Evaluation of various unit types in the unit selection approach for the Czech language using the Festival system

Martin Grůber, Daniel Tihelka, Jindřich Matoušek

Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic

gruber@kky.zcu.cz, dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

Abstract

The present paper focuses on the utilization of concatenative speech synthesis, aiming to determine and compare the influence on the synthesized speech quality when various unit types are used in the unit selection approach. There are several unit types which can be used for this purpose. This work deals with those most widely used, i.e. *halfphones*, *diphones*, *phones*, *triphones* and *syllables*. Speech was synthesized using these unit types and the outcome was listened to a by number of listeners, whose task was to evaluate the quality of synthetic speech. The result of the listening test performed for the Czech language is presented. However, it can be assumed that the results would be probably equal for other languages with similar structure, as we made no language-dependent modification in the Festival system. No research of a similar character has been conducted yet, so this unique evaluation should suggest what unit types are appropriate for general TTS systems.

Index Terms: speech synthesis, unit selection, various unit types

1. Introduction

The unit selection approach is one of the possibilities of the concatenative speech synthesis. Today, the method is extensively used due to its simplicity and the increasing quality of the speech produced.

The main principle of concatenative speech synthesis is the concatenation of segments of natural speech signal, which is stored in a *speech corpus* in the form of utterances. It is assumed that speech is composed of *acoustical (speech) units*. The real speech signal is by means of automatic or hand-made segmentation divided into *segments* which correspond to the speech units. These segments are stored in a *unit inventory* as a list of all units, which can be used for synthesis. The synthesized speech is produced as a concatenation of appropriate units from this inventory. It is evident that the synthetic speech, generated in this way, reproduces the voice of the speaker who recorded the speech corpus.

As was mentioned above, the cornerstone of speech is a speech unit. It is an absolute term for marking the same type of speech sound. The specific realization of the specific unit is called *candidate of the speech unit*. However, there is an issue of what the length of the unit should be. The maximum coverage of coarticulation effects and trouble-free concatenation (neither spectral nor prosody discontinuities) are the requirements to meet in this task. In this respect, we would like to choose long units, e.g. words or sentences. On the other hand, we need to keep the unit inventory as small as possible, i.e. to use only a limited number of different units. This requirement makes us use shorter units. In the course of choosing unit type, a trade-off has to be made.

Although we have our own system for speech synthesis [1], *The Festival Speech Synthesis System*[1] [2] was used in order to compare the speech synthesized by various unit types. It would be more difficult to implement the application of various unit types into our system than into the Festival system, which is used for experiments like this. Afterwards, we are planning to apply the achieved results and findings in our system as well.

The Festival system is an environment which was developed at The Centre for Speech Technology Research at The University of Edinburgh. One of its purposes is to allow the researcher to focus on his own problem in terms of speech synthesis instead of developing a whole complex system. Festival is composed of modules which can be modified independently. We adapted those that were originally used for standard diphone unit selection speech synthesis in such a way that it allows the application of four more unit types.

First of all, in section 2, a brief description of the Festival system is stated. Section 3 is dedicated to the application and implementation of various unit types (diphones, phones, triphones, halfphones and syllables) in the Festival system. There are described modifications which were needed to be performed in order to use these units in Festival and the achieved results are also shown. In section 4, the synthesized speech quality using different unit types is evaluated and compared by means of a listening test.

All of the units in the present paper are named according to the Czech version of SAMPA phonetic alphabet.

2. The Festival system

2.1. Introduction

The Festival system is an environment which is suitable for the development of speech synthesizers. It is being used for synthesis in a number of languages, but the basic version contains only data for English and Spanish. The system is intended for 3 groups of users:

- Users who want to generate high quality speech from general text without any knowledge of speech synthesis and without a need to intervene in the process.
- Users who design dialogue systems or any other systems and need to use the output of the speech synthesis. In this case, some changes need to be performed, e.g. particular voice or phrasing selection.
- Researchers developing new methods and approaches to speech synthesis. Indeed, we are among these users, aiming at improving speech synthesis quality. We modified the Festival system so that we could reveal features

[1] free download at http://www.cstr.ed.ac.uk/downloads/

which affect speech quality and make changes to the process of synthesis in order to be able to test various unit types for the purposes of this paper.

