Current State of Czech Text-to-Speech System ARTIC[⋆]

Jindřich Matoušek, Daniel Tihelka, and Jan Romportl

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz, rompi@kky.zcu.cz

Abstract. This paper gives a survey of the current state of ARTIC – the modern Czech concatenative corpus-based text-to-speech system. All stages of the system design are described in the paper, including the acoustic unit inventory building process, text processing and speech production issues. Two versions of the system are presented: the single unit instance system with the moderate output speech quality, suitable for low-resource devices, and the multiple unit instance system with a dynamic unit instance selection scheme, yielding the output speech of a high quality. Both versions make use of the automatically designed acoustic unit inventories. In order to assure the desired prosodic characteristics of the output speech, system-version-specific prosody generation issues are discussed here too. Although the system was primarily designed for synthesis of Czech speech, ARTIC can now speak three languages: Czech (both female and male voices are available), Slovak and German.

1 Introduction

This paper gives a survey of the current state of the text-to-speech (TTS) system ARTIC. ARTIC (Artificial Talker in Czech) has been built on the principles of concatenative speech synthesis, i.e. it primarily consists of three main modules: acoustic unit inventory (AUI), text processing module and speech production module [1]. Moreover, it is a corpus-based system; to our knowledge ARTIC is the only Czech TTS system using large carefully prepared corpora [2] as the ground for the automatic definition of speech synthesis units and the determination of their boundaries and also for unit selection technique. The block diagram of the ARTIC TTS system is shown in Fig. 1.

The paper is organised as follows. Section 2 briefly describes the phonetic inventory used in our system. Section 3 deals with the acoustic units actually employed in the system and the process of their automatic preparation. In Sections 4.1 and 4.2 two versions of our system are presented. Finally, Section 5 concludes the paper and outlines our future work.

[⋆] Support for this work was provided by the Academy of Sciences of the Czech Republic, project No. 1ET101470416, and the Grant Agency of the Czech Republic, project No. 102/05/0278.

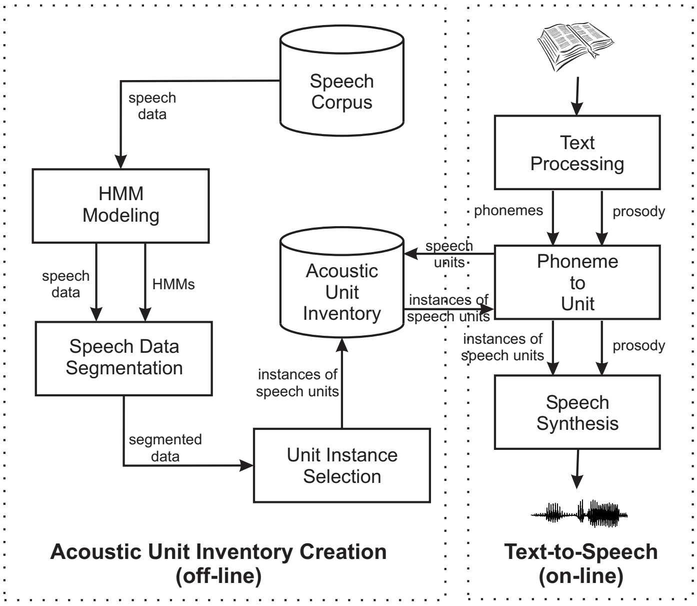

Fig. 1. A simplified scheme of the Czech text-to-speech system ARTIC.

2 The Phonetic Background

Phonemes, or phones respectively, are the basic phonetic units that represent the spoken speech of each language. Naturally, the phonetic inventories constitute a ground of each speech synthesis system. In our system we currently use 41 “basic” phones. The set of phones is shown in Table 1.

