Symbolic Prosody Driven Unit Selection for Highly Natural Synthetic Speech

Daniel Tihelka

Department of Cybernetics
University of West Bohemia in Pilsen, the Czech Republic
dtihelka@kky.zcu.cz

Abstract

In the effort to obtain synthetic speech with the quality near to natural, and potentially, to be able to build expressive synthesis, the unit selection approach has become very important. To increase the naturalness of our native TTS system ARTIC we employed the specific version of the approach. It is driven by the high–level symbolic prosody description, defined according to the phenomena of prosodic synonymy and homonymy. The concrete prosody of a synthesized phrase is not explicitly set here, but emerges on the basis of the target and concatenation costs. Although this is our first treatment requiring some simplification, and for the synonymy/homonymy phenomena only the basics are defined, the first results have already shown that there is a significant shift towards high quality. Listening tests comparing speech from single–instance version to selection–based version of ARTIC showed clear preference of the selection–based version. In addition, the level of naturalness was on average assessed as "close to natural".

1. Introduction

Since 1999 the TTS system ARTIC (ARtificial Talker In Czech) has been developed at our department. It is a multilingual, concatenative method–based synthesizer employing some of the latest speech technologies [1]. As the very first among all institutions dealing with speech synthesis in the Czech Republic, special balancing of sentences selected into a speech corpus, the automatic segmentation of the corpus, or pitch–marks detection based on the glottal signal have been employed in the synthesizer. Thanks to this, although ARTIC uses only one candidate for each unit (we will further call it *single–instance* version), the system achieves a high level of intelligibility and a fair degree of naturalness. However, the latter is an area where some limits still exist.

The recent research [2] helped us to reveal those factors which contribute to the deterioration of naturalness most significantly. We let listeners determine as many not–naturally perceived artefacts in synthetic speech as they were able to, and then we looked for factors which caused them. Not surprisingly, the result was that the most significant factor was the modification of the signal of candidates – 38% of artefacts. Much less significant were segmentation inaccuracies (19%) and the discontinuity of spectral characteristics at the point of concatenation (11%). Those are intuitively expected results, but the concrete "levels of significance" of particular naturalness deterioration factors have not been examined in such a way. Although the results are related to ARTIC TTS, we assume that they are applicable to any other synthesizer which uses the same (or similar) technologies.

This research has been supported by the Ministry of Education of the Czech Republic, project no. LC536

To reduce those problems and, naturally, to increase naturalness, it is an obvious choice to employ the *unit selection* approach [3] within ARTIC. The results of mentioned research were very useful for the determination of the concrete approach to the selection – we require to minimize the need of the signal processing of selected candidates. Results presented in this paper show us that the selection driven by symbolic prosody is an appropriate choice. Employing the unit selection approach for the Czech language is a unique act, it is the first achievement so far. However, for us it is also an essential step in the effort to reach the next level in speech synthesis research – expressive speech synthesizer.

The paper is organized as follows. Section 2 shortly summarizes the basics of unit selection. Section 3 then describes and discusses specificities related to symbolic prosody driven unit selection; the problem of prosodic synonymy and homonymy is also introduced here. Section 4 presents the listening tests procedure and summarizes the results of the tests. It also outlines the future work.

2. Unit selection in general

As a great deal has already been written about unit selection in various papers, we will only shortly summarize the basic ideas, which, however, will be useful for the establishing of the terminology used in this paper. In this approach, each candidate of each unit in the corpus[1] is generally described by the vector of P target features related to the contextual and prosodic properties of the candidate as well as by the vectors of Q concatenation features related to signal properties at the beginning and the end of the candidate. The selection algorithm, as was defined in [3] and as we have also used, is based on the minimizing of two costs. The first, *target cost*:

$$TC_i(c) = \sum_{p=1}^{P} tw_p \, \mathcal{F}_p(u_i(c), t_i) \qquad (1)$$

measures the difference between the features t_i of a unit i required in synthesized phrase and the features of the candidate c of unit u_i which can be used for the generation of the signal. tw_p is the weight which can adjust the significance of the feature p and function $\mathcal{F}_p$ computes feature–dependent difference. The second, *concatenation cost*:

$$CC_{i-1,i}(c,d) = \sum_{q=1}^{Q} cw_q \, \mathcal{G}_q(u_{i-1}(c), u_i(d)) \qquad (2)$$