2.2. Unit selection

In the unit selection approach, synthetic speech is produced by concatenating speech units selected from a unit inventory.

Each target speech unit has its own list of candidate units. The naturalness of the synthetic speech is then affected by both unit types chosen and candidates selected to build speech. However, once a unit type is chosen it cannot be varied (except for the use of hybrid units, which is not our case), so the only way of controlling speech quality is the criterion of candidate selection. Usually, it consists of two costs.

The first, called *target cost* reflects how each candidate meets the requirements for communication function (what the synthesized phrase is supposed to express or communicate), which also includes the prosodic and phonetic context. The differences between the desired target unit and the real features of a candidate are crucial. In the Festival system, the following features were chosen to describe the communication function: emphasis, position in a syllable, position in a word, position in a phrase, left and right context. Each of these features has a different weight (weights are determined ad hoc) and an overall cost is calculated. It is clear that the application of various unit types requires various features. Some of those mentioned above cannot be used for all of the unit types. For example, the determination of the feature 'position in syllable' is absurd for syllable units and, therefore, it is useless. Other unit types need other modifications in the unit selection algorithm, so we had to make changes to the Festival system in order to be able to use all of them, as described in section 3.

The second one, *join cost*, means how the candidate unit meets the requirements for perceptual smoothness. The differences between the following features of two successive units affect the join cost in the Festival system: F0 and spectral discontinuity (computed as Euclidean distance of vectors composed of z-score normalized 12 MFCC coefficients and energy). These features are also weighted unlikely. The spectral characteristics are determined in instants of time when the first unit of the concatenation ends and the second one begins in their original utterances. It is not guaranteed that the MFCC coefficients are appropriate for the characterization of a unit or computing the join cost; however, they are still widely used for speech synthesis. There is no proof of which features currently examined are the best ones and could be used instead of these coefficients. Thus, we also used them for all unit types in order to be able to compare results correctly.

The best sequence of units is then found using the Viterbi algorithm through the whole unit inventory. It attempts to minimize a cost function which combines the two costs mentioned above.

2.3. Unit inventory

In order to use the unit inventory, it is necessary to create it in such a form that the Festival system needs. In our approach, automatic segmentation (see [1] and [3]) is made by using HTK tools. Moreover, we are also trying to improve it by new methods so that it is able to determine the boundaries of phones more accurately [3]. The current segmentation process produces a file that is not directly usable in Festival. As its output are segments in the form of triphones, several modifications have to be made, and new files (one file for one utterance in a database) are created. For the testing of various unit types, it is easier to adapt these files for all the desired types rather than to make significant modifications in Festival, but some changes in the unit handling modules in the system are still required.

Each file with an utterance has a specific structure and contains the following items: phrases, words, syllables and segments from the utterance, and the relations between these items are also saved there (e.g. which syllable is contained in which word, etc.). These files are then used as a part of the unit inventory. Exactly in this form they can only be used for triphones; for the other unit types they have to be modified. Especially segments need to be renamed and times of their beginning and end have to be determined according to the unit type.

As mentioned above, the MFCC coefficients are used for join cost calculation, so they have to be included in the unit inventory. For the synthesis, the LPC coefficients and residual signal are used. Therefore, it is essential to store these coefficients as well. This is a standard setting in the Festival system; however, the application of different coefficients for join cost computation as well as different coefficients for storing the waveform could be used.

The unit inventory for every single unit type contains all the items mentioned and it is loaded by the Festival system before the synthesis.

3. Application of various unit types in the Festival system

The effort to improve the quality of synthesized speech leads us to the question which unit type is suitable for speech synthesis and under what conditions. Nowadays, there are debates regarding the best unit type selection. It is difficult to determine what type is best for speech synthesis in a TTS system, each having its own advantages and disadvantages. In this paper we attempt to conduct some experiments and establish the strengths and weaknesses, thus contributing to answering this question. By comparing the results achieved we should draw a conclusion what unit type appears to be the best one. Eventually, we could take advantage of every particular unit type and suggest the use of this type in a special system, e.g. any speech synthesizer in a limited domain, which is also as very current topic. This research is unique in comparing the unit types under the same conditions. For all unit types, there is used the same speech corpus, segmentation, features for cost computation, etc.