In addition to the basic set of phones, some significant allophones could be also utilised (see the the row “Allophones” in Table 1). Currently we use glottal stop [?] as it was shown to improve the quality of the synthesised Czech speech [3]. The nasal allophones [N] and [F] are employed very often as well. Two symbols representing pause are utilised too: the short inter-word pause and long silence. The former is very important for speech synthesis as it handles the word-to-pause or pause-to-word coarticulation and helps to maintain the correct speaking rate during speech synthesis. The latter is used mainly for modelling both the leading and trailing silences presented in the utterances of the source speech corpus. On the whole, 46 phone-like units are currently used in our system.

It is generally known that context independent phones are not suitable for speech synthesis tasks because they do not respect the phonetic features of the adjacent phones. Using phones without any complementary information about their neighbours would result in hardly intelligible synthetic speech that would suffer from phone-to-phone coarticulation problems. In speech synthesis systems more precise acoustic unit inventories (AUIs), which respect the phone-to-phone coarticulation phenomenon, are used. These inventories enrich the phone set

Table 1. Czech phonetic inventory used in our TTS system (in SAMPA [4] notation).

Basic Set	Vowels	[a], [a:], [e], [e:], [i], [i:], [o], [o:], [u], [u:]
	Diphthongs	[o_u], [a_u], [e_u]
	Plosives	[p], [b], [t], [d], [c], [J\], [k], [g]
	Nasals	[m], [n], [J]
	Fricatives	[f], [v], [s], [z], [Q\], [P\], [S], [Z], [x], [h], [j]
	Liquids	[r], [l]
	Affricates	[t_s], [d_z], [t_S], [d_Z]
Allophones		[F], [N], [?], [G], [r=], [l=], [m=], [@]

either by adding the information about the neighbours (the case of context dependent phones, so called triphones) or by shifting the phone boundaries so that the signal of an resulting acoustic unit partly covers the signals of more phones (in the case of diphones two halves of adjacent phones are captured). In our system both approaches are used. In the single unit instance system, described in Section 4.1, triphones are exclusively employed. On the other hand, in the multiple unit instance system presented in Section 4.2 both diphones and triphones can be used (currently, there is a slight preference for diphones). Some experiments with other unit types (e.g. halfphones or syllables [5]) were also conducted.

3 Acoustic Unit Inventory Creation

In concatenative corpus-based speech synthesis the source speech corpus forms the basis the speech synthesis system is built on. For our purposes, two speech corpora were designed very carefully in order to contain phonetically balanced sentences [2]. They comprise 5,000 to 10,000 sentences (about 13 to 18 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). In the current system 12 MFCCs plus normalised energy together with corresponding first, second and third differential coefficients (52 coefficients in total) are used.

Due to the very large corpora, automatic techniques have been searched for in order to create AUI. We have designed a statistical approach (using three-state left-to-right single-density model-clustered crossword-triphone hidden Markov models, HMMs) to the automatic construction of AUI of Czech language. As a result, all the speech available in the corpus was segmented into phones, or triphones respectively. Knowing the unit boundaries, AUI can be relatively easily built from the segmentation scheme (either using single or multiple instances per each speech unit) by collecting features needed in speech synthesis (e.g. speech samples or parameters, pitch-marks [2], duration, F0, etc.). The automation of the whole AUI process allows us also to control the size of the resulting AUI (by tuning up the clustering process, see Section 4.1 and/or working with single/multiple unit instances). Several experiments with the baseline segmentation system were carried out in order to find as precise segmentation as possible, e.g. the removal of the unit boundaries offset caused by HTK (the hidden Markov model toolkit) parameterization mechanisms [6], glottal stop modelling [3], various speech parameterization schemes [7], various HMM initialisation methods [6], correction of pause alignments [1], etc. When some pre-segmented data are available (by an expert in acoustic phonetics, preferably, or by a speaker-independent speech recognition system run in forced-alignment mode), a more accurate HMM initialisation method, so-called bootstrap, could be utilised to get slightly better segmentation results, and/or to use them to adjust the automatic segmentation [1]. Our segmentation system achieves the segmentation accuracy of 96 % (in tolerance region 20 ms) or 86 % (in tolerance region 10 ms) when compared to the reference manual segmentation [6].

