ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Recent Improvements on ARTIC: Czech Text-to-Speech System

Jindřich Matoušek, Jan Romportl, Daniel Tihelka, Zbyněk Tychtl

Department of Cybernetics
University of West Bohemia in Pilsen
{jmatouse,rompi,dtihelka,tychtl}@kky.zcu.cz

Abstract

This paper presents recent improvements on ARTIC – the modern Czech corpus-based text-to-speech system. As a statistical approach (using hidden Markov models) was applied to create an acoustic unit inventory, several improvements concerning acoustic unit modelling, clustering and segmentation have been accomplished to increase the intelligibility of the resulting speech. Two approaches to the generation of prosodic features were also proposed and implemented to increase the naturalness of synthetic speech. To produce as smooth synthetic speech as possible, a multiple unit instance scheme with on-line unit candidate selection was proposed as well. Our work on an alternative harmonic/noise-based speech production method is also mentioned. In addition, an important step towards multilinguality was achieved as German and Slovak language modules were implemented besides two Czech voices within the framework of ARTIC TTS system.

1. Introduction

In today's world, speech technologies play an important role. Since they aim to make our lives more pleasant, we can encounter them more often in our everyday lives. For example, we can utilize a *speech recognition* system to dictate a letter or to summarize talks using keywords, or we can employ a *text-to-speech* system (TTS) to read SMS in handhelds, e-mails and other e-documents aloud. There are usually two criteria how to evaluate TTS systems. Firstly, synthetic speech must be *intelligible*. Secondly, it should sound *natural*. Additionally, with the increasing importance of global communication, another, *multilingual*, criterion arises: it is useful for TTS systems to be able to speak more languages.

In our previous work, we have designed ARTIC, the modern Czech corpus-based TTS system [1]. As a fully working *concatenative* TTS system, it consists of three main components: acoustic unit inventory (AUI), text processing and speech production modules. Based on a carefully designed speech corpus [2], *statistical approach* (using three-state left-to-right single-density state-clustered crossword-triphone hidden Markov models, HMMs) was employed to create AUI of Czech language in a fully automatic way. As a part of this approach, decision-tree-based clustering of similar states of corresponding triphone HMMs was utilized to define the set of basic speech units (clustered states in this case) used later in speech synthesis. As a result, all the speech available in the corpus was segmented into these clustered states. Then, the most suitable instance of all candidates of each state was selected off-line and used as a representative of the unit during synthesis. Simplified text processing was carried out, limiting itself to punctuation-driven sentence clauses detection, simple text normalization (transcribing digits and abbreviations), and detailed rule-based phonetic transcription. As intelligibility was our first goal, no prosody generation was implemented. The synthetic speech was produced by a modified OLA method (both in time and frequency domain). Although monotonous, it achieved a high degree of intelligibility. Hereafter, this system will be referred to as the baseline system (its scheme is shown in Figure 1).

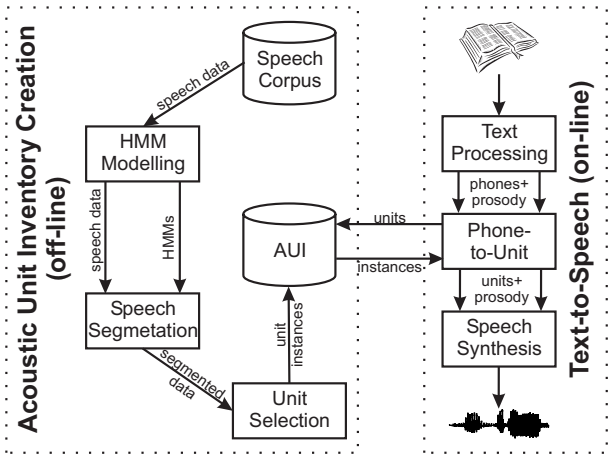

Figure 1: *A scheme of the baseline Czech TTS system ARTIC.*

In this paper[1], we describe several improvements to the baseline TTS system mentioned above. The improvements concern the increase of both intelligibility (by proper modelling, clustering and segmentation of acoustic units – see Section 2) and naturalness (by generating and synthesizing prosodic features in Section 3, and proposing an on-line unit candidate selection scheme in Section 4) of synthetic speech. In Section 5, the use of a harmonic/noise approach is discussed as an alternative speech production method. An important step towards multilinguality is mentioned in Section 6, as German and Slovak language modules were implemented beside two Czech voices within the framework of ARTIC TTS system.