[1]The candidate is a particular token of a unit which can be used for the creation of synthetic speech. Thus, units appear on the input of synthesizer, and their best candidates are searched in the corpus. Let us further consider all unit tokens in the corpus as candidates – we have not been concerned with the reduction of candidate number so far.

measures of the quality of the join of two juxtaposed candidates c and d for units u_{i-1} and u_i. Similarly, αw_q is a weight and $\mathcal{G}$ returns concatenation feature–dependent difference. For synthesized phrase with I units, the sequence of such candidates which minimizes the total cost:

$$C(t_{1,\ldots,I}, u_{1,\ldots,I}) = \sum_{i=1}^{I} TC_i + \sum_{i=2}^{I} CC_{i-1,i} \qquad (3)$$

is then concatenated in order to create output speech.

The basic and the most obvious requirement for the selection algorithm is that when a phrase which is presented in the corpus appears at the input of the selection algorithm, candidates from that phrase have to be selected – the phrase is hence simply played back, and so the best possible quality is obtained. From the target and concatenation costs point of view it means that both costs have to be 0 for all pairs target–candidate c and for each pair of consecutive candidates c and d from the phrase:

$$TC_i(c) = 0 \qquad \forall i = 1, \ldots, I \qquad (4)$$

$$CC_{i-1,i}(c, d) = 0 \qquad \forall i = 2, \ldots, I \qquad (5)$$

Generally, when a phrase not presented in the corpus is synthesized, those conditions can be applied to sub-phrases, words or clusters of candidates.

3. Symbolic prosody in unit selection

The idea of using symbolic prosody only to drive the selection (or more precisely the target cost) was firstly published in [4] for Mandarin – however, the authors used very simplified treatment. The independent generalization of the idea was then published in [5]. It was also mentioned in [6] that the use of only symbolic prosody description helped to increase the naturalness. In the next sections we will therefore show why it is advantageous to drive unit selection by symbolic prosody only. We will also outline the phenomena of prosodic synonymy and homonymy, whose understanding and incorporation in the selection is essential for high naturalness.

3.1. Why symbolic prosody only

It is clear that different phrases have different prosody renditions[2], based on their types, length, phrasing, structure, etc. However, it is also clear that even phrases with the same structure and spoken in the same style have different renditions. Therefore, in TTS systems the prosody is usually described at the high–level which is not dependent on the rendition. Such kind of description allows us to encourage the basic prosodic character of the phrase; e.g. ToBI or the phrase and accent component of Fujisaki model are among those used for English. For Czech we defined *prosodic grammar* [7] which is advantageous in its relation to linguistic knowledge; we are, moreover, convinced that the formalism can be used for other languages as well.

The synthesized phrase is firstly described by a higher–level description which is then used for the generation of low–level prosody, manifesting itself in the required prosodic contours, i.e. rendition of the synthetic phrase. However, together with some kind of symbolic description, this low–level prosody is also used in most approaches to drive the selection algorithm by means of target cost.

To show the disadvantage of such treatment, let us for now consider a prosody generator which is able to estimate perfectly

[2]Rendition here means the concrete shape of prosodic contours.

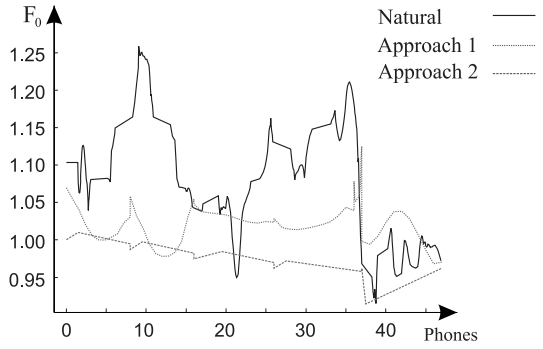

Figure 1: *The shape of F_0 (relative to mean) in natural phrase compared with contours estimated by two of our approaches for that phrase.*

natural low–level prosodic contours. When a phrase which is also present in the corpus appears at the input of unit selection, it can be supposed that each candidate from that phrase has the target cost 0 for all high–level features (see further). However, the rendition given by estimated contours can differ significantly from the rendition inherent to the phrase (see the Figure 1). As a result, the conditions in Equations (4) and (5) will not be reached, and candidates from different phrases (which have also 0 target cost for high–level features) can appear in the synthesized phrase. One can object that the low–level features can be used for the control of the global prosodic character of the phrase, but our results show that it becomes pointless.