In order to use all the below-mentioned unit types, we had to make some additional changes in Festival. One of the major modifications consists in the integration of our system of phonetic transcription for the Czech language. The other considerable modification was adding a syllabification algorithm [4]. Both these changes were needed to be performed in order to be able to process any Czech text incoming into the Festival system.

In the following subsections, the application of diphones, phones, triphones, halfphones and syllables is subsequently presented.

3.1. Diphones

We used diphones in this experiment as it is a commonly used unit type in speech synthesizers. Diphones are also the basic units which are used in the Festival system, requiring minimum amount of effort to implement them.

A diphone is a unit beginning in the middle of one phone and ending in the middle of the subsequent phone. The bound-

ary of the diphone is then in the area with stationary signal, which should improve the quality of concatenation. Each unit contains a transition between phones, thereby also including a coarticulation effect which is very important for the naturalness of the synthetic speech.

As mentioned in section 2.3, modification was needed to be performed in the files with utterances for this unit type. It consisted in the renaming of the segments from triphone form to the phone form, because Festival is ready for working with diphones implicitly on the basis of phone names. Festival is able to generate diphone names within the system.

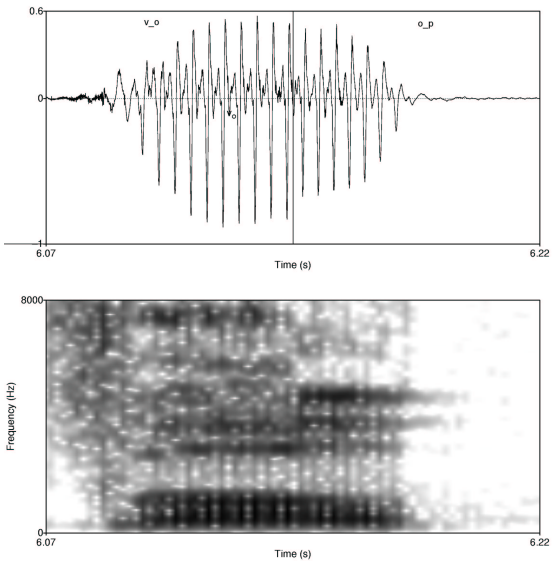

Figure 1: Concatenation of two diphones, originally non consecutive, but continuous in synthesized speech. Waveform and spectrogram.

In fig. 1 the waveform of two concatenated diphones [v_o] and [o_p] is shown. This concatenation was produced as a result of synthesis. It seems to be almost smooth, in spite of the fact that the units were selected from different utterances and they were not originally consecutive. In the spectrogram, the point of concatenation is still visible in the area of higher frequencies (about 4-5 kHZ), but it was not perceived at all.

In fig. 2, there is presented another concatenation of two diphones, [h_a] and [a_#] ([#] denotes pause). Again, they were selected from different original utterances and were non-consecutive. This time, the point of concatenation is extremely visible in the waveform as well as in the spectrogram and it was reported to cause speech degradation in the middle of the phone [a]. For solving this problem, there should be some correction (e.g. some type of normalization) to ensure that there will be at least approximately the same amplitude level. But there is no simple solution because by amplifying the signal of one diphone, we could need to amplify another, and energy accumulation could occur.

One of the advantages of diphones is their relatively small amount. Taking into account the fact that Czech language has 43 different phones, plus 3 types of pauses (loud breath, break and boundary break) and glottal stop, in the sum we have 47 different phone units, it means we have $47^2 \approx 2200$ different diphone units. In addition, some of them don't practically appear

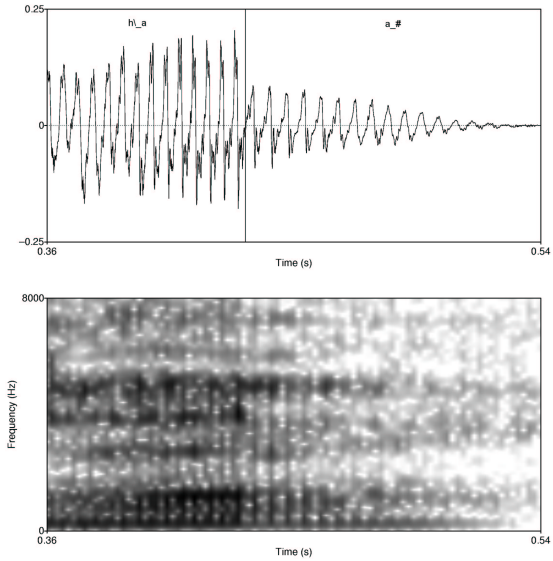

Figure 2: Concatenation of two diphones, originally non consecutive, visibly non continuous in synthesized speech. Waveform and spectrogram.

in the common text, see table 1 in section 4.

