Hybrid Syllable/Triphone Speech Synthesis

Jindřich Matoušek, Zdeněk Hanzlíček, Daniel Tihelka

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

Abstract

In this paper, the syllable, an alternative phonetic unit to the phone, is researched in the context of speech synthesis. Several approaches to syllable modelling within the statistical approach (using hidden Markov models) to the acoustic unit inventory creation are proposed and evaluated. To be able to synthesize an arbitrary text, the syllable inventories were supplemented with triphones resulting in hybrid syllable/triphone inventories. Listening tests were accomplished both to assess the quality of the resulting synthetic speech produced using the hybrid syllable/triphone inventories and to choose the best approach to syllable modelling. The resulting synthetic speech is highly intelligible and fluent. Although the synthetic speech generated using the baseline triphone inventory was assessed slightly better, the results of the very first experiments with syllable modelling are very promising.

1. Introduction

Nowadays, concatenative synthesis is the most widely used approach to speech synthesis. This approach employs an acoustic unit inventory (AUI) which should comprise all relevant sounds of a language to be synthesized. Traditionally, context-dependent phones (triphones), diphones or even sub-phone units like half-phones, fenemes or senones are employed [1]. AUIs built on such short units are general and flexible enough to be efficiently used in text-to-speech (TTS) synthesis tasks while being of reasonable size, which allows them to be employed also in low-resource devices (e.g. pocket PCs, mobile phones, etc.). On the other hand, there is a need of relatively many concatenation points when producing the resulting synthetic speech. Since these points are sources of possible discontinuity problems (caused mainly by coarticulation phenomenon or prosody mismatch) in the synthetic speech, special attention (e.g. spectral and prosody smoothing or unit-selection technique) should be paid to these points to ensure the synthetic speech to be as "fluent" as possible. However, spectral smoothing and other signal modifications could result in the increased computational requirements and potentially introduce some degradation to the resulting speech. Alternatively, the unit selection technique bypasses the need of signal modifications by using very large AUIs with many representatives of each unit in different spectral and prosodic contexts, having enormous requirements on memory of TTS devices. Although these methods result in high-quality synthetic speech, their usage in low-resource devices is not possible. However, without treating the concatenation points the synthetic speech could suffer from the deterioration of the quality of speech.

One solution could be to use larger units (words or phrases), which usually results in synthetic speech of very high quality, even if no special treatment is applied. Such long units do catch the coarticulation phenomena inside their speech segments. Moreover, the coarticulation is not so strong at the boundaries of words as it is at the boundaries of phone-like units. However, employing such long units is not practical from the point of view of TTS synthesis, where an arbitrary text could appear at the input of the system, i.e. any speech could be synthesized (not only the given words or phrases).

A compromise between short and long units is examined in this paper. Another phonetic unit – *the syllable* – is proposed to be the basic unit for speech synthesis. A syllable usually consists of more phones (a typical Czech syllable has 2–3 phones), preserving coarticulation between phones inside the syllable [1]. Thus, synthetic speech using syllables should result in fewer concatenation points and sound more "fluent". Moreover, the syllable is generally considered to be the basic speech unit expressing the prosodic characteristics of speech. As a result, the syllable seems to be a suitable unit in speech synthesis tasks, because its usage is straightforward and justifiable from the phonetic point of view and could result in AUIs of reasonable sizes. In this paper the statistical approach to AUI creation is modified to enable syllable modelling as well.

The paper is organized as follows. In Section 2 the baseline TTS system is briefly described. Section 3 deals with several proposals how to model syllables and how to segment speech into syllables in a fully automatic way. The automatic syllabification process in Czech speech is discussed here too. In Section 4 we present the results of our experiments. Finally, Section 5 contains the conclusion and outlines our future work in this field.

2. Baseline speech synthesis system

In our previous work we have designed a modern Czech TTS system ARTIC [2] based on concatenation of context-dependent phones (i.e. triphones). A technique for the automatic construction of the triphone inventories was proposed and utilized in the system. Based on a carefully designed speech corpus [3], *statistical approach* (using three-state left-to-right single-density model-clustered crossword-triphone hidden Markov models, HMMs) was employed to create AUI of the Czech language in a fully automatic way. As a part of this approach, decision-tree-based clustering of similar triphone HMMs was utilized to define the set of basic speech units (i.e. clustered triphones) used later in speech synthesis. As a result, all the speech available in the corpus was segmented into these triphones. Then, the most suitable instance of all candidates of each triphone was selected off-line and used as a representative of the unit during synthesis.

