Czech HMM-Based Speech Synthesis*

Zdeněk Hanzlíček

Department of Cybernetics, University of West Bohemia,
Univerzitní 8, 306 14, Pilsen, Czech Republic
zhanzlic@kky.zcu.cz
http://www.kky.zcu.cz/en

Abstract. In this paper, first experiments on statistical parametric HMM-based speech synthesis for the Czech language are described. In this synthesis method, trajectories of speech parameters are generated from the trained hidden Markov models. A final speech waveform is synthesized from those speech parameters. In our experiments, spectral properties were represented by mel cepstrum coefficients. For the waveform synthesis, the corresponding MLSA filter excited by pulses or noise was utilized. Beside that basic setup, a high-quality analysis/synthesis system STRAIGHT was employed for more sophisticated speech representation. For a more robust model parameter estimation, HMMs are clustered by using decision tree-based context clustering algorithm. For this purpose, phonetic and prosodic contextual factors proposed for the Czech language are taken into account. The created clustering trees are also employed for synthesis of speech units unseen within the training stage. The evaluation by subjective listening tests showed that speech produced by the combination of HMM-based TTS system and STRAIGHT is of comparable quality as speech synthesised by the unit selection TTS system trained from the same speech data.

Keywords: HMM-based speech synthesis, TTS, Czech language.

1 Introduction

Nowadays beside concatenative unit selection method, HMM-based speech synthesis [1] is one of the most researched synthesis methods. In this synthesis method, hidden Markov models (or possibly other parametric models) are trained from natural speech database. Spectral parameters, fundamental frequency, duration and eventually some excitation parameters are modelled simultaneously by the corresponding HMMs. For a more robust model parameter estimation, HMMs are clustered by using decision tree-based context clustering algorithm. For this purpose, phonetic and prosodic contextual factors are taken into account. They respect phonetic, prosodic and linguistics characteristics of given language. The created clustering trees are also employed for synthesis

* This work was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989 and by the Ministry of Education of the Czech Republic, project No. 2C06020. Author would also like to thank Prof. Hideki Kawahara from Wakayama University for his permission to use the STRAIGHT analysis/synthesis method [2]. The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is also highly appreciated.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 291–298, 2010.

of speech units unseen within the training stage. During synthesis, trajectories of all speech parameters are generated from these trained models in the maximum likelihood sense. A final speech waveform is reconstructed from those speech parameters.

Originally, this method was developed for Japanese. However, as well as other synthesis methods, it is mainly language independent. Thus, TTS systems for many other languages (e.g. English [3]) have been successfully implemented. For a more detailed language listing see e.g. [1]. The expansion of that synthesis method was possible thanks to the HTS toolkit [4] that provides statistical methods for HMM manipulation (including training, context clustering, parameter trajectory generation etc.).

This paper describes first experiments on statistical parametric HMM-based speech synthesis for the Czech language. For building of our experimental TTS system, we also employed the HTS toolkit. Within our experiments, two different speech analysis/synthesis methods and speech representations were compared:

1. A simple representation by Mel cepstrum coefficients. Methods for parameter extraction and speech synthesis by using MLSA filter were provided by SPTK toolkit [5].
2. More sophisticated speech representation by the high-quality analysis/synthesis method STRAIGHT [2]. It has been already used in HMM-based speech synthesis framework, e.g. by Zen et al. [6].

A large listening test was organized for the evaluation and comparison of various settings of our experimental TTS system. Results showed that speech produced by the combination of HMM-based TTS system and STRAIGHT is of comparable quality as speech synthesised by the unit selection TTS system trained from the same speech data.

The paper is organized as follows. Section 2 deals with the phonetic and prosodic characteristics of the Czech language. According to them, a suitable set of contextual factors is proposed for the purposes of Czech HMM-based speech synthesis. In Section 3 a brief description of HMM-based speech synthesis system and its settings for our experiments are presented. Results of performed listening tests are shown in Section 4. Finally, Section 5 summarizes the paper and outlines our future work.

