Prosody Model and its Application to Czech TTS System

Jan Romportl, Jindřich Matoušek, Daniel Tihelka

University of West Bohemia in Pilsen, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
rompi@students.zcu.cz, jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz

ABSTRACT

The first part of this paper[1] proposes a formal theoretical framework for prosody description. This framework is based on empirically acquired axioms (following the linguistic structuralism) and in terms of the mathematical set theory it presents prosody as a relation between abstract sentence underlaying structures and intonation (together with timing). On the basis of this framework one can utilise various system description and analysis methods as well as pattern processing techniques to model the aforementioned relation. The second part of this text introduces the application of this framework to the Czech text-to-speech system ARTIC. It uses rules to place abstract intonation schemes (melodemes and cadences) depending on the position of an intonation centre of an utterance.

1. INTRODUCTION

Probably all text-to-speech (TTS) concerned papers agree that naturalness and also intelligibility of synthetic speech strongly depends on its prosodic quality. However, if one asks what such "prosodic quality" means one usually gets an answer vaguely summarising the goal of all prosody research to make TTS sound "human-like", no matter methods involved and long-term perspective offered. Indeed, this is under certain circumstances true but we feel that more comprehensive insight to this problematics is needed if the aforementioned dream of all TTS designers should be fulfilled. In this paper we present results of the initial stage of our research on prosody which tries to employ some results achieved by the functional approach of Prague Linguistic School. Couple of our results were utilised and applied to the state-of-art Czech TTS system being developed at Cybernetics department of the University of West Bohemia in Pilsen.

2. PROSODY PROPERTIES

Although "prosody" is generally known as a sort of synonym for suprasegmental features of human speech, we will try to give more formal conception. First of all we must cope with some empirical observations as well as set up what we want to achieve. The latter is explicated by an assumption, that "acceptably natural prosody" means such suprasegmental properties of synthetic speech that would be produced by a human speaker in a clear and intelligible utterance without excessive emotional concern. The former can be briefly summarised by following:

Axiom 1 (based on [1]):

I. Every continuous speech is divisible into smaller units (we will call them "phonemic clauses").
II. These units have their own specific intonation (e.g. melody and intensity, or contour of fundamental frequency and volume) and timing. The word "intonation" is used for the attribute constituted by melody and intensity.
III. These units can be separated by pauses.

Axiom 2:

"Prosodic quality" of every utterance can be fully described by intonation and timing.

Axiom 3 (for discussion see [1], [2], [3]):

Intonation and timing are "functionally involved"; they are constituted by elements where some of them have linguistic function(s) meanwhile some of them do not. Most relevant functions are delimitative and semantical.

The semantical function is (at least) of two kinds: it helps a listener create a notion of an utterance's *meaning* (as a linguistic concept) and participates in creating an utterance's (ontological) *content* (or factual knowledge) which can poorly be derived from the text itself.

Axiom 4 (see [1]):

The principle of tolerance and relevance. Each element of a certain functional layer can realise itself freely within boundaries given by a system.

We do not have enough space to discuss the above axioms and we are aware they might be somehow modified when facing future research. Further in the text we will use this notation: $\langle x_1, x_2, \ldots x_n \rangle$ is an ordered n-tuple of objects $x_1 \ldots x_n$ (in this order); relation R is a set of all 2-tuples $\langle y, x \rangle$ such that objects y and x (in this order) are in a relation R (e.g. yRx); for a relation R symbol *dom*(R) is a set of all x such that $\langle y, x \rangle \in R$; for a relation R symbol *rng*(R) is a set of all y such that $\langle y, x \rangle \in R$; for a set A symbol *pot*(A) is a set of all

1 This research is supported by the Grant Agency of Czech Republic no. 102/02/0124 and the Ministry of Education of Czech Republic, project no. MSM235200004.

subsets of A; for sets A and B operator A | B produces a set of <y, x> such that $\langle y,x\rangle \in A|B \Leftrightarrow \langle y,x\rangle \in A \wedge y \in B$ (notation is based on [10] and slightly modified).

Now we can propose the following definition:

Definition 1:

Be relation P', $<IT_{SJ}, <TR_S, MR_S, A>> \in P'$, called *prosody*. TR_S is the tectogrammatical representation of a sentence S, MR_S is the morphonological representation of a sentence S, A stands for stochastic attributes (perhaps properties that cannot be modeled otherwise), IT_{SJ} is the set whose elements adequately describe intonation and timing of sentence S realised as an uttered sentence token J.

Tectogrammatical representation is such a (non-linear) representation of a sentence which describes both its meaning and its syntax in terms of so-called *tectogrammatical layer* of language description. In detail it is elaborated and described in [4], [5]. It can be also called a *semantical structure*. Morphonological form is understood as a form which represents a sentence as it appears in its very surface structure (more specially for our purposes we can call this a graphemical form). "Adequate description of intonation and timing" is a very vague term but this way we leave the question of suitability of various techniques of intonation and timing description open.