3.2. The basis of prosodic synonymy/homonymy

Although the background of this paper and results presented here are connected with the synonymy and homonymy phenomena, the paper is not exclusively focused on it. Moreover, the phenomena require further intensive research, and thus only the general idea will be shortly outlined here.

Let us suppose that $C = \{c_1, c_2, \ldots, c_N\}$ is the set containing all candidates of a unit. Be further $P = \{p_1, p_2, \ldots, p_M\}$ the set of all hypothetic symbols (or features) of the higher–level prosody description, e.g. *stressed/unstresses*, *phrase type*, *prosodeme type*, but also *position in phrase* and others. Then the relation:

$$S : p_i \rightarrow \mathcal{C} \subseteq C \qquad \forall i = 1, \ldots, M \qquad (6)$$

assigns to one symbol from P such a sub-set $\mathcal{C}$ of candidates from C which may have different prosodic rendition, but are nevertheless equal from the meaning point of view (e.g. the declarative phrase cannot be changed into the interrogative). Such a relation shall be called *prosodic synonymy of candidates*; each of the candidates in the set $\mathcal{C}$ can be used for the rendition of prosody described by symbol p_i, and the meaning of synthesized phrase will be kept unchanged.

However, some, possibly rare, prosodic events corresponding to symbols from P cannot be presented in given corpus. Therefore we define the relation:

$$H : c_j \rightarrow \mathcal{P} \subseteq P \qquad \forall j = 1, \ldots, N \qquad (7)$$

assigning to one candidate the sub-set of symbols from P. This relation shall be called *prosodic homonymy of candidates* and it corresponds to the fact that particular prosodic rendition of the candidate can be used for the expression of different meanings.

Having a unit in a synthesized phrase represented by $\{p_1, p_2, \ldots, p_k\} \subseteq P$ prosodic description (built on the basis

of both synonymy and homonymy), the target cost is 0 for such a set of corresponding candidates $\mathcal{C}'$ which is given by:

$$\mathcal{C}' = \bigcap_{i=1}^{k} S(p_i) \tag{8}$$

where S is the relation given by Equation (6). Thus, it is necessary to choose such a set of features P to Equation (8) wont be empty for any combination.

The fact that each prosodic description symbol can be related to many different naturally–sounding renditions, which are, however, equal from the point of view of phrase meaning, allows us to completely exclude the low–level prosody from the target cost. It then gives higher freedom to the selection and increases the chance to found optimal, i.e. naturally sounding, candidate sequence. The target cost using only symbolic features will result in more (even overlapping) candidate sub-sequences, potentially with very different prosodic contours, but reaching Equation (4) (and Equation (5) is reached for the sequence as well).

However, even if each sub-sequence contains natural prosody, the mere concatenation of them does not ensure the natural prosody of whole phrase; moreover, as we mentioned, they can overlap and thus the mere concatenation is impossible. Therefore, the concatenation cost is now responsible for the finding of such an intersection of sub-contours which will contain smooth[3] prosody rendition. Thus, there is no further modification of speech signal required.

In this way, the particular prosodic rendition *emerges* on the basis of symbolic description, and the style of emerged prosody is very similar to the speaker, as will be seen in Section 4.4. It also elegantly bypasses the need to build a prosody generator, which is in itself quite a difficult task.

3.3. Cost features used

As we are only beginning to deal with unit selection driven by symbolic prosody, the problem of obtaining the symbolic description according to prosodic synonymy requires further intensive research. Therefore, we made a step aside towards some simplification in symbolic features in this very first experiment. The homonymy was not considered for the time being.

