarter and in the top three quarters of speaker's range respectively (Gussenhoven 2005:13). If the level of the trailing tone of the nuclear accent is sustained to the end of the phrase, the % boundary tone is marked which is interpreted as *a half-completed rise* (e.g. after LH*) or *a half-completed fall* (e.g. after L*HL nuclear accent) (Gussenhoven 2005:12).

4. In general, no + sign is used between the starred tone and the trailing tone. Such notation signals that a starred and a trailing tone of a bitonal pitch accent do not occur closely in time. This is due *tone linking* which causes that trailing tones may spread over many unaccented syllables until next pitch accent or phrase boundary. The + symbol between the starred tone and trailing tone is used to indicate closer temporal realization of the two tones. Figure 22 illustrates two phrases of the same nuclear contour but containing contrastive prenuclear pitch accents: H*+L (a) and H*L (b). It can be seen in a) that the two tones of the prenuclear H*+L accent are aligned close in time which results in a steep fall from the high pitch target on the accented syllable. Figure 5b) shows the effect of tone linking: the trailing tone of the H*L prenuclear accent spreads until the H*L nuclear pitch accent.

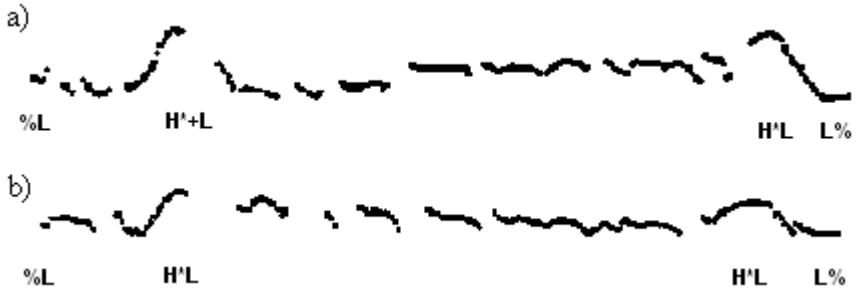


Figure 22 (adopted from Gussenhoven 2005:18): Examples of two different realizations of a prenuclear H*(+)L pitch accent: without (a) and with (b) tonal spreading of the trailing L tone.

More examples of tone linking are illustrated in the Figure 23. It can be seen that tone linking makes it possible to account for the pitch preceding the accented syllable without the use of a leading tone.

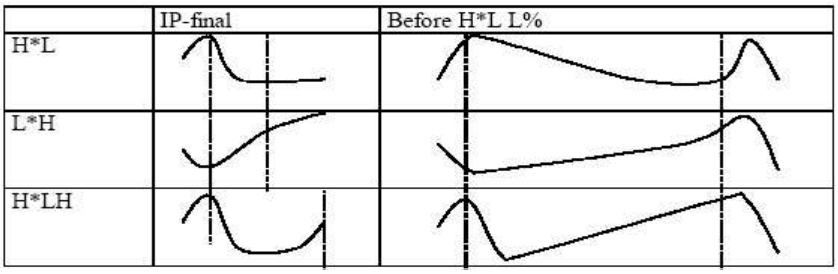


Figure 23 (adopted from Gussenhoven 2005:11): The effect of partial linking of the trailing tones of bitonal pitch accents (first column) in the nuclear position (second column) and prenuclear position (third column)

5. Unlike in other AM-based models where H and L tones of monotonal H* and L* pitch accents correspond to targets from which some pitch movement starts in ToDI system they describe a high and a low level pitch which spreads until subsequent tonal target.

6. Downstep is treated as an optional modification of H tones. It refers to a stepwise lowering of the level of H tones in a sequence of H pitch accents. It was shown in (van den Berg et al. 1992) that in a sequence of downstepped H tones the value of each subsequent peak is a constant proportion of the previous peak. Downstep is marked by a ! diacritic.

To sum up, it is assumed that all intonational tunes occurring in Dutch can be described in ToDI as sequences of the tonal categories listed in Table 5. Tonal categories must always occur in the order indicated by the arrows.

INITIAL BOUNDARY		PRENUCLEAR ACCENT		NUCLEAR ACCENT		FINAL BOUNDARY
		(!)H*		H*		
%H	→	(!)H*L	→	(!)H*L	→	H%
%HL		H*+L		L*(H)		L%
%L		L*(H)		L*(!)HL		%
		L*!HL		H*!H		

Table 5: (on the basis of Gussenhoven 2005): Inventory of tonal elements in ToDI

3.7. Intonation modeling and analysis in Polish

This section is dedicated to presentation of the approaches proposed to Polish intonation modeling. Apart from that an overview of some of the most important studies on Polish intonation is given. These studies include previous works on Polish intonation which deal primarily with description of intonation and seek for acoustic correlates of accents in Polish as well as recent studies whose results are effectively used or are applicable in speech technology systems.

3.7.1. Speech technology-oriented analysis of Polish intonation

A comprehensive analysis of Polish intonation oriented towards application in speech technology systems is presented in (Demenko 1999). Among others, the analyses investigate distinctive features of accent in Polish, acoustic correlates of phrase boundaries and realization of basic intonation patterns. The results are used in automatic detection of accents and phrase boundaries and types of prenuclear/nuclear accents by means of statistical modeling methods including neural networks and discriminant function analysis. The performance of the models designed in (Demenko 1999) is reported in sec. 6.1.2. Apart from coding of f0 contours, the

issue of intonation generation is addressed as well. Two rule-based approaches are proposed. The first one uses the Fujisaki model (Fujisaki 1983) in which intonational tunes result from superposition of two components of a different temporal domain: phrase and accent component (see sec. 3.3.1). For the purpose of specification of the values of amplitudes and timing of accent/phrase commands and damping factors of the accent/phrase control mechanism a number of acoustic and statistical analyses was carried out. In accordance with the model of phrase structure proposed in (Jassem 1984) which distinguishes between anacrusis, pre-head (pre-ictic intonation), head (ictic intonation) and tail (post-ictic intonation), the phrase component was divided into three parts. The first one included unaccented syllables and the first accented syllable, the second one included subsequent accented syllables and the last one started with the nuclear accented and finished at the phrase boundary. The results were not very satisfying: only two out of nine nuclear accents (HL and ML, see Table 6) were properly modeled. Another difficulty was found in modeling the temporal alignment of f0 peak relative to the start/middle/end of the accented vowel. These results indicate low flexibility of the superpositional model in generating f0 contours in Polish. In view of the drawbacks of the superpositional approach a sequential model was defined in the framework of the phrase structure model described above. The author reports on a greater flexibility of the sequential model in generation of f0 contours.

In both models a description of intonation consisting of an inventory of nine nuclear and two prenuclear accent types was used. The prenuclear accents are monotonal (H vs. L), whereas the nuclear accents are bitonal and they are described as a combination of two tones: L, M and H. The exception is the accent labeled as xL which is realized as a movement from a low pitch target to even a lower target and the LHL accent which is realized as a rising-falling pitch movement.

The types of accents are distinguished on the basis of perceptually significant properties of pitch movements including: direction, range, position relative to speaker's voice range and timing relative to the onset of the accented vowel. the results of a perception study showed that tunes described by different types of nuclear accents can be assigned different meanings. They are illustrated in Table 6.

NUCLEAR MELODY	MEANING	NUCLEAR MELODY	MEANING
ML	statement without any emotional content	**MH**	continuation, politeness
HL	surprise	**MM**	boredom
LM	flattery	**xL**	dislike
HM	persuasion, calming sb down	**LHL**	effusiveness
		LH	question, surprise

Table 6 (on the basis of Demenko 1999:77): Interpretation of the meaning of tunes described by nuclear accents.

The conclusion that can be drawn from the results reported in (Demenko 1999) for the current study is that a sequence-based approach is more useful for modeling intonation contours

in Polish than a superpositional approach. Apart from that the results of automatic detection of accent and phrase boundary position as well as classification of accents reported in sec. 3.3.1 show that these tasks can be performed with a high accuracy by neural network models on the basis of vectors of acoustic features.

3.7.2. Testing the PaIntE model for Polish

The PaIntE model usefulness for modeling Polish intonation was evaluated objectively in a resynthesis of the intonation contours from the PaIntE parameters. The results were reported in the master thesis of the author (Hałupka 2004) and here they will be only briefly discussed.

The experimental material consisted of 200 short phrases (including at most 3 minor phrases) read by a professional male speaker. The phrases presented basic mode types, i.e. declarative and interrogative (both why- and yes/no questions). As explained in the previous section for the purpose of the PaIntE parameterization the location of accented and phrase-boundary syllables have to be provided. The most important part of labeling consisted in determining pitch accent and boundary tone types. For this task two intonation annotation systems were used: American English ToBI (Beckman & Ayers 1997) and the system presented in (Demenko 1999, see also the section 3.7.1) and further referred to as PToBI.

In the PaIntE model the analysis of f0 contours is carried out in two steps. In the first place, the extraction of the PaIntE parameters is carried out within a three syllable window around the accented/phrase boundary syllable. The second step consists in reconstruction of the contour from the parameters. The parameterization of intonational events is based on especially designed approximation function (see 3.2.2). Every parameterized intonation event is the sum of the rising and falling curve. The f0 contour results from linear interpolation between successive curves. An example of a Polish intonation contour resynthesized from the PaIntE parameters is depicted in Figure 24.

In order to investigate the accuracy of PaIntE stylization correlation and RMSE between the original and stylized contours was measured. The results are summarized in Table 7. The high value of the correlation coefficient and low value of RMSE indicate that the stylization based on the PaIntE parameterization produces f0 contours which are very similar to the original ones. Unfortunately, no subjective evaluation of the stylization accuracy was carried out, so it is hard to which extent the original and stylized contours are perceptually equivalent.

Annotation scheme	Mean correlation	Mean RMSE
ToBI	0.837	14, 21 Hz
PToBI	0.88	12, 45 Hz

Table 7: Quantitative results of the PaIntE stylization.

In general, the application of the PaIntE model to generation of intonation contours in Polish proved that the PaIntE approximation function is flexible and universal. The results of

the parameterization were used further in the analyses on acoustic realization of specific accent types in Polish (Wagner 2005). They showed that the PaIntE parameters provide a very useful representation of intonation on the phonetic level and can distinguish between different types of pitch accents. However, it seems that good resynthesis result can not be regarded as the only adequate measure of model's usefulness for application in speech synthesis, because the main difficulty in generating intonation in synthesis consists in estimation of the model's parameters from symbolic input describing utterance's features. The discussion in section 7.1 gives more insight into these issues and explains why another approach to contour generation is proposed in this thesis.

Figure 24 shows the editing window of *ESPS XWaves*; from top to bottom: waveform (phoneme boundaries are marked with vertical lines), the original f0 contour, the resynthesized contour, transcription and ToBI annotation. The utterance consisted of two minor intonational phrases: "są wszyscy, ale pójdą" (everybody is here, but they will go). More examples of the original and resynthesized contours are given in (Hałupka 2004).

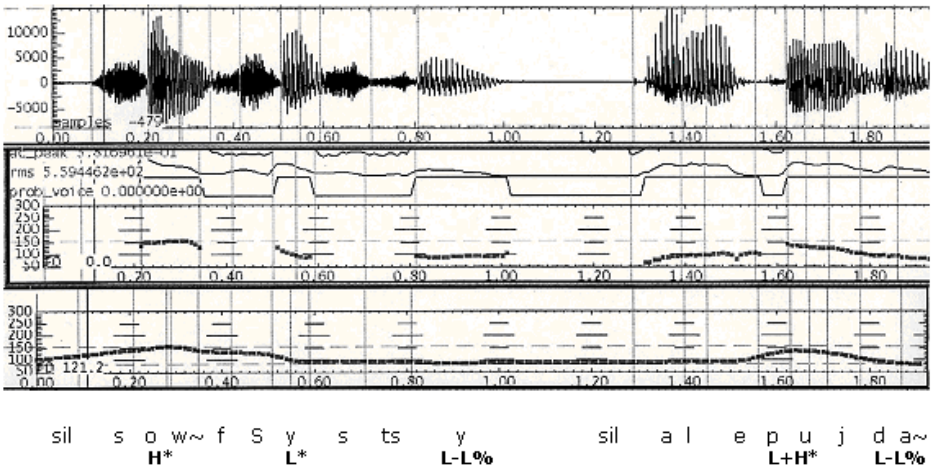


Figure 24: Example of the PaIntE resynthesis of a Polish utterance.

3.7.3. Polish intonation modeling in Festival TTS system

The intonation model applied in the Polish module of Festival TTS system is described in (Oliver & Clark 2005). As regards description of intonation it was analyzed on the basis of analysis of intonation patterns found in the PoInt corpus (Karpiński & Kleśta 2000) designed for the purpose of Polish intonation analysis and including a diverse speech material (fragments of read literary texts, quasi-spontaneous monologues, map task-based dialogues). For the purpose of analysis the contours were stylized using the Momel algorithm (Hirst & Espesser 1993). The final inventory consists of three accent types and results from clustering of pitch contours described in (Oliver 2005): rising-falling (LHL), rising (R) and falling (F) accents. No solution to deriving this description from f0 contour is proposed, but instead a method capable of accent location and type prediction from text is designed. It uses CART trees and among prediction

features there are (Oliver & Clark 2005): position of syllable in phrase/word, strength of break, stress, number of (stressed, unstressed, accented) syllables since/till last phrase break, number of syllables since/till last accented syllable, number of minor phrase breaks since last major phrase break, a simplified part of speech (content/function word). The features which contributed to the prediction accuracy the most include: syllable distance to/from the phrase break, syllable position in word, strength of break after the next syllable, stress on the syllable and distance to phrase break measured in the number of accented syllables. A high overall accent prediction accuracy is obtained - 83.3%.

For the purpose of contour generation a linear regression model is built which estimates three targets per syllable: at the start, in the middle and at the end of the syllable. These targets are then smoothed and interpolated through to produce a continuous f0 curve. Among other features used for pitch target estimation there is the information on accent location and type given by the CART model. As regards the performance of the regression model on the training subset the overall correlation between the observed and estimated targets is 0.68 and RMSE=42; on the test subset r=0.71 and RMSE=41.89.

Figure 25 illustrates a contour generated with the methods proposed in the study. The top panel contains the waveform, the bottom panel shows the generated f0 contour. The utterance was: "Florentynka przenikliwie spojrzała w oczy Kłoski" (Florentynka piercingly looked Kłoska into the eyes).

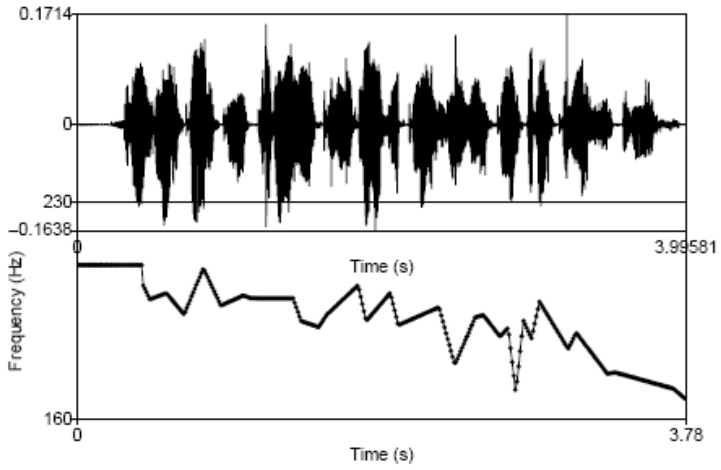


Figure 25 (adopted from Oliver 2005:5): Example of a pitch contour generated by interpolation between pitch targets estimated with a LR model.
The utterance was: "Florentynka przenikliwie spojrzała w oczy Kłoski"
(Florentynka piercingly looked Kłoska into the eyes).

The quality of intonation modeling was evaluated also in a perception study in which listeners compared pairs of stimuli consisting of a stimulus with a contour generated with the previous rule-based model used for Polish intonation generation in Festival (Oliver 1998) and the other stimulus with the contour generated by interpolation through targets estimated by the

LR model. In general, the results indicated that the listeners preferred the new intonation model to the previous rule-based one.

The conclusion which can be drawn on the basis of these results for the current study is that the approach adopted in (Oliver & Clark 2005) can be successfully applied to generation of intonation contour in Polish and for a variety of speech styles (which were represented in the PoInt database).

3.7.4. Pitch line

The method of stylization of intonation contours called *PitchLine* was presented for the first time in (Demenko & Wagner 2006). The main reason for developing our own stylization method rather then using an existing one (e.g. Momel or Prosogram) was the full control of the stylization process and parameterization of f0 contours.

1. The general purpose

F0 curves consist of two layered elements: "a microprosodic component caused by the nature of the individual phonematic segments of the utterance and a macroprosodic component reflecting the choice of intonation pattern for the utterance" (Hirst & Espesser 1993:76). Even though perturbations resulting from segmental effects on f0 contours may contribute to the perceived naturalness of the synthesized speech and thus, sometimes are accounted for in intonation modeling (Kohler 1995, van Santen & Möbius 1997), they do not contribute to conveying of intonational messages. For this reason, most often the first step in intonation modeling consists in contour stylization which aims at factoring of the raw f0 contour into the microprosodic and macroprosodic component; in the later stages of analysis only the latter component is taken into account.

On the phonetic level intonational features are described in terms of continuous parameters. The choice of parameters depends on the theoretical framework. It can be assumed that the parameters should refer to the inherent features of pitch accents and boundary tones e.g. duration and amplitude of the pitch movement (as in the Tilt or PaIntE model) or position of the f0 peak relative to some segmental landmark (as in phonological models).

2. The approximation function and parameterization scheme

PitchLine is based on the following theoretical framework. As in the Tilt intonation model it is assumed that intonational tunes can be regarded as *strings* of *events* (pitch accents, boundary tones) associated with the segmental structure of the utterance. The events are modeled as rising, falling or rising-falling pitch movements. They are delimited by target points in the contour (f0 minima and maxima) which define their start, peak and end; some of the targets are effectively corresponding to phonological tones (H, L). In the Tilt model the position of the target points is detected automatically. In PitchLine it is carried out manually.

The parts of f0 contour corresponding to the events are approximated with functions described by Equation 6.

$$y = 1 - 2^{\gamma-1}.x^{\gamma} \quad 0 < x < 0.5$$
$$y = 2^{\gamma-1}.(1-x)^{\gamma} \quad 0.5 < x < 1.0$$

Equation 6: The approximation functions used in the PitchLine stylization.

The stretches of contour between subsequent events are called *connections* and are approximated with straight lines.

In PitchLine the approximation is carried out semi-automatically: the choice of the approximation function i.e., R-rising, F-falling, or C-connection and the alignment of the function with the segmental string depend on the human labeler and are decided upon by clicking in the appropriate location on the approximation panel. Taking into account the findings concerning segmental anchoring of tonal targets (e.g. Ladd et al. 1999, Caspers & van Heuven 1993, Xu 1998, Arvaniti & Ladd 1995, Arvaniti, Ladd & Mennen 1998, Ladd, Mennen & Schepmann 2000) it is assumed that the start and end of the approximation functions have to be aligned with some segmental landmark located on the pre-accented, accented or post-accented syllable. During the approximation the normalized mean square error can be controlled: it is displayed on the approximation panel.

At the output the program[4] provides a file containing the values of the stylized f0 curve (which can be used for pitch resynthesis in *Praat*) and another file with parameters describing the events: slope (describing the steepness of the f0 curve), Fp (f0 value at the point of the alignment of the approximation function), amplitude of the pitch movement and shape coefficient of the curve.

The usefulness of the approach adopted in PitchLine was tested on a subset of the expressive speech corpus used in the current thesis and described in sec. 4.1.2. The speech material included recordings of a male (MW) and female (AW) speakers, altogether 1000 phrases. For each phrase phonetic transcription and segmentation was provided automatically using the program Creatseg (Szymański & Grocholewski 2005). For prosodic labeling a simplified annotation was adopted similar to that used in the Tilt model (Taylor 2000): accented syllables were marked with *a* and phrase boundaries with *b*, silences were marked with *sil*. Pitch was extracted every 10 milliseconds using the ESPS method available in Wavesurfer.

The stylization accuracy was evaluated objectively by measuring the NMSE value between original and stylized f0 contours and subjectively in a perception study. The average NMSE value for the two speakers is 0.003, which indicates that the proposed method provides an accurate approximation of f0 contours. As regards the perception test results the general impression of the listeners was that the phrases resynthesized with the stylized f0 contours sounded very natural (for details see Demenko & Wagner 2006, Wagner 2006). It means that PitchLine stylization is well capable of extracting of the macroprosodic component of f0 curves reflecting the choice of the intonation pattern for the utterance.

[4] The computer implementation of the PitchLine stylization method was done by J. Ogórkiewicz from the Laboratory of Language and Speech Technology.

Figure 26 illustrates fragment of the PitchLine editing window. The upper panel contains the waveform, the panel below shows the original f0 contour (dotted line), the stylized contour (solid line), approximation functions (R,L,C) used for stylization of the intonational events and NMSE value. The vertical lines show approximate phoneme boundaries, accented syllables are marked in grey. The bottom panel includes SAMPA transcription of the utterance: "czy sylwetki te naprawdę wyleciały przez okno" (whether the figures really flew out of the window).

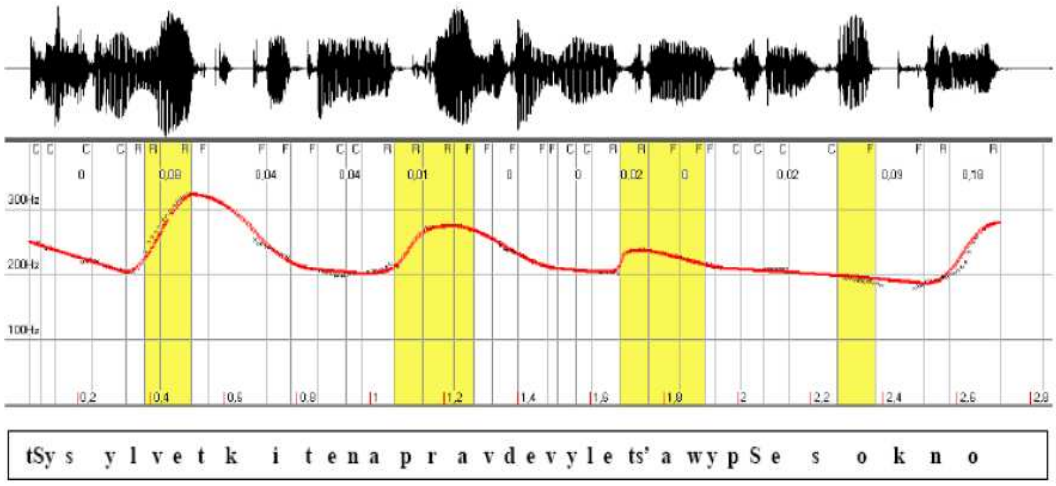


Figure 26: Example of the stylization of intonation contour in the PitchLine program. The utterance was: "czy sylwetki te naprawdę wyleciały przez okno" (whether the figures really flew out of the window).

3. Clustering of intonational events

On the basis of the acoustic parameters obtained in the stylization the classification of pitch accents and boundary tones was carried out on the basis of the female voice (AW) speech material (Wagner 2006). The k-means clustering algorithm available in Statistica was used for this task. The method requires that the user defines the number of clusters. Taking into account the classification of pitch accents given in (Demenko 1999) and on the basis of visual and auditory analyses of the speech data six clusters were defined for the classification of pitch accents. The resulting classification had the lowest variance among objects within the same group and the maximal variance among the groups. All the acoustic parameters significantly discriminate among the groups.

Figure 27 illustrates the prototypical pitch accents found in each cluster: accented syllable boundaries are marked in grey, the vertical lines show approximate phoneme boundaries. The solid line marks the stylized f0 contour, whereas the dotted line shows the original contour. Above the contours the approximation functions are marked (F, R, C) and the value of the NMSE is given.

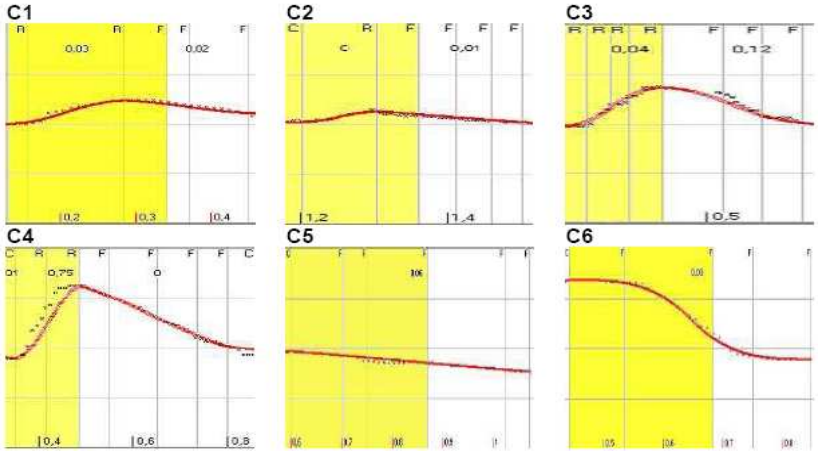


Figure 27: Pitch accent classes obtained in k-means clustering.
Accented syllable boundaries are marked in grey, female speaker.

As it can be seen in the classification 4 types of rising-falling accents and 2 types of falling accents were distinguished. The rising-falling accents differ with respect to:

a) amplitude of the rising and falling pitch movement
b) steepness of the rising/falling slope of the movement
c) alignment of the f0 peak relative to accented syllable onset
d) level of f0 peak

As regards the falling pitch accents they differ with respect to the following features:

a) the target level from which the fall starts and thus, the amount of the fall
b) the bend of the f0 curve
c) the steepness of the falling slope

Pitch accents grouped in the cluster 1 (C1, 176 instances) are characterized by a gentle rise followed by a gentle fall starting on the accented syllable and a significant range of the f0 change. They usually occur at the beginning or in the middle of continuation phrases (56%) and statements (ca. 30%). Pitch accents grouped in the second cluster (C2, 202 instances) can be found in the middle or towards the end of continuation phrases (50%), statements (35%) and exclamations (10%). The slopes of the rising and falling f0 curves are very gentle as the range of the f0 change is small. Cluster 3 (C3) also groups rising-falling pitch accents (110 instances). They are realized by a very steep rising f0 curve followed by a steep fall starting at the onset of the post-accented syllable. These pitch accents occur almost always phrase-initially and in continuation phrases (55%), statements (20%) and exclamations (25%). 27 rising-falling pitch accents have been classified in the Cluster 4 (C4). They have very steep f0 slope and are followed by very steep fall on the post-accented syllable; there is also a very big range of the f0 change. 37% of these pitch accents occur at the beginning of exclamations. Pitch accents grouped in the cluster 5 (C5, 183 instances) occur at the end of statements (48%) and are realized by gently falling f0 curve on the accented syllables and followed by a very gently falling boundary tone. Smaller percentage of these pitch accents occurs at the end of continuation phrases (35%). The cluster 6 (C6) has 51 members; these are sharply falling pitch accents of a final position in exclamations and continuation phrases.

As regards the classification of boundary tones four clusters have been found, they are illustrated in Figure 28.

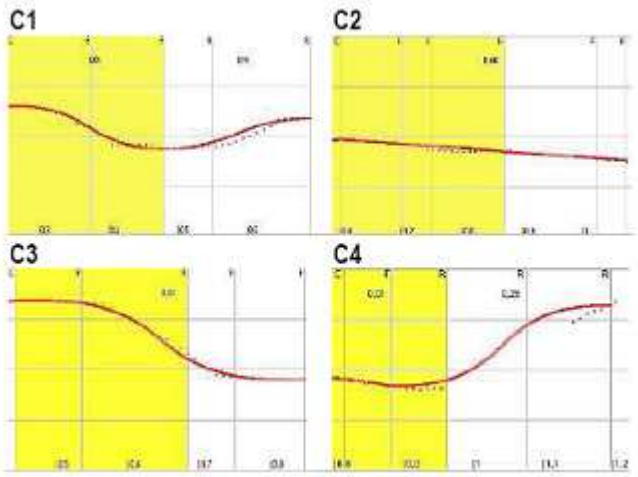


Figure 28: Boundary tone classes obtained in k-means clustering. Accented syllable boundaries are marked in grey, female speaker.

The first distinction that can be drawn between various boundary tones regards the direction of the final pitch movement. the members of the cluster C1 and C4 are examples of rising boundaries, whereas the end tones grouped in the clusters C2 and C3 are examples of falling boundaries. Since the clustering was based on the parameters derived for the accented and post-accented syllables the resulting classification illustrates not only types of boundary tones, but also examples of different nuclear tones. This explains why the falling end tones belong to two different clusters: in C2 they are part of a nuclear tone that could be described by the sequence L*L-L%, whereas in C3 - H*L-L%.

The difference between the two rising end tones is in the level of the f0 target at the boundary: in C1 it has lower level than in C4 (or in other words: the two boundaries have different range of the final pitch movement).

The 95 members of the first cluster (C1) are characteristic for continuation phrases. The second cluster (C2) counts 106 members which usually occur at the end of statements (45%) and sometimes exclamations (21%). The boundary tones grouped in the third cluster (C3, 37 instances) are characteristic for exclamations (54%) and statements (30%). In all the questions (and less often continuation phrases) the end tone of the cluster 4 (C4) occurred.

4. Discussion and conclusions

The classification of intonational events on the basis of the parameterization obtained in the approximation of the f0 contours provide a description of intonational events which is in accordance to that presented in (Demenko 1999).

The approximation function used in the PitchLine method produces smooth f0 contours free from microprosodic effects. Both the objective and subjective evaluation of the stylization results i.e., the low NMSE value (average=0,003) and high ratings of the perceived naturalness of the stylized f0 contours are promising.

There are two main reasons why the stylization and parameterization provided by PitchLine were not used in the analyses carried out in the scope of this thesis. First of all, the stylization is semi-automatic: for each unit (phoneme or syllable) marked in the transcription and segmentation panel a specific approximation function has to be selected manually. It is a very time-consuming and laborious task and it seemed impossible to adopt this method to a large speech corpus, like the one used in the current thesis. Secondly, it is still uncertain what kind of parameters could be the most useful from the point of view of intonation modeling. As mentioned before, the parameters should reflect inherent acoustic properties of pitch accents and boundary tones: this can be investigated by looking for acoustic correlates of accents and phrase boundaries which can be regarded as such inherent properties. These issues are addressed in the Chapter 6.

3.7.5. Previous studies

In (Jassem 1961) a number of analyses was carried out in order to determine acoustic correlates of accent in Polish. The analyses investigated the effect of accentual prominence on variation in fundamental frequency, intensity, duration, and spectral characteristics of vowels. The results prove that *variation in pitch* is the main correlate of accent, whereas intensity and duration can be regarded as secondary features distinguishing between accented and unaccented vowels. No significant effect of accentual prominence on spectral characteristics was found. On the basis of these results the author concludes that accent in Polish can be considered as melodic instead of dynamic as assumed by other authors (e.g. Dłuska 1947).

Perceptually-motivated description of Polish intonation in terms of intonemes was presented in (Steffen-Batóg 1973, 1996). Basic assumptions concerning Polish intonation system presented in the latter work are the following:

1. The hierarchy of prosodic domains consists of intonational phrase, intoneme, tonal unit and intonation unit/segment.

2. Intonation units are considered as distinctive pitch movements associated with syllables. Their inventory consists of one level, two monotonal and four bitonal units. The monotonal units are distinguished on the basis of the pitch movement direction and bitonal units differ with respect to both direction and range of the pitch movement. In the monotonal class there are level, fall and rise intonation units; the bitonal units include: weak fall–rise, strong fall–rise, weak rise–fall and strong rise– fall, where weak and strong refer to the range of the movement.

3. Tonal unit is defined as a class of homotonal accent groups, i.e. AGs of the same intonational-accentual structure which is determined by the position of the accented syllable and the type and position of intonation units within the accent group. Accent group is defined as a sequence of an accented word followed by unstressed word.

4. Intonational tunes are considered as strings of intonemes and the latter are regarded as classes of functionally equivalent tonal units. The functional equivalence is described as the possibility of a substitution of one tonal unit by another in some position in an intonational tune

of an utterance which results in a new tune that belongs to the intonation system of the language. Altogether there are 26 classes of intonemes distinguished on the basis of the following perceptually significant features:

a) the type of tonal units associated with the accented syllable and the following unaccented syllables,
b) the level of pitch on the accented syllable relative to the level of pitch on the following unaccented syllables,
c) the level of pitch on the previous unaccented syllable relative to pitch level on the accented syllable.

Apart from that a distinction is drawn between neutral vs. other intonemes on the basis of the position of accented syllable in the accent group: neutral intonemes start with the accented syllable.

The intoneme system provides a framework to describe Polish intonation. It was used for the purpose of the analysis of the relation the form of the tune and the meaning it conveys (Steffen-Batóg 1996:). The results show that for the interpretation of emotional content of the utterance and its modality the knowledge of the situational context is very important. One of the conclusions that can be drawn from this result is that rather than investigating the relation between the form and meaning the tune conveys one should investigate the contribution of intonation to signaling different situational contexts. Apart from the author points out that the nuclear part of the contour is the most important for the interpretation of the meaning conveyed by intonation.

In the end, it should be mentioned that in the previous works on Polish intonation the issue of segmental effects on pitch variation was also investigated (e.g. Steffen-Batóg 1973, Matuszkina 1976).

3.7.6. Recent studies

In (Karpiński 2006) the structure and intonation of map task dialogues was investigated. In a number of analyses the categorization of dialogue moves was established which was used in the description of dialogue structure and definition of a dialogue model. Apart from that, intonational tunes and nuclear melody associated with selected dialogue moves was analyzed as well as intonation of sequences of phrases linked by various discourse relations. Intonational features were analyzed in terms of *intonograms* created by means of a *Praat* script and/or tonal units resulting from perceptual stylization of intonation contours with *Prosogram*. The results can be effectively used in various speech applications, especially in dialogue systems.

A recent approach to pitch stylization and automatic labeling of intonation is presented in (Wypych 2005, 2006). In the first place, f0 extraction based on comb-filtering takes place. On the basis of information on syllable boundaries (which are aligned automatically) representation of the pitch contour over the length of a syllable is provided in terms of two stylization segments: one associated with the beginning -to-middle part of syllable and the other associated with the middle-to-end part of the syllable. Each segment is described with a vector of acoustic

features such as pitch height, glissando and duration. After this stage, pitch tracking algorithm is used which works on a window containing at most 7 stylization segments and is expected to reduce the number of errors in f0 extraction. The f0 curve obtained in this way serves as a basis for automatic recognition of intonation according to the systems proposed in (Jassem, in press). After identification of the intonational phrase structure, top-down processing of the stylized pitch contour is carried out which uses the knowledge provided by the automatic (HMM-based) intonation recognizer to reduce the contextual variability observed in the contour?

So far, no evaluation of the proposed methods has been reported.

3.8. Discussion

In this chapter basic assumptions underlying different types of intonation models were discussed and an overview of the study of Polish intonation was given. It could be observed that in comparison to other languages (e.g. English or German) for which various intonation models within different theoretical frameworks were proposed there are few analyses of Polish intonation whose results can be effectively used in speech applications. For this reason, the research carried out in the scope of this thesis (and presented in the next chapters) can be regarded as a contribution to the state-of-the art knowledge on Polish intonation and development of methods and solutions which are useful for various speech applications.

The chapter started with an overview of typologies according to which intonation models can be classified. Generally, a distinction is drawn between four types of models: phonetic vs. phonological, sequential vs. superpositional, generative vs. analytical and data-driven vs. rule-driven. The first distinction can be regarded as primary, because it is based on the levels of representation and analysis for description of intonation, which is fundamental for the formulation of a model.

In phonetic models intonation is described on the phonetic level which is intermediate between acoustic/physical parameters and surface-phonology/phonology (see sec. 1.1.2). Intonational features are described in terms of continuous parameters and only melodic aspects of intonation are taken into account. Consequently, as Ladd points out: "'Acoustic correlates' have been sought for a variety of meaningful aspects of utterances including surface constituent structure, the discourse status of referring expressions and speaker emotion and attitude" with "no attempt to identify phonological categories" (Ladd 1996:20). On the contrary, on the phonological level functional aspects of intonation are described in terms of distinctive abstract categories. Sometimes a parallel is drawn between intonational phonology and segmental phonetics (op.cit.:35): tonal categories correspond in a way to phonemes, because like phonemes they also constitute *discrete units* related by *phonological opposition* or *contrast*. Thus, a substitution of one tonal category by another in a given tune should result in a new tune conveying a different intonational meaning.

An *intonational grammar* determines how tonal categories combine to produce units of a higher order (intonational phrases), which resembles the role that morphology and syntax play in segmental phonology.

Apart from the inventory of tonal categories and intonational grammar a phonological system requires also specification of *mapping rules* which determine how categories are realized on the phonetic level. In this way functional and melodic aspects of intonation are analyzed and

represented on different levels namely, the phonological and phonetic levels respectively. This is closely related to the most important assumption underlying phonological approach i.e., to tell apart linguistic and paralinguistic meanings conveyed by intonation: "The most basic task of any phonological analysis is to identify the categorical or quantal elements in a phonological system and to account for the ways in which the realization of these elements varies. If intonational and paralinguistic messages are indeed distinct, then one of the sources of variation in the realization of intonational categories is paralinguistic modification, and the basic task of analyzing intonational phonology is to tell intonation and paralanguage apart"(Ladd 1996:39). This feature of phonological intonation systems is considered as their advantage and at the same time it is pointed out that in phonetic models all intonational meaning is treated as paralinguistic (see discussion in Ladd 1996:33ff). Since on the acoustic/physical level both linguistic and paralinguistic intonational meanings are realized by means of properties such as loudness, voice quality and pitch range, it is difficult but not impossible to tell them apart if the abstract, cognitive, phonological level of description of intonation is accounted for. It should be noted that phonological descriptions of intonation are defined in analyses based on a very carefully controlled speech material (e.g. Arvaniti 2001, 2002, Arvaniti, Ladd & Mennen 1998). Therefore, it seems that in the framework of a strictly phonological systems such as ToBI both detection/classification of prosodic constituents as well as prediction of tonal targets for the purpose of contour generation may pose problems, because of multiple and diverse factors affecting the scaling and alignment of tonal targets. The drawbacks of phonological descriptions are mentioned by various authors and are summarized in the discussion below.

1. Models that are based on discrete representations are not able to represent the whole variety of naturally occurring intonational tunes, because the number of categories is limited (Taylor 2000). As a matter of fact, the phonological level in intonation modeling introduces a quantization error (Möbius et al. 2000, Batliner et al. 2000) as "the whole variety of f0 values available in acoustics is reduced to a mere binary opposition Low vs. High, and to some few additional, diacritic distinctions".

2. The distribution of tonal categories is very uneven. It is reported by various authors (e.g. Black & Hunt 1996, Clark 2003) that almost 60% of pitch accents labeled in the Boston University News corpus (Ostendorf, Price & Shattuck-Hufnagel 1995) fell into H* category, whereas the frequency of other accent types (!H*, L+H*, L+!H* and others) did not exceed 15%. The unequal distribution of tonal categories was also the reason for merging accent and boundary types in the studies dealing with design of automatic methods for prosody labeling based on ToBI-transcribed speech corpora and described in the sec. 6.1.2. Since tunes which convey distinct intonational meanings (linguistic, paralinguistic) are perceptually different, but phonological descriptions account for linguistic meanings solely, perceptually different contours can be grouped together on the phonological level (Taylor 2000, Mixdorff 2002b). This may provide an explanation for the fact that tonal categories have unequal distribution.

3. Some authors point out (Taylor 2000) that phonological descriptions pretend to be of a categorical nature but the boundaries between the categories are not strict and easily identifiable. This remark refers to the apparel drawn between segmental and intonational phonology (Ladd 1996) according to which tonal categories, like phonemes, are identified on

the basis of and are related by phonological opposition/contrast. Thus, substitution of one tonal category by another results in a new intonational tune and changes the message it conveys. But as a matter of fact, distinct messages are not always encoded in the tonal description of the tune, because two tunes of the same tonal representation may be perceptually different if they convey distinct paralinguistic meanings.

4. It is claimed in (Batliner et al. 2000) that in some speech applications like ASR or ASU better results can be obtained without referring to a phonological description: "one always gets better results if one can do without such an intermediate level i.e., if one can establish a direct link between syntactic/semantic function and phonetic form. After all, if such a mapping can be done automatically, we can map level A (phonetic form) onto level C (linguistic function) without an intermediate (phonological) level B; with such a level, we have to map A onto B, and B onto C. If this can be done automatically, we do not need B any longer" (op.cit.:2). However, the results of a perception study investigating the quality of intonation generated by various intonation models reported in (Syrdal et al. 1998, Möhler 2001) show that better results (in terms of perceived naturalness of synthetic speech or higher correlation between original and generated f0 contours) are obtained if more precise information regarding prosody is provided i.e., not only accent/boundary position but also their type. The results of sensitivity analysis showing the importance of factors used for estimation of pitch targets for generation of f0 contours presented in the current study (sec. 7.2.3) are in accordance with the results presented in (Syrdal et al. 1998).

In view of the drawbacks of strictly phonological models discussed here it is assumed within the framework of a comprehensive intonation model a surface phonological description should be used instead of a phonological one. The first reason is that perceptual differences between tunes can be the effect of distinct linguistic as well as paralinguistic messages conveyed by intonation. Since in speech synthesis both meanings are taken into account a surface phonological description of the G-ToBI type can be very useful, because it accounts for both melodic and functional aspects of intonation and thus, encodes both linguistic and paralinguistic intonational meaning.

Secondly, a surface phonological description of intonation can be to some extent language independent, because specific intonation functions are encoded in the same forms in different languages (cf. sec.2.2.5 & 2.4). Consequently, solutions and methods based on this kind of description can find application in multilingual speech technology systems.

The surface phonological description proposed in the current study is described in detail in the Chapter 5.

Chapter 4. Tools and resources

This chapter presents tools and resources used in the analyses presented in this thesis.

As explained in the sec. 1.2 the main goal of the research carried out in the scope of the thesis is development of a comprehensive intonation model. The main application domain of the model is speech synthesis, but other applications such as speech recognition are not excluded. Moreover, model should account for various speech styles. These two requirements determined the choice of speech material for analyses described in the following chapters. Two speech databases were used: a subset of the unit selection corpus applied in the Polish module of BOSS TTS system (Breuer et al. 2000) and an especially designed corpus of expressive speech. The structure and features of the speech corpora are described in sec. 4.1. In sec. 4.2 the segmental and prosodic annotation of the speech material is presented. In the next sections of this chapter contour preparation and data collection for analyses are described. In the end, the statistical analyses and modeling methods used in the current work are described.

4.1. Speech material

The speech material used in the analyses and experiments presented in this thesis consisted of two speech corpora representing two different speech styles. The subset of the unit selection corpus created for the Polish module of BOSS system contains speech which can be described as unemphatic and news-commentary style. The expressive speech corpus designed especially for the purposes of this thesis includes emphatic speech. The features of the two corpora are described in the next sections.

4.1.1. Polish unit selection corpus

The entire linguistic material was read by a single professional radio speaker (male) during several recording sessions supervised by an expert phonetician. The resulting database consists of 4 hours of speech and is still under development. Therefore, the methods of automatic detection of accentual prominence and prosodic breaks as well as identification of pitch accent and boundary tone types (see Chapter 6) will be very useful and will enable not only faster, but also more consistent prosodic annotation of the new speech material.

The first part of the speech corpus was recorded in 2005 and includes recordings of read dialogues, fragments of fiction and reportage, railway information and examples of expressive speech.

In the second part of the corpus (recorded in 2006) six datasets are distinguished. They include utterances devised by linguists or extracted from various text sources. The utterances contain:

a) the most frequent consonant structures (367)
b) all Polish diphones (114)
c) triphones in CVC syllables in voiced context and different intonation patterns (676)
d) triphones in CVC syllables in sonorant context and different intonation patterns (1923)
e) the most frequent words (6000)

In the analyses carried out in the scope of this thesis only a subset of the whole corpus was used. In the selection of speech material two criteria were used and they were related to prosodic features and labeling consistency. As regards prosodic features it was important that the speech material for analyses contained examples of a variety of tune types and different prosodic structures. As regards labeling consistency, 4 phoneticians were in charge of the segmental and prosodic annotation of the corpus. In sec. 6.1.3 the results of studies investigating inter-labeler consistency in prosody transcription are presented which indicate that some discrepancies in the annotation provided by different labelers can be observed. On the one hand no such study has been carried out on the Polish unit selection corpus. On the other it is unclear to which extent inconsistency in labeling could affect the design of methods of automatic prosody labeling. For these reasons, major part of the speech material for the current analyses was selected from the part of the corpus annotated by a single labeler (the author of this thesis).

From the second part of the unit selection corpus (recorded in 2006), the whole dataset e) which consists of utterances with the most frequent words was used. It contains the most natural speech from among all the datasets and examples of a variety of intonational patterns. These utterances represent various prosodic structures, because they contain not only isolated sentences but also examples of discourse. Apart from that, a couple of utterances from each of the other datasets was selected as well: they are characterized by a significantly less variable prosody and represent only basic intonation patterns. Moreover, as regards specific text material (e.g. sentences with triphones) during recordings more attention was paid to segmental than suprasegmental features, which resulted in a "flat" prosody.

In order to obtain a more representative speech material, from the first part (recorded in 2005) of the database a number of utterances was selected which are examples of realizations of various tune types and different prosodic structures, as well as various speech styles The structure of the subset of Polish unit selection corpus used in the analyses described in this thesis is summarized in Table 8. The parts d) and e) of the database were annotated by the author of this thesis.

dataset	fiction	railway info	emotional	a)	c)	d)	e)	total:
no of utterances	12	56	13	8	69	76	818	1052

Table 8: Structure of the database extracted from the unit selection corpus.

4.1.2. The expressive speech corpus

This corpus contains recordings of 4 speakers (AD, AW - female, MW, WI - male) reading a chapter of M. Bulhakov's novel "The Master and Margaret". This text was selected for recordings because of its features: it includes diverse examples of dialogues, monologues, discourses, different modes and expressivity. The speakers had time to prepare for the reading. They were instructed to read in a moderate tempo and to give the text an emotional, but natural interpretation. The latter requirement did not pose a very difficult task, as the text if full of humor, irony and grotesque, and all speakers admitted that Bulhakov's novel belongs to their favorite books. Of course, there are as many unemphatic as emphatic sentences, because not all the text could be given an emotional interpretation. From each speaker ca. 20 minutes of speech was obtained. The recordings were carried out in a professional recording studio of one of the Polish commercial radio stations.

From this database the material from three speakers was used in the analyses described in the current thesis: altogether 1758 phrases. The material from one of the male speakers (WI) was excluded, because it contained too many disfluencies which significantly affected the realization of suprasegmental features.

4.2. Annotation

The next two sections present the annotation of the speech corpora.

4.2.1. Phonetic transcription and segmentation

The computer coding conventions are based on SAMPA for Polish (Wells 1997) and IPA alphabet. Two sets of characters were defined for the exact grapheme to phoneme mapping for the Polish language – an input set of characters and an output phonetic/phonemic alphabet. The input set of symbols for Polish was defined as a set of the following symbols: X={a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, q, r, s, ś, t, u, v, w, x, y, z, ź, ż, #, ##}. One hash substitutes interword spaces in a string of orthographic symbols, and two hashes denote sentence final punctuation marks. An inventory of 39 phonemes was employed for broad transcription and a set of 87 allophones was established for the narrow transcription of Polish (Demenko Wypych & Baranowska 2003). The modifications introduced into the original Polish version of SAMPA are described below.

1. The palatal phonemes /c/ (as in "kiedy") and /J/ (as in "giełda") are necessary for describing acoustical differences between 1) velar /k/ (as in: "kat") and /c/; 2) velar /g/ (as in "gad") and /J/. /c/ is similar to English "k" in "ski": it is never aspirated and occurs only before /j/, /i/, /e/ and /o/./k/ like English "k" in "sky" is never aspirated. /J/ is similar to English "g" in "geese" and is the voiced counterpart to /c/: it occurs only before /j/, /i/, /e/ and /o/./g/ like English "g" in "go" is the voiced counterpart to /k/.

2. The rules for the transcription of Polish graphemes: *ą* and *ę* have not been precisely defined for a long time and the current transcription systems makes assumptions according to the synchronic or asynchronic pronunciation on the basis of theoretical considerations. For *ę* possible transcriptions are: /ew̃/ /e/ /em/ /en/ /eń/ /eŋ/. For *ą* possible transcriptions are: /ow̃/ /o/ /om/ /on/ /oń/ /oŋ/. j~ is a nasal counterpart to /j/. It may replace /w̃/ before palatalized spirants, especially after /e/. w~ is a nasal counterpart to /w/; it occurs word-finally only after /o/ (spelt "ą") and /e/ (spelt "ę" - in emphatic context). Besides, it occurs only before spirants.

Phonetic labeling was done automatically with the program CreatSeg (Szymański & Grocholewski 2005) which uses HMM models. The features of the program include:

- calculating segment boundaries based on phonetic transcription
- context-dependent phoneme duration models
- considering "forced" transition points for semi–automatic segmentation
- accepting triphone statistical models trained with HTK tools
- tools for duration models calculation
- orthographic-to-phonetic conversion
- evaluation of decision trees to synthesis of unknown triphones
- accepting wave or MFCC files (plus several label formats) as input
- posterior triphone-to-monophone conversion

The segmentation accuracy is about 80-90% depending on the type of transcription. As accurate phoneme segmentation has fundamental importance for speech processing some manual corrections were introduced both into the automatic transcription and segmentation.

4.2.2. Prosodic annotation

An annotation system for unit selection speech corpus requires information about segmental (phoneme boundaries and identity) and prosodic structure (location of stress, accent, phrase boundary, boundary strength). The rules of segmentation at a syllable level were based on the assumption of a relationship between *sonority* and *syllable structure*. According to it, a continuous phoneme string can be converted into syllables by locating *syllable boundaries* at positions in the string immediately preceding a point of minimum sonority. Slavic languages are known for their variety of word-initial consonant clusters. This issue is important, because most theories of a syllable are based on the *Maximal Onset* rule (e.g. Clements & Keyser 1981) which says to syllabify as many consonants as possible into onsets. The definition of a "possible Onset" is governed by 1) the Sonority-Sequencing Principle ("within an Onset, sonority must not decrease") and 2) a language-specific feature that defines allowed Onset-clusters according to existing word-initial sequences ("CC is a possible Onset-cluster if it occurs word-initially"). The rules for determination of syllable boundaries in Polish were defined on the basis of a lexicon containing 10 mln different syllable structures (Śledziński 2007) and fully automatically implemented in a program *Annotation Editor* (Klessa 2006).

The automatically phonetically labeled speech corpus was annotated with prosodic features by 4 phoneticians on the basis of perceptual and acoustic analyses of the speech signals.

On the phrase level information on sentence modality and strength of the phrase break were provided. On the word level (in the sense that only one such feature can be marked per word) verification of the automatically inserted stress markers was carried out and pitch accent types were labeled. An inventory of 7 accent types was defined. The accents were distinguished on the basis of melodic properties similar to those distinguished in the IPO system (t'Hart, Collier & Cohen 1990): direction and range of the pitch movement, and its temporal alignment with accented vowel. The resulting inventory includes 2 rising, 2 falling 1 rising-falling pitch accents and two *level* accents distinguished by increased duration. Details on the acoustic realization of these accents and their distinguishing features is given in sec. 5.3.2 and 5.2.1 respectively.

As regards description of intonation at a phrase level only basic types of intonational tunes were distinguished and annotated: statements, questions, exclamations, and continuation phrases. For needs of the current thesis some changes were introduced into interpretation of this phrase-level description. Details are given in sec. 5.2.1 and 5.3.3.
Other information above the segmental level included:

a) word boundaries with distinction between orthographical words (e.g. #v_#do.mu) and prosodic words (e.g. #do.mu)
b) pause location: pauses were marked with #$p
c) syllable (initial) boundaries were marked with a dot
d) #$j - inappropriate segments for synthesis (marked only at word level)
e) ./ - vocal fry or creaky voice (marked only at syllable/word level) characterized by irregular, very low f0 and intensity

Examples of fully annotated utterances are given in Figure 29 & Figure 30.

Figure 29 illustrates a fully annotated utterance: "od kilku dni dokucza mi wręcz niemożliwy ból głowy (for several days I've been having an incredible headache). The utterance comes from the unit selection corpus. The top panel depicts the waveform, below it the pitch contours is visualized. The bottom panel contains segmentation (vertical lines indicate phoneme boundaries) and annotation. In the Figure 30 the utterance from the expressive speech database (male speaker) is depicted: "Skąd?! Do was wszystkich kieruję to pytanie!" (Where from?! I address this question to all of you).