This research was supported by the Academy of Sciences of the Czech Republic, project No. 1ET101470416, and the Ministry of Education of the Czech Republic, project No. MSM235200004.

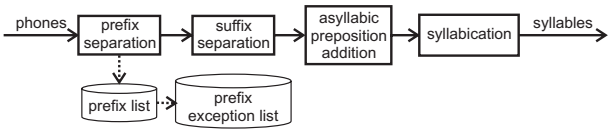

Figure 1: *A block diagram of the automatic syllabification process.*

The speech corpus comprises 5,000 phonetically balanced sentences (about 13 hours of speech). Each sentence is described by linguistic and signal representations of speech. As for linguistics, both orthographic and phonetic transcriptions of each sentence are used. Speech signals are represented by their waveforms and their spectral properties are described by vectors of Mel Frequency Cepstral Coefficients (MFCCs) calculated using 20 ms windowed speech signal with 4 ms shift. In the current system 12 MFCCs plus normalized energy together with corresponding first, second and third differential coefficients (52 coefficients in total) are used.

As for text-to-speech, simplified text processing is carried out, limiting itself to punctuation-driven sentence clauses detection, simple text normalization (transcribing digits and abbreviations), and detailed rule-based phonetic transcription. Both rule-based and data-driven prosody generation could be utilized. The synthetic speech is produced by a modified OLA method (both in time and frequency domain). More details about the baseline system could be found in [2, 3, 4].

3. Experiments

In this section, the experiments with syllable modelling are described in more detail. After a series of experiments three approaches to syllable modelling emerged. The first one re-segments the original triphone segmentation, thus obtaining syllable segmentation (see Section 3.2). The other approaches described in Sections 3.4 and 3.5 are more systematic and create syllable inventories by explicit syllable modelling and segmentation of speech into syllables. The approaches differ in context modelling. The same speech corpus as described in Section 2 was used throughout this work. All the experiments were proposed for the Czech language, although the proposed approaches to syllable modelling are supposed language-independent.

3.1. Automatic syllabification

In TTS systems, the input text is typically processed and converted into a sequence of phones – the basic pronunciation units of speech. When syllables are about to be used as the basic speech units in speech synthesis, a process called *syllabification* should be applied on the sequence of phones, converting it into the sequence of syllables.

We have proposed an algorithm for the automatic syllabification of the sequence of Czech phones (see Fig. 1). It respects the morphological structures of words and prevents detecting syllables across word stems and prefixes or suffixes. Since the detection of syllables in the sequence of phones is ambiguous, especially in consonant clusters where the correspondence between written and pronunciation forms of words is less evident, the algorithm makes the process of syllabification consistent. Our algorithm tends to detect open syllables when no consonant clusters are present. In the case of consonant clusters, the syllable boundary is located between the consonants.

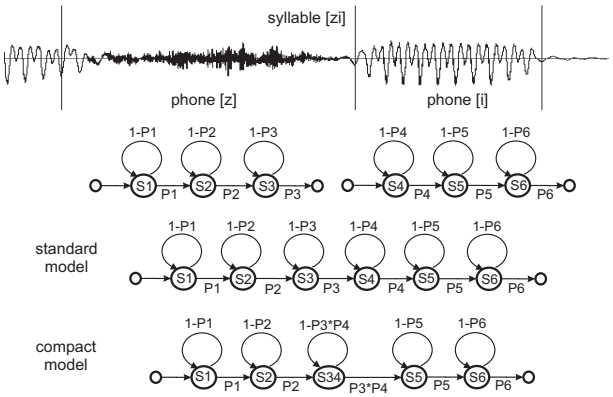

Figure 2: *An example of a syllable HMM. Si denotes the states and Pj the transitional probabilities.*