The "real" nature of the aforementioned relation is somehow *infinite* (note the original meaning of this word is closely related to *indeterminate*) thus it would be of great benefit to utilise some mathematical theory for indeterminacy description (see very promising [10]). So far we must settle for the following (maybe contra-intuitive) assumption which helps us establish some sort of "discourse universe":

Axiom 5:

For each 3-tuple $<TR_S, MR_S, A>$ all possible uttered sentence tokens have been brought to existence.

Theorem 1:

$\forall S \in dom(P') \exists IT_{SI}, IT_{SJ} \in rng(P'):$

$\langle S, IT_{SI}\rangle \in P' \wedge \langle S, IT_{SJ}\rangle \in P' \Rightarrow IT_{SI} \neq IT_{SJ}$

Proof is a direct consequence of axioms 4 and 5. In short it formalises the fact that one sentence can be uttered in more ways (concerning intonation and timing).

Definition 2:

Acceptably natural prosody is the relation P = P' | Q where

$Q \in pot(rng(P'))$ such that

$\forall S \in dom(P') \forall q \in rng(P'):$

$q \in Q \Leftrightarrow q = \underset{q_i:\langle S,q_i\rangle \in P'}{argmin}\ J(S, q_i)$

where $J(S, q_i)$ is a criterial function such that (in case of suitable indexing) $J(S, q_1) < J(S, q_2) < ... < J(S, q_n)$.

Theorem 2:

P is a function.

Proof results from the definition 2 and from the requirements posed on the criterial function $J(S, q_i)$ because for each $S \in dom(P)$ exactly one IT_S is given.

The responsibility for constituting functional (not in terms of linguistics but in terms of mathematics) dependency between prosody and text is thus put on the criterial function that can be represented for example by some subjective perceptional tests. Basically its goal is to choose one realization of intonation and timing that is best in terms of given criteria. Technically for the sake of artificial prosody generation (in TTS systems for instance) we may outline the criterial function so as to select such realizations which are as close to specific real data as possible.

Through the above approach we postulated and formalised functional dependency between intonation (plus timing) and abstract linguistic representation of a sentence which can be derived from graphemical (i.e. generally morphonological) representation of this sentence and its context. The reason we considered the tectogrammatical level of representation as underlying instead of some "more surface" level is straightforward: the tectogrammatical representation (TR) is able to describe contextual information with regard to the relevance of *topic-focus articulation* (TFA) which reflects communicative function of a sentence (this can deal for instance with communicational aspects of word ordering which has a significant relevance in Slavic languages). For details see [4], [5], [6], [8]. We are convinced that ignoring all these aspects of text could bring some short-term advantages but seems to be shortsighted when facing the long-term goal of cognitive sciences.

Our framework allows us to understand prosody in terms of the system theory and thus model it using methods based on system description and analysis (in practical applications a lot of work can be done by pattern processing techniques). For the importance of language structures formalising see for example [7].

3. APPLICATION FOR CZECH TTS

Concerning the above mentioned we distinguish two levels of prosody description (see [1], [2], [3], following the Prague linguistic structuralism): functionally motivated and acoustically motivated (though this term might not be best fitting). We are far from thinking some great progress can be achieved without co-operation of these two constituents so we adopted and slightly changed terminology as it is presented in [3].

Continuous text - in written form - is segmented into sentences which we understand to be particular utterances (to make the problem easier) - in spoken form. Each utterance is divided into major and (optionally) minor phonemic clauses (also called prosodic phrases) which are then rhythmically segmented into phonemic words (a phonemic word is one or more words subordinated to one word-stress). We can actually say it

is almost a rule in Czech to place a word stress in the beginning of a phonemic word and for the sake of TTS we can postulate it (in this case we must cope with sometimes occurring "pre-phonemic words"). An example:

'V posledních **'dnech** /$_1$ **'naší 'do**volené /$_2$ jsem **'ko**nečně **'po**chopil //$_2$ že **'ne**snáším **'ces**tování..
In the last days /$_1$ of our holiday /$_2$ I finally realised //$_2$ that I hate traveling.

Major phonemic clauses are divided by /$_2$, minor phonemic clauses by /$_1$ (these are quite optional and strongly depends on speech rate) and // signs pauses. Bold characters show stressed syllables and ' stands for phonemic word beginnings. So far we have implemented rather simple rule-based method of phonemic words and word stresses placement but it has proved to be efficient enough because of relative simplicity of this area in Czech language (word stress has only a delimitative function and does not change the meaning anyway). Phonemic clauses detection is being developed and has not been implemented sufficiently yet.