We made no changes in the target cost and join cost computation algorithm for diphones since the Festival system implicitly treats them in the desired way.

3.2. Phones

A phone is considered to be one of the fundamental phonetic units of speech. The application of this unit type then could seem to be very natural. However, as the boundaries of a phone unit are determined directly by segmentation, it is necessary for the segmentation to be made very accurately. Otherwise, one phone is likely to contain a part of another, which is absolutely undesirable and affects the synthetic speech quality.

Since the triphone segmentation was used, as described in 2.3, triphone labels needed to be renamed to phones which were then stored in Festival's utterance files. This time, changes were made also in the Festival system because otherwise it wouldn't be able to interpret the segment names properly. We had to edit the part of unit handling module that stores the units in Festival's unit inventory.

To illustrate the effect of segmentation inaccuracy, there is shown a concatenation of two units, [v] and [a], that were non-consecutive in the original utterance in fig. 3. The first one ([v]) has a different right context in the original utterance. It is phone [o] and it is easy to see that this phone affects the unit chosen for synthesis. The quality of the synthesized speech is worsen by this effect. Apparently, in this particular case, the cost penalizing incorrect right context was outweighed by other costs.

The seeming advantage is the count of the phone units. For the Czech language we have 47 phones, as was mentioned in the previous section. However, this means that there is a huge amount of candidates for the target unit. Therefore, the enumeration of the best candidate sequence is computationally very exacting and time-consuming. On the other hand, in a very specialized limited domain speech synthesizer (e.g. on the basis of

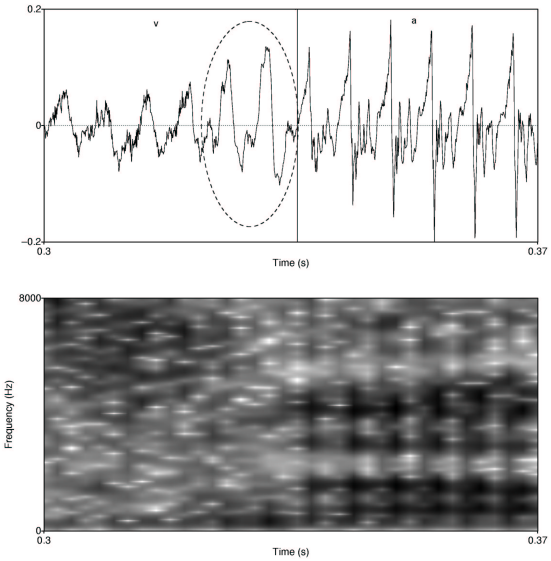

Figure 3: Transition between non-consecutive phones. Right context of the [v] unit was [o] in the original utterance, whereas in the sythesized speech it is phone [a]. This is also visible in the waveform.

sentence unit type), the phones may be advantageous to be used for connecting the sentences in a meaningful way. In that case, diphones would be inappropriate due to their quantity.

We made again no changes in the target cost computation algorithm but a modification was made in the join cost computation. Measurement of the difference between F0 in the joint of two units has sense only in such a case, that we concatenate two voiced or contrariwise two unvoiced units (where the difference should be zero as well as the values of F0). So the algorithm was edited accordingly. When there is a concatenation of a voiced unit with an unvoiced one, this cost is set to zero.

3.3. Triphones

On the basis of good experience with this unit type in the ARTIC speech synthesizer [1], [5], we included it in this research as well. By its principle, it should suppress the disadvantages of phones regarding the transition between units; however, there is still the segmentation problem.

Boundaries of a triphone are the same as for a phone, but the unit includes information about its context. Thus, instead of considering e.g. phone [o], we consider triphone [l-o+r], which means the phone [o] is preceded by phone [l] and followed by phone [r].

