Unit Selection and its Relation to Symbolic Prosody: a New Approach

Daniel Tihelka and Jindřich Matoušek

Department of Cybernetics
University of West Bohemia, Czech Republic
dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

Abstract

Aiming at the improvement of the quality of synthetic speech generated by our native TTS ARTIC, we adopted the unit selection method. Our unit selection module is driven by prosody described solely by high-level symbolic features which are linked to the prosody of synthesized phrases through the phenomena of prosodic synonymy and homonymy. It was confirmed that such an approach not only generates speech with high naturalness but also keeps the richness of prosody. Our first version of this approach significantly increased the quality of the output speech, which was assessed by listeners as very close to natural.

The concept of prosodic synonymy and homonymy is, therefore, further extended and formally described in this paper, and its importance to the unit selection treatment is demonstrated. In addition, the difference of this concept from the concepts most frequently used is shown. Moreover, the first experiment following the formal definition of the problem presented in this paper has been carried out, proving that the whole concept is feasible.

1. Introduction

The unit selection speech synthesis has become a frequently used technique in concatenative speech synthesis. The treatment of this technique in our native TTS ARTIC [1] uses target specification described exclusively on a high symbolic level, i.e. the target only defines the communication function required to be expressed, but does not define any explicit low-level prosodic requirements expressing that function [2]. It was shown that the low-level requirements are not necessary (or even desirable) for the selection, as the perceived naturalness of speech generated by this approach was assessed very high (as *almost natural*); moreover, the style of the emerged prosody was mostly perceived as the same as the style recorded by the speaker in the corpus. In the above-mentioned article we introduced the phenomena of prosodic synonymy and homonymy. As will be shown further, we regard these as essential for the treatment of unit selection in agreement with human perception and we expect them to lead to the decrease in the size of speech corpora required for unit selection approach while maintaining (or even increasing) the quality of generated speech. Our concept is, therefore, further extended and formally described in this paper, and it is linked together with the framework of *prosodic grammar* [3], also developed at our department.

The paper is organized as follows. Section 2 shortly summarizes the prosodic grammar, and formally describes our concept of unit selection as well as the phenomena of prosodic synonymy and homonymy. It also points out the main difference between the described concept and the concepts most frequently used. Section 3 then presents an experimental realization of the proposed concept, aiming to verify the correctness of the proposed formal definition. In Section 4 the results of listening tests comparing the experiment and our original selection module are shown, while Section 5 summarizes the paper and outlines our future work and expectations.

2. Formal Problem Description

The somewhat simplified attempt to drive unit selection by prosody described on a high level was published in [5], and the idea was (independently) generalized in [6]. However, our aim is not to use a set of features designed “from experience”, but to approach the unit selection technique following the vague nature of human perception and to give the whole concept a formal framework in which the synonymy and homonymy are essentially important. Our rationales are inspired by the alternative set theory [7] incorporating vagueness into its basis.

2.1. Prosodic Grammar

Each TTS system must derive prosody of synthetic speech from the text representation of an utterance at its input. The majority of the approaches used (e.g. ToBI, Tilt, Fujisaki model, etc.) treat prosody as the composition of duration, pitch (F_0) and intensity, and aim at generating the courses of those characteristics across the utterance directly from the text. This is, however, not very suitable for our purposes, as this treatment reduces the rich nature of prosody to the three contours only, suppressing microprosody, differences in expression, or other phenomena, even those yet unknown.

Therefore, we proposed the *prosodic phrase grammar* [3, 4] which is related to linguistic knowledge and which allows a detailed description of the prosody, while not limiting the richness of prosody in any way. The grammar builds a hierarchical tree structure above the synthesized phrase, where the relations in the tree describe the prosodic relations in the underlying phrase. To be more specific, the grammar consists of the following alphabet:

prosodic sentence (PS) prosodic manifestation of a syntactically consistent unit

prosodic clause (PC) linear unit in speech delimited by pauses

prosodic phrase (PP) segment of speech containing a certain continuous intonation scheme

prosodeme (P0), (Px) abstract unit describing communication function – we defined null prosodeme and functionally involved prosodeme specifying intended communication

This research is supported by the Grant Agency of the Czech Republic, project no. GACR 102/06/P205, and by the Academy of Sciences of the Czech Republic, project no. 1ET101470416.

function (to put it simply, distinguishing a declarative phrase from a question, etc.)

prosodic word (PW) group of words belonging to one stress, often considered as a basic rhythmic unit

semantic accent (SA) the prosodic word expressing some emphasis

The simple illustration of the phrase described by the prosodic grammar is shown in Figure 1.

