German and Czech Speech Synthesis Using HMM-Based Speech Segment Database*

Jindřich Matoušek[1], Daniel Tihelka[1], Josef Psutka[1], and Jana Hesová[2]

[1] University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz, psutka@kky.zcu.cz
[2] University of West Bohemia, Department of Applied Linguistics,
Riegrova 11, 306 14, Plzeň, Czech Republic
jhesova@kaj.zcu.cz

Abstract. This paper presents an experimental German speech synthesis system. As in case of a Czech text-to-speech system ARTIC, statistical approach (using hidden Markov models) was employed to build a speech segment database. This approach was confirmed to be language independent and it was shown to be capable of designing a quality database that led to an intelligible synthetic speech of a high quality. Some experiments with clustering the similar speech contexts were performed to enhance the quality of the synthetic speech. Our results show the superiority of phoneme-level clustering to subphoneme-level one.

1 Introduction

This paper presents an experimental German speech synthesis system based on an automatically built speech unit database. In our previous work, such a system was successfully designed for the Czech language (see Figure 1) [2–4]. The synthetic speech of the Czech synthesizer sounds very intelligibly. This system can employ various concatenation-based speech synthesis techniques such as linear prediction (LP), PSOLA, Harmonic+Noise Model (HNM) or their combinations (e.g. LP-PSOLA). Being a concatenative speech synthesis system there is a need to employ a speech segment database (SSD) during the synthesis process. This database was built in a fully automatic way using a statistical approach based on modeling and segmentation of speech by Hidden Markov models (HMMs). Our experience in such kind of speech modeling and the language independent nature of the statistical approach let us try to apply the same speech segment database construction method to another language. During centuries our country (and especially the western part) has been influenced by our German spoken neighbors, there are relatively many German loanwords in Czech language, so German was selected for the first experiments with our non-native speech synthesis.

The paper is organized as follows. Section 2 describes the main differences between Czech and German languages in view of speech synthesis. In section

* This research was supported by the project no. MSM235200004 of the Ministry of Education of Czech Republic and the firm SpeechTech.

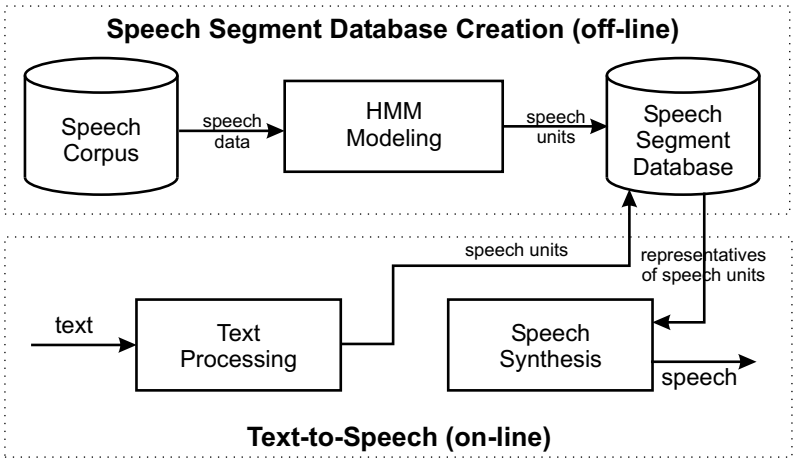

Fig. 1. A simplified scheme of a Czech TTS system which uses an automatically built speech segment database.

3 a detailed description of the baseline German text-to-speech (TTS) system is given. Section 4 then presents our experiments with clustering similar speech contexts and shows the success of statistical approach to SSD modeling for the German language. Finally, Section 5 contains the conclusion and outlines our future work.

2 The main differences between Czech and German

This section is dedicated to the main differences between Czech and German languages. Since there are so many differences resulting from the diverse nature of these languages (Czech is a Slavic language, German is a Germanic one), we will limit ourselves to the most important ones in view of speech synthesis [10].