3.2. Re-segmenting triphone boundaries (SYL0)

In fact, no explicit syllable modelling is performed in this approach. The original triphone segmentation is used [2] and then re-segmented into syllables. The syllables are detected as described in Section 3.1. Thus, the inter-syllabic boundaries are given by the corresponding triphone boundaries and boundaries between phones inside syllables are removed. However, the resulting syllable inventory does not contain all Czech syllables, because the original AUI was built to comprise all Czech (clustered) triphones. Therefore, to ensure that an arbitrary Czech syllable could be synthesized, the syllable inventory is extended with the original triphone inventory and syllables not present in the syllable inventory (more specifically, syllables with low occurrences – see Section 3.4) are synthesized as a sequence of corresponding triphones. Although this approach represents the most intuitive change from the original triphone inventories to syllable ones, the resulting synthetic speech is far from the quality of the original triphone synthetic speech. This may be caused by the non-explicit syllable modelling (syllables are composed of triphone segments modelled by separate triphone HMMs). Then, when the concrete instance of a syllable segment is to be selected, the average triphone-based score within syllable is used [2] to assess every syllable segment in the corpus. The score computed in such a way does not have to necessarily describe the syllable segment in the best way. So, other solutions for syllable modelling were researched and are described in the next subsections.

3.3. Syllable HMM definition

To enable the exact syllable modelling, we modified the statistical approach based on triphone HMMs modelling [2]. The key task here was to define a syllable HMM. We propose to create a syllable model as a composition of each phone (or triphone) models (as shown in Fig. 2). The "standard" syllable model results from the straight concatenation of the individual phone HMMs. So, on the assumption that the phone HMMs consist of 3 emitting states, the composite syllable HMM has $3p$ states (where p denotes the number of phones in the syllable). We also experimented with "compact" syllable models. In these models, the outside states of the neighbouring phone HMMs, which represent transitional speech signals from one phone to another, were put together, reducing the number of states in a syllable model to $2p + 1$. Since no audible differences were

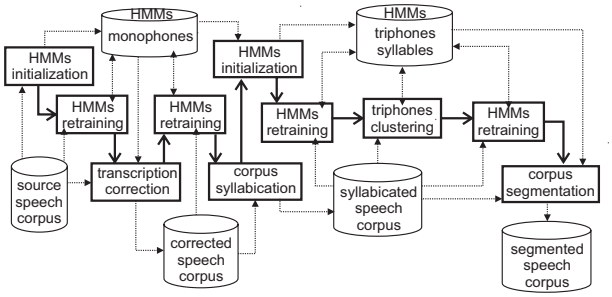

Figure 3: *Schematic illustration of context-independent syllable modelling.*

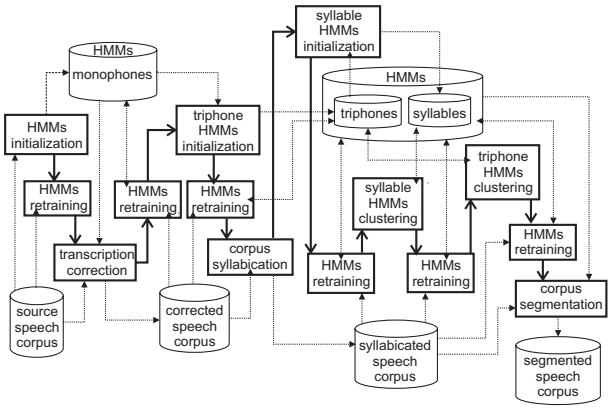

Figure 4: *Schematic illustration of context-dependent syllable modelling.*

perceived in the synthetic speech when both models were used, the compact syllable model was exclusively employed in the experiments described in the next subsections.

3.4. Context-independent syllable modelling (SYL1)

The statistical approach to AUI creation typically trains and clusters the models of speech units. After the final models are available, they are used to segment the speech waveforms present in the source speech corpus into the speech units.

First of all, from the point of view of TTS synthesis, it is important to mention that it is not possible to use syllable HMMs only unless it is ensured all Czech syllables are present in the corpus. In our case, this requirement was not guaranteed, as sentences to be recorded were selected with respect to triphones [3]. To cope with this, only syllables that occurred more than n-times (we chose somewhat arbitrarily $n = 20$) were left in the system and stored into the resulting AUI. Syllables with number of occurrences less than n were replaced by the corresponding triphones; such triphones were (after some clustering, see below) stored into the AUI as well. Thanks to such a hybrid syllable/triphone inventory, an arbitrary text could be synthesized.