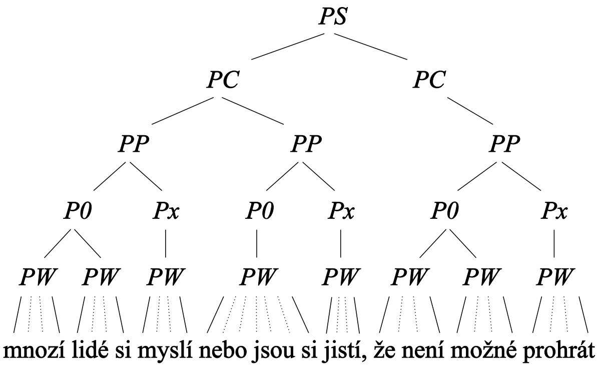
PS
PC
PC
PP
PP
PP
P0
Px
P0
Px
P0
Px
PW
PW
PW
PW
PW
PW
PW
PW
mnozí lidé si myslí nebo jsou si jistí, že není možné prohrát

Figure 1: *The illustration of the tree build using the prosodic grammar for the czech phrase "Many people think or are convinced that it is impossible to lose".*

To establish the relation of the prosodic grammar with speech units being concatenated in order to create the synthetic speech, let, similarly to [4], P^l_{NS} be a set of all nodes of *prosodic structure* (i.e. a particular tree produced by the prosodic grammar), and l be the index in the hierarchy of the structure ($1 = PS, \dots, 5 = PW$). Since the level of prosodic words is not enough for the purposes of speech units concatenation – not only prosodic, but also phonetic information must be kept in the sequence of units – let us extend P_{NS} by the set of all phonetic symbols of the utterance underlying the prosodic structure, staying on level $l = 6$.

Each unit in the synthesized phrase at the input of the synthesizer and each candidate in a speech corpus can, in general, be described by symbolic target features defined as:

$$t_l = \left(\mathcal{F}_l(P^l_{NS}, P^{l-1}_{NS}), t_{l-1}\right), \qquad l = 2, \dots, 6 \qquad (1)$$

$$t_1 = \emptyset \qquad (2)$$

where $\mathcal{F}$ is a function defining the relation between levels l and $l-1$ in the prosodic structure (e.g. the relation of units to the prosodic word) which can differ for individual levels. The recursion allows us to fully describe the whole hierarchy.

Although prosody is a suprasegmental feature not appearing on individual phone-like units, the proposed treatment allows us to link each unit with the expression of a certain communication function given by the prosody of juxtaposed units and described by the prosodic structure. Moreover, prosody expressed by units thus described will be preserved in all its richness, as there is no reduction or simplification of prosody modelling at all. Each candidate in the corpus will, consequently, be described by one (or more in the case of homonymy, see further) corresponding target feature t, one t will also be assigned to each unit in a synthesized phrase – this t will further be called *target specification*.

2.2. The Concept of Prosodic Synonymy and Homonymy

In [2] we introduced the phenomena of *prosodic synonymy* and *prosodic homonymy* from the point of view of unit selection, linking individual candidates with the communication function expressed by them (a very simplified illustration of the phenomena is shown in Figure 2). We will extend and more formally describe the phenomena here, in order to establish some formal apparatus describing the unit selection approach from our point of view. Obviously, finding a real relation defining these phenomena in agreement with human perception is not trivial, and will be the objective of our further intensive research. Let us also note that the t is the whole target specification as introduced in the previous section, not only one of the target features as was stated in the last paper.