One of the most important differences are phonetic features and phonetic transcription of both languages. Phonetic forms of Czech words are very similar to their orthographic forms. This is a feature common to all Slavic languages. Relatively simple phonetic transcription rules can be employed to convert orthographic form (i.e. letters) to phonetic form (i.e. phonemes). On the other hand, the correlation between phonemes and letters is much more complicated in German. The phonetic transcription rules are then more complex and more difficult to describe (e.g. the same phoneme /f/ is pronounced in different letter contexts as in words *Film, Philosophie, vier*, etc).

Due to very different nature of both languages there are many differences in their phonetic inventories, of course. In fact, German phonetic inventory is very large because there are many loanwords in German and the pronunciation of these non-German words introduces new phonemes to the German phonetic inventory. There are also relatively many loanwords in Czech but when pronounc-

ing these non-Czech words foreign phonemes are replaced by "the most similar" Czech ones. German has more vowels (including e.g. schwa) than Czech. Both languages distinguish short and long vowels. However the pronunciation is rather different. While most of German long vowels are closed and all short vowels open, all Czech vowels are pronounced rather in a neutral way. An example of different vowel systems is that Czech has just two vowels to pronounce the letters "*e, é*": short /e/ and long /e:/. In German there are five vowels to pronounce sounds similar to these two Czech vowels: /E, 9, e:, E:, 2:/ [5]. As for consonants, there are more consonantal phonemes in Czech language. Virtually all German consonants except for /C/ and /pf/ are present in Czech language too. Czech consonants are influenced by the characteristic voiced/unvoiced much more than German in which the intensity of consonantal pronunciation has more distinguish feature. There are also aspirated consonants /p, t, k/ in word-initial position in German. On the other hand no aspiration occurs in Czech. The German phonetic inventory used in our system is described in Section 3.1.

Some differences could be also found in prosodic features of both languages. Since prosody is ignored in the first version of our system, only the most distinct difference concerning stress will be mentioned. Contrast between stressed and unstressed syllables is more emphasized in German. German allows more reductions and elisions in unstressed positions. Stress is always on the first syllable in Czech. On the other hand, in German the stress is variable and is dependent on a word stem.

3 Baseline System

The baseline German speech synthesis system will be described in this section. The way the system was built was almost the same as in the case of a Czech TTS system ARTIC designed in our previous work [2–4].

3.1 Phonetic Inventory

When modeling or synthesizing speech the first step usually consists of defining the basic phonetic inventory of a language in focus. All speech units used in synthesis are then derived from this inventory. In our Czech TTS system 45 phonemes and significant allophones (including two kinds of pauses) were used [2]. The German phonetic inventory was defined in the extent of German SAMPA comprising 46 phonemes.

Some simplification was taken into account when defining the inventory. As mentioned in Section 2 there are relatively many loanwords in German. The pronunciation of these foreign words copies the pronunciation in the original language introducing new phonemes into German phonetic inventory (e.g. /æ/ in an English word *Cat*). If we respected this fact, the phonetic inventory would be augmented to include all phonemes possibly present in German pronunciations. To reduce the inventory size, domestic German phonemes were taken into account only. As in the case of Czech language the foreign phonemes were

replaced by phonetically the most immediate German phonemes. From similar reasons nasalized vowels loaned from French (e.g. /õ/ in *Fondue*) are not supported by our phonetic inventory. Non-syllabic vowels and syllabic consonants are also ignored.

3.2 Speech Corpus

Speech corpus consists of important speech material needed to model speech units and to create speech segment database. In case of our Czech TTS the corpus comprised a large number of sentences described by their orthographic and phonetic forms, speech waveforms, glottal signals and parametric representations [2, 3]. Glottal signals were measured by a device called electroglottograph and were used for the detection of the moments of principal excitation of vocal tract (usually the moments of glottal closure – so called pitch-marks) [4]. These pitch-marks are often used in contemporary standard speech synthesis techniques (e.g. PSOLA or some methods of harmonic synthesis). In [4] we experimented with speech corpus construction process and we found that a large number of carefully selected sentences spoken almost in a monotonous way and very precise pitch-mark detection are a need for a high-quality synthetic speech.