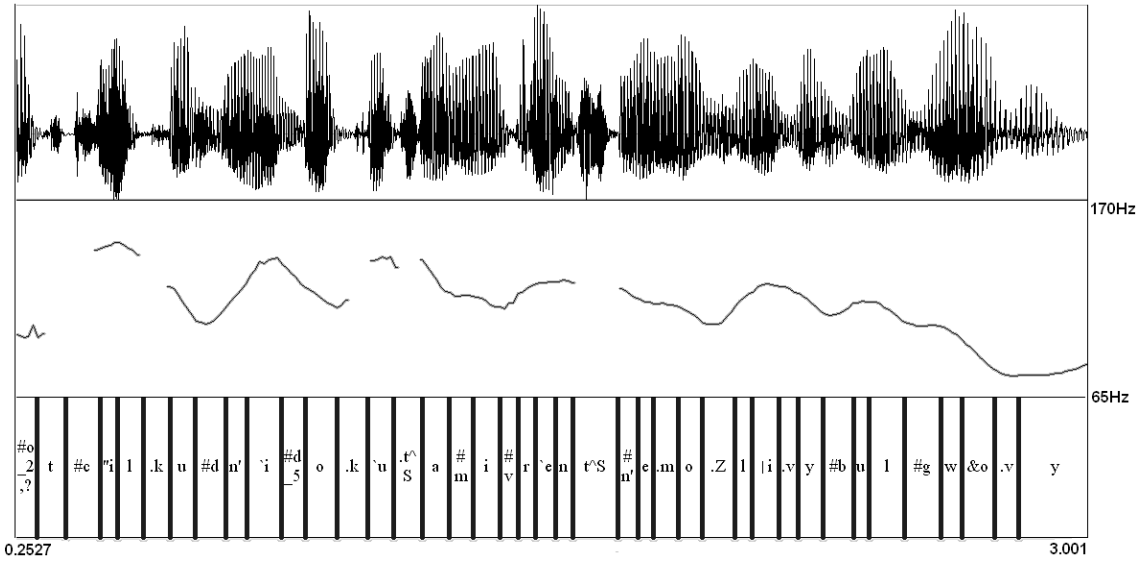


Figure 29: Example of a fully annotated utterance:
"Od kilku dni dokucza mi wręcz niemożliwy ból głowy." (For several days, I've been having an incredible headache); male speaker, unit selection corpus.

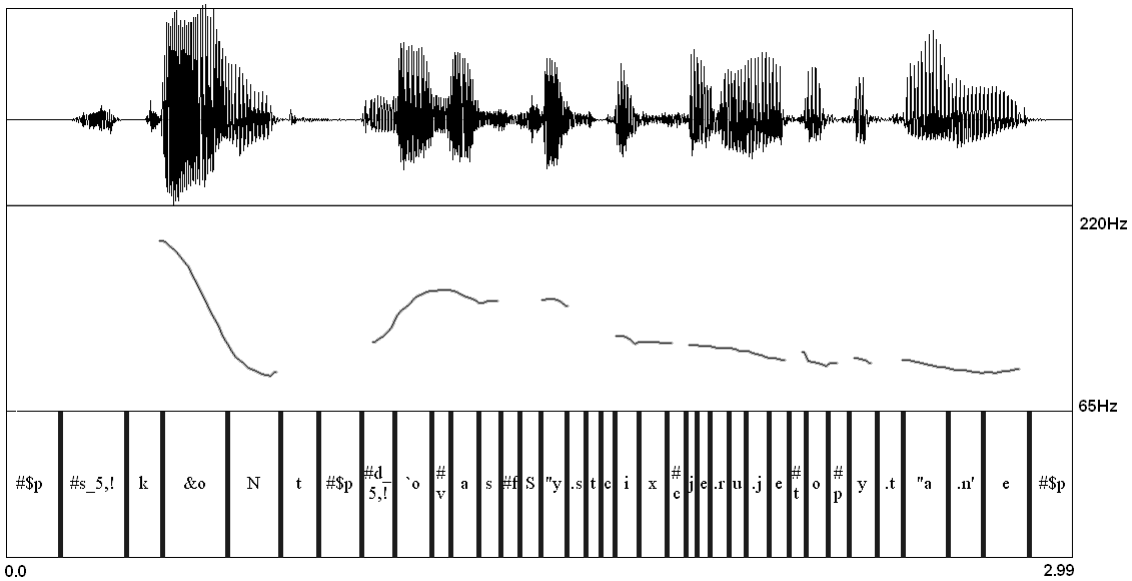


Figure 30: Example of a fully annotated utterance:
"Skąd! Do was wszystkich kieruję to pytanie." (Where from?! I address this question to all of you!); male speaker, expressive speech corpus.

4.3. Contour preparation and data collection for analyses

In this section the methodology adapted to f0 contour preparation and data extraction for analyses are discussed.

4.3.1. F0 extraction and processing

The determination of pitch of speech signals can be a difficult task (e.g. Hess 1983) and automatic methods of pitch extraction yield errors which can affect the results of pitch analysis. Preprocessing of fundamental frequency aims at reduction of errors resulting from automatic f0 extraction, so that the resulting f0 data can be regarded as a reliable basis for parameterization and analysis of f0 contours. f0 tracking errors occur most of all due to microprosody i.e., segmental effects on f0 contours, and to a lesser extent – due to voice quality and incorrect estimation of glottal pulses size. Erroneous f0 values can be defined as values given by a pitch tracker which have no confirmation in a perceptual analysis of pitch height or in a manual f0 measurement. The most common errors involve:

a) tracking of f0 values (1-4 frames) in unvoiced regions. Such situations require attention: on the one hand they might be faulty f0 values (so, they are consequence of erroneous pitch tracking and should be deleted), on the other hand they might signal presence of a very short vowel (e.g. of a phrase-initial position and surrounded by fricatives) and then the values should remain and be corrected if necessary
b) perturbations in the f0 course at the transitions from voiced to unvoiced parts and vice versa
c) presence of local outlying values (1-2 frames) in the f0 course over the voiced part of an utterance – this is characterized by jumps in frequency from one value to the next
d) absence of f0 values (1-2 frames) in a continuous f0 contour over the length of a voiced region
e) tracking of unrealistic f0 values (1-4 frames) below or above speaker's range: this usually results from voice quality e.g. at the end of phrases laryngelization produces very high f0 values, but the perceptual impression is that the pitch is low
f) pitch doubling or halving (1-4 frames) is when a pitch tracker gives the actual f0 values doubled or halved. This occurs if the size of the glottal pulses is incorrectly estimated.

Some of these problems are consequence of the glottal production mechanism (e.g. laryngelisation), while others (e.g. faulty f0 values at transitions from voiced to unvoiced regions) result from segmental effects (microprosody). They cause short-term deviations in the path of the contour and their elimination is important for a proper determination of the acoustic parameters used in the analysis of f0 contours.

There are various methods to eliminate these effects and at the same time to preserve the long term features which reflect macroprosodic component of an f0 curve which reflects the choice of the intonation pattern for the utterance: median/mean/exponential smoothing, various types of filters (e.g. Savitzky-Golay) or transformations (e.g. Hamming window) can be applied. Attention has to be paid not to transform the contour too heavily, so that 'meaningful' pitch variation (the one which is related to intonation) is not eliminated.

Another issue which has to be dealt with in f0 contour preparation is treatment of unvoiced parts of the contour. It is claimed that listeners perceive a contour with an unvoiced region as similar to one which contains no unvoiced region (Kohler 1991). This finding suggests that interpolation across unvoiced regions should produce a continuous f0 contour which contains the same information as the original f0 contour.

Bearing these considerations in mind in the current study the following methodology to f0 extraction and contour preparation was adopted.

A *Praat* (Boersma & Weenink 2005) script *getf0.psc* was written which performed the extraction and preprocessing of f0 data. Pitch was extracted every 10ms (based on *autocorrelation* method, Boersma 1993) and optimal values of thresholds for detection of silent and unvoiced intervals, as well as various costs (octave, octave-jump, voiced-unvoiced cost, for details see *Praat* manual) were specified. In order to get a reliable extraction and reduce errors it was decided that all values that fall outside the defined pitch range should be considered as erroneous. After a visual inspection of a subset of original f0 tracks a decision was made to set pitch range for the male speakers between 65 and 220Hz and for the female speakers between 120 and 450Hz. All values detected below and above these thresholds were treated as missing. In order to eliminate pitch perturbations and segmental effects a number of transformation methods was tested including median smoothing, Hamming window and others available in *Time Series* module in *Statistica* package. Finally, it was decided to use smoothing available in *Praat* which is based on a low-pass filtering of a pitch contour; filtering with a 5Hz bandwidth yielded the best results in terms of elimination of short-term deviations with simultaneous preservation of intonational features. It was also a more economic solution, as both extraction and smoothing could be performed by means of a single script. As regards unvoiced regions they were interpolated through, which is a commonly adopted solution (e.g. Möhler 1998, Taylor 2000, Braunschweiler 2003, Reichel 2007). The waveforms were resynthesized with the smoothed f0 contours using PSOLA (Charpentier & Moulines 1990). The *Praat* script *getf0.psc* used for f0 extraction and smoothing (male voice data) is given below.

```
###getf0.psc

form pitch extraction and manipulation
comment Enter directory where files are kept:
sentence soundDir C:\Documents and Settings\lacrimosa\Pulpit\baza E\
endform
Create Strings as file list... list 'soundDir$'\*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
  select Strings list
    fileName$ = Get string... ifile
    name$ = fileName$ - ".wav"
Read from file... 'soundDir$'\'name$'.wav
To Pitch (ac)... 0 65 15 no 0.03 0.55 0.01 0.35 0.14 220
Smooth... 5
Interpolate...
select Sound 'name$'
plus Pitch 'name$'
To Manipulation
Get resynthesis (PSOLA)
Write to WAV file... 'soundDir$'\'name$'new.wav
endfor
```

Figure 31 and Figure 32 illustrate pairs of *default* (i.e., extracted with default parameters) f0 contours (marked with a dotted line) and contours *extracted and preprocessed* with the script *getf0.psc* (marked with a solid line, bottom panel). The top panel contains the waveform. The vertical dashed lines indicate word boundaries. In the Figure 31 the f0 contour of a sentence from the unit selection corpus is depicted: "ponieważ Piotr jest tak bardzo leniwy, nie będzie mu się chciało iść do sklepu" (as Peter is so much lazy, he won't feel like going to the shop). In the Figure 32 the utterance was: "czy sylwetki te naprawdę wyleciały przez okno" (whether the figures really flew out of the window"), female speaker AW, expressive speech corpus.

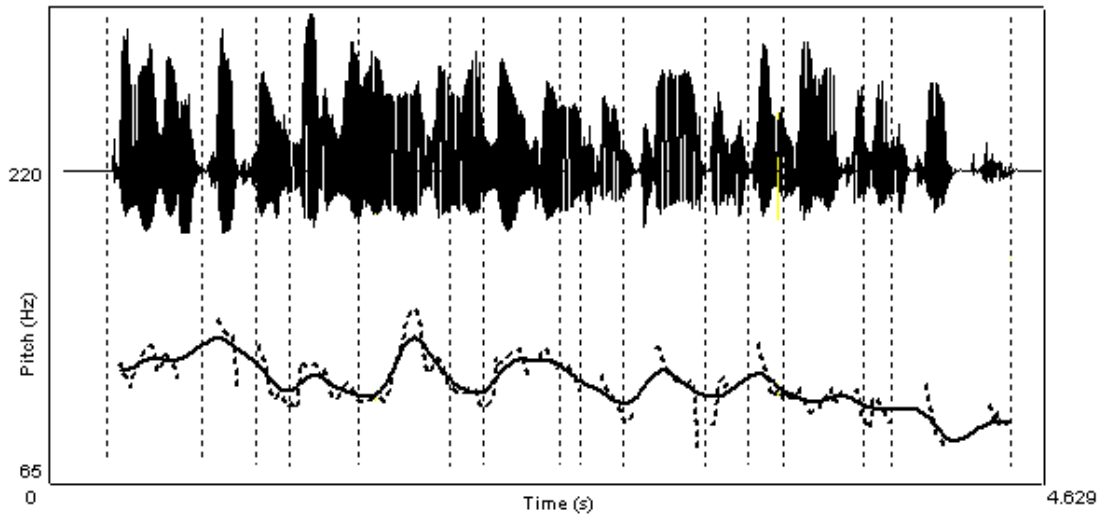


Figure 31: Example of two contours extracted with the default parameters (dotted line) and with the getf0 script (solid line). The dashed vertical lines indicate word boundaries.

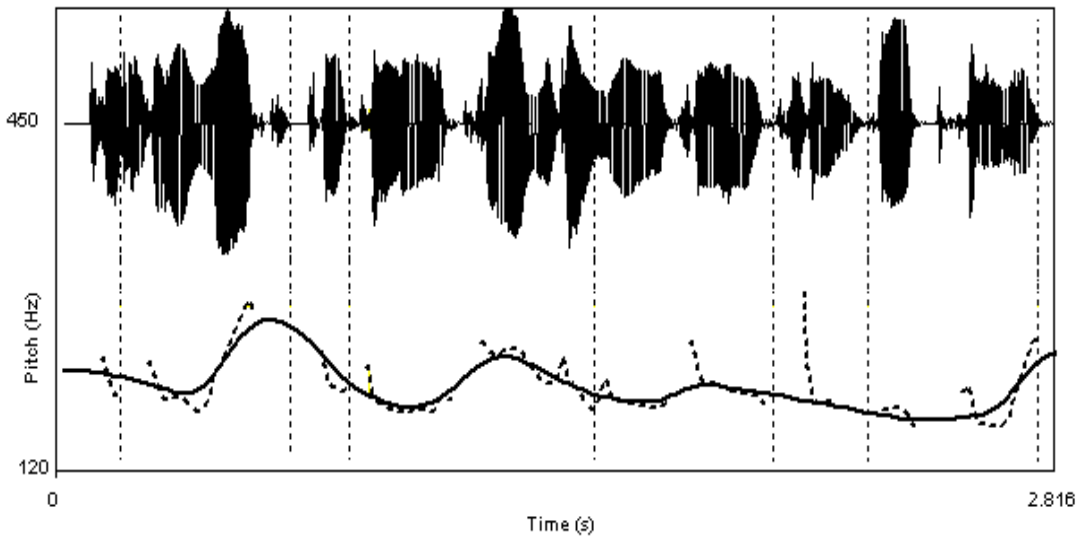


Figure 32: Example of two contours extracted with the default parameters (dotted line) and with the getf0 script (solid line). The dashed vertical lines indicate word boundaries.

4.3.2. Data collection for analyses

For needs of acoustic and statistical analyses of intonation presented in this thesis the f0 data listed in Table 9 was collected for each syllable and vowel; the specific measurements are given in the script *collectf0data.psc*.

parameter	description
f0start	f0 value at the start of a syllable/vowel
start	syllable/nucleus start in ms
f0end	f0 value at the end of a syllable/vowel
end	syllable/nucleus end in ms
dur	syllable/nucleus duration
nucl_mid	the middle of the nucleus (in ms)
f0nucl_mid	f0 value in the middle of the nucleus
f0mean	overall pitch level (in Hz)
stdev	standard deviation from the mean
f0max	maximum pitch (in Hz)
maxloc	absolute position of f0 peak (in ms)
f0min	minimum pitch
minloc	absolute position of f0 minimum (in ms)
c1, c2	amplitude of the rise/fall
a1, a2	steepness of the rise/fall
d1, d2	duration of the rise/fall
z1, z2	data needed for calculation of the slope feature (see the script below)
c3, c4, d3, d4	data necessary for calculation of tilt amplitude and duration (see the script below)
Amp	tilt amplitude
Dur	tilt duration
Tilt	tilt parameter describing shape of a pitch contour on the
slope	describes the amount of pitch variation on the syllable/vowel

Table 9: F0 and duration parameters extracted with the collectf0data.psc script and used in the analyses presented in this thesis.

The definition of the *slope* parameter needs a comment.

Slope is regarded as a measure of variability in pitch calculated as a ratio of the sum of rising and falling amplitude to the sum of their duration. This parameter is expected to be a good predictor of pitch accent location, because pitch variation is the main acoustic cue of accentual prominence (cf. sec. 6.1.1). The amplitudes/durations are summed together because it is assumed that both rising and falling amplitude/duration contribute to the perceived prominence. Their ratio is used to make it possible to compare the amount of pitch variation on syllables of a different duration.

The data listed in the table above was derived automatically from f0 contours (extracted and smoothed as described in the previous section) and phonetic segmentation of the waveforms with a *Praat* script *collectf0data.psc*.

```
###collectf0data.psc

form syllable-based pitch analysis
comment Enter directory where sound files and text grids are kept:
sentence soundDir C:\Documents and Settings\lacrimosa\Pulpit\diss_latest\analysis
comment Enter full path of the resulting text file:
sentence output C:\Documents and Settings\lacrimosa\Pulpit\diss_latest\analysis\results.txt
comment Enter the number of tier to be analyzed
integer tier_number 1
endform
Read Strings from raw text file... C:\Documents and Settings\lacrimosa\Pulpit\diss_latest\analysis\list1.txt
     numberOfFiles = Get number of strings
for ifile to numberOfFiles
     select Strings list1
     fileName$ = Get string... ifile
     name$ = fileName$ - ".wav"
Read from file... 'soundDir$'\'name$'.wav
Read from file... 'soundDir$'\'name$'.Pitch
Read from file... 'soundDir$'\'name$'.Textgrid
select TextGrid 'name$'
nlabels = Get number of intervals... 'tier_number'
   for label from 1 to nlabels
select TextGrid 'name$'
  label$ = Get label of interval... tier_number label
  if (label$ <> "")
     start = Get starting point... tier_number label
     end = Get end point... tier_number label
     dur = end - start
     nucl_mid = start + (dur/2)
select Pitch 'name$'
f0start= Get value at time... 'start' Hertz Nearest
f0end = Get value at time... 'end' Hertz Nearest
f0max = Get maximum... 'start' 'end' Hertz Parabolic
maxloc = Get time of maximum... 'start' 'end' Hertz Parabolic
c1 = f0max - f0start
c2 = f0max - f0end
d1 = maxloc - start
d2 = end - maxloc
a1 = c1/d1
a2 = c2/d2
c3 = c1 - c2
c4 = c1 + c2
d3 = d1 - d2
d4 = d1 + d2
z1 = c3 + c4
z2 = d3 + d4
Amp = c3/z1
Dur = d3/z2
Tilt = Amp + Dur
slope = z1*z2
f0min = Get minimum... 'start' 'end' Hertz Parabolic
minloc = Get time of minimum... 'start' 'end' Hertz Parabolic
f0nucl_mid = Get value at time... 'nucl_mid' Hertz Nearest
meanf0 = Get mean... 'start' 'end' Hertz
stdev = Get standard deviation... 'start' 'end' Hertz
echo 'fileName$' 'label$' 'start' 'f0start' 'end' 'f0end' 'dur' 'nucl_mid' 'f0nucl_mid' 'meanf0' 'stdev' 'f0max'
'maxloc' 'c1' 'c2' 'd1' 'd2' 'a1' 'a2' 'Amp' 'Dur' 'Tilt' 'slope' 'f0min' 'minloc' 'newline$'
  fappendinfo 'output$'
endif
endfor
select Pitch 'name$'
plus TextGrid 'name$'
plus Sound 'name$'
Remove
endfor
```

At the input the script required a waveform with corresponding pitch file (f0 contour) and a text grid file with phonetic segmentation. The original annotation files were all blfs (BOSS label format) and they had to be converted to a text grid format first. A Perl script was written for that purpose (by S.Breuer, IKP, Uni Bonn). Since the goal was to collect data for vowels and syllables, the script was run twice: first, on the original annotation file including all segment boundaries (later only the information regarding vowels was left) and for the second time - on a file with phoneme boundaries removed and only syllable boundaries left (which was done with another Perl script). For each interval (phoneme/syllable) marked in a text grid file the data listed above was collected and saved in a text file. This file was later exported to *Statistica* spreadsheet where further calculations were made. In the resulting database for each syllable/vowel the information on the features of the previous and next syllable/vowel were provided. Data describing the current nucleus/syllable was given (SAV)/(SA) extension, the previous nucleus/syllable - (PAV)/(PA) and the next - (FAV)/(FA).

Apart from the parameters derived automatically from utterance's acoustics and listed above a number of other features which could be useful for definition of the description of pitch accents and boundary tones on the phonetic level was determined. The necessary calculations were carried out on the basis of the parameters extracted with the script shown above; the features include:

a) *peak(norm)* - f0max position on the syllable expressed as a distance from the nucleus onset: values <1 indicate a peak which occurs before the onset
b) *direction* - calculated as a difference between overall pitch level on the previous nucleus: f0mean(PAV) and on the current nucleus: f0mean(SAV)
c) *tilt* - it is calculated in the same way as the Tilt parameter extracted with the script, but unlike the former it describes the shape of the pitch contour in a two-syllable window including accented (SA) and post-accented syllable (FA). It is expected that this features will help to distinguish between accents of a different direction and different range.
d) *f0peak(relative)* - f0max determined in a two syllable window including the current (SA) and next syllable (FA) relative to overall pitch level (f0mean) on the phrase
e) *f0mean(relative)* - f0mean on vowel relative to overall pitch level (f0mean) on the phrase
f) *f0min(relative)* - f0min determined in a two syllable window including the current (SA) and next syllable (FA) relative to overall pitch level (f0mean) on the phrase
a) *nucl_dur(expected)* - mean duration of the nucleus determined for each vowel type
g) *nucl_dur(relative)* - nucleus duration calculated as the ratio of the absolute duration (extracted with the script) and expected duration. It indicates the degree to which a vowel is longer or shorter as expected. This duration measure was proposed in (Rapp, 1996).
h) *syl_dur(expected)* - mean duration of the syllable of a given structure determined by the number of phones in the onset and coda (i.e., onset and coda length), segment type (voiceless/voiced obstruent or sonorant) and vowel type
i) *syl_dur(relative)* - syllable duration calculated as the ratio of the absolute duration (extracted with the script) and expected duration. It indicates the degree to which a syllable is longer or shorter as expected.

4.4. Statistical analyses and statistical modeling techniques

In the current section a general overview of statistical methods and modeling techniques used in intonation analysis and modeling is given. The presentation is confined to this information which is important from the point of view of the design of classification and regression models without going into detail in the computational approaches underlying the presented methods.

In order to carry out the necessary statistical analyses and design the classification and regression models a professional software *Statistica 6.0* (StatSoft, Inc. 2001) was used. It offers a wide range of basic and advanced analytic procedures and is characterized by a flexible and easily customizable user interface. The overview given in this section is based most of all on the information provided in the *Statistica* electronic manual.

In sec. 4.4.1 the goal, assumptions and interpretation of the results of basic analyses (correlation, ANOVA and discriminant function analysis) are discussed. With these analyses one can a) investigate associations between features, b) search for features which discriminate the best between groups and can be used as predictor variables in classification and regression tasks.

The application of neural networks and decision trees to classification and regression problems is discussed in sec. 4.4.2 and 4.4.3 respectively. In this thesis these two methods (and also discriminant function analysis) are used to detect location of accented syllables and phrase boundaries and to identify types of pitch accents and boundary tones (Chapter 6). Regression models are used for estimation of f0 targets from which f0 contours are generated (Chapter 7).

4.4.1. Basic analyses

1. Correlation

Whenever continuous data is dealt with relation between variables has to be investigated in order to avoid redundancy. *Correlation* can be used for this purpose - it is a measure of the relation between values of two or more variables, or in other words - it determines the extent to which values of the variables are "proportional" to each other. Such dependencies are expressed in terms of a *correlation coefficient*: the most popular one is *Pearson's r.* It can range from -1.00 to +1.00: the value of -1.00 represents a perfect negative correlation and a value of +1.00 represents a perfect positive correlation. In order to avoid redundancy only those variables should be selected for the analyses for which the value of *r* is the closest to 0 (i.e., little correlation). Data redundancy may have serious consequences for the results of a discriminant function analysis (see description below).

2. Analysis of variance (ANOVA/MANOVA)

The purpose of analysis of variance is to test for significant differences between means in different groups of variables by comparing variances. The analysis results provide an answer to the question whether two or more groups differ significantly with respect to the mean of one

specific dependent variable (*one-way ANOVA*) or more than one dependent variable at a time (*multivariate analysis of variance* or *MANOVA*). Whenever means of a given variable(s) differ significantly between groups (which is indicated by the value of the *F* test) it can be concluded that this variable *discriminates* between these groups. In general, the higher the value of the F test, the bigger the discriminative power the variable has.

Whenever more then two groups are compared *Scheffe's test* can be carried out in order to investigate which groups in particular are different from each other with respect to the mean of a specific dependent variable. Scheffe's test is also used whenever one finds unexpected results in an experiment to test their statistical significance. In case of data sparseness *Tukey unequal number HSD test* can be applied instead of Scheffe's test, because it provides a reasonable test of differences between group means if the number of cases in the groups is unequal but at the same time not too discrepant.

3. Discriminant function analysis

Its goal is to determine which variables discriminate the best between two or more naturally occurring groups. Variables used in a discriminant function analysis should be *continuous* - even though there is a possibility to include *categorical* dependent variables the results may be *statistically questionable* due to the computational approach taken to coding the categorical variables. An important assumption is that variables used to discriminate between groups are not redundant, which can be concluded by measuring the correlation coefficient reflecting the association between the variables.

From the point of view of the computational approach discriminant analysis is very similar to the analysis of variance. The value of the F test indicates whether groups differ significantly with respect to the mean of a specific dependent variable: this variable can be then used to predict whether a new case belongs to the group or not. For example, in the context of pitch accent analysis initially, a number of features describing the accents can be provided and then, a *stepwise* discriminant analysis can be applied in order to distinguish the features that discriminate between the accent types the best. When a *forward stepwise* analysis is chosen dependent variables are moved into the model on the basis of their discriminative power. In the *backward stepwise analysis* the model initially includes all predictor variables and effects, which are then removed. In the end, a model with only significant predictor variables is obtained.

During the analysis *discriminant functions* are estimated. There are as many functions as groups and each function is used to distinguish between two groups. The functions are defined in terms of standardized coefficients which reflect the amount of the individual contribution of the dependent variables into the distinction defined by the discriminant function.

Another popular application of the discriminant analysis is to prediction of a case's membership of a given group. *Classification functions* are estimated during the analysis (again, there are as many functions as groups) and they make it possible to calculate the *classification values* for cases. Each case is then classified into the group for which it has the biggest

classification value. In order to assess the performance of classification functions it is necessary to carry out the classification for the cases which were not used for the estimation of the functions, because post-hoc predictions will always yield better results then a-priori predictions. For this purpose cases should be divided into two subsets: one used for computing classification functions and the other - for their assessment in a cross-validation test.

4.4.2. Neural networks

Neural networks represent a powerful modeling technique inspired by the structure and learning performed by the human brain. In speech technology, next to classification and regression trees neural networks belong to the most popular techniques used to solve classification and regression problems. Information concerning the design of neural network models can be found in (Tadeusiewicz 1993, 1998).

1. Classification

In classification problems, the goal is to assign each case to one of a number of classes (or, more generally, to estimate the probability of membership of the case in each class) on the basis of the information provided by the input variables which can be numeric or nominal.

Prior to the design of a network it has to be ensured that none of the assumptions listed below is violated:

a) the variables can be continuous and nominal, but neural networks do not function well if the nominal variable has a significant number of values
b) as regards the amount of data in the training set required for a successful training of the neural network it depends on the size of the network and the complexity of the problem to be solved. It is assumed that the number of cases should be 10 times greater then the number of the connections in the network. The number of cases should also be proportional to the number of variables
c) only influential variables should be used (this can be determined in the analysis of variance, discriminant analysis, sensitivity analysis)
d) outliers should be removed, because they may affect the training and the networks are not always resistant to noises in the data
e) specification of how to handle missing data: missing values can either be substituted with the mean value or cases including missing values are excluded from the training

Design of a neural network model involves the following steps (Patterson 1996):

a) selection of the type of the network: for classification linear networks, multilayer perceptrons (MLP), radial basis function (RBF) and probabilistic neural networks (PNN) can be used. In non-linear networks *complexity* has to be specified (the number of hidden layers and hidden neurons)
b) sampling of cases. Data is divided into 2 or 3 subsets: training, test and selection. A neural network is optimized using a training subset. Test subset is used to perform an unbiased estimation of the network's likely performance. Additionally, a selection subset can be

distinguished which is used to halt training, to mitigate over-learning, or to select from a number of models trained with different parameters.

c) a number of experiments is carried out iteratively during which various networks are trained: the performance of the model is assessed on the basis of predictive accuracy on the test subset and value of network's error on the test/selection subset
d) when training of the network is in progress the training error is decreasing and so is the selection error. If, on the contrary, the selection error is increasing it means that the network starts to overfit the data and is no longer capable of generalizing the results of the training. In this case the training should be ceased
e) once an optimal network configuration is found the model's performance should be estimated on new data. It is assumed that in order to get a reliable indicator of the model's performance the training should be repeated a number of times, each time using new training, selection and test cases drawn from the population. Then, the test set performances of the individual networks can be averaged. In order to avoid data loss for training and to keep a reasonable size of the three sets resampling of the input data can be carried out Among resampling techniques there are *cross-validation*, *Monte Carlo* and *bootstrapping.*
f) final selection of the network should be based on its predictive accuracy and error value

2. Regression

The objective of building regression models is to estimate the value of the output variable(s), given the known input variables (both discrete and continuous). Unlike in the classification problems the output variable(s) is continuous instead of nominal. Regression problems can be solved using the following network types: linear, MLP, RBF and GRNN (generalized regression neural networks).

Most of the assumptions underlying the design of the network for classification purposes are also valid for regression networks. What differs is assessment of model's performance. In regression, the most important indicator of performance is *standard deviation of the prediction errors*. If this is no better than the training data standard deviation, then the network has performed no better than a simple mean estimator. So, if the ratio of the prediction error SD to the training data SD is significantly below 1.0 it means that the network is capable of performing good regression and making a reasonable prediction. The performance of the network can also be assessed by analyzing the residual errors generated for the observed values and the values estimated by the model. Another measure of the performance of a regression networks is Pearson's *r* coefficient which shows to which extent the observed and estimated values are correlated. The results of the network's performance can also be displayed graphically, which makes comparison of the distribution of observed vs. estimated values easier.

4.4.3. Decision trees

Decision trees (also: classification & regression trees, C&RT) are the state-of-the-art method applied to solve classification and regression problems. In this thesis, decision trees were used only for classification purposes.

In a number of respects classification trees are significantly different from other classification methods. First of all, a decision tree can be regarded as a set of logic conditions. Trees employ a hierarchy of predictions, with many predictions sometimes being applied to particular cases, to sort the cases into predicted classes. Traditional methods use simultaneous techniques to make one and only one class membership prediction for each and every case.

In this thesis the *QUEST* classification tree program (Loh & Shih 1997) available in *Statistica* was used. It has a number of innovative features for improving the reliability and efficiency of the classification trees that it computes. It is faster and less unbiased then other programs (e.g. C&RT, Breiman et al. 1984) particularly in situation when 1) predictor variables have dozens of levels or 2) some predictor variables have many levels while other have only few. Moreover, the speed of the Quest program does not affect the predictive accuracy.

Chapter 5. The description of intonation

In this chapter the description of intonation on the surface phonological and phonetic level is proposed. The former is based on the prosody labeling system used defined for and used in the Polish unit selection corpus. As explained in the sec. 1.2 the specification of description of intonation can be regarded as the first task when developing an intonation model, because it is used by the components of the model which deal with coding and generation of intonation contours.

The surface-phonological description proposed here is in terms of discrete distinctive categories of pitch accents and boundary tones and encodes not only melodic, but also functional aspects of intonation. The usefulness of this description for generation of intonation contours (Hypothesis 1 & Hypothesis 1b) will be investigated in Chapter 7.

Apart from that, description of intonation on the phonetic level is defined. On this level intonation is represented and analyzed in terms of continuous parameters which describe the macroprosodic component of f0 contour and encode melodic aspects of intonation. In the definition of the phonetic description the following issues are taken into account (Taylor 2000):

a) the description should be *compact* and free from *redundancy* i.e., it should use a small number of features which can not be derived from one other
b) the description be should be such that it can be easily *derived from an utterance's acoustics*
c) it should have *wide coverage* i.e., be able to express distinctions in utterances which are perceptually different

The usefulness of this description for automatic recognition of the type of pitch accents and phrase boundaries (Hypothesis 1 & Hypothesis 1a) will be investigated in Chapter 6.

Another issue investigated in this chapter is prosodic structure and particularly the number of phrasing levels and classification of phrases. It was shown in (Clark 2003) that the information related to prosodic structure can be effectively used in intonation modeling as well as in within- and across-speaker normalization of pitch variation. Therefore, it is expected that the description of prosodic structure determined in the analyses presented in the next section provides an important information for estimation of pitch targets for contour generation and thus, affects the quality of f0 contour generation in speech synthesis (Hypothesis 1c).

The analyses presented in this chapter are based on the subset of Polish unit selection corpus described in sec. 4.1.1.

5.1. Prosodic structure

The analyses presented in the following sections aim at the definition of description of prosodic structure. It was shown in (Clark 2003) that information provided by such description

can be used as a framework for control of pitch variation caused by extrinsic factors (such as phrase position in the utterance, phrase length) and proved that it improves the results of f0 prediction and in effect: the quality of f0 contour generation. It is so, because prosodic structure affects pitch range, which means that it affects the distribution of pitch targets on an f0 scale.

Pitch range serves very often as a reference point for description of the scaling of pitch targets (this approach is called *normalizing*, see Ladd 1996:256). The representation of the intonational events proposed in this thesis describes pitch accents and boundary tones in terms of pitch movements between specific target levels: H and L. Since this is a surface phonological description some of these targets will effectively correspond to phonological tones, but others not. H targets are located at the top of or high in the current pitch range, and L targets are located at the bottom of or low in the range. Therefore, for the purpose of both recognition and modeling of intonational events it is necessary to get to know what kind of factors and to which extent cause pitch range modifications.

Following the methodology adopted in (Clark 2003) the effect of prosodic structure on pitch range is analyzed and on the basis of the results the description of prosodic structure is proposed. Prior to that, basic assumptions concerning prosodic structure in Polish are given in the next section.

5.1.1. Preliminary remarks

There is no general consensus on the number of levels of phrasing and different theories give different proposals. For example, the studies presented in (Wightman et al. 1992) and (Ladd & Campbell 1991) showed that there exist acoustic cues to distinguish among four levels of phrasing, but most often two levels are accepted. They are defined and referred to in various ways in different theories: *intonation* and *intermediate phrases* (Beckman & Pierrehumbert 1986), *major phrases* and *tone groups* (Ladd 1986), *double* and *single bar boundaries* (O'Connor & Arnold 1961), *major* and *minor tone groups* (Trim 1959). Without going into detail, the general assumption is that the smaller constituent includes a nucleus and the bigger constituent consists of at least one smaller constituent and has an audible break associated with it (for discussion on issues related to phrasing and prosodic structure see sec. 2.2.2 & 2.2.3).

In (Demenko 2000, Francuzik, Karpiński & Kleśta 2002) acoustic analyses of realization of intonational phrases in Polish were carried out. Their results showed that prolongation of the pre-pause syllable (and especially the vocalic nucleus), occurrence of a final fall, additional pitch movement at the end of a phrase and a pause following it belong to the cues that help listeners to detect phrase boundaries. On the basis of these results and analyses of text corpora rules for automatic detection of phrase boundaries in written texts were formulated (Baranowska et al. 2003, Karpiński et al. 2003).

The phrasing scheme applied in the Polish module of BOSS system distinguishes between two levels, i.e. *major* and *minor intonational phrase.* Major phrases (IPs, intonational phrases) are equivalent to the length of a sentence or paragraph and they have to include at least one minor intonational phrase (ip). IP is the domain over which a full pitch range reset occurs. Minor phrases are equivalent to the length of a clause (subordinate, coordinate) or syntactic

phrase and have to include a nuclear accent. The two phrase types are associated with boundaries of a different strength which are marked with two different break indexes: 5 (major phrase, IP boundary) and 2 (minor phrase, ip boundary). It is assumed that there is no need to define more than two levels of phrasing and that this kind of system can capture pitch range phenomena very well. Acoustic analyses presented in the following sections are expected to confirm this hypothesis. A comparison of the three phrasing systems for Polish mentioned above is given in Figure 33. In general, the phrasing levels do not always correspond to each other between the systems, but these differences are not very significant.

ACOUSTIC CUES	BREAK INDEX		
	BARANOWSKA ET AL. 2003	KARPIŃSKI ET AL. 2003	POLISH BOSS
extra strong, obligatory, signaled by a very long pause and one of the features: fundamental frequency movement, increased energy, increased duration		[1] paragraph boundary	
strong, non-optional, signaled by at least two prosodic features (viz. pause, fundamental frequency movement, increased energy, increased duration, preferably one of them is a pause	[1] paragraph/sentence/ parenthetical clause/quotation boundary	[2] sentence/parenthetic also clause/quotation boundary	5 paragraph/sentence boundary
weak, non-optional, signaled by one of the features: pause, fundamental frequency movement, increased energy, increased duration, preferably not a pause	[2] intra-sentential boundary; separate parts of complex sentences	[3] intra-sentential boundary; separate parts of complex sentences	2 intra-sentential boundary; separate parts of complex and compound sentences, clauses
weak, optional, can be signaled by fundamental frequency movement, increased energy, increased duration or none of them	[3] intra-sentential boundary; depending on factors such as speech tempo and phrase length to separate parts of compound sentences	[4] intra-sentential boundary; depending on factors such as speech tempo and phrase length to separate parts of compound sentences	2 intra-sentential boundary; depending on factors such as speech tempo and phrase length, at boundaries of syntactic phrases

Figure 33: Comparison of three different phrasing systems

As mentioned before, major phrases (IPs) have to include at least one minor phrase (ip). Above the level of a major intonational phrase, there is an *utterance* which can include more than one major phrase. At a lower level, each phrase consists of *prosodic words* consisting of *syllables*. Prosodic structure defined in this way is an example of a *hierarchical system* and follows the assumptions of the *Strict Layer Hypothesis* according to which constituent at a specific level in the structure can *dominate only constituents at a lower level*. There exists a rich experimental material (e.g. Cooper & Sorensen 1981, Ladd 1988) showing that prosodic structure defined in the SLH theory is too restricted (see also sec. 2.2.3). In spite of this it is

assumed that within the framework of this theory a model of prosodic structure can be defined which makes effective control of pitch variation at a phrase level possible.

5.1.2. Hypotheses

In the previous section only general assumptions concerning prosodic structure in Polish were presented. Now, more detailed analyses will be carried out in order to provide a finer model of prosodic structure which can be used as a framework for control of pitch variation at the phrase level. Following the methodology adopted in (Clark 2003) the effect of the following factors on pitch range modifications will be investigated:

a) phrase type (minor vs. major)
b) minor phrase (ip) position (initial, medial, final) in the major phrase (IP)
c) IP length measured in the number of ips it includes (single vs. complex)

In order to define the prosodic structure representation the following hypotheses will be tested.

Hypothesis 1. *Pitch variation on the phrase level can be effectively controlled within a two-level phrasing system distinguishing between major(IP) and minor phrases (ip).*

In order to confirm this hypothesis it has to be proven first that major (IP) and minor (ip) phrases indeed constitute different phrasing levels. If IP constitutes a different phrasing level than ip there should occur a pitch range reset at the end of IP to a default starting pitch at the start of the next IP. This can be confirmed by investigating differences in the pitch range and the level of phrase-initial and/or phrase-final pitch between IP-final and IP-initial ips. Moreover, the analysis of ips of the same position in IPs of a different length is expected to give an insight into the alignment of IP-initial and IP-final ips. If IP is the domain of a full pitch reset then initial ips should align with each other irrespective of IP length and so should IP-final ips. If that is the case then it can be assumed that IP constitutes a different phrasing level than ip.

Hypothesis 2. *The primary distinction should be drawn between intonational phrases (IPs) consisting of a single ip (i.e. single phrases) and those which include more than one ip (complex phrases). The latter have a specific internal structure in which phrases (ips) of initial, medial and final position are distinguished.*

This hypothesis is based on the assumption that since major phrases can include one or more minor phrases a question arises whether there are differences in the pitch range between IPs of a different length. Another issue concerns IPs which include several minor phrases: what kind of sub-categorization of minor phrases should be adopted so that no significant variation in pitch is lost. As regards distinctions based on IP length it is assumed that *single phrases* should constitute a class of their own. As regards the sub-categorization of phrases it is assumed that depending on their position in IP minor phrases should be classified as *initial, medial* or *final*. This hypothesis will be tested by investigating the statistical effects of different classifications of phrases on the distribution of pitch targets describing pitch range.

Hypothesis 3. *Structural elements of intonational phrases (i.e. classes of phrases distinguished in the analyses) should not be affected by IP length.*

In this hypothesis the effect of major phrase (IP) length on pitch range characteristics of minor phrases it consists of will be investigated. If the proposed classification of phrases is adequate, then it is expected that IP length should have no statistically significant effect. More importantly, if IP is the domain of a full pitch range reset, then ips of initial position in IP should align with each other irrespective of IP length, and so should medial and final ips. In order to test this hypothesis differences in pitch range between ips of the same position in IPs of a different length will be investigated

The effects discussed above will be investigated with *analysis of variance* (ANOVA) and *Scheffe's tests*. The purpose of ANOVA is to test for significant differences between means by comparing variances. Scheffe's tests show which means contributed to the effect: that is, which groups are particularly different from each other. Scheffe's tests are also used whenever one finds unexpected results in an experiment to test their statistical significance. In case of data sparseness *Tukey unequal number HSD test* will be applied instead of Scheffe's test, because it provides a reasonable test of differences between group means if the number of cases in the groups vary but is not too discrepant.

5.1.3. Methodology

The speech material used in the analyses presented in the next sections consists of 1137 major intonational phrases (1061 statements, 59 wh-questions and 14 exclamations) and 2513 minor phrases. It must be noted, that the number of phrases in each dataset created for the purpose of particular experiments varied, because the least numerous classes were excluded in order to avoid results based on a sparse data.

It is assumed that pitch range can be described by the following variables:

a) f0max represents the top of speaker's range and is the level where H pitch targets are located
b) f0min represents the bottom of the range and is the level of L targets
c) f0mean determines the middle of the range and is the level where M targets are located
d) S.D. (standard deviation) describes the amount of pitch variation

As explained in sec. 2.2.7 pitch range involves two dimensions: overall level and span. In the current study f0max, f0min, meanf0 and S.D. variables will be used to express modifications in the overall level (i.e., rising or lowering of the pitch scale) and span (i.e., expansion or compression of the pitch scale). Following the methodology adopted in (Clark 2003) *f0start* and *f0end* parameters will be used additionally to investigate pitch variation in phrases and especially - their alignment. Analysis of variation in f0end which describes phrase-final pitch will make it possible to test Hypothesis 1: If IP constitutes a different phrasing level than ip and constitutes the domain of full pitch reset then IP-final pitch should reach the bottom of speaker's range and be significantly lower in comparison to f0end of non-final ips.

The complete set of variables used in the analyses is listed below. These parameters were derived from syllable-based data automatically collected with a *Praat* script *collect f0 data* (sec.4.3.2):

a) *f0start* - the first f0 value in a phrase - the first f0 value on a phrase-initial syllable
b) *f0end* - the last f0 value in a phrase - the last f0 value on a pre-boundary syllable
c) *f0max* - maximum f0 in a phrase derived from the highest f0 peak associated with accented syllable in a given phrase
d) *f0min* - minimum f0 in a phrase derived from the lowest f0min associated with accented syllable in a given phrase
e) *f0mean* - mean f0 over the length of a phrase
f) *S.D.* - standard deviation from the mean f0

F0mean and S.D. were extracted from original f0 contours using a modified version of collectf0data.psc script which worked on text grids containing only phrase type and boundary location information.

On the basis of phonetic alignment of utterances for each phrase a number of features was extracted.

Information concerning minor phrases (ips):

a) *ip distance to IP start (dist_start):* Subsequent ips were numbered from *0* (at the start of IP) to *n* (at the end of IP). In this way IP-initial ips were grouped together, but not IP-final ips. This grouping makes it possible to investigate whether there is a significant difference in the overall pitch range between initial vs. non-initial ips. If it is so, it means that IP-initial ips share properties and should constitute a class of their own (Hypothesis 2).

b) *ip distance to IP end* (*dist_end*): Subsequent ips were numbered from *0* (at the end of IP) to *n* (at the start of IP), in this way IP-final ips were grouped but not IP-initial ips. This grouping makes it possible to investigate whether there is a significant difference in pitch range between final vs. non-final ips. If it is so, it means that IP-final ips share properties and should constitute a class of their own (Hypothesis 2).
c) *ip position in phrase* (*initial, final, single* classification): Distinction is made between IP-initial, IP-final and single phrases. This grouping makes it possible to investigate whether there are significant differences in pitch range between initial vs. final vs. single vs. other ips. If it is so, the Hypothesis 2 can be accepted. This grouping makes it also possible to test Hypothesis 1: If IP and ip constitute two different phrasing levels it is expected that final fall in pitch to the bottom of speaker's range occurs at the end of IP and is followed by pitch range reset to default starting pitch at the beginning of the next IP. This hypothesis can be tested by investigation of differences in pitch range and f0start and f0end between ips of initial and final position in IP.

Information concerning major phrases (IPs):

a) *IP length:* It is calculated as the total number of ips that a given IP includes (a sum of dist_start, dist_end and 1). The investigation of the effect of the interaction of ip position and IP length is expected to confirm Hypothesis 1 which says that ip and IP constitute different phrasing levels, and Hypothesis 2 according to which depending on ip position in IP, initial, medial and final ips are distinguished and depending on IP length single ips are

distinguished. Moreover, if no significant effect of IP length on the distribution of pitch targets describing pitch range of different types of phrases is observed Hypothesis 3 will be confirmed.

On the basis of statistically significant effects of various categorizations of phrases on pitch range a model of prosodic structure will be proposed.

5.1.4. Analysis of phrases of a different position

1. Analysis and results for dist_start dataset.

At first the effect of *dist_start* categorization on pitch range characteristics of phrases was investigated. Phrases of the same position in IP measured in a number of ips from IP start were grouped together. The least numerous ip classes (5 and 6) were excluded from the analysis. Figure 34 illustrates distribution of pitch targets (f0max, f0min, meanf0, f0end and f0start) in various ip classes: 0 indicates IP-initial ips, 1 - ips of the second position in IPs, 2 - ips of the third position in IPs, etc. Table 10 shows ANOVA results.

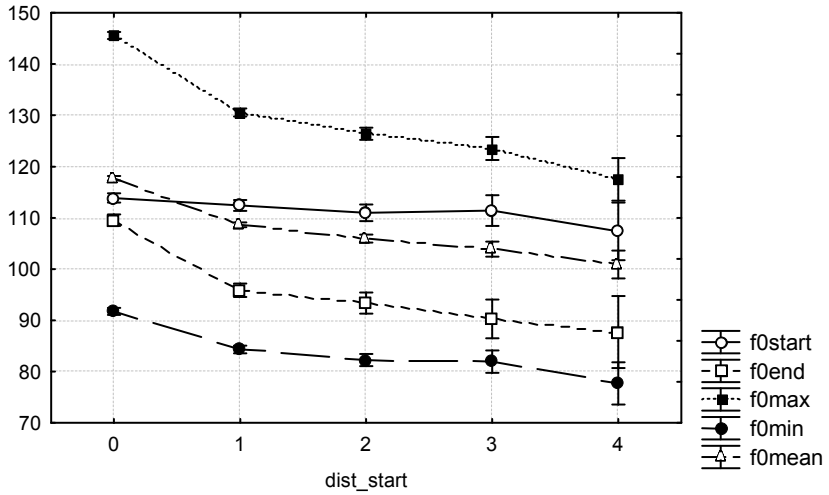


Figure 34: Representation of pitch variation in dist_start dataset.

variable	f0start	f0end	f0max	f0min	f0mean	S.D.
F test	3,74	93,19	352,82	84,43	311,43	29,91
p			<0.01			

Table 10: Multivariate ANOVA results: the effect of dist_start grouping on pitch range.

In the graphical representation of pitch range in the dist_start dataset depicted in the Figure 34 it can be seen that IP-initial ips have higher pitch range than non-initial ips. This is indicated by significantly higher average values of f0max, f0min and meanf0 in initial phrases compared to other phrases. In general, the distribution of average values of f0 variables in the

Figure 34 shows that overall pitch range gets lower as the phrase distance from IP start increases, which is indicated by the downward trend in f0max, meanf0 and f0min.

ANOVA results prove that the dist_start classification causes statistically significant variation in all f0 variables describing pitch range and the strongest effect is found for f0max (F=352,82) and meanf0 (F=311,43). These effects are confirmed in Scheffe's test with the exception of that found for f0start variable.

It can be concluded that the greatest difference in the pitch range is that between initial vs. non-initial ips: they differ significantly from non-initial ips with respect to all f0 parameters. Pitch range between non-initial ips differs to a lesser degree: the only statistically significant variation is found in f0max and meanf0 between ips of a second position in IP (dist_start=1) vs. other ips and in f0max between ips of third (dist_start=2) and fifth (dist_start=5) position in IP.

To sum up, the effect of dist_start categorization of phrases affects significantly scaling of f0 targets describing pitch range and suggest that on the basis of similar pitch range features ips of initial position in IP should be grouped together.

2. Analysis and results for dist_end dataset.

In the dist_end categorization ips were grouped according to their distance from IP end. Like in the previous analysis based on dist_start classification, the least numerous ip classes (5 and 6) were not taken into account. In the graphical representation of pitch range in dist_end dataset in it can be seen that IP-final ips have much lower overall pitch level and final pitch than non-final ips. This is indicated by significantly lower average values of f0max, f0min, meanf0 and ipf0end in final than in non-final phrases. It can also be observed that the distribution of f0 targets is similar between non-final phrases. Figure 35 illustrates distribution of pitch targets (f0max, f0min, meanf0, f0end and f0start) in various ip classes: 0 marks IP-final ip, 1 - the second ip from IP end, 2 - the third ip from IP end, etc.

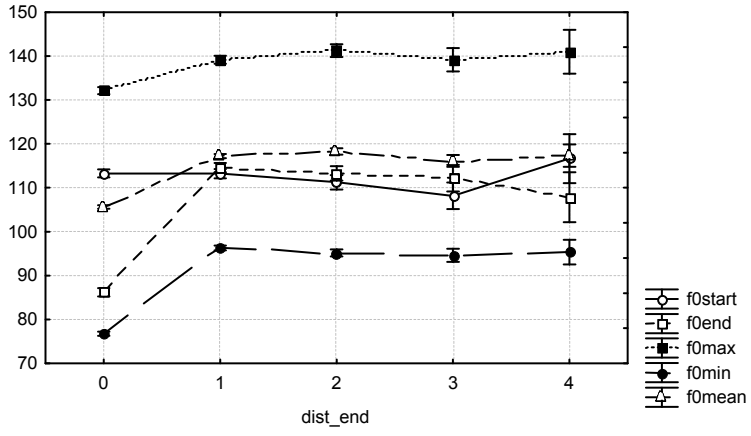


Figure 35: Representation of pitch variation in dist_end dataset.

Table 11 shows ANOVA results based on the dist_end dataset.

variable	f0start	f0end	f0max	f0min	f0mean	S.D.
F test	4,01	472,66	47,29	944,44	430,75	102,19
p	<0.01					

Table 11: Multivariate ANOVA results: the effect of dist_end grouping on pitch range.

ANOVA results prove that dist_end grouping causes statistically significant variation in all f0 parameters and the strongest effect is found for f0min (F=944,44), ipf0end (F=472,66) and meanf0 (F=430,75). These effects are confirmed in Scheffe's test which additionally shows that the greatest difference in pitch range is between final vs. non-final ips. This is indicated by statistically significant difference in all f0 variables describing pitch range as well as f0end variable between final vs. other ips. Scheffe's test results confirm also the observation that pitch range between non-final ips differs to a lesser extent than between final vs. non-final ips. The only statistically significant difference between various non-final ips is in S.D. (between ips of the second and third position from IP end) and in f0start (between ip of the second and fourth position from IP end).

To sum up, on the basis of the distribution of f0 data as illustrated in the Figure 35 and results of ANOVA and Scheffe's test it can be concluded that IP-final ips differ significantly from non-final ips in respect to pitch range and these differences are statistically significant. These results indicate that phrases of a final position in IP have some special status and should be grouped together.

