Czech HMM-Based Speech Synthesis: Experiments with Model Adaptation*

Zdeněk Hanzlíček

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
zhanzlic@kky.zcu.cz

Abstract. This paper describes some experiments on model adaptation for statistical parametric speech synthesis for the Czech language. For building an experimental TTS system, HTS toolkit was utilised. Speech was represented by using high-quality analysis/synthesis system STRAIGHT. For definition of speech unit context, a new reduced set of contextual factors was proposed. During model clustering, some missing contextual factors, that were not included in this set, can be simulated by using combined context-related clustering questions. The model transformation was performed by a combination of CMLLR and MAP adaptation. Speech data from 3 male and 3 female speakers was used in our experiments. In the performed listening test, speech generated from regularly trained and adapted models was compared. Both voices were evaluated as identical and of a similar quality.

Keywords: HMM-based speech synthesis, speaker adaptation.

1 Introduction

Nowadays, statistical parametric (HMM-based) speech synthesis [1] is one of most researched synthesis methods. A great advantage of this method is the possibility to generate new voices by an adaptation of models trained for another speaker or even of models trained by using data from several different speakers [2]. Many adaptation methods have been developed – for an overview see [3].

Some basic experiments on HMM-based speech synthesis applied to the Czech language was already presented in [4]. This paper describes some consecutive experiments on model adaptation, which was performed by using a combination of 2 methods: CMLLR (constrained maximum likelihood linear regression) and additional MAP (maximum a posteriori) adaptation.

* This work was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989 and by the Technology Agency of the Czech Republic, project No. TA01011264. Author would also like to thank Prof. Hideki Kawahara from Wakayama University for his permission to use the STRAIGHT analysis/synthesis method [5]. The access to the MetaCentrum computing facilities, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic, is appreciated.

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 107–114, 2011.

For speech representation the analysis/synthesis method STRAIGHT [5] was utilized. Our experimental TTS system was built by using the well-known HTS toolkit [7]. Prosodic and linguistic characteristics of particular language are captured in a rich context of context-dependent units (models). Speech or language properties taken into account are called contextual factors. For the Czech language, we define a new reduced set of those factors.

For a more robust model parameter estimation, models are clustered using a decision tree-based context clustering algorithm. This process is controlled by simple context-related questions. To compensate the absence of some factors in our reduced set, we proposed the combined clustering questions which test the combinations of more factors.

Speech produced by regularly trained and adapted models was compared in a listening test. Speech data from 3 male and 3 female speakers were used. Results showed that speech produced by adapted models nearly of the same quality as speech generated from regularly trained models. Moreover, both voices were evaluated as identical.

The paper is organized as follows. In Section 2 a description of our HMM-based speech synthesis system and its settings are presented. Experimental evaluation is presented in Section 3. Finally, Section 4 summarizes the paper and outlines our future work.

2 System Overview

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

- **STRAIGHT** (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed, version 4.0) [6]
- **SPTK** (Speech Signal Processing Toolkit, version 3.3) [8]
- **HTS** (HMM-based Speech Synthesis System, version 3.4.1) [7]

2.1 Training Stage

Training stage can be roughly divided into 3 main parts:

1. **Parameter extraction** – Speech signal was sampled at 16 kHz. STRAIGHT analysis method used Gaussian F_0 adaptive window with 5 ms shift. Composed parameter vector contained 40 mel cepstral coefficients, logF_0 value and 5 band aperiodicity coefficients, again with their delta and delta-delta.
2. **Model training** – Model parameters were estimated from speech data by using maximum likelihood criterion. We employed 5-state left-to-right MSD-HSMM with single Gaussian output distributions. First, robust models for particular monophones are trained. Then, context-dependent models are derived and retrained. The definition of unit context for our experiments is described in Section 2.4.
3. **Context clustering** – For a more robust model parameter estimation, context clustering based on MDL (Minimum Description Length) criterion was performed. Decision trees were separately constructed for cepstral, logF_0, aperiodicity and duration parts of models.