4 Speech Synthesis System

Two versions of speech synthesis system are currently supported: single unit instance system and multiple unit instance system. Each version has its pros and cons. The single unit instance system uses a compact acoustic unit inventory (there is only one instance of each speech unit present in the inventory) and thus it is suitable for low-resource devices (mobile phones, pocket PCs, etc). On the other hand, the multiple unit instance system takes more instances of each speech unit into account and selects the optimal instances dynamically during synthesis runtime. Consequently, the resulting synthetic speech is of a higher quality, but at the expense of enormous memory requirements. More details will be explained in the following subsections.

Text processing forms an important part of a text-to-speech system. The input text is typically a subject of a thorough analysis and processing. The task of text processing module is then to get a unique phonetic representation of the written text. A punctuation-driven sentence clauses detection is performed to estimate the structure of a synthesised utterance. Since the input text generally contains “non-standard words” (abbreviation, acronyms, numbers and numerals written as figures, dates, hours, currency amounts etc.), so-called text normalisation based on a tagger performing context-dependent morphological disambiguation of each word is implemented in ARTIC [8]. Then, a detailed rule-based phonetic transcription takes care of the conversion from the normalised written (i.e. orthographic) to the pronunciation (i.e. phonetic) form. Finally, the acoustic units actually used in the system (i.e. triphones or diphones) are derived from the fundamental phonetic representation.

4.1 Single Unit Instance System

In the single unit instance system each triphone is represented and synthesised using a single instance. However, "pure" triphones could not be used directly in speech synthesis – their number is enormous (in the case of 46 phones it is 97,336 triphones) and a substantial part of them need not appear in the speech corpus even if a sophisticated sentence selection procedure is employed. There are also a number of triphones which appear very rarely in the corpus. For these reasons, it is good to cluster the set of all triphones and obtain a set of clusters with "similar" triphones present in each cluster. In our system, decision-tree based clustering of similar models of corresponding triphone HMMs has been utilised within the framework of the automatic speech segmentation process mentioned in Section 3 to define the set of clustered triphones. The clustered triphones are then the basic speech units used later in speech synthesis.

After the clustering and the final segmentation are done, many instances of each speech unit exist in the acoustic unit inventory. Unlike the multiple unit instance system, an off-line unit instance selection scheme was proposed to select a single, "most representative" instance per each speech unit. The selection scheme is based on a statistical analysis of all unit instances available in the segmented speech corpus (outliers are removed and the instance with the highest segmentation score is selected [9]).

Having a single instance of each speech unit, there is a need to modify the signals of the representative instances in order to meet the characteristics of the synthesised utterance, mainly the prosodic and spectral features. A modified OLA technique (both in time and frequency domain) is employed for these purposes and also for the concatenation of the modified units into the resulting synthetic speech. Some experiments with a harmonic/noise-based speech signal generation method were also conducted, especially in the context of the reduction of the AUI storage requirements [10].

The intelligibility and naturalness of synthetic speech is highly influenced by its suprasegmental features – i.e. prosody. In the case of the single unit instance system, an explicit prosody estimator/generator is needed.

The prosody model used with the single unit instance system (called data-driven prosody model) is conceptually similar to the approach of concatenative synthesis (enriched with a unit selection approach): it concatenates elementary prosody units derived from real speech data contained in a specially designed and annotated prosody corpus (unlike the rule-based model used in the previous versions of our TTS system [1]). The prosody units can have either one representative for a parametrisation of a specific portion of a text, or more representatives and a prosody generation module chooses the best fitting one according to a particular criterion, as it is analogically in a TTS unit selection approach.

The data-driven model comprises of two basic components – prosodic structures and a surface prosodic characteristics generator. Prosodic structures formally describe the linguistic functions of certain prosodic phenomena in terms of derivation trees produced by a generative prosodic grammar [11]. In other words it can be called a formal suprasegmental phonology where each word of a sentence is described by its distinctive features based on the position of the word within a certain prosodic structure (i.e. derivation tree). Prosodic structures consist of abstract units such as prosodic clauses, prosodic phrases, prosodic words, prosodemes, semantic accents. Each word is thus described by its relation to these units – so called description array.