3. Analysis and results for initial/final/single dataset.

The classification scheme adopted in the analyses in previous sections made it possible to investigate variation in pitch between IP-initial ips vs. non-initial ips and IP-final ips vs. non-final ips. At the same time, in the two datasets major phrases (IPs) consisting of one ip (from now on referred to as *single phrases*) were grouped with initial and final ips respectively. The results of statistical analyses have shown that initial and final ips differ significantly from ips of other position and that non-initial and non-final ips have similar pitch range. At the moment, the goal is to investigate to which extent these results are influenced by a) grouping of single phrases with initial/final ips and b) grouping of non-initial/non-final ips together with phrases that are actually final/initial in an IP. Now, an attempt will be made to determine whether single phrases should be treated like initial or final phrases or maybe constitute a separate class (hypothesis 2). For this purpose in the current analysis a distinction will be drawn between single, initial and final phrases.

The graphical representation of distribution of pitch targets (f0max, f0mean, f0min, f0start and f0end) in single/initial/final dataset is given in Figure 36.

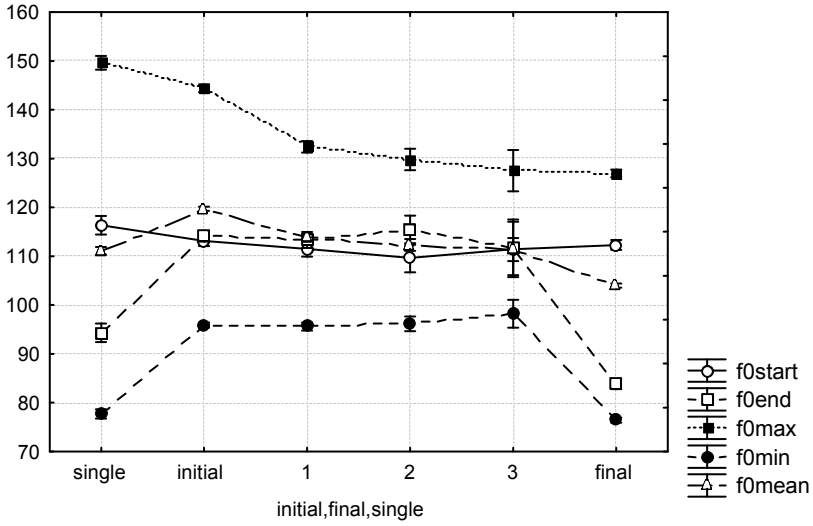


Figure 36: Representation of pitch variation in initial/final/single dataset.

variable	f0start	f0end	f0max	f0min	f0mean	S.D.
F test	408,79	286,35	761,32	528,01	199,28	408,79
p	<0.01					

Table 12: Multivariate ANOVA results: the effect of initial/final/single grouping on pitch range.

The distribution of f0 targets in Figure 36 shows that the categorization of ips adopted in the current analysis has significant effect on pitch range and it is statistically significant as showed by ANOVA results Table 12. In general, single ips are characterized by the widest pitch range (span) which is indicated by very high average f0max and very low average f0min. It can be seen that f0mean of single ips is similar to that of medial ips. Scheffe's test results confirm this observation (no statistically significant difference), which suggests that the difference between single vs. medial ips is in span rather than overall pitch level. On one hand, single phrases are characterized by significantly lower final (f0end) and minimum pitch (f0min) in comparison to other (non-final) ips which makes them similar to final ips. On the other hand they have very high f0max, which is the feature of initial ips.

As regards initial and final ips it can be seen that they differ with respect to overall pitch level which is much lower in final than in initial and medial ips. The distribution of average values of f0max, meanf0 and f0min indicates the lowest overall pitch level of final ips and the highest level of IP-initial ips which proves that *at the end of an intonational phrase there occurs a pitch range reset* to default starting pitch at the start of the next IP. This is an evidence that *major and minor intonational phrases constitute different phrasing levels* (Hypothesis 1).

In general ANOVA and Scheffe's test results prove that the classification of ips adopted in the current analysis is based on statistically significant variation in pitch: differences in pitch range can be observed between phrases of initial, final, medial position in IP and single phrases. At the same time, medial ips (i.e. ips of the second, third and fourth position from IP start

marked with digits 1, 2, 3) are characterized by similar pitch range irrespective of their distance from IP start: this observation is confirmed in Scheffe's test which show no statistically significant differences in pitch variation between them. These results confirm the Hypothesis 2 according to which minor phrases should be classified relative to their position in IP.

For the final definition of the prosodic structure representation within which pitch range can be controlled the alignment of ips of the same position in IPs of a different length needs to be analyzed, which is the subject of the next section.

5.1.5. Analysis of phrases of a different length

The goal of the analyses presented in this section is to answer the question whether IP length has influence on pitch range of ips. The other goal is to confirm the findings of analyses based on classification of ips with respect to their position in IP, namely that:

a) major and minor phrase constitute different phrasing levels, because the former is the domain of pitch range reset and the latter is not (Hypothesis 1)
b) distinction should be made between IPs depending on whether they *consist of one or more ips* (single vs. complex intonational phrases (Hypothesis 2)
c) complex IPs can be modeled as sequences of *initial, medial and final* ips (Hypothesis 2)

In order to finally confirm these findings and test the Hypothesis 3 differences between ips of the same position in IPs of a different length measured in the number of ips are analyzed. It is expected that on the one hand IP-initial phrases exhibit different properties than phrases of initial in IPs consisting of a single ip. On the other hand, initial, medial and final phrases in complex IPs (i.e., IPs consisting of more than one ip) should have the same pitch range characteristics irrespective of IP length.

For the purpose of the analyses four datasets were created: in each dataset ips of the same position in IP of a different length were collected. Phrase position is indicated by digits 0-3 which expresses the distance from IP start (as it was done in the dist_start classification). IP length is marked by letters from a for single phrases to e for IPs including 4 ips. Phrases grouped in classes 1b (second in an IP of including 2 ips), 2c, 3d were IP-final. Altogether, 2457 ips were analyzed: 1132 from dataset 0 (i.e. including IP-initial ips), 873 from dataset 1, 353 from dataset 2 and 99 from dataset 3.

1. Results for dataset 0.

In this dataset ips of initial position in the intonational phrases of different length (1-5) are compared. This grouping makes it possible to compare pitch range of initial vs. single phrases. It is expected that the analyses based on this dataset prove that single ips have different pitch range than initial ips and therefore, they should be distinguished from among other phrases and constitute a phrase class on their own. The other goal is to show that ips of IP-initial position in complex IPs have similar pitch range irrespective of IP-length.

Figure 37 illustrates pitch range of phrases analyzed in the dataset 0; single phrases are marked as 0a (0 stands for IP-initial position and a indicates IP length=1).

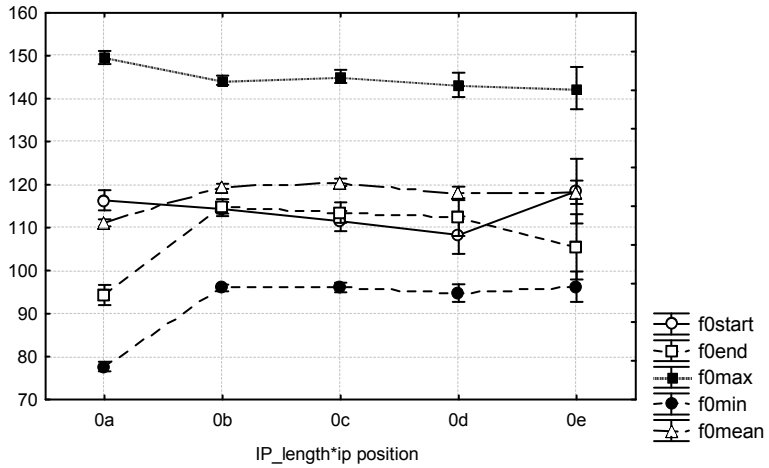


Figure 37: Representation of pitch variation in the dataset 0.

variable	f0start	f0end	f0max	f0min	f0mean	S.D.
F test	4,08	53,84	9,09	202,80	78,91	66,51
p			<0.01			

Table 13: Multivariate ANOVA results: dataset 0.

ANOVA results depicted in Table 13 above indicate statistically significant variation in all f0 variables representing pitch range as well as the start and final pitch (f0start, f0end). Scheffe's test results confirm the results of previous analyses and show significant difference in pitch range between single (0a) and initial phrases, which once more proves that they should be regarded as different phrase types. At the same time it can be observed that initial ips in complex phrases have nearly the same pitch range irrespective of IP length (this is also indicated in Scheffe's tests). This finding confirms the hypothesis that ips of initial position in complex IPs should be grouped together. The fact that the bottom, middle and the top level of pitch range are constant between initial phrases shows that *under roughly similar conditions speakers control pitch level* (see Ladd 1996:66).

2. Results for dataset 1, 2 and 3

Datasets 1, 2 and 3 include phrases of the second, third and fourth position in IPs of different length (2-5). This grouping makes it possible to compare pitch range of final vs. medial phrases. It is expected that the analyses based on this dataset prove that final ips and medial ips have different pitch range characteristics and therefore represent different sub-phrase types. The other goal is to show that ips of the same position in IP (medial or final) have similar pitch range irrespective of IP length. The figures 38-40 illustrate pitch range of ips analyzed in the three datasets; final phrases are marked as 1b, 2c and 3d.

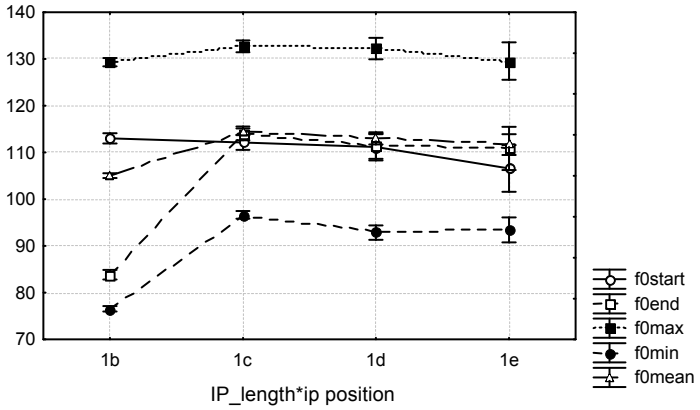


Figure 38: Representation of pitch variation in the dataset 1.

variable	f0start	f0end	f0max	f0min	f0mean	S.D.
F test	2,42	437,64	7,13	560,37	175,46	162,89
p	<0.01					

Table 14: Multivariate ANOVA results: dataset 1.

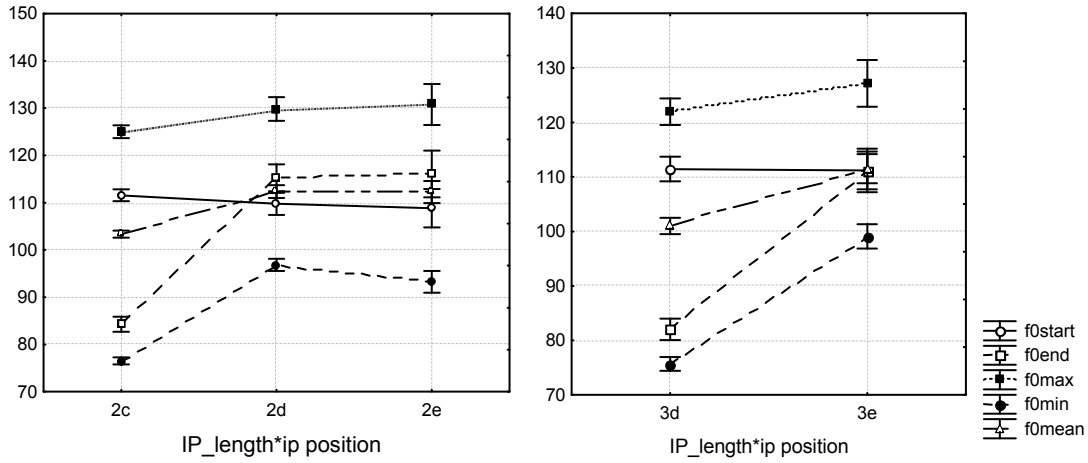


Figure 39: Representation of pitch variation in the dataset 2 (left) and 3 (right).

variable	F test	p
f0start	0,01	0.93
f0end	210,43	
f0max	4,31	
f0min	331,94	<0.01
f0mean	46,03	
S.D.	67,44	

variable	F test	p
f0start	1,48	0,23
f0end	220,26	
f0max	7,54	
f0min	405,55	<0.01
f0mean	80,16	
S.D.	70,11	

Table 15: Multivariate ANOVA results for dataset 2 (left table) and 3 (right table).

ANOVA results show that the grouping of ips adopted in the dataset 1 (Table 14) has the strongest effect on f0min (F=560,37) and f0end (F=437,64); f0mean and S.D. are less affected (F=175,46 and 162,89 respectively). ANOVA based on datasets 2 and 3 (

Table 15) show the same effects. It can be seen (Figure 38 & Figure 39) that at the end of the intonational phrase (f0end) there occurs a low pitch (f0min) that reaches almost the bottom of the range (so called *final fall*). As in the previous analyses only very small and insignificant differences in pitch range between medial ips can be observed, which proves that they should be distinguished from among other phrases and regarded as a different ip type in the same way as initial, single and final ips.

The results of Scheffe's test (much the same for the three datasets) confirm the difference in the pitch range between final and medial ips found in the previous analyses. In general, medial ips are characterized by significantly higher overall pitch level and narrower span than final ips.

The comparison of the distribution of f0 data describing pitch range of final ips in different datasets shows that they have the same average f0max, f0mean and f0min. It means that irrespective of extrinsic factors such as IP length, IP-final ips have the same pitch range, which is a reason for grouping them together. The graphical representation of pitch range in datasets 1, 2 and 3 also shows that within the group of medial and final phrases the three target levels described by f0max, meanf0 and f0min variables are consistently aligned irrespective of IP length, which is again the evidence that *under roughly similar conditions speakers control pitch level* (e.g. Ladd 1996).

5.1.6. Conclusions

The analyses presented in the previous sections proved that *prosodic structure is an extrinsic factor* in the variation of pitch, because it *modifies the 'vertical' scale of pitch*. This is manifested in raising and lowering of the top, middle and bottom levels of the range between phrases of different position in IP and IPs of a different structure.

It was shown that on the basis of significant differences in pitch range features of phrases a distinction should be made between minor phrases of *initial, medial* and *final* position in the major phrase which constitute *different sub-phrase types* in the structure of intonational phrase (Hypothesis 2). The other distinction that should be drawn is between single and complex IPs. IPs which include only one minor phrase have significantly different pitch range from initial or final ips, and therefore should be regarded as a different phrase type, namely *single phrases* (Hypothesis 2). They are characterized by the widest span from among all phrase types and have overall pitch level similar to that of medial ips.

Initial and final phrases are also characterized by wide span: it is narrower than span of single phrases, but wider than span of medial ips. From the beginning to the end of an intonational phrase a gradual lowering of overall pitch level can be observed: this is manifested by the highest average values of f0max, f0mean and f0min (representing the top, middle and bottom of pitch range respectively) on initial ips and the lowest average values of these parameters on final ips. The difference in pitch range between initial and final ips reflects the pitch range reset which is the feature of major intonational phrases and distinguishes them from minor phrases, and proves that *ips and IPs constitute different phrasing levels* (Hypothesis 1).

The analyses in which ips of the same position but belonging to IPs of a different length were investigated showed that in complex intonational phrases IP length has generally no effect on pitch range of ips (Hypothesis 3), but a distinction should be made between phrases which consist of a single ip vs. more ips. The results of ANOVA and Scheffe's tests indicated significant differences in pitch range between single, initial, medial and final phrases and at the same time showed that phrases of the same type have similar pitch range (i.e. f0max, f0mean, f0min and S.D.) irrespective of IP length (Hypothesis 3). This also proves that *under roughly similar conditions speakers control pitch level.* Finally, the categorization of phrases proposed in this thesis into major (IP) and minor (ip) on the one hand and initial, medial, final and single on the other was shown to group together prosodic structure constituents of similar pitch range and this confirms Hypothesis 1, namely that pitch range can be controlled within a two-level phrasing system.

In view of these results it is proposed that phrases should be classified into *initial, medial, final* and *single*. The information concerning prosodic structure is expected to be helpful in the estimation of pitch targets for contour generation (see: Hypothesis 1c), which will be confirmed in the analysis presented in Chapter 7. Apart from that, the information concerning phrase type will be used in pitch normalization across speakers (sec. 7.3).

5.2. Surface phonological description

In this section a description of intonation on the surface phonological level is proposed. The main theoretical assumption is that intonational tunes are represented and analyzed as *strings of distinctive elements also referred to as intonational events which are linked with the elements of the segmental string*. Two types of elements are distinguished, namely *pitch accents* and *boundary tones.*

Pitch accents are associated with metrically strong, stressed syllables and are realized by distinctive pitch movements. As regards boundary tones in a strict sense they are defined as single distinctive tones (H or L) associated with the end of intonational phrases (Ladd 1996:80), but in this thesis they are considered as distinctive, non-prominence lending pitch movements occurring at phrase boundaries. The choice of the pitch accents/boundary tones of a particular type determines the interpretation of the meaning conveyed by the tune.

In the framework defined here the stretches of contour between the events (sometimes referred to as connections e.g. Taylor 2000) are irrelevant, because they do not contribute to conveying of intonational meaning.

Tunes are made up of one or more pitch accents and an obligatory boundary tone. They have specific structure in which constituents such as *pre-head, head, nucleus and tail* can be distinguished.

Pre-head is the stretch of unstressed syllables preceding head - the part of the contour which starts with a major stressed syllable. Nucleus is the only obligatory constituent and it "occurs on the most prominent stressed syllable which is normally also the last stressed syllable" (Ladd 1996:209). The nuclear part of the tune is the most important one: the choice of a specific combination of a pitch accent and boundary tone which make up a *nuclear tone* (also: *nuclear melody*) determines sentence modality.

The shape of tail which is defined as "the stretch of contour following the nuclear syllable" (op.cit.:209) determines the post-nuclear intonation.

It is assumed that accents of a given types may have various structural roles in the tune, but as it will be seen in sec. 5.2.2 where the distribution and structural roles of accents are discussed some accents occur in the prenuclear or nuclear position more often then others. There are also no restrictions concerning the possible combinations of pitch accents and pitch accents and boundary tones, but again, some combinations are used more often then others.

The surface-phonological description of pitch accents and boundary tones proposed in this thesis is in terms of discrete distinctive categories and encodes not only melodic, but also functional aspects of intonation. As mentioned at the beginning of this chapter the description is based on the prosody annotation system used originally in the unit selection corpus for the Polish module of BOSS TTS system (see sec. 4.2).

In that system an inventory of 7 accent types was defined which were given symbolic representation in the current study. The accents were distinguished on the basis of melodic properties similar to those distinguished in the IPO system.

In the system used for prosodic annotation of Polish unit selection corpus a single label was used to mark the strength of prosodic break and direction of the distinctive pitch movement associated with the phrase boundary. Consequently, a distinction was drawn between seven types of phrase boundaries and they were given a symbolic representation. Two boundary tones are associated with phrase-initial boundaries and the other five with phrase-final boundaries.

The usefulness of the surface phonological description of boundary tones proposed in the current study for generation of intonation contours (Hypothesis 1 & Hypothesis 1b) will be investigated in Chapter 7.

5.2.1. Pitch accents

On the basis of perceptual and acoustic analyses seven types of accents have been distinguished and used in the prosodic labeling of the Polish unit selection corpus.

Two accents are realized only by duration: the syllables with which they are associated are characterized by "flat" f0 contour (a *plateau*) and increased duration in comparison to unaccented syllables. Other accent types which include two rising, two falling and one rising-falling accent can be considered as *"proper" pitch accents*, because they are realized by pitch variation (pitch movements). They are distinguished on the basis of the following perceptually significant features:

a) direction of the pitch movement (rise vs. fall) on the accented and post-accentual vowel
b) amplitude of the movement
c) position (timing) of the peak/minimum relative to accented vowel/syllable boundaries
d) the amount of pitch variation on the accented and post-accentual vowel

Figures 40-42 show prototypical accents in each group. From top to bottom the waveform, spectrogram, transcription/segmentation panel and pitch contour panel are depicted. The position of accented vowel is marked in color on the transcription panel. Additionally, the horizontal lines at the bottom of the pitch contour panel mark the position of the accented and post-accentual vowels. The symbols in brackets are the labels used in the pitch accent type

annotation in the Polish unit selection corpus. It is assumed that for the purpose of intonation description on the surface phonological level they were replace with ToBI-like labels, where H indicates pitch target located in the top of or high in the current pitch range and L indicates targets located at the bottom of or low in the range; asterisk indicates the alignment of the target with the accented syllable. As regards the level accents it is proposed to label them as LI (level, interval) and LD (level, duration). There seem to be some inconsistency in the classification of the accents into rising and falling: the rising accent labeled as LH* illustrated in Figure 41 actually involves a fall in the overall pitch level between the accented and post-accentual vowels. The reason why it is described as rising is that unlike other accents is involves a distinctive rise in pitch realized on the accented vowel.

In Figure 40 two types of falling pitch accents are illustrated: H*L and HL*.

a) H*L (")

b) HL* (&)

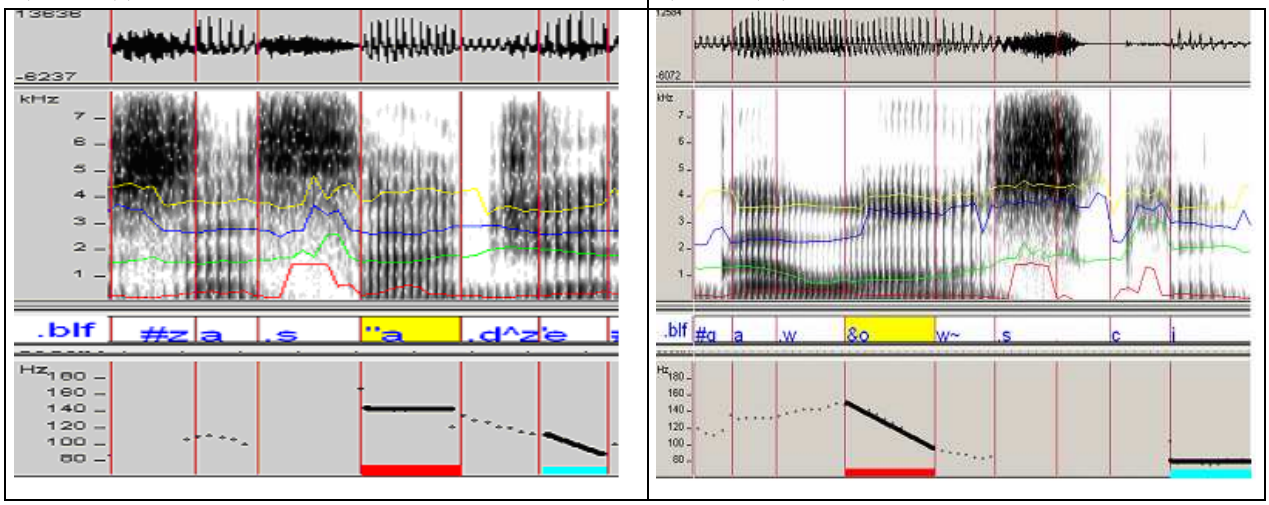


Figure 40 (adapted from Demenko & Wagner 2007): Examples of falling accents.

The H*L accent (labeled originally as ",
Figure 40 a) has the following distinguishing features:

a) there occurs a fall in the overall pitch level between the accented and post-accentual vowels
b) a falling pitch movement is realized on the post-accentual vowel, whereas on the accented vowel there occurs little variation in pitch or a plateau is realized
c) the fall can be realized as a "jump" between two pitch levels/plateaus on the accented and post-accentual vowels
d) f0 peak (local maximum) is located on the accented vowel/towards the end of the accented syllable, local minimum occurs on the post-accentual syllable

The other falling accent HL* (labeled originally as &,
Figure 40b) has the following features:

a) there occurs a fall in the overall pitch level between the accented and post-accentual vowels
b) a falling pitch movement is realized on the accented vowel, whereas on the post-accentual vowel there occurs little variation in pitch or a plateau is realized

c) f0 peak (local maximum) is located at the beginning/towards the beginning of the accented vowel/syllable, local minimum occurs on the post-accentual syllable

Figure 41 illustrates prototypical rising pitch accents: L*H and LH*.

a) L*H (%) b) LH* (`)

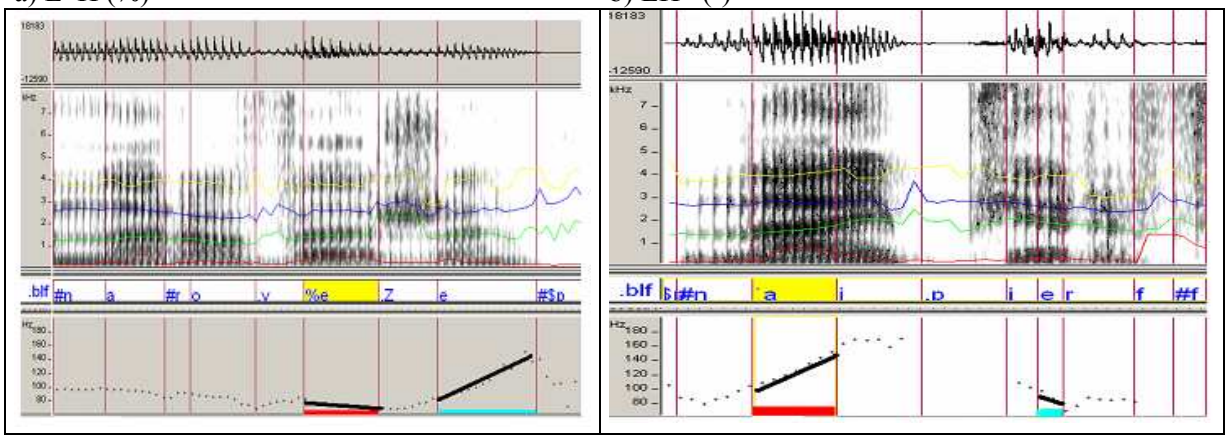


Figure 41 (adapted from Demenko & Wagner 2007): Prototypical rising pitch accents.

The rising L*H accent (labeled originally as %, Figure 41a) has the following features:

a) there occurs a rise in the overall pitch level between the accented and post-accentual vowels
b) a rising pitch movement is realized on the post-accentual vowel, whereas on the accented vowel there occurs little variation in pitch or a plateau is realized
c) the rise can be realized as a "jump" between two pitch levels/plateaus on the accented and post-accentual vowels
d) local f0 minimum is located on the accented vowel/towards the end of the accented syllable
e) f0 peak (local maximum) is located on the accented vowel/towards the end of the accented syllable, local maximum occurs on the post-accentual syllable

The other rising accent LH* (originally `, Figure 41b) is distinguished by the following features:

a) there occurs a fall in the overall pitch level between the accented and post-accentual vowels
b) a rising pitch movement is realized on the accented vowel, whereas on the post-accentual vowel there occurs little variation (fall) in pitch or a plateau is realized
a) f0 peak (local maximum) is located on the accented vowel/towards the end of the accented syllable, local minimum occurs on the post-accentual syllable

The class of rising-falling pitch accents LH*L (labeled originally as |, Figure 42a) is distinguished on the basis of the following features:

a) there occurs a fall in the overall pitch level between the accented and post-accentual vowels
b) both rising and falling pitch movements are realized on the accented vowel; on the post-accentual vowel there occurs little variation (fall) in pitch or a plateau is realized
b) f0 peak (local maximum) is located somewhere in the middle of the accented vowel, while local minimum occurs on the post-accentual syllable

The level accents have the following characteristics:

a) LI accent (labeled originally as *, Figure 42b) is realized by an interval (either rise/fall) between the overall pitch level on the pre-accentual vs. accented/post-accentual vowels; on the three vowels and corresponding syllables there is very little or no variation in pitch at all
b) LD accent (labeled originally as <) is realized solely by increased duration as compared to unaccented syllables; no variation in pitch is observed on the pre-accentual/accented/post-accentual vowel/syllable and they have the same overall pitch level

Figure 42 gives examples of prototypical accents LH*L and LI.

a) LH*L(|) b) LI (*)

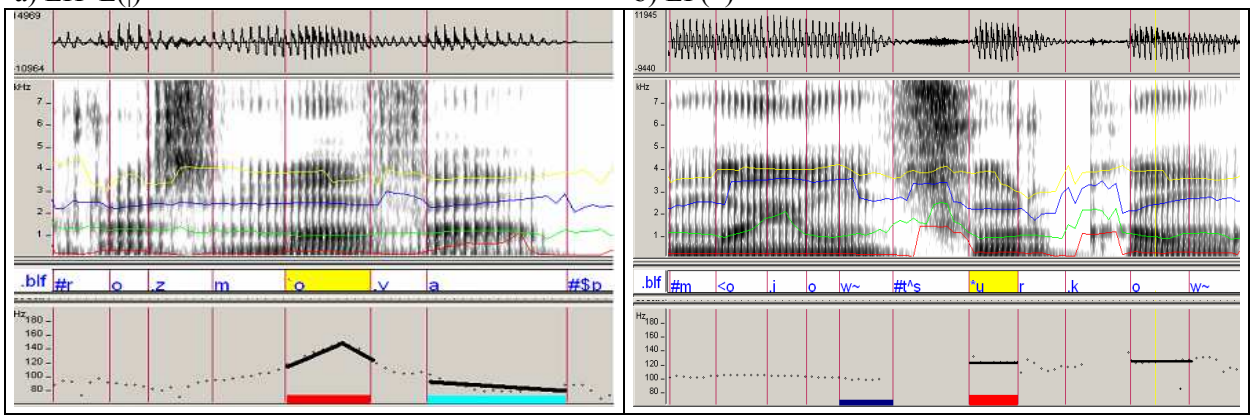


Figure 42 (adapted from Demenko & Wagner 2007): A prototypical rising-falling pitch accent H*L (a) and level accent LI (b)

One important remark that has to be made concerning the inventory of accents that will be used in further analyses.

Prosodic prominence involves two different phonetic features - pitch accents and stress. As indicated in a number of studies (Jassem 1961, Rietveld & Gussenhoven 1985, Demenko 1999, Tamburini 2006, see sec. 6.1.1) pitch variation is the primary acoustic correlate of pitch accents, while duration, intensity and spectral emphasis can be regarded as secondary features. Stress, on the contrary, exhibits a strong correlation with syllable and nuclei duration (e.g. Ortega-Llebaria, Prieto & del Mar Varnell 2007). In view of these facts, the types of rising, falling and rising-falling accents described in this section can be regarded as *pitch accents*. What is unclear is the role of the two *level accents*: LD and LI. Their main feature is increased duration and they do not involve any significant pitch variation. These accents were identified and marked by labelers, because they are associated with perceptually prominent syllables, but this prominence seems to be stress-related rather then accent-related. Therefore, it seems justified to exclude the LI and LD accents from the surface phonological description of pitch accents. Consequently, the inventory of pitch accents used in the analyses dedicated to detection of accent position and recognition of accent type (sec. 6.4 & 6.5) contains: H*L, HL*, LH*L, L*H and LH* accents.

5.2.2. Distribution and structural roles of accents

The goal of the discussion presented in this section is to provide information concerning the distribution and structural roles of the pitch accents distinguished in the surface phonological description. Some conclusions can be drawn on the basis of the data given in the Table 16.

accent type	overall count	overall percentage	count (nuclear position)	percentage (nuclear position)
H*L	3175	38,42	310	18,20
L*H	1237	14,97	348	20,43
LH*	2271	27,48	269	17,79
HL*	1334	16,14	707	42,04
LH*L	246	2,98	60	1,53
total:	8263	100	1694	100

Table 16: General distribution and frequency in the nuclear position of different pitch accent types.

In the second and third column overall frequency of different pitch accent types is given calculated on the two corpora used in the analyses presented in the following chapters. The two next columns show frequency of different pitch accent types in the nuclear position.

As regards overall frequency, it can be seen that H*L accents have the highest frequency from among all pitch accent types. This can be attributed to the fact, there are significantly more prenuclear (6569) then nuclear accents (1694) in the database and H*L accents occur almost always (90% of cases) in the prenuclear position.

The second most frequent accent type is LH*. Like H*L it occurs significantly more often in the prenuclear than nuclear position and this concerns almost 87% instances of LH* accents.

HL* accents have similar overall frequency with L*H accents, but they differ as regards the structural roles in tunes that they play most often. HL* accents occur much often in the nuclear position (50% instances), whereas L*H accents occur most often in the prenuclear position. However, in comparison to H*L and LH* the L*H accents occur in the position of the nucleus relatively often (28% of cases).

LH*L is clearly the least frequent accent type. Instances of LH*L accents constitute only 2.98% of all the accents. It occurs in the nuclear position with a similar frequency to LH* accents (24% of cases).

In view of this discussion it can be concluded that accents of a given type may have various structural roles in the tune. However, it can be observed that accents of a given type occur in the nuclear (e.g. HL* and H*L) or prenuclear (e.g. LH* and H*L) position more often then other accents.

As mentioned before, it is assumed that there are no restrictions concerning possible combinations of pitch accents and boundary tones which make up nuclear tones, but at the same time it is expected that some combinations occur more often then others. In order to investigate this issue frequency of various nuclear tones was analyzed: It is depicted in the Figure 43.

The basic declarative intonation is almost always realized by combination of HL* accent and subsequent 5,. boundary tone. Only small percentage (11.48%) of this kind of contour is realized by the nuclear tone consisting of the other falling pitch accent H*L followed by the 5,. boundary tone.

As regards the interrogative mode it is almost always (in 94% of cases) realized by the nuclear tone consisting of L*H pitch accent and subsequent 5,? boundary tone. There are only six instances of questions ending in the combination of the other rising pitch accent LH* and 5,? edge tone.

In the two corpora used in the experiments presented in this thesis the nuclear tune in exclamations is always realized by the combination of HL* pitch accent and subsequent 5,! boundary tone.

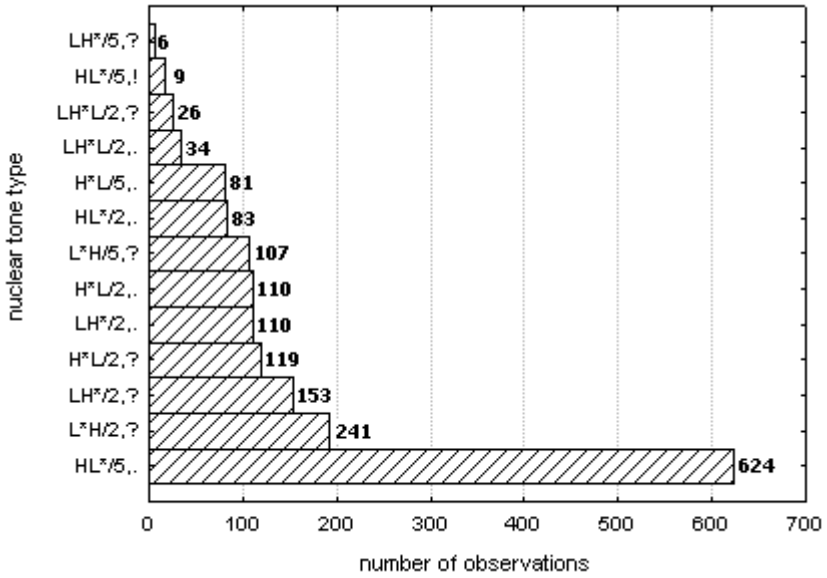


Figure 43: Frequency of different types of nuclear tones.

This summary shows that not all combinations of pitch accents and boundary tones are allowed, which contradicts the previous assumption.

As regards nuclear tones in minor phrases a greater diversity in the possible combinations of pitch accents and boundary tones can be observed.

In continuation phrases the nuclear tune is most often realized by combination of some rising pitch accent and 2,? boundary. There are 241 instances of nuclear tones consisting of L*H accent followed by 2,? boundary, 153 instances of nuclear tones consisting of the other rising pitch accent, namely LH* and subsequent 2,? boundary tone. The falling H*L accent occurs less often in the nuclear position in continuation phrases in comparison the two rising pitch accents (119 instances) and finally, there are 26 instances of nuclear tones consisting of LH*L accents followed by 2,? boundary tone.

The nuclear contour in minor phrases inside complex sentences which end with cadence are always realized by combination of all pitch accent types except for the L*H accent and 2,. boundary tone. LH* and H*L accents occur in this position with the same frequency (110 instances each). LH*L accents play this role less often: there 34 instances of nuclear tone in this type of tunes consisting of this particular accent type followed by 2,. boundary tone. It should

also be noted that LH*L accent type is the only one which never occurs in the nuclear position on major intonation phrases.

The discussion presented here shows that not all possible combinations of pitch accents and boundary tones occur in the speech corpora, which may have serious implications for the definition of intonational grammar, but this is not the goal of this thesis.

The purpose of the analysis presented here was to show that the surface phonological description of intonation proposed in this thesis encodes not only melodic, but also functional aspects of intonation. The current analysis was rather limited, but some preliminary conclusions can be drawn. It could be seen that various accent types can have various roles in the tunes and that basic intonational meanings are conveyed by different types of nuclear tones.

5.2.1. Boundary tones

The annotation of phrase-boundary phenomena in the Polish unit selection corpus involves simultaneous marking of the type of pitch movement occurring at the phrase end as well as the strength of phrase break. The former are indicated by punctuation marks signaling the direction and range of the movement, whereas the latter are indicated by digits 2 and 5 which signal minor and major phrase boundaries respectively.

This kind of description can have some advantages from the point of view of automatic prosody labeling, because it makes it possible to perform two tasks (i.e., boundary tone type and break strength) simultaneously with a single model.

On the basis of perceptual and acoustic analyses seven types of boundary tones were distinguished for the purpose of prosodic annotation of Polish unit selection corpus: two of them are associated with phrase-initial boundaries and the other five with phrase-final boundaries.

Summary of the guidelines specified for phrase boundary labeling is given below.

BOUNDARY	DEFINITION, CONTEXT, USE
-5,.	Intonation on the first accented word in sentence with a falling (H*L, HL*), level (LI, LD) or rising-falling accent (LH*L). In most cases it is used to mark initial boundaries of statements or wh questions.
-5,?	Intonation on the first accented word in sentence with a rising accent (LH*, L*H). It marks initial boundaries of minor phrases inside complex sentences.
5,!	Intonation on the last word in sentence with a falling accent (HL*, H*L). In most cases it is used to mark the boundary of exclamatory sentences.
5,?	Intonation on the last word in sentence with a rising accent (LH*, L*H). In most cases it is used to mark the boundary of yes-no questions.
5,.	Intonation on the last word in sentence with a falling (H*L, HL*)or level (LI, LD) accent. In most cases it is used to mark boundaries of declarative sentences or wh questions.
2,?	Intonation on the last word in the phrase with a rising accent (LH*, L*H). In most cases it is used to mark boundaries of for continuation phrases.
2,.	Intonation on the last word in a phrase with a falling (H*L, HL*) or level (LI, LD) accent. In most cases it is used to mark boundaries of minor phrases inside complex sentences.

Figure 44: Labels used for intonation annotation at phrase level in Polish unit selection corpus

For the purpose of the current study only the initial boundaries are excluded from the boundary tone inventory. As opposed to final boundary tones they can be considered as less significant to intonation analysis and modeling, because they do not contribute to conveying of intonational messages. the resulting inventory consists of phrase types of boundaries which are re-defined in the following way:

a) ***2,.*** ***5,.*** and ***5,!*** are examples of *falling boundary tones*. They are realized by a falling pitch movement from a higher target level on the penultimate/ultimate syllable in the phrase to a lower f0 target associated with the phrase boundary. The three boundary types differ with respect to the amplitude of the fall (***5,!*** has the greatest fall amplitude and ***5,.*** the smallest) and scaling of the f0 targets (in case of ***5,.*** they are positioned significantly lower in the speaker's range).

b) ***2,?*** and ***5,?*** are examples of *rising boundary tones*. They are realized by a rising pitch movement from a lower target level on the penultimate/ultimate syllable in the phrase to a higher f0 target associated with the phrase boundary. They differ with respect to the amplitude of the rise which is greater for ***5,?*** than ***2,?*** boundary and scaling of the targets f0 targets (the rise starts higher in the range and ends lower in case of 2,? latter boundary tone).

The description of boundary phenomena proposed here is based on two different types of information which include the strength of prosodic break and type of pitch movement occurring at the end of the phrase. Yet, it seems that the term *boundary tone* (interchangeably with *phrase boundary type*) can be used to refer to these phenomena without introducing any serious terminological fuss.

Figure 45 and Figure 46 illustrate prototypical boundary tones in each class. From top to bottom the waveform, pitch contour and transcription/segmentation panels are shown.

a) 2,? b) 5,?

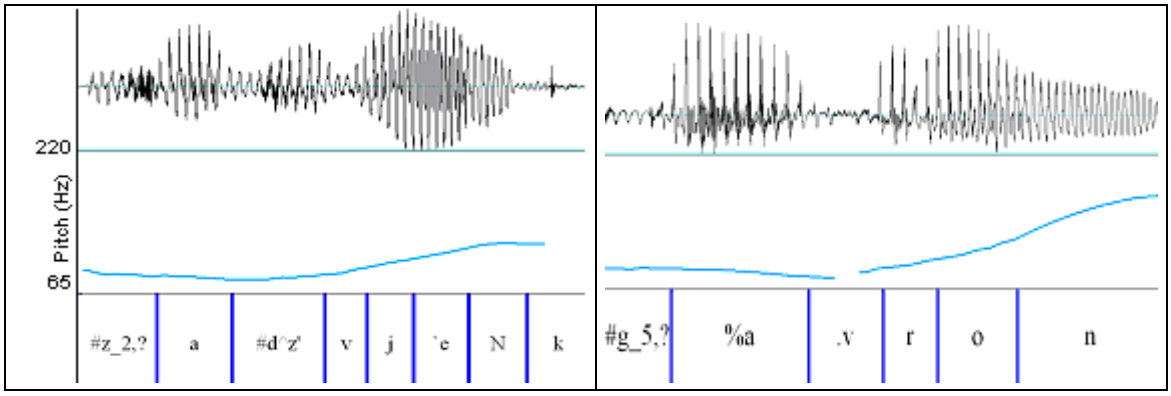
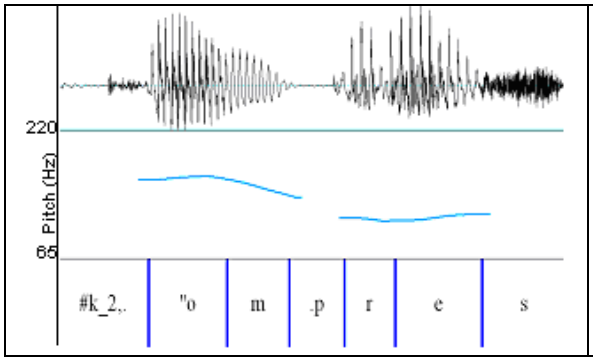
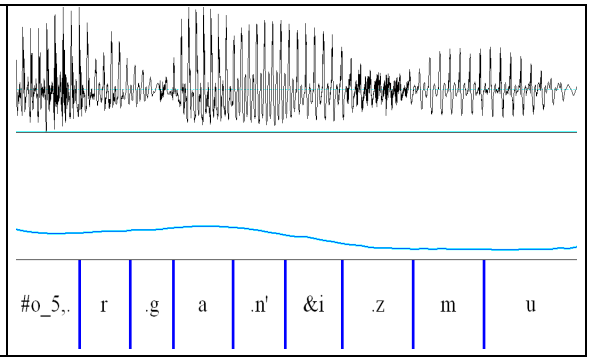


Figure 45: Prototypical rising boundary tones 2,? (a) and 5,? (b).

a) 2,.



b) 5,.



c) 5,!

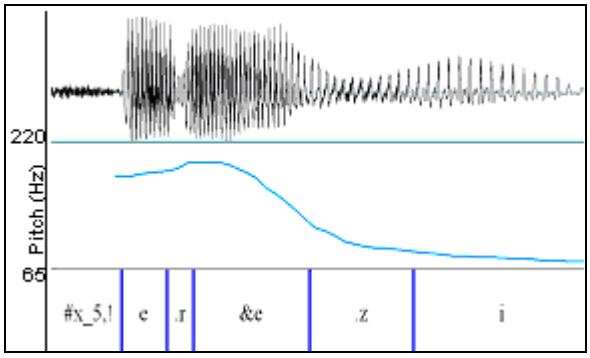


Figure 46 (a-c): Prototypical falling boundary tones.

5.3. Phonetic description

The phonetic description of intonation proposed in this thesis is defined in a number of statistical analyses in which the contribution of a number of acoustic features to the distinction between various pitch accent/boundary tone types was investigated. The resulting description is compact as it uses only small feature vectors to express the fine differences between different types of intonational events and can be easily derived from utterance's acoustics.

On the basis of the phonetic description it should be possible to extract the higher-level surface phonological description of intonation. This issue will be investigated in Chapter 6 where a number of models will be designed to perform this task. The results are expected to prove Hypothesis 1a according to which the phonetic description of intonation proposed in this thesis provides information which is significant to the detection and classification of the elements of intonational tunes: pitch accents and boundary tones.

5.3.2. Pitch accents

The phonetic description of pitch accents will be used in the automatic recognition of pitch accent type (see sec. 6.5). In order to define a description which is adequate for this task a

number of ANOVA and discriminant function analyses was carried. In the analyses the effect of pitch accent type on variation in the acoustic parameters listed in sec. 4.3.2 was investigated. The optimal feature set that was found in the analyses consists of:

b) ***direction:*** it is calculated as a difference between mean f0 on the accented and post-accented vowel. It describes direction of a pitch movement and distinguishes rising accents from falling accents.
c) ***f0peak(relative):*** measured as a difference between f0max in a two syllable window including accented and post-accented syllable and mean f0 on a phrase
d) ***f0min(relative):*** calculated as a difference between f0mean on a phrase and f0min in a two-syllable window including accented and post-accented syllable
e) ***f0mean(relative):*** meanf0 on accented vowel relative to mean f0 on a phrase. This feature is used to distinguish between L* and H* accents e.g. L*H vs. LH*
f) ***Amp(SAV):*** amplitude on the accented vowel, a value of -0,5 indicates fall, a value of 0,5 indicates rise, values in between indicate some amount of rise and fall occurring on the accented vowel. It distinguishes not only between falling vs. rising vs. LH*L accents, but also between two different types of falling (H*L vs. HL*) and rising pitch accents (L*H vs. LH*)
g) ***c1(SAV):*** rising amplitude on the accented vowel;
h) ***c2(SAV):*** falling amplitude on the accented vowel; together with c1(SAV) it describes the amount of pitch variation on the vowel and helps to distinguish accents with level pitch perceived on the vowel from those with a pitch movement. Besides, c1(SAV) and c2(SAV) have similar function to Amp(SAV).
i) ***tilt:*** the tilt parameter used in the current analysis is not calculated for a single syllable, but in a two-syllable window including accented and post-accented syllable. This feature describes the shape of the pitch accent and is highly correlated with position of f0 peak (also defined within a two syllable window). Like c1, c2 and Amp it discriminates between falling vs. rising vs. LH*L accents and also between types of rising and types of falling accents

The results of ANOVA depicted in Table 17 show that pitch accent type has the greatest effect on amplitude and tilt parameters, therefore it can be expected that these features will be the best discriminators of pitch accent type.

variable	F	p
direction	650,55	<0.01
f0peak(relative)	198,21	
f0min(relative)	551,01	
f0mean(relative)	591,38	
c1(SAV)	631,91	
c2(SAV)	1537,86	
Tilt	1258,64	
Amp(SAV)	1042,20	

Table 17: One-way ANOVA results.

The associations between the f0 parameters are shown in the table below. It can be seen that there are some significant correlations, which may indicate that part of this description is redundant. However, at this moment it is not certain which of the parameters could be excluded, because it is not known to which extent they contribute to pitch accent type recognition.

variable	direction	f0peak (rel.)	f0min (rel.)	f0mean (rel.)	c1 (SAV)	c2 (SAV)	Tilt	Amp (SAV)
direction	1,00							
f0peak(relative)	0,02	1,00						
f0min(relative)	-0,16	0,37	1,00					
f0mean(relative)	0,48	0,67	0,59	1,00				
c1(SAV)	-0,05	0,31	0,12	0,31	1,00			
c2(SAV)	0,29	-0,23	-0,61	-0,40	-0,41	1,00		
Tilt	-0,57	0,40	0,31	0,12	0,35	-0,52	1,00	
Amp(SAV)	-0,14	0,27	0,38	0,39	0,69	-0,65	0,49	1,00

Table 18: Correlation matrix showing associations between parameters describing pitch accents.

The box-whiskers plots depicted in Figure 47 show the distribution of means and standard deviations of f0 parameters for which the highest values of the F test are observed:
a) direction, b) f0mean(relative), c) tilt, d) c1(SAV), e) c2(SAV) and f) Amp(SAV).

The description of pitch accents proposed in this section is compact and can be easily derived from utterance's acoustics. Whether the information that it provides is significant to the automatic classification of pitch accent types (Hypothesis 1a) will be investigated in sec. 6.5.

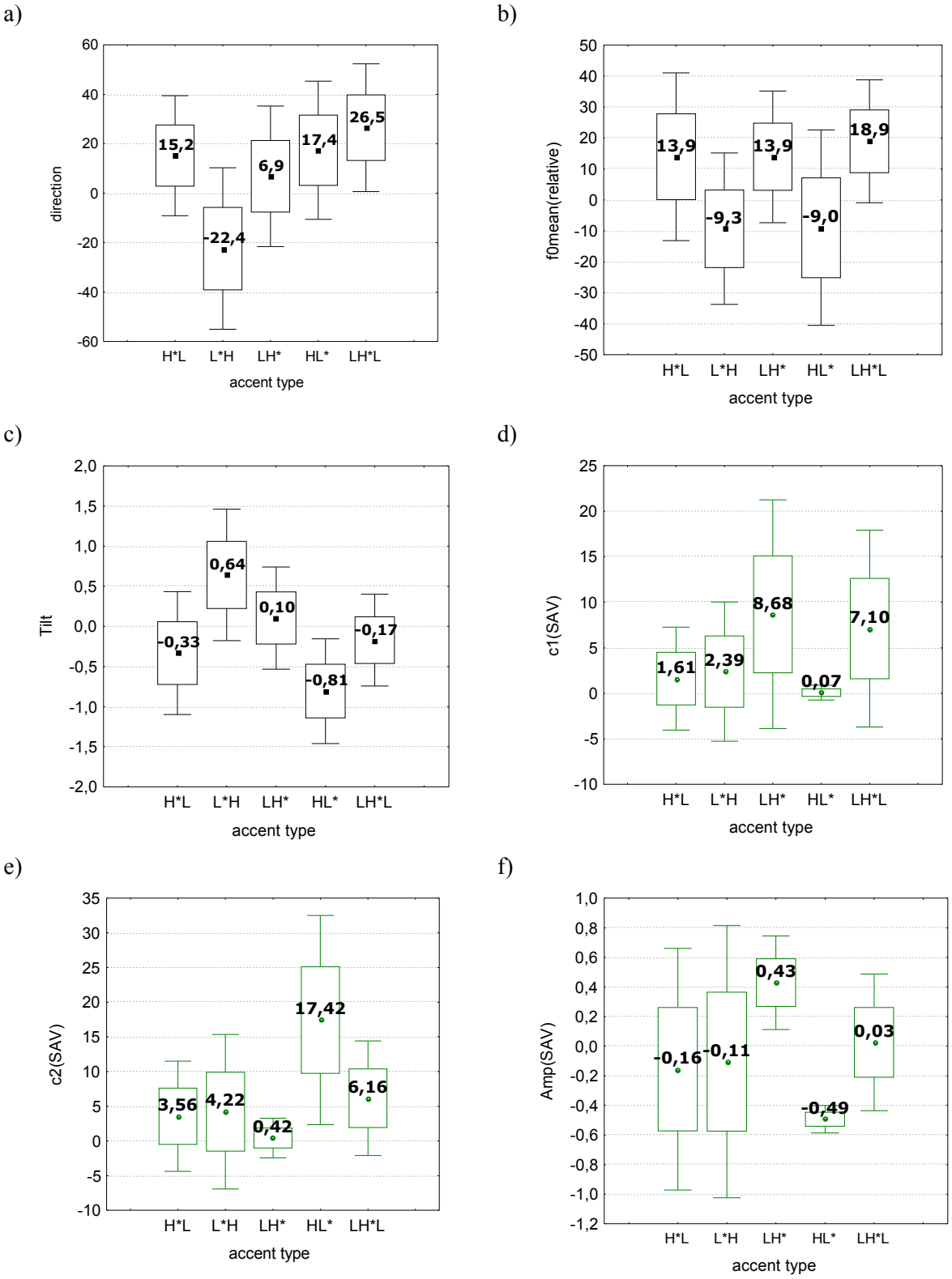


Figure 47: Distribution of means and D.S. of parameters describing pitch accents.

