Is Unit Selection Aware of Audible Artifacts?

Jindřich Matoušek, Daniel Tihelka, Milan Legát

University of West Bohemia, Faculty of Applied Sciences,
New Technologies for the Information Society,
Univerzitní 8, 306 14, Plzeň, Czech Republic
{jmatouse,dtihelka,legatm}@kky.zcu.cz

Abstract

This paper presents a new analytic method that can be used for analyzing perceptual relevance of unit selection costs and/or their sub-components as well as for tuning of unit selection weights. The proposed method is leveraged to investigate the behavior of a unit selection based system. The outcome is applied in a simple experiment with the aim to improve speech output quality of the system by setting limits on the costs and their sub-components during the search for optimal sequences of units. The experiments reveal that a large number (36.17%) of artifacts annotated by listeners are not reflected by the values of the costs and their sub-componets as currently implemented and tuned in the evaluated system.

Index Terms: speech synthesis, unit selection, concatenation cost, target cost, audible artifacts

1. Introduction

Despite the increasing popularity of HMM based and hybrid speech synthesis methods, unit selection concatenative systems still represent the mainstream in many real life applications, especially in limited domains where synthesized chunks are combined with prerecorded prompts. In such applications, the ability of the unit selection to deliver highly natural and to the recordings well fitting output are the key factors. Not surprisingly, the unit selection also remains the first choice for eBook reading applications, which have been acquiring a lot of interest over the recent years. This is due to a better acceptance of the unit selection synthetic speech output quality by end users.

Nevertheless, the unit selection method has seemed to be getting abandoned as a research topic over the last few years. There is no question that a huge amount of efforts have already been invested in improving its speech output quality since the introduction of the method [1]. It has been analyzed from almost all possible angles. Many works have dealt with experiments introducing different speech parameterizations and distances, which could be used for measuring the quality of concatenations [2], [3]; the target cost sub-components; pruning of the large unit databases; tuning weights of the costs [4]; and last but not least optimizing the unit search to lower computational costs of the method [5], [6], to name some.

Still, we believe that the most important problem related to the unit selection—the haphazard presence of audible artifacts—has not been investigated thoroughly enough. Generally speaking, there are three main sources of these quality drops. First, any database, no matter how thoroughly it is verified, contains mislabelings at different levels. Second, the costs that are used when searching for the optimal sequences of units are not always well correlated with human perception. Third, the traditional implementation of the search algorithm allows, as long as the cost of the whole sequence of units is minimum, for selecting units that should locally be avoided according to their assigned costs. This can especially be observed when the unit database is small. In theory, the same behavior can however be observed in large footprint systems as well.

Little has also been invested in analyzing the audible artifacts and real understanding of the latent constructs that influence human perception of them. This is predominantly a consequence of not having reliable objective methods for TTS quality evaluation as well as large costs and labor intensiveness of the subjective methods. In this paper, we present a method, the goal of which is to provide more insight into the two latter sources of audible artifacts mentioned above. The proposed method represents in its nature an analytic complement to the traditional TTS quality evaluation techniques (e.g. MOS or ABX tests).

The rest of this paper is organized as follows. The next section briefly describes our implementation of the unit selection. We put stress on explaining the individual costs used in our system as they are important for understanding the results of the presented experiments. The proposed analytic method as such does not however depend on their implementation. Section 3 deals with the first perceptual experiment, the goal of which was to detect synthetic units/chunks of utterances that contain audible artifacts. In Section 4, we describe the second perceptual experiment showing to what extend can setting of limits on values of costs and their sub-components im-

prove the quality of the system's output. In Section 5, we briefly discuss the obtained results, and finally, in Section 6, we draw conclusions and outline the intension for our future work.

2. ARTIC TTS System

2.1. Overview

ARTIC (Artificial Talker in Czech) is a Czech text-to-speech system developed since 1997. It is a corpus based system, which makes use of a large carefully designed speech inventory annotated at orthographic, phonetic and prosodic levels. Two speech synthesis methods—fixed-inventory synthesis (a.k.a. diphone synthesis) and unit-selection synthesis using diphones as basic units—had originally been implemented [7]. The system has recently been extended by the HMM-based synthesis [8].

The experiments described in this paper are mainly related to our unit selection implementation. The current target and concatenation cost design is described in the following subsections. The total cost is then a simple sum of the two costs.