2.2 Adaptation

A lot of modification and combination of basic adaptation methods (MLLR and MAP) have been developed – see e.g. [3]. For our experiments, we select CMLLR (constrained maximum likelihood linear regression) combined with additional MAP (maximum a posteriori) adaptation.

During the adaptation, multiple linear transformation functions are estimated by using speech data from target speaker. Since it is not possible to estimate a transform for each (clustered) context-dependent model, each transform is usually shared by a group (class) of related models.

In our experiments, classes sharing one transform were derived from context-independent units, i.e. models containing the same central monophone were adapted by using the same transform.

2.3 Synthesis Stage

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 models which are not available (were not trained). The final speech waveform is reconstructed from the generated parameters by using STRAIGHT-based vocoding.

2.4 Contextual Factors

In the HMM-based speech synthesis method, the phonetic and prosodic characteristics of a given language are respected by the specification of so called contextual factors. A speech unit (and the corresponding model) is given as a phone with its phonetic and prosodic context information. In this manner, the language prosody is modelled implicitly – in various contexts different units/models can be used. The context description is usually very rich [9]. Contextual factors are mostly defined as

- the position of the current phone/syllable/word in the parent syllable/word/ phrase
- the length of the current syllable/word/phrase in phones/syllables/words etc.

For a richer context description, the prosodic properties of speech are more precisely captured in the models. On the other hand, for a greater amount of contextual factors and wider range of their values, more training data is necessary to ensure a sufficient occurrence of all feasible combinations of defined factors.

In [4] a basic set of contextual factors was proposed. Contrary to other languages [9], this set was quite reduced. Based on some informal listening test, a new set of contextual factors was designed.

It encompasses the following prosodic features:

- *Prosodic clause* – a linear segment of speech delimited by pauses.
- *Prosodeme* – a rather abstract unit describing communication function. In the Czech language, it is usually connected with the last prosodic word in the phrase. In our experiments, only 4 main prosodeme types were distinguished: terminating satisfactorily (TS), terminating unsatisfactorily (TU), non-terminating (NT) and a formal null prosodeme.

– *Prosodic word* – a group of words belonging to one stress, often considered as a basic rhythmic unit.

For a more detailed description of Czech prosody see e.g. [10]. Compared to the default factor set proposed in [4], information on syllable boundaries was excluded from our new set. Our informal experiments revealed a low influence of that information. Moreover, the syllabification is ambiguous for the Czech language, in some cases syllable margins cannot be strictly determined. Thus, most contextual factors based on syllables would not be very precise. A more thorough study on the significance of particular contextual factors is planned to be performed in the future.

The set of contextual factors is summarized in Table 1. All factors, that define positions within the prosodic structure, were limited to values between 1 and 4, further positions are denoted 5+. The main motivation for this is the presumption that only several first marginal positions are prominent. Positions deeper inside the units become less distinguishable and the accurate position determination is supposed to be irrelevant.

A context-depended unit (or model) can be represented by a string

$$a_1\text{–}a_2\text{+}a_3\text{@P:}b_1\text{\_}b_2\text{@W:}c_1\text{\_}c_2/d_1$$

where all subscripted lower case letters are contextual factors defined in Table 1. The other characters in this string help to refer to particular factors (e.g. during model clustering).

Table 1. Contextual factors

Factors		Possible values
a_1, a_2, a_3	Previous, current and next phoneme	Czech phoneme set (see e.g. [4])
b_1, b_2	Phone position in prosodic word (forward and backward)	1, 2, 3, 4, 5+
c_1, c_2	Prosodic word position in clause (forward and backward)	
d_1	Prosodeme type	TS, TU, NT, null

2.5 Combined Context-Related Clustering Questions

The clustering algorithm utilises predefined set of context-related questions to build a decision-tree. By definition of more complex questions, some contextual factors, that were not included in the basic set, can be partly simulated, e.g.

– Phone position in the clause can be compensated by a combination of *phone position in the prosodic word* and *prosodic word position in the clause*. Obviously, only several marginal positions (related with the first and last word in the clause) can be reasonably defined this way. However consistently with position values definition in Table 1, only the marginal positions are prominent and are beneficial to be determined accurately.