2. Acoustic unit inventory

In this section we describe the most important improvements to the AUI we have made recently. These enhancements concern mainly using glottal stop, clustering acoustic units and some issues of automatic speech segmentation.

[1]This research was supported by the project No. 102/02/P134 of the Grant Agency of the Czech Republic and the project No. MSM235200004 of the Ministry of Education of the Czech Republic.

2.1. Glottal stop

Although our baseline system produced highly intelligible speech, certain distortions were observed in some speech contexts, especially in words starting with a vowel. In natural speech, such contexts are characterized by the presence of so-called *glottal stop*. Since glottal stop is not considered as a phoneme in the Czech language, it was not included in our baseline phonetic inventory. However, our findings showed the importance of the correct modelling of glottal stop for the intelligibility of the Czech speech.

After a series of experiments (with respect to the observations of Czech phoneticians) we proposed a phonetic transcription rule (in the form described in [1]) for inserting glottal stop into the sequence of Czech phones

$$\mathrm{VOW} \rightarrow \mathrm{?VOW} \;/\; \langle |, \mathrm{PREFV}- \rangle \;\_\,, \qquad (1)$$

where "VOW" stands for a vowel (or diphthong), "?" is a symbol for glottal stop, "PREFV" is a prefix or a part of a compound word ending with a vowel, and the symbol "−" marks the morphological juncture. The symbol "|" marks the word boundary. The text before the symbol "→" describes the input text to be transcribed, the phones after "→" express the result of the transcription. The symbol "_" separates the left and right context of the input text. If various contexts are allowed (denoted by "<" and ">"), individual components are separated by a comma.

Informal intelligibility tests (17 listeners were involved) were carried out to evaluate glottal stop modelling using Rule (1). Results shown in Fig. 2 (section GSA) confirm that the listeners did prefer the synthetic speech with glottal stops.

2.2. Clustering

Clustering is an important part of HMM-based AUI creation. Based on the source speech corpus and decision-tree-based technique [1], it divides the spectrum of speech sounds (represented by triphones) into groups. In each group, triphones with similar phonetic and acoustic features are located. Thus, clustering ensures more robust speech modelling and emulates an allophonic description of speech. In HMM-based AUI creation, the *clustering level* defines the types of speech units used later in synthesis. When clustering is performed on the *state level* (a default option for speech recognition purposes), so-called clustered states (also termed senones or fenemes [3]) can be used directly in speech synthesis. From the point of view of signal, a feneme represents a small subphoneme unit. It is a flexible unit that can effectively stand for an arbitrary speech context and thus it is very suitable for clustering purposes. However, concatenating such small units results in many concatenation points (three points per phone in the case of three-state HMMs). These points constitute possible sources of discontinuity problems that can be heard as some glitches degrading otherwise very intelligible and quite natural synthetic speech.

In our research we tried to retain the same speech context quality modelling while reducing the inherent number of concatenation points in synthesis. To do that, clustering was performed on the *model level*. As a result, it was possible to use the whole triphone (i.e. phone-sized unit) directly in speech synthesis. So, the number of concatenation points dramatically decreased to one point per phone.

Comparison Category Rating (CCR) listening tests were proposed to compare both kinds of clustering from the point of view of synthesis. The results in Fig. 2 (section CLU) show that two-thirds of 88 listeners evaluated the synthetic speech as more natural, fluent and with fewer intrusive elements if whole triphones were concatenated.

2.3. Segmentation

Substantial attention was also paid to the issues of *automatic segmentation* of speech into triphones. Employing the HMM-based approach to AUI creation, the resulting (clustered) triphone HMMs are aligned with the speech data, producing time stamps corresponding to the boundaries between triphones. The experiments are described in more detail e.g. in [4]. A very important finding was that when using HTK, the hidden Markov model toolkit, to segment speech, an *offset* $o = (w - s)/2$ caused by HTK speech parametrization mechanism (w is the size and s shift of the parametrization window) is introduced into the resulting segmentation. Hence, the boundaries in the final segmentation are shifted by the offset o.