The whole process starts with HMMs initialization. In the case of context-independent syllable HMMs, there are two options how to initialize syllable HMMs. Syllable HMMs could be initialized directly from the speech data or on the basis of phone HMMs. Since the phone HMMs are very robust and well-trained, reasonable estimates of initial syllable HMMs could be expected. Moreover, the correction of phonetic transcription (so-called *realignment* [2]) of each utterance in the speech corpus could be accomplished without affecting the syllable models. On the other hand, less accurate syllable HMMs were obtained when they were initialized directly from the speech data. In addition, during the realignment process some syllable HMMs in some utterances could be replaced by phone/triphone models corresponding to an alternative phonetic transcription, reducing the robustness of such syllable HMMs.

The phone-based initialization method was used throughout our work to initialize both syllable and triphone HMMs. After some retraining of both syllable and triphone HMMs, the clustering of similar triphone HMMs was carried out both to get more robust triphone HMMs and to enable synthesizing triphones not present in the source speech corpus (the triphone clustering was essentially the same as described in [2]). The syllable HMMs were not touched by this operation. After clustering both the syllable and clustered triphone HMMs were retrained. Finally, the resulting syllable and triphone HMMs were aligned with the speech data, producing time stamps corresponding to the boundaries between syllables and/or triphones (this process is known as the *segmentation* of speech into the given units). The segmented syllables and triphones can be then stored into the hybrid syllable/triphone inventory and used later in speech synthesis. The detailed scheme of the AUI construction process can be seen in Fig. 3.

3.5. Context-dependent syllable modelling (SYL2)

Although the synthetic speech produced using AUI based on SYL1 approach was of much higher quality than SYL0 approach, some distortions or nonfluencies were still notable. Such glitches were present especially at the syllable boundaries. Hence, in our next research we introduced *context-dependent syllable HMMs*.

The process of modelling and segmentation of context-dependent syllable and triphone HMMs is shown in Fig. 4. It is an extension of context-independent syllable modelling (SYL1) described in Section 3.4. In this approach, context-dependent syllable HMMs were initialized with corresponding triphone HMMs. Again, only syllable HMMs that occurred more than n-times in the speech corpus were left in the system, other syllables were replaced by the corresponding triphone HMMs. Due to the miscellaneous speech contexts it is not practical to take individual phones as the context of syllable HMMs. It would lead to an enormous number of undertrained syllable HMMs. Instead, more general contexts were used. Each context was formed by a cluster of acoustically similar phones (namely each vowel in his short and long version, voiced and unvoiced plosives, voiced and unvoiced fricatives, and voiced and unvoiced affricates).

The Table 1 shows some interesting statistics about the described methods to syllable modelling (TRI denotes the baseline triphone inventory). The comparison of both context-independent and context-dependent syllable modelling is given in Section 4. We also experimented with the number of occurrences n defining syllables in the system.

Table 1: *Statistics of different inventories (AUI size is in MB).*

Method	# Syllables	# Triphones	Total	AUI size
SYL0	947	7,103	8,050	37
SYL1	957	2,254	3,211	19
SYL2	5,771	2,462	8,233	60
TRI	0	7,103	7,103	22

4. Results

In this section the quality of the proposed hybrid syllable/triphone inventories is assessed, mainly with respect to the quality of the resulting synthetic speech (see Section 4.1). Another option how to evaluate the quality of AUI is to express the accuracy the units are segmented with. This could be done by comparing the automatic segmentation to the reference manual segmentation. The segmentation accuracy could be expressed as the percentage of the automatically detected boundaries which lie within a tolerance region (10 ms in our case) around the reference boundary. The segmentation accuracy of all approaches to syllable modelling described in this paper was similar, ranging between 70% – 75% (with slightly better results for SYL2 approach).

4.1. Listening tests

A number of listening tests were carried out to evaluate different approaches to syllable modelling with respect to the quality of the resulting synthetic speech. For instance, the tests help us to find the "optimal" steps when proposing the syllable modelling scheme SYL1 (described in Section 3.4) and SYL2 (described in Section 3.5). The tests also revealed the poor quality of SYL0 approach to syllable modelling.