2 Czech Language Characteristics

For naturally sounding synthesized speech, phonetic and prosodic characteristics of a particular language should be taken into account. In HMM-based speech synthesis method, these language characteristics are respected by definition of so called contextual factors. Then, a speech unit is given as a phoneme with its phonetic and prosodic context information. In this manner, the language prosody is implicitly modelled, because for various contexts different units/models can be used. Contextual factors for the Czech language are summarized in Section 2.3.

2.1 Phonetic Characteristics

The set of Czech phones is defined in Table 1. In our experiments, phonemes from the basic set were used. In addition we also employed glottal stop `[?]`, inter-word pause

[#] and long silence [$]. Other allophones listed in Table 1 could also be utilised. However they usually correspond to basic phonemes in a special phonetic context, thus in a system with context depended units, allophones and basic phonemes are implicitly distinguished.

Table 1. Czech phonetic inventory used in our TTS system (in SAMPA [7] 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=], [@]

Sometimes, the syllable is considered to be an alternative phonetic unit to the phone in the Czech language. Syllabification and syllable utilisation in context of concatenative speech synthesis were researched e.g. in [8]. Though, they seem not to be suitable basic units for speech synthesis, the information on syllable boundaries in text should be taken into account, because it has obviously some influence on (human) speech production.

2.2 Prosodic Characteristics

The prosodic structure of the Czech language can be described with the prosodic phrase grammar [9,10] which defines a hierarchical tree structure above a synthesised utterance. The following functionally relevant structures are distinguished:

- *Prosodic sentence* – syntactically consistent unit that usually corresponds to the whole utterance.
- *Prosodic clause* – linear unit in speech delimited by pauses.
- *Prosodic phrase* – segment of speech containing a certain continuous intonation scheme.
- *Prosodeme* – a rather abstract unit describing communication function. In the Czech language, this function is usually connected with the last prosodic word in the phrase. For other prosodic words, a formal null prosodeme is defined. For the complete prosodeme description see [9].
- *Prosodic word* – group of words belonging to one stress, often considered as a basic rhythmic unit.
- *Semantic accent* – emphasis of a prosodic word.

For the present, the prosodic phrases and semantic accents were not employed in our experimental setup. Their detection and modelling is not an easy task. However, a statistical method for their assignment in speech data has been already developed and described in [11]. Application of prosodic phrases and semantic accents is planned in future experiments.

2.3 Contextual Factors

Regarding the characteristics of the Czech language, a set of suitable contextual factors was defined. It is presented in Table 2. In comparison with other languages, e.g. English [3], this set is very reduced. However, greater amount of contextual factors or also greater amount of their possible values result in higher computational requirements. Thus within our primary experiments, we decided for a reduced set of factors. Significance of particular contextual factors is planned to be researched in the future.

Table 2. Contextual factors

Factor	Possible values
Previous and next phoneme	see Table 1
Phone location in syllable	first, inner, last, single
Syllable location in prosodic word	
Prosodic word location in clause	
Clause location in sentence	
Prosodeme type	terminating satisfactorily, terminating unsatisfactorily, nonterminating, null

3 HMM-Based TTS System

A thorough description of an HMM-based TTS system, including utilised statistical methods, appeared in many publications, e.g. [1]. This section gives only a brief overview, because these methods are not the object of our contribution. For building of our experimental HMM-based TTS system, the following tools were utilised

- tools for speech analysis and reconstruction
  - SPTK - Speech Signal Processing Toolkit [5]
  - STRAIGHT - Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram [12]
- tools for HMM manipulation (training, parameter generation etc.)
  - HTS - HMM-based Speech Synthesis System [4]