For this unit type, no modifications were made in the files with utterances. These files already contain the names of the segments as it is required for the application of this unit type. On the other hand, a modification had to be implemented in the Festival system in order to be able to read the unit inventory from the files correctly.

It is clear that by using this approach we have a large number of different triphone units. In the place of 47 phone units we have $47^3 \approx 100000$ triphone units. As well as in the case of diphones, not all the units appear in the real utterance. However, it is still a great deal of triphones and it is almost impossible to have such a unit inventory that would contain all of them. To avoid this problem, there is an algorithm that groups together the units with similar context. It is made by virtue of the acoustic similarity.

In order to find out which units should be in the same group, we need to take a look at their acoustic signal and a phonetic similarity. Well-suited combination of these two aspects divides the phones for potential left context into 15 groups, and for right context into 14 groups.

For example, phones [p], [t] and [k] are in the same group for the left context. For the right context, these phones are also in the same group, but, in addition, there are also phones [t_s], [t_S] and all 3 types of pauses along with them.

Using this grouping we have only approximately 10.000 units, but it is still possible that during the synthesis there will be a missing unit. In the Festival system, so-called backoff rules (see [2]) can be used. These rules enable the replacement of a unit which is not in the inventory by another unit which is similar to the missing one. It is obvious that this method can be applied mainly to triphones. For other unit types, this kind of replacement could change a sense of synthesized utterances.

The target cost computation did not need to be modified. In the join cost computation algorithm, the same changes as described in previous section for determination of F0 difference were performed.

3.4. Halfphones

The application of halfphones was presented by AT&T Labs in [6] and the results shown there are very promising. Thus, we attempted to compare also this unit type with the others in order to prove or disprove their qualities.

Halfphones are units which start at the beginning of a phone (or in the middle of it) and end in the middle of the same phone (or at the end of it) - they are created by cutting a phone into two halves. Thus, the phone [a] is divided into a sequence of two halfphones, [a1] and [a2].

For the application of this unit type, we had to adapt both the unit inventory in the form of files with utterances and also the Festival system. The main modification was renaming of segments in the unit inventory and editing Festival so that it was able to use this unit type.

The tendency to use the halfphone units could partially replace the application of hybrid diphone-phone (diphone-triphone) unit types. When the halfphones, which are selected during synthesis time, were originally consecutive in an utterance, it means that they could be concatenated into phones, diphones or even longer units. The point of concatenation is sometimes in the middle of a phone and sometimes on the border. As it is noted in [6], the halfphones should be promising units because they could maintain the advantages of phones and diphones. However, they also have disadvantages. One of them is the fact that they are very short, so in a synthesized utterance there is a large number of concatenations. As it is known, at the point of concatenation there could arise many problems which, however, were not reported in [6].

The number of halfphone units should be doubled as compared to the number of phone units, but we didn't cut into halves the units representing pauses. It means that there are 91 different units. This simplification shouldn't affect the final quality of the synthetic speech.

At computation of target cost for these units, there is an anomalous situation. One of the costs which penalizes differ-

ent left or right context will always be zero (except pause units, because they are treated as phone units). The unit [a1] will always have the unit [a2] as its right context and vice versa, [a2] will always have [a1] as its left context. The algorithm computing this cost could also be modified in such a way that it would consider as a context one more unit following (preceding) the immediate neighbouring unit. The other features affecting the target cost remained the same as for previous unit types. In the join cost computation algorithm, there was made a modification in order to measure F0 difference meaningfully, as described in section 3.2.

3.5. Syllables

Syllables are taken in this experiment as the only representative of longer unit types. It is interesting to confront the previous phone-like unit type with syllables, which include more than one phone (a typical Czech syllable has 2-3 phones).

Syllables are often considered the phonological building blocks of words with boundaries aligned to phones. There again can arise the problem of segmentation inaccuracy.

For this unit type, the files with utterances didn't need to be edited. Segments were ignored and only syllables were used. The modification of the Festival system was in this case more extensive than before. Firstly, we needed to adapt the system, so that it could accept the correct names for syllable units. It was performed the same way as it was performed for previous unit types, by editing the unit handling module.