Let us describe some of the tests in more detail. In the first test (TEST1) the impact of the syllable context modelling on the quality of the resulting speech was evaluated. Comparison Category Rating (CCR) test [5] was used in this case. Two versions of the synthetic speech of the same sentence (10 different sentences were employed), one using context-independent syllables (SYL1) and the other using context-dependent syllables (SYL2), were played to the listeners, who were asked to compare the quality of those versions on a 7-point scale: "much better", "better", "slightly better", "about the same", "slightly worse", "worse", "much worse". 15 listeners participated in the test. The results (generalized to "better/worse/same" scale) shown in Fig. 5 (section TEST1) demonstrate the clear preference for the synthetic speech produced using context-dependent syllables (SYL2). All listeners preferred SYL2 to SYL1 with the average evaluation "better".

The second test (TEST2) probed the influence of the minimum number of occurrences that defines syllables in the system (see Section 3.4). Three minimum numbers of occurrences were taken into account. Hybrid syllable/triphone inventories based on SYL2 approach were employed. The same 15 listeners participated in the tests. The results in Fig 5 (section TEST2) show that there was a slight but consistent preference for syllables with more number of occurrences. So, robust syllable modelling seems to be important and leads to both better segmentation and speech synthesis.

To compare the hybrid syllable/triphone inventory SYL2 with our baseline triphone inventory, another CCR test (with the same setup as in TEST1) was carried out. 11 listeners took part in the test. There was a slight preference for the baseline triphone inventory (the evaluations in CCR test ranged between "about the same" and "slightly better").

5. Conclusion

In this paper, the syllable, an alternative phonetic unit to the phone, is researched in the context of speech synthesis. Several approaches to syllable modelling within the statistical approach (using HMMs) to the AUI creation were proposed. The basic units used are syllables, either context-independent or context-dependent ones. Since not all syllables could be present in the source speech corpus the inventories are built from, the syllable inventories were supplemented with triphones, resulting in hybrid syllable/triphone acoustic unit inventories. This is important for TTS synthesis where an arbitrary text could appear at the input. The text is first converted into the sequence of phones and then expressed as a sequence of syllables. When a syllable is not present in the inventory, corresponding triphones are used instead.

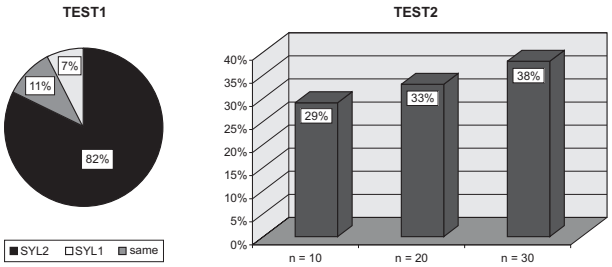

Figure 5: *The results of the listening tests.*

Listening tests were accomplished both to assess the quality of the resulting synthetic speech produced using hybrid syllable/triphone inventories and to choose the best approach to syllable modelling. Although the synthetic speech generated using the baseline triphone inventory was assessed slightly better, the results of the very first experiments with syllable modelling are very promising. The synthetic speech produced using the hybrid syllable/triphone inventory is of acceptable quality, highly intelligible and fluent. Thus, the hybrid syllable/triphone inventory could be viewed as an alternative to the baseline triphone inventory.

As only a few initial experiments with the explicit syllable modelling have been conducted so far, there is still room for improvement. More refinement and various parameters tuning (e.g. more precise context definition and syllable clustering, various syllable/triphone ratios, some modifications of the underlying syllable HMM topology and the training scheme, etc.) of the syllable modelling process could be proposed in the future work. A new speech corpus containing sentences selected with respect to syllables could also contribute to the more precise syllable modelling. Using syllables as both acoustic and prosodic units (and thus ensuring synchronization between acoustic and prosodic modules in TTS systems) opens up new possibilities for prosody modelling as well.

6. References

[1] Huang, X., Acero, A., Hon, H., "Spoken Language Processing", Prentice Hall PTR, New Jersey, 2001.

[2] Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z., "Recent Improvements on ARTIC: Czech Text-to-Speech System", Proc. ICSLP, vol. 3, Jeju, Korea, 2004, pp. 1933–1936.

[3] Matoušek, J., Psutka, J., Krůta, J., "Design of Speech Corpus for Text-to-Speech Synthesis", Proc. Eurospeech, vol. 3, Ålborg, 2001, pp. 2047–2050.

[4] Matoušek, J., Tihelka, D., Psutka, J., "Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction", Proc. Eurospeech, Geneva, 2003, pp. 301–304.

[5] "Methods for Objective and Subjective Assessment of Quality", ITU-T Recommendation P.800 (08/96), 1996.