– *Prosodic word length in phones* can be determined by combination of forward and backward *phone position in the prosodic word*. The accurate length can be determined only for words of length 1–4 phones. For longer words, at least one of position takes the value 5+, therefore the accurate length cannot be determined from contextual factors of one unit. However, the marginal values are expected to be most important again. *Clause length in prosodic words* can be expressed analogously.

A more thorough study on using such composed context-related questions for model clustering will be performed in the future. In our current experiments, the aforementioned simple combinations were employed.

3 Experiments and Results

3.1 Experimental Data Description

For our experiments speech data from 6 different speakers were utilised. This data was originally recorded for the purposes of a unit selection TTS system [11]. Thus, the overall quality was guaranteed. For simplification, male speakers are denoted M_{AJ}, M_{JS}, M_{TF} and female speakers F_{KI}, F_{MR}, F_{PP} – subscripted letter are initials of their names.

For our experiments, we selected one hour of speech from each speaker. Though the quality rises with the amount of training data, our previous experiments [4] showed that one hour of data is enough for synthesised speech of an acceptable quality.

Since recorded utterances from speakers M_{AJ}, M_{JS}, F_{KI} and F_{MR} were equal, we decided to use those speakers for training of initial models for adaptation. Thus, the differences in results for particular speakers should be caused by the variance between their voices and not by the selection of training utterances. Data from remaining speakers were employed for adaptation.

For pragmatic reasons, were preferred adaptation between speakers of the same gender, i.e. female-to-female or male-to-male. No cross-gender adaptation was performed in our experiments.

Equal utterances from all speakers should be also ideal for training of an average voice [2] – both gender dependent and independent. However, our informal experiments did not reveal any noticeable improvement – probably more data from more speakers should be used. Thus, we decided for a simpler experimental setup with one-speaker initial models. More thorough average voice experiments are planned in the future.

3.2 Quality Evaluation

In the first test, the quality of speech synthesised from adapted HMMs was evaluated. Two main questions were inspected

1. Naturally, the overall quality of speech generated from adapted models is expected to be lower when compared to speech synthesised from regularly trained models. How significant is the quality degradation caused by the adaptation process?

2. In case the amount and quality of speech data is sufficient, a regular model training could be performed. How distinctive is the difference between speech trained and adapted from the same amount of data?

To answer the first question, 1 hour of pure speech data (i.e. computed without pauses) was used to train models for particular speakers. Then, models were adapted by 10 minutes of speech from another speaker. In the test, participants listened to pairs of utterances generated from models

- trained from 1 hour of data from speaker X
- trained from 1 hour of data from another speaker and adapted with 10 minutes of data from speaker X

Listeners should select an utterance which of higher quality. The following 5-point scale was used:

1. utterance A is better than B
2. utterance A is slightly better than B
3. both utterances are similar
4. utterance A is slightly worse than B
5. utterance A is worse than B

The results are presented in upper half of Table 2. Utterances synthesised from regularly trained models were mostly evaluated as equal or slightly better. However, some listeners nearly continually preferred the adapted voice.

10 minutes of speech, that were utilised for the adaptation, could be also used for an independent training of a new model set. Naturally, no quality results could be expected from such a low amount of training data. However, we wanted to know whether the quality will be really poor or still acceptable. Thus, we synthesised pairs of utterances by using models

- trained from 10 minutes of data from speaker X
- trained from 1 hour of data from another speaker and adapted with 10 minutes of data from speaker X

Those pairs of utterances were inserted into the aforementioned test. The results are presented in the lower part of Table 2. Speech synthesised by using 10 minutes of training data was mostly evaluated as slightly worse than speech generated from adapted models. However, some utterances were evaluated as really worse and some as qualitatively equal.