In addition, some changes in the target cost computation were needed to be carried out, especially the left and right context penalization. It is not necessary to take into account the whole syllable adjacent to the target unit. It is assumed that the whole syllable which forms the context doesn't affect it. Thus, only the last phone of the preceding syllable was treated as the left context and the first phone of the following syllable was treated as the right context. Moreover, these phones were divided into groups in the same way as was done for left and right part of triphone name in section 3.3. The reason is the high number of different syllable units.

The next thing to change in the target cost computation was the feature called position in a syllable. It was removed because it is pointless to use this feature.

As well as for the previous unit type, the join cost computation algorithm was modified. The F0 difference was measured only in such cases when it was meaningful, i.e. when the concatenation occurred in the transition between two voiced or two unvoiced phones.

The problem of the application of syllables is the amount of units. It is not easy even to make a list of all syllables in the Czech language. We use an automatic syllabification [4], which is performed for the phonetically transcribed text, and some syllables are thereby different from the case when it would be implemented for orthographical form of the same text. In addition, the syllabification is not always unambiguous in the Czech language.

In spite of these problems, the list containing about 14.000 syllables which should be included in the unit inventory was generated. There have to be all possible units, and this requirement is almost impossible to achieve. In the application of the syllable units, the backoff rules included with the Festival system are unusable. So in a real TTS system, there has to be another way of synthesizing utterances containing unavailable syllables, e.g. some combination of shorter units. However, like phones, a limited domain synthesis can profit from the advantages that syllables have.

4. Conclusion

In order to compare the results of application of various unit types, we used our speech corpus for synthesizing a listening test. The corpus, recorded in a consistent news-like style by a semi-professional female speaker with some radio-broadcasting experience, contains approximately 12.5 hours of natural speech, stored in 5000 utterances. During synthesis, statistical data about units were collected and are presented here.

Units	Number of different units
Diphones	1528
Phones	47
Triphones	3023
Halfphones	91
Syllables	5684

Table 1: Number of different units in unit inventory for each unit type

In table 1, there is the number of different units in the unit inventory for each unit type. It can be seen that in our fairly large corpus, we covered only 70% of diphones, 30% of triphones and 40% of syllables. phones and halfphones were covered completely, because the number of different units is very low for these unit types. When synthesizing the sentences, we encountered a problem with missing units for triphones and syllables. Therefore, we had to choose such sentences to synthesize which contain only the units we have. For this experiment it is conceivable as we aimed to prove the behaviour of units, not to build a real TTS system where this would have to be solved by another way. For example, in the Festival system the backoff rules could be more adapted to this problem when using triphones or any type of hybrid synthesizer [4] could be used for syllables.

Units	Maximum number of candidates	Minimum number of candidates	Average number of candidates
Diphones	5004	3	1519
Phones	38451	309	17618
Triphones	9994	15	552
Halfphones	38451	309	17693
Syllables	3317	1	788

Table 2: Statistics about units used during the synthesis of utterances for the listening test

In table 2, there is stated maximum, minimum and average number of candidates for each unit type used during the synthesis. You can see, that phones and halfphones have the highest maximum and minimum number of candidates, and these numbers are the same for both of them. The average number differs because in our approach we used the same pause units for phones as well as for halfphones, we didn't cut them into halves. Although the results display the statistics obtained for units used for synthesis of the testing sentences, the results for whole corpus will be very similar.

Taking into account the number of units in a synthesized sentence, which was approximately 150, the number of possible concatenations for phones, diphones and triphones is

about n^{150}, where n is the average number of candidates for particular unit types. For halfphones, it is approximately n^{300} because the number of units in the synthesized utterance is doubled. Finally, for syllables it is about n^{60}. It is evident that for phones and halfphones, the algorithm computing the best units sequence needs to perform lots of operations and the whole process of synthesizing is highly computationally exacting. The synthesis of one utterance for the listening test using phones and halfphones takes approximately 24 hours. The fact that it takes the same time for both unit types, in spite of there being more possible concatenations for halfphones, may be explained by any kind of optimalization used by the Festival system, which needs to be more verified. The synthesis using the other unit types takes only a few minutes, but it was still out of real time. However, it does not matter for our experiment because we examined qualities of unit types rather than possibilities of speech synthesis acceleration.