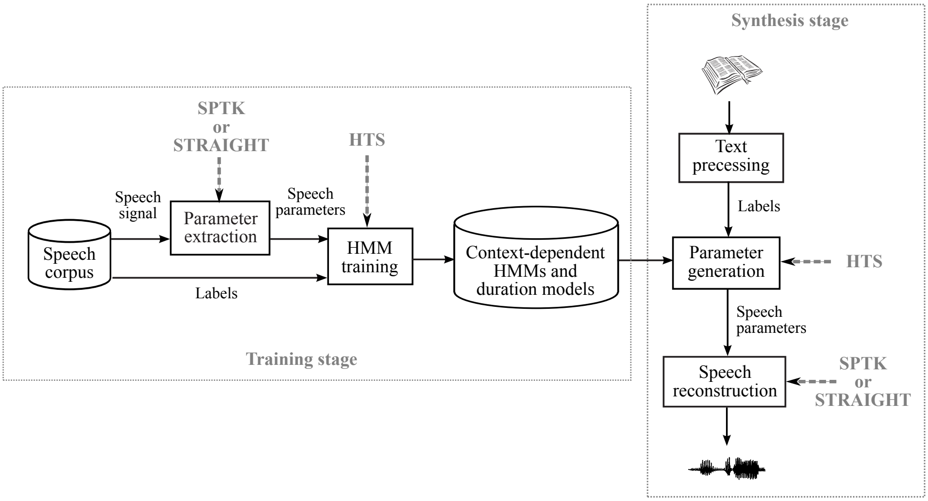

Fig. 1. A simplified scheme of our HMM-based TTS system

Training stage can be roughly divided into 3 main parts:

1. **Parameter extraction** – speech signal was sampled at 16 kHz.
   - For the mel cepstral analysis a 25 ms Blackman window with 5 ms shift was employed. Each speech frame was represented by a composed parameter vector containing 25 mel cepstral coefficients and F_0 value with their delta and delta-delta.
   - The STRAIGHT analysis method used Gaussian F_0 adaptive window with 5 ms shift. Composed parameter vector contained 40 mel cepstral coefficients, F_0 value and 5 aperiodicity coefficients, again with their delta and delta-delta.
2. **Model training** – model parameters are estimated from speech data by using maximum likelihood criterion. First, robust models for particular single phoneme are trained. Then, models for all particular combinations of contextual factors within training data are estimated.
3. **Context clustering** – for a more robust model parameter estimation, clustering of contextual factors is performed.

In the synthesis stage, trajectories of speech parameters are generated directly from the trained HMMs. Clustering trees from the training stage are utilised to find a suitable substitute for units unseen in training data. The final speech waveform is reconstructed from the generated parameters by using appropriate synthesis methods, i.e. MLSA filter excited by pulses/noise or STRAIGHT-based vocoding.

4 Experiments and Results

For our experiments, two different voices (male and female) were employed. Both were professional speakers with broadcast experiences. Speech data were originally recorded for purposes of a concatenative TTS system utilizing unit selection method [13].

One large MOS (mean opinion score) listening test was conducted to evaluate the quality of speech produced by our experimental TTS system. In this test, all combinations of speaker (male or female), speech representation (by using SPTK or STRAIGHT) and amount of training data (10 minutes, 1 hour and 5 hours without pauses) appeared. In addition, natural utterances from source speech corpus and utterances produced by a concatenative TTS system were also mixed in this test. The concatenative TTS system [10], that employed unit selection synthesis method, was trained from the same speech data.

18 listeners took part in this test, they listened to single utterances (96 sentences in sum) and evaluated them according to the overall quality (acceptance) by using the standard MOS scale:

1. bad quality
2. poor quality
3. fair quality
4. good quality
5. excellent quality

Table 3. MOS test results (mean score ± standard deviation)

Speech generation method	Training data amount	Notation	Score	
			Male	Female
Natural speech		NAT	4.76 ± 0.06	4.04 ± 0.19
MLSA + pulses/noise	10 minutes	M–10m	2.20 ± 0.24	1.70 ± 0.29
	1 hour	M–1h	2.62 ± 0.34	2.18 ± 0.31
	5 hours	M–5h	2.73 ± 0.39	2.29 ± 0.48
STRAIGHT	10 minutes	S–10m	2.75 ± 0.21	1.68 ± 0.34
	1 hour	S–1h	3.56 ± 0.24	2.83 ± 0.15
	5 hours	S–5h	3.83 ± 0.28	3.38 ± 0.31
Unit selection		US	3.57 ± 0.62	3.14 ± 0.66

