Advanced Prosody Modelling*

Jan Romportl, Jindřich Matoušek, Daniel Tihelka

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
rompi@students.zcu.cz, jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz

Abstract. A formal prosody model is proposed together with its application in a text-to-speech system. The model is based on a generative grammar of abstract prosodic functionally involved units. This grammar creates for each sentence a structure of immediate prosodic constituents in the form of a tree. Each prosodic word of a sentence is assigned with a description vector by a description function and this vector is used by a realization function to create appropriate intonation for the prosodic word. Parameters of the model are automatically set up using real speech data from a prosody corpus, which is also described.

1 Introduction

Prosody is not only a very important element contributing on naturalness of synthetic speech but also almost indiscerptible constituent of a spoken message structure. As a consequence, modelling of prosody has been already for a significant period of time treated as apparently one of the crucial areas of text-to-speech system design.

Our prosody model presented in the following text is conceptually similar to the approach of concatenative synthesis: it concatenates elementary prosody units derived from real speech data contained in a specially designed and annotated prosody corpus. This approach can achieve significantly higher naturalness of resulting synthetic speech, similarly to the phenomenon when concatenative synthesis achieves better naturalness than formant synthesis.

The model is also underlied by a formal apparatus which leads to interesting results concerning a language system functioning.

2 Prosody Description Framework

In order to be able to adequately describe prosody functioning and its relation to text we propose following framework which can more formally describe systemic behaviour of prosody as a language phenomenon.

Prosody can be formally underlied by a generative grammar with special terminal and non-terminal symbols based on functionally relevant structures which

* This research was supported by the Grant Agency of Czech Republic No. 102/02/0124.

can be uncovered in speech material. The following theory (yet it should be mentioned that only a fragment of it can be presented here due to the space limitations) could thus be called as a kind of formal suprasegmental phonology. We distinguish following suprasegmental functionally relevant structures:

Prosodic sentence (PS)
Prosodic sentence is actually a prosodic manifestation of a sentence (e.g. an utterance) as a syntactically consistent unit, yet it can also be unfinished or grammatically incorrect.

Prosodic clause (PC)
Prosodic clause is such a linear unit of a prosodic sentence which is delimited by pauses.

Prosodic phrase (PP)
Prosodic phrase is such a segment of speech where a certain intonation scheme is realized continuously. A single prosodic clause often contains more prosodic phrases.

Prosodeme (P0),(Px)
Prosodeme is an abstract unit (sort of a "suprasegmental phoneme") established in a certain communication function within the language system. We have postulated that any single prosodic phrase consists of two prosodemes: so called "null prosodeme" and "functionally involved prosodeme" (where (Px) stands for a type of the prosodeme chosen from the table shown below), depending on the communication function the speaker intends the sentence to have. We distinguish the following prosodemes (for the Czech language; other languages may need some modifications):

- P0 – null prosodeme
- P1 – prosodeme terminating satisfactorily
  - P1-1 – no indication
  - P1-2 – indicating emphasis
  - P1-3 – indicating imperative
  - P1-4 – indicating interjection
  - P1-5 – indicating wish
  - P1-6 – specific
- P2 – prosodeme terminating unsatisfactorily
  - P2-1 – no indication
  - P2-2 – indicating emphasis
  - P2-3 – indicating "wh-" question
  - P2-4 – indicating emphasised "wh-" question
  - P2-5 – specific
- P3 – prosodeme nonterminating
  - P3-1 – no indication
  - P3-2 – indicating emphasis

P3-3 – specific

Prosodic word (PW)
Prosodic word (sometimes also called phonemic word) is a group of words subordinated to one word accent (stress).

Semantic accent (SA)
By this term we call such a prosodic word attribute, which indicates the word is emphasised (using acoustic means) by a speaker. The relevancy of "semantic accent" is discussed in [1].