The same corpus, as described earlier, was used to synthesize a listening test. It consists of 5 sentences, each of them was synthesized in 5 various versions. The versions were different in the unit type that was used for synthesis. The sentences were not originally in the corpus and they were selected from newspaper articles.

The listeners were asked to evaluate the synthesized sentences in all versions by marks 1 to 5 (optimally to sort them by quality from the worst one to the best one), where the 5 means the best, this mark always having to be used for the best sentence in terms of naturalness, fluency, intelligibility and prosodic consistency. Sentences which seemed to be equal could be evaluated by equal mark. Afterwards, normalization was performed in order to take advantage of the whole scale. The resulting average marks and standard deviations are shown in table 3.

Units	Average mark	Standard deviation
Diphones	3.61	1.22
Phones	1.88	0.91
Triphones	3.57	1.40
Halfphones	3.81	1.35
Syllables	2.24	1.33

Table 3: The average marks and standard deviations for various unit types

Halfphones with the average mark 3.81 were evaluated as the best unit type. Diphones and triphones have more or less equal marks, as when compared in [5]. However, after performing a statictical one-way analysis of variance (ANOVA), it was proved that there is no significant difference between triphones, diphones and halfphones. During statistical comparison of the results of these three unit types, the p-value for the null hypothesis, that there is no difference among means, reached the value 0.65.

Syllables with the average mark 2.24 were rated a little better than phones, which were identified as the worst ones with the average mark 1.88. This occurs even though the algorithm looking for the best phone sequence theoretically had the best opportunity to select the most appropriate units due to the highest number of candidates. However, in this case as well, the difference between the means of the marks for these two unit types is not statistically significant, which was proved by the ANOVA test. The p-value was determined as 0.094.

On the other hand, between these two groups (diphones, triphones and halfphones on one side, and phones and syllables on the other side) a significant difference was detected. The p-values were equal or near-equal to zero when comparing unit types from one group with those from the other group.

There are further factors which affect unit selection and which can be changed. One of them are weights, used for computation of the target cost and the join cost. In this experiment, Festival implicit setting of these weights was applied. The balancing of the weights should influence the final synthetic speech quality and this setting might be dissimilar for each unit type. However, we attempted to maintain equal conditions for all the tested unit types and in that way achieve a consistent result.

The conclusion may suggest that halfphones, diphones and triphones are comparable regarding the synthetic speech quality. However, taking into account the fact that the synthesis using halfphones was multiple with respect to computational complexity, the application of diphones or triphones seems to be more profitable.

5. Acknowledgements

We thank all those that have supported our research work and development of the synthesis of the Czech language.

This paper has been funded by GAČR 102/06/P205, the Academy of Sciences of the Czech Republic project No. 1ET101470416 and the EU 6th Framework Programme IST-034434.

We also greatly appreciate the help of the computation centre available at our university, which allows us to perform this computationally exacting and time-consuming speech synthesis experiments.

6. References

[1] Matoušek J., Romportl J., Tihelka D., Tychtl Z.: "Recent Improvements on ARTIC: Czech Text-to-Speech System", *In Proceedings of INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing. Jeju, Korea, vol. III, pp. 1933-1936. ISSN 1225-441x*

[2] Clark R. A. J., Richmond K., King S.: "Festival 2 - Build Your Own General Purpose Unit Selection Speech Synthesiser", *CSTR, The University of Edinburgh*

[3] Matoušek J., Tihelka D., Psutka J.: "Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction", *In Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH 2003. Geneva, Switzerland, pp. 301-304, ISSN 1018-4074*

[4] Matoušek J., Hanzlíček Z., Tihelka D.: "Hybrid Syllable/Triphone Speech Synthesis", *In Proceedings of Interspeech 2005 - Eurospeech, Lisbon, Portugal, s. 2529-2532, ISSN 1018-4074, 2005*

[5] Tihelka D., Matoušek J.: "Diphones vs. Triphones in Czech Unit Selection TTS", *TSD 2006. Lecture Notes in Artificial Intelligence 4188, Springer-Verlag, Berlin, Hiedelberg, 2006, pp.531-538., 2006*

[6] Conkie A.: "Robust Unit Selection System for Speech Synthesis", *In Proceedings of the Eurospeech '99 Conference, Budapest, Hungary, 1999.*