2.2. Concatenation Cost Implementation

The concatenation cost consists of three sub-components—the difference in energy, the difference in $F0$ and the Euclidean distance of 12 MFCC coefficients [9]. All the values are z-score normalized in order to align their ranges. Moreover, the $F0$ sub-component is only computed when concatenating diphones at voiced ends. In case that voiced/unvoiced segments are to be concatenated, the $F0$ sub-cost is set to 1. Unvoiced segments are concatenated at zero $F0$ cost. Values of all features are calculated pitch-synchronously and the total concatenation cost is calculated as an average of the values of the three sub-components.

2.3. Target Cost Implementation

To compute the target cost, the following features are evaluated:

- *suitability for prosodic word position*. The feature evaluates the difference in position within prosodic word by a non-linearly increasing penalization [10]. This allows to avoid discrete *initial, middle, final* features and to non-linearly model the positions in a continuous space.
- *type of prosodeme (a sort of a prosodic phrase)* [11]. This feature uses simple binary match criterion.
- *left and right phonetic context*. This feature, also often used as a sub-component of the concatenation cost, penalizes disagreements in left and right phonetic contexts of a given diphone. Similarly to the prosodemes, this feature is binary with all the disadvantages of it. However, some analyses have recently been undertaken to overcome this limitation [12].

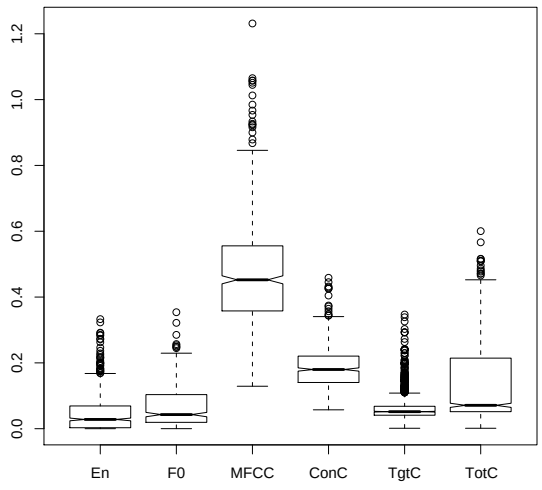

Figure 1: *Boxplots of costs of the units forming the optimal sequences found by the unit selection search algorithm.*

Each feature is weighted by a heuristically set weight (prosodeme the most prominent, the phonetic context the least), and the value of the target cost is then given by the weighted average:

$$TgtC = \frac{\sum_{t=1}^{T} F(t) \times w(t)}{\sum_{t=1}^{T} w(t)}, \qquad (1)$$

where $F(t)$ is the feature value, and T is the number of features.

3. Perceptual Annotation Experiment

3.1. Outlier Detection

As already mentioned in the introduction, this work is aiming at the audible artifacts haphazardly appearing in the output of the unit selection systems. If we start with an assumption that the costs correlate reasonably well with human perception, most of the selected units of extreme costs should lead to audible artifacts.

In order to see whether or not such units are being selected at all, the box-and-whisker diagrams (boxplots) were used. The boxplots of values of all concatenation cost sub-components and also of the costs as such of the units forming the optimal sequences of units in our test set of utterances are shown in Fig. 1. The plots indeed show that some units of rather outlying costs tend to appear in the selected sequences of units.

The goal of the next step is to investigate whether these, in terms of the costs, outlying units coincide with audible artifacts. An "annotation" perceptual experiment

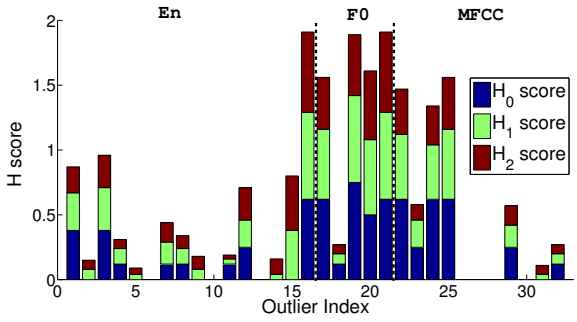

Figure 2: *The H_L scores of the outliers of the concatenation cost sub-components.*

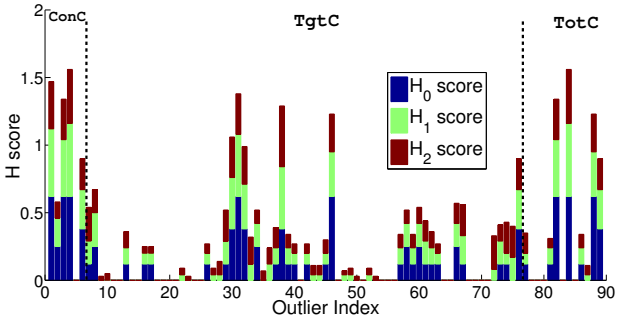

Figure 3: *The H_L scores of the outliers of the unit selection costs.*

was conducted using a set of 50 sentences. At least one unit of an outlying cost or sub-cost was found in approximately 80% of the selected sentences. The test sentences were selected randomly from a large news text corpus. All sentences were manually checked to make sure that they do not contain foreign language inclusions, out-of-vocabulary words or complex tokens that could be wrongly handled by the front end (text analysis) module of our system.