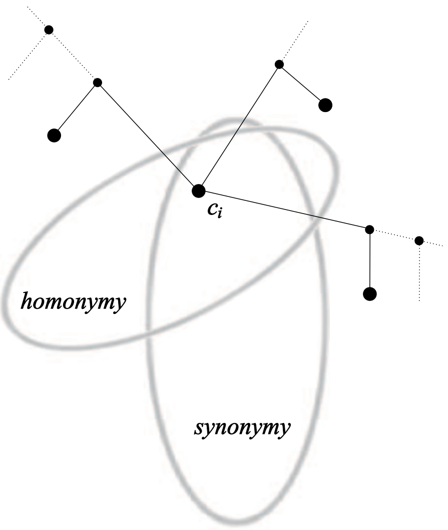

Figure 2: *The simplified illustration of the correspondence between the sets of synonymous and homonymous candidates. The relations to the different trees of prosodic grammar are also denoted.*

Let $C = \{c_1, c_2, \dots, c_N\}$ be the set containing all candidates of a unit – this set is *sharply determined* by the particular candidates in it. Be further $t(m)$ a particular target specification (one symbol of all possible ones, let us suppose a finite set) given by function in Equation (1). Then the relation:

$$S : t(m) \rightarrow \mathcal{C} \subseteq C \qquad \forall m = 1, \dots, M \qquad (3)$$

assigns to $t(m)$ a *not sharply determined* sub-set $\mathcal{C}$ of candidates which express the communication function given by $t(m)$; in other words, each candidate from $\mathcal{C}$ can be used for the rendition of prosody, expressing the function $t(m)$, no matter how different the prosody-related features of the candidates are. The sub-set $\mathcal{C}$ must not be sharply determined in order to be in agreement with the vague nature of human perception (and production), as well as in agreement with the blurred relation between the intended communication function and the form of prosody realizing that function (e.g. different courses of F_0 can express the same communication function, as it was indirectly confirmed in [2] or [6]). Moreover, for the suprasegmental level it is often contrast which is more important than the absolute values.

In general, there is no need for target cost to be 0 for all synonymic candidates (although it will usually be met when a phrase occurring in the corpus is synthesized). Therefore, let target cost be more generally defined as a similarity function $\mathcal{G}$ between target features $t(c)$ of a candidate c and target specification t required for the candidate:

$$TC(c,t) = \mathcal{G}(t(c),t) \tag{4}$$

and for the correct function of the selection algorithm, the following equation must be met:

$$\forall c \in \mathcal{C} \text{ and } \forall d \in (C - \mathcal{C}) : TC(c,t) < TC(d,t) \tag{5}$$

Closely related to the synonymy is the phenomenon of *prosodic homonymy*. Let $T = \{t(1), t(2), \ldots, t(M)\}$ be *sharply determined* set containing all possible target features, then the relation:

$$H : c_n \rightarrow \mathcal{T} \subseteq T \qquad \forall n = 1, \ldots, N \tag{6}$$

assigns to a candidate c_n a *not sharply determined* sub-set $\mathcal{T}$ of target features describing different communication functions – the candidate c_n can be used for the rendition of all communication functions in $\mathcal{T}$. The sub-set $\mathcal{T}$ is not sharply determined for the same reasons as the synonymy (and because the phenomena are related).

We have not yet dealt with homonymy in depth, and so we did not formally define any explicit requirements for the target cost from the point of view of this phenomenon. There is only the requirement given by Equation (5), specifying the relation of the homonymous candidate to all its synonymous partners (see Figure 2).

2.3. Why a Vagueness is Profitable

In most of the unit selection TTS, discrete features are used in the target cost (we also did this in [2]). The sub-cost assigned to a particular feature then acquires value 0 if the value of the feature matches the required target, or 1 (or some fixed value) if the feature differs from the target. Let us outline the shortcomings of such treatment.

When distinct 1/0 values are used, the set C can be split into two *sharply determined* sets $\mathcal{C}_1$ (matching candidates) and $\mathcal{C}_2$ (not matching candidates), such that $\mathcal{C}_1 \cup \mathcal{C}_2 = C$ and $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. The same result is also obtained for each combination of features. If there are more values possible in the target cost, the set C is split into more sub-sets, but the principle remains the same. Let us note that the sharp split would also be obtained if we supposed 0 target cost in Equation (5) for all synonymous candidates.