5.3.3. Boundary tones

The description on the phonetic level will be used in the automatic recognition of boundary tone types (see sec. 6.3). In order to define a description which is adequate for this task a series of ANOVA and discriminant function analyses was carried. In the analyses the effect of boundary tone type on variation in various acoustic parameters was investigated as well as their contribution to the distinction between various boundary tone types. The optimal feature set that was found in the analyses consists of:

a) ***f0end:*** final f0 value on a phrase boundary syllable. It distinguishes between types of rising and falling boundaries: 2,? vs. 5,? and 2,. vs. 5,.
b) ***f0mean(PAV):*** mean f0 on the nucleus of the syllable preceding the current syllable. It is used because of its high correlation with other variables which are important for distinction between phrase boundary types (see Table 19).
c) ***dist_#$p:*** distance of the current syllable to the next pause (#$p) marked in the annotation file. It is measured in syllables and distinguishes between boundaries of a different strength (i.e., 5 vs. 2).
d) ***direction:*** is calculated as a difference between mean f0 on the vowel of the syllable preceding the phrase-final syllable and on the nucleus of the phrase-final syllable. It describes direction of a pitch movement and distinguishes rising from falling boundaries.

Table 19 shows correlations between f0mean(PAV) and duration parameters as well as variables describing syllable distance to the next pause. All these features are important cues of boundary type. Duration and distance to the pause have the same function: they distinguish between boundaries associated with major (5) vs. minor intonation phrases (2). The f0mean(PAV) parameter distinguishes between various types of falling vs. rising boundaries. Significant correlations can be seen between all the variables: from among them those features were which discriminated between boundary tone types the best i.e., f0mean(PAV) and dist_#$p.

feature	syl_dur (relative)	nucl_dur (relative)	f0mean (PAV)	dist_next #$p	dist_#$p
syl_dur(relative)	1,00				
nucl_dur(relative)	0,79	1,00			
f0mean(PAV)	-0,47	-0,39	1,00		
dist_next#$p	-0,52	-0,44	0,55	1,00	
dist_#$p	-0,55	-0,45	0,56	0,97	1,00

Table 19: Correlation matrix depicting the association between f0 parameters.

Table 20 shows ANOVA results: it can be seen that variation in the acoustic features due to boundary tone type distinctions is statistically significant ($p<0.01$). The greatest effect of boundary tone type can be observed for syllable final pitch (f0end variable), whereas dist_#$p variable is the least affected by boundary tone type.

variable	F	p
f0mean(PAV)	566,68	<0.01
f0end	914,15	
dist_#$p	341,77	
direction	408,46	

Table 20: ANOVA results: The effect of boundary type on meanf0 on preceding vowel, syllable final pitch, distance to the next pause and direction of a pitch movement.

In Figure 48 the distribution of means and standard deviations of the acoustic parameters describing boundary tones is illustrated: a) f0mean(PAV), b) f0end, c) dist_#$p, d) direction.

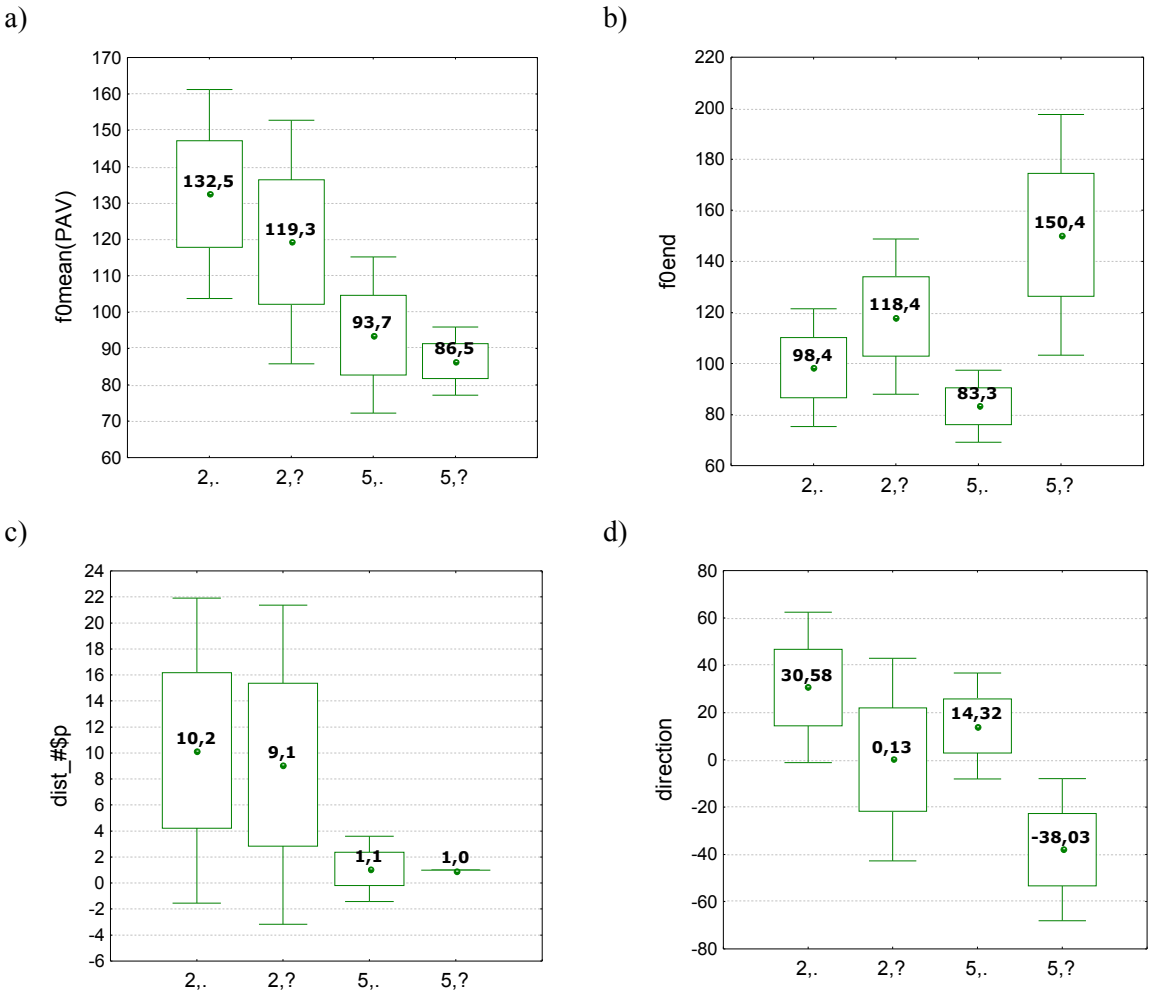


Figure 48: Distribution of means and D.S. of parameters describing boundary tones.

It can be seen in (a) that ***2,.*** boundaries are characterized by the highest mean f0 on the preceding vowel described by *f0meanf0(PAV)*, whereas ***5,?*** boundaries are characterized by the lowest mean f0.

In (b) the distribution of mean values of syllable final pitch is depicted. As expected, rising boundaries (***2,?*** and ***5,?***) have higher final pitch (variable *f0end*) than falling boundaries (***2,.*** and ***5,.***), but it can also be observed that ***2,?*** boundary "ends" at a significantly lower pitch than the other rising boundary (***5,?***).

The feature *dist_#$p* does not describe the realization of boundary tones but helps to draw distinction between minor (2) and major phrase boundaries (5). As mentioned before, in the Polish unit selection speech corpus used in the current study only 20% of minor phrase breaks is signaled by pauses. On the contrary, all major phrase breaks are followed by a pause; dist_#$p variable reflects these properties of the speech material.

The variable *direction* is the inverse of f0end. It describes the direction of the pitch movement at a phrase boundary; values below 0 indicate rise, values near 0 indicate that both falling and rising pitch movement occurs and values above 0 indicate fall. As mentioned before the boundary tone ***2,?*** is regarded as a rising boundary, but on the preceding syllable there may occur some fall in pitch as well and this property is depicted in the graph d).

The description of boundary tones proposed in this section is compact and can be easily derived from utterance's acoustics. In sec. 6.3 of the next chapter it will be investigated whether information provided by this description is significant to the automatic classification of types of phrase boundaries (Hypothesis 1a).

Chapter 6. Prosodic labeling. Coding of f0 contours

In this chapter three hypotheses will be tested. The first is related to the problem of definition of intonation description and the other two are related to the issue of automatic prosody labeling.

In the first place, the goal of the experiments presented in this chapter is to prove that the phonetic description of intonation proposed in this thesis provides information which is significant to the detection and classification of the elements of intonational tunes: pitch accents and phrase boundaries (Hypothesis 1a).

Secondly, the goal is to prove that the methods of automatic prosody labeling proposed in the thesis give results comparable the inter-labeler consistency in manual transcription of prosody (Hypothesis 2a). Automatic labeling of prosody consists in coding of intonation contours into a higher-level description which in case of the current study is the surface phonological description proposed in the Chapter 5.

The third goal is to confirm the Hypothesis 2b according to which a high accuracy in the automatic detection and classification of intonational events can be achieved even if only a small vector of features derived from utterance's acoustics and information extracted from utterance's transcription/segmentation are used at the input to the model.

Most of the existing approaches to automatic prosody labeling rely on both acoustic and text-based features, and often higher-level linguistic information such as POS is required. On the contrary, the methods designed in the current study use only small sets of acoustic features (e.g. four features are used to recognize boundary tone types), but achieve comparable results. The only information distinct from utterance's acoustics includes phonetic segmentation of the speech signal. It is assumed that the reduced amount of information necessary for the automatic detection/recognition of prosodic constituents is the advantage of the methods proposed in this thesis.

For the purpose of automatic classification of the type of accents and phrase boundaries it is necessary to know the exact position of accented syllables and phrase boundaries in the utterance. In most works presented in sec. 6.1.2 the same feature set is used to perform both the detection of the location of accented syllables and phrase boundaries, as well as the recognition of accent/boundary types. In the current study a different approach is adopted in which these tasks are performed by different models. This approach was chosen on the basis of the results of a number of statistical analyses aiming at selection of acoustic features (listed in sec. 4.3.2) which are the best cues of accent/boundary presence and which discriminate the best between various accent/boundary types. The analyses showed that different feature sets are "optimal" for these tasks.

To sum up, in order to test the hypotheses presented above various models will be designed which are capable of performing the following tasks:

- detection of phrase boundary location
- detection of accentual prominence (i.e., the location of accented syllables)
- recognition/classification of pitch accent types
- recognition/classification of boundary tone types

It is assumed that high accuracy achieved by the models can be regarded as a confirmation of Hypothesis 1a, because it indicates that the acoustic features which constitute description of intonational events on the phonetic level adequately express the differences between types of pitch accents and boundary tones distinguished on the surface phonological level of analysis and representation.

In order to confirm the other two hypotheses the performance of the designed models will be compared to the performance of other models (an overview is given in sec. 6.1.2) and to inter-labeler consistency in manual transcription of prosody (6.1.3). If the accuracy achieved by the models designed in the experiments presented in this thesis are comparable to those reported in the literature, the hypotheses will be proven.

All the experiments presented in this chapter are based on the subset of Polish unit selection system described in sec. 4.1.1. As it will be seen in the next section where existing approaches are presented it is very common to develop methods of automatic prosodic labeling on the basis of a single speaker database.

As regards methods used for the automatic detection/recognition of intonational events, following the approach taken in other studies linear (discriminant analysis function, linear neural networks) and non-linear statistical modeling techniques (decision trees, MLP and RBF networks) will be applied.

This chapter is organized as follows.

In the next section an overview of the existing approaches and solutions to the problem of automatic prosody labeling is given. In the first place, the state-of-the-art knowledge on the acoustic correlates of accentual prominence and prosodic boundaries is presented (6.1.1). This information will be taken into account while building models for automatic detection of the location of accented syllables and phrase boundaries. Next, existing approaches to automatic detection/recognition of prosodic constituents will be discussed (sec. 6.1.2): the information provided in this discussion will be taken into account while building models for automatic classification of pitch accent and boundary tone types. Apart from that the results reported by other authors will be used as a reference for evaluation of the models designed in the current study. The same goal has the overview of studies investigating inter-transcriber consistency in labeling of prosody given in sec. 6.1.3.

In sec. 6.2 & 6.3 a number of solutions will be proposed to the problem of automatic detection of phrase boundary location and recognition of boundary tone type.

In sec. 6.4 a number models will be designed for the purpose of automatic detection of the location of accented syllables, which can also be referred to as detection of accentual

prominence. In sec. 6.5 different statistical modeling techniques will be applied to the task of pitch accent type recognition.

6.1. Approaches and solutions presented in the literature

In this section an overview of the existing approaches and solutions to the problem of automatic prosody labeling is given. The number of studies presented here suggests that the issues related to automatic prosody labeling are in the range of research interests of many authors, which also shows the significance of this subject.

6.1.1. Acoustic cues of prominence and prosodic breaks

Prominence can be attributed to the occurrence of a lexical stress or accent: they both have prominence-cueing function i.e., they cause that *a word or its part stands out from its environment* (Terken 1991).

The issue of prominence detection was investigated in a number of studies (e.g. Rietveld & Gussenhoven 1985, Streefkerk 1997, Sluijter & van Heuven 1996, Portele & Heuft 1997, Tamburini & Caini 2005, Tamburini 2005, 2006). The main acoustic correlates of *accentual prominence* indicated in the literature are pitch variation (pitch movements) and overall intensity. On the contrary, duration and spectral emphasis are identified as the main cues signaling *stress*. Below, two of the latest works in this area are briefly discussed.

In (Ortega-Llebaria, Prieto & del Mar Varnell 2007) a study identifying direct acoustic correlates of stress is presented. The authors report on strong effect of duration on perception of stress. This effect is enhanced if increased nucleus duration is accompanied by increased intensity, but at the same time if there are no duration cues speakers rely solely on intensity to predict stress. The same analysis carried out for a different nucleus type ([i], the previously analyzed one was [a]) showed contrastive results. Listeners rely more on intensity than duration in prediction of stress, but in case of no duration differences listeners stop using intensity as a cue for stress. The results of the analysis of the interaction between duration and spectral tilt showed that the latter factor does not contribute to perception of stress.

An example of implementation of the findings of prominence-related studies is automatic identification of prominence presented in a number of papers by Tamburini (2005, 2006). The author proposes a function capable of identifying prominence which uses only acoustic features and does not require any phonetic information (such as transcription and segmentation if the utterance). For identification of prominence features such as nucleus duration, spectral emphasis, pitch variation measured with Tilt parameters, and overall intensity are used. The duration of syllabic nuclei is calculated from nucleus boundaries which are automatically identified. The Tilt parameters (amplitude and duration) are summed up as they "both contribute to support prominence in a reinforcement fashion" (Tamburini 2006:). The author reports on 80-90% accuracy on identification of prominence.

As regards the types of acoustic cues for prosodic boundary location and strength conflicting views can sometimes be found in the literature. Some authors prove that duration of

silent pause following phrase boundary belong to the most important features signaling boundary strength (e.g. Horne, Strangert & Heldner 1995, Bulyko & Ostendorf 2001). In the latter work dedicated to automatic labeling of prosodic features the recognition accuracy of phrase break location using pause duration alone achieved 95,8%, which proves importance of this feature. Other authors show that silent pause is a weak cue, given that sometimes less than 50% of phrase boundaries are followed by a silent interval (Yoon, Cole & Hasegawa-Johnson 2007). More agreement can be found as regards the role of pre-boundary lengthening as a cue for prosodic boundary (e.g. Wightman et al. 1992).

In (Horne, Strangert & Heldner 1995) gradual increase in silent interval duration (SI) from 0 for no-boundary condition up to 900ms for prosodic utterance of a final position in the text is observed. This effect of boundary strength on SI duration is unaffected by focus distinction (i.e., by presence vs. absence of focal accent on the penultimate syllable). The results of analysis of the effect of boundary strength on final lengthening (whose domain is rhyme of phrase-final syllable) are less explicit, but in general they prove that final lengthening exists at higher-ranked boundaries (i.e., above prosodic word) and affects most of all the final segment.

In (Yoon, Cole & Hasegawa-Johnson 2007) gradual increase in duration of the nucleus in the final syllable in prosodic word, ip and IP is reported. The occurrence of lexical stress on the penultimate syllable induces this effect, but only if the pre-boundary syllable is IP-final.

More acoustic cues to phrase boundary strength are given in (Yoon, Kim & Chavarria). Apart from the final lengthening of pre-boundary syllable rhyme the authors investigated the effect of boundary strength (distinction was drawn between minor of major phrase boundaries) on rhyme-final f0 level, f0 drop and slope, and intensity at the start and end of the sonorant part of the pre-boundary syllable's rhyme. They found that significant differences can be observed in f0 and intensity features between rhymes of syllables of a final position in minor vs. major phrases, but to some extent these effects depend on speech style.

The role of f0 features as acoustic cues for boundary strength was also analyzed in (Carlson & Swerts 2003). The authors sought for acoustic features which listeners use to predict upcoming phrase breaks and their strength. In a perception test listeners classified words and 2-sec. long stimuli (extracted from utterances, the following context e.g., the pause after the boundary, was unknown) as being followed by weak boundary, strong boundary or no boundary. A significant correlation between listeners' ratings of boundary strength and presence/absence of final creak as well as f0 median of last voiced 50ms of the word/stimuli proves that f0 features constitute acoustic correlates of phrase boundaries. Another finding of the experiments described in (Carlson & Swerts 2003) is that in the absence of cues such as final lengthening and silent pause duration listeners rely on other acoustic features (f0) in prediction of phrase boundary and its strength.

The results of the studies discussed in the section show that in the recognition of phrase boundaries and their strength the key role play the following features: duration of the silent pause following the boundary, lengthening of pre-boundary syllable (or nucleus, or rhyme), f0 - final pitch and pitch variation at the phrase boundary. It can be concluded that these features can be used in automatic prediction of prosodic phrasing.

6.1.2. Automatic recognition of prosodic constituents

Over the years, a lot of research has been dedicated to automatic recognition of prosodic constituents. In this section an overview of existing approaches is given with the aim of getting an insight into methods and features used for this task and accuracy with which prosodic constituents can be automatically recognized. As mentioned before, the results of the studies presented here will be used as a reference for the evaluation of the performance of the models designed in this thesis.

1. In (Rapp 1996) an automatic procedure that recognizes phonological tonal categories of pitch accents and boundary tones (Fery 1993) from symbolic input is presented. The system distinguishes between 5 underlying pitch accent types and 4 further accents available in the surface structure, and 4 boundary tones. The system makes joint prediction of accent type and boundary tone (i.e., some accents are followed by a boundary tone and others not).

The input to the system is phonetic information: the speech data is automatically segmented into words, syllables and phones and transcribed.

A decision tree is trained to recognize the phonological categories from the following input features:

a) *duration*: including syllable and nucleus duration. Three duration measures are used: actual duration resulting from the segmentation, expected - calculated by summing the mean durations over all the phonemens in a syllable or giving the mean duration of syllable's nucleus and relative - calculated as a ratio of actual and expected durations
b) *pause*: distance to the next pause and its duration
c) *f0 features*: f0 is extracted with ESPS pitch tracker and subject to approximation in a 2 syllable window. The f0 features describe the pitch movement within that window. For each syllable the following information is provided as a result of the approximation: mean f0, height, alignment and steepness of the peak and rise/fall
d) *intensity*: for syllable nuclei which is determined with the help of phonetic segmentation the information on mean energy and spectral tilt are provided
e) *lexical features*: they are derived from the transcription and segmentation of the speech data and include vowel type (short, long, etc.) and phoneme, word stress, distance to the next word boundary and POS

The input to the recognizer included one hour of news stories read by a professional German radio speaker: 10455 syllables was used for training and the remaining 2436 syllables was used for testing.

In order to optimize the recognition performance a number of experiments using various feature sets was carried out. The best performance i.e., 78,6% accuracy in joint prediction of pitch accents and boundary tones was achieved with a feature set including: f0 parameters, distance plus length of the following pause, syllable-based duration parameters, lexical stress on the syllable and distance to the word boundary. For comparison, the recognition of accented vs. unaccented syllables had 86,9% accuracy.

These results show that from the point of view classification of intonational events intensity and POS are less significant than duration and f0 features.

2. In (Kießling et al. 1996) the solution to the problem of automatic recognition of prosodic constituents for needs of Verbmobil project (Wahlster 1993) which aimed at automatic speech-to-speech translation in appointment scheduling dialogues is presented. For classification of prosodic constituents: pitch accents (marked with A), prosodic boundaries (marked with B) and syntactic-prosodic boundaries (M) a 4-layer MLP network is used. The classification is binary i.e.,
   a) accented (groups primary, secondary and emphatic accents) vs. unaccented
   b) boundary (major intonational phrase boundary) vs. no boundary (groups intermediate phrase boundary, word boundary and aggramatical boundary like hesitation or repair)
   c) as for syntactic-prosodic boundaries a distinction is drawn between strong vs. weak/no boundary

The database used for training pitch accent and phrase boundary recognizer consisted of 30 dialogues (100 minutes of speech) by 53 male and 4 female speakers. In case of syntactic-prosodic boundaries 13 hours of speech including 322 male and 203 female speakers were used for training. All models were tested on a database consisting of 3 dialogues (12 minutes) by 3 male and 3 female speakers was used.

The classification is performed at syllable level (i.e., word-final syllables are classified). Only acoustic features are used: duration, f0, pause length, energy, speaking rate. Additionally, for each syllable the information is given on lexical word accent on the syllable and syllable position in the word.

In a number of experiments MLP networks are trained with different feature sets and various complexity of the networks is tested. The best classification results are reported when all features are used: the overall recognition rate for boundaries is 88.3% and for accents - 82.6%. In boundary recognition the most important are *f0 features*, *energy* and *pause length,* whereas in accent recognition - *f0 features*. As regards syntactic-prosodic boundaries (M) the overall recognition accuracy is 86.7%. For this task another MLP network was designed and the recognition was based on a different feature set (for details see the paper).

3. Demenko (1999) automatic classification of accent types in Polish with MLP networks is presented. Three experimental settings are described.

The first experiment is carried out on a database consisting of 1535 mono- and bi-syllabic words on which one of 9 nuclear contours was realized. They were distinguished on the basis of perceptually significant features such as direction of f0 movement, its range and timing with accented syllable's onset and position of the pitch targets for the rise/fall in speaker's range. The classification of syllables was based on a 5-feature vector including f0 features only. In the training subset the overall recognition accuracy achieved 85.5%, whereas in the test set - 82.6%.

In the second experiment a database consisting of read sentences (polysyllabic structures) was used. 1630 sentences were used for training of an MLP network and 300 for testing. This time the distinction was made between 2 prenuclear (L and H) and 9 nuclear pitch accents. The classification was performed at foot level. An 11-feature vector was used for recognition of accent type: it consisted of 9 f0 features, normalized energy and duration of the vowel of a foot-final position. The overall recognition accuracy for the training subset was 83% and for the test subset - 80%.

The second experiment consisted in recognition of accent type in continuous speech. The inventory of accent types used in the previous experiment was modified based on a perception study results: 3 classes of rising nuclear accents were collapsed in a single class R, 4 classes of falling nuclear accents were also collapsed into a single class F. So the distinction was made between H (high prenuclear), L(low prenuclear), R, F and MM ("level" nuclear accent, not signaled by intonation) accents. The overall recognition accuracy was between 79-83% depending on the accent type. The recognition was performed by the MLP network designed in the experiment 2.

The automatic classification of vowels using MLP network and a vector of 8 features (7 f0 features plus duration), resulted in 91% recognition accuracy of unaccented vowels, 86% of accented and 84% of pre-boundary vowels.

4. In (Wightman et al. 2000) an approach to prosodic labeling based on a perceptually motivated inventory of ToBI labels is presented. A low inter-transcriber agreement for some tonal categories observed by the authors motivated a re-definition of the labeling system. For that purpose a study was carried out in which perceptual prominence of various tonal categories was investigated. As a result the inventory of EToBI labels was reduced to 2 accent categories marked with ** and * (the former groups bi-tonal accents perceived as more prominent and the latter groups other accents), whereas all major phrase boundaries were collapsed into a single category (%).

The task of the recognizer is to assign each syllable to one of the classes: unaccented, accented **, accented *, accented ** plus boundary tone %, accented * plus boundary tone %, boundary tone %.

The labeling algorithm uses a decision tree based VQ designed jointly with a HMM. At the input it requires segmentation and transcription of the speech signal, and 24 features determined for each syllable. The features include: lexical stress, position in the word, phoneme of the nucleus, normalized duration, maximum/average pitch ratio.

The recognizer was trained on a speech corpus consisting of texts representing different prosodic styles read by a professional female speaker. A subset of the speech data (42 utterances) was kept for testing.

The recognition of accented ** vs. unaccented syllables was performed with 84% accuracy. Worse results were achieved for the recognition of prominent (both ** and *) vs. unaccented syllables - 69,3%. The highest accuracy is reported for detection of phrase boundaries: on average 93,4%.

5. In (Bulyko & Ostendorf 2001) automatic annotation of prosody using a decision tree is presented. The labeling scheme is based on a simplified ToBI system (Pitrelli, Beckman & Hirschberg 1994) which distinguishes between 3 pitch accent types and 4 boundary tones marking the boundaries of major intonational phrases. For the recognition of prosodic labels phonetic segmentation and transcription of the speech data is required. The input to the recognizer includes text-based and acoustic features:

a) *f0 features*: mean and peak f0 for the word normalized by the utterance peak, mean f0 and f0 slope of the word's last 60 ms normalized by the utterance peak

b) *duration*: measured for the stressed vowel and the last phone in the word normalized by the corresponding phone's mean duration
c) *pause*: the length of the pause following the word
d) *lexical features*: POS and syntactic structure, semantic class of the word

The speech database used in the experiment contained system responses from a travel planning dialog read by a female speaker. One part of the database (2752 words) was used for training of the tree and the other part was used for testing (688 words). The results show that the use of acoustic and text-based features results in a better performance in comparison to the accuracy achieved by the system when only acoustic or text-based features are used. Prosodic breaks were recognized with 96,2% accuracy and the most important predictor variable was length of the following pause. Accent type was correctly classified in 65,2% and the prediction was predominantly based on normalized f0 peak and vowel duration, and POS. The most important feature for the recognition of boundary tone type (93,9% accuracy) was normalized f0 peak, mean and word-final f0, and normalized vowel duration. The distinction between accented vs. unaccented syllables was correct in 80,9%.

6. In (Sridhar, Bangalore & Narayanan 2007) a method of automatic prosody labeling in a maximum entropy framework is presented. The method was developed and tested on subsets of two annotated (with POS and ToBI) speech corpora: Boston University Radio Corpus, BU (Ostendorf, Price & Shattuck-Hufnagel 1995) and Boston Directions Corpus, BCD (Nakatani, Hirschberg & Grosz 1995). The speech material used by the authors consists of 8 speakers (female and male) and 375 utterances in BC and 37 utterances in BDC corpus. The model relies on the following input features:

a) lexical and syntactic features: The former are simply the words in a given utterance. The latter include: POS (provided in the corpora annotation), function vs. content word (derived from POS), *supertags* which "encapsulate predicate-argument information in a local structure" (Sridhar, Bangalore & Narayanan 2007:6) and thus, provide a richer syntactic information than POS, they are obtained in a process called *suppertagging* (Bangalore & Haffner 2005). The features discussed here are provided per word plus for 3 next and 3 previous words.
b) acoustic-prosodic features: They include f0 and RMS energy (e) features extracted over 10ms frames, both f0 and (e) are normalized across speakers using z-score transformation; no duration features are used. A feature vector comprising f0 and (e) and their delta and acceleration coefficients is created and subjected to quantization (for details see the text:7). A quantized feature vector serves then as input to the maximum entropy model.

Two maximum entropy models are tested. The first one uses only lexical and syntactic features to perform a binary classification of accents (accented vs. unaccented) and boundaries (boundary tone presence vs. no boundary). The accuracy of pitch accent detection is about 84% on the BU corpus and 79.81% on the BDC corpus. Boundaries are correctly identified in 90,28% on the BDC corpus and in about 91% on the BU corpus.

The second model uses only acoustic-prosodic features. The accuracy of pitch accent detection is 80.09% on the BU corpus and 74.51% on the BDC corpus. Boundaries are correctly

identified in 84.10% on the BU corpus and in 85.53 % on the BDC corpus. For comparison, using the same feature set a HMM-based recognizer achieves a much lower accuracy (about 8-10%). When adding syntactic features the performance of both HMM and maximum entropy models gets significantly better.

In the recognition of break indices the same feature sets are used as in the detection of pitch accents and boundary tones.

As regards the recognition using the full inventory of ToBI break indices (0 - word-internal, 1- word boundary, 2-disjuncture, possible intermediate boundary, 3-intermediate phrase boundary, 4- major phrase boundary) the accuracy on the BU and BDC corpora are 64.73% and 66.56% respectively using acoustic features only and increases to 72.90% and 69.81% when syntactic features are used as well. In general, the results show that syntactic features are more significant for detection of pitch accents, boundaries and strength of prosodic break.

As regards the recognition using only a binary classification (boundary vs. no boundary) the accuracy on the BU and BDC corpora are 73.98% and 75.94% resp. using acoustic features only and increases to 84.01% and 87.58% when syntactic features are used as well. In general, the results show that syntactic features are more significant for detection of pitch accents, phrase boundaries and strength of prosodic breaks.

The authors compare the performance of their models with the accuracy of prosody prediction in Festival and AT&T TTS systems. The results show that the maximum entropy models are superior to the existing approaches.

7. In (Rosenberg 2005) a method of automatic prosody labeling for elicited spontaneous speech is presented. The database used in the experiments is subset of Boston Directions Corpus, BCD (Nakatani, Hirschberg & Grosz 1995) and consists of 65 minutes of speech (14014 words) of 3 male and 1 female speaker. The corpus is ToBI-labeled by hand. The recognition of pitch accents and boundary tones is performed using a toolkit offering a number of machine learning techniques (e.g. decision trees). As regards pitch accents, three classes are distinguished, 2 phrase accents and 2 boundary tones, for each of these a zero class is also defined (no accent/no phrase accent/no boundary). The approach assumes that word identity/boundaries, and break index location/type are given. The classification is performed at the word level and uses duration, f0 and intensity features, POS, break indices, word length (measured in syllables) and word position in the intermediate phrase.

Pitch accent detection (i.e., a binary classification: accented vs. unaccented) achieved 85.7% accuracy, whereas identification of pitch accent type (where *unaccented* also counts as a type) achieved 79.2% accuracy. As regards phrase accents they were correctly detected in 72.4%, whereas for boundary tones 73.2% detection accuracy is reported. The joint prediction of phrase accents and boundary tones had a very low accuracy - only 54.7%.

Some of the previous and recent approaches not described in the current are presented in the following works: (Pierrehumbert 1983), (Wightman 1992), (Wightman & Ostendorf 1992, 1994), (Hirschberg 1993), (Strom 1995), (Ross & Ostendorf 1996), (Syrdal, Hirschberg, McGory & Beckman 2001), (Braunschweiler 2003).

6.1.3. Inter-transcriber consistency in labeling intonational events

In this section inter-transcriber consistency in annotation of prosodic features is briefly discussed. This issue is our of interest, because inter-transcriber labeling consistency can be regarded as a reference point for assessment of the results of automatic labeling of intonation.

1. The study presented in (Pitrelli, Beckman & Hirschberg 1994) reports on the results of evaluation of inter-transcriber reliability of prosody labeling in the ToBI framework (Beckman & Ayers 1997, Beckman & Hirschberg 1994). The inventory of labels included more than six pitch accent types (H*, L*+H, !H*, L+H*, H*+!H, L*, L*+!H), two phrase accents (H, L-), and two boundary tones (H%, L%).

The authors report on 80.6% inter-transcriber agreement on the presence vs. absence of a pitch accent and 68% agreement on presence and choice of accent type. The overall consistency in labeling of pitch accent types when both transcribers agreed on accent presence is 64%. The overall agreement on the presence vs. absence of a phrase accent is 89.8%, whereas the agreement on presence and choice of the phrase accent is 85%. The overall consistency in labeling of phrase accent types when both transcribers agreed on its presence is 72.9%.

As regards the consistency of boundary tone labeling the agreement of boundary tone presence vs. absence reported in the study is 93.4%, whereas the agreement on presence and choice of the boundary tone type is 90.9%. The agreement on a boundary tone type when both transcribers agreed on boundary presence is 78.8%.

2. In (Grice et al. 1996) a corpus of 733 words including read texts and dialogue speech was labeled with GToBI pitch accents and boundary tones by 13 transcribers with various experience. 9 pitch accent types, 2 phrase accents and 5 edge tones (i.e., a phrase accent followed by a boundary tone) were distinguished in the transcription. Inter-transcriber consistency was measured "by comparing labels placed by transcribers on potential site for a tonal element" (Grice et al. 1996:7). The comparison was made pairwise: the measure of inter-transcriber consistency is "the percentage of transcriber-pair-words which exhibit agreement on a particular elements potential site" (after Pitrelli et al. 1994:125).

The overall inter-transcriber consistency for pitch accents reported in (Grice et al. 1996) is 71%: each word was provided with a label indicating one of the 9 pitch accent type mentioned above or alternatively, zero accent for unaccented words. When only a binary decision on accent presence/absence (i.e., accented vs. unaccented) is taken into account the inter-transcriber agreement increases to 87%. Ignoring the distinction between non-downstepped vs. downstepped accents gives a consistency of 74%.

The overall inter-transcriber consistency for edge tones is 86% (a "zero boundary" is also taken into account) and the same for boundary strength (i.e., distinction between ip and IP boundary).

3. In (Reyelt 1996) the results of two inter-transcriber consistency labeling experiments are reported. In the first experiment the consistency of labeling *functional* aspects of prosody was investigated on the basis of a corpus consisting of 480 read utterances. 5 inexperienced labelers marked secondary and main accents and boundaries (with no further distinctions). The

transcription was based on auditory analysis solely. The overall consistency achieved 40% for secondary, 72% for main accents and 76% for boundaries. The measures were the same as in the previously described experiments in (Grice et al. 1996) and (Pitrelli, Beckman & Hirschberg 1994).

In the second experiment trained subjects participated and apart from functional labeling they also marked pitch accent type, boundary tone type and boundary strength. The database used in the experiment consisted of 233 spontaneous dialogues. Unlike in the first experiment, this time labelers could inspect pitch contours. The overall consistency of labeling of the functional prosodic features (main vs. secondary accent, phrase boundary) achieved 91%, pitch accents were labeled with overall consistency of 85%, boundary tones - 88% and prosodic breaks - 94%. Significant differences in inter-transcriber agreement regarding specific pitch accent and boundary tone types could be observed. In comparison to untrained labelers, the trained ones have achieved greater consistency in labeling main accents (86%) but much lower consistency in secondary accent labeling (only 32%). The overall agreement on major phrase boundaries achieved 90%, whereas on minor (intermediate) phrase boundaries - only 44%.

4. A recent study on inter-transcriber reliability of prosodic labeling using ToBI is presented in (Yoon et al. 2007). The study is based on a subset of the Switchboard corpus which consists of recordings of spontaneous telephone conversations. Altogether 181 files from the corpus containing utterances from 79 different speakers were transcribed by two labelers using a simplified version of the ToBI system (Beckman & Ayers 1997, Beckman & Hirschberg 1994). This simplified annotation system included three pitch accent types (H*, L* and X* for "elusive" cases), two phrase accents (H- and L-) and three boundary tones (H%, L% and r% for boundaries resulting disfluencies).

The authors report on 89.14% inter-transcriber agreement on the presence vs. absence of pitch accents regardless of pitch accent type and 86.57% when the type of accent is taken into account. The consistency in labeling of pitch accent types when both transcribers agreed on accent presence is 94.6%. The overall agreement on the presence vs. absence of a phrase accent is 88.39%, whereas the agreement on presence and choice of the phrase accent is 85.63%. The consistency in labeling of phrase accent types when both transcribers agreed on its presence is 88.07%.

As regards the consistency of boundary tone labeling the agreement of boundary tone presence vs. absence reported on the study is 90.04%, whereas the agreement on presence and choice of the boundary tone type is 89.33%. The agreement on a boundary tone type when both transcribers agreed on boundary presence is 88.7%.

These results show higher inter-transcriber consistency in labeling of prosodic features in comparison to that reported in (Pitrelli, Beckman & Hirschberg 1994) and (Syrdal & McGory 2000), but the authors admit that this might be due to considerably smaller inventory of categories used in their study.

6.2. Detection of phrase boundary location

In this section methods capable of detecting the location of phrase boundaries are proposed.

In some studies (e.g. Horne, Strangert & Heldner 1995) *pauses* are identified as the most important acoustic cues signaling phrase boundaries. For example in (Bulyko & Ostendorf 2001) the accuracy of prediction of phrase break location using pause duration alone were 95,8%. However, quite contradictory view can also be found. For example, in (Yoon, Cole & Hasegawa-Johnson 2007) it is reported that only 40% of phrase boundaries in the Boston University Radio News Corpus (Ostendorf, Price, Shattuck-Hufnagel 1995) is signaled by silent pauses. Apart from that, it should be noted that not all pauses signal prosodic breaks, some of them result from hesitations and thus do not have this function.

The experiments presented in the current study are based on speech material from the Polish unit selection corpus which most of all includes isolated sentences. For that reason it seems useless to measure duration of pauses marking sentence (thus, major intonational phrase) boundaries.

For this practical reason and the reasons mentioned above no information on pause presence/absence or duration will be used in the current study.

The detection of boundary location will be performed on the word-level which ensures that no word will be assigned more than one boundary - this can happen when prediction is made on the syllable level. The information on word boundaries is given in the utterance's segmentation/annotation.

Detection of boundary location is an example of a classification problem: each word can either be phrase-final or not. Therefore, for this task statistical methods are applied which are capable of solving classification problems: discriminant function analysis, decision trees and neural networks.

6.2.1. Features

Firstly, a number of features which could serve as cues for boundary location were selected on the basis of the information provided in the literature (sec. 6.1.1). Then, in a series of ANOVAs the effect of boundary presence vs. absence on variation in the acoustic features was investigated. It was found that the following features can be regarded as the best acoustic correlates of prosodic boundary presence:

a) ***nucl_dur(previous)*** - relative duration of the nucleus of the previous syllable
b) ***syl_dur(relative)*** - relative duration of the syllable
c) ***nucl_dur(relative)*** - relative duration of the nucleus
d) ***tilt(SAV)*** - tilt value on the vowel
e) ***c2(SAV)*** - the rising amplitude on the vowel
f) ***f0mean(SAV)*** - overall f0 level on the vowel
g) ***slope(PAV)*** - the amount of variation in pitch on the vowel of the previous syllable. The use of this feature is motivated by the fact that pre-boundary syllables are most often directly preceded by accented syllables and accented vowels exhibit greater pitch variation (both in terms of the amount of rising and falling pitch) than the unaccented ones.

Table 21 presents means and S.D. from the means for each class of vowels/syllables: +b (pre-boundary), -b (word-final). It can be seen that in the -b class the duration of the pre-boundary syllable, vowel and the preceding vowel are significantly shorter than in the +b class. As regards f0 features it can be observed that in the +b class there occurs significantly less falling pitch (indicated by higher average tilt value) than on the word-final vowels that do not precede a boundary, which is the effect of rising boundaries. This can also be concluded from significantly higher rising amplitude (c1) on the pre-boundary vowels. These vowels are also characterized by much lower overall pitch (f0mean), which can be attributed to falling boundaries. As regards the amount of pitch variation on the preceding vowel described by the slope(PAV) variable is it significantly greater in the +b than -b class. This effect could be expected, because most of pre-boundary syllables are preceded by accented syllables, which results from the fact that stress in Polish has a fixed position in word (penultimate syllable, but obviously, there are some exceptions). As it will be seen in the next sections pitch variation (here described by *slope* feature) is one of the acoustic correlates of accentual prominence, which the difference between vowels preceding/not preceding pre-boundary vowels/syllables.

class:	nucl_dur(previous)		syl_dur(relative)		nucl_dur(relative)	
	mean	S.D.	mean	S.D.	mean	S.D.
-b	0,90	0,26	0,90	0,21	0,90	0,32
+b	1,23	0,31	1,34	0,35	1,47	0,62
average all:	0,99	0,31	1,02	0,32	1,06	0,49

class:	tilt(SAV)		c1(SAV)		f0mean(SAV)		slope(PAV)	
	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.
-b	-0,75	0,57	0,38	1,63	114,09	13,06	88,12	67,78
+b	-0,04	0,94	6,13	11,96	99,81	21,83	129,40	76,81
average all:	-0,56	0,76	1,98	6,94	110,12	17,22	99,64	72,80

Table 21: Means and S.D. from means of f0 and duration features of word-final (-b) and pre-boundary (+b) vowels/syllables .

The differences discussed here are statistically significant as shown by ANOVA results. It can be seen in Table 22 that the most affected by boundary presence/absence are syllable duration and previous vowel duration. From among f0 features the most affected are tilt and f0mean of the vowel.

feature	F	p
nucl_dur(previous)	399,26	
syl_dur(relative)	504,01	
nucl_dur(relative)	23,64	
tilt(SAV)	92,61	<0.01
c1(SAV)	25,31	
f0mean(SAV)	81,07	
slope(PAV)	64,03	

Table 22: ANOVA results: the effect of prosodic boundary presence.

Some quite significant correlations between the variables were found, but a preliminary discriminant analysis showed their elimination would deteriorate the accuracy of phrase boundary detection. Consequently, it was decided to use all of them.

variable	nucl_dur (previous)	syl_dur (relative)	nucl_dur (relative)	tilt(SAV)	c1(SAV)	f0mean (SAV)	slope (PAV)
nucl_dur(previous)	1,00						
syl_dur(relative)	0,36	1,00					
nucl_dur(relative)	0,35	0,76	1,00				
tilt(SAV)	0,30	0,40	0,44	1,00			
c1(SAV)	0,26	0,44	0,53	0,53	1,00		
f0mean(SAV)	-0,29	-0,37	-0,35	-0,27	0,00	1,00	
slope(PAV)	0,26	0,18	0,18	0,15	0,06	-0,16	1,00

Table 23: Correlation matrix showing associations between the variables.

6.2.2. Detection with discriminant analysis function

This section reports on the results of detection of boundary location with discriminant analysis function. Altogether 7780 instances of word-final syllables/vowels were used in the analysis and they were split into 2 samples: one (learning) was used to compute the classification functions. It included 4159 instances of word-final (-b) and 1623 pre-boundary (+b) syllables/vowels. The other sample (cross-validation) consisted of 1398 word-final (-b) and 528 pre-boundary (+b) syllables/vowels. It was used for the evaluation of the classification functions in a cross-validation test. The results computed for the learning and cross-validation samples are given in Table 24.

class:	learning	cross validation
-b	90,45	90,05
+b	74,63	74,04
overall%:	86,04	85,68

Table 24: Accuracy of boundary location detection for learning sample and resulting from the cross-validation test.

It can be seen that word-final vowels which are non-pre-boundary at the same time (-b class) are detected with a higher accuracy than those of a phrase-final position (+b). The results computed in the cross-validation test prove an overall high rate of detection of boundary location.

In general, the results achieved in the current study are comparable to those reported by other authors (e.g. Wightman & Ostendorf 1994, Kießling et al. 1996, Sridhar, Bangalore & Narayanan 2007, Rosenberg 2005). Given the small number of acoustic features that is required

to achieve a detection accuracy of about 86% the results reported here can be regarded as satisfying.

6.2.3. Detection with decision trees

Like in the previous experiment when boundary location was detected with discriminant analysis function the database used in the current experiment included 7780 instances of word-final syllables/vowels. They were split into 2 samples: one (learning) was used to train the decision tree. It included 4159 instances of word-final (-b) and 1623 pre-boundary (+b) syllables/vowels. The other sample (test) consisted of 1398 word-final (-b) and 528 pre-boundary (+b) syllables/vowels and was used for the evaluation of the performance of the tree. Additionally, a cross-validation test was carried out on the training sample in order to evaluate the performance of the designed model.

Figure 49 illustrates the importance ranking of predictor variables used in the classification of word-final vowels into pre-boundary (+b) and non-pre-boundary (-b). The importance is depicted on a 0-100 scale: a value of 100 indicates high importance and values near 0 indicate low importance. It can be seen that the most important features are *duration of the previous nucleus* and *the current syllable*, whereas with the exception of *tilt(SAV)* the f0 features are significantly less important for the detection of a phrase boundary location. To some extent the results shown here are in accordance with the effects shown by ANOVA analysis (sec.6.2.1).

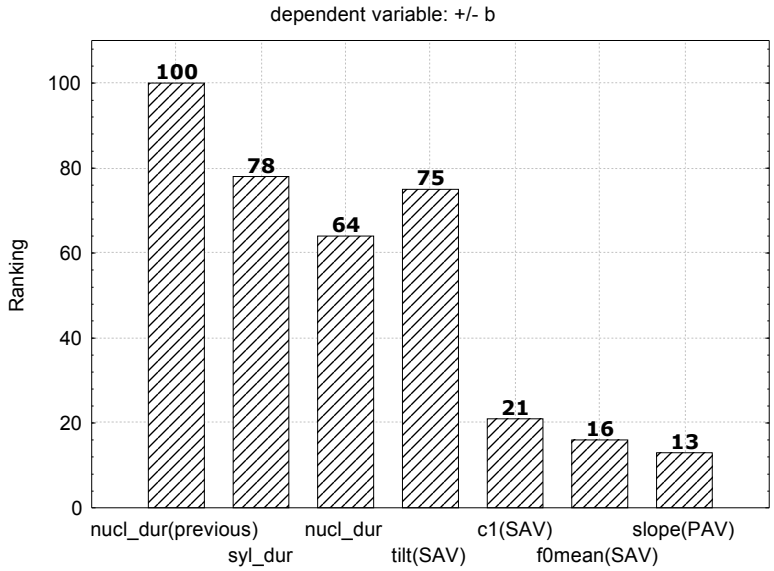


Figure 49: Importance ranking of predictor variables used in the detection of phrase boundary location.

The graph in Figure 50 illustrates the structure of the decision tree designed in the current experiment. The tree is very small: it has 5 splits and 6 terminal nodes. It can be seen that the top-level splits were based on conditions referring to the most important features:

syl_dur(relative) and *nucl_dur(previous)*. But the *c1(SAV)* variable which was indicated as having small importance is referred to in a split condition as well, as opposed to *tilt(SAV)* or *nucl_dur* which were given high position in the ranking, but are not used in the classification. In the graph the following information is shown:

a) the number of cases in each observed class that are sent to the node
b) the predicted class to which cases sent to the node is assigned
c) split condition for a split node

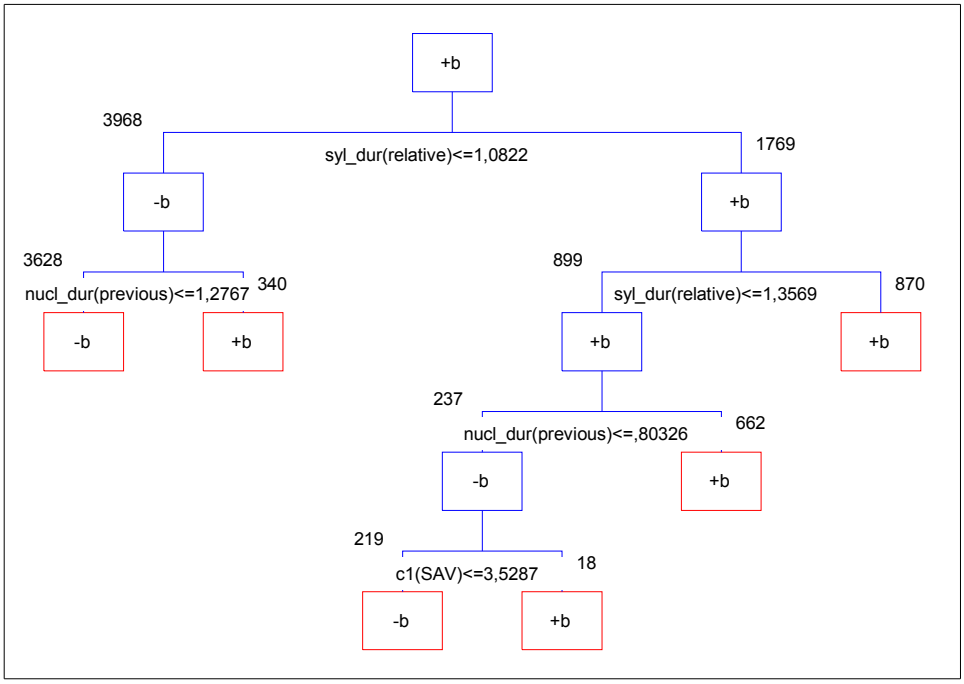


Figure 50: Decision tree designed for detection of phrase boundary location.

The results of detection of prosodic boundary location with the tree designed in the current experiment are presented in Table 25. In a) misclassification matrix computed for the training sample is presented, in b) the results based on the test sample are shown and c) gives the results of the cross-validation test.

a)

class	-b	+b
-b		352
+b	642	
% correct	84,48	78

b)

class	-b	+b
-b		113
+b	219	
% correct	84,33	78,6

c)

class	-b	+b
-b		343
+b	747	
% correct	81,94	78,56

Table 25: Misclassification matrix for the learning a), test b) and c) cross-validation sample.

The average accuracy of detection prosodic boundaries is comparable to that obtained with discriminant analysis function. What differs is the accuracy achieved in specific classes. It can be seen that +b vowels are detected with higher accuracy with the tree designed in the current experiment than with the discriminant function analysis. On the contrary, the detection

of -b vowels is lower in the current study. But like previously, the results can be regarded as satisfying. They are generally comparable to those reported by other authors, but our approach requires only a few acoustic features, which is an advantage. Moreover, the usefulness of the approach presented here is proved by the fact that the accuracy of automatic detection of boundary location achieved in this study is comparable to consistency among human labelers (e.g. Reyelt 1996, Grice et al. 1996, see sec.6.1.3).

6.2.4. Detection with neural networks

For design of the classification networks *Intelligent Problem Solver* program available in SNN package was used. Linear, MLP and RBF networks were chosen for a preliminary training (with networks of a different complexity and with various case samplings) and assessment. All MLP networks had 3 layers and the number of units in the hidden layer varied between 5 and 20. The number of units in the RBF networks varied between 5 and 90. The networks were trained using the whole set of independent (predictor) variables (36) and alternatively, allowing the network to automatically select a subset of variables. The database used in the experiments contained 17481 syllables. They were split into 3 subsets:

- training (5844) used to optimize the network
- selection (800) used to halt training to mitigate over-learning and/or to select from a number of models trained with different parameters
- test (1000) used to perform an unbiased estimation of the network's likely performance

The train and select subsets were resampled, but the test set was maintained the same to allow comparison of results.

The best results were achieved for an MLP network with 14 units and RBF network with 23 units in the hidden layer; linear networks performed much worse as regards detection of +b vowels (about 60% accuracy) and thus, they are not taken into account in the statistics. The results obtained on the selection test will be skipped in the discussion as they are of less interest. The summary of the two networks is given in Table 26. Starting from the left column the following information is provided:

a) network type and structure (i.e., the number of input, hidden and output units)
b) performance of a network on the training/test sample; the performance measure is the proportion of cases correctly classified
c) the error (RMS) of the network on the training/test sample

network	perf. training	perf. test	RMS training	RMS test
MLP 4:4-14-4:1	0,82	0,80	0,68	0,75
RBF 7:7-23-1:1	0,81	0,80	0,31	0,31

Table 26: Model summary details.

In general, both MLP and RBF networks achieved high overall detection accuracy about 80%. It can be seen that the networks' performance is nearly the same on the training and test samples, which means that they are able to generalize to new cases very well. Table 26 and Table 27 summarize the networks' performance on the training and test sample: In Table 27 results achieved with the MLP network are presented; In Table 28 the performance of the RBF network is summarized.

MLP 4:4-14-4:1	training		test	
class:	-b	+b	-b	+b
total:	4229	1615	722	278
correct	3462	1337	592	212
missclass.	767	278	130	66
correct%	81,86	82,79	81,99	76,26
missclass.%	18,14	17,21	18,01	23,74

Table 27: Summary statistics: MLP network.

RBF 7:7-23-1:1	training		test	
class:	-b	+b	-b	+b
total:	4235	1609	729	271
correct	3409	1317	578	221
missclass.	826	292	151	50
correct%	80,50	81,85	79,29	81,55
missclass.%	19,50	18,15	20,71	18,45

Table 28: summary statistics: RBF network.