The task of listeners was to mark segments, which they found unnatural or containing any sort of distortion. The shortest segment which could be marked was a phoneme. Most of the participants were typically marking segments of approximate length of 3–5 phonemes. The test was conducted using a web interface allowing the listeners to work from home. It was, however, stressed in the test instructions that the annotations shall be done in a silent environment and using headphones. The listeners were only presented with audio, they did not have access to any visual information like spectrograms or oscilograms of the test sentences. Since the annotation of audible artifacts is not a simple task, only experienced listeners were invited to participate. In total, 8 listeners finished the listening test, 5 of them being TTS researchers.

Generally speaking, the annotations can also be obtained from naive listeners. In that case, a larger pool of listeners is needed, and the perceptual relevance threshold thr defined in the following section needs to be increased.

3.2. Listening Test Evaluation

Generally speaking, it is not a simple matter, to evaluate an annotation listening test. One of the concerns always is how to identify non-reliable listeners. This particular issue was not a problem in our study as all participants were highly motivated to provide good quality annotations.

Another issue is the different sensitivity of each participant to various kinds of artifacts. In order to evaluate the perceptual relevance of the outliers, keeping the sensitivity issue in mind, the H_L *score* (2) was introduced:

$$H_L(i) = \frac{\sum_{n=i-L}^{i+L} D_n}{(2L+1) \times N}, \tag{2}$$

where L stands for a tolerance interval length, i is the index of a given outlier, N is the number of listeners. D_n, the number of annotations of a particular phoneme, is defined as follows:

$$D_n = \sum_{i=1}^{N} h_n(j), \tag{3}$$

where $h_n(j)$ is an annotation of the phoneme n defined as:

$$h_n(j) = \begin{cases} 1 & n \in A_j \\ 0 & n \notin A_j \end{cases} \tag{4}$$

where the set A_j is the list of indeces of phonemes annotated by the j-th listener. The H_L *score* in fact represents the number of annotations obtained for each phoneme and its close neighborhood.

Having the H_L *score* defined, each position of a unit of an outlying cost or sub-cost (hereafter referred to as "outlier") was assigned its value. Fig. 2-3 show all outliers and their H_L *scores* sorted by groups corresponding to the concatenation cost sub-components and the costs themselves. Note that the length of the tolerance interval was set to $L = 2$, which was motivated by the above mentioned observation that most listeners used 3–5 phoneme long segments for annotating.

To further quantify the perceptual relevance of the outliers, we have defined a perceptual threshold[1] $thr = 0.5$ for the sum of H_{0-2} *scores* of a particular phoneme (hereafter referred to as $S_2(i)$), and calculated hit/false alarm rates. Summing the H_L *scores* up to the length L allows for normalizing the relevance of artifacts annotated exactly at a particular phoneme with those annotated less precisely. The *Hit Rate* was defined as:

$$Hit\ Rate = \frac{N_{hit}}{N_{outl}} \times 100\,[\%], \tag{5}$$

[1]Experiments showing the impact of different settings of the perceptual threshold thr are presented in [13]. Based on those experiments, the value $thr = 0.5$ used in the current paper was set for the sake of clarity of the method explanation.

Table 1: *The perceptual relevance of the outliers of the concatenation cost sub-components.*

	En	F0	MFCC
Hit Rate [%]	31.25	80.00	45.45
False Alarms [%]	68.75	20.00	54.55
Missed Rate [%]	82.98	91.49	89.36

Table 2: *The perceptual relevance of the outliers of the unit selection costs.*

	Join Cost	TgtCost	TotCost
Hit Rate [%]	83.33	40.79	33.33
False Alarms [%]	16.67	59.21	66.67
Missed Rate [%]	89.36	59.57	91.49

where N_{hit} is a number of outliers of a given cost or a cost sub-component for which the condition $S_2(i) \geq thr$ is fulfilled, and N_{outl} stands for a number of all outliers found for a given cost or cost sub-component. Analogically, the the *Missed Rate* can be defined as:

$$Missed\ Rate = \frac{N_{mis}}{N_{annot}} \times 100\,[\%]\,, \qquad (6)$$

where N_{mis} is a number of *annotated artifacts*, i.e. phonemes fulfilling the condition $S_2(i) \geq thr$, that do not match any outlier position, and N_{annot} is the total number of annotated artifacts.