In the following generative rule description we use two more terminal symbols ("$" and "#") which stand for pauses differing in their length. The rules should be understood this way: $(PC) \longrightarrow (PP)\{1+\}\ \#\{1\}$ means that the symbol (PC) (prosodic clause) generates one or more (PP) symbols (prosodic phrases) followed by one # symbol (pause).

$$(PS) \longrightarrow (PC)\{1+\}\ \$\{1\} \tag{1}$$

$$(PC) \longrightarrow (PP)\{1+\}\ \#\{1\} \tag{2}$$

$$(PP) \longrightarrow (P0)\{1\}\ (Px)\{1\} \tag{3}$$

$$(P0) \longrightarrow (PW)\{0+\} \tag{4}$$

$$(P0) \longrightarrow (SA)\{1\}\ (PW)\{1+\} \tag{5}$$

$$(Px) \longrightarrow (PW)\{1\} \tag{6}$$

$$(Px) \longrightarrow (SA)\{1\}\ (PW)\{2+\} \tag{7}$$

If we apply these rules on a sentence, they create a tree of immediate constituents consisting of the terminal and non-terminal symbols used. We define $V^T = \{(PW), (SA), \$, \#, \emptyset\}$ a set of all terminal symbols and $V^N = \{(PS), (PC), (PP), (P0), (Px)\}$ a set of all non-terminal symbols (note that (Px) is just an "abbreviation" for all symbols (P1-1), (P1-2), etc.). Indeed $V = V^N \cup V^T$ is a set of all symbols, e.g. the whole alphabet.

Note: the rule (5) is used for "wh-" questions (such as Czech "Kdy dnes večer přijedete?", in English "When will you come this evening?") where (SA) stands as an attribute of "kdy" ("when") which functionally underlies the intonational form with two "intonational centres" – one expressed by (SA) and the other one by a functionally involved prosodeme (P2-3) or (P2-4) at the end of the sentence (generating one or more last prosodic words).

Now we can define a *description function*

$$\mathfrak{D} : V \to \mathcal{D} \tag{8}$$

This description function "describes" quite uniquely in terms of prosodic units each symbol (node) of a certain prosodic tree. Detailed information about this

function and other formalisms concerning it (such as the structure of the class $\mathcal{D}$) are presented in a monograph [3].

For the sake of this text we settle for an easier explanation and simplifying representation, which is now (temporarily – until new algorithms are efficient enough) used in our computer realization of this model – each prosodic word of a sentence is described by a vector with the following values: number of prosodic clauses of the sentence, index of the prosodic clause the prosodic word appears in, prosodeme type the prosodic word appears in, prosodeme length (measured in prosodic words), index of the prosodic word in its prosodeme, the number of syllables of the prosodic word, the number of phones of the prosodic word, index of the stressed vowel in the prosodic word.

Once more mentioned, the formal representation of the description function is far more complex, but in our experiments we have realized the simplified description is often quite sufficient for practical purposes.

The prosodic word description is then used in a *realization function*

$$\mathfrak{R} : \mathcal{D} \rightarrow \mathcal{I} \times \mathcal{C} \tag{9}$$

where $\mathcal{I} = \{i_1, \ldots, i_l\}$ is a set of *initial conditions* and $\mathcal{C} = \{c_1, \ldots, c_m\}$ is a set of *cadences*. A cadence is a real intonational pattern which fits into a range of a single prosodic word and the set $\mathcal{C}$ can be also called a "cadence inventory". Initial conditions say where a cadence chosen for each prosodic word should start.

To be more concrete: our text-to-speech system works so far only with melody (e.g. fundamental frequency, $F0$) when using this prosody model. In such a case we have $i_j \in \mathbb{R}$ and $c_k \in \mathbb{R}^x$ (where the dimension x ranges from 10 to 20, optimal value seems to be 15), i_j represents an initial $F0$ value at the beginning of a prosodic word while the vector c_k describes the $F0$ contour of this prosodic word in terms of i_j multiples.

This all means that once we have a prosodic tree of a sentence, we can construct its intonation (and timing, if this is included in the cadence formalism) by the following operation:

$$\mathfrak{R}(\mathfrak{D}(w_1)) \circ \mathfrak{R}(\mathfrak{D}(w_2)) \circ \ldots \circ \mathfrak{R}(\mathfrak{D}(w_n)) \tag{10}$$

where $\circ$ is an operation of juxtaposition (simply placing one element next to each other) and $w_i \in V^T$ are prosodic words and pauses of a sentence with such a suitable indexing which reflects the (left to right) linear ordering of the symbols.