The set P contains here symbolic features like *stressed, unstressed unit* (we suppose synonymy in stress prosody rendition), *phrase declarative, interrogative* and *unterminated* related only to units in the last word, and *other* for all other units except the last word (we suppose synonymic rendition anywhere within the phrase except for the last word; candidates rendering the last word are synonymous according to the phrase type). *Onset, nucleus, coda* features were also used, since the syllable is often considered as a prosody expressing component, and thus the rendition can depend on the position in it. In addition, features related to left and right context phone were also used in order to ensure correct coarticulation. During the target cost computing, all possible candidates were considered for each unit (whole set C in Section 3.2). Each unit in the synthesized phrase was described by set $\{p_1, p_2, \ldots, p_k\} \subseteq P$, e.g. *unstressed*, *in the last word of declarative phrase*, *onset*, *phone [d] in left context*, etc. The function $\mathcal{F}_p, p = 1, \ldots, k$ from Equation (1) then returned 0 for those candidates reaching Equation (7), and 1 for all others.

As the concatenation cost is responsible for the resulting prosody rendition, the difference in F_0 at the boundaries of units was considered – the discontinuities in F_0 are the most perceived.

[3]Smoothness is sufficient, as the naturalness is ensured within sub-contours.

In addition MFCC coefficients were used for spectral smoothness, as they are widely used in unit selection based TTS systems. Those features were normalized for each phrase by means of z–score. Surprisingly, even if only those two features were responsible for the overall prosody of the phrase (neither duration nor intensity was considered), the results were assessed very highly by listeners (see Section 4.4).

The notable thing is that candidates are synonymous from from the point of view of their position in the phrase (no such feature was used, except the affiliation with the last word). The results shown in Section 4.4 are very encouraging even for the described simplifications; it gives us certainty that we are on the right way.

4. Experiments

In this section, we will describe our experiments with the first version of symbolic prosody driven unit selection, as well as the conditions under which the experiments were carried out.

4.1. Corpus used

As the recording of corpus specially designed for unit selection is quite expensive, we decided to use the already existing one. Although the corpus has been recorded for the single–instance version of ARTIC, it is suitable at least for the very first version of unit selection. The corpus consists of 5000 sentences (about 13 hours of speech) recorded by a female speaker in news–like style. It is important that the speaker was able to keep broadly consistent style of prosody during the recording. Speech generated on the basis of this corpus was also used in experiments about the deterioration of naturalness [2] mentioned in Section 1.

4.2. The choice of units

Although the single–instance version of ARTIC uses triphones (phone–sized units with the information about context in their names), we decided to use diphones, as the experiments in [2] showed the that 19% of problems were caused by segmentation inaccuracy. Shifting the boundary into the half of phone, the precision of the segmentation loses its importance. On the other hand, we lost some triphone advantages, and so we intend to subject the possibility of the use of triphones to further research, and to explore their advantages and possible disadvantages in unit selection algorithm.

4.3. Listening tests

To reflect the improvement (or possibly the deterioration) of ARTIC's employing symbolic prosody driven unit selection technique, we designed three types of listening tests, each consisting of 10 phrases. 14 listeners took part in the tests, most of them having no previous experience with speech synthesis.

The first test was focused on the measure of improvement obtained by using the selection technique within ARTIC against the single–instance version of ARTIC – although there are many researches into unit selection, no explicit comparison showing the amount of improvement against single–instance has been carried out. Therefore, this is quite a unique comparison which, moreover, uses the same speech corpus. CCR (Comparison Category Rating) test [8] was used in this case; two versions A and B of the same phrase, not presented in the corpus, were played to listeners, one from single–instance and one from unit selection. Listeners were asked to compare the quality of those versions on a 7–point scale:

A much better	...	3	A slightly worse	...	-1
A better	...	2	A worse	...	-2
A slightly better	...	1	A much worse	...	-3
about the same	...	0			

The second test was intended to test the naturalness of synthesized phrases (not from the corpus). For the evaluation, modified MOS (Mean Opinion Score) test [8] was used; the phrase synthesized using unit selection was played to listeners, who evaluated the naturalness on a modified 5–point scale:

Completely natural	...	5
Almost natural	...	4
Something in the middle	...	3
Rather artificial	...	2
Completely artificial	...	1

One phrase within this test was fake: the phrase from corpus was used in unit selection, and thus candidates from the phrase were selected (see Equations (4) and (5)). Listeners who rated the phrase worse than 4 were excluded from the results; it occurred in one case. Unfortunately, such a kind of test has never been carried out for single–instance ARTIC and thus we cannot compare the amount of naturalness increase; carrying out that test within the described tests was rejected, as it would be confusing for listeners.

