Acta Cybernetica **19** (2010) 715–731.

Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis*

Bálint Tóth† and Géza Németh†

Abstract

Statistical parametric, especially Hidden Markov Model-based, text-to-speech (TTS) synthesis has received much attention recently. The quality of HMM-based speech synthesis approaches that of the state-of-the-art unit selection systems and possesses numerous favorable features, e.g. small runtime footprint, speaker interpolation, speaker adaptation. This paper presents the improvements of a Hungarian HMM-based speech synthesis system, including speaker dependent and adaptive training, speech synthesis with pulse-noise and mixed excitation. Listening tests and their evaluation are also described.

Keywords: Hungarian HMM speech synthesis, speaker adaptation, pulse-noise excitation, mixed excitation

1 Introduction

Several TTS methods were created in the last decades, including rule based articulatory [1] and formant synthesis [2], which try to model the speech production mechanism; diphone, triphone based concatenative synthesis [3] and corpus-based unit selection synthesis [4], which are based on recordings from a speaker; and statistical parametric synthesis, which became a focused research are in the past few years.

The voice characteristics of automatic rule based articulatory and formant models can be widely modified, although the quality of these systems is not satisfactory, as the applied rules are not precise enough. Diphone and triphone based methods produce constant quality and the voice characteristics can be modified to some degree, but they still sound unnatural. Corpus-based unit selection systems produce high quality, natural sounding voice, but the quality is not constant, the voice

*The work was partially supported by the Hungarian National Office for Research and Technology (Teleauto project - OM-00102/2007, BelAmi project - ALAP2-00004/2005) and by ETOCOM project (TÁMOP-4.2.2-08/1/KMR-2008-0007) through the Hungarian National Development Agency in the framework of Social Renewal Operative Programme supported by EU and co-financed by the European Social Fund.

†Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, E-mail: {toth, nemeth}@tmit.bme.hu

characteristics cannot be modified and for the best quality large runtime databases are required.

In statistical parametric synthesis usually the hidden Markov Model paradigm is used [5]. It has numerous advantages compared to other methods: it has comparable voice quality to that of the state-of-the-art unit selection methods, the runtime database is small (2-10 MB) [6], the voice characteristics can be changed by speaker adaptation [7][8] and interpolation [9] and emotions can also be expressed [10].

HMM-based TTS is categorized as a kind of unit selection speech synthesis, although in this case the units are not waveform samples, but spectral and prosody parameters extracted from the waveform. HMMs are responsible for selecting those parameters which most precisely represent the text to be read and a vocoder generates the synthesized speech from these parameters. Different vocoder techniques can be applied, generally pulse-noise or mixed excitation is used (the latter has better quality, but its runtime database and computational cost is higher).

The first section of the current paper gives an overview about the architecture of HMM-based speech synthesis (that is the basis for our TTS system). It investigates the two basic training (speaker dependent, speaker adaptive) and the two basic synthesis methods (pulse-noise excitation, mixed excitation) that are applied in order to improve the systems quality. In the second part of the paper Hungarian specific solutions of the system are discussed and a listening test and its evaluation are carried out, which involves diphone-based, corpus-based unit selection and HMM-based Hungarian TTS systems.

2 HMM-based text-to-speech synthesis

Hidden Markov models are often used to simulate the behavior of physical processes based on observations. In speech technology HMMs can successfully model the behavior of human speech. Both in speech recognition and synthesis descriptive parameters of a speech corpus are used as observations, which is much more efficient than wave sample based observations. HMMs have already been applied in speech recognition for a long time [11]. In the last decade HMM-based speech synthesis became a focused research area. It differs from the method applied in speech recognition in three main parts:

- In case of speech synthesis at the last step instead of "pattern matching" "pattern selection" is executed, so the most likely parameters (e.g. spectral coefficients, pitch, state duration) are selected. Speech is generated by a vocoder from the selected parameters.
- Prosody is also modeled in speech synthesis, including pitch and phoneme durations.
- In speech synthesis a more complex acoustic model is used instead of triphones, which involves segmental and supra-segmental information. This is described by context dependent labels (see subsection 2.1.3).

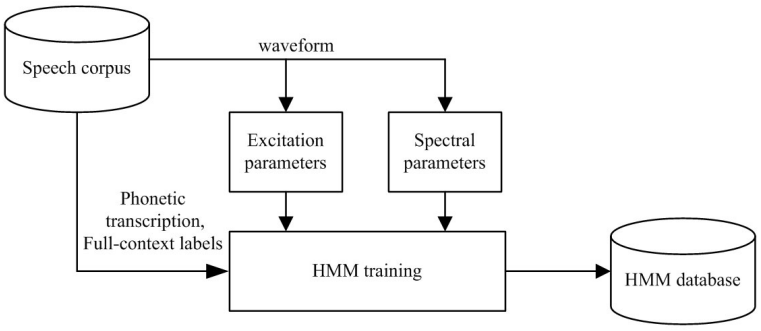

Figure 1: Block diagram of speaker dependent training.

HMM-based TTS consists of two main tasks: the training and the speech synthesis task. In the training task the HMM parameters are trained by a large, precisely labeled speech corpus. As a result a small HMM database is created, which includes the representative parameters of the speech corpus (training). From this database the best matching parameters of the text to be read are selected and the utterance is generated by a vocoder (synthesis).

2.1 Training

There are two main types of HMM training: the speaker dependent and the speaker adaptive training methods.

2.1.1 Speaker dependent training

For speaker dependent training (see Figure 1) a rather larger speech corpus (minimum 1-1.5 hours of speech) from a given speaker, the phonetic transcription and precise phoneme boundary labeling are required. The spectral parameters (e.g. features derived by linear prediction analysis), their first and second derivatives, the pitch, its first and second derivatives are extracted from the waveform.

As the next step phonetic transcriptions are extended to context dependent labels (see subsection 2.1.3.). When all these data are prepared, the training procedure is started. During training the HMMs learn the spectral and excitation parameters according to the context dependent labels of the given corpus. To be able to model parameters with varying dimensions multi-space distribution HMMs (MSD-HMMs) are used [12] (e.g. logF0 in case of voiced/unvoiced regions is modeled by 2 dimensional HMMs). To model the rhythm of the speech state duration densities are calculated for each phoneme. The set of state durations of each phoneme HMM is modeled by a multi-dimensional Gauss distribution.