Alternatively, when some pre-segmented speech data are available (preferably by an expert in acoustic phonetics), the automatic segmentation can be adjusted by comparing it with the reference segmentation. In our approach, so-called *boundary-specific correction* was employed to correct a particular type of boundaries (i.e. boundaries between different types of phones, e.g. vowels and fricatives, etc.). In this case, a more accurate HMM initialization method, so-called *bootstrap* could be utilized to get slightly better segmentation results [4].

A detailed analysis of the results of the automatic segmentation revealed some erroneous insertion of short pauses. A short pause is often incorrectly aligned with the closure of a plosive or affricate at the beginning of a word [3]. To cope with these problems, a rule $\# + exp > dur$ was proposed. This rule defines the minimum combined duration dur of a short pause $\#$ and plosive or affricate exp. If the combined duration is lower than dur, then short pause # is removed from the segmentation. The minimum combined duration was set individually for each plosive or affricate using statistics of their durations in the speech corpus; we actually use $dur = q_{95}$, where q_{95} is a 0.95-quantile of a particular exp. In our system $q_{95} \approx 140$ ms for plosives and $q_{95} \approx 200$ ms for affricates.

3. Prosody generation

The intelligibility and naturalness of synthetic speech is highly influenced by its suprasegmental features – i.e. prosody. We have incorporated two different prosody models into our TTS system: *rule-based* and *data-driven*.

3.1. Rule-based approach

The rule-based approach to prosody modelling applies a set of rules derived from phonetic research and description of the suprasegmental speech phenomena. Our prosody model consists of 21 various parameters (coefficients) which have been set up experimentally. These coefficients control e.g. a baseline $F0$ contour, a slope of an overall declining melody tendency, shapes of all declining/ascending cadences, intensity modulations, duration changes, influence of word/sentence stress, etc. (see [5] for more details).

An input sentence is segmented into *prosodic clauses* (or *phonemic clauses*) which are further divided into *prosodic words*. A prosodic clause is an "intonationally coherent" part of a sentence and a prosodic word is a group of words subordinated to one word stress (accent). Since the first syllable of a prosodic word of the Czech language is in most cases the stressed one, prosodic words and stresses can be designated us-

ing several quite simple rules.

The melody modulation (within a prosodic clause) is based on the superimposing of an overall declining $F0$ tendency with prosodic-word-sized $F0$ patterns. These $F0$ patterns are called *cadences*. Our cadence inventory consists of three basic cadences named according to their overall $F0$ slope – flat, ascending, descending. Each cadence also has its stressed variant and the ascending and descending ones have another three variants distinguished by the magnitude of $F0$ increase or decrease respectively (we refer to this aspect of cadences as "tendency" – low, mid, high). This means we have 14 different cadences.

Each prosodic word except for the last one is assigned a flat cadence. The last one is assigned one of the ascending or descending cadences according to the modality of the sentence (indicative, interrogative, etc.).

The intensity modulation increases the volume of stressed syllables by a pre-set coefficient and the timing modulation increases the duration of stressed vowels as well as modifies the overall duration of the last prosodic word in a clause according to the number of its phones.

3.2. Data-driven approach

This approach is being developed due to the limitations (concerning speech naturalness) of the rule-based approach. The idea is to set up model parameters automatically, using real speech data from a corpus.

Based on sentences from the source speech corpus, the speech data are segmented into prosodic words by the same method used in the rule-based approach. This way we have about 50,000 prosodic words with detailed representation of $F0$ shapes of these words. The cadence inventory is then created using an *agglomerative clustering* algorithm which creates n clusters whose *centroids* (or other representatives) are considered to be cadences representing the variability of intonation patterns over prosodic words. We experiment with n ranging from 10 up to 200.

The text is segmented into abstract categories (prosodic clauses, prosodic phrases, prosodemes) describing its structural prosodic properties. For each prosodic word of a sentence to be synthesized, a vector representing these categories together with quantitative characteristics of the word (number of its phones and syllables, index of a stressed vowel, etc.) is estimated. On the basis of this vector, the appropriate cadence and the initial $F0$ value of the prosodic word is chosen (to be as similar as possible to the same sentence configuration occurring in the corpus).