The last test was used to prove our assumption that the prosody rendition which emerged in synthesized phrases imitates the style of the speaker who recorded the corpus. The phrase which is presented in the corpus was synthesized, but all candidates from that phrase were excluded from the selection. Listeners have at their disposal both versions of the phrase, the original one from the corpus and the synthesized one. The question was: does the synthesized version have the same kind of prosody as the original phrase, with *yes/no* possibilities. We know that this test was very difficult for lay listeners and the results do not have to be precise; on the other hand, they correlate with our assessment.

4.4. Results

Detailed results are shown in Figure 2, where the mean scores for each phrase across all listeners are depicted. Although the speech generated by single–instance version of ARTIC is highly intelligible, the quality of the speech increases rapidly when symbolic prosody driven selection is used. When compared to the single–instance version, the increase is almost 2.7 point on the testing scale. Moreover, the naturalness is only about 1 point distant from the listener's idea of naturalness. Listeners also perceived the emerged prosody in 77% similar to natural prosody. It should be noted that all listeners preferred speech generated using the selection.

4.5. Future work

As can be seen, the results of the listening tests are very encouraging. However, we are only at the beginning and much remains unresolved. Our main aim is, therefore, to make the phenomena of prosodic synonymy/homonymy more precise, and to define the features used for the target cost based on our prosodic grammar, while taking the phenomena into consideration.

Moreover, if the selection is not able to find candidates with smooth F_0 transition, further post–processing can be applied to smoothen the transition and thus to prevent unpleasant "jumps" in F_0. Even if it can possibly deteriorate the quality, we expect

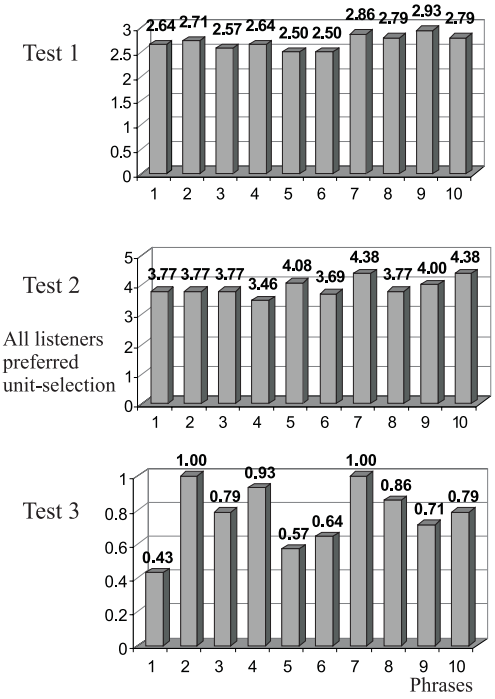

Figure 2: *The results of listening tests for all experiments.*

that this mechanism will be invoked only several times and the modification will not be large.

Special thanks are due to Jan Romportl for his valuable comments on the prosody–related topics.

5. 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., Matoušek, J. "The Analysis of Synthetic Speech Distortions", Proc. of Czech–German Workshop on Speech Processing, Czech Academy of Sciences, pp. 124–129. Prague, 2004.

[3] Hunt, A.J., Black, A.W. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", Proc. of ICASSP, vol. 1, pp. 373–376, Atlanta 1996.

[4] 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.

[5] 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.

[6] Balestri, M., Pacchiotti, A., Quazza, S., Salza, P.L., Sandri, S. "Choose the Best to Modify the Least: A New Generation Concatenative Synthesis System", Proc. of Eurospeech, vol. 5, pp. 2291–2294, Budapest 1999.

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

[8] "Method for Objective and Subjective Assessments of Quality", ITU-T Recommendation P.800, 1996.