The results are summarized in Tab. 1-2. We also present the percentage of the annotated audible artifacts missed by each of the costs and their sub-components. In total, 36.17% of the annotated artifacts are not identified by either of the outliers. It is interesting to compare for example the results obtained for the $F0$ and $MFCC$ sub-components of the concatenation cost. It can be seen that both sub-components miss about the same number of *annotated artifacts*, but the $F0$ sub-component shows considerably higher *Hit Rate*.

4. Perceptual Preference Experiment

Having the results of the annotation experiment, it was interesting to speculate how the quality of the synthetic utterances change (if at all) when a limit is set on the costs and their sub-components during searching for the optimal sequences of units forming the test sentences. In other words, what impact has pruning of the search beam based on a pre-set maximum allowed value for the costs and their sub-components.

Obviously, too radical pruning of the search space can lead to inability of the search algorithm to deliver the target sequence of phonemes. Nevertheless, having a large unit database on hand, such an experiment can be conducted. Each concatenation cost sub-component, as well as the costs themselves, were assigned a maximum threshold equal to the value of the upper whiskers of the respective boxplots shown in Fig. 1 (note that the whiskers are placed using 1.5 times the interquartile range; more details can again be found in [13]).

Table 3: *The impact of setting a limit on the concatenation cost sub-components.*

	En	F0	MFCC
Improvement [%]	31.25	41.67	33.33
Deterioration [%]	18.75	16.67	16.67
No impact [%]	50.00	41.67	50.00

Table 4: *The impact of setting a limit on the unit selection costs.*

	Join Cost	TgtCost	TotCost
Improvement [%]	50.00	66.67	66.67
Deterioration [%]	10.00	0.00	33.33
No impact [%]	40.00	33.33	0.00

To evaluate the impact of the modification of the search algorithm, the ABX preference test was conducted. The test sentences were re-synthesized using the modified system and presented to listeners in randomized pairs together with their original versions. The test participants were the same as in the annotation listening test. The task of the listeners was to express their preference regarding the overall quality of the samples. The test also contained sentences that were identical due to not containing any outliers. These sentences were used to check the reliability of the ratings as no preference was expected for the pairs containing them. Again, no visual information was provided to the listeners.

The following results were obtained: 5-prefer original, 9-no preference and 10-prefer modified version. The figures represent ratings for which 60% of listeners found an agreement, also the pairs containing the identical sentences are not included.

The obtained results show a slight preference to the modified system. Despite the fact that the preference is clearly not statistically significant, it is still interesting to analyze removal of which outliers lead to the largest improvement rate. The result of this analysis is shown in tables Tab. 3-4 and will be discussed in the section to follow.

5. Discussion

Let us first take a look at the results obtained for the target cost. It can be seen that removing the related outliers seems to lead to improvements of the system. This is in contrast to the perceptual importance of the target cost outliers obtained in the first test. We believe that this discrepancy is due to the different nature of the two perceptual experiments. While the first one poses implicitly the requirement on the listeners to mark as short segments as possible, the target cost would actually require the opposite as it is rather a supra-segmental cost. On the other hand, setting a limit on the target cost has bigger effect

on the behavior of our system. This is because the target cost outliers appear in larger quantities due to to a large extend binary nature of the cost, and also because when the target cost is "violated", our system stays with this "violation" as long as the concatenations are believed to be smooth according to the concatenation cost or not needed at all.

If we turn next to the concatenation cost and its sub-components, it can be seen that a better consistency was found between the two experiments. In line with the discussion in the previous paragraph, this is perfectly understandable result. Also, it comes as no surprise that $F0$ is the most important sub-component of the concatenation cost in our current implementation. This observation is supported by previous experiments showing fine-grained $F0$ contours as powerful predictors of concatenation discontinuities [14].

Finally, setting the limits on the costs and their sub-components is only one of the potentially possible ways of avoiding units of outlying costs in the selected sequences of units. An interesting alternative could be to tune the unit selection weights using zero number of units with outlying costs as a tuning objective.