3.3 Voice Identity Evaluation

Within the HMM-based speech synthesis framework, the only purpose of model adaptation is a change of voice identity. Thus, the similarity between the adaptation data and the synthesised speech . However, the comparison between regularly trained and adapted models is maybe more useful, because these are two main alternatives for obtaining a new voice identity. Since we wanted to prove that voice obtained by model

Table 2. Comparison of speech quality

Compared utterances (A – B)		Score
regular training	adaptation (10 minutes)	(mean ± std)
M_{TF} (1h)	$M_{AJ} \rightarrow M_{TF}$	2.91 ± 0.83
	$M_{JS} \rightarrow M_{TF}$	2.46 ± 0.67
F_{PP} (1h)	$F_{MR} \rightarrow F_{PP}$	2.75 ± 0.72
	$F_{KI} \rightarrow F_{PP}$	2.88 ± 0.66
M_{TF} (10m)	$M_{AJ} \rightarrow M_{TF}$	3.88 ± 0.86
	$M_{JS} \rightarrow M_{TF}$	3.97 ± 0.73
F_{PP} (10m)	$F_{MR} \rightarrow F_{PP}$	4.01 ± 0.71
	$F_{KI} \rightarrow F_{PP}$	3.76 ± 0.89

adaptation is close to the voice obtained by regular training, we decided for a preference test with corresponding setup.

11 participants took part in our test. They listened to 12 pairs of utterances and evaluated their similarity according the following scale

1. both voices are identical
2. voices are very similar
3. voices are slightly similar
4. voices are totally different

Again 1 hour of speech was used for training of initial models and 10 minutes for their adaptation. The results are presented in Table 3. Notation M_X or F_X means, that results are calculated for both male (M_{AJ} and M_{JS}) or female (F_{MR} and F_{KI}) speakers. In most cases, both voices were evaluated as equal or rarely as very similar.

Table 3. Comparison of voice identity

Compared utterances		Score
regular training (1 hour)	adaptation (10 minutes)	(mean ± std)
M_{TF}	$M_X \rightarrow M_{TF}$	1.26 ± 0.31
F_{PP}	$F_X \rightarrow F_{PP}$	1.18 ± 0.25

4 Conclusion and Future Work

In this paper, first experiments on speaker adaptation for HMM-based speech synthesis for the Czech language were presented. For building an experimental TTS system, HTS toolkit was utilised. For speech representation, STRAIGHT analysis-synthesis methods was used.

Our experiments proved, that for a small amount of training data (10 minutes) the adaptation of a different speaker's model set is preferable to the regular training of a

fully new model set. Moreover, listening tests also revealed that speech generated from adapted models is of a similar quality as speech produced by models regularly trained by the same amount of speech data as the initial models for the adaptation. The difference in voice identity was also appreciated as insignificant.

In our future experiments, we will mainly focus on using speech data which is generally problematic to employ in the speech synthesis, e.g. data recorded in a common environment by non-professional speakers. Our aim is to be able to utilise speech data uttered in a spontaneous style. This would significantly increase the amount of voices that could be synthesised.

References

1. Zen, H., Tokuda, K., Black, A.W.: Review: Statistical parametric speech synthesis. Speech Communication 51, 1039–1064 (2009)
2. Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Guan, Y., Oura, K., Tokuda, K., Karhila, R., Kurimo, M.: Thousands of Voices for HMM-Based Speech Synthesis. In: Proceedings of Interspeech 2009, pp. 420–423 (2009)
3. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Transactions on Audio, Speech, and Language Processing 17, 66–83 (2009)
4. Hanzlíček, Z.: Czech HMM-Based Speech Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 291–298. Springer, Heidelberg (2010)
5. 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. In: Speech Communication, vol. 27, pp. 187–207 (1999)
6. STRAIGHT, a speech analysis, modification and synthesis system, `http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html`
7. HMM-based Speech Synthesis System (HTS), `http://hts.sp.nitech.ac.jp`
8. Speech Signal Processing Toolkit (SPTK), `http://sp-tk.sourceforge.net`
9. Tokuda, K., Zen, H., Black, A.W.: An HMM-based Speech Synthesis System Applied to English. In: Proceedings of IEEE Workshop on Speech Synthesis, pp. 227–230 (2002)
10. 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)
11. 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)