In order to investigate the importance of the input variables for the detection of phrase boundary location a sensitivity analysis was carried out. Table 29 shows importance ranking of the predictor variables computed in the analysis: 1 indicates the highest importance.

network	nucl_dur (previous)	syl_dur (relative)	nucl_dur (relative)	tilt(SAV)	c1(SAV)	f0mean (SAV)	slope (PAV)
MLP 4:4-14-4:1	3	4		1		2	
RBF 7:7-23-1:1	4	1	6	2	7	3	5

Table 29: Sensitivity analysis results.

The importance of the features describing boundary tones presented here differs to some extent from that shown by discriminant analysis results and in the importance ranking of predictor variables computed in the decision tree analysis, in that the f0 features are of high importance for detection of boundary location. The rankings differ also between the two networks, which means that they use different features to detect +b and -b vowels. The MLP network performs the classification on the basis of 4 features only: the most important is tilt(SAV) and f0mean(SAV) and unlike in the previous experiments, the duration features have lower importance. The RBF network uses most of all syllable relative duration and also tilt and f0mean as predictors.

6.2.5. Conclusions

The goal of the analyses presented in the previous sections was to propose methods capable of identifying the location of phrase boundaries: this information is used by the recognizer of boundary tone type.

In a series of statistical analyses the effect of boundary presence/absence on variation in acoustic features which can be useful for detection of boundary location was investigated. It was found that features such as *relative duration* of syllable/nucleus and previous syllable nucleus, as well as the amount of *tilt, rising amplitude* and *overall pitch* level on the accented vowel, and the *amount of pitch variation* occurring on the previous syllabic nucleus constitute the most important cues for boundary location. They can be regarded as part of the phonetic description of intonation.

The high prediction accuracy obtained with various statistical classification methods including discriminant function analysis, decision trees and neural networks prove that these features are highly significant for identification of accented syllables. What's more, the results of sensitivity analysis showed that pitch variation is as important as duration for detection of phrase boundary location. These results prove Hypothesis 1a.

The models designed in the current study achieve high accuracy in detection of phrase boundary location. The worst results were achieved with the linear model (discriminant analysis function) where the accuracy of boundary detection was about 74%. The RBF network had the best performance: phrase boundaries were correctly identified in 81.55%.

The performance of the models designed in the current study is in between the worst and the best results achieved by other models presented in the literature (cf. sec.6.1.2). The same conclusion can be drawn on the basis of the comparison with inter-transcriber consistency in identification of phrase boundary location reported in a number of works presented in sec. 6.1.3.

In general, the results presented here confirm Hypothesis 2a & Hypothesis 2b and show that a high accuracy in the identification of phrase boundary location can be achieved even with a small feature vector, which can be regarded as an advantage of the methods proposed in this study compared to 276 features used in (Kießling et al. 1996).

6.3. Automatic classification of boundary tone types

In this section methods for automatic recognition of phrase boundary type are designed. They use the acoustic features distinguished for description of boundary tones on the phonetic level (sec. 5.3.3) to derive the surface phonological description. These features include:

e) ***f0end:*** final f0 value on a phrase boundary syllable. It distinguishes between the two types of rising and falling boundaries (2,? vs. 5,? and 2,. vs. 5,.).
f) ***f0mean(PAV):*** mean f0 on the nucleus of the syllable preceding the current syllable. It is used because of its high correlation with other variables which are important for distinction between phrase boundary types such as syllable/nucleus relative duration and distance to the next pause (in ms).

g) ***dist_#$p:*** distance of the current syllable to the next pause (#$p) marked in the annotation file. It is measured in syllables and distinguishes between boundaries of a different strength (i.e., 5 vs. 2).
h) ***direction:*** is calculated as a difference between mean f0 on the vowel of the syllable preceding the phrase-final syllable and on the nucleus of the phrase-final syllable. It describes pitch movement direction and distinguishes between rising and falling boundaries.

The methods designed in the next sections assume that the location of phrase boundaries is known: in the current experiments this information is known from the prosodic annotation of the speech material, but in could also be provided by the models designed in the previous experiments.

The current study is based on 1502 boundary tones and among them there are 629 instances of ***5,.***, 462 instances of ***2,?***, 304 instances of ***2,.*** and 107 of ***5,?*** boundaries; the third type of falling boundary tones labeled as ***5,!*** was excluded from the analyses due to sparseness of data (only 7 instances of 5,! boundary tones were found in the speech material).

6.3.1. Classification using discriminant analysis function

In this section the results of recognition of boundary tone type with discriminant analysis function are reported. Altogether 1502 instances of boundary tons were used in the analysis and they were split into 2 samples: one (learning) was used to compute the classification functions and the other (cross-validation) for their evaluation in a cross-validation test.

The results computed for the learning and cross-validation samples are given in Table 30.

accent	learning	cross-validation
2,.	75,77	70,13
2,?	67,26	70,63
5,.	97,74	99,30
5,?	92,11	83,87
average%:	83,84	82,45

Table 30: Accuracy of boundary type recognition for learning sample and resulting from the cross-validation test.

It can be seen that the highest recognition accuracy was obtained for 5,. boundary tone. In general, the boundary tones marking the end of major phrases (5) are recognized with a higher accuracy then those signaling minor phrase breaks (2). It can be explained by the fact that the differences between 2,? and 2,. boundaries are less salient than between 5,. and 5,?. This can also be concluded on the basis of analysis of the misclassification matrix which showed that in the cross-validation test 18% of 2,. boundaries were recognized as 2,? and 16,6% of 2,? boundaries as 2,. The 2,? boundary tones are a bit problematic, because they also happen to be

erroneously recognized as 5,. (7%) or 5,? boundaries (5,6% in the CV test). These results show that there are some mismatches regarding the opposition of boundary tones.

The results of discriminant analysis prove that the most important cue of boundary tone type are mean f0 on the pre-boundary vowel (f0mean(SAV), F=176,8) and syllable final f0 (f0end, F=145,2).

The accuracy of boundary tone type prediction achieved with the discriminant function in the current study is comparable to that reported by other authors. In (Ross & Ostendorf 1996) the distinction between boundary tone types was binary i.e., boundaries were classified either as falling or rising. The prediction accuracy of rising boundaries was 62% and of falling boundaries - 88%.

6.3.2. Classification using decision trees

The data available for building and testing the decision tree included 1502 instances of boundary tones. They were split into two samples: training (1132) and test (377); the proportion of training to test cases is about 2:1 in every boundary tone class.

Figure 52 illustrates the importance ranking of predictor variables used in the boundary tone type classification. The importance is depicted on a 0-100 scale: a value of 100 indicates high importance and values near 0 indicate low importance. It can be seen that the most important variables are *f0end* and *f0mean(SAV)*, whereas distance to the following pause and direction of the pitch movement described by *dist_#$p* and *direction* variables are less important features for boundary tone type classification, which is in accordance with discriminant analysis results presented in the previous section.

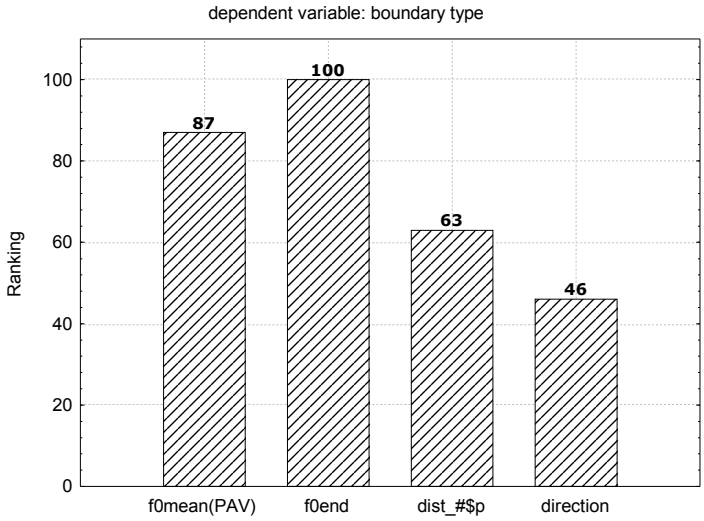


Figure 51: Importance ranking of predictor variables used in the recognition of boundary type

Figure 52 illustrates the structure of the decision tree designed in the current experiment. The tree is very small: it has 9 splits and 10 terminal nodes. It can be seen that the

top-level splits were based on conditions referring to the most important features: f0mean(PAV) and f0end. In the graph the following information is shown:

a) the number of cases in each observed class that are sent to the node
b) the predicted class to which cases sent to the node is assigned
c) split condition for a split node

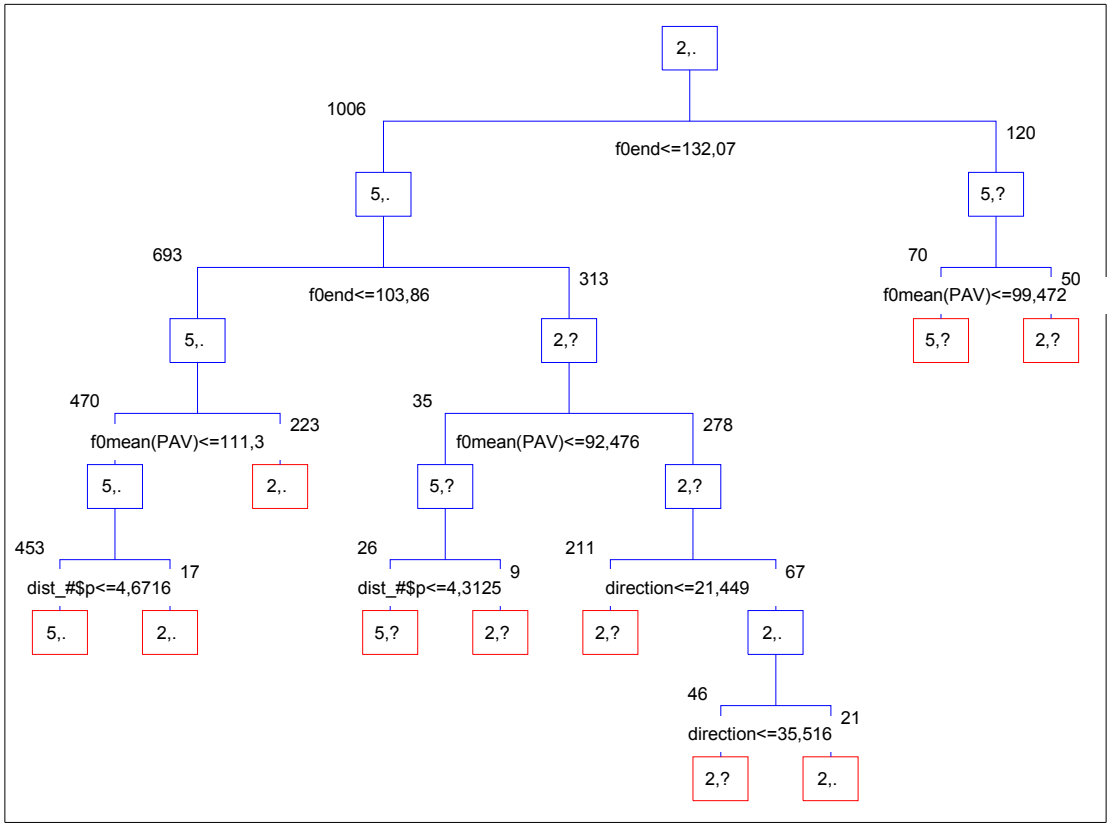


Figure 52: Decision tree designed for phrase boundary type recognition

The results of boundary tone recognition with the tree designed in the current experiment are presented in Table 31. In a) misclassification matrix computed for the training sample is presented , in b) the results based on the test sample are shown and c) gives the results of the cross-validation test.

a) learning sample

boundary	2,.	2,?	5,.	5,?
2,.		58	33	0
2,?	56		1	0
5,.	1	2		0
5,?	0	17	3	
average%: 86,09	74,89	77,08	92,4	100

b) test sample

boundary	2,.	2,?	5,.	5,?
2,.		15	11	0
2,?	19		0	1
5,.	4	5		0
5,?	0	6	0	
average%: 84,63	70,13	79,37	92,25	96,77

c) cross-validation test

boundary	2,.	2,?	5,.	5,?
2,.		67	33	0
2,?	56		1	0
5,.	4	5		0
5,?	1	23	3	
average%: 84,31	73,13	71,73	92,4	100

Table 31: Results of boundary type recognition for learning (a) and test sample (b) and achieved in the cross-validation test (c) carried out on the learning sample

In comparison to the results achieved with the discriminant function analysis, the tree designed in the current experiment performs better as regards the recognition accuracy of 2,? and 2,. boundary tones. It can be seen that the average accuracy of boundary tone type recognition in the training sample is 86% and in the test sample 84,64%; this high accuracy is also confirmed by the results of cross-validation test where the overall average accuracy is 84,31%.

The performance of the decision tree varies with boundary tone type: the best results are achieved for 5,?, whereas 2,. boundaries are correctly recognized in 75%in the training sample and 70% in the test sample. The number of misclassifications of 2,? boundaries (they happen to be recognized as 5,. or 5,? boundary tones) again suggests that some mismatches regarding the opposition of boundary tones is present. But anyway, it should be stated that a misclassification of the sort when a 2,? (rising) boundary is classified as 5,? (also rising) is significantly less serious than if 2,? is recognized as 5,. (falling). In the work by (Ross & Ostendorf 1996) mentioned in the previous section 7% of falling boundaries were recognized as rising, and 24% rising boundaries were recognized as falling. In the current study the percentage of faulty recognized falling boundaries as rising is about 8%, but rising boundaries are classified as falling significantly less often: 1,5% in the training and 1,3% in the test sample.

6.3.3. Classification using neural networks

Like in the previous sections where boundary tones were recognized with discriminant analysis function and decision tree, the dataset used in the current analysis consisted of 1502 instances of boundary tones split into training (1132) and test (377) samples.

The networks summarized in Table 32 were created with the help of a neural network designer available in the *Statistica* program. MLP, linear and RBF networks were selected for the task of boundary tone recognition.

In order to determine an effective configuration for the networks a number of experiments was carried out in which various configurations were tested and the results were computed for different sample settings. The highest recognition accuracy was achieved for MLP network with 20 hidden units and RBF network with 54 hidden units.

The summary of the two networks is given in Table 32. Starting from the left column the following information is provided:

a) network type and structure (i.e., the number of input, hidden and output units)
b) performance of a network on the training/test sample; the performance measure is the proportion of cases correctly classified
c) the error (RMS) of the network on the training/test sample

network	perf. training	perf. test	RMS training	RMS test
RBF 4:4-54-4:1	0,88	0,88	0,21	0,23
MLP 4:4-20-4:1	0,87	0,85	0,55	0,97

Table 32: Model summary details

It can be seen that the RBF network performs better than the MLP network: the average accuracy of boundary tone type recognition is 88% in the training and in the test sample. The network has also significantly lower value of the RMS error. The MLP network also achieves high recognition accuracy: 87% in the training and 85% in the test sample, but the network error computed for the test sample is much higher in comparison to the error on the training sample, which may indicate over-learning.

Table 33 and Table 34 summarize the networks performance based on the test sample: in Table 33 results achieved with the MLP network are presented; in Table 34 the performance of the RBF network is summarized.

It can be observed that with the two networks the best recognition accuracy was obtained for 5,. boundaries. The other boundary tone types are more often correctly identified with the RFB than MLP network. Unlike in the previous classification with discriminant function analysis the percentage of correctly recognized 2,? boundaries are very high and much higher than that of 2,. boundaries. As a matter of fact, for the latter boundary tone type the worst recognition accuracy can be observed.

MLP 4:4-20-4:1	2,?	5,.	2,.	5,?
total:	126	143	77	31
correct	101	141	51	29
missclass.	25	2	26	2
correct%	80,16	98,6	66,23	93,55
missclass.%	19,84	1,4	33,77	6,45

Table 33: Summary statistics: MLP network.

RBF 4:4-54-4:1	2,?	5,.	2,.	5,?
total:	126	143	77	31
correct	107	141	54	30
missclass.	19	2	23	1
correct%	84,92	98,6	70,13	96,77
missclass.%	15,08	1,4	29,87	3,23

Table 34: Summary statistics: RBF network.

In order to investigate the importance of the input variables for the boundary tone recognition results a sensitivity analysis was carried out. Table 35 shows importance ranking of the predictor variables computed in the analysis: 1 indicates the highest importance.

network	f0mean (PAV)	f0end	dist_#$p	direction
RBF 4:4-54-4:1	1	3	2	4
MLP 4:4-20-4:1	1	3	4	2

Table 35: Sensitivity analysis results.

The importance of the features describing boundary tones presented here is similar to that shown by discriminant analysis results and in the importance ranking of predictor variables computed in the decision tree analysis. The most important feature is mean f0 on the pre-boundary vowel - *f0mean(PAV)* variable. As mentioned before this feature distinguishes not only rising boundaries from the falling ones, but also the two types of rising (2,? vs. 5,?) and falling (2,. vs. 5,.) boundaries. Syllable-final f0 (*f0end*) is less important than in the previous analyses, and gave priority to *dist_#$p* (in the RBF network) and *direction* (in the MLP network). The analysis of a misclassification matrix showed that like in the previous analyses most of misclassifications in the 2,? and 2,. boundary tone groups result from faulty recognition of 2,? as 2,. boundary and vice versa.

6.3.4. Conclusions

The goal of the analyses presented in the previous sections was to propose methods for automatic recognition of boundary type. The recognition is performed at the word level and on the basis of a small vector of acoustic features which constitute description of boundary tones on the phonetic level. From this information a surface phonological description is automatically derived. The features used by the models include:

a) *f0end* - pitch level at phrase boundary
b) *f0mean(PAV)* - overall pitch level on the nucleus of the previous syllable
c) *direction* - direction of the pitch movement occurring at the phrase end
d) *dist_#$p* - distance to the next pause measured in the number of syllables.

The importance of particular features for identification of boundary tone types differs between the models. For example, the classification tree relied mostly on the *f0end* parameter, whereas this feature was of a secondary importance for the MLP and RBF networks which used predominantly the *f0mean(PAV)* parameter to identify boundary tone type.

The high recognition accuracy achieved with different statistical classification methods (including discriminant analysis function, decision trees, neural networks) confirms Hypothesis 1a, because it proves that these features are highly significant for identification of boundary tone types.

As regards performance of the designed models the worst results were achieved with the linear model (discriminant analysis function): overall accuracy of boundary tone identification was 82.45% in the cross-validation test. The RBF network had the best performance: boundary tones were correctly identified in almost 88%. These results are comparable to the consistency

in boundary type identification achieved by human labelers (cf. sec. 6.1.3), which confirms Hypothesis 2a.

The performance of the models is also comparable to that reported by other authors (e.g. Wightman & Ostendorf 1994, Kießling et al. 1996, Sridhar, Bangalore & Narayanan 2007, Rosenberg 2005), even though the recognition of boundary tone types was based on four features solely. These results confirm Hypothesis 2b according to which a high accuracy in the automatic detection and classification of intonational events can be achieved even if only a small vector of features derived from utterance's acoustics and transcription/segmentation is used.

6.4. Detection of accented syllable location (accentual prominence)

The goal of the analyses presented in this section is to propose a method capable of identifying the location of accented syllables on the basis of a small feature vector which can be easily derived from utterance's acoustics. Apart from that, the models designed in the current study rely on information regarding syllable and vowel boundary, and lexical stress. Two experimental settings are taken into account. At first, the detection of accentual prominence is carried out on the syllable level: each syllable can be identified as accented or not on the basis of the features provided at the input to the model. In the second experiment, the detection is made on the word level: only stressed syllables are taken into account, the information on stress is provided in the automatic prosodic annotation (see sec.4.2.2),. The second approach has this advantage that only one syllable per word can be identified as accented.

In order to identify the best acoustic cues of accentual prominence a number of analyses was carried out which investigated the effect of accent presence/absence on a syllable on a number of acoustic features. The results are presented in the next section. Then, the results of classification of accented/unaccented syllables using various statistical classification techniques (discriminant function analysis, decision trees and neural networks) are reported.

6.4.1. Features

Accents serve as cues to syllable's prominence, therefore the first step towards detection of accents in the string of syllables is determination of the acoustic correlates of *accentual prominence*. As mentioned in the sec. 6.1.1 pitch variation (pitch movements) and overall intensity are usually indicated as the main acoustic correlates of accents (Jassem 1961, Streefkerk 1997, Sluijter & van Heuven 1996, Tamburini 2006). Therefore, f0 features are the primary ones used in the detection of accents. Since accents are associated with stressed syllables in the segmental string of the utterance, duration features are important for this task as well. Apart from that, overall *syllable energy* is commonly used, but the results reported in (Rapp, 1996) show that this factor is not essential and does not improve the classification results significantly.

In view of these findings in a series of ANOVA and discriminant function analyses the effect of accent presence vs. absence on variation in acoustic features listed in sec. 5.3.2 was

investigated. It was found that the following features can be regarded as the best acoustic correlates of the presence of accentual prominence:

a) ***syllable and nucleus duration:*** the results of a comprehensive analysis of vowel duration presented in (Demenko 1999) proved that accented vowels are significantly longer then the unaccented ones (on average 15-17ms). The relative measures of duration are proposed to eliminate the effect of syllable structure and vowel type on the observed duration.
b) ***f0max:*** the level of the maximum pitch on the syllable. The height of the peak correlates positively with perceived prominence of the syllable. Since pitch accents are prominence-lending this feature is expected to discriminate well between accented vs. unaccented syllables. It is significantly correlated with the f0mean(SAV) variable (mean f0 on the vowel).
c) ***tilt:*** at first sight, there may seem to be no reason for distinction between accented/unaccented syllables based on the tilt value of the syllable/vowel. Yet, since rising pitch movements are more important for perception of accentual prominence (see e.g. Tamburini 2005, 2006) then falling pitch movements, tilt which describes shape of an f0 contour and the amount of rise and fall in pitch may be very useful. Besides, it is significantly correlated with a number of other "meaningful" parameters including normalized peak position, overall f0 level on the vowel, rising amplitude on the syllable/vowel. Tilt representation was used as a measure of pitch variation in (Tamburini 2003): by multiplying the event amplitude (EvAmp) by its duration (EvDur) and a further factor that expresses the relevance of the event along the utterance (EvRel) good prominence detection can be obtained. The tilt parameter used in this study is calculated as described in sec. 5.3.2.
d) ***slope:*** is a measure of the overall pitch variation on the syllable/vowel. Greater pitch variation is expected to occur on accented than unaccented syllables and vowels.

Firstly, the analyses were performed on the syllable level i.e. both stressed and unstressed syllable were taken into account, altogether 15566 instances. As mentioned in the sec. 6.4.1 the features discussed here are determined for each syllable and vowel. Since they are highly correlated a decision has to be made which one is a better acoustic cue of accent. Both ANOVA and discriminant analysis results showed that syllable-based features (i.e., features determined for a given syllable) are affected to a greater extent then vowel-based features by the presence of accent and also discriminate better between unaccented and accented syllables. As regards duration features relative measures are used. In view of the findings concerning the effect of syllable structure and vowel type on duration reported in (Demenko 1999) and mentioned earlier in this section it seems important to use the relative measures which abstract away from these effects. ANOVA results shows the presence/absence of an accent has significant effect ($p<0.01$) on pitch variation on the syllable (described by *tilt, slope* and *f0max* parameters) and duration of the syllable/vowel (here expressed by the variables *syl_dur(relative)* and *nucl_dur(relative).*

The correlations between the variables were investigated in order to avoid redundancy: the results are depicted Table 36. Values of the correlation coefficient (*r*) close to 1.0 indicate positive correlation, values near -1.0 indicate negative correlation and values close to 0.0 indicate little correlation. It can be seen that with the two exceptions (slope and f0max, syllable

and nucleus duration) other values of the correlation coefficient are close to 0: this means that there is no redundancy in the representation of pitch and duration variation and consequently, that information provided the selected features will contribute to the detection of accents.

feature	syl_dur (relative)	nucl_dur (relative)	tilt	slope	f0max
syl_dur(relative)	1,00				
nucl_dur(relative)	0,59	1,00			
tilt	0,12	0,09	1,00		
slope	0,03	0,06	0,10	1,00	
f0max	-0,06	-0,14	0,06	0,61	1,00

Table 36: Correlation matrix of f0 and duration parameters used in the detection of accented syllables

Table 37 shows distribution of mean values and standard deviations of f0 and duration parameters. It can be seen that accented syllables and vowels are on average 10 ms longer than the unaccented ones. They have also significantly higher average slope, which indicates more pitch variation on the accented syllables. Besides, accented syllables have higher f0 peaks and higher value of the tilt parameter (0.11 vs. -0.41), which shows that most pitch accents involve both rising and falling pitch movement and the amount of rise is greater then that occurring on unaccented syllables. Since, in general rising pitch movements are more relevant for prominence then pitch falls the tilt parameter may be an important predictor of accentual prominence. The effects discussed here are statistically significant ($p<0.01$).

a)

+/-acc	N	syl_dur(relative)		nucl_dur(relative)	
		mean	ST	mean	ST
+acc	3926	1,04	0,25	1,08	0,34
-acc	11640	0,93	0,34	0,97	0,43
in total:	15566	0,96	0,32	1,00	0,42

b)

+/-acc	tilt		slope		f0max	
	mean	ST	mean	ST	mean	ST
+acc	0,11	0,82	131,18	84,12	125,74	18,55
-acc	-0,41	0,77	78,57	63,01	115,54	16,79
in total:	-0,28	0,81	91,84	72,63	118,12	17,81

Table 37: Mean and ST of duration a) of syllables/vowels and pitch variation b) depending on the accent presence/absence on the syllable

As mentioned at the beginning of this section, in the analyses on the word level only stressed syllables are taken into account. In this experimental setting the comparison of features of unaccented vs. accented syllables shows significant differences in all f0 and duration parameters ($p<0.01$). Again, no significant correlations can be observed between the variables. Table 38 shows means and standard deviations of f0 and duration parameters describing features of accented vs. unaccented syllables. In general, the distribution of means of f0 parameters is almost the same as that observed in the previous experiment where the analysis

was performed on the syllable level. But in comparison to the previous analysis, in this experiment a greater difference in syllable/vowel duration between accented vs. unaccented syllables (20 ms) can be observed.

a)

+/-acc	N	syl_dur(relative)		nucl_dur(relative)	
		mean	ST	mean	ST
+acc	3926	1,04	0,25	1,08	0,34
-acc	2491	0,83	0,32	0,88	0,34
in total:	6417	0,96	0,30	1,00	0,35

b)

+/-acc	tilt		slope		f0max	
	mean	ST	mean	ST	mean	ST
+acc	0,11	0,82	131,18	84,12	125,74	18,55
-acc	-0,41	0,77	73,39	58,05	115,40	12,92
in total:	-0,09	0,84	108,74	80,18	121,73	17,34

Table 38: Mean and ST of duration of syllables/vowels a) and pitch variation b) depending on the accent presence/absence on the stressed syllable

Irrespective of the experimental setting ANOVA results prove that the most significant difference between accented vs. unaccented syllables is in f0 features described by tilt, slope and f0max. Additionally, in the dataset including only stressed syllables the effect of accent position on syllable duration can also be observed, which confirms the findings reported in (Demenko 1999). These effects are depicted in Table 39.

a)

variable	F	p
syl_dur(rel.)	329,53	
nucl_dur(rel.)	184,91	
tilt	1288,26	<0.01
slope	1709,22	
f0max	1025,49	

b)

variable	F	p
syl_dur(rel.)	770,27	
nucl_dur(rel.)	489,45	
tilt	648,47	<0.01
slope	902,65	
f0max	591,40	

Table 39: ANOVA results: the effect of accent presence on variation in f0 and duration features of syllables; a) syllable-level analysis, b) word-level analysis.

The features represented by variables: *syl_dur(relative), nucl_dur(relative), tilt, slope* and *f0max* are used as predictors of accentual prominence presented in the following sections.

6.4.2. Detection with discriminant function analysis

In this sections the results of detection of accentual prominence with a linear model - discriminant function analysis are reported. Discriminant function analysis offers a method for fitting linear models with categorical dependent variables and continuous predictors. Two experimental settings are taken into account. At first, the detection is performed on the syllable level and both stressed and unstressed syllables are classified. Secondly, the model is designed which performs the detection on the word level (only stressed syllables are used).

1. Detection on the syllable level.

Altogether 15566 syllables were taken into account in the analysis and they were divided into 2 samples: one (learning) was used to compute the classification functions and the other (cross-validation) for their evaluation in a cross-validation test. Prior classification probabilities were specified as equal for each dependent group (i.e., +acc, -acc). The results computed for the training and testing samples (using cross-validation test) are given in Table 40:

a)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	63,15	1232	719
-acc	75,63	1424	4420
in total:	72,51	2656	5139

b)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	64,05	1265	710
-acc	74,46	1476	4303
in total:	71,81	2741	5013

Table 40: Classification matrix learning a) and cross-validation sample b)

These results are well below those reported in the literature (e.g. Tamburini 2006, Rapp 1996, Wightman et al. 1994, Bulyko & Ostendorf 2001) but as a matter of fact, in the current study a different classification method is used.

High percentage (ca. 80%) of correct classifications of syllables into accented and unaccented in continuous speech based on vowel features (including duration, intensity and f0 variation) are reported in (Demenko 1999). Therefore, another discriminant analysis was carried out in which the syllable-based features were replaced by vowel-based parameters. Apart from syllable and vowel duration for the current classification the following features were selected:

a) overall amplitude: computed as in the Tilt model (see sec. 5.3.2). Alternatively, the tilt value on the vowel could be used, but it shows higher correlation with the other two parameters, which are:
b) a1: steepness of the rise
c) a2: steepness of the fall; like the *slope* feature steepness is the measure of pitch variation on the vowel
d) f0mean: was used for the same reasons as *f0max* in the previous analysis, what is more it shows no significant correlation with other variables

All these features differ significantly ($p<0.01$) between accented and unaccented vowels, which was proven in the ANOVA analysis. The most significant effect of accent presence is observed for *amplitude* ($F=439$) and *f0mean* ($F=292.05$) which shows that the main acoustic correlate of accent is pitch variation. There are no significant correlations between them and syllable-based features used in the previous analysis (the exception is f0mean and f0max, but the latter is more strongly correlated with the other parameters used in the current classification).

The results are depicted in Table 41 and are comparable to those obtained in the previous analysis, thus the use of vowel- instead of syllable-based features does not affect the classification.

a)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	63,15	1232	719
-acc	75,63	1424	4420
in total:	72,51	2656	5139

b)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	64,05	1265	710
-acc	74,46	1476	4303
in total:	71,81	2741	5013

Table 41: Classification matrix for learning a) and cross-validation sample b) based on 17317 cases, missing data was deleted pairwise.

2. Detection on the word level.

The detection of accentual prominence described here was based on 6417 stressed syllables available in the speech corpus. The stress position was assigned from rules and it was done automatically using the *Annotation Editor* software.

The data was divided into two samples used for computing the classification functions and for evaluation of classification results. The variables used in the analysis included: *relative nucleus* and *syllable duration* and syllable-based pitch features: *tilt, slope* and *f0max*. The most significant effect of the presence of accent is observed for *tilt* (F=201.77) and *duration* (syllable: F=125.23, nucleus: F=103.39). The large difference between the F test values for the f0 and duration features shows that pitch variation is a stronger acoustic correlate of accentual prominence then duration.

The percentage of correct classifications in the learning (a) and cross-validation samples (b) is given in the table below.

a)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	73,45	1433	518
-acc	79,79	251	991
in total:	75,92	1684	1509

b)

class	%correct	+acc p=0.05	-acc p=0.05
+acc	74,89	1479	496
-acc	80,94	238	1011
in total:	77,23	1717	1507

Table 42: Classification matrix for learning a) and cross-validation sample b)

The results are much better than those obtained in the previous analysis where both unstressed and stressed syllables were used. They are also closer to those reported in the literature, but as a different method is used here to detect the accentual prominence it is hard to make any comparisons. It can be expected that non-linear methods such as classification trees and neural networks that are extensively used to solve classification problems will yield better results.

6.4.3. Detection with classification trees

Classification trees are the state-of-the-art method applied to solve classification (and regression) problems.

In the current study the QUEST classification tree algorithm (Loh & Shih 1997) available in *Statistica 6.0* (StatSoft, Inc. 2001) is used. QUEST uses *discriminant-based univariate splits* as a split method. It has a number of innovative features for improving the reliability and efficiency of the classification trees that it computes. It is faster and less unbiased

then other programs (e.g. C&RT, Breiman et al. 1984) particularly in situation when 1) predictor variables have dozens of levels or 2) some predictor variables have many levels while other have only few. Moreover, the speed of the QUEST algorithm does not affect its predictive accuracy.

Like in the previous section at first, an attempt is made to identify accented syllables from among stressed and unstressed syllables and then, only stressed syllables are taken into account.

1. Detection on the syllable level.

The data available for building and testing the classification tree included 15549 syllables. They were randomly divided into 2 samples: learning (including 6269 and 2139 unaccented/accented syllables respectively) - used to build the tree, and testing (3869 unaccented and 1308 accented syllables) - used for evaluation. Additionally, a cross-validation test was run on the learning sample.

As a result of a tree building a tree with 65 splits and 66 terminal nodes was obtained. Table 43 displays the information on the 10 first splits:

a) The left branch and right branch child nodes to which cases or objects are sent if they satisfy, or do not satisfy, respectively, the split condition at a split node.
b) The number of cases or objects in each observed class that is sent to the node.
c) The predicted class to which cases sent to the node is assigned.
d) Information detailing the split condition for a split node

node	left branch	right branch	n.class +acc	n.class -acc	predicted class	split condition	split variable
1	2	3	2618	7754	+acc	-109,35	slope
2	4	5	1188	5896	-acc	-127,84	f0max
3	6	7	1430	1858	+acc	0,81	tilt
4	8	9	838	5312	-acc	-86,87	slope
5	10	11	350	584	+acc	0,91	tilt
6	12	13	297	1174	-acc	-1,14	nucl_dur
7	14	15	1133	684	+acc	0,38	tilt
8	16	17	623	4712	-acc	-1,00	syl_dur
9	18	19	215	600	+acc	-0,97	nucl_dur
10	20	21	95	314	+acc	-58,99	slope

Table 43: Table displaying the classification tree structure

It can be seen that the first five split conditions for the top split nodes are based on f0 features of syllables, which means that in the identification of accent pitch variation plays much more important role then duration. Similar conclusion can be drawn on the basis of importance ranking depicted on a 0-100 scale for each predictor variable in Figure 53 - value of 100 indicates high importance and values near 0 indicate low importance. It can be seen that f0 features are significantly higher in the ranking then duration features, but the latter are important too.

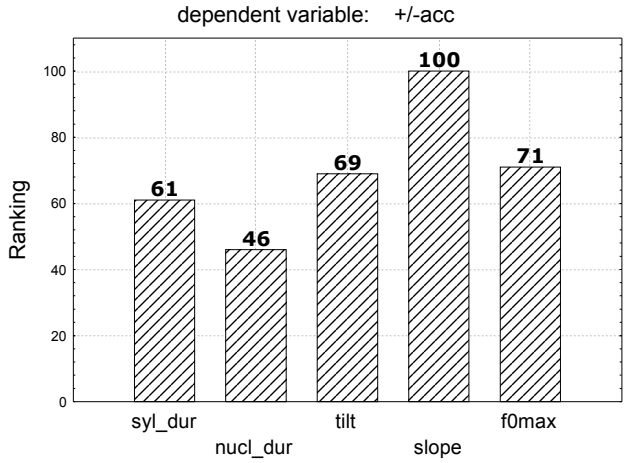


Figure 53: Importance ranking of accentual prominence predictors (syllable level)

The identification of an accent with the classification tree is accurate in almost 78%: this result is much better then that obtained with a linear model (general discriminant function) presented in the previous section. The results are also comparable to those reported by other authors e.g. in (Wightman et al.2000) the recognition accuracy for accented vs. unaccented syllables varied from 69.3% to 82,8% depending on the experimental setting.

The misclassification matrix for learning a) and test b) samples and results of cross-validation test based on the learning sample c) are depicted below. Global cost indicates the amount of misclassifications. It can be observed that % of correct classifications are the highest in the test sample; the results computed for the learning sample are slightly better then cross-validation results.

a) global cost=0,23 ST=0,004

class	+acc	-acc
+acc		1485
-acc	479	
% correct	77,61	76,31

b) global cost=0,21 ST=0,006

class	+acc	-acc
+acc		780
-acc	289	
% correct	77,91	79,84

c) global cost=0,21 ST=0,004

class	+acc	-acc
+acc		1619
-acc	567	
% correct	73,49	74,17

Table 44: Misclassification matrix for the learning a) and test sample b). In c) results of global cross-validation based on the learning sample are given.

The results presented in this section show that pitch variation is the main acoustic correlate of accentual prominence and that non-linear modeling techniques are more efficient in the identification of accentual prominence from acoustic features then linear models.

2. Detection on the word level.

This section reports on the results of building a classification tree capable of making distinction between stressed accented and unaccented syllables. The data used for building and testing of the tree consisted of 6417 syllables. Like previously, they were divided into 2 samples: learning (including 2618 accented and 1660 unaccented syllables) - used to build the

tree and testing (1308 and 831 accented and unaccented syllables respectively) - used for evaluation. Additionally, a cross-validation test was run on the learning sample.

As a result of the tree building a tree including 34 splits and 35 terminal nodes was obtained. Table 45 displays the information on the 10 first splits:

Node	left branch	right branch	n.class +acc	n.class -acc	predicted class	split condition	split variable
1	2	3	2618	1660	+acc	-112,89	slope
2	4	5	1235	1353	-acc	-1,02	syl_dur(relative)
3	6	7	1383	307	+acc	0,07	tilt
4	8	9	628	1050	-acc	-130,70	f0max
5	10	11	607	303	+acc	-35,39	slope
6	12	13	387	221	+acc	-1,07	nucl_dur(relative)
7	14	15	996	86	+acc	-105,12	f0max
8	16	17	432	996	-acc	0,19	tilt
9	18	19	196	54	+acc	0,11	tilt
10	20	21	120	106	-acc	-123,35	f0max

Table 45: Table displaying the classification tree structure

It can be seen that unlike in the previous analysis where both stressed and unstressed syllables were included, in the current tree the second split condition is based on syllable duration. As mentioned before (sec. 6.1.1) increased syllable/vowel duration can be regarded as one of the most important acoustic correlates of stress, but the effect of lengthening is even greater if the syllable is accented: in our database the accented syllables/vowels are on average 20 ms longer then the stressed unaccented ones and this difference is statistically significant ($F=443.58$, $p<0.01$). On the other hand, it can be seen that most of the splits are conditioned by f0 features of syllables i.e., features describing the amount of pitch variation which is the main correlate of accentual prominence. Figure 54 shows importance ranking of the predictor variables on a 0-100 scale; a value of 100 indicates high importance and values near 0 indicate low importance. It can be seen that from among f0 features *f0max* (describing height of f0 peak) is the most important predictor of accentual prominence.

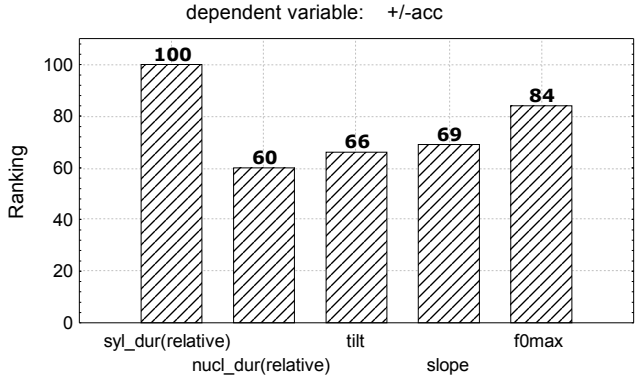


Figure 54: Importance ranking of accentual prominence predictors (word level)

The misclassification matrix for learning a) and test b) samples and results of cross-validation test based on the learning sample c) are depicted below. Global cost indicates the amount of misclassifications. The highest accuracy of accent detection was obtained in cross-validation (c); in the learning sample the accented syllables were correctly identified in 79,14% and in the test sample - in 77,06%.

a) global cost=0,2; ST=0,006

class	+acc	-acc
+acc		350
-acc	546	
%correct	79,14	78,9

b) global cost= 0,21; ST=0,009

class	+acc	-acc
+acc		156
-acc	300	
%correct	77,06	81,2

c) global cost =0,21; ST = 0,006

class	+acc	-acc
+acc		282
-acc	503	
%correct	80,79	83,0

Table 46: Misclassification matrix for the learning a) and test sample b).
In c) results of global cross-validation based on the learning sample are given.

The results presented here are again very close to those reported by other authors: for example, in (Bulyko & Ostendorf 2001) the prediction accuracy of accent location is around 74% when only acoustic- or only text-based features are used and increases to 80,9% when using both. For comparison, inter-human agreement when manually tagging prominence in continuous speech is around 80% (Pitrelli, Beckman & Hirschberg 1994, Syrdal & McGory 2000). It means that the automatic identification of accentual prominence with classification trees which has only slightly lower accuracy is very promising and can find application in automatic labeling of speech corpora.

6.4.4. Detection with neural networks

Another method applied to identification of accented syllables is neural networks. Like classification trees they belong to non-linear modeling techniques, their features have been discussed in section 4.4.2.

Like before, two datasets were used: 1) including stressed accented/unaccented, and unstressed syllables and 2) including only stressed syllables. The features used for accent prediction include like previously relative syllable (syl_dur) and nucleus duration (nucl_dur), tilt, slope and maximum f0 on the syllable (f0max).

Three types of networks have been selected for classification of syllables into accented vs. unaccented:

a) LN - linear network which in *Statistica* is trained using the standard pseudo-inverse (SVD) linear optimization algorithm (Golub and Kahan, 1965). The reason behind using linear models is that a simple model should always be chosen in preference to a complex model if the latter does not fit the data better
b) MLP - a multilayer perceptron with 3 layers and varying number of units in the hidden layer
c) RBF network with a varying number of units in the hidden layer

The networks were built using *Intelligent Problem Solver* program available in *Statistica*. It guides user through design of a network and offers possibility of selecting from among a number of options (e.g. specifying types and complexity of networks to be created or classification thresholds), which gives a finer control over the design. From among the networks created with the *Intelligent Problem Solver* from each network type the one with the lowest error on the test sample was retained. In this way it was possible to compare the performance of networks of a different type.

1. Detection on the syllable level.

The dataset used for design of the networks consisted of 15549 syllables. They were divided into the training and test samples in a proportion of 2:1. The summary of the designed networks and their performance is given in Table 47. Starting from the left column the following information is provided:

a) network type and structure (i.e., the number of input, hidden and output units)
b) performance of a network on the training/test sample; the performance measure is the proportion of cases correctly classified
c) the error (RMS) of the network on the training/test sample

network	perf. training	perf. test	RMS training	RMS test
MLP 5:5-17-1:1	0,82	0,81	0,65	0,71
Linear 5:5-1:1	0,69	0,69	0,39	0,39
RBF 5:5-82-1:1	0,83	0,81	0,32	0,33

Table 47: Model summary details.

It can be seen that MLP and RBF networks perform much better then the linear model. The overall classification accuracy in the training sample was 83% (MLP) and 82% (RBF) and 81% in the test sample. These results are better than those obtained with the decision tree (sec. 6.4.3) and show efficiency of neural networks in classification tasks. The linear model yielded worse results: the overall accuracy in both training and test samples 69%. It can also be observed that the RBF network has the lowest RMS error in both training and test sample, while the MLP network has the highest error. The error of the network on the test sample carries a very important information: if it is considerably higher compared to the error of the network on the training sample it may indicate that the network starts to fit too much to the data and loses the capability to generalize (so called over-learning or over-fitting); in such case the results should be treated carefully. Table 48 shows detailed breakdown of misclassifications and prediction accuracy in the accented (+acc) and unaccented (-acc) class. In the training sample the MLP and RBF networks have higher accuracy of prediction of accented syllables then of unaccented syllables; the linear network yields the same % of correct classifications in the accented and unaccented class. In the test sample, the MLP and RBF networks perform better for identification of unaccented then accented syllables.

a) training sample

model	MLP 5:5-17-1:1		Linear 5:5-1:1		RBF 5:5-82-1:1	
class:	+acc	-acc	+acc	-acc	+acc	-acc
total:	2618	7754	2618	7754	2618	7754
correct	2180	6346	1786	5330	2176	6393
missclass.	438	1408	832	2424	442	1361
correct%	83,27	81,84	68,22	68,74	83,12	82,45
missclass.%	16,73	18,16	31,78	31,26	16,88	17,55

b) test sample

model	MLP 5:5-17-1:1		Linear 5:5-1:1		RBF 5:5-82-1:1	
class:	+acc	-acc	+acc	-acc	+acc	-acc
total:	1308	3869	1308	3869	1308	3869
correct	1031	3157	915	2662	1034	3182
missclass.	277	712	393	1207	274	687
correct%	78,82	81,60	69,95	68,80	79,05	82,24
missclass.%	21,18	18,40	30,05	31,20	20,95	17,76

Table 48: Model summary statistics based on the test sample: a) MLP, b) RBF network

In order to investigate the importance of the input variables for the classification results sensitivity analysis was carried out. Table 49 shows importance ranking of the predictor variables computed in the analysis: 1 indicates the highest importance.

network	tilt	slope	f0max	syl_dur	nucl_dur
MLP 5:5-17-1:1	2	1	3	4	5
Linear 5:5-1:1	1	2	3	4	5
RBF 5:5-82-1:1	1	3	2	4	5

Table 49: Sensitivity analysis results

It can be seen that from among f0 features *tilt* is the most important predictor of accentual prominence; the duration features are low in the ranking. These results are in accordance with those obtained in the classification tree analysis and show that pitch variation is the most important acoustic cue of accentual prominence.

2. Detection on the word level.

The three networks (MLP, linear and RBF) trained on the dataset including stressed and unstressed syllables were used for identification of accented syllables in the dataset consisting of all stressed syllables (6417 cases). In this way the performance of the previously designed networks was tested, the results are presented in Table 50.

model	MLP 5:5-17-1:1		Linear 5:5-1:1		RBF 5:5-82-1:1	
class:	+acc	-acc	+acc	-acc	+acc	-acc
total:	3926	2491	3926	2491	3926	2491
correct	3211	2034	2701	1856	3210	2046
missclass.	715	457	1225	635	716	445
correct%	81,79	81,65	68,8	74,51	81,76	82,14
missclass.%	18,21	18,35	31,2	25,49	18,24	17,86

Table 50: Model summary statistics.

It can be seen that like in the previous analysis MLP and RBF networks yielded much better results then the linear model. The overall classification accuracy of the two networks is about 82%, whereas the linear model performs with overall 72% accuracy (as it can be seen correct identification of unaccented syllables is much higher then that of accented syllables). These results show high efficiency of the designed networks in solving classification problems.

Table 51 shows importance ranking of the predictor variables resulting from sensitivity analysis: 1 indicates the highest importance. It can be seen that features describing pitch variation (tilt, slope and f0max) have higher position in the ranking compared to duration features, which means that they play much more important role in the identification of accentual prominence.

network	tilt	slope	f0max	syl_dur	nucl_dur
MLP 5:5-17-1:1	2	3	1	5	4
Linear 5:5-1:1	1	2	4	3	5
RBF 5:5-82-1:1	1	3	2	4	5

Table 51: Sensitivity analysis results.

6.4.5. Conclusions

As stated at the beginning of sec. 6.4 the goal of the analyses presented here was to propose a method capable of identifying accentual prominence from acoustic parameters derived from the speech signal. For that purpose in a series of statistical analyses it was investigated which acoustic parameters can be useful for detection of accents. It was found that syllable and nucleus duration, tilt value describing the shape of the f0 contour on a syllable, height of the peak and slope describing the amount of pitch variation on a syllable can be regarded as the main acoustic correlates of accentual prominence. These features can be regarded as a part of description of intonation on the phonetic level. The high prediction accuracy obtained with various statistical classification methods including discriminant function analysis, decision trees and neural networks prove that these features are highly significant to the identification of accented syllable location in the utterance, which confirms Hypothesis 1a. What's more, the importance ranking of predictor variables and sensitivity analyses confirmed the findings presented in the literature (cf. sec. 6.1.1), namely that pitch variation is a more important acoustic cue of accentual prominence then duration.

The comparison of the performance of various classification methods shows that MLP and RBF networks yield the best results with overall prediction accuracy between 81%-82% in the test sample. The results are the same for the two experimental settings.

Classification trees achieve lower accuracy: the best result is 80.79% accuracy in the detection of accented syllables (cross-validation test, word-level prediction). The syllable-level prediction yields lower accuracy: 77.91% in the test sample.

Linear models have the worst performance and the results depend on whether the prediction is made on the syllable or word level (the latter yields 77.23% accuracy in the global cross-validation test). It can be concluded that non-linear modeling techniques are more effective in solving classification problems then linear ones and that word-level predictions yield better results.

6.5. Automatic classification of pitch accent types

The goal of the analyses presented in this section is to provide a method capable of deriving a surface phonological description of pitch accents from the phonetic description. On the surface phonological level five pitch accent types are distinguished (cf. sec. 5.2.1): LH*, L*H, LH*L, H*L and HL*. The accent types are distinguished on the basis of perceptually significant features including:

a) direction of a pitch movement - rise vs. fall (e.g. LH* vs. HL*)
b) timing of an f0 peak relative to accented vowel onset (e.g. H*L vs. HL*)
c) range of an f0 change on the vowel: a change in the fundamental frequency which is smaller than a given threshold (e.g. *glissando threshold*, see sec. 3.4.2) will be perceived as a level tone rather than as a pitch movement. A distinction is made between two types of falling and rising accents depending on whether a level pitch (e.g. H*L, L*H) or pitch movement is perceived on the vowel (HL*, LH*)

These features are encoded in the phonetic description of accents (see sec. 5.3.2) which is in terms of the following continuous parameters:

a) ***direction:*** it is calculated as a difference between mean f0 on the accented and post-accented vowel. It describes direction of a pitch movement and distinguishes rising accents from falling accents.
b) ***f0peak(relative):*** measured as a difference between f0max in a two syllable window including accented and post-accented syllable and mean f0 on a phrase
c) ***f0min(relative):*** calculated as a difference between f0mean on a phrase and f0min in a two-syllable window including accented and post-accented syllable
d) ***f0mean(relative):*** meanf0 on accented vowel relative to mean f0 on a phrase. This feature is used to distinguish between L* and H* accents e.g., L*H vs. LH*
e) ***Amp(SAV):*** amplitude on the accented vowel, a value of -0,5 indicates fall, a value of 0,5 indicates rise, values in between indicate some amount of rise and fall occurring on the accented vowel. It distinguishes not only between falling vs. rising vs. LH*L accents, but also between two different types of falling (H*L vs. HL*) and rising pitch accents (L*H vs. LH*)
f) ***c1(SAV):*** rising amplitude on the accented vowel

g) ***c2(SAV):*** falling amplitude on the accented vowel; together with c1(SAV) it describes the amount of pitch variation on the vowel and helps to distinguish accents with level pitch perceived on the vowel from those with a pitch movement. Besides, c1(SAV) and c2(SAV) have similar function to Amp(SAV).
h) ***tilt:*** the tilt parameter used in the current analysis is not calculated for a single syllable, but in a two-syllable window including accented and post-accented syllable. This feature describes the shape of the pitch accent and is highly correlated with position of f0 peak (also defined within a two syllable window). Like c1, c2 and Amp it discriminates between falling vs. rising vs. LH*L accents, and also between types of rising and types of falling accents

In the next sections a number of models are designed to recognize the type of accent on the basis of the description on the phonetic level presented above. The recognition will be performed on the word level and it is assumed that the location of accented syllables is given: in the current experiments this information is known from the prosodic annotation of the speech material, but it could also be provided by the models designed in the previous experiments.

Table 52 summarizes the features of the database used in the analyses presented in the following sections. The database consists of 3671 instances of accents; it can be seen that accent types are disproportionately distributed: the most numerous class (H*L) includes 1401 instances, whereas the least numerous one (LH*L) is represented by 96 cases. As a matter of fact, to some extent it reflects the distribution of the classes in the population: there are more prenuclear than nuclear accents and some accent types play specific structural roles in the tune more often then others e.g. H*L and LH* accents occur predominantly in the prenuclear position, whereas HL* and L*H accents can be found most of all in the nuclear position. The data was split into two samples: training (2754) and test (917); the proportion of training to test cases is 2:1 in every pitch accent class.

accent type	training	test	all:
H*L	1063	338	1401
L*H	275	94	369
LH*	789	265	1054
HL*	559	192	751
LH*L	68	28	96
total:	2754	917	3671

Table 52: Distribution of pitch accent types in the database.