The surface prosody generator is a classifier assigning each description array occurred in training data with a cadence – an intonation (and rhythmic) scheme (pattern) fitting into an interval of a single prosodic word. A cadence inventory is constructed using a suitable agglomerative clustering algorithm over vectors representing sampled F0 contours of prosodic words occurring in the training data. The classifier uses a mapping from the space of possible description arrays to the space of cadences while the so called relation of prosodic homonymy [12] solves cases when the particular description array is not in the training data and thus it is "unobserved" from the point of view of the model.

This classifier is solved and implemented quite satisfactorily (yet not the case for the prosodic homonymy and synonymy formalisation and detection, which still need to be thoroughly explored) and the current research in the field of prosody is focused mainly on a parser producing the prosodic structures of input sentences. Currently we use only a rule-based prosodic parser but an HMM-based and a probabilistic grammar-based parser are being developed and tested.

The prosodic structures – as a universal system of prosody formalisation – are also to be used as a symbolic prosody description in the system with multiple unit instances (described further in the text) where no explicit surface prosody modifications are carried out.

4.2 Multiple Unit Instance System

In the recent years we have started to deal with the multiple unit instance approach (mostly known as unit selection), and the first fully working version of the unit selection module has lately been integrated into the ARTIC TTS system.

Contrary to many other unit selection systems, our unit selection is driven by the target specification described only at a high-level by symbolic features (so-called deep structure). Consequently, no explicitly set low-level prosodic contours described in Section 4.1 (so-called surface structure) are required here [13]. Although one can object that the use of low-level features is advantageous for the control of the global prosodic character of the synthesised phrase, our results show that it is not so important. Moreover, this treatment allows us to avoid the necessity of the prosodic and spectral modifications of the selected candidate sequence, which cause the most significant degradations of speech quality (as shown in [14]).

The whole idea of the symbolic-driven unit selection is based on the fact that although it is possible to generate very natural explicit contours of prosodic characteristics (see section 4.1) used to drive TTS, it is very hard (if at all possible) to generate such explicit prosody which would guarantee the basic requirement for unit selection – to select adjacent units from an original phrase in the corpus if the phrase appears at the input of TTS. Moreover, our experience suggests that it is not straightforward to find a sequence of units which is both natural-sounding and following the explicit contours. Therefore, we try to adapt the concepts of *prosodic synonymy* and *homonymy* to our approach of unit selection. We expect these concepts to mimic the prosodic style of the original speaker on their own. Our preliminary results, described further, seem to suggest that our suppositions have been correct.

We carried out informal listening tests in order to evaluate the quality of the speech synthesised using the first version of the unit selection [13]. They were divided into 3 groups, the first comparing the unit selection and single unit instance versions, using a 7-point CCR test (3 for unit selection much better, 2 unit selection better, ..., -3 single candidate much better). The second test was a modified MOS evaluating naturalness (5 for completely natural, 4 almost natural, ..., 1 completely artificial), and the last test was used to assess the similarity of synthesised and natural phrases. The results, computed as the average of assessments produced by 14 – mostly lay – listeners, were very encouraging. The unit selection version was assessed as *much better* (average score 2.66) in the CCR test (while using the same corpus for the building of both systems!). The level of naturalness was evaluated as *almost natural* (score 3.95) in the MOS test, and the prosody style was perceived as *equal to the original* in 77% of evaluations (more details about the tests and their evaluations could be found in [13]).

5 Conclusion & Future Work

An overview of the current state of the text-to-speech system ARTIC was given in the paper. All substantial aspects of the system were outlined here. Two versions of the system were presented, each of them being suitable for different tasks. Having low demands on memory (up to 10 MB) and yielding the moderate output speech quality, the single unit instance system could be mainly used in low-resource devices (mobile phones, pocket PCs, etc.). On the other hand, a noticeably higher quality at the expense of the markedly increased computational requirements (hundreds of megabytes of RAM are required) could be obtained by the multiple unit instance system. As such, the multiple unit instance version is suitable for servers or powerful PCs.