Further in this text we will speak mostly about intonation (and more specifically about melody) since for Czech it is the most important attribute of prosody (see [1], [3]) and the only one implemented in the TTS system ARTIC so far (see [9]). There is also no tectogrammatical parser available for us which means that at this stage of research and development we cannot fulfil the goals appointed by the theoretical part of this paper. However, they (together with the above formalization) should be kept in mind as well as the fact we want to develop a prosody model for TTS, not describe and formalise language system.

Prosody of an utterance is described (as it was already mentioned) in two levels. The first level consists of so called *melodemes*. Melodeme is an abstract intonational pattern established in certain function within a language system. The second level is characterised purely by acoustic attributes irrespective to their function. If we stay under "protective wings" of *acceptably natural prosody* we can describe this level in terms of *cadences*. We can understand cadence to be a model generating elements of IT_{SJ} set, but - an important note - the input of such model must be a structure much more simple than <TR_S, MR_S, A>. In short - cadence describes "real" acoustic properties of intonation without regard to a meaning or content of an utterance.

So far we use simple rule-based model of cadences of this inventory (in terms of the F0 contour slope): flat, ascending, descending, ascending-descending, descending-ascending. Each cadence can be enhanced by the attribute "stressed" (e.g. it models F0 contour with a word stress) and is presented (except for the flat cadence) in three forms (simply distinguished by numbers 1, 2, 3) which vaguely express "how quick F0 moves up or down" (we will refer to it as the tendency of a cadence). Each cadence places a number of F0 values (in Hertz) at certain timestamps and all such values are then linearly interpolated producing the final F0 contour. These cadences are intended to underlay phonemic words. It means that a cadence is quite a simple unit covering only one phonemic word and thus concerning just morphonological form of a phonemic word (e.g. a subset of MR_S). For example the flat stressed cadence places three values: first at the beginning of the first phoneme of a phonemic word, then local maximum at the beginning of the nucleus of the stressed syllable, then local minimum at the end of the last phoneme of a phonemic word. The concrete numbers are calculated using referential F0 value (supplied by the superordinate layer - melodemes, as will be seen later) and coefficients specific to particular cadence (indeed the coefficients are dependent on a speaker and are set up experimentally).

Melodeme placement depends on the position of so called *intonation centre*. Normal position of an intonation centre in Czech is on the last phonemic word of an utterance. This position is automatically understood to designate emotionally neutral utterance. However, contextual factors of TFA (especially word order) may change the position and this can be expressed by TR_S (for details see [4], [6], [8]). Since we do not have a tectogrammatical parser so far (as it was already mentioned) we must actually omit the structure given by TR_S and we place intonation centre always on its "normal" position. There are also phonemic clause intonation centers and these are automatically placed on the last phonemic word of phonemic clauses.

Intonation centre - because of its function - is a border between two neighbouring melodemes (indeed it is just a conceptual point of view - the function and the intonation centre itself is actually realised by particular melodemes or cadences respectively). This means that an intonation centre starts a new melodeme of the type dependent on a sentence modality (declarative, interrogative, imperative, etc.) and other aspects (not realised in this phase yet). We currently use following melodemes (based on [2], [3]):

M0 - null melodeme
M1 - terminating descending melodeme
  M1-1 - neutral
M2 - terminating ascending melodeme
  M2-1 - neutral
M3 - nonterminating melodeme
  M3-1 - neutral
  M3-2 - neutral, pause preceding

Indeed we should distinguish more melodemes but those written above are quite sufficient for emotionally neutral utterance description (when almost omitting TR_S, and accounting only for a sentence modality as it is our case at this stage of research and prosody application). M0 is understood as a formal description of such a segment of a phonemic clause where actually no melodeme is realised. It ranges from the beginning of a phonemic clause to the phonemic word right before the intonation centre of the phonemic clause and it describes slightly descending tendency of phonemic clause melody. M1 is used for an utterance terminating in case the sentence is declarative, imperative or interrogative-