6.5.1. Classification using discriminant function analysis

In this section the results of classification of pitch accent types with discriminant function analysis are presented.

The results of a multivariate significance test showed that pitch accent type affects most of all the features described by c2(SAV), tilt and f0max(relative) parameters (F= 259,7, F=156

and F=110,7 respectively). It can be expected these features serve as the best predictor variables of pitch accent type.

The accuracy of pitch accent type classification was assessed in a cross-validation test. The results computed for learning and cross-validation sample are given in Table 53.

accent	learning	cross-validation
H*L	66,98	71,89
L*H	70,91	70,21
LH*	79,09	80,38
HL*	88,01	88,54
LH*L	63,24	85,71
average:	75,02	78,08

Table 53: Classification accuracy computed for the learning sample and resulting from the cross-validation test.

The results of cross-validation test are better (average accuracy: 78.8%) in comparison to those obtained for the learning sample (75.2% correct classifications). In the learning sample the best recognition accuracy was achieved for HL* and LH* accents, whereas LH*L accents are characterized by the worst recognition accuracy. In the cross-validation test the highest percentage of correct classifications was obtained for HL* and LH*L accents.

Analysis of misclassification matrix computed for the training sample showed that in 212 cases (out of 1063, ca. 20%) H*L accents were recognized as LH* accents and in 80 cases (out of 789, ca. 10%) LH* accents were recognized as H*L. It indicates that the distinction between these two accent types may be problematical - either the parameters selected for description of pitch accents are incapable of conveying of the subtle (most of all perceptual) differences between the accents, or there is some inconsistency in the annotation of accent types. Even though the difference between H*L and LH* accents is salient perceptually, in some cases the substitution of one of these accents by the other is not "damaging", because very often H*L and LH* accents have the same role in the tune (they occur predominantly in the prenuclear position and both involve a local maximum) and contribute to conveying of the intonational meaning in a similar way.

As regards misclassifications of other accent types, 4% of the HL* accents was recognized as LH*L accent, 7% as H*L accent and only 1% as L*H accent (which is the most serious misclassification). The L*H accent was most often misclassified as LH* (21%), besides 10 (4%) instances of L*H accents were recognized as H*L accents and 9 (3%) as HL* accents. LH*L accents were most often misclassified as H*L (22%). The misclassification matrix for the cross-validation test shows very similar effects and will not be commented here in detail.

In general, these results show that there are not many serious misclassifications, which occurs when rising accents are recognized as falling and vice versa. Most of the misclassifications occur between accent types which are similar with respect to their realization at the acoustic level and the structural role they play in tunes.

6.5.2. Classification using decision trees

The data available for building and testing the decision tree included 3671 instances of pitch accents. They were split into two samples: training (2754) and test (917); the proportion of training to test cases is 2:1 in every pitch accent class.

Figure 55 illustrates the importance ranking of predictor variables used in the pitch accent type classification. The importance is depicted on a 0-100 scale: a value of 100 indicates high importance and values near 0 indicate low importance. It can be seen that the most important variables are tilt and amplitude: c2(SAV) and Amp(SAV), whereas relative peak and minimum height described by f0max(relative) and f0min(relative) are the least important features for pitch accent type classification.

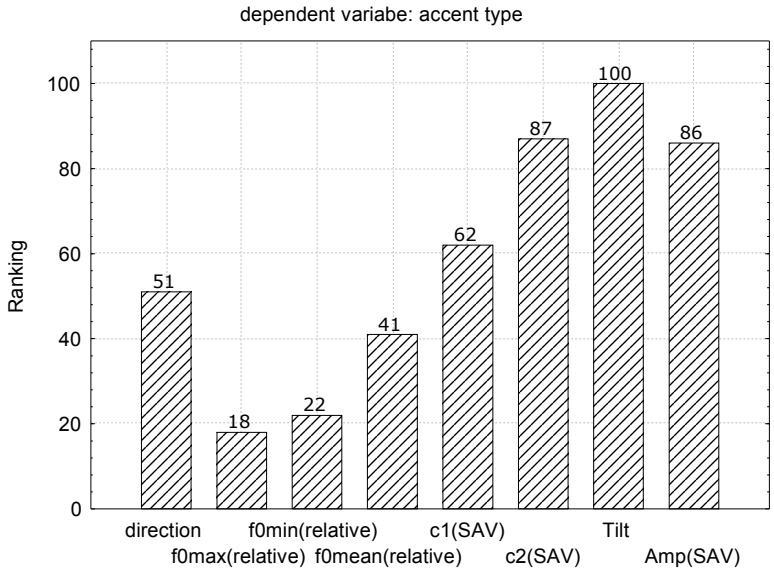


Figure 55: Importance ranking of pitch accent type predictors.

The fact that a specific variable is present in the ranking does not mean that it was used as a predictor in the classification. This can be concluded from the analysis of the tree graph depicted in Figure 56. The graph illustrates the structure of the decision tree designed in the current experiment: the tree has 27 splits and 28 terminal nodes. It can be seen that no split condition was based on the least important variables: f0max(relative) and f0min(relative). In the graph the following information is shown:

a) the number of cases in each observed class that are sent to the node
b) the predicted class to which cases sent to the node is assigned
c) split condition for a split node

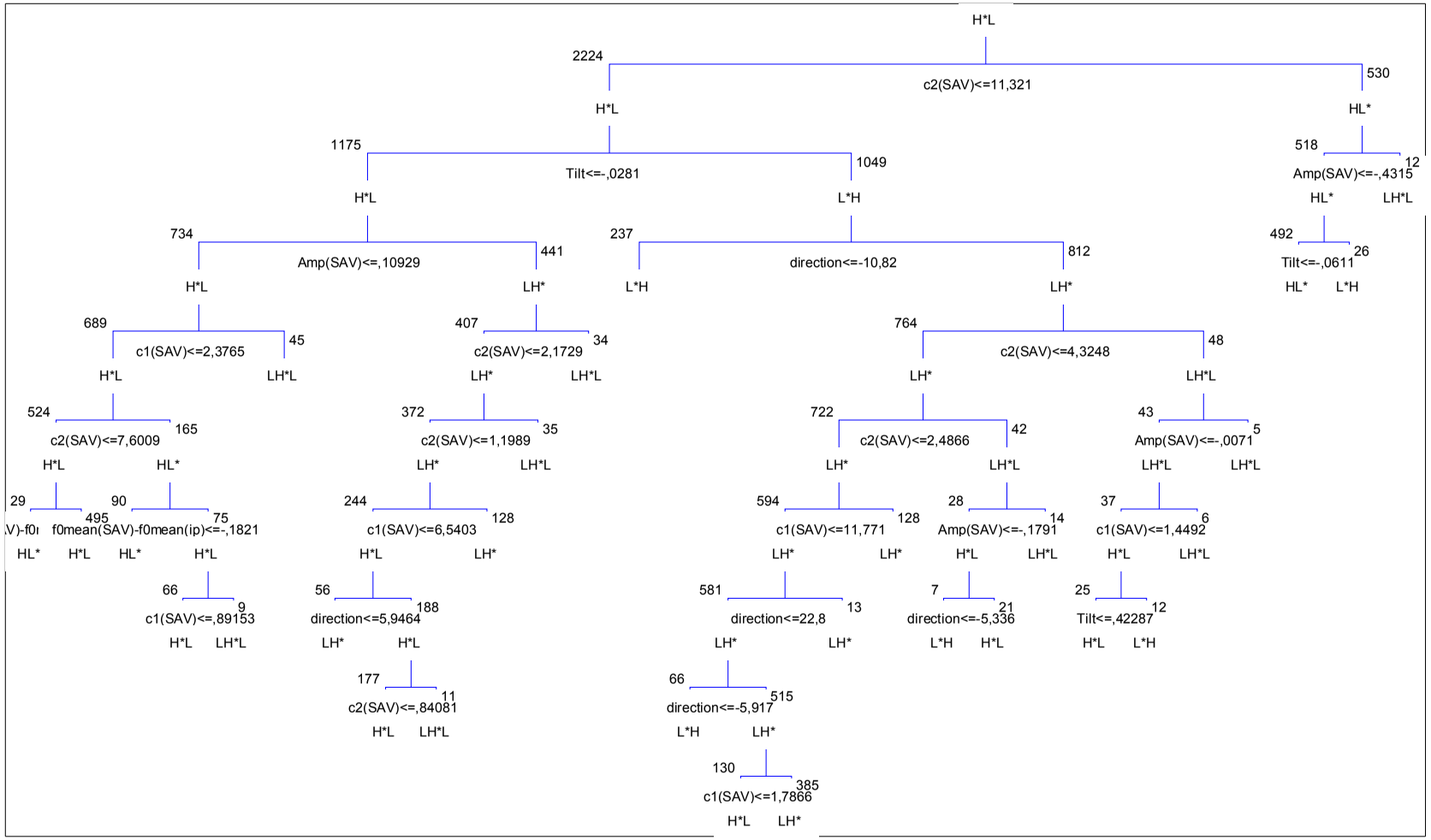


Figure 56: Decision tree used for prediction of pitch accent type. The tree has 27 splits and 28 terminal nodes.

The results of pitch accent type recognition with the tree designed in the current experiment are given in Table 54 which contains misclassification matrix computed for the training a) and test sample b).

a) learning sample

accent:	H*L	L*H	LH*	HL*	LH*L
H*L		7	125	34	5
L*H	14		83	11	0
LH*	161	18		1	2
HL*	91	9	2		1
LH*L	54	1	51	5	
average%:	69,9	87,27	66,92	90,88	88,24

b) test sample

accent:	H*L	L*H	LH*	HL*	LH*L
H*L		2	29	6	2
L*H	2		35	6	0
LH*	48	6		0	1
HL*	26	4	0		0
LH*L	22	1	15	4	
average%:	71,01	86,17	70,19	91,67	89,29

Table 54: Misclassification matrix for the learning a) and test sample b).

In comparison to the results achieved with the discriminant function analysis, the tree designed in the current experiment performs much better. It can be seen that the average accuracy of pitch accent type recognition in the training sample is 80% and in the test sample 81%. The performance of the decision tree varies with pitch accent type: the best results are achieved for HL*, whereas H*L and LH* accents are correctly recognized only in 70% in the test sample and even less than that in the training sample. But, as explained in the previous section, this kind of misclassification is considered as less serious than e.g., identification of LH* accent as HL* or H*L accent as L*H.

These results are comparable to those presented in (Demenko 1999, Jassem & Demenko 1999) where overall classification accuracy was between 79-83% depending on pitch accent type.

6.5.3. Classification using neural networks

Like in the previous sections where pitch accent types were recognized with discriminant analysis function and decision tree, the dataset used in the current analysis consisted of 3617 instances of pitch accents split into training (2754) and test (917) samples.

For the purpose of building the models presented here a neural network designer available in *Statistica* package was used. Following the methodology adopted in (Demenko 1999) MLP, linear and RBF networks were applied to the task of pitch accent type recognition.

In order to determine an effective configuration for the networks a number of experiments was carried out in which various configurations were tested and the results were

computed for different sample settings. The highest recognition accuracy was achieved with an MLP network including 15 hidden units and RBF network with 82 hidden units.

The summary of the two networks is given inTable 55. Starting from the left column the following information is provided:

a) network type and structure (i.e., the number of input, hidden and output units)
b) performance of a network on the training/test sample; the performance measure is the proportion of cases correctly classified
c) the error (RMS) of the network on the training/test sample

It can be seen that the MLP and RBF networks have the same performance as regards the average accuracy of pitch accent type recognition: 79% in the training and 81% in the test sample. What differs is the value of the RMS error which is significantly higher for the MLP network than for the RBF network. Additionally, the MLP network error computed for the test sample is much higher in comparison to error on the training sample, which may indicate over-fitting. Therefore, these results should be treated with caution.

network	perf. training	perf. test	RMS training	RMS test
MLP 7:7-15-5:1	0,79	0,81	0,86	1,02
RBF 8:8-82-5:1	0,79	0,81	0,24	0,24

Table 55: Model summary details

Table 56 gives a summary on the networks performance based on the test sample: in a) results achieved with the MLP network are presented; in b) the performance of the RBF network is summarized. It can be observed that with the two networks the best recognition accuracy was obtained for HL* and LH* accents. L*H accents are more often correctly identified with the MLP than RBF network. The worst recognition accuracy can be observed for LH*L accents and the percentage of misclassifications of LH*L accents with the RBF network is inverse of the recognition accuracy of this accent type with the MLP network. Yet, in both cases the results are significantly worse than those obtained with the decision tree (see previous section, 6.5.2).

a)

MLP 7:7-15-5:1	H*L	L*H	LH*	HL*	LH*L
total:	338	94	265	192	28
correct	259	73	220	172	17
missclass.	79	21	45	20	11
correct%	76,63	77,66	83,02	89,58	60,71
missclass.%	23,37	22,34	16,98	10,42	39,29

b)

RBF 8:8-82-5:1	H*L	L*H	LH*	HL*	LH*L
total:	338	94	265	192	28
correct	267	66	227	172	11
missclass.	71	28	38	20	17
correct%	78,99	70,21	85,66	89,58	39,29
missclass.%	21,01	29,79	14,34	10,42	60,71

Table 56: Summary statistics computed on the basis of the test sample.

In order to investigate the importance of the input variables for the pitch accent recognition results a sensitivity analysis was carried out. Table 57 shows importance ranking of the predictor variables computed in the analysis: 1 indicates the highest importance.

network	Amp(SAV)	c2(SAV)	c1(SAV)	Tilt	f0mean (relative)	direction	f0max (relative)	f0min (relative)
MLP 7:7-15-5:1	3	2	5		4	1	6	7
RBF 8:8-82-5:1	1	2	3	4	6	5	7	8

Table 57: Sensitivity analysis results

The importance of the features describing pitch accents differs depending on the model i.e., MLP vs. RBF network. As regards the former the most important feature is *direction* of the pitch movement which draws distinction between L*H vs. other accents and explains higher recognition accuracy of this pitch accent type with the MLP than with the RBF network (where *direction* has 5th position in the ranking).

As regards the RBF network the most important feature is Amp(SAV) which is also highly significant in accent type classification with the MLP network. The features describing f0 amplitude c2(SAV) and c1(SAV) are positioned high in the ranking too. Interestingly, *tilt* which was at the top of the importance ranking in the recognition of pitch accent type with the decision tree presented in sec. 6.5.2 is at 4th position in the RBF network ranking and disappeared completely from the MLP network ranking. The lack of the tilt feature or its limited "participation" in the identification of pitch accent type may explain the low accuracy of LH*L accent recognition obtained with the neural networks as compared to the very good results achieved with the decision tree (cf. sec. 6.5.2). This hypothesis is additionally supported by the results of ANOVA and discriminant analyses (sec. 6.5.1) which showed that next to amplitude parameters (Amp c1, c2) tilt is the one which differs most significantly between various accent types and also discriminates between them the best.

What is similar between the ranking presented here and in the sec. 6.5.2 is that the relative peak and minimum height described by the variables f0max(relative) and f0min(relative) play the least significant role in the recognition of pitch accent type.

6.5.4. Conclusions

As stated at the beginning of sec. 6.5 the goal of the analyses presented here was to propose a method capable of recognizing the surface phonological description of pitch accents pitch from acoustic parameters which constitute the description of accents on the phonetic level. This description consists of the following features:

a) *tilt* - describing the "shape" of the accent
b) *Amp(SAV), c1(SAV)* and *c2(SAV)* describing amplitude changes on the accented vowel
c) *f0mean(relative)* - the overall pitch level on the accented vowel
d) *direction* - describing the direction of the pitch movement

The high recognition accuracy achieved with different statistical classification methods including discriminant function analysis, decision trees and neural networks prove that these

features are highly significant for identification of pitch accent types, which confirms Hypothesis 1a. The importance ranking of features showed that tilt and amplitude affect the accuracy of accent type recognition to a greater degree than other features - f0mean(relative) and direction. It might explain worse performance of neural networks (where tilt feature was omitted or positioned low in the ranking) in comparison to the decision tree (where tilt was the first in the ranking).

In general, the average accuracy of pitch accent type recognition achieved in the current study which is between 78% and 81,7% can be regarded as satisfying. It is similar to results achieved by other models presented in sec. 6.1.2 (e.g. Rapp 1996, Kießling et al. 1996, Demenko 1999, Sridhar, Bangalore & Narayanan 2007) or even better as compared to 65.2% accuracy reported in (Bulyko & Ostendorf 2001). It should also be noted that most of these models rely on higher-level information such as POS. On the contrary, the models designed in the current study do not require such information, which makes them more universal. In the current study a vector of eight features describing the accents on the phonetic level was proposed. The features can be easily derived from utterance's acoustics. The models depend on the information on syllable and phoneme boundary, but this is not considered as their disadvantage, because in most speech applications such information has to be provided anyway. The high accuracy of pitch accent type recognition achieved by the models confirm Hypothesis 2a according to which a high accuracy in the automatic detection and classification of intonational events can be achieved even if only a small acoustic feature vector is used at the input to the classification model.

The performance of the models designed in this study is also higher in comparison to the consistency achieved by human labelers in manual annotation of pitch accent types (e.g. Pitrelli, Beckman & Hirschberg 1994, Grice et al. 1996, Reyelt 1996), which confirms Hypothesis 2b and proves the usefulness of the models.

Chapter 7. Generation of f0 contours

In this chapter two hypotheses will be tested which are related to the issue of intonation generation. In the first place, the goal is to prove that the approach to f0 generation in which f0 contours result from interpolation between pitch targets of a pre-defined position in syllable structure (at syllable start, in the middle of syllabic nucleus and at syllable end) whose values are estimated by means of a regression model is capable of providing contours which are very similar to those occurring in natural speech (Hypothesis 2c).

In general, most of the existing TTS systems are capable of generating a naturally sounding speech in a news commentary style. However, a high-quality expressive speech is still a challenge and many studies are dedicated to this problem (e.g. Campbell 1998, 2004, Iida 2004). The goal of the study presented here will be to prove that the framework proposed to modeling of unemphatic speech can be successfully used to generate f0 contours in expressive speech (Hypothesis 3).

In the following sections the steps taken to achieve the two goals are described.

Multilayer perceptrone neural networks are trained to perform regression and estimate three target values per syllable: f0 value at the start of the syllable, in the middle of the nucleus and at the end of syllable, on the basis of 36 input features. The f0 contours result from interpolation between predicted target f0 values. Since with the exception of 1) syllables of a phrase-initial position, 2) preceded by pause or disjuncture resulting from hesitation or disfluency, or 3) preceded by a syllable marked with a break index higher than 1, syllable-initial f0 value is effectively the same as f0 value at the end of the previous syllable, it is assumed that interpolation can be done between nucleus-middle and syllable-final f0 values and estimation of f0start values is necessary only in cases such as 1)-3) listed above. This hypothesis is confirmed in the analyses presented in this chapter.

It will be shown that the method of f0 contour generation proposed in this thesis is capable of producing naturally sounding f0 contours, which is the result of providing the models with meaningful and significant input features. The features used for pitch target estimation are derived from utterance's segmentation and annotation. A lot of them refers to prosodic features such as stress or boundary strength or is somehow related to prosodic features (e.g. phrase type, distance to the previous stressed syllable, etc.). Therefore, in this chapter two more hypotheses will be tested. The first one (Hypothesis 1b) can be confirmed by showing that the surface phonological description proposed in this thesis which reflects melodic and functional aspects of intonation provides information of a high significance for the estimation of pitch targets and thus, for the results of contour generation in speech synthesis. The second one (Hypothesis 1c) can be confirmed by showing that the incorporation of a finer information on phrase structure increases the accuracy of pitch target estimation by regression model. In order to test these hypotheses a sensitivity analysis will be carried out which results in importance ranking of predictor variables: if the variables related to prosodic features and phrase structure have high position in the ranking it can be assumed that they hypotheses are proven.

In the next section a brief explanation is given for the choice of the approach to contour generation proposed in this thesis. This is followed by description of building regression models for unemphatic speech and expressive speech. In the end the results of perception test are presented proving the hypotheses tested in this chapter.

7.1. Preliminary remarks

In the Chapter 3 the most influential approaches to intonation modeling which provide different methods of f0 contour generation were presented. The comprehensive intonation model developed in this thesis proposes a framework for f0 contour generation which consists in interpolation between pitch targets of a pre-defined position in the syllable (at syllable start, in the middle of nucleus and at syllable end) and are estimated with a regression model. In this section the motivation for this kind of approach to contour generation is given. As a matter of fact, two issues have to be accounted for: Firstly, the choice of the sequence-based approach and secondly, the choice of pitch targets for the f0 contour generation.

The reason for choosing the sequential approach to intonation modeling is that it was proven to yield better results in Polish then the superpositional approach (Fujisaki model) which appeared to be too constrained (Demenko 1999, cf. sec. 3.7.1). Similar observation can be found in (Taylor 1993, 2000) where the superpositional approach to contour generation also appeared to lack flexibility to model a wide variety of intonation contours. Apart from that, the results of resynthesis of intonation contours from the PaIntE parameters reported in (Hałupka 2004, Wagner 2004, cf. sec. 3.7.2) confirmed the view on the usefulness of the sequence-based approach presented in (Demenko 1999). The results additionally proved that intonation contours can be effectively generated by interpolation between pitch targets (f0 and time of the start, peak, bottom and end of the event) derived from PaIntE parameters.

As regards the issue of the type of targets for generation of intonation contours two approaches can be taken: the targets can be either anchored in syllable structure or derived from the parameters describing intonational events on the phonetic level (e.g. tilt, amplitude, duration, etc.).

The first reason for choosing the former approach is that previous attempts to generation of Polish intonation proved its usefulness (Oliver & Clark 2005) and generally, good results were achieved: The overall correlation between the observed and estimated f0 targets: syllable-start, syllable-middle and syllable-end f0 values was 0.68/71 in the training/test sample. The similarity between original and generated f0 contours was also investigated in a perception test which showed that the proposed model generates more acceptable intonation in comparison to the former rule-based model.

The second reason is that generation of f0 contours from pitch targets of a pre-defined position in the syllable is more straightforward (as it involves fewer steps), but at the same time brings results comparable to those obtained with theory-based models. This can be concluded when comparing the results reported in various studies.

In (Dusterhoff & Black 1997) three different experimental settings were used to evaluate the quality of f0 contour generation in the framework of the Tilt theory (Taylor 1998). the speech material used in the experiments was Boston Radio News Corpus (Ostendorf, Price & Shattuck-Hufnagel 1995). It should be noted, that in this framework intonation contours are

generated from tilt parameters which have to be first converted into RFC representation and each event is decomposed into its rise and fall components. Then, the level and position of pitch targets (at the start, peak and at the end of the event) are calculated.

In the first experiment intonation contours were generated from automatically labeled Tilt parameters. For the purpose of the estimation of the five tilt parameters (startf0, amplitude, duration, tilt and peak position) ten regression models are built - five for pitch accents and the other five are used for boundary tones. Apart from that, two regression models are trained to predict the targets necessary for the generation of connection elements and pauses. The authors report on correlation=0.6 and RMSE=32.5Hz between the original smoothed and generated contours. The results using manually labeled Tilt events are slightly worse with correlation=0.57 and RMSE=33.9Hz. The contour generation from Tilt parameters derived from ToBI labels yielded the worst results with correlation=0.55 and RMSE=34Hz.

In (Dusterhoff, Black & Taylor 1999) the same approach was adopted to f0 generation of different speech styles. On the news commentary corpus the correlation between the original and generated (from automatically labeled Tilt parameters) f0 contours was r=0.6 and RMSE=34.3Hz. On the corpus consisting of isolated sentences r=0.74 and RMSE=9.1 was obtained. The most problematic was estimation of Tilt parameters for instructional text utterances and consequently, the correlation between the contours is r=0.53 and RMSE=21.1.

In the latest study dedicated among other things to f0 generation using the Tilt theory (Taylor 2000) the DCIEM corpus consisting of map task dialogues was used (Bard et al. 1995). Depending on the representation used to generate the contours the correlation between the original smoothed and generated contours varied from 0.77 to 0.84 with RMSE between 6.82 and 7.51Hz.

The latest results of f0 generation in the PaIntE approach are reported in (Möhler 2001). Different experimental settings were used determined by the type of parameterization and the size of the codebook. In the first experiment six regression trees were designed to estimate the values of the PaIntE parameters (Möhler 1998, see sec. 3.2.2). In the second experiment the parameters were normalized with respect to the pitch range of the phrase (mapped onto the interval 0 to 1 describing the bottom and top of the range respectively). In the third experiment apart normalization with respect to pitch range the parameters are anchored in syllable structure e.g., the position of the peak is mapped onto one of the intervals corresponding to the onset/nucleus/coda of the current/previous/next syllable. Consequently, contour generation from the third type of representation involves the greatest number of steps, but the results reported in (Möhler 2001) indicate that this approach yields the best results.

The correlation between the original contours and contours generated from the six PaIntE parameters estimated with the regression model varies between 0.59 and 0.62 depending on the parameterization scheme.

As regards the results of f0 generation from the vector-quantized PaIntE parameters the best results are obtained for the codebook including 8 codewords: r=0.69, RMSE=14.0Hz.

The study also shows that the quality of contour generation drops when no information on the type of pitch accents/boundary tones is available and only information on their location is used in the estimation of PaIntE parameters.

Quite different representation is used for contour generation in (Black & Hunt 1996). It consists of three targets anchored in the syllable structure: at the start and end of the syllable and in the middle of syllabic nucleus. The targets are estimated for each syllable in the utterance

from a number of linguistic features. In this way it is possible to reduce the number of models which estimate the representation of intonation contours to three, as opposed to twelve or six regression models used in the studies presented above. The speech material used in (Black & Hunt 1996) was the same as that used in (Dusterhoff, Black & Taylor 1999) and (Dusterhoff & Taylor 1997), which makes comparison of the results easier. The results reported by the authors are comparable to those reported in the previously described studies: the correlation between the original and generated contours is 0.62 with RMS=34.8.

This result as well as that reported in a recent study for Polish intonation modeling (Oliver & Clark 2005) shows that the approach in which f0 contour result from interpolation between pitch targets of a constant position in the syllable structure gives results comparable to those reported for theory-based approaches. At the same time the latter are less straightforward and involve a number of intermediate steps between target/parameter estimation and contour generation.

Further evidence supporting the approach proposed in this thesis will be given in the following sections of this chapter.

7.2. F0 contour generation in unemphatic speech

In this section the Hypothesis 2c is tested which says that an approach to f0 generation in which f0 contours result from interpolation between pitch targets of a predefined position in the syllable structure (at the onset start, in the middle of the nucleus and at the end of the coda) whose values are estimated by means of a regression model provides a high quality speech characterized by natural intonation.

In order to test this hypothesis the performance of the regression model designed in this study will be evaluated in an objective and subjective manner. The latter evaluation involves a perception test and will be the subject of sec. 7.4. The former evaluation involves calculating the correlation between original and estimated pitch targets and/or between original and generated f0 contours. Apart from high correlation, in order to prove Hypothesis 2c the results achieved by the model presented here should compare favorably with or be at least as good as the results reported by other authors.

A sensitivity analysis will be carried out to see whether information provided by the surface phonological description of intonation and phrase structure model proposed in this thesis is significant for the estimation of pitch targets for contour generation. If so, it confirms the Hypothesis 1b and Hypothesis 1c.

7.2.1. Determination of features for pitch variation control

On the basis of the methodology presented by other authors and discussed in the previous section a number of syllable- and phrase-related features as predictor variables for estimation of the level of f0 targets: f0start (syllable-initial f0), f0mid (nucleus-middle f0) and f0end (syllable-final f0). The features are divided into syllable- and phrase-level features. In the first place, syllable-level features are discussed which include stress, accent and boundary tone

features as well as information concerning position of the syllable and prosodic break following it.

1. Stress feature: it is a binary feature - a syllable is either lexically stressed or not. Stress was assigned automatically from rules and verified by labelers during prosodic annotation of the corpus. The information on lexical stress is provided for the current syllable (SA) and the previous (PA) and second previous syllable (PA2) and the following (FA) and second following syllable (FA2). With very few exceptions the stress features determined in this 5-syllable window has significant effect on the level of pitch targets describing pitch contour of a syllable.

2. Accent features: syllable can be either unaccented or accented in which case it is labeled with one of the accent types: H*L, HL* LH*L, LH*, L*H, LD or LI. In sec. 5.3.2 distinctive features of these accents were discussed and description of their realization in terms of acoustic features was presented. It could be seen that accent type significantly affects pitch variation on syllable, which means that it is an important predictor of syllable's f0 contour. On the basis of manually labeled pitch accent types information on the accent type on the current syllable and two preceding and two following syllables is provided.

3. Boundary tone features: a distinction is drawn between 5 boundary tone types marking the end of statements (5,.), questions (5,?), exclamations (5,!), continuation phrases (2,?) and declarative phrases inside complex sentences (2,.). Each word-final syllable can be marked with a boundary tone or not, thus a 6-way opposition is possible. It could be seen in sec. 5.3.3 that boundary tone type affects pitch variation on a syllable, which means that it is an important predictor of a syllable's f0 contour. On the basis of manually labeled boundary tone types information on the boundary tone on the current syllable and two following syllables is provided.

4. Prosodic break (BI). In a 5-syllable window around the currently analyzed syllable the break index on the syllable is described:

a) 0 - marks boundaries of syllables or words inside clitic groups
b) 1 - phrase-medial, prosodic word boundary
c) 2 - corresponds in a sense to 2 break index in EToBI where it marks "a strong disjuncture marked by pause or virtual pause, but with no tonal marks, i.e., a well-formed tune continuous across the juncture" (Beckman & Ayers 1997:35). In the database used in this study this break index was marked only a couple of times whenever a disjuncture resulting from a hesitation or disfluency occurred. Utterances containing this break index were excluded from the speech material used in the study presented in this thesis.
d) 3 - marks minor phrase boundary
e) 4 - marks major phrase boundary

In one-way ANOVA analyses the effect of prosodic break strength on pitch variation on syllable was investigated and it was found that break on the current syllable - BI(SA) and on the next syllable BI(FA) affect pitch variation most significantly, whereas prosodic break at the preceding syllable - BI(PA) has the least significant effect (but it is still statistically significant, $p<0.01$).

5. Syllable position and other features

Various measures were used to describe syllable position.

a) dist_start(+s) - this feature gives syllable distance to the start of the phrase measured in the number of stressed syllables
b) dist_start(+acc) - gives distance of the syllable to phrase start measured in the number of accented syllables
c) dist_end(+s) - gives distance of the syllable to phrase end measured in the number of stressed syllables
d) dist_end(+acc) - gives distance of the syllable to phrase end measured in the number of accented syllables
e) dist(W_start) - gives the number of syllables to the start of the current word
f) dist(W_end) - gives the number of syllables to the end of the current word
g) W_length - describes the length of the current word measured in syllables
h) dist(ip_end) - syllable position in the phrase is expressed in the number of syllables up to the phrase (ip) boundary
i) dist_ip(start) - syllable position in the phrase is expressed in the number of syllables from the phrase (ip) start
j) dist_next#$p(syllables) - gives distance of the syllable to the next pause (#$p)
k) dist_next#$p(ms) - like above, but measured in millisec.
l) dist_next(+acc) - distance to the next accented syllable
m) dist_prev(+acc) - distance to the previous accented syllable
n) nucl_dur - gives duration of the nucleus
o) syl_dur - gives duration of the syllable
p) syl_structure - this feature describes syllable structure in terms of the number and type of segments in the onset and coda; onsets and codas were classified as sonorants, voiceless obstruents or voiced obstruents. This classification is based on the effect of segment type on pitch variation (e.g. van Santen, Möbius 1997)

In a series of one-way ANOVA analyses the effect of the features listed above on the level of pitch targets located in the syllable (f0start, f0mid and f0end) was investigated. The results show that all these features affect pitch variation on a syllable. The variables of the greatest effect on f0 values include: dist_next(+acc), dist_start(+acc) and dist_end(+acc). It should also be marked that the effect of these factors (and other as well) is usually less significant on f0start value in comparison to nucleus-mid and syllable-final pitch. The variables: syllable structure, word length and distance to the word start affect the level of the three pitch targets the least.

But, in general, all the features selected for predictor variables in the regression model have statistically significant effect on the values of f0start, f0mid and f0end ($p<0.01$). The only exception is lack of significant variation in syllable-initial pitch between syllables positioned at a different distance from the word start (described by dist(W_start).

The use of the duration features can be questioned: neither nucleus nor syllable duration shows significant correlation with values of the three pitch targets (f0start, f0mid and f0end). However, it was decided to include them in the predictor feature set, because they are highly correlated with factors of a significant effect on the level of the pitch targets i.e., end tone (ET) and break index (BI) on the current syllable (SA). Besides, no damage is done by including

features of a low significance for prediction, because during the training the network performs selection of the input features itself: it identifies variables that can be safely ignored in subsequent analyses and retains key variables.

In sec. 5.1 a representation of prosodic structure was defined on the basis of statistically significant effect of a number of phrase-level factors on pitch variation in phrases. In particular it was analyzed how various categorizations of phrases affect f0 parameters describing pitch range (f0max, f0min, f0mean, f0end and st_dev). It was found out that a distinction between two phrasing levels i.e., major and minor phrases and distinctions based on ip position in IP and IP length provide a framework within which pitch variation at phrase level can be effectively controlled. The prosodic structure representation is incorporated into the f0 prediction and generation model and used in the normalization of pitch.

For each syllable in the database the information concerning features of the phrase in which the syllable occurs is provided.

1. Phrase type

A distinction is drawn between the following phrase types:

a) single - major intonational phrase (IP) contains a single minor intonational phrase (ip)
b) initial: minor phrase has initial position in the major phrase
c) medial: groups ips of a medial position in IP (medial = non-initial and non-final)
d) final: groups ips of a final position in IP

ANOVA results showed that phrase type has statistically significant effect on f0start (F=412,28), f0mid (F=562,4) and f0end values (F=690,3).

2. Phrase length

This factor was used mainly to control pitch variation on phrases of a medial position in ips of a different length, but only a weak (but still, statistically significant) effect of this factor on the level of the pitch targets can be observed: F(f0start)=3,7, F(f0mid)=3,5 and F(f0end)=5,7.

3. Tune type

The effect of tune type on distribution of pitch targets on the f0 scale is comparably significant to that of phrase type. In the prosodic annotation of Polish unit selection corpus the information on basic intonational meaning conveyed by intonational tunes is provided. The basic meanings are: statement, interrogation, continuation and exclamation and they are encoded in the nuclear melody determined by a sequence of a pitch accent followed by a boundary tone e.g. in our prosodic labeling scheme statement contour is signaled by a HL* accent followed by a 5,. boundary tone, whereas interrogation is conveyed by a L*H accent followed by a 5,? boundary. Tune type affects most significantly syllable-final pitch, while syllable-initial pitch is affected to a lesser degree (F=227,6) than syllable-mid (F=292,3) and -final pitch (F=358,6).

To sum up, all features chosen for input variables to the regression network have statistically significant effect on the level of pitch targets: syllable-initial, nucleus-middle and -final f0 values, but among them pitch at syllable start is the least affected by the variables. These results suggest that the input features should have high contribution into the estimation of pitch target values.

7.2.2. Selection, training and testing of the regression network

The first step towards model building consisted in the specification of the type and complexity of neural networks to be trained. *Statistica Neural Networks* (SNN) software was used in which the following network types can be applied to regression problems: linear, multilayer perceptrone (MLP), general regression neural network (GRNN) and radial basis function (RBF). In order to determine the best network type for our task a preliminary experiment was carried out. Due to limitations of GRNN and RBF networks which are incapable of performing *extrapolation* only MLP and linear networks were taken into account. Extrapolation occurs when an output value is estimated by projecting the trend of the curve fitted to the input values onwards. It can be expected that using extrapolation may sometimes have undesirable effects, but on the other hand, inability of the network to estimate output values which lie outside the range of training data may have even more serious consequences.

Intelligent Problem Solver program available in SNN package was used to design regression networks. Linear and MLP networks were selected for the preliminary training (using back propagation and/or conjugate gradient descend method) and assessment. All MLP networks had 3 layers and the number of units in the hidden layer varied between 1 and 20. The networks were trained using the whole set of independent (predictor) variables (36) and alternatively, allowing the network to automatically select a subset of variables. The database used in the experiments contained 17481 syllables. They were split into 3 subsets:

- training (6011) used to optimize the network
- selection (6019) used to halt training to mitigate over-learning and/or to select from a number of models trained with different parameters
- test (5451) used to perform an unbiased estimation of the network's likely performance

The train and select subsets were resampled, but the test set was maintained the same to allow comparison of results.

The performance of the designed networks was evaluated in a number of ways. They include:

a) analysis of the correlation between observed and predicted values: in general, correlation coefficient can be regarded as a good indicator of performance (but, cf. the comment in g), sec. 7.2.3)
b) analysis of the ratio of the prediction to data standard deviations: if this is 1.0, then the network has performed no better than a simple mean estimator. A ratio below 1.0 indicates a good regression performance
c) analysis of the residuals: they are computed as a difference between the observed and predicted values, thus the greater the residual value the less accurate the prediction
d) analysis of a scatterplot illustrating the distribution of observed vs. predicted values: if the values show similar distribution it can be assumed that they are highly correlated

In general, MLPs performed better than linear models in terms of correlation between observed and predicted f0 values and prediction error SD. Moreover, it was observed that better results were achieved when all input features were in use while the model was trained than

when a subset of features was selected automatically. On the basis of these results only MLP networks were subject to further training in which all input variables were used.

Depending on the network features correlations (generated for all selected cases) between the observed and predicted values of 1) syllable-initial pitch (f0start) varied between 0.59 and 0.67, 2) nucleus-medial f0 (f0mid) varied between 0.74 and 0.78, and 3) syllable-final f0 (f0end) varied between 0.73 and 0.79.

The best results in terms of correlation, as well as SD prediction error were achieved for an MLP network with 14 neurons in the hidden layer. The summary of the model is given in the table depicted in Table 58. It can be seen that the performance is similar on the training, selection and test subsets. The low network error indicates that no over-learning occurred.

network type/config.	perf. train.	perf. select.	perf. test.	train. error	select. error	test. error	training	input feat.	hidden(1)
MLP 36:14:3	0,72	0,76	0,74	0,07	0,08	0,08	BP72b	36	14

Table 58: Model training summary: regression network, unemphatic speech.

7.2.3. Results

Table 59 presents the regression statistics computed for the best MLP network designed in the experiment described in the previous section. For each of the subsets (training, selection and test) the following information is presented:

a) ***Data Mean.*** Average value of the target output variable
b) ***Data S.D.*** Standard deviation of the target output variable.
c) ***Error Mean***. Average error (residual between target and actual output values) of the output variable.
d) ***Abs. E. Mean***. Average absolute error (difference between target and actual output values) of the output variable.
e) ***Error S.D.*** Standard deviation of errors for the output variable.
f) ***S.D. Ratio***. The error to data standard deviation ratio.
g) ***Correlation***. The standard Pearson-R correlation coefficient between the predicted and observed output values. Correlation shows how well the predicted contour's variation follows a target contour.

As it is pointed out in (Clark 2003:123): *"The use of these scores for intonation is somewhat ungrounded as they do not correlate well with perceptual ratings of the difference between contours"*. It is so, because it happens that perceptually similar contours have high error values or low correlation coefficient and the other way round. But up to now, no better measures have been found and as a result, one has to refer to these scores in the assessment of regression models' performance.

subset	training			selection			test		
statistics	**f0start**	**f0mid**	**f0end**	**f0start**	**f0mid**	**f0end**	**f0start**	**f0mid**	**f0end**
data mean	112,37	109,89	111,91	112,45	110,40	112,45	109,81	108,28	109,88
data S.D.	16,43	17,20	15,97	16,19	17,04	15,62	16,40	17,64	16,19
error mean	-0,25	-0,02	0,14	-0,24	-0,09	-0,06	2,30	2,15	2,27
Abs.error mean	11,87	9,93	9,21	12,32	10,78	9,87	12,18	11,38	10,35
error S.D.	8,99	7,47	7,03	9,33	8,11	7,46	9,51	8,70	8,12
S.D. ratio	0,72	0,58	0,58	0,76	0,63	0,63	0,74	0,65	0,64
Correlation	0,69	0,82	0,82	0,65	0,78	0,78	0,67	0,77	0,77

Table 59: Regression statistics, unit selection corpus.

It can be observed that the performance on the training subset is the best in terms of both correlation and SD ratio which is well below 1 and indicates good regression performance. The performance of the model on the test subset is only slightly worse than on the selection subset as regards prediction of f0mid and f0end values but better as regards prediction of f0start values. The SD ratio is comparable in the two subsets and below 1 - this means that the network is capable of providing good estimation of the output values.

The performance of the regression model designed in the current study is depicted in the graphs presented below. In the graphs the observed values of the pitch targets: *f0start, f0mid* and *f0end* are plotted against the values estimated by the regression model. The results computed for the training set are marked with circles and for the test set with crosses.

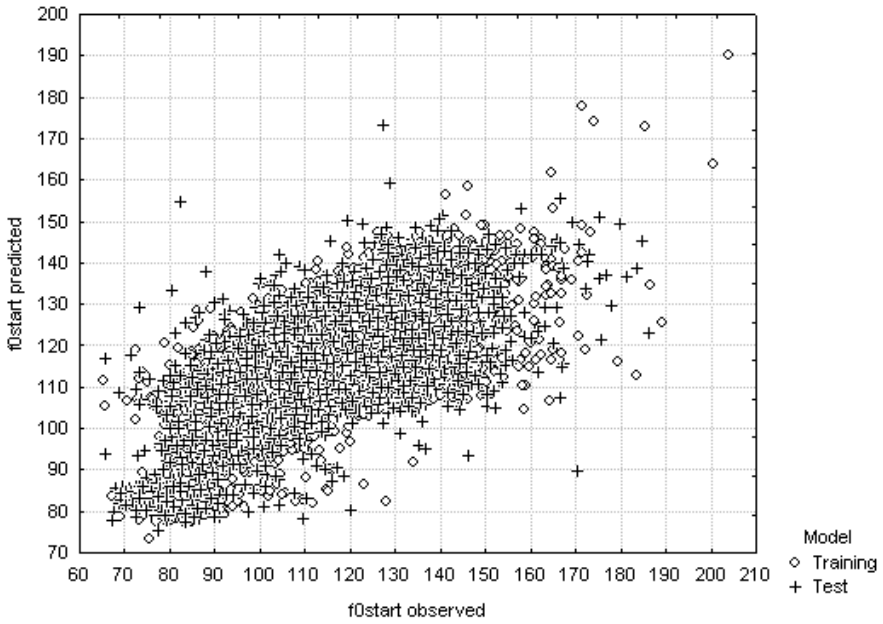


Figure 57: Scatterplot of observed vs. estimated f0start target values.

As regards the results of f0start target estimation it can be seen that the range of predicted f0start values (between 80-150Hz not taking into account some outliers) is more constrained than that of original f0 values (in the range 70-160Hz with a considerable amount of outliers between 160-180Hz). This implies some deficiency of the regression model and may be attributed to the fact that from among the three pitch targets f0start was the least affected by the factors used as input features to the model. As explained at the beginning of this chapter (p.176) f0start value constitutes a less significant pitch target than f0mid or f0end. It is so, because with the exception of syllables of a phrase-initial position or preceded by syllables marked with a break index larger than 1, f0start value can be replaced with f0end value of the previous syllable. For that reason, the estimation performed by the model should be sufficient for generation of naturally-sounding f0 contours.

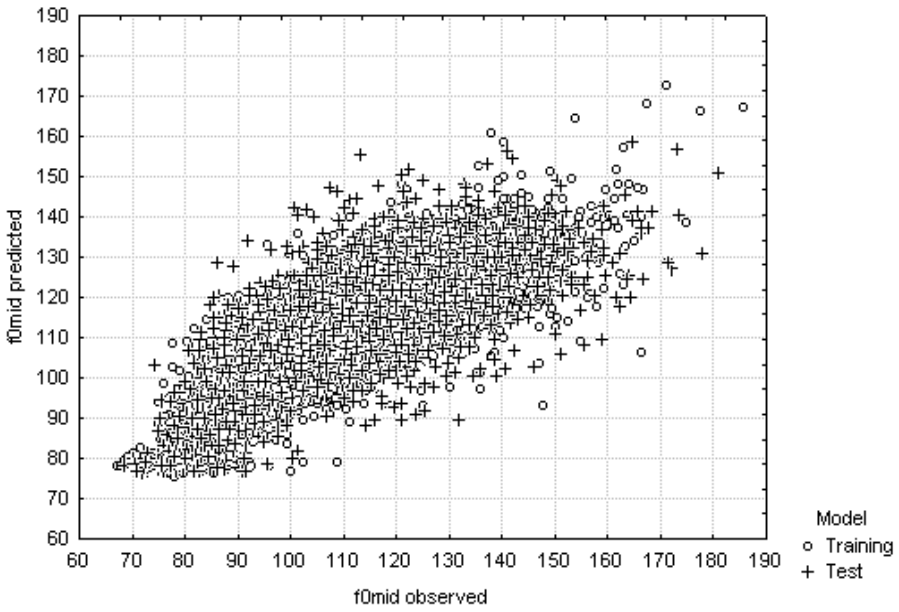


Figure 58: Scatterplot of observed vs. estimated f0mid target values.

The correlation between observed and estimated nucleus-medial f0 values is higher in comparison to that between observed vs. predicted values of syllable-initial pitch. It can be concluded on the basis of the analysis of scatterplots of these values presented in the graphs a) and b). It can be seen that the bottom of the range is similar between the original and estimated f0mid values (about 75Hz) ignoring a couple of outlying original f0 values (below 70Hz). As regards top of the range, it is about 10-15Hz lower for the estimated values than for the original values. This effect can also be observed in the difference between mean and S.D. of nucleus-middle pitch: 111.4 Hz in the original database and 109.9Hz/108.8Hz in the estimated training/test subset respectively.

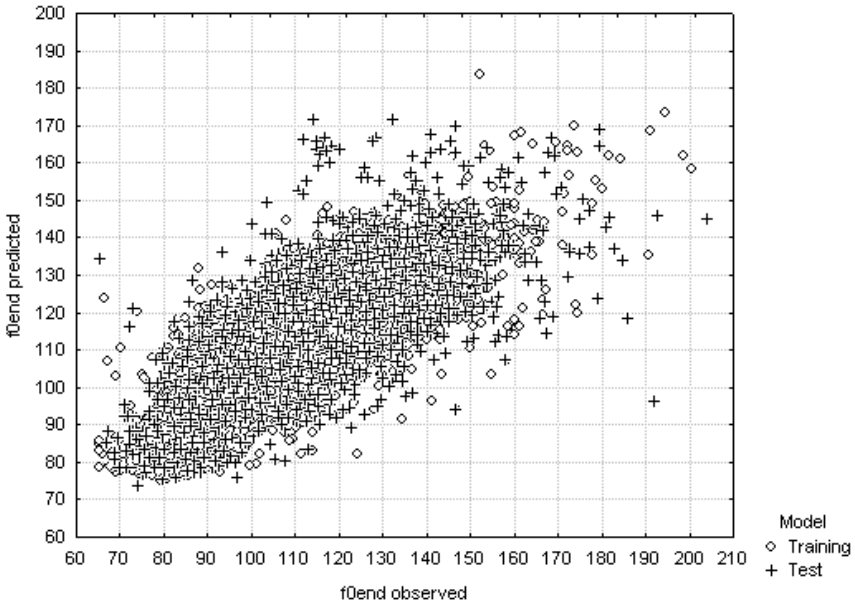


Figure 59: Scatterplot of observed vs. estimated f0end target values.

The correlation between observed vs. predicted f0end values is the same as that between observed vs. predicted f0mid values (r=0.77). The estimation of f0end values shows opposite effects to those observed for f0mid: the difference in the top of the range between observed vs. predicted values is smaller (observed: about 160Hz without outlying values, predicted: 150Hz with a considerable number of outliers up to 170Hz) than that in the bottom of the range (observed: about 65Hz, predicted: about 75Hz). It can be seen that the main difference in the estimation between training and test subsets regards the level and number of outlying values: they are more "extreme" in the training subset, but more numerous in the test subset.

A sensitivity analysis was carried out to investigate the contribution of particular input features to the estimation of output f0 values. The results computed for the test subset show that the 10 most important features are:

- *ET(SA)* - boundary tone on the current syllable
- *ACC(SA)* - accent type on the current syllable
- *tune type*
- *ip position*
- *ET(FA1)* - end tone on the next syllable
- *dist_end(+s)* - distance to the phrase end in the number of stressed syllables
- *ACC(PA1)* - accent type on the previous syllable
- *nucleus duration*
- *BI(FA1)* - break index on the next syllable
- *dist_next(+acc)* - distance of the current syllable to the next accented syllable (measured in the number of syllables)

The fragment of the importance ranking of predictor features shows that the key features are those related to *accent, boundary tone* and *prosodic structure*. These results indicate that the surface phonological description as well as the phrase structure model proposed in this thesis provide information of a high significance for the estimation of pitch targets and thus, for

the results of contour generation in speech synthesis. Consequently, it can be assumed that this proves Hypothesis 1b and Hypothesis 1c.

At the other end of the ranking (no displayed here) there are features describing *distance of the syllable from ip start, distance to the next pause, ip length, syllable structure* and *stress features.*

7.2.4. Conclusions

In general, the performance of the regression model designed for estimation of pitch targets in unemphatic speech (unit selection corpus) is comparable or even better than the performance of the models designed in other studies, which is indicated by the value of the correlation coefficient between the observed vs. estimated f0 targets or between the original vs. generated f0 contours.

In the current study the overall correlation between observed vs. estimated values is 0.69/0.67 for the f0start target, 0.82/0.77 for f0mid and 0.82/0.77 for f0end target in the training/test set respectively. These results compare favorably with those reported in (Black & Hunt 1996) where the correlation between observed vs. estimated f0 targets was 0.53/0.55 for syllable-initial pitch, 0.66/0.68 for vowel-mid pitch and 0.56/0.55 for syllable-final pitch in the training/test sets respectively. The performance of the model designed in the current study can be regarded as better than the performance the LR models described in another study dedicated to Polish intonation modeling (Oliver & Clark 2005). In that study the overall correlation between observed vs. estimated targets in the training subset was 0.68 as opposed to r=0.78 obtained in the current study, and on the test subset r=0.71 as opposed to r=0.74.

When the correlation between whole intonation contours is taken into account the results presented in the current study are better compared to those reported by other authors as well.

In (Dusterhoff & Black 1997) the correlation between original smoothed f0 contours and contours generated from estimated Tilt parameters (startf0, amplitude, duration, tilt, peak position) in the test set varied between 0.55 (for prediction using ToBI labels) and 0.6 (for prediction using automatic Tilt labels), while in the current study this correlation achieves 0.74 in the test set.

The overall 0.75 correlation (computed for all datasets) between generated f0 contours vs. original ones (i.e., obtained with the getf0.psc script, see sec. 4.3.1) achieved in the current study compares favorably also with the correlation of 0.6/0.53/0.74 between the contours from news commentary/instructional text/isolated sentences corpus (Dusterhoff, Black & Taylor1999).

Apart from the high correlation coefficient which is regarded as a good indicator of performance, the good estimation achieved with the regression network designed in the current experiment is also confirmed by the low value of prediction error S.D. (between 0.58 and 0.74), which indicates that the network's output is not constrained to lie within the range of observed (input) values, in other words: the model performs better than a simple mean estimator, because it is capable of performing a "reasonable" extrapolation.

The comparison with other models presented here shows that the model designed in the current study yields better results than those reported in other studies. Consequently, it can be

assumed that this partly confirms Hypothesis 2c and proves that the framework adopted in the comprehensive intonation modeling generates f0 contours which are very similar to natural contours. Therefore, it can be expected that the intonation generated in this framework will be characterized by high naturalness. In order to asses the naturalness of the generated contours and confirm completely the Hypothesis 2c a perception study will be carried out (see sec. 7.4).

7.3. F0 contour generation in expressive speech

The reason why in the current study two different regression models are built for the two speech corpora i.e., unit selection and expressive speech instead of a single one is that in this way problems related to different data types can be avoided. The same approach was used in (Dusterhoff, Black & Taylor 1999) where separate regression models were built for the estimation of Tilt model parameters for three corpora including recordings of different speakers. The first corpus consisted of news commentary, the second one included isolated sentences and the third one - an instructional text. The authors explain their decision in the following way: "cross-data training was avoided so that each individual speaker and style could be modeled without the difficulties caused by the different data types" (op.cit.:1).

The current section starts with some preliminary remarks, then the design of the regression model is described. In the end the results are briefly discussed.

7.3.1. Preliminary remarks

In this section the Hypothesis 3 is tested which says that the intonation model developed in this thesis can be regarded as comprehensive if it provides a framework for generation of a high-quality, naturally sounding intonation of expressive speech.