6. Conclusions and Future Work

In this paper, two perceptual experiments were presented aiming at the analysis of the audible artifacts present in synthetic speech produced by the unit selection based system. The first experiment forms together with the detection of the outlying values of the unit selection costs and their sub-components a powerful analytic method for the unit selection based TTS systems. The method is driven by the actual costs of an evaluated system, which allows for leveraging the method for the analysis of different systems. We would like to invite interested readers to cooperate on measuring their unit selection implementations using the proposed method.

It has been found that only marginal system improvement can be achieved for our system by setting a limit on the costs during search for the optimal sequences of units. This can be due to data scarcity in our acoustic database. The bigger concern however is the rather low perceptual relevance of the currently used costs and their sub-components. In order to achieve a bigger quality improvement, a rigorous analysis of perceptual cues have to be undertaken. The need for this analysis is further amplified by the observation that 36.17% of the artifacts annotated by the listeners remain unindetified by either of the currently used costs and their sub-components. At the same time, a large number of units with outlying costs do not correspond to audible artifacts annotated by the listeners.

We plan to further experiment with the proposed approach by its extension into more voices and larger sets of data. It is also our intention to conduct the experiment proposed in the last paragraph of the previous section, i.e. to tune our system against the criterion of minimum number of outlying units.

Finally, we also want to look more closely at the audible artifacts that do not correspond to extreme values in the currently used costs and investigate whether or not they can be due to mislabelings in our acoustic database.

7. Acknowledgements

Support for this work was provided by the TA CR, project No. TA01011264 and by the European Regional Development Fund (ERDF), project "New Technologies for the Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

8. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP '96*, vol. 1, Atlanta, Georgia, May 1996, pp. 373–376.

[2] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," in *ICSLP '98*, Sydney, Australia, 1998, pp. 1983–1986.

[3] J. Vepa, "Join cost for unit selection speech synthesis," Ph.D. dissertation, University of Edinburgh, 2004.

[4] J.-D. Chen and N. Campbell, "Objective distance measures for assessing concatantive speech synthesis," in *EUROSPEECH '99*, Budapest, Hungary, September 1999, pp. 611–614.

[5] D. Tihelka, J. Kala, and J. Matoušek, "Enhancements of Viterbi search for fast unit selection synthesis," in *INTERSPEECH '10*, Makuhari, Japan, 2010, pp. 174–177.

[6] S. Sakai, T. Kawahara, and S. Nakamura, "Admissible stopping in Viterbi beam search for unit selection in concatenative speech synthesis," in *ICASSP '08*, Las Vegas, USA, 2008, pp. 4613–4616.

[7] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text–to–speech system ARTIC," in *Proc. of the 9th International Conference TSD 2006, Lecture Notes in Artificial Intellingence*, vol. 4188. Springer Berlin / Heidelberg, 2006, pp. 439–446.

[8] Z. Hanzlíček, "Czech HMM-based speech synthesis," in *Proc. of the 13th International Conference TSD 2010, Lecture Notes in Artificial Intelligence*, vol. 6231. Germany: Springer Berlin / Heidelberg, 2010, pp. 291–298.

[9] D. Tihelka, "Symbolic prosody driven unit selection for highly natural synthetic speech," in *INTERSPEECH '05*, Lisbon, Portugal, 2005, pp. 2525–2528.

[10] D. Tihelka and J. Matoušek, "Unit selection and its relation to symbolic prosody: A new approach," in *INTERSPEECH '06*, vol. 1, Pittsburgh PA, USA, 2006, pp. 2042–2045.

[11] J. Romportl and J. Matoušek, "Formal prosodic structures and their application in NLP," in *Proc. of the 8th International Conference TSD 2005, Lecture Notes in Artificial Intelligence*, vol. 3658. Germany: Springer Berlin / Heidelberg, 2005, pp. 371–378.

[12] M. Legát, "Impact of phonetic context mismatches on quality of vowel concatenations," in *Proceedings of 2012 IEEE 11th International Conference on Signal Processing*, Beijing, China, October 2012, pp. 523–526.

[13] ——, "Configuring TTS evaluation method based on unit cost outliers detection," in *Proc. of the 16th International Conference TSD 2013, Lecture Notes in Artificial Intelligence*, 2013, p. (accepted).

[14] M. Legát and J. Matoušek, "Pitch contours as predictors of audible concatenation artifacts," in *Proceedings of the World Congress on Engineering and Computer Science 2011*, San Francisco, USA, 2011, pp. 525–529.