The German speech corpus was created under the same circumstances as the Czech one. The only exception was that no sentence selection was performed since limited amount of German text was available. About 6 000 German sentences were available in textual form. Some unsuitable sentences were excluded and all remaining 5 255 sentences were used. The comparison of Czech and German speech corpus is given in Table 1.

3.3 Speech Unit Modeling

The speech corpus is used as a basis for speech unit modeling. German speech units were modeled in the same way as the Czech ones, i.e. HTK system was used to model three-state left-to-right single-density crossword-triphone HMMs [2, 3, 8]. To make more robust models and to enable modeling triphones not present in the speech corpus, a clustering procedure is employed to tie similar triphones. This is very important for TTS synthesis since clustering ensures that an arbitrary triphone, i.e. arbitrary text, could be synthesized. In the baseline system the clustering was performed on model's state level. The more detailed information about modeling can be found in [2, 3].

3.4 Speech Segment Database

The resulting triphone HMMs were employed to segment the speech corpus into the basic speech units (in fact Viterbi search was realized to find the boundaries between these speech units in each sentence of the speech corpus). In the baseline system the basic speech unit represents a state of a crossword-triphone state-clustered HMM again. After segmenting there are many representatives

Table 1. A comparison of Czech and German speech corpora used for SSD construction. Number of clustered states is given for state-level clustering and number of triphones corresponds to model-level clustering.

	Czech	German
Number of sentences	5 000	5 255
Amount of speech data [min]	772	738
Number of phonemes	43	46
Number of clustered states	9 097	6 876
Number of triphones	6 258	4 687

(speech segments) of each speech unit in the speech corpus. The most representative segment of each unit is then selected and stored in the speech segment database. The same simple segment selection procedure as for Czech database was implemented [2, 3]. These representative segments are used in the synthesis stage.

3.5 Text-to-Speech

In fully automatic TTS process an arbitrary input text is converted to the corresponding output speech. This is the case of the Czech TTS system ARTIC where phonetic transcription rules are applied to the input orthographic text producing a sequence of phonemes. This sequence is then converted to a sequence of speech units (in the baseline system these units are the clustered states) and finally a concatenative speech synthesis technique is employed to join the speech units in the resulting speech [2]. In fact all standard concatenative techniques can be used. A time domain synthesis method is implemented in the baseline system in the time of writing this paper. No text-to-prosody module has been implemented so far, so the synthetic speech exhibits constant prosodic features and sounds in monotone.

Once again, the German text-to-speech process copies the Czech one. However, there is one significant exception. Since the implementation of precise German phonetic rules is very difficult and exceeds our knowledge of German, a phonetic dictionary was created manually. This dictionary includes all words present in the speech corpus now and may be arbitrarily augmented to cover more and more words. The text is then firstly segmented into words, which are then looked up in the dictionary. Simple phonetic rules were also proposed to phonetically transcribe words not found in the dictionary. The form of phonetic rules is the same as for Czech [9]. Here is an example of such a rule for German:

$$i \rightarrow i : /\_ \langle e \rangle \ . \tag{1}$$

This rule is applied e.g. in a word *diese* [di:z@].

4 Clustering Issues

Several experiments were made as an extension of the baseline system. They concerned mainly the fluency and the overall quality of the synthetic speech. Since the number of resulting speech units strongly depends on the clustering procedure (see Section 3.3), the attention was focused on clustering issues. Various clustering thresholds also affect clustering results and they were already analyzed in [3]. In research described in this paper the phoneme/subphoneme level of clustering was examined.