In order to prove this hypothesis the methodology used for estimation of pitch targets and contour generation in unemphatic speech presented in the previous sections is now adapted to development of a corresponding model for expressive speech.

The above hypothesis will be proven if the model designed for expressive speech is capable of generating contours which are similar to the original f0 contours and if its performance is comparable to the performance of models designed in other studies. Apart from that, the result of the perception study described in sec. 7.4 should confirm high quality of the generated intonation.

The study presented in the current section is based on the expressive speech corpus described in sec. 4.1.2. It contains recordings of three different speakers (2 female and 1 male) reading a literary text including diverse examples of dialogues, monologues, discourses, different modes and expressivity. The text if full of humor, irony and grotesque, which resulted in a very expressive interpretation by the speakers. From each speaker ca. 20 minutes of speech was obtained. Not all the material was used in the study presented in this section: all utterances including disfluencies, repetitions or other features which significantly affect the realization of prosodic features were excluded. Consequently, 1758 phrases were used in the analyses. As the speech material comes from different speakers some f0 normalization was necessary. In the current study the pitch targets were normalized relative to the mean and S.D. determined for a

given speaker and phrase (initial, medial, final, single). Consequently, the targets predicted by the regression model will be z-score normalized f0 values and in contour generation they will be re-scaled and then smoothed and interpolated through to produce a continuous f0 contour.

In the following section the training and testing, and performance of the model are described.

7.3.2. Selection, training and testing of the regression network

The first step towards model building consisted in the specification of the type and complexity of neural networks to be trained. Like in the previous experiment *Statistica Neural Networks* (SNN) software was used in which the following network types can be applied to regression problems: linear, multilayer perceptrone (MLP), general regression neural network (GRNN) and radial basis function (RBF). For the reasons described in the previous section only MLP and linear networks were taken into account for the preliminary training and assessment of regression models. This task was carried out with the help of *Intelligent Problem Solver* program available in the SNN package. The MLP networks were trained using back propagation and/or conjugate gradient descend method. All networks had 3 layers and the number of units in the hidden layer varied between 1 and 20. The networks were trained using the same variable set as in the previous experiment based on the unit selection corpus, see sec. 7.2. In the training either the whole set of independent (predictor) variables (36) was used or alternatively, the network was allowed to automatically select a subset of the most useful variables. The database used in the current experiments contained 9722 syllables. They were split into 3 subsets:

- training (6892) used to optimize the network
- selection (1430) used to halt training to mitigate over-learning and/or to select from a number of models trained with different parameters
- test (1430) used to perform an unbiased estimation of the network's likely performance

The train and select subsets were resampled, but the test set was maintained the same to allow comparison of results.

In order to assess the performance of the models the correlation between the observed and estimated f0 values was analyzed (this can be done in a number of ways, see p. 183) as well as the ratio of the prediction to data standard deviations: if this is 1.0, then the network has performed no better than a simple mean estimator. A ratio below 1.0 indicates a good regression performance.

In general, MLPs performed much better than linear models in terms of correlation between observed and predicted f0 values, and prediction error SD. Moreover, it was observed that better results were achieved when only a subset of the input variables was used in the training of the models instead of all variables. On the basis of these results only MLP networks were subject to further training with a subset of input variables automatically selected by the network designer.

Depending on the network features correlations (generated for all selected cases) between the observed and predicted values of syllable-initial pitch (f0start) varied between 0.57 and 0.63, 2) the nucleus-medial f0 (f0mid) varied between 0.61 and 0.7, and 3) syllable-final f0 (f0end) varied between 0.62 and 0.7.

The best results in terms of correlation, as well as SD prediction error were achieved for an MLP network with 15 neurons in the hidden layer and 10 input features. The summary of the model is given in Table 60. It can be seen that the performance is similar on the training, selection and test subsets. The low network error indicates that no over-learning occurred.

network type/config.	perf. train.	perf. select.	perf. test.	train. error	select. error	test. error	training	input feat.	hidden(1)
MLP 15:10:3	0,77	0,80	0,77	0,07	0,08	0,07	BP88b	15	10

Table 60: Model training summary: regression network, expressive speech corpus.

7.3.3. Results

Table 61 presents the regression statistics computed for the best MLP network designed in the current experiment. For each subset (training, selection and test) the following information is presented:

a) ***Data Mean.*** Average value of the target output variable
b) ***Data S.D.*** Standard deviation of the target output variable.
c) ***Error Mean***. Average error (residual between target and actual output values) of the output variable.
d) ***Abs. E. Mean***. Average absolute error (difference between target and actual output values) of the output variable.
e) ***Error S.D.*** Standard deviation of errors for the output variable.
f) ***S.D. Ratio***. The error to data standard deviation ratio.
g) ***Correlation***. The standard Pearson-R correlation coefficient between the predicted and observed output values. Correlation shows how well the predicted contour's variation follows the target contour.

subset	training			selection			test		
statistics	**f0start**	**f0mid**	**f0end**	**f0start**	**f0mid**	**f0end**	**f0start**	**f0mid**	**f0end**
data mean	0,05	0,02	-0,04	0,05	0,02	-0,04	0,06	0,05	-0,02
data S.D.	0,94	0,97	1,03	0,96	0,99	1,01	0,93	0,96	1,01
error mean	-0,02	-0,03	-0,04	-0,02	-0,01	-0,02	-0,02	-0,04	-0,07
Abs.error mean	0,72	0,69	0,74	0,77	0,74	0,73	0,72	0,69	0,72
error S.D.	0,54	0,52	0,54	0,58	0,55	0,54	0,54	0,53	0,54
S.D. ratio	0,77	0,72	0,72	0,80	0,74	0,72	0,77	0,72	0,72
Correlation	0,64	0,70	0,70	0,61	0,67	0,70	0,63	0,70	0,70

Table 61: Regression statistics, expressive speech corpus.

Like in the previous experiment based on the unit selection corpus it can be seen that the model's performance on the training subset is the best in terms of both correlation and SD ratio which is well below 1 and indicates good regression performance. The performance on the test subset is better than on the selection subset as regards estimation of f0start and f0end values. The SD ratio is comparable in the two subsets and it is below 1 - this means that the network is capable of providing good estimation of the output values. The performance of the regression model designed in the current study is depicted in the graphs below where observed vs. predicted syllable-initial (f0start), nucleus-medial (f0mid) and syllable-final (f0end) f0 values estimated on the basis of the training and test subsets are plotted. On the x axis estimated z-score speaker and phrase type (initial, medial, final single) normalized f0 values are depicted, whereas on the y axis - z-score normalized observed f0 values.

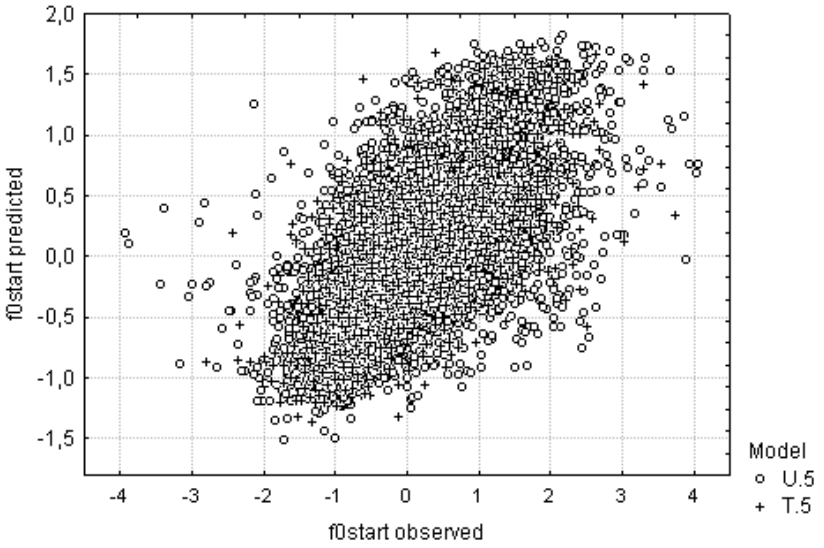


Figure 60: Scatterplot of the observed vs. estimated f0start target values.

Like in the unit selection corpus it can be seen that the range of predicted f0start values (between 1,3-1,8) is more constrained than that of original f0 values (in the range between -2 and 2,5 with a considerable amount of outliers below and above these values). It may imply some deficiency of the regression model, but on the other hand it should not affect the quality of generated intonation too much, because as explained at the beginning of this chapter (p.176) f0start value constitutes a less significant pitch target than f0mid or f0end. It is so, because with the exception of syllables of a phrase-initial position or preceded by syllables marked with a break index larger than 1, f0start value can be replaced with f0end value of the previous syllable. For that reason, the estimation performed by the model should be sufficient for generation of naturally-sounding f0 contours.

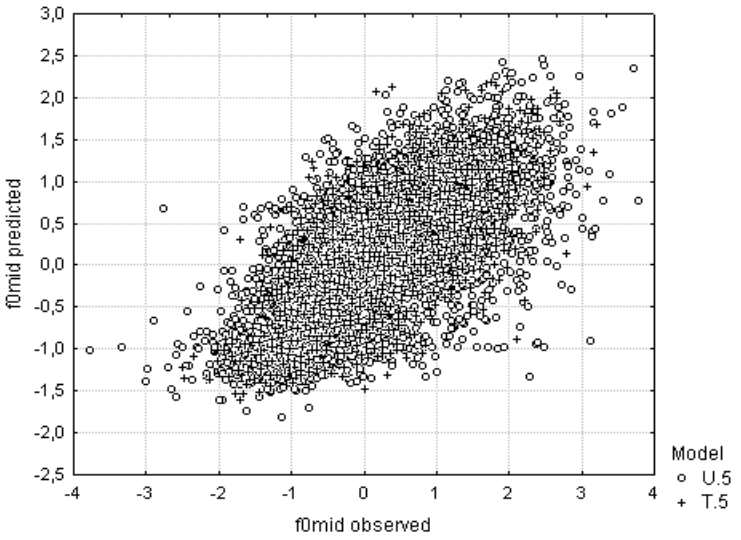


Figure 61: Scatterplot of the observed vs. estimated f0mid target values.

The correlation between observed and estimated nucleus-medial f0 values is higher in comparison to that between observed vs. predicted values of f0start, because they differ to a lesser extent. This can be concluded on the basis of the analysis of scatterplots of observed vs. estimated f0 values presented in the graphs a) and b). It can be seen that the bottom of the range is at the level of -1.5 for the estimated values as opposed to -2 (with some outlying values of even lower level). As regards the top of the range, most of f0mid values estimated with the model do not exceed the level of 2.0, whereas for the observed f0mid values the top limit seems to be 3.

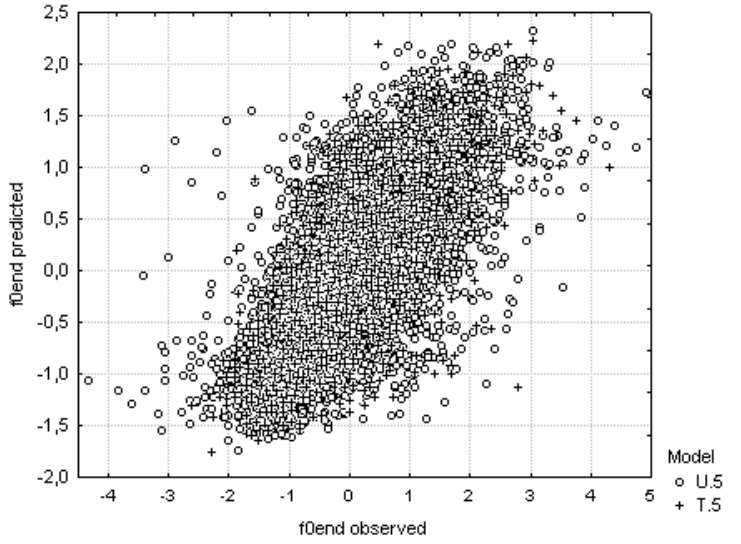


Figure 62: Scatterplot of observed vs. estimated f0end target values.

The correlation between observed vs. predicted f0end values is the same as that between observed vs. predicted f0mid values (r=0.7) and in general, the f0end data is distributed in a similar way to f0mid: It can be seen that the bottom of the range is at the level of -1.6 for the estimated f0end values as opposed to -2 (with some outlying values of even lower level). As

regards the top of the range, most of f0end values estimated with the model do not exceed the level of 2.0, whereas the majority of the observed f0end values reach the level of 3.

The regression model designed in this experiment performs the target estimation using a combination of 15 input features which were automatically selected during the network training. A sensitivity analysis was carried out to investigate the contribution of particular input features to the estimation of output f0 values. The list below shows feature importance ranking starting from the most important feature and ending at the least important one.

a) *ACC(PA1)* - accent type on the previous syllable
b) *ACC(SA)* - accent type on the current syllable
c) *ACC(FA1)* - accent type on the next syllable
d) *dist(ip_start)* - syllable position in the phrase expressed in the number of syllables from the phrase (ip) start
e) *dist(ip_end)* - syllable position in the phrase expressed in the number of syllables to the phrase (ip) end
f) *dist_start(+acc)* - gives the distance of the syllable to phrase start measured in the number of accented syllables
g) *ET(SA)* - boundary tone on the current syllable
h) *ACC(FA2)* - accent type on the second next syllable
i) *dist_next(+acc)* - distance of the current syllable to the next accented syllable (measured in the number of syllables)
j) *dist_end(+acc)* - gives the distance of the syllable from phrase start measured in the number of accented syllables
k) *ET(FA2)* - boundary tone on the second next syllable
l) *ET(PA2)* - boundary tone on the second previous syllable
m) *dist_prev(+acc)* - distance of the current syllable from the previous accented syllable (measured in the number of syllables)
n) *BI(FA2)* - break index on the second next syllable
o) *dist_start(+s)* - distance to the phrase start measured in the number of stressed syllables

It can be seen that 8 out of 15 features is directly related to prosodic features: accent, boundary tone and break strength described on the surface phonological level in terms of discrete distinctive categories. With the exception of syllable features dist(ip_start) and dist(ip_end) other features describing syllable position refer to prosodic features to some extent too.

These results confirm Hypothesis 1b which says that the surface phonological description proposed in this thesis which reflects melodic and functional aspects of intonation provides information of a high significance for the estimation of pitch targets and thus, for the results of contour generation.

7.3.4. Conclusions

In general, the performance of the regression model designed for estimation of pitch targets in the expressive speech corpus is not as good as that of the regression model based on the unit selection corpus. This is indicated by lower correlation between the observed and estimated values of f0 targets obtained for the former than for the latter model. However, the performance of the regression model designed in the current study is comparable to the results reported by other authors e.g., in (Black & Hunt 1996) the correlation between observed vs. estimated f0 targets was 0.53/0.55 for syllable-initial pitch, 0.66/0.68 for vowel-mid pitch and 0.56/0.55 for syllable-final pitch in the training/test sets respectively.

In the study presented in (Oliver & Clark 2005) which deals with Polish intonation modeling the average correlation between observed vs. estimated targets in the training subset was 0.68 which is the same as in the current study, and on the test subset r=0.71 as opposed to r=0.68 obtained in this study. The results presented in the current study compare favorably with those reported by other authors when the correlation between whole contours is taken into account. In (Dusterhoff & Black 1997) the correlation between original smoothed f0 contours and contours generated from estimated Tilt parameters (startf0, amplitude, duration, tilt, peak position) in the test set varied between 0.55 (for prediction using ToBI labels) and 0.6 (for prediction using automatic Tilt labels), while in the current study this correlation achieves 0.68 in the test set. This result can be considered as a good one taking into account the fact that the speech corpus used in (Dusterhoff & Black 1997) was the Boston Radio News Corpus (Ostendorf, Price & Shattuck-Hufnagel 1995) thus, the speech material represented news commentary style which is characterized by significantly less variable prosody than expressive speech and consequently, it should be easier to model.

The overall 0.68 correlation (test sample) between generated f0 contours vs. original ones (i.e., obtained with the getf0.psc script, see sec. 4.3.1) achieved in the current study compares favorably also with the results reported in (Dusterhoff, Black & Taylor 1999) with only one exception. The authors report on 0.6 correlation between f0 contours generated from originals and predicted Tilt model parameters in the new commentary corpus and 0.53 in the instructional text corpus. The exception is r=0.74 between original and generated contours in the isolated sentence corpus.

Apart from the high correlation coefficient which is regarded as a good indicator of model's performance, the quality of target estimation achieved with the regression network designed in the current experiment is also confirmed by the low value of prediction error S.D. (between 0.52 and 0.58). This value indicates that the network's output is not constrained to lie within the range of observed (input) values and proves that the model is capable of performing a "reasonable" extrapolation.

The comparison with other models presented here shows that the performance of the model designed in the current study achieves comparable or even better results. Consequently, it can be assumed that this partly confirms Hypothesis 3 and proves that the intonation model developed here is comprehensive and provides a framework for generation of intonation contours which are very similar to the natural contours.

In order to asses the naturalness of the generated contours and confirm completely the Hypothesis 3 a perception study was carried out which is reported in the next section.

7.4. Perceptual evaluation

As explained at the beginning of this chapter, an objective evaluation of the results of contour generation is not an entirely reliable indicator of model's performance (Clark & Dusterhoff 1999) and therefore, a subjective evaluation is necessary as well.

In this section the results of a perception test which aimed at a subjective evaluation of pitch contour modeling with the regression models designed in the previous experiments are described. The test results are expected to finally prove the hypotheses tested in the previous sections, namely that the approach to f0 generation proposed in this thesis provides a high quality speech characterized by natural intonation and that it can be successfully applied to various speech styles represented in the current study by the speech material included in the unit selection corpus and expressive speech corpus.

In order to prove these hypotheses intonation contours of utterances not included in the training or testing of the regression models were generated from syllable-based f0 targets estimated with the models and resynthesized with the original waveforms.

The f0 generation and resynthesis procedure, the perceptual test procedure and results are discussed in detail in the following sections.

7.4.1. Preliminary remarks

Among the methods used for the perceptual evaluation of intonation models performance *mean opinion score* (MOS) is probably the most popular one (e.g. Syrdal et al. 1998, Mixdorff & Jokisch 2003). An alternative to the absolute rating of the quality of generated intonation is *comparison category rating* (CCR). In this approach the evaluation is based on the comparison of pairs of stimuli consisting of original and synthesized stimuli. The comparison may be relative to the similarity between the signals (the subjects are supposed to rate whether the signals are the same, different, very different, etc., see e.g. Clark & Dusterhoff 1999) or their quality (in this case the subjects are aksed to judge the quality of the second signal in the pair relative to the quality of the first signal, see e.g. Brinckmann 2006). It seems that from the two comparison-based methods of evaluation the former one presents an easier task for the listeners, but its results do not always provide an adequate information on the quality of the generated intonation. The fact that the resynthesized signal differs from its natural counterpart does not necessarily mean that it has an unacceptable intonation. The question about the similarity between the signals concerns melodic properties i.e., the *form of intonation*. Maybe, it would be more appropriate to concentrate on the *functional aspects of intonation* instead and to investigate the source of the perceived differences e.g., by asking whether "different" utterances can be used alternatively in the same situational context. A positive response could be interpreted as a confirmation that the generated intonation is equivalent to the original one as regards the meaning it conveys.

Taking these considerations into account in the current study the perceptual evaluation of the quality of f0 contour generation will be based on two measures. In the first place, a pairwise comparison of similarity between stimuli with the natural and generated intonation contour will be made. Apart from comparing melodic properties of stimuli the test participants

will be asked to indicate whether the perceived differences affect interpretation of the utterances. Secondly, mean opinion score (MOS) will be calculated based on the judgments of the quality of the stimuli resynthesized with the intonation contours generated with the model proposed in this thesis. It is assumed that with this methodology a more meaningful information on the quality of intonation modeling can be obtained than with a simple same/different/etc. distinction which does not take functional aspects of intonation into account.

The results of the two tests indicating similarity between the stimuli with the original and generated intonation contours on the one hand and high naturalness of the generated intonation contours on the other hand, will be regarded as a confirmation of the hypotheses tested in this section.

7.4.2. Preparation of stimuli

In order to be able to carry out the perception test the stimuli had to be created first, which involved the following steps:

a) extraction of original f0 contours (see sec. 4.3.1 for details)
b) extraction of data describing the temporal alignment and scaling of f0 targets in the original speech files: syllable start (in ms) plus f0 at the start, the middle of the nucleus plus f0 value at that point, syllable end plus f0 value at that point (it should be noted that this step was already taken at the stage of database preparation for the design of the regression models)
c) creation of pitch tiers from the data extracted in the previous step: the original (observed) f0 targets were interpolated through to give continuous f0 contours
d) resynthesis (with PSOLA) of waveforms with the pitch tiers including the original (observed) f0 targets
e) substitution of the original f0 targets with the predicted ones in pitch tier files - in the expressive speech corpus the targets had to be first rescaled, because the model was trained on a z-score normalized (with respect to mean and S.D. of a given speaker and phrase type) f0 data
f) smoothing and interpolation between the pitch targets
g) resynthesis (using PSOLA) of the waveforms with the f0 contours created in the 2 previous steps

The f0 extraction, manipulation and resynthesis with the waveform were carried out in *Praat* using the script *getf0.psc* and *resynthesis.psc*. The latter script is a modified version of the former script (cf. sec. 4.3.1) and it is below in the form that it was used for male voice data (WI and MW).

```
###resynthesis.psc

form resynthesis
comment Enter directory where files are kept:
sentence soundDir C:\Documents and Settings\lacrimosa\Pulpit\baza E\
endform
Create Strings as file list... list 'soundDir$'\*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
  select Strings list
    fileName$ = Get string... ifile
    name$ = fileName$ - ".wav"
Read from file... 'soundDir$'\'name$'.wav
To Manipulation... 0.01 120 450
Read from file... 'soundDir$'\'name$'.PitchTier
select PitchTier 'name$'
To Pitch (ac)... 0 65 15 no 0.03 0.55 0.01 0.35 0.14 220
Smooth... 5
Interpolate
select Pitch 'name$'
Down to PitchTier
select PitchTier 'name$'
plus Manipulation 'name$'
Replace pitch tier
select Manipulation 'name$'
Get resynthesis (PSOLA)
Write to WAV file... 'soundDir$'\'name$'res.wav
endfor
```

As a result of the procedure described above two sets of stimuli were obtained which varied only in the intonation keeping other acoustic features intact. One set consisted (A) of the stimuli resynthesized with the f0 extracted from the original utterance and the other (B) included stimuli resynthesized with the f0 contours generated from the pitch targets estimated with the regression models. For the purpose of a pairwise comparison of stimuli with the original and generated f0 contours a similar methodology to that presented in (Brinckmann 2006) was adopted and it follows the specifications for perception evaluation given in (ITU-T 1994, 1996). For every utterance the original sample (A) was paired with corresponding generated sample (B). Half of the stimuli was presented in the A-B order and the other half in a reversed order - in this way it is possible to investigate to which extent the subjects participating in the test are consistent in their decisions. Apart from that, eight control stimuli consisting either of the original (A-A) or generated samples (B-B) were created with the aim of investigating to which extent the subjects are able to discriminate between the presented stimuli.

Altogether 98 pairs of stimuli were generated on the basis of: 31 utterances from the unit selection corpus and 66 utterances from the expressive speech corpus. The latter included: 8 utterances read by the female speaker AD, 27 read by the speaker AW and 31 read by the male speaker MW. The utterances were selected randomly from among those utterances which were not used in the training or testing of the regression models. The stimulus pairs were saved in a single file (*concatenate* command available in *Praat* was used for that purpose) and the two stimuli were separated with a pause.

As regards the mean opinion score test only the stimuli with the generated f0 contours were used (i.e., 98 stimuli altogether).

The figures below illustrate examples of the stimuli used in the perception study. The top panel contains the waveform. Below it f0 contours are depicted: the original f0 contour is marked with a dashed line and the f0 contour resulting from interpolation between estimated pitch targets is marked with a solid line. The vertical dashed lines indicate word boundaries; the first boundary stands for a pause preceding the phrase start. In the Figure 63 the utterance from the expressive speech corpus (female speaker AD) is depicted: "Paragon! Gdzie paragon!" (The receipt! Where is the receipt!). In the Figure 64 the utterance from the unit selection corpus is depicted (WI, male speaker): "ponieważ Piotr jest bardzo leniwy, nie będzie mu się chciało iść tak daleko do sklepu." (As peter is so much lazy, he won't feel like going to the shop.). It can be seen that generally, the generated f0 contours are similar to the original ones.

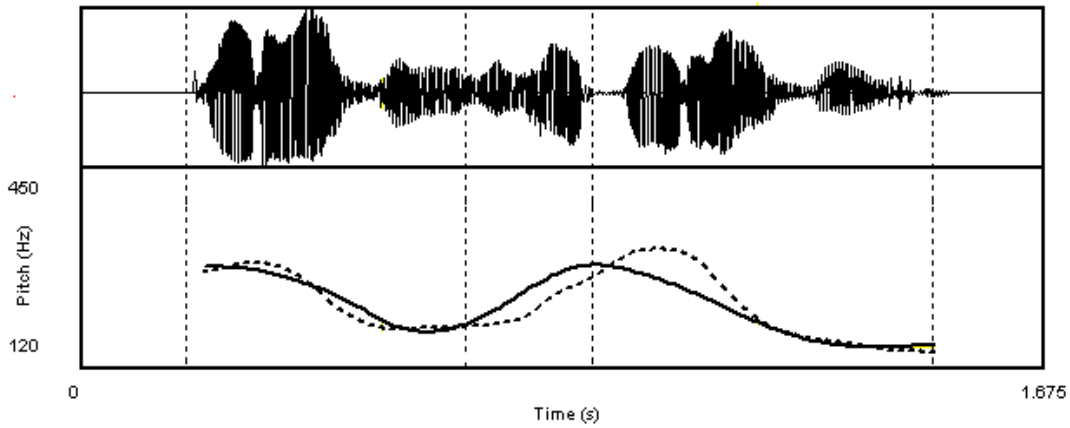


Figure 63: Example of a speech signal used in the perception test.
The original contour is marked with a dotted line, the generated f0 contour - with a solid line.
The dashed vertical lines indicate word boundaries.

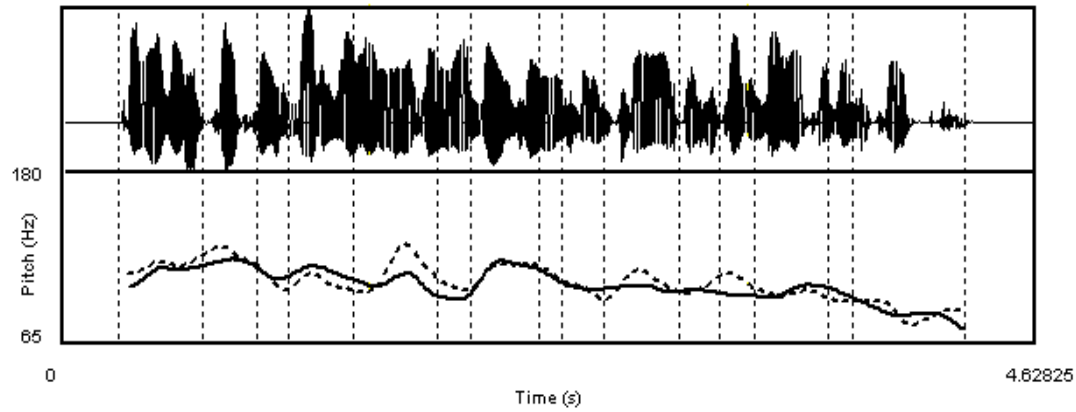


Figure 64: Example of a speech signal used in the perception test.
The original contour is marked with a dotted line, the generated f0 contour - with a solid line.
The dashed vertical lines indicate word boundaries.

7.4.3. Task, presentation and subjects

As mentioned before, the test was divided into two parts. In the first place, the pair-wise comparison of similarity between stimuli was evaluated on a 5-point scale, where 0 reflects no audible difference in intonation and 4 reflects no audible similarity. The same procedure was adopted in (Clark & Dusterhoff 1999) where correlation between objective (r, RMSE) and subjective (perceptual) evaluation of the quality of intonation generation was investigated. For the reasons given at the beginning of this chapter (see p.196) apart from decision as to whether the presented stimuli are similar or not, each time a difference was perceived (i.e., a rate <1 was chosen) the subjects were asked to indicate whether it involves modification of utterance's interpretation or not. If it does, it means the two stimuli can not be used alternatively in the same situational context and consequently, it can be assumed that the original and generated intonation contours are not equivalent as regards intonational meaning.

In the second part of the test (further referred to "intonation quality" test) the subjects judged the quality of the generated intonation by listening to the utterances resynthesized with the generated f0 contours. The assessment was done on a 5-point scale, where -2 indicates bad quality (the stimulus "sounds" unnatural), -1 stands for poor, 0 for fair, +1 indicates good and +2 indicates an excellent quality.

Prior to the start of each of the two experiments the subjects were given appropriate instructions and a preliminary training was carried out during which examples of stimuli including A-B, B-A, A-A, B-B pairs occurring in the similarity test as well as examples of stimuli occurring in the experiment investigating the quality of the generated intonation with example scores were presented.

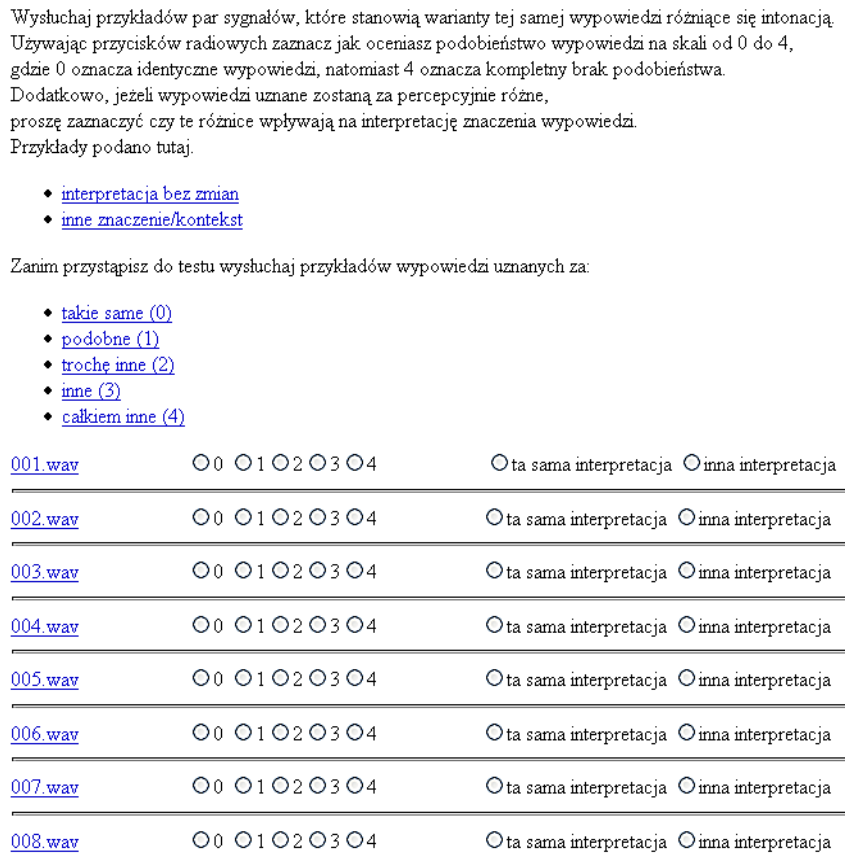
The test was conducted with a web-based interface which consisted of ten pages and designed by the author of this thesis. Six pages included forms with data for the similarity test and the other four included forms with data used in the second experiment.

For each stimulus pair a link was created and after clicking it the stimuli were played. Five ranking buttons were aligned to the right of the link corresponding to the ranking values specified for a given test. In the similarity experiment two additional buttons were created to mark the decision whether the stimuli that are perceived as different have different interpretation (i.e., they can not be used alternatively in the same situational context) or not.

There was no limit as regards how many times each stimulus could be played. The decision on the rank was made by marking appropriate radio button. After finishing the experiment, the subjects were asked to send the forms to the author. A fragment of the web interface used in the similarity experiment is depicted in Figure 65.

Ten subjects participated in the test. They were all native speakers of Polish and second year philology students. They had general knowledge on intonation and were familiar with some speech synthesis systems. The perception test took place in a computing laboratory; The stimuli were presented over headphones using standard audio software on Microsoft Windows XP workstations. The similarity test was ca. 40 minutes long and after a 20 minutes' break the MOS experiment was conducted which lasted about 30 minutes.

After collecting the forms sent by the participants the data included in the forms was exported into a Statistica spreadsheet where the mean of all scores (MOS) was calculated for each stimulus and statistical analyses were performed.


Wysłuchaj przykładów par sygnałów, które stanowią warianty tej samej wypowiedzi różniące się intonacją.
Używając przycisków radiowych zaznacz jak oceniasz podobieństwo wypowiedzi na skali od 0 do 4,
gdzie 0 oznacza identyczne wypowiedzi, natomiast 4 oznacza kompletny brak podobieństwa.
Dodatkowo, jeżeli wypowiedzi uznane zostaną za percepcyjnie różne,
proszę zaznaczyć czy te różnice wpływają na interpretację znaczenia wypowiedzi.
Przykłady podano tutaj.

- interpretacja bez zmian
- inne znaczenie/kontekst

Zanim przystąpisz do testu wysłuchaj przykładów wypowiedzi uznanych za:

- takie same (0)
- podobne (1)
- trochę inne (2)
- inne (3)
- całkiem inne (4)

001.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
002.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
003.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
004.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
005.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
006.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
007.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja
008.wav	○0 ○1 ○2 ○3 ○4	○ta sama interpretacja ○inna interpretacja

Figure 65: A screenshot of the web-interface used in the similarity test.

7.4.4. Results of the similarity test

In the first place, it was checked whether all subjects were able to adequately perform their task which consisted in judging the similarity between the stimuli with the original intonation contour and stimuli resynthesized with the generated f0 contours. For that purpose the answers provided by the subjects were analyzed: it was assumed that only those subjects can be taken into account who used at least four different ratings and who rated the control pairs (A-A, B-B) with 0 or 1. As a matter of fact, all the subjects effectively used the 5-point scale. As regards the second requirement only one in each seven control pair of stimuli was ranked with 2 (a little different), but no difference in the interpretation of the utterances was indicated by the listeners. Moreover, the average rating for A-A stimuli pairs was 0.01 and for B-B pairs 0.04. Consequently, all the answers were used in statistical analyses.

In a number of ANOVA and Scheffe's tests the effect of factors such as speech style, order of presentation (i.e., A-B vs. B-A, see sec. 7.4.2), speaker, sentence length and tune type on the perceived similarity between the stimuli was investigated.

ANOVA results show that the effect of speech style on the average similarity ratings received by the stimuli is not statistically significant ($p=0.67$, $F=0.19$). MOS for the stimuli pairs from the unit selection is 1.7, whereas for the stimuli from the expressive speech corpus MOS=1.78 (a little different). These results indicate that the listeners can hear differences between the stimuli with the original and generated intonation contours.

As regards the order of presentation it has no significant effect on the perceived similarity between the stimuli ($p=0.3$, $F=1.09$): for A-B pairs the average MOS is 1.78, whereas for B-A pairs it is only slightly higher and equals 1.94.

As the expressive speech corpus consists of recordings from three speakers, the interaction between speaker and rating of the stimuli was analyzed. ANOVA results show that the effect of speaker is statistically significant ($p<0.01$, $F=13.5$). In general, the male voice (MW) stimuli received the highest similarity ratings with the average MOS=1.5 (i.e., in between similar and a little different). The worst result in terms of the perceived similarity between the stimuli with the original vs. generated intonation was obtained for the female speaker AD with the average MOS=2.9 signaling that the stimuli are distinct. As regards the other female speaker (AW) the average MOS is 2.07 (a little different).

Neither tune type ($p=0.6$, $F=2.4$) nor sentence length ($r=0.12$) had significant effect on listeners' ratings.

In separate analyses the interaction between speech style, order of presentation, speaker, sentence length, tune type and listeners' judgments concerning the interpretation of the stimuli perceived as different (ratings >1) was investigated.

In the first place, the interaction between average MOS and percentage of stimuli pairs which were judged as having different interpretation was investigated. There is a strong correlation between them ($r=0.65$), which suggests that the greater the perceived difference between the stimuli the more distinct their interpretation becomes.

ANOVA results show that the effect of speech style is statistically significant ($p<0.01$, $F=19.45$). In general in the emphatic speech (AW, AD and MW speakers, expressive speech corpus) almost half of the stimuli pairs rated as a bit different, different or completely different was at the same time judged as having different interpretation, whereas in the unemphatic speech (WI speaker, unit selection corpus) this percentage was significantly lower (21%).

As regards the effect of speaker it is also statistically significant ($p<0.01$, $F=6.65$); the worst results were obtained for AD female speaker from the expressive speech corpus. Almost 70% of the stimuli pairs which were scored 2, 3 or 4 was at the same time judged as having different interpretation. This percentage was the smallest for the MW speaker (34.4%) and for the other female speaker (AW) is was in between with half of the stimuli judged as having different interpretation.

ANOVA shows that the order of presentation of stimuli (A-B vs. B-A) could affect listener's judgments: the average percentage of stimuli judged to have different interpretation is higher for B-A pairs (mean=43.3) in comparison to A-B pairs (mean=31.8%).

Like in the previous analysis, no statistically significant effect of tune type ($p=0.19$, $F=1.57$) or sentence length ($r=-0.03$) on listener's ratings of the interpretation of the stimuli was observed.

The results discussed here show that there are perceptual differences between stimuli with the natural intonation contours and those resynthesized with the intonation generated from pitch targets. On average, the stimuli were perceived as a little different from each other, but as it could be seen listener's judgments depended on the speaker.

Apart from that, the test results confirm the observation made at the beginning of this section (p. 196) that evaluation of the quality of intonation modeling in terms of the perceived similarity between speech signals with natural and generated f0 contours is not very informative, because it does not take into account the functional aspects of intonation. This can be concluded on the basis of the listener's judgements concerning the interpretation of perceptually different stimuli. It was shown that in such cases the percentage of utterances which indeed have different interpretation varies between 21% and 70% depending on factors such as speech style, speaker and order of stimuli presentation. This clearly shows that the fact that two stimuli differ with respect to their melodic features does not necessarily mean that they can not be used interchangeably in the same situational context. If so, it means that the original and generated intonation contours are equivalent.

Examples of stimuli which received different scores in the test are depicted in Figures 66-69.

In Figure 66 stimuli judged as perceptually very similar (average rating=0.8) are illustrated. The example comes from the expressive speech corpus, female speaker (AW). The utterance was: "na lewej nowe, lśniące czółenko" (on the left, a new, shiny shoe). The top panel contains the waveform, below the original (dashed line) and generated pitch contour (solid line) are depicted, at bottom panel contains the transcription: vertical lines indicate phoneme boundaries.

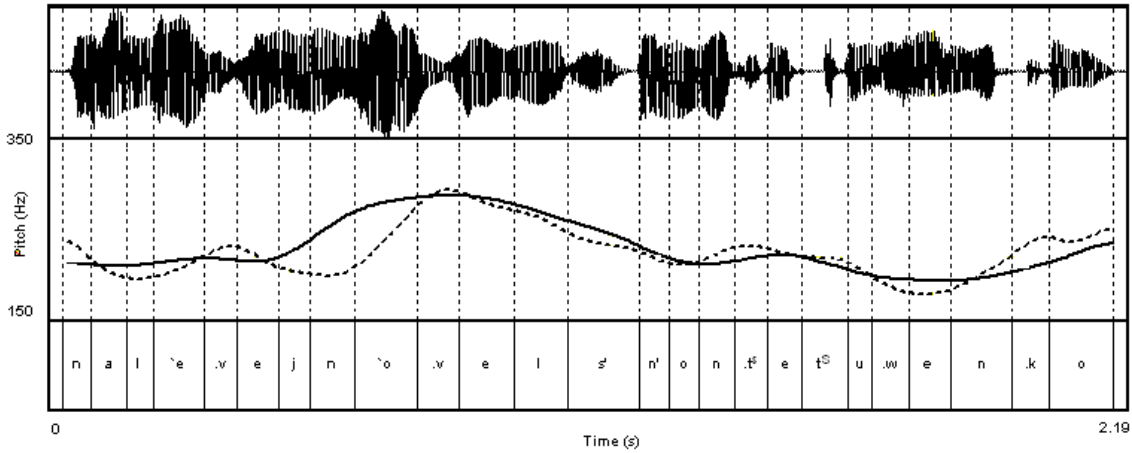


Figure 66: Example of two tunes judged as perceptually very similar (rating=0.8).
The original contour is marked with dashed line, the generated contour - with a solid line; expressive speech corpus, female speaker (AW).

The figures below illustrate stimuli which received low similarity ratings (<3) and were indicated as having different interpretation. From top to bottom the waveform is depicted, the original (dashed line) and generated pitch contour (solid line), the transcription panel: vertical lines indicate phoneme boundaries.

In Figure 67 & Figure 68 the original contour is emphatic, but not the generated one; both examples come from the expressive speech corpus. In the Figure 67 the utterance was: "Wspaniały sklep!" (A wonderful shop!), female speaker (AW). In the Figure 68 the utterance was: "Duszko, serdeńko, ślicznotko!" (My soul, my heart, my beauty!), male speaker (MW).

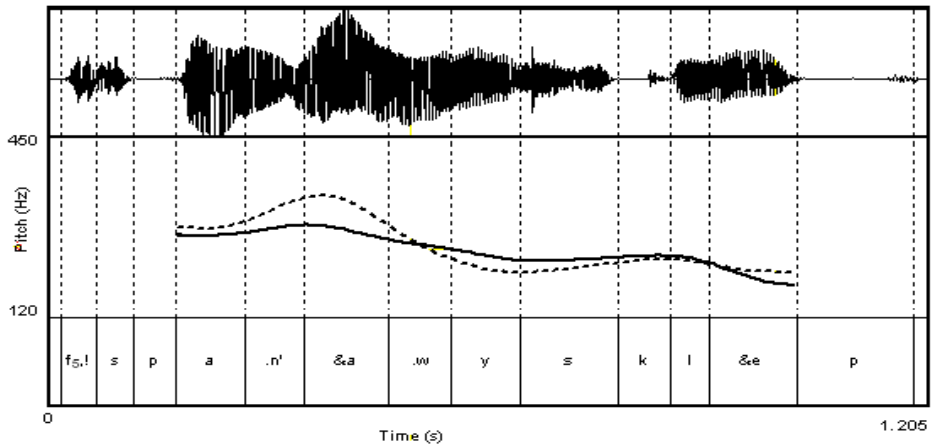


Figure 67: Example of two perceptually different contours conveying different meaning.
The original contour is marked with dashed line, the generated contour - with a solid line; expressive speech corpus, female speaker (AW).

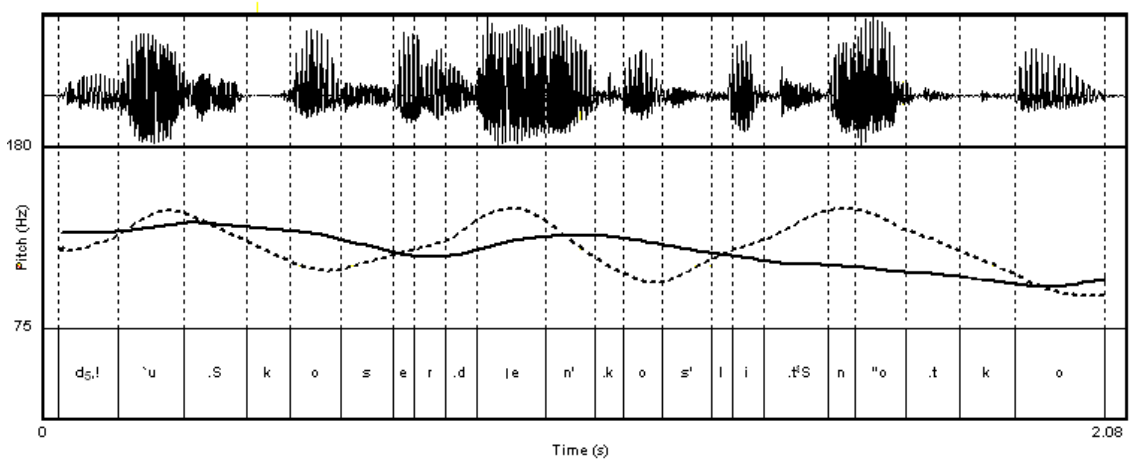


Figure 68: Example of two perceptually different contours conveying different meaning:
The original contour is marked with dashed line, the generated contour - with a solid line; expressive speech corpus, male speaker (MW).

The two tunes illustrated in the Figure 69 were judged as having the same interpretation, even though they were perceived as a little different considering their melodic properties (the average rating was 2). In spite of that, the stimulus resynthesized with the generated contour received a high rating (average=1,5) in the second perception test in which listeners judged the quality of the modeled intonation contours.

The example comes from the unit selection corpus and the utterance was: "Mój organizm jest bardzo czuły na wszelkie zmiany pogody". (My organism is very sensitive to all weather changes); male speaker (WI).

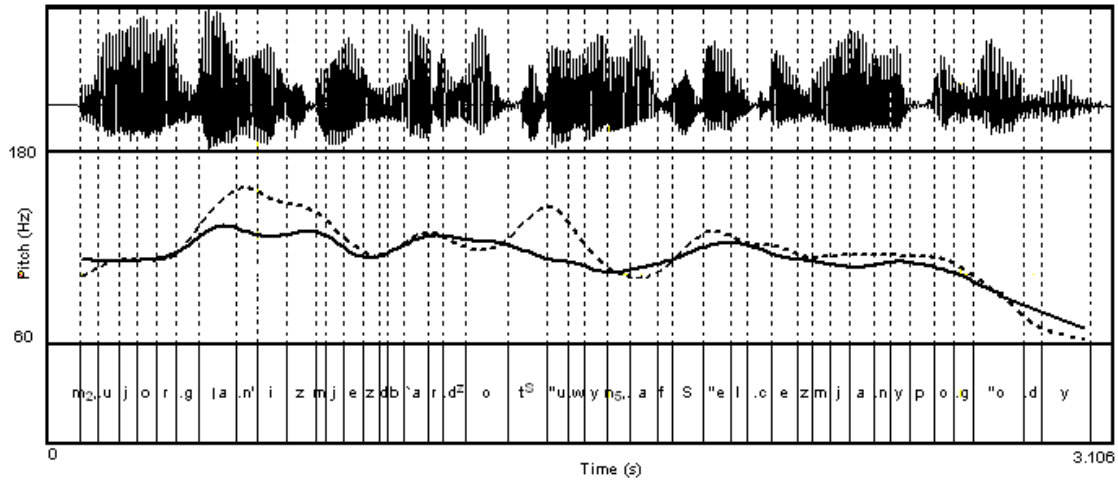


Figure 69: Example of two perceptually different tunes having the same interpretation: The original contour is marked with dashed line, the generated contour - with a solid line; unit selection corpus.

7.4.5. Results of the "intonation quality" test

Like in the previous experiment, at first, it was checked whether all subjects were able to adequately perform their task which consisted in judging the quality of the stimuli resynthesized with the generated f0 contours. For that purpose the answers provided by the subjects were analyzed: it was assumed that only those subjects can be taken into account that used at least four different ratings. Apart from that, As a matter of fact, all the subjects effectively used the 5-point scale and consequently, all the answers were used in statistical analyses.

In a number of ANOVA and Scheffe's tests the effect of the following factors on the perceived quality of the stimuli was investigated: speech style, speaker, sentence length and sentence type.

ANOVA results show that the effect of speech style is statistically significant ($p<0.01$, $F=40.98$). In general the unemphatic speech (WI speaker, unit selection corpus) received higher ratings than the emphatic speech (AW, AD and MW speakers, expressive speech corpus) i.e., MOS=0.88 vs. MOS= 0.17. These results show that the approach to f0 generation proposed in the this thesis is more appropriate for intonation modeling in unemphatic than emphatic speech. However, as the average ratings for the stimuli representing the two speech styles are *fair* and *good*, the results can be generally regarded as satisfying.

There is also an effect of tune type on the average rating of the perceived stimulus quality ($p<0.01$, $F=5.12$), but Scheffe's test results show that it is statistically significant only between statements vs. continuation phrases. In general, statements are perceived as the most

naturally sounding with the MOS=0.61, and continuation phrases received the lowest ratings: MOS=0.04.

As the expressive speech corpus consists of recordings from three speakers, the interaction between speaker and rating of the stimuli was analyzed, but ANOVA results show that the effect of speaker on MOS received by the stimuli is not statistically significant (p=0.69, F=0.36).

As regards the sentence length no significant correlation between this factor and ratings of the stimuli is observed (r=0.01).

The figures below illustrate stimuli which received the different average ratings in the perception test. In the figures the top panel contains the waveform. Below it, the generated pitch contour is depicted (solid line) and for comparison the original f0 contour (dotted line) is given. The bottom panel contains orthographic transcription; vertical lines indicate word boundaries.

Figure 70 illustrates example of a high-quality generated intonation (average rating=1.5). The utterances was: "Pijany turysta rzucił się do ucieczki przed rozwścieczonym gospodarzem zajazdu" (~the drunk tourist fled from the furious owner of the roadhouse). The utterance comes from the unit selection corpus, male speaker (WI).

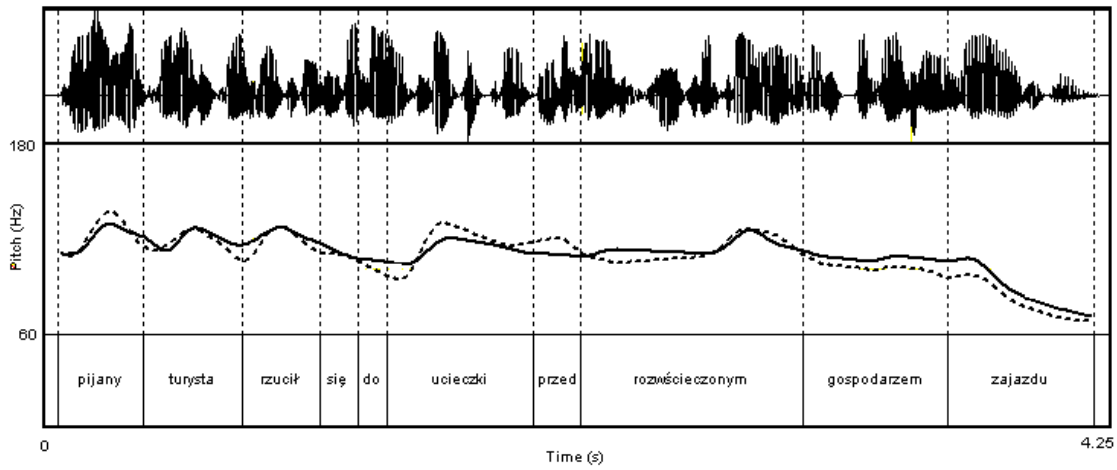


Figure 70: Example of a high-quality generated intonation.
The generated contour is marked with solid line, for comparison - the original contour is depicted (dotted line); unit selection corpus, male speaker (WI).

Figure 71 gives example of a lower-quality generated intonation (average rating=-0.8). The utterances was: "Z kotami nie wolno!" (Cats are not allowed!). The utterance comes from the expressive speech corpus, female speaker (AW).

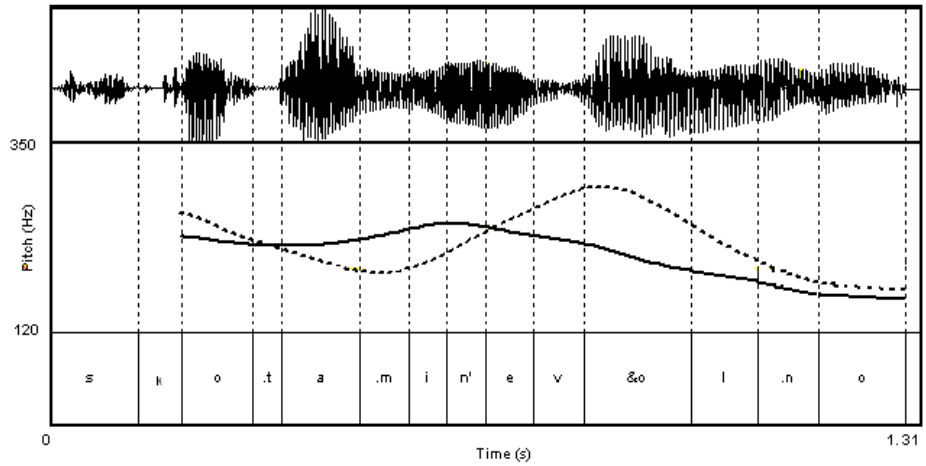


Figure 71: Example of a lower-quality generated intonation, expressive corpus, female speaker.