Again, listening tests (16 listeners participated) were carried out to decide which approach produces better prosody for synthetic speech in terms of its naturalness. Both female (PrF) and male (PrM) voices were taken into account. The results in Fig. 2 (sections PrF and PrM) show that the data-driven model is quite significantly preferred. Moreover, the data-driven model is still in its very first version and we expect its further improvements.

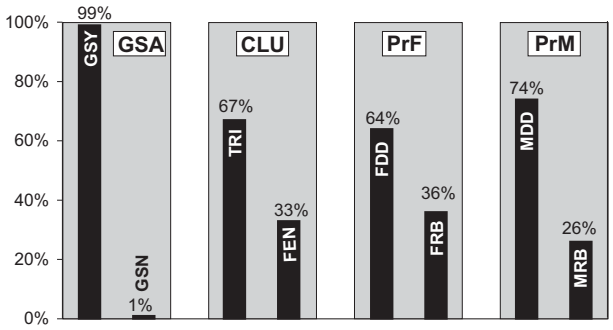

Figure 2: *The results of listening tests: glottal stop analysis (GSA) – using (GSY) or not using (GSN) glottal stop, clustering level (CLU) – triphones (TRI) or fenemes (FEN), prosody model for female voice (PrF) – rule-based (FRB) or data-driven (FDD), prosody model for male voice (PrM) – rule-based (MRB) or data-driven (MDD). Listeners' preferences are expressed in percentage.*

4. Unit selection

During synthesis, there is a need to modify the representatives of concatenated units in order to meet the requirements of the prosody generator (also called *target specifications*), especially duration and fundamental frequency contour. However, these modifications usually cause deterioration of the quality of generated speech, and therefore an approach called *unit selection* (or *selection-based synthesis*) has been researched. Its goal is to choose such a unit representative that is used for speech generation (when many candidates are available), which is the closest to target specifications (the closeness being reflected by *target cost*); and moreover, such a sequence of representatives has to result in spectrally smooth concatenation (being reflected by *concatenation cost*) [6].

Although ARTIC produces speech of very high intelligibility, it sounds slightly "buzzing-like". Therefore, we have been faced with the increase in naturalness, as a result of which we started to experiment with the unit selection approach. We used the corpus mentioned in Section 1, as it is quite large [2], recorded in news-style and precisely segmented (see Section 2.3).

As triphones have been successfully used in our baseline system, contrary to other research we decided to use triphones as basic units for unit selection. Owing to the fact they contain context information directly in their identifiers, target cost can aim at prosodic features only, such as duration, fundamental frequency, stress, etc. (in the case of phones, additional information connected with phonetic context must be taken into account). *Linear regression* is utilized to estimate weights for target cost, with the possibility of fine-tuning by hand using listening tests. As for concatenation cost, we use the same MFCC coefficients as those used for HMMs in the segmentation stage for the present, though we are aware of the lack of more sophisticated measuring for both regression training and concatenation cost measurement which would better correspond to human perception.

However, even with perfectly tuned unit selection, a certain amount of modification is still necessary: when speech corpus of limited size has to be used, the sequence of selected candidates does not match exactly the target and some candidates do not connect smoothly. There is some space for methods allowing high-quality modifications and high spectral smoothing, for example undermentioned harmonic/noise approach to speech production.

5. Harmonic/noise-based speech production

Besides all mentioned steps for the improvements of the synthetic speech, other speech signal generation schemas were also

dealt with as an alternative to the OLA speech synthesis method. *Harmonic plus noise modelling* (HNM) techniques [7] were found very promising in our goal of reaching high-quality synthetic speech. Moreover, using frequency domain modelling, it is possible to reach a higher compression rate for AUI storage (except the phase components).

In the common time-domain approaches the only implicit *phase* manipulation is employed by the "pitch-synchronous" processing of the waveforms. The phase incoherence problems arise not only at the concatenation points of the units but also on the inter-frame level. Those are usually perceived as artefacts disturbing the notion of speech fluency. The phase incoherence can be overcome by proper speech unit selection from a huge unit database. However, the requirement of the huge speech unit database is at variance with the intention of implementing quality speech synthesis on low-resource appliances (phones, handhelds, etc.).