Although the baseline system produces a very intelligible high-quality speech, some audible glitches can degrade the speech. These glitches appear at unit boundaries, especially in long sections of voiced speech. These problems can be possibly minimized by using a parametric domain synthesis technique which enables controlling spectral features of speech, i.e. spectral smoothing especially at unit boundaries. However, time domain has also some advantages like almost no signal processing, i.e minimum degradation of speech. When still staying in the time domain, reducing the number of concatenation points is an alternative solution. The basic speech unit used in the baseline system is so-called clustered state (or feneme [1, 2]), which corresponds to a state of a state-clustered triphone HMM. On the signal level feneme represents a small subphoneme unit. It is a flexible unit that can effectively stand for an arbitrary speech context. However, concatenating such small units results in many concatenation points – possible discontinuity problems [7].

In our next research we tried to retain the same speech context quality modeling while reducing the number of concatenation points in synthesis. To do that, clustering was performed on model's level (in contrast to previous state-level modeling). The basic speech unit used was then the whole triphone (a phoneme-sized unit) and the number of concatenation points dramatically decreased to one point per phoneme. The same number of concatenation points is achieved also for diphones which are traditionally used in most of today's synthesis systems. Unlike diphones triphones take into account a context of both preceding and successive phonemes. The principles of these two different speech clustering processes are shown in Figure 2.

Let us to compare both kinds of clustering from the synthesis point of view. As for speech recognition tree-based clustering on the state level was referred to outperform clustering on the model level [6]. As for speech synthesis our expectations came true. Indeed, the synthetic speech of a triphone-based synthesis system sounds more naturally and fluently and with less intrusive elements. It was evaluated to be better than synthetic speech of feneme-based system by the listeners. Spectral discontinuities can be easily identified when using fenemes (see Figure 3 for the differences). Indeed, it seems that triphone-based synthetic speech is superior to feneme-based one. However, some more detailed listening tests should be performed to be sure about it. Maybe some problems with triphone-based synthesis can appear when synthesizing a "very unknown" speech context – a triphone very different from speech data available in training speech corpus. Fenemes may model some rare contexts more precisely.

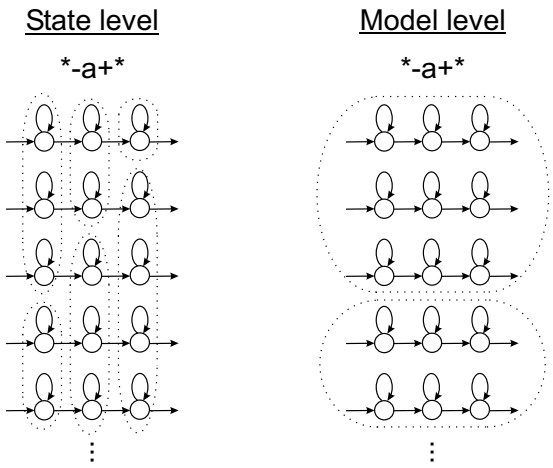

Fig. 2. An illustration of state-level (left) and model-level clustering (right) of all triphones derived from the phoneme /a/. Dotted lines show the examples of the resulting clusters.