7.4.6. Conclusions

The results of the two perception experiments presented in sec. 7.4.4 & 7.4.5 show that the utterances resynthesized with the intonation contours generated with the model proposed in this thesis are on average perceived as a little different from the utterances having original intonation, but only in 37.7% of cases the differences in the melodic properties of the utterances cause difference in the interpretation of the intonational meaning. Moreover, the utterances with the generated intonation contours were perceived as naturally sounding. This conclusion can be drawn on the basis of the average ratings of the stimuli which are in between fair and good depending on the speech style and to lesser extent - tune type. This confirms hypothesis Hypothesis 2c that the approach to f0 contour generation proposed in this thesis provides a high quality speech characterized by natural intonation. The fact that the stimuli from the expressive speech corpus also received positive ratings (MOS=0.17) confirms the Hypothesis 3 and proves that the intonation model developed in this thesis is comprehensive, because it provides a framework for generation of a naturally sounding intonation in expressive speech.

It should also be noted that the results prove one more thing, namely they show that the normalization relative to mean and S.D. determined for a given speaker and phrase (initial, medial, final, single) is adequate and can be effectively applied to eliminate across and within-speaker pitch variation.

Chapter 8. Conclusions

This chapter summarizes the results of the research carried out in the scope of this thesis. At first, the main findings of the analyses and experiments are discussed and results of testing the hypotheses formulated in the beginning are presented. Then, the possibilities of application of the methods designed in the dissertation to speech synthesis and recognition are sketched. The chapter ends with an outlook on future work.

8.1. Summary of the main findings

The research presented in this dissertation was focused on the development of the *comprehensive intonation model for speech synthesis*.

As explained in sec. 1.2 the term *comprehensive* as it was used in this thesis refers to three aspects:

a) the levels of representation and analysis for description of intonation taken into account in the model
b) the structure of the model and the tasks it performs
c) speaking styles taken into account in the model

Consequently, three major hypotheses were formulated which refer to these aspects and which were tested in the thesis. The first hypothesis was the following:

Hypothesis 1. *The intonation model developed in this thesis can be regarded as comprehensive if it provides a framework for description of intonation at various levels of analysis which is useful for coding and generation of f0 contours.*

As explained in sec. 1.1.4 in the comprehensive intonation modeling two levels of representation and analysis of intonation are taken into account: *phonetic* and *surface phonological level*.

The surface-phonological description of intonation proposed in this thesis (sec. 5.2) was based on the prosody labeling system used in the annotation of unit selection corpus for Polish module of BOSS TTS system (Breuer et al. 2000). The description is in terms of discrete distinctive categories and encodes not only melodic, but also functional aspects of intonation.

On the surface phonological level intonational tunes are defined as strings of distinctive elements (intonational events) which are linked with the elements of the segmental string. Two types of elements are distinguished, namely pitch accents and boundary tones.

An inventory of five pitch accent types was proposed. The accents were distinguished on the basis of the following perceptually significant properties:

a) direction of the pitch movement (rise vs. fall) on the accented and post-accentual vowel

b) amplitude of the movement
c) position (timing) of the peak/minimum relative to accented vowel/syllable boundaries
d) the amount of pitch variation on the accented and post-accentual vowel

The resulting inventory consists of two falling accents (HL*, H*L), two rising (L*H, LH*) and one-rising-falling accent (LH*L). It was shown that the accents may have various structural roles in the tunes, but there are some restrictions concerning the possible combinations of pitch accents and boundary tones. As regards the boundary tones they are defined as distinctive, non-prominence lending pitch movements occurring at phrase boundaries. An inventory of boundary tones was proposed on the basis of features such as direction of the pitch movement, amplitude of the movement and scaling of the pitch targets at the start and end of the movement. The resulting inventory consists of three falling (5,. 5,! 2,.) and two rising boundary tones (5,? 2,?). Unlike in other systems (e.g. ToBI) the information on the boundary type encodes also the information on the strength of prosodic break.

The reason behind definition of abstract and symbolic descriptions of intonation like the surface phonological description proposed here is that the information they provide improves the prediction and generation of f0 contours in speech synthesis (e.g. Syrdal et al. 1998, Möhler 2001).

On the phonetic level intonation is described in terms of continuous parameters and only melodic aspects of intonation are taken into account. The phonetic description of intonation proposed in this thesis (see sec. 5.3) was defined in a series of statistical analyses in which the contribution of different acoustic features to the distinction between types of pitch accents and boundary tones was investigated. The resulting description is compact - it uses only small feature vectors to express the fine differences between various types of intonational events and can be easily derived from utterance's acoustics. On the basis of the phonetic description it is possible to extract the higher-level surface phonological description of intonation, which is related to the problem of coding of f0 contours.

Apart from the description of intonation on the phonetic and surface phonological levels a finer description of prosodic structure was proposed on the basis of the results presented in sec. 5.1. In a series of analyses the effect of various classification of phrases on the variation in the pitch range was investigated. As a result, a description of prosodic structure was defined in which a distinction was drawn between two phrasing levels (major and minor intonational phrase), and phrases were categorized relative to their length (single vs. complex phrases) or position in the major phrase (initial, medial, final). The motivation for the analyses came from the study presented in (Clark 2003) which showed that incorporation of a more detailed information on prosodic structure improved the results of f0 contour generation in speech synthesis.

In order to investigate whether the description of intonation proposed in the thesis is useful for coding and generation of f0 contours (Hypothesis 1) a number of fine-grained hypotheses had to be tested first. To begin with the following hypothesis was formulated:

Hypothesis 1a. *The phonetic description of intonation proposed in this thesis provides information which is significant to the detection and classification of the elements of intonational tunes - pitch accents and boundary tones.*

This hypothesis was tested in Chapter 6 where a number of models (including linear and non-linear models such as decision trees, MLP and RBF neural networks) was designed for the purpose of automatic detection of the location of accented syllables and phrase boundaries, and classification of pitch accent and boundary tone types. The detection and classification were performed on the basis of the acoustic features distinguished in the phonetic description of intonation. Thus, the task of the models was to derive the surface phonological description from the phonetic one. The models achieved high overall accuracy, which proved that the acoustic features adequately express the fine differences between types of pitch accents and boundary tones. Consequently, it can be said that the information which is provided in the phonetic description of intonation is significant to the detection and classification of the elements of intonational tunes, which confirms the Hypothesis 1a.

The next two hypotheses were formulated as follows:

Hypothesis 1b. *The surface phonological description proposed in this thesis which reflects melodic and functional aspects of intonation provides information of a high significance to the estimation of pitch targets and thus, to the results of contour generation in speech synthesis.*

Hypothesis 1c. *The description of prosodic structure proposed in this thesis provides an important information for the estimation of pitch targets and thus, affects the performance of the regression model and the overall quality of intonation modeling.*

These hypotheses were tested in Chapter 7 where the approach to f0 contour generation in speech synthesis was proposed.

In this approach intonation contours are generated by interpolation between three pitch targets which are anchored in the syllable structure: at the start of the syllable, in the middle of nucleus and at the end of the syllable. Two regression models (one for unemphatic and the other for expressive speech) were designed to estimate the scaling of these targets on the basis of text-based and prosodic features. The latter included the information provided in the surface phonological description of pitch accents and boundary tones and in the description of prosodic structure. In order to check to which extent this kind of information affects the results of pitch target estimation sensitivity analyses were carried out. The purpose of sensitivity analysis is to identify which input variables are considered most important by a particular model. The results showed that the variables which referred to prosodic features (e.g. the type of phrase, pitch accent or boundary tone) had high sensitivity, which confirms the Hypothesis 1b and Hypothesis 1c.

In view of the findings discussed here Hypothesis 1 can be accepted.

The next hypothesis referred to the structure of and tasks performed by the intonation model. It was formulated as follows:

Hypothesis 2. *The intonation model proposed in this thesis can be regarded as comprehensive if it is bi-directional i.e., makes coding and generation of intonation contours possible and performs these tasks with a high accuracy.*

This hypothesis was tested in Chapter 6 and Chapter 7. As mentioned before, in Chapter 6 methods capable of automatic detection and classification of the elements of intonational tunes were designed, whereas in Chapter 7 an approach to generation of f0 contours in speech synthesis was proposed. In order to examine the performance of the designed models and confirm Hypothesis 2 three further hypotheses were formulated, namely:

Hypothesis 2a. *Automatic detection and classification of intonational events can yield accuracy comparable to the inter-labeler consistency in manual transcription of prosody.*

Hypothesis 2b. *A high accuracy in the automatic detection and classification of intonational events can be achieved even if only a small vector of acoustic parameters and information extracted from utterance's transcription/segmentation are used as input features to the model.*

Hypothesis 2c. *An approach to f0 generation in which f0 contours result from interpolation between pitch targets of a predefined position in the syllable structure (at the onset start, in the middle of the nucleus and at the end of the coda) whose values are estimated by means of a regression model provides a high quality speech characterized by natural intonation.*

In order to examine the performance of the models designed for automatic detection and classification of intonational events the accuracy achieved by the models was compared to the results reported for inter-transcriber consistency in manual labeling of prosody. In general, the accuracy achieved by the proposed models which is in between 77% and 88% depending on the specific task and type of the model is comparable to the consistency in manual prosody annotation, which proves the Hypothesis 2a. It is also comparable to the accuracy achieved by other models designed for automatic prosody labeling (an overview is given in sec. 6.1.2). Most of these models rely on higher-level linguistic information (e.g. POS) and define large feature sets (e.g. 276 features are used in Kießling et al. 1996). On the contrary, the models proposed in this thesis use only small feature vectors (eight features at most) which can be easily derived from utterance's acoustic (assuming that phoneme/syllable/word boundaries are known), but the models achieve comparable accuracy in the detection and classification of prosodic constituents. This proves Hypothesis 2b.

As mentioned before in this section, in the comprehensive intonation model developed in this thesis generation of f0 contours is carried out in two stages. Firstly, three pitch targets per syllable are estimated with a regression model on the basis of text-based and prosodic features. Secondly, the targets are smoothed and interpolated through, which results in a continuous f0 contour.

Two regression models were designed in the study presented in Chapter 7. One model was trained on the material representing unemphatic speech (see sec. 7.2) and the other was trained to predict targets for expressive speech (see sec. 7.3). The effectiveness of the proposed

approach to f0 contour generation was assessed in two different ways. Firstly, by comparing the objective measures of the models performance (correlation between the observed and estimated f0 targets) with the results presented in the literature. Secondly, a perception study (see sec. 7.4) was carried out in which the listeners judged the similarity between the stimuli synthesized with the original and generated f0 contours, as well as the overall quality and naturalness of the synthesized intonation.

The overall correlation of 0.75/0.68 between generated and original intonation contours in the unemphatic/expressive speech was comparable to the results reported in the literature (see sec. 7.1).

The results of perception experiments showed that the utterances resynthesized with the intonation contours generated with the approach proposed in this thesis were generally perceived as a little different from the utterances having original intonation, but only in 37.7% the differences in the melodic properties involved different interpretation of the intonational meaning. Moreover, the utterances with the generated intonation contours were perceived as naturally sounding, which could be concluded on the basis of the average ratings of the stimuli which varied from fair and good.

The results presented here show that the approach to f0 contour generation proposed in this thesis provides a high quality speech characterized by natural intonation (Hypothesis 2c).

In view of the findings on the performance of the classification and regression models Hypothesis 2 can be accepted.

The last hypothesis refers to speaking styles that are taken into account in the modeling. It was assumed that the comprehensive approach to intonation modeling should not be confined to a single speaking style, but on the contrary it should provide methods and solutions which can successfully be applied to intonation generation of unemphatic as well as expressive speech. On the basis of this assumption the following hypothesis was formulated:

Hypothesis 3. *The intonation model can be regarded as comprehensive if it provides a framework for generation of a high-quality, naturally sounding intonation of expressive speech.*

The hypothesis was confirmed by the results of the objective and perceptual evaluation of the quality of intonation contour generation in the proposed framework presented above.

As the speech corpus used to design the model for expressive speech contained material from three different speakers f0 normalization similar to that proposed in (Clark 2003) and adopted in (Oliver & Clark 2005) was used. Pitch targets were scaled relative to the mean and standard deviation determined for a given speaker and phrase (initial, medial, final, single according to the results presented in sec. 5.1). Consequently, the targets predicted by the regression model were z-score normalized f0 values and in contour generation they had to be re-scaled and then smoothed and interpolated through to produce a continuous f0 contour.

The fact that the generated intonation received high overall ratings in the perception test proved not only the usefulness of the approach to intonation generation in emphatic speech proposed in the thesis, but also indicated the adequacy of the method adopted to normalization of within and across-speaker pitch variation.

To conclude, all the hypotheses discussed above presented specific requirements for a comprehensive intonation model and determined the research goals. The summary given in this section showed that the intonation model developed in this thesis provides a framework for description of intonation at different levels (phonetic, surface phonological) which is useful for both coding and generation of intonation contours. Moreover, the model is bi-directional and provides solutions to the problem of automatic prosody labeling and intonation generation in speech synthesis. Finally, it was shown that the model can be effectively applied to various speech styles. All that proves that *the model developed in the current thesis is comprehensive*, thus it can be concluded that the major goal of the thesis was successfully accomplished.

8.2. Future work

The solutions proposed in this thesis make it possible to automatically describe, predict and generate intonation contours for a given input text. However, there still remain some important issues which have not been investigated in the scope of the study presented here and which can be regarded as future tasks.

First of all, it is necessary to get to know to which extent the information provided by the surface phonological description of intonation contributes to the accuracy of pitch target prediction, and to which extent it affects the quality of generated intonation.

On the one hand the sensitivity analysis showed that information related to pitch accent/boundary type and prosodic structure (e.g. strength of break following the constituent, phrase type) is the most important from the point of view of estimation of pitch targets by regression models (see sec. 7.2.3). On the other hand, the results of study presented in (Brinckmann 2006) suggest that there is no significant difference in the quality of generated speech irrespective of whether the symbolic prosodic features are taken into account or not. It is so, because "the error introduced by symbolic prosody prediction perceptually equals the amount of error produced by the direct method that does not exploit any symbolic prosody features" (op.cit.107). In the current study the prosodic features used by the regression models were obtained from the manual prosody transcription of the speech corpora. Since in the scope of this thesis methods of automatic labeling of intonational features and phrase structure were developed it seems reasonable to repeat the experiments on the basis of this information obtained automatically. It is interesting to which extent the accumulation of error in the prediction of prosodic features and estimation of pitch targets will affect the quality of generated intonation.

Another future task is related to the module of automatic labeling of prosody. The models presented in the thesis were designed on the basis of analysis of a single speaker corpus. The same approach is taken in the studies by other authors as well (examples were given in the sec. 6.1.2), but its drawback is that the models are speaker-dependent. Therefore, in the next step the same methodology will be applied to the design of speaker-independent models. For this purpose the expressive speech corpus may be used.

As mentioned above, prosodic features (see sec. 7.2) had a high position in the importance ranking of predictor variables used in the estimation of pitch targets by regression models. However, some facts suggest that this information is not entirely sufficient to generate a

high-quality intonation in expressive speech. In general, the results of objective evaluation (i.e., in terms of RMSE and correlation) and subjective evaluation (in a perception test) obtained for the expressive corpus are comparable to those obtained for the unit selection corpus (see sec. 7.2.3). However, in the expressive corpus examples of utterances resynthesized with the automatically generated contours can be found which have intonation resembling more that of a news-reading style than that of expressive/emphatic speech. This suggests that a more functional description of prosody may be required for generation of a naturally sounding expressive speech. An example of such a functional system applied to labeling of prosody in a spontaneous speech database is given in (Kießling et al. 1996) where pitch accent types are distinguished on the basis of their structural roles in tunes (primary, secondary, emphatic).

Apart from that, some higher-level linguistic information describing the syntactic and/or semantic structure of an utterance could be integrated into the models proposed in this thesis. The existing approaches make a limited use of this kind of features, nevertheless they might contribute to the performance of the intonation models. At present, the most commonly used higher-level information is POS: it appears in a number of studies dedicated to both automatic prosody labeling (e.g. Sridhar, Bangalore & Narayanan 2007) and intonation generation (e.g. Mixdorff & Jokisch 2003).

Another interesting research issue regards evaluation of intonation models. It seems that objective measures (RMSE, correlation) of similarity between natural intonation contours and those generated by the models do not reflect precisely their perceptual similarity (e.g. Clark & Dusterhoff 1999). Among the methods used for the perceptual evaluation of intonation models performance *mean opinion score* (MOS) is probably the most popular one (e.g. Syrdal et al. 1998, Mixdorff & Jokisch 2003). An alternative to the absolute rating of the quality of generated intonation is *comparison category rating* (CCR). In this approach the evaluation is based on the comparison of pairs of stimuli consisting of original and synthesized stimuli. The comparison may be relative to the similarity between the signals (the subjects are supposed to rate whether the signals are the same, different, very different, etc., see e.g. Clark & Dusterhoff 1999) or their quality (in this case the subjects are asked to judge the quality of the second signal in the pair relative to the quality of the first signal, see e.g. Brinckmann 2006). Yet, as suggested in the discussion in sec. 7.4.1 the most serious drawback of the methods of perceptual evaluation is that they primarily focus on melodic aspects of intonation and only secondarily on the functional aspects. The methodology used for the perceptual evaluation of f0 contour generation proposed in the current thesis was the first step towards developing of comprehensive methods of evaluation. The results of the perception tests in which the quality of the generated intonation was evaluated proved that apart from the measure based on the similarity between original and synthetic intonation the quality of the latter and its particular function should also be assessed.

Finally, it remains a future task to investigate to which extent the methods and solutions developed in the scope of this thesis are language-independent, as the next goal of the comprehensive approach is to propose an intonation model which can be easily adapted to new languages and which makes a cross-language analysis of intonation possible.

References

Allen J., Hunnicut S & Klatt D. (1987) *From Text To Speech, The MITTALK System*. Cambridge University Press, 1987

Anderson M., Pierrehumbert J., & Liberman M. (1984) *Synthesis by rule of English intonation patterns*. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2.8.2-2.8.4

Arvaniti A. & Ladd R.D. (1995) *Tonal alignment and the representation of accentual targets.* Proceedings of ICPhS 1995', Stockholm, Sweden, vol.4, pp. 220-223

Arvaniti A., Ladd R.D. & Mennen I. (1998) *What is a starred tone? Evidence from Greek.* In Broe & Pierrehumbert [editors], Papers in Laboratory Phonology V: Acquisition and the Lexicon. Cambridge University Press, pp. 119-131

Arvaniti A. & Baltazani M. (2000) *Greek ToBI: A system for the annotation of Greek speech corpora*. Proceedings of LREC, Athens, Greece, vol.2, pp.555-562

Arvaniti A. (2001) *The intonation of wh-questions in Greek.* Studies in Greek Linguistics 21:57-68. Thessaloniki.

Arvaniti A. (2002) *The intonation of yes-no questions in Greek*. In M. Makri-Tsilipakou [ed.], Selected Papers on Theoretical and Applied Linguistics, Thessaloniki: Department of Theoretical and Applied Linguistics, School of English, Aristotle University, pp. 71-83

Atterer M. & Ladd R.D. (2004) *On the phonetics and phonology of "segmental anchoring" of F0: evidence from German.* Journal of Phonetics 32:177-179

Auran C. (2004) Momel-INTSINT [Praat script] downloadable from: http://www.univ-. lille3.fr/silex/equipe/auran/english/index.html

Bangalore S. & Haffner P. (2005) *Classification of large label sets*. Proceedings of Snowbird Learning Workshop 2005

Baranowska E., Francuzik K., Karpiński M. & Kleśta J. (2003) *Determining phrase boundaries in written texts for the purpose of Polish speech synthesis*. Speech and Language Technology 7:71-78

Bard E.G., Sotillo C., Anderson A.H. & Taylor P. (1995) *The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment*. Proceedings of ESCA-NATO Tutorial and Workshop on Speech under Stress, Lisbon, Portugal, pp.25-28

Batliner A., Nöth E., Möbius B. & Möhler G. (2000) *Prosodic models and speech recognition: towards the common ground*. Proceedings of Prosody 2000: Speech Recognition and Synthesis (Kraków, Poland), pp. 15-20.

Baumann S., Grice M. & Benzmueller R. (2000) *GToBI - a phonological system for the transcription of German intonation*. Proceedings of Prosody 2000: Speech Recognition and Synthesis Workshop, Cracow, pp. 21-28.

Beckman M.E. & Pierrehumbert J. (1986) *Intonational structure in English and Japanese*. Phonology Yearbook 3:255-310

Beckman M.E. & Hirschberg J. (1994) *The ToBI Annotation Conventions*. Unpublished manuscript, Ohio State University and AT&T Bell Telephone Laboratories

Beckman M.E. & Ayers M. (1997) *Guidelines for ToBI labeling*. Department of Linguistics, The Ohio State University (http:www.ling.ohio-state.edu/phonetics/E_Tobi)

Berkovits R. (1994) *Durational effects in final lenghtening, gapping and contrastive stress*. Language and Speech, 37(3):237-250

Black A.W. & Hunt A.J. (1996) *Generating F0 contours from ToBI labels using linear regression.* Proceedings of ICSLP 96', Philadelphia, USA, 3:1385-1388

Boersma P. (1993) *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. Proceedings of the Institute of Phonetic Sciences, 17:97–110. University of Amsterdam

Boersma P. (1998) *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* The Hague: Holland Academic Graphics

Boersma P. & Weenink D. (2005) *Praat: doing phonetics by computer* (Version 4.3.14) [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/

Bolinger D. (1978) *Intonation across languages.* In: Greenberg [ed.],Universals of human language, vol. 2: Phonology, Palo Alto, CA: Stanford University Press, pp. 471-524

Botinis A., Granström B. & Möbius B. (2001) *Developments and paradigms in intonation research.* Speech Communication, vol.33 (4), pp.263-29

Braunschweiler N. (2003): *Automatic Detection of Prosodic Cues*. PhD thesis, University of Konstanz, Germany

Breiman L., Friedman J. H., Olshen R. A., & Stone C. J. (1984) *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software

Breuer S., Stober K., Wagner P. & Abresch J. (2000) *Dokumentation zum Bonn Open Synthesis System BOSS II*. Unveroffentliches Dokument, IKP, Bonn, 2000

Brinckmann C. (2006) *Improving prosody prediction for speech synthesis with and without symbolic prosody features*. Research report, Phonus no.10, University of Saarland,February 2006

Brown G. (1983) *Prosodic structure and the given/new distinctions*. In Cutler & Ladd [editors], Prosody: models and measurements. Heidelberg: Springer, pp.67-77

Bruce G. (1977) *Swedish Word Accents in Sentence Perspective*. Lund: Gleerup

Bulyko I. & Ostendorf M. (2001) *Joint prosody prediction and unit selection for concatenative speech synthesis*. Proceedings of ICASSP 2001

Campbell N. (1998) Acoustic Nature and Perceptual Testing of Corpora of Emotional Speech. Proceedings of ICSLP'98

Campbell N. (2004) *Expressive Speech - Simultaneous indication of information and affect*. In: G.Fant, H.Fujisaki, J.Cao & Y.Xu [editors], From Traditional Phonology to Modern Speech Processing, pp.49-58

Campione, E., Hirst, D., & Véronis, J. (2000). *Stylisation and symbolic coding of F0: comparison of five models*. In A. Botinis [ed.], Intonation: Models and Theories (pp. 185-208). Dordrecht: Kluwer Academic Publisher

Carlson R. & Swerts M. (2003) *Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials*. Proceedings of ICPhS 2003'

Caspers J. & van Heuven V.J. (1993) *Effects of time pressure on the phonetic realisation of Dutch accent-lending pitch rise and fall*. Phonetica 50:161-171

Chen A., Rietveld T. & Gussenhoven C. (2001) Language-specific effects of pitch range on the perception of universal intonational meaning. Proceedings of 9th *Eurospeech*, 2:91-94

Clark R. & Dusterhoff K. (1999) *Objective methods for evaluating synthetic intonation*. Proceedings of Eurospeech 1999 4:1623-1626

Clark R. (2003) *Generating Synthetic Pitch Contours Using Prosodic Structure.* PhD thesis, The University of Edinburgh, 2003

Clements G.N. & Keyser S.J. (1981) *CV Phonology: A Generative Theory of the Syllable*. Cambridge, MA: MIT Press.

Cooper W. & Sorensen J. (1981) *Fundamental frequency in sentence preduction*. Heidelberg: Springer

Cruttenden A. (1997) *Intonation*. Cambridge University Press

Crystal D. (1969) *Prosodic systems and intonation in English*. Cambridge University Press

d'Alessandro Ch., Mertens P.& Beaugendre F. (1994) *Automatic Stylization of Intonation: Application to Speech Synthesis*. Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis

De Pijper J.R. & Sanderman A. (1993) *Prosodic cues to the perception of constituent boundaries*. Proceedings of *Eurospeech*'93, pp. 1210-1214

De Pijper J.R. & Sanderman A. (1994) *On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues*. J. Acoust. Soc. Am. 96(4): 2037- 2047

Demenko G. (1999) *Analysis of Polish suprasegmentals for needs of Speech Technology*. Adam Mickiewicz University Press, Poznań

Demenko G. (2000) *Automatic analysis of phrase in Polish*. Speech and Language Technology 4:13-22

Demenko G., Wypych M. & Baranowska E. (2003) *Implementation of Polish grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis*. Speech and Language Technology 7:79-96

Demenko G. & Wagner A. (2006) *The Stylization of Intonation Contours.* Proceedings of Prosody 2006, Dresden, Germany

Demenko G. & Wagner A. (2007) *Prosody annotation for unit selection text-to-speech synthesis*. Archives of acoustics, 32(1):.25-40

Dłuska M. (1947) *Prosody of Polish*. Kraków

Dusterhoff K. & Black A.W. (1997) *Generating F0 contours for speech synthesis using the Tilt intonation theory*. Proceedings of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications, Athens, Greece, pp.107-110

Dusterhoff K. Black A.W. & Taylor P. (1999) *Using decision trees within the tilt intonation model to predict f0 contours*. Proceedings of *Eurospeech* 99', Budapest, Hungary, pp. 1627–1630

Dutoit T. & Stylianou Y (2003) *Text-to-Speech Synthesis***.** The Oxford Handbook of Computational Linguistics, Mitkov [ed.], Oxford University Press, 2003, Chapter 17, p.323-338

Féry C. (1993) *German intonational patterns*. Tuebingen: Niemeyer

Francuzik K., Karpiński M. & Kleśta J. (2002) *A preliminary study of the intonational phrase, nuclear melody and pauses in Polish semi-spontaneous narration.* Proceedings of Prosody 2002, Aix-en-Provence, ProSig and Universite de Provence

Fujisaki H. & Hirose K. (1982) *Modelling the dynamic characteristics of voice fundamental frequency with application to analysis abd synthesis of intonation.* Proceedings of 13th International Congress of Linguistics, pp.57-50

Fujisaki H. (1983) *Dynamic characteristic of voice fundamental frequency in speech and singing*. In: MacNeilage [editor], The Production, of Speech, Springer: New York, pp. 39-55

Golub G. & Kahan W. (1965). *Calculating the singular values and pseudo-inverse of a matrix*. SIAM Numerical Analysis, B 2 (2), 205-224

Grabe E. (1998) *Comparative intonational phonology: English and German*. MPI Series of Psycholinguistics 7, Wageningen: Ponsen & Looijen

Grice M. & Savino M. (1995) *Intonation and communicative function in a regional variety of Italian*. Phonus 1:19-32

Grice M., Reyelt M., Benzmueller R. & Mayer J., Batliner A. (1996) *Consistency in transcription and labeling of German intonation with GToBI.* Proceedings of 4th ICSLP, Philadelphia, pp.1716-1719

Grice M., Baumann S. & Benzmueller R. (2005) *German intonation in autosegmental-metrical phonology*. In S.A. Jun [ed.], Prosodic Typology - The Phonology of Intonation and Phrasing, Oxford: OUP, pp. 55-83

Gussenhoven C. (1983). *Testing the reality of focus domains.* Language and Speech 26:61-80.

Gussenhoven C. (1984) *On the grammar and semantics of sentence-accents*. Dordrecht: Foris.

Gussenhoven C. & Rietveld T. (1998) *Fundamental frequency declination in Dutch: testing three hypotheses.* J. Phon., 16:355-369

Gussenhoven C. (2002). *Intonation and interpretation: Phonetics and Phonology*. Proceedings of Prosody 2002, Aix-en-Provence, ProSig and Universite de Provence

Gussenhoven C. (2005). *Transcription of Dutch Intonation*. In Sun-Ah Jun [ed.ś Prosodic Typology: The Phonology of Intonation and Phrasing. *Oxford:* Oxford University Press.118-145

Gussenhoven C. (2006) *The Phonology of Intonation*. In Paul de Lacy [ed.] The Cambridge Handbook of Phonology. Cambridge University Press

Halliday M.A.K. (1967) *Intonation and grammar in British English*. The Hague: Mouton

Halliday M.A.K. (1970) *Language structure and language function*. In: John Lyons [ed.] New Horizons in Linguistics, Harmondsworth, England: Penguin, pp. 140-164.

Hałupka A. (2004) *Intonation modeling for speech synthesis application.* Master Thesis, Adam Mickiewicz University Poznań

Harris M. O., Umeda N. & Bourne J. (1981) *Boundary perception in fluent speech*. J.Acoust.Soc.Am. 8(4):1139-1145

Hermes D.J. & van Gestel J.C. (1991) *The frequency scale of speech intonation*. J.Acoust.Soc.Am. 90:97–102

Hess W. (1983) *Pitch determination of speech signals*. Springer-Verlag: Heidelberg

Hirschberg J. (1993) *Pitch accent in context: predicting intonational prominence from text*. Aritificial Intelligence 63(1-2):305-340

Hirst D.J. & Espesser R. (1993) *Automatic modelling of fundamental frequency using a quadratic spline function.* Travaux de l'Institut de Phonétique d'Aix 15, 71-85.

Hirst D.J. & Di Cristo A. (1998) *Intonation Systems: A Survey of. Twenty Languages*. Cambridge University Press, Cambridge

Hirst, D.J., Di Cristo, A. & Espesser, R. (2000) *Levels of representation and levels of analysis for intonation*. In M. Horne [ed] Prosody : Theory and Experiment. Kluwer: Dordrecht

Hirst D.J. (2001). *Automatic analysis of prosody for multilingual speech corpora.* In E.Keller, G.Bailly, J.Terken & M.Huckvale [eds.] Improvements in Speech Synthesis, Wiley.

Hirst, D.J. (2004) *The phonology and phonetics of Speech Prosody: Between Acoustics and Interpretation*. Proceedings of Prosody 2004, Nara, Japan

Horne, M., Strangert, E. & Heldner, M. (1995) *Prosodic boundary strength in Swedish: final lengthening and silent interval duration.* Proceedings of Int. Cong. Phon. Sci. 95, Stockholm, 1:170-173

Iida A., Higuchi F., Campbell N. & Yasumura M. (2004) *A Corpus-based Speech Synthesis System with Emotion*. Speech Communication, 40(1-2):161 – 187

Jassem W. (1961) *Akcent języka polskiego*. Prace Językoznawcze 31, Komitet Językoznawstwa PAN, Kraków, 1961

Jassem W. (1984) *Automatic segmentation of the Speech signal into phone-length elements.* Proceedings of 10th ICPhS, Fortis Publications, Dordrecht, pp.318-321

Jassem W. & Demenko G. (1999) *Modelling Intonational Phrase Structure with Artificial Neural Networks*. Proceedings of Eurospeech 1999, pp.711-714

Jilka M. (1996) *Regelbasierte F0 Generierung der Intonationsmuster des Amerikanischen Englisch*. Magisterarbeit. Universität Stuttgart

Jilka M., Möhler G. & Dogil G. (1999) *Rules for the Generation of ToBI-based American English Intonation*. Speech Communication 28:83 - 108.

Karpiński M. & Kleśta J. (2000) *The Project of an intonational database for the Polish language*. Proceedings of Prosody 2000: Speech Recognition and Synthesis, Kraków, Poland

Karpiński M. (2006) *Struktura i intonacja polskiego dialogu zadaniowego*. Adam Mickiewicz University Press, Poznań

Karpiński M., Kleśta J., Baranowska E. & Francuzik K. (2003) *Interphrase pause realization rules for the purpose of high quality Polish speech synthesis*. Proceedings of Speech Analysis, Synthesis and Recognition in Technology. Szczyrk, Poland, pp.85-89

Kießling, R. Kompe, A. Batliner, H. Niemann, E. Nöth (1996) *Classification of Boundaries and Accents in Spontaneous Speech*, Verbmobil project annual report.

Klatt D.H. (1980) *Software for a cascade/parallel formant synthesizer.* JASA 67(3):971-995.

Klessa K. & Klessa W. (2006) *Annotation editor*. Program do anotacji segmentalnej i suprasegmentalnej korpusów mowy.

Kohler K. (1987) Categorical pitch perception. Proceedings of 11th ICPhS, 5:331-333

Kohler K. (1991). A model of german inotnation. In: K. Kohler [ed] Studies in German Intonation. Universitat Kiel.

Kohler K. (1995) *The Kiel Intonation Model (KIM), its Implementation in TTS Synthesis and its Application to the Study of Spontaneous Speech*. Document downloadable from: www.ipds.uni-**kiel**.de/kjk/forschung/**kim**.de.html

Kohler K. (2005) *Timing and communicative functions of pitch contours*. Phonetica 62(2-4):88-105

Kröger B. J. (1998) *Ein phonetisches Modell der Sprachproduktion*. Niemeyer: Tübingen

Ladd D.R. & Campbell N. (1996) *Theories of prosodic structure: evidence from syllable duration*. Proceedings 12th ICPhS, Aix-en-Provence, Universite de Provence

Ladd D.R. (1996) *Intonational Phonology.* Cambridge University Press, Cambridge

Ladd D.R. (1998) *Segmental anchoring of pitch movements: autosegmental phonology or speech production?* In: Quene, H. & van Heuven, V. [eds], Speech and Language Studies for Sieb G. Noteboom, pp. 123-131. Utrecht, LOT.

Ladd D. R., Faulkner D., Faulkner H. & Schepman A. (1999) *Constant "segmental anchoring" of F0 movements under changes in speech rate.* J. Acoust. Soc. Am. 106: 1543-1554.

Ladd D. R., Mennen I. & Schepmann A. (2000) *Phonological conditioning of peak alignment of rising pitch accents in Dutch*. J. Acoust. Soc. Am. 107: 2685-2696.

Lea W.A. (1979) *Prosodic aids to speech recognition*. In: W.A. Lea [ed] Trends in speech recognition, Prentice-Hall, Englewood Cliffs NJ, pp.166-205

Liang H., Qian Y.& Soong F.K. (2007) *An HMM-Based Bilingual (Mandarin-English) TTS*. Proceedings of SSW6

Lieberman M. & Pierrehumbert J.(1984) *Intonational invariance under changes in pitch range and length*. In: Aronoff & Oehrle [editors], Language, Speech and Structure, M.I.T Press, Cambridge, pp.157-233

Liberman, M. (1975) *The Intonation System of English*. PhD dissertation, MIT. [IULC edition, 1978]

Loh W.Y, & Shih Y.S. (1997) *Split selection methods for classification trees*. Statistica Sinica, 7:815-840

Matuszkina O. (1976) *Wpływ artykulacji spółgłoskowej na przebieg częstotliwości podstawowej w sygnale mowy polskiej*. Prace IPPT vol. 37

Mayer J. (1995) *Transcription of German intonation: the Stuttgart System*. Technical report, IMS University of Stuttgart

Mennen I., Schaeffler F. & Doherty G. (2007) *Pitching it differently: A comparison of the pitch ranges of German and English speakers*. Proceedings of ICPhS 2007'

Mertens P. & d'Alessandro Ch. (1995) *Pitch Contour Stylization Using A Tonal Perception Model*. Proceedings of 13th ICPhS, 4:228-231

Mertens P. (2004) *The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model*. Proceedings of Speech Prosody 2004, Nara, Japan

Mishra T., van Santen J. & Klabbers E. (2006) *Decomposition of pitch curves in the general superpositional model*. Proceedings of Speech Prosody 2006, Dresden, Germany

Mixdorff H.& Fujisaki H. (1997) *Automated Quantitative Analysis of F0 Contours of Utterances from a German ToBI-Labeled Speech Database*. Proceedings of Eurospeech 1997', 1: 187-190

Mixdorff H.& Fujisaki H. (1998) *The Influence of Syllable Structure on the Timing of Intonational Events in German*. Proceedings of the ICSLP '98, Sydney, Australia.

Mixdorff H. (1998): Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F0-Contours. PdD thesis submitted to TU Dresden.

Mixdorff H.& Fujisaki H. (2000) *A quantitative description of German prosody offering symbolic labels as a by-product.* Proceedings of the ICSLP 2000, 2:98-101

Mixdorrf H. (2001) *MFGI, a Linguistically Motivated Quantitative Model of German Prosody*. In Keller, Bailly, Monaghan, Terken & Huckvale [editors], Improvements in Speech Synthesis, Wiley Publishers, pp.134-143

Mixdorff H. (2002a) *An Integrated Approach to Modeling German Prosody*. Habilitation thesis submitted to TU Dresden. Vol. 25, Studientexte zur Sprachkommunikation, w.e.b Universitätsverlag, Dresden

Mixdorff H. (2002b) *Speech technology, ToBI and making sense of prosody*. Proceedings of Prosody 2002, Aix-en-Provence, ProSig and Universite de Provence

Mixdorff H. & Jokisch O. (2003) *Evaluating the Quality of an Integrated Model of German Prosody.* International Journal of Speech Technology 6(1):45-55

Möbius B., Paetzold M. & Hess W. (1993) *Analysis and synthesis of German f0 contours by means of the Fujisaki model*. Speech Communication 13:53-61

Möbius B. (1995) *Components of a quantitative model of German intonation*. Proceedings of the 13th International Congress of Phonetic Sciences, 2:108-115

Möbius B., Möhler G., Schweitzer A., Batliner A. & Nöth E. (2000) *Prosodic models and speech synthesis: towards the common ground*. Proceedings of Prosody 2000: Speech Recognition and Synthesis (Kraków, Poland), 155-160.

Möbius B. & van Santen J. (2000) *A quantitative model of f0 generation and alignment.* In: A. Botinis [editor], Intonation - Analysis, Modeling and Technology, Kluwer: Dordrecht, pp. 269-288

Möhler G. (1998) *Describing intonation with a parametric model*. Proceedings of ICSLP 98', Sydney, Australia

Möhler G. & Conkie A. (1998) *Parametric Modeling of Intonation Using Vector Quantization*. Prooceedings of 3rd International Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 331-314

Möhler G. (1999) *Comparing two different principles of parametric F0 modeling*. Proceedings of the Joint ASA/DAGA Meeting, Berlin, Germany

Möhler G. (2001) *Improvements of the PaIntE model for F0 parametrization*. Technical report. IMS, University of Stuttgart, September 2001

Moulines, E. & Charpentier, F (1990) *Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones*. Speech Communication, (9):453–467

Nakatani C., Hirschberg J. & Grosz B. (1995) *Discourse structure in spoken language: Studies on speech corpora*. Working Notes of AAAI-95 Spring Symposiom on Empirical Methods in Discourse Interpretation

Narayanan S. (2004) *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall

Nooteboom, S.G. & Kruyt J.G. (1987) *Accents, Focus Distribution, and the Perceived Distribution of Given and New Information*. J. Acoust. Soc. Am 82(5):1512-1524

O'Connor J.D. & Arnold G.F. (1973) *Intonation of Colloquial English*. Longman: London

Oliver D. (1998) *Polish Text to Speech Synthesis*. Master Thesis, Edinburgh University

Oliver D. (2005) *Deriving pitch accent classes using automatic f0 stylization and unsupervised clustering techniques*. Proceedings of 2nd Baltic Conference on Human Language Technologies, Tallinn, Estonia, pp. 161-166

Oliver D. & Clark R. (2005) *Modeling pitch accent types for Polish speech synthesis*. Proceedings of Interspeech 2005, 1965-1968

Ortega-Llebaria M., Prieto P.& del Mar Varnell M. (2007) *Perceptual Evidence for Direct Acoustic Correlates of Stress in Spanish*. Proceedings of ICPhS 2007'

Ross K. & Ostendorf M. (1996) *Prediction of abstract prosodic labels for speech synthesis.* Computer Speech and Language (10):155-185

Ostendorf M., Price P.J. & Shattuck-Hufnagel S. (1995) *The Boston University Radio News Corpus*. Technical report ECS-95-001, Boston University

Palmer H. (1922) *English Intonation, with syntactic exercises*. Cambridge: Hefffer

Patterson D. & Ladd (1999) *Pitch Range Modeling:. Linguistic dimensions of variation.* International. Congress of Phonetic Sc., San Francisco

Patterson D. (1996) *Artificial Neural networks*. Singapore: Prentice Hall

Pellom B., Ward W. & Pradhan S. (2000) *The CU communicator: an architecture for dialogue systems*. In ICSLP-2000, (2):723-726

Pierrehumbert J. (1980) *The Phonology and Phonetics of English Intonation*. PhD dissertation, MIT. [IULC edition, 1987]

Pierrehumbert J. (1983) *Automatic recognition of intonation patterns*. ACL Proceedings of 21st Annual Meeting, pp.85-90

Pitrelli J.F., Beckman M.E. & Hirschberg J. (1994) *Evaluation of prosody transcription labeling reliability in the ToBI framework*. Proceedings of ICSLP 1994', pp.123-126

Portele T. & Heuft B. (1997) *Towards a prominence-based synthesis system*. Speech Communication, (21):61-72

Prieto P., van Santen J., and Hirschberg J. (1995) *Tonal alignment patterns in Spanish*. J.Phon. vol.23

Rapp S. (1996) *Automatic labeling of German prosody*. Proceedings of ICSLP 96'

Reichel U. (2007) *Data-Driven Extraction of Intonation*. Proceedings of SSW6

Reyelt M. (1996) *Consistency of Prosodic Transcriptions Labelling Experiments with Trained and Untrained Transcribers*. Proceedings of the 13th ICPhS, (4): 212-215

Rosenberg A. (2005) *Automatic Prosody Labeling*. Final Project Report for EE6820-Spring '05. Columbia University

Ross K. & Ostendorf M. (1996) *Prediction of abstract prosodic labels for speech synthesis*. Computer Speech Lang (10):155-185

Schmerling (1976). *Aspects of English Sentence Stress*. University of Texas Press, Texas

Silverman K. & Pierrehumbert J. (1990) *The Timing of Prenuclear High Accents in English*. Papers in Laboratory Phonology I, Cambridge University Press

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P.J., Pierrehumbert J. & Hirschberg J. (1992) *ToBI: A standard for labeling English prosody*. Proceedings of ICSLP 92', pp. 867-870

Sluijter A.M.C. & van Heuven V.J. (1996) *Acoustic correlates of linguistic stress and accent in Dutch and American English*. Proceedings of In ICSLP 96'

Sproat R. (1997) *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic: New York

Sridhar R., Bangalore V.K. & Narayanan S.S. (2007) *Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework*. IEEE transactions on audio, apeech and language processing, 16(4):797-811

Statistica 6.0, StatSoft, Inc. (2001). Statistica for Windows [Computer program manual]. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK, 74104-4442, (918) 749-1119, fax: (918) 749-2217, e-mail: info@statsoft.com, WEB: http://www.statsoft.com.

Steffen-Batóg M. & Katulska K. (1984) Individual differences *in the perception of main stress group boundaries in Polish*. Lingua Posnaniensis (27):107-115

Steffen-Batóg M. (1966) *Versuch einer Strukturellen Analyse der polnischen Aussagemelodie. Zeitschrift fuer Phonetik, Sprachwissenschaft und Kommunikationsforschung.* (19):397-440

Steffen-Batóg M. (1973) *The effect of consonant articulation and intonation of fundamental frequency in consonants*. Speech Analysis and Synthesis, 3:121-134

Steffen-Batóg M. (1996) *Struktura przebiegu melodii polskiego języka ogólnego*. Adam Mickiewicz University Press, Poznań

Streefkerk B. M. (1997) *Acoustical correlates of prominence: A design for research*. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 21:131-142

Strom V. (1995) *Detection of accents, phrase boundaries and sentence modality in German with prosodic features.* Proceedings of Eurospeech 95', Madrid, Spain, pp. 2039-2041

Syrdal A. & McGorg J. (2000) *Inter-Transcriber Reliability of ToBI Prosodic Labeling*. Proceedings of ICSLP 2000, 3:235-238

Syrdal A.K., Möhler G., Dusterhoff K., Conkie A. & Black A.W. (1998) *Three Methods of Intonation Modeling*. Proceedings of SSW3, pp.305-310

Syrdal K., Hirschberg J., McGory J. & Beckman M. (2001) *Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody*. Speech Communication, vol.33(1), pp. 135-151(17)

Szymański M. & Grocholewski S. (2005) *Transcription-based automatic segmentation of speech. Archives of Control Sciences*, 15:465-472

Śledziński D. (2007) *Analiza cech akustycznych sylab języka polskiego na potrzeby technologii mowy*. PhD thesis, Adam Mickiewicz University, Poznań

Tadeusiewicz R. (1993) *Sieci neuronowe*. Akademicka Oficyna Wydawnicza, Warszawa

Tadeusiewicz R. (1998) *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami.* Akademicka Oficyna Wydawnicza, Warszawa

Tamburini F.& Caini C. (2005) *An automatic system for detecting prosodic prominence in American English continuous speech.* International Journal of Speech Technology, 8:33-44

Tamburini F. (2005) *Automatic Prominence Identification and Prosodic Typology.* Proceedings of InterSpeech 2005, pp. 1813-1816

Tamburini F. (2006). *Reliable Prominence Identification in English Spontaneous Speech*. Proceedings of Speech Prosody 2006, Dresden, PS1-9-19

Taylor P. (1992) *A phonetic model of English intonation*. PhD thesis, University of Edinburgh

Taylor P. (1993) *Synthesizing intonation using the RFC model*. Proceedings of ESCA Workshop on Prosody

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

Taylor P. (1998) *The Tilt intonation model*. Proceedings of ICSLP 98'

Taylor P. (2000) *Analysis and synthesis of intonation using the tilt model*. J. Acoust. Soc. Am 107(3):1697-1714

Terken J. (1991) *Fundamental frequency and perceived prominence of accented syllables*. J. Acoust. Soc. Am. 89:1768-1776

t'Hart J., Collier R. & Cohen A. (1990) *A Perceptual Study of Intonation.* Cambridge University Press, Cambridge

Tokuda K., Zen H. & Black A. (2002) *An HMM-based Approach to English Speech Synthesis.* Proceedings of Autumn Meeting of the Acoustical Society of Japan, 3-10-15, Sep.

2002.Trim J. (1959) Major and minor tone groups in English. Le Maitre Phonetique 112, pp.26-29

Uhmann S. (1991) *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*. Tuebingen: Niemeyer

van der Hulst H.G. (1999) *Word accent*. In: H. van der Hulst [ed.] Word prosodic systems in the languages of Europe. Mouton de Gruyter, Berlin & New York, 3-116.

van Santen J. & Möbius B. (1997) *Modeling pitch accent curves*. Proceedings of ESCA Workshop Intonation: Theory, Models and Applications, pp.321-324

van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (1998) *Progress in Speech Synthesis*. New York: Springer-Verlag

van Santen J., Kain A., Klabbers E. & Mishra T (2005) *Synthesis of prosody using multi-level unit sequences. Speech Communication* 46:365-375

Wagner A. (2005) *A phonological model of intonation and intonation transcription system ToBI for Polish - a preliminary study.* Speech and Language Technology, 8:137-162

Wagner A. (2006) *A comprehensive model of intonation for application in speech technology.* Proceedings of 8th International PhD Workshop OWD 2006, 22:91-96

Wahlster W. (1993) *VERBMOBIL: Translation of Spontaneous Face-to-Face Dialogs*. Proceedings of 3rd *Eurospeech*, pp. 29-38

Wells J.C. (1997) *SAMPA computer readable phonetic alphabet*. In: Gibbon, D., Moore, R. & Winski, R. [editors] *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Wightman C.W. (1992) *Automatic Detection of Prosody for Speech Recognition and Parsing*. PhD thesis, Boston University

Wightman C.W., Shattuck-Hufnagel S., Ostendorf M. & Price P. (1992) *Segmental durations in the vicinity of prosodic phrase boundaries*. J. Acoust. Soc. Am., March 1992

Wightman C.W. & Ostendorf M. (1994) *Automatic Labeling of Prosodic Patterns*. IEEE Transactions on Speech and Audio Processing. October,1994.

Wightman C.W., Syrdal A., Stemmer G., Conkie A. & Beutnagel M. (2000) *Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Synthesis*. In Proceedings of Interspeech 2000

Wightman C.W. & Ostendorf M. (2002) *Automatic Recognition of Prosodic Features*. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, San Francisco, May, 1992.

Wypych M. (2005) *An automatic intonation recognizer for the Polish language based on machine learning and expert knowledge*. Proceedings of Interspeech 2005, Lisboa, Portugal

Wypych M. (2006) *Automatic Pitch Stylization Enhanced with Top-Down Processing*. Proceedings of Speech Prosody 2006, Dresden, Germany

Xu Y. (1998) *Consistency of tone-syllable alignment across different syllable structures and speaking rates*. Phonetica, 55:179-203

Yamagishi J., Kobayashi T., Renals S., King S., Zen H., Toda T. & Tokuda K. (2007) *Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV*. Proceedings of SSW6

Yoon T.J., Cole J. & Hasegawa-Johnson M. (2007) *On the edge: Acoustic cues to layered prosodic domains.* Proceedings of ICPhS 2007'

Yoon T.J., Heejin K. & Chavarría S. (2004) *Local Acoustic Cues Distinguishing Two Levels of prosodic Phrasing: Speech Corpus Evidence*. Lab. Phon. 9, University of Illinois at Urbana-Champaign

Appendix A: List of figures

Appendix B: List of tables

Appendix C: Published work

Demenko G. & Wagner A. (2005) *Analysis of accented syllables in different prosodic contexts for use in unit selection speech synthesis*. Proceedings of SASR, Cracow, Poland, September 19-23, 2005

Jassem K. & Wagner A. (2005) *A Conceptual Ontology for Machine Translation from/into Polish*. Proceedings of 2nd Language and Technology Conference, Poznań, Poland, April 21-23, 2005, pp. 437-440.

Demenko G. & Wagner A. (2006) *The stylization of intonation contours*. Proceedings of Speech Prosody 2006, Dresden, Germany, May 2-6, 2006

Wagner A. (2006) *A comprehensive model of intonation for application in speech synthesis.* Proceedings of VIII International PhD Workshop OWD 2006, vol.2, pp.91-96

Demenko G., Grocholewski S., Wagner A., Szymański M. (2006) *Prosody annotation for corpus based speech synthesis.* Proceedings of 11th Australasian International Conference on Speech, Science and Technology, Auckland, New Zealand, December 06-08, 2006

Demenko G. & Wagner A. (2006) *Prosody annotation for unit selection TTS synthesis*. Archives of Acoustics, vol. 32 (1), pp. 25-40

Jassem K. & Wagner A. (2006) *Semantic disambiguation in a MT system based on a bilingual dictionary.* Archives of Acoustics vol. 32 (1), pp. 75-88

Jassem K., Graliński F., Wagner A. & Wypych M. (2006) *Text Normalization as a Special Case of Machine Translation*. Proceedings of the International Multiconference on Computer Science and Information Technology, Volume 1, XXII Autumn Meeting of Polish Information Processing Society, Wisła 2006. (Best paper award)

Graliński F., Jassem K., Wagner A. & Wypych M. (2006) *Linguistic aspects of text normalization in a Polish text-to-speech system*. Systems Science, vol.32(4) pp.7-15

Wagner A. (2007) *Analysis of peak contours in different segmental and suprasegmental contexts.* Proceedings of 19th International Congress on Acoustics, Madrid, Spain, September 2-7, 2007

Demenko G., Wagner A., Jilka M. & Möbius B. (2007) *Comparative investigation of Peak Alignment of Peak Contours in Polish and German Unit Selection Corpora*. Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, Germany, August 22-24, 2007

Demenko G., Grocholewski S., Klessa K., Ogórkiewicz J., Wagner A., Śledziński D., Lange M. & Cylwik N. (2008) *Jurisdic - Polish speech database for taking dictation of legal texts.* To appear in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008

Wagner A. (2008) *Automatic labeling of prosody*. To appear in: Proceedings of ISCA tutorial and research workshop on experimental linguistics, Athens, Greece, August 25-27, 2008