The crucial importance for the realization function has the following *principle of an exchange*:

$$\forall D_i, D_j \in \mathcal{D}, D_i \neq D_j : \mathfrak{R}(D_i) = \mathfrak{R}(D_j) \Leftrightarrow R(D_i, D_j) \tag{11}$$

$R(\cdot, \cdot)$ is a *relation of indistinguishableness*, as it is defined in *Alternative Set Theory* described in [2]. We cannot analyse and discuss this principle and the form of the relation R here any further due to space limitations but it is done so in [3].

In short: this principle allows us to substitute under some extent (e.g. as long as two different prosodic word descriptions are in the relation of indistinguishableness) one prosodic word description by another one while the prosodic representation remains untouched. The advantage of it will be shown in the next section of this text.

3 Prosodic Data Retrieval

Obviously all parameters of the aforementioned formal relations must be set up using real prosodic data. Thus we have chosen four most frequent speakers (in the radio part) from the *Czech TV & Radio Broadcast News Corpus* – which is almost 4,000 sentences – and their utterances were manually annotated using XML tags to represent occurrences of the abstract prosody units described above (e.g. semantic accents, prosodemes, prosodic words, phrases and clauses). The text was segmented into communicationally coherent parts (*turns* – each consisting of 2-5 sentences) which reflect also the aspect of *topic-focus articulation*. The new prosody corpus created this way is used also for speech recognition purposes and is described in [4].

This corpus is now used mostly as training data for designing a suitable text parser capable of parsing a text in terms of the prosodic structures described in the previous section. However, cadences (as concrete $F0$ patterns) are derived from different speech data: we use the same speech corpus which is used for speech unit (triphones) retrieval in the TTS system ARTIC – e.g. the same system the prosody model is used with, which brings great advantages when the prosody model is combined with "unit selection", also tested with this TTS system.

This speech data consist of 5,000 sentences uttered by a female speaker. Glottograph data, e.g. full $F0$ contours, are included for all utterances. These contours were segmented into parts extending over prosodic words and then represented as vectors of the dimension x (as it was introduced in the previous section). This way we acquired 55,655 detailed representations of $F0$ shapes of prosodic words.

The set $\mathcal{C} = \{c_1, \ldots, c_m\}$ (the cadence inventory) is created by an agglomerative clustering algorithm (with various parameters – depending on a type of an experiment) applied on the aforementioned $F0$ vectors. The elements of $\mathcal{C}$ (e.g. cadences) are constructed as either centroids of the clusters, or there is one (or more) vector chosen from each cluster as its representant (using diverse methods, for example "elimination of outliers" by Mahalanobis' distance). We experiment with various values of m (the number of cadences) ranging from 3 up to 200.

Good results are achieved for example for the number of clusters $m = 30$. In such a case the smallest cluster consists of 911 vectors ($F0$ patterns) and the largest of 3571. Figure (1) shows 30 cadences created from the clusters by choosing the vectors (one vector from one cluster) with the smallest distance from cluster centroids.

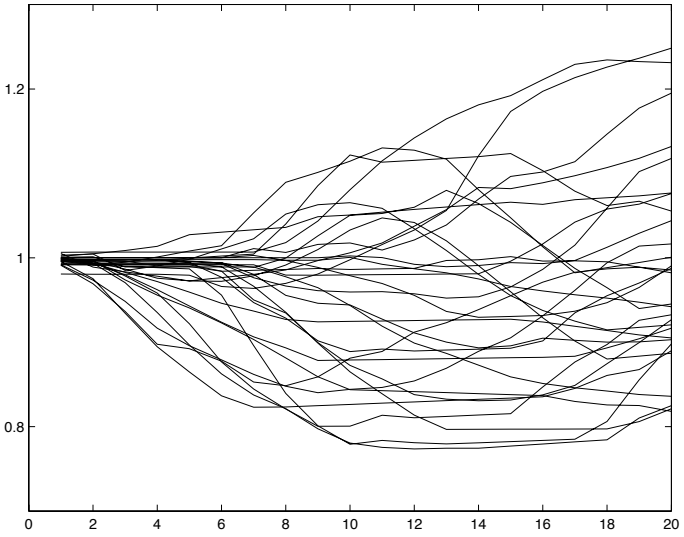

Fig. 1. Cadences – $F0$ patterns for prosodic words

The key procedure – implementation of the $\mathfrak{R}$ function on the basis of real data – is far from being trivial and so far it is in an early stage of the research. However, the first results with the synthesised speech using $F0$ generated this way are more than encouraging. The goal is to implement the function according to the theoretical framework presented in [3] as well as to find some possible modifications of this framework based on results of experiments with real data.