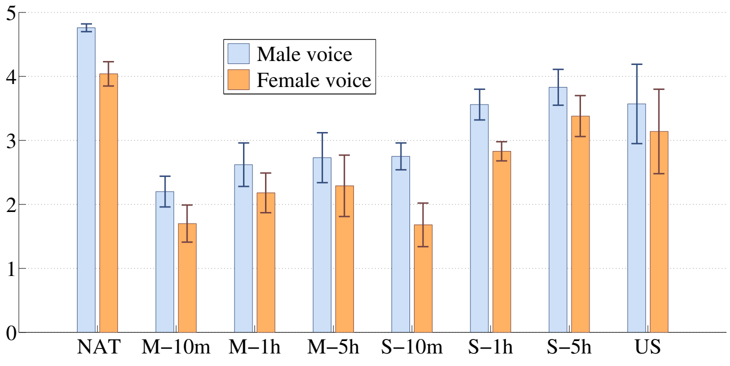

Fig. 2. MOS test results

The results of that test are presented in Table 3 and Figure 2. Expectably, speech quality increased with the amount of training data. A complex excitation modelling by STRAIGHT proved also very significant for quality perception. These findings are in accordance with other research works, e.g. [6]. For a comparable amount of training data HMM-based synthesis system with STRAIGHT produces speech of similar quality as system with unit selection method.

5 Conclusion

In this paper, first experiments on statistical parametric HMM-based speech synthesis for the Czech language were presented. For building an experimental TTS system, HTS toolkit was utilised. For speech representation, two different speech analysis/synthesis methods were used: Mel cepstral analysis + synthesis by using MLSA filter and STRAIGHT. The evaluation by subjective listening tests showed that speech produced by the HMM-based TTS system with STRAIGHT is of comparable quality as speech synthesised by the unit selection TTS system trained from the same speech data.

5.1 Future Work

In our future experiments, we will mainly focus on two important task:

- *Contextual factors* - in our experiments only a simple set of contextual factors was employed. The influence of other prosodic and linguistics characteristics of the Czech language (e.g. prosodic phrase and semantic accent) should be also analysed.
- *Excitation representation* - the comparison of MLSA filter excited by pulses or noise and STRAIGHT synthesis method confirmed that the proper excitation modelling has a significant influence on resulting speech quality. Thus we plan to research more methods and models for excitation representation (e.g. ML excitation method [14]).

References

1. Zen, H., Tokuda, K., Black, A.W.: Statistical Parametric Speech Synthesis. Speech Communication 51, 1039–1064 (2009)
2. Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A.: Restructuring Speech Representations using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication 27, 187–207 (1999)
3. Tokuda, K., Zen, H., Black, A.W.: An HMM-based Speech Synthesis System Applied to English. In: Proc. of IEEE Workshop on Speech Synthesis, pp. 227–230 (2002)
4. HMM-based Speech Synthesis System (HTS), `http://hts.sp.nitech.ac.jp`
5. Speech Signal Processing Toolkit (SPTK), `http://sp-tk.sourceforge.net`
6. Zen, H., Toda, T., Nakamura, M., Tokuda, K.: Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. IEICE Transactions on Information and Systems E90-D, 325–333 (2007)
7. Czech SAMPA, `http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm`

8. Matoušek, J., Hanzlíček, Z., Tihelka, D.: Hybrid Syllable/Triphone Speech Synthesis. In: Proc. of Interspeech 2005, Lisbon, Portugal, pp. 2529–2532 (2005)
9. Romportl, J., Matoušek, J., Tihelka, D.: Advanced Prosody Modelling. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 441–447. Springer, Heidelberg (2004)
10. Tihelka, D., Matoušek, J.: Unit Selection and its Relation to Symbolic Prosody: A New Approach. In: Proc. of Interspeech 2006 – ICSLP, Pittsburgh, Pennsylvania, vol. 1, pp. 2042–2045 (2006)
11. Romportl, J.: Prosodic Phrases and Semantic Accents in Speech Corpus for Czech TTS Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 493–500. Springer, Heidelberg (2008)
12. STRAIGHT, a speech analysis, modification and synthesis system, `http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html`
13. Matoušek, J., Romportl, J.: Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 326–333. Springer, Heidelberg (2007)
14. Maia, R., Toda, T., Zen, H., Nankaku, Y., Tokuda, K.: An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modelling. In: Proc. of 6th ISCA Workshop on Speech Synthesis, pp. 131–136 (2007)