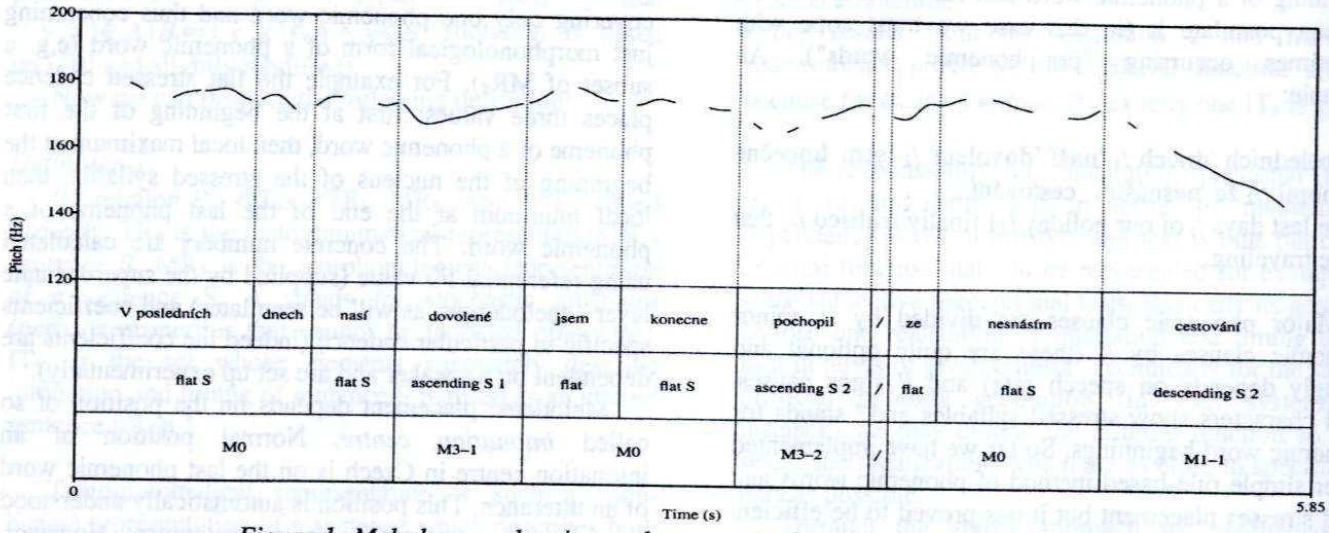

Figure 1: Melodeme and cadence placement in the resulting synthesised speech

supplementary (like English "wh-" questions). It starts at the intonation centre and reaches the end of the utterance. M2 is very similar to M1 with the difference that it is used to terminate interrogative-inquiring sentence (e.g. questions with a "yes/no" answer). M3 is used in those phonemic clauses which are not last in particular utterances.

Melodemes are chosen to fulfil particular function given by the communicative intention and cadences can be understood as tools intended for their concrete realization. Figure 1 transparently illustrates the usage of cadences and melodemes on the example of the above introduced sentence (without minor phonemic clauses).

4. CONCLUSION

The formalization of prosody in the part 2 should not be taken linguistically rigorous. It is certainly purpose build, however, it tries to establish solid ground on which one can base prosody models applied in many systems such as TTS. Moreover, it implies the dependency between the semantical structure and a context of a sentence which is necessary to take into account to build as much as possible human-like sounding TTS system.

The presented rule-based method of prosody description is a very simplified realization of this formalization. However, the results - as can be evaluated with Czech TTS ARTIC system - are quite satisfactory, taking the simplicity of so far used method into account. The intonation is quite far from being perfectly human-like but this can be much improved by the use of more sophisticated cadence models (for example stochastic or neural network based). Moreover, since the prosody description is separated into two autonomous parts (functional and non-functional), changes made to one part almost do not influence the other one and the formalization sets up a solid area for a further research.

The future work will focus on more elaborate cadence models (based on the stochastic system description) and will enhance timing properties, especially syllable duration. More thorough insight into the intensity attribute of intonation will be undertaken too (this means modelling intensity contour not only at stressed syllables).

5. REFERENCES

1. Daneš, F. "Intonace a věta ve spisovné češtině (Sentence Intonation in Present-Day Standard Czech)" , Nakladatelství Československé akademie věd, Prague, 1957
2. Romportl, M. "Základy fonetiky (Basics of Phonetics)", Prague, 1973
3. Palková, Z. "Fonetika a fonologie češtiny (Phonetics and Phonology of Czech)", Karolinum, Prague, 1994
4. Sgall, P. - Hajičová, E. - Panevová, J. "The Meaning of the Sentence in Its Semantic and Pragmatic Aspects", Reider, Dordrecht, 1986
5. Panevová, J. "Formy a funkce ve stavbě české věty (Forms and Function in Czech Syntax)", Academia, Prague, 1980
6. Sgall, P. - Hajičová, E. - Buráňová, E. "Aktuální členění věty v češtině (Topic and Focus in Czech)", Academia, Prague, 1980
7. Hajič, J. - Hajičová, E. - Rosen, A. "Formal Representation of Language Structures", In TELRI Newsletter No.3, pp. 12-19, June 1996
8. Hajičová, E. "The Prague Dependency Treebank: Crossing the Sentence Boundary", In Proceedings of the Second Workshop on Text, Speech, Dialogue, pp. 20-27, Marianske Lazne, Czech Republic, 1999
9. Matoušek, J. - Psutka, J. "ARTIC: A New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction", In The Proceedings of the 6th International Conference on Spoken Language Processing ICSLP2000, vol. IV. Beijing, China, 2000, pp. 612-615.
10. Vopěnka, P. "Úvod do matematiky v alternatívnej teórii množín (Introduction to Mathematics in Alternative Set Theory)", Alfa, Bratislava, 1989