If we determine $\mathfrak{D}(w_i)$ for each prosodic word w_i which occurred in the corpus we cannot create the function $\mathfrak{R}$ since there are some (actually the majority of) prosodic words which occur more times with the same description (e.g. the vector $\mathfrak{D}(w_i)$ is same for each occurrence of that word) but realized with different $F0$ patterns (cadences). It means one out of more cadences must be chosen as a functional value of $\mathfrak{R}$ for a particular description. In the present version of the prosody module for a TTS synthesis the most frequent cadence for a particular description is chosen. The same approach is used also for choosing appropriate initial conditions – e.g. the value of $F0$ at the beginning of a prosodic word.

However, there is still another obstacle – how to define the $\mathfrak{R}$ function for such a $D_i \in \mathcal{D}$ (e.g. prosodic word description) which does not occur in a corpus. A solution for this can give the principle (11): the unknown value $\mathfrak{R}(D_i)$ is set to be equal to a known value $\mathfrak{R}(D_j)$, where D_j occurs in the corpus and is in the relation of indistinguishableness with D_i, e.g. $R(D_i, D_j)$.

Obviously it still is not easy to find out when two descriptions are in the relation of indistinguishableness. One of the goals of our further research is to derive this criterion formally. So far we make use of experimentally gained knowledge which shows that it is often sufficient to cause slight perturbations to the values of D_i (the least important ones, such as exact length of a prosodic word in phones, index of a prosodic phrase, etc.) until we get such a description D_j which occurs in the corpus. After this procedure D_i and D_j are still most likely to be in the relation R.

4 Conclusion

As it was already mentioned, this approach to prosody modelling has already been successfully tested with the text-to-speech system ARTIC. The results show significantly better performance and speech naturalness than the rule-based prosody model used so far.

The intonation naturalness was evaluated by a MOS test with the scale 1 (best) – 5 (worst). During this test participants were listening to and evaluating various synthesised sentences with intonation generated by different models – monotonous (no intonation), rule-based, data-driven (presented in this article) and real (acquired by electroglottograf measuring of real speech). The test results are shown in Table 1. Further details and tests are presented in [3].

model	monotonous	rule-based	data-driven	real intonation
average evaluation	4,41	3,39	2,48	1,90
st. deviation	1,13	1,04	1,12	0,99

Table 1. The results of various prosody models evaluation according to MOS tests

Yet there is still much work to do, particularly in the improvements of the way the cadence inventory is created and indistinguishable descriptions are recognised. Moreover, further research is concerned with a suitable and reliable prosodic parser producing prosodic trees of input sentences.

Our prosody model is also based on processing of real prosodic data and this means the research also tries to answer the question whether it is possible to create a prosody model with data of one speaker and then use it with a voice of a different speaker (the current results show it is – under some constraints – possible, even if the prosody model is set up using female speech data and the synthesised voice uses male speech data). Another topic is the influence of the structure and extent of the prosodic corpus used to create the model. Concerning the extent it has shown that one can use significantly smaller corpus than we had expected.

It should be mentioned too, that the mathematical formalisms we use to describe prosody functioning not only can answer some questions about rela-

tions lying beyond the humans sight, but they more importantly help ask new questions which lead to new interesting experiments never thought of before.

References

1. Palková, Z.: Fonetika a fonologie češtiny (Phonetics and Phonology of Czech). Karolinum, Prague (1994).
2. Vopěnka, P.: Úvod do matematiky v alternatívnej teórii množín (Introduction to Mathematics in Alternative Set Theory). Alfa, Bratislava (1989).
3. Romportl, J.: Generování prozodie z textu pro účely syntézy řeči (Generating Prosody from Text for Speech Synthesis Purposes). Západočeská univerzita v Plzni (University of West Bohemia in Pilsen), Pilsen (2004).
4. Kolář, J., Romportl, J., Psutka, J.: Czech Speech and Prosody Database Both for ASR and TTS Purposes. Proceedings of Eurospeech 2003, vol. 2. Geneve (2003) 1577-1580.
5. Matoušek, J., Psutka, J.: ARTIC: a New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction. Proceedings of ICSLP 2000, vol. IV. Beijing (2000) 612–615.