However, this treatment tries to model phenomena naturally blurred by vagueness by sharp distinction. Obviously, there must exist a set of cases (again, not sharply determined) where a distinct criterion errs[1] – i.e. a distinct criterion determines a sharp set of candidates where all are supposed to express an equal communication function (e.g. accent), but when a particular candidate is used in synthetic speech, humans do not perceive that function (no accent at all), or even worse, they perceive another not required function. Although prosody is a suprasegmental feature and cannot be expressed by one individual unit, the misinterpretation by a distinct criterion occurs in each set of candidates. The concatenation cost can then prefer a sequence of candidates from the misinterpreted parts, resulting in undesirable expression.

3. Experiment

To put the formal definition into the practice and to verify that the whole concept is correct, we need a sample experiment which would follow the formal definition. However, as the realization of the whole concept requires further intensive research (which we are planning to focus on), we adopted some simplifications in this experiment. We restrict ourselves only to the level of t_6 (relation of speech units with prosodic words), and, contrary to our aim when the phenomena are supposed to be obtained by examining the relations among data in corpus, the synonymy relation was ad hoc explicitly defined by a windowing function, as described further.

We covered each prosodic word, often declared to be a basic rhythmic unit, by the three von Hann windows (also known as Hanning):

$$w_n = 0.5 \left(1 - \cos(\frac{2\pi n}{N-1}) \right) \tag{7}$$

These windows can be considered as the "suitability" of the candidates for the particular position in the prosodic word. Each candidate is described by three real numbers obtained from the value of the corresponding window, as illustrated in Figure 3. From the point of view of the "classic" approach, the synonymy thus defined covers features like position in word, stress (in Czech fixed, but can be extended to non-fixed stress handling), and partly also word length. However, as there are no explicit threshold values fixing the suitability to a certain number of fixed levels (as in the classic approach), there cannot be found any sharp determination of the set C.

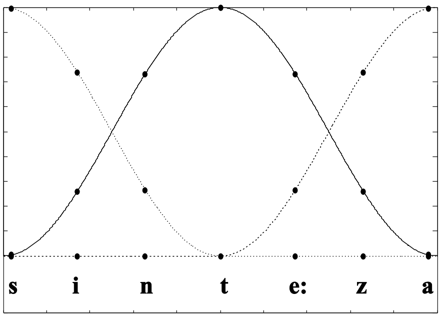

Figure 3: *The illustration of the correspondence of windowing functions to a prosodic word "synthesis". Individual windows are distinguished by line style, points correspond to the values describing the candidates.*

The target cost is defined for this experiment simply as the sum of the differences between the expected (given by target) and available (given by the examined candidate) values. In this way, the requirement defined by Equation (5) is easily met.

In order to obtain comparable results, we used the same corpus as in [2]. It consists of 5,000 sentences (about 13 hours of speech) recorded in news-like style by a female voice talent with some radio broadcasting experience, so the style of prosody was kept broadly consistent during the recording. We also used diphones and the same features for concatenation cost (F_0 and MFCC, all *z*-score normalized) as explained in [2].

[1] As a part of the proposed concept, we are planning to establish a more formal proof.

4. Results

The speech generated using the concept discussed was compared to the speech from our first unit selection described in [2]. Several news reports from the Internet were synthesized by both approaches, and 10 shorter phrases (from 2 to 4 secs) were randomly selected for CCR listening tests (shorter phrases are easier to remember and compare for listeners). Two versions A and B of the same phrase were played to 14 listeners, who were asked to compare the quality of those versions on a 3-point scale – A better (1), about the same (0), A worse (-1). As the order of the versions was altered in the tests, the assessments were normalized in order for A to correspond to the described concept and for B to be the original version. Detailed results are shown in Figure 4, where the mean scores for each phrase across all listeners, as well as the overall average score, are depicted together with the corresponding standard deviations.