After two Czech voices (male and female) were built on the principles described above, two other languages (Slovak [15] and German [16]) have been successfully implemented within the framework of ARTIC TTS system. Our text-to-speech system has been recently applied also in the area of audiovisual speech synthesis – the first computer 3D Czech talking head with realistic face animations was designed [17].

In our next work we will continuously aim at improving the quality of the synthetic speech produced by our TTS system. Beside other aspects (e.g. enhanced prosody generation or dynamic unit selection) a substantial attention will be paid to the improvements in the quality of the automatically designed acoustic unit inventories. We will focus mainly on the increase of the accuracy of the automatic segmentation of speech, and on minimising the size of the inventories while maintaining the quality of the resulting speech. Research in the field of the automatic voice conversion, which will enable to change the voice the system "speaks" with no need to record a huge number of new speech data, has been launched recently as well.

References

1. Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z.: Recent Improvements on ARTIC: Czech Text-to-Speech System. Proc. ICSLP, vol. III. Jeju Island, Korea (2004) 1933–1936.
2. Matoušek, J., Psutka, J., Krůta, J.: On Building Speech Corpus for Concatenation-Based Speech Synthesis. Proc. Eurospeech, vol 3. Ålborg, Denmark (2001) 2047–2050.
3. Matoušek, J., Kala, J.: On Modelling Glottal Stop in Czech Text-to-Speech Synthesis. Proc. TSD. Springer, Berlin (2005) 257–264.
4. Czech SAMPA. http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm.
5. Matoušek, J., Hanzlíček, Z., Tihelka, D.: Hybrid Syllable/Triphone Speech Synthesis. Proc. Interspeech. Lisboa, Portugal (2005) 2529–2532.
6. Matoušek, J., Tihelka, D., Psutka, J: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction. Proc. Eurospeech. Geneva (2003) 301–304.
7. Matoušek, J., Tihelka, D., Psutka, J: Experiments with Automatic Segmentation for Czech Speech Synthesis. Proc. TSD. Springer, Berlin (2003) 287–294.
8. Kanis, J., Zelinka, J., Müller, L.: Automatic Numbers Normalization in Inflectional Languages. Proc. SPECOM. Moscow (2005) 663–666.
9. Donovan, R. E., Woodland, P. C.: A Hidden Markov-Model-Based Trainable Speech Synthesizer. Computer Speech and Language 13:223–241 (1999).
10. Tychtl, Z.: Phase-Mismatch-Free and Data Efficient Approach to Natural Sounding Harmonic Concatenative Speech Synthesis. Proc. EUSIPCO. Wien, Austria (2004) 1027–1030.
11. Romportl, J., Matoušek, J.: Formal Prosodic Structures and their Application in NLP. Proc. TSD. Springer, Berlin (2005) 371–378.
12. Romportl, J.: Structural Data-Driven Prosody Model for TTS Synthesis. Proc. Speech Prosody, vol II. Dresden, Germany (2006) 549–552.
13. Tihelka, D.: Symbolic Prosody Driven Unit Selection for Highly Natural Synthetic Speech. Proc. Eurospeech. Lisbon (2005) 2525–2528.
14. Tihelka, D., Matoušek, J.: The Analysis of Synthetic Speech Distortions. Proc. Czech-German Workshop on Speech Processing, Czech Academy of Sciences. Prague (2004) 124–129.
15. Matoušek, J., Tihelka, D.: Slovak Text-to-Speech Synthesis in ARTIC System. Proc. TSD. Springer, Berlin (2004) 155-162.
16. Matoušek, J., Tihelka, D., Psutka, J., Hesová, J.: German and Czech Speech Synthesis using HMM-Based Speech Segment Database. Proc. TSD. Springer, Berlin (2002) 173-180.
17. Krňoul, Z., Železný, M.: Realistic Face Animation for a Czech Talking Head. Proc. TSD. Springer, Berlin (2004) 603–610.