For the purposes of designing a quality speech synthesis system (not only) for low-resource appliances, the HNM approach can be employed with a smaller unit database. It is also due to its capability to manipulate the phase components during the synthesis. In [8] we proposed the method of "phase substitutions" which allows us to dramatically reduce the storage space of phase components in the harmonic AUI.

In the proposed method, a small number of phase vectors (so-called *representative phase vectors*) are stored to the AUI. They are not only copies of the ones obtained by the analysis; during the off-line analysis, the starting candidate for the phase representative is selected from the spectrally stable part of the voiced segment in the unit and then it is extended by appending the additional elements to its end. The new elements are obtained by the analysis of the neighbouring frames in the analyzed segment [8]. The method stores only one phase vector for every uninterrupted voiced sequence of frames (voiced segment) instead of storing the phase vector for every frame. A dramatic reduction of the database storage space demands is achieved in this way. It corresponds with the rate n_{vs}/n_{vf} (where n_{vs} is a number of voiced segments and n_{vf} is a number of voiced frames). When using short speech units (like triphones) the voiced units mostly do not contain more than one uninterrupted sequence of the voiced frames. So the number of the voiced speech segments can be considered to be comparable to the number of the voiced speech units.

To give a concrete example, our AUI contains 6,258 speech units. 5,826 of them contain the sequence of the voiced frames. All the voiced segments consist of 95,132 frames. Instead of 95,132 phase vectors just 5,826 of their representatives are stored in AUI. For this particular database we save 93.9% of the space required for the storage of the phase vectors. The final size of "harmonic" AUI is for this example 8MB (without any further compression) comparing to 25MB size of original time domain AUI.

6. Towards multilinguality

After two Czech voices (male and female) were built on the principles described above, two other languages (Slovak and German) have been successfully implemented within the framework of ARTIC TTS system. In fact, the language modules differ mainly in the text processing component (i.e. phonetic transcription, prosody generation, etc.). The techniques for automatic AUI design and speech signal production remain the same as those proposed for the Czech language.

7. Conclusions

Several improvements to the baseline TTS system ARTIC have been discussed in this paper. Due to the complexity of such a system, only a brief explanation of every enhanced component was given. As for AUI, extending a Czech phonetic inventory with glottal stop was shown to increase the intelligibility of the synthetic speech. Model-level clustering during HMM-based AUI creation, enabling triphones to be used directly in speech synthesis, was evaluated to outperform state-level clustering. Some refinements of automatic speech segmentation were also presented. Two prosody generation models were further introduced to obtain more natural synthetic speech – the data-driven approach was evaluated as better than the rule-based one. Listening tests were carried out to assess the contribution of each individual component. Further, a unit selection scheme was proposed to achieve more fluent and distortion-free speech. Finally, the use of an alternative harmonic/noise-based speech signal generation method was discussed, especially in the context of the reduction of the AUI storage requirements by efficient modelling of the phase component.

In our next work we will continuously aim at improving the synthetic speech produced by our TTS system. All concepts elaborated here will be followed, especially those related to more precise segmentation, data-driven prosody modelling, unit selection and harmonic/noise-based speech production. More languages (especially Slavic) are also planned to be synthesized in the near future.

8. References

[1] Matoušek, J., Psutka, J., "ARTIC: a New Czech Text-to-Speech Synthesis System Using Statistical Approach to Speech Segment Database Construction", Proc. of ICSLP2000, Beijing, 2000, pp. 612–615.

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

[3] Donovan, R. E., Woodland, P. C., "A Hidden Markov-Model-Based Trainable Speech Synthesizer", Computer Speech and Language 13:223–241, 1999.

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

[5] Romportl, J., Matoušek, J., Tihelka, D.,"Prosody Model and its Application to Czech TTS System", Proc. of UkrOBRAZ 2002, Kyjiv, 2002, pp. 93–96.

[6] Hunt, A. J., Black, A. W., "Unit Selection in Concatenative Speech Synthesis System Using a Large Speech Database", Proc. of ICASSP'96, Atlanta, 1996, pp. 373–376.

[7] Stylianou, Y., "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis", IEEE Trans. Speech and Audio Proc., 9(1), 2001, pp. 21–29.

[8] Tychtl, Z., Matouš, K., "The Phase Substitutions in Czech Harmonic Concatenative Speech Synthesis", TSD2003, Springer Verlag, 2003, pp. 333-340.