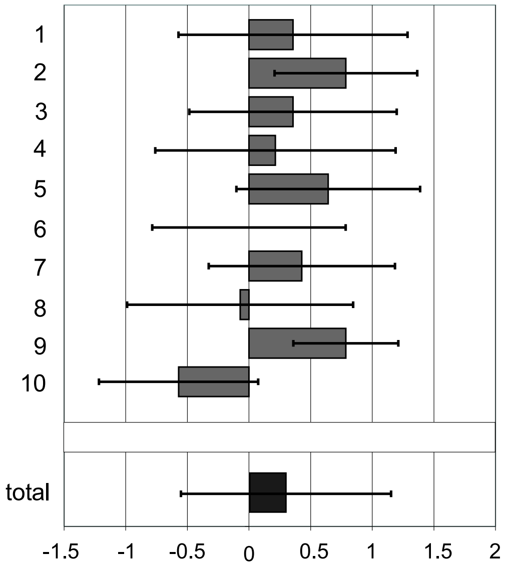

Figure 4: *The results of the listening tests evaluating the overall quality of speech generated according to the described concept (score going to 1), and the speech generated by our previous version of unit selection (score going to -1). The standard deviation is shown for each score.*

It can be seen that the proposed experiment does not perform worse than the original approach using distinct features (in fact, the experiment has a slightly higher score, but due to high standard deviations it cannot be considered as statistically significant). Moreover, if we take into consideration the fact that the experiment does not exploit the whole power of prosodic structure, the results are very encouraging, and the whole concept appears valid.

5. Conclusion

We introduced the formal description of the unit selection approach driven by symbolic prosody described by means of our phrase grammar. The proposed concept directly incorporates vagueness, which is a phenomenon obviously present in human perception as well as between prosody and the corresponding communication function. Therefore, we formally defined the phenomena of synonymy and homonymy which are the link between the low level, represented by individual candidates, and the communication function which they can express, while the concept of not sharply determined sets was used. The finding of those phenomena and the use of the mathematical framework defined in [7] can reveal other interesting relations among speech units as well as their relation to human perception. We expect that all of this will result in the decrease in the size of speech corpora required for unit selection approach, while maintaining (or even increasing) the quality of generated speech. Naturally, as the concept is very recent, some modification or elaboration may be required in the course of our further research.

We have also reached a state where the quality of speech does not differ very much among the versions, but there are still unnatural artefacts perceived, and the use of standard listening tests is not very profitable (which is also confirmed by the results of the listening tests in Section 4). Therefore, we are planning to utilize a special methodology of speech assessment introduced in [8]. The tendency of grouping similar types of error artefacts allows much finer evaluation and comparison of the tested versions.

6. References

[1] Matoušek, J., Romportl, J., Tihelka, D., and Tychtl, Z. "Recent Improvements on ARTIC: Czech Text-to-Speech System", Proc. of ICSLP. vol. III, pp. 1933–1936. Jeju, Korea, 2004.

[2] Tihelka, D. "Symbolic Prosody Driven Unit Selection for Highly Natural Synthetic Speech.", Proc. of Interspeech 2005, pp. 2525–2528. Lisbon, Portugal, September 2005.

[3] Romportl, J., Matoušek, J., Tihelka, D. "Advanced Prosody Modelling", Proc. of TSD, pp. 441–447, Brno 2004.

[4] Romportl, J., "Structural Data-Driven Prosody Model for TTS Synthesis", Accepted to Speech Prosody. Dresden, Germny 2005.

[5] Chu, M., Peng, H., Yang, H., Chang, E. "Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer", Proc. of ICASSP, vol. 2, pp. 785–788, Salt Lake City 2001.

[6] Clark, R.A.J., Richmond, K., King, S. "Festival 2 – Build Your Own General Purpose Unit Selection Speech Synthesizer", Proc. of ISCA Speech Synthesis Workshop, pp. 173–178, Pittsburgh 2004.

[7] Vopěnka, P. "Úvod do matematiky v alternatívnej teórii množín (Introduction to Mathematics in Alternative Set Theory)". Alfa, Bratislava 1989.

[8] Mayo, C., Clark, R. A. J., King, S. "A Multidimensional Scaling of Listener Responses to Synthetic Speech", Proc. of Interspeech 2005, pp. 1725–1728. Lisbon, Portugal, 2005.