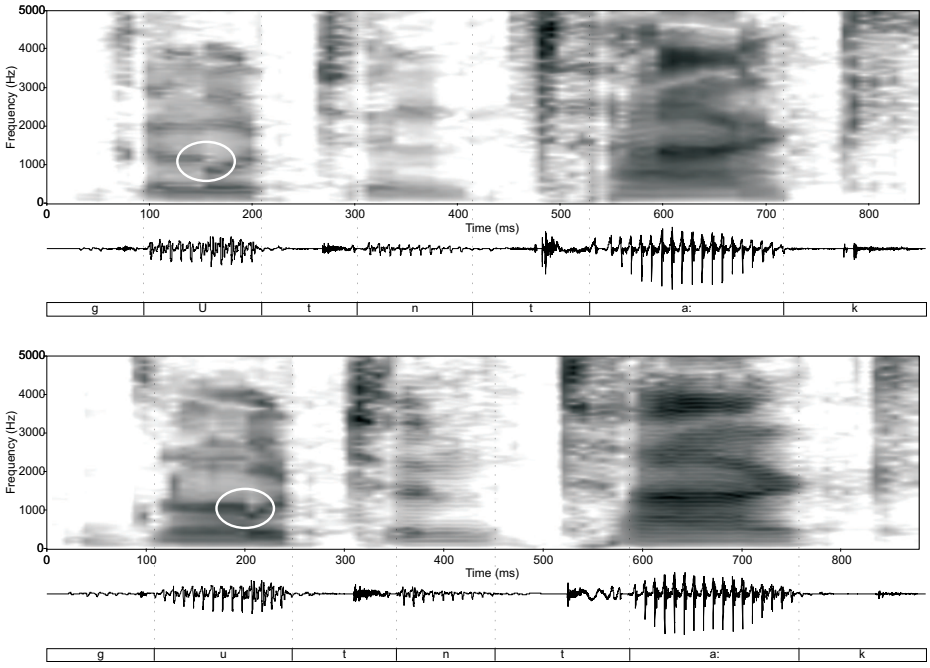

Fig. 3. A comparison of synthetic waveforms and spectrograms of a German sentence "*Guten Tag.*" when using fenemes (top) or triphones (bottom) as speech units. White oval shows an example of a within-phoneme formant discontinuity typical for feneme-based synthesis. Dotted lines show phoneme boundaries in the synthetic speech.

5 Conclusion and Future Work

In this paper a new experimental German TTS system was presented. When building the system we took advantage of our experience with the Czech TTS system. The system uses automatically designed SSD. The synthetic speech is of a high quality and sounds very intelligibly. So, HMM-based approach to SSD construction implemented in [1–3] was shown to be language independent. Experiments with level of clustering were also performed and clustering on phoneme level was judged to outperform subphoneme-level clustering from the synthesis point of view. Samples of synthetic speech are available on http://artin.zcu.cz/people/matousek/research/tts/TSD2002/samples.htm.

Since basic German synthesis system has been designed so far, there are many parts which could be improved. There is no doubt synthetic speech could be even better. Phonetic dictionary should be augmented to cover more words and/or more precise phonetic transcription rules should be defined to describe the pronunciation of words out of dictionary more properly. Some experiments with modeling could be also realized to achieve more precise automatic segmentation and synthesis. Various synthesis methods could be also examined to find out the method that would lead to the highest quality of synthetic speech. Of course, there are two important parts not taken into account so far: text preprocessing and prosody generation. These issues should be also managed to have a TTS system of the highest quality.

References

1. Donovan R.E., Woodland P.C.: A Hidden Markov-Model-Based Trainable Speech Synthesizer. Computer Speech and Language, 13. (1999) 223–241.
2. Matoušek J., and Psutka J.: ARTIC: a New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction. Proceedings of ICSLP2000, vol. IV. Beijing (2000) 612–615.
3. Matoušek J.: Text-to-Speech Synthesis Using Statistical Approach to Automatic Speech Segment Database Construction (in Czech). Ph.D. thesis, Pilsen (2001).
4. Matoušek J., Psutka J., and Krůta J.: On Building Speech Corpus for Concatenation-Based Speech Synthesis. Proceedings of Eurospeech2001, vol 3. AAlborg (2001) 2047–2050.
5. Gibbon D., Moore R., and Winski T.: Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter. Berlin (1997).
6. Young S.: Tree-Based State Tying for High Accuracy Acoustic Modelling. Proceedings of the ARPA Workshop on Human Language Technology. Plainsboro, New Jersey (1994) 307–312.
7. Hon H., Acero A., Huang X., Liu J., and Plumpe M.: Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems. Proceedings of ICASSP'98, vol. 1, Seattle (1998) 293–296.
8. Young S. et al.: The HTK Book. Entropic Inc. (1999).
9. Psutka J.: Communication with Computer by Speech (in Czech). Academia, Prague (1995).
10. Duden. Aussprachenwörterbuch (in German). Max Mangold, Duden-Verlag, vol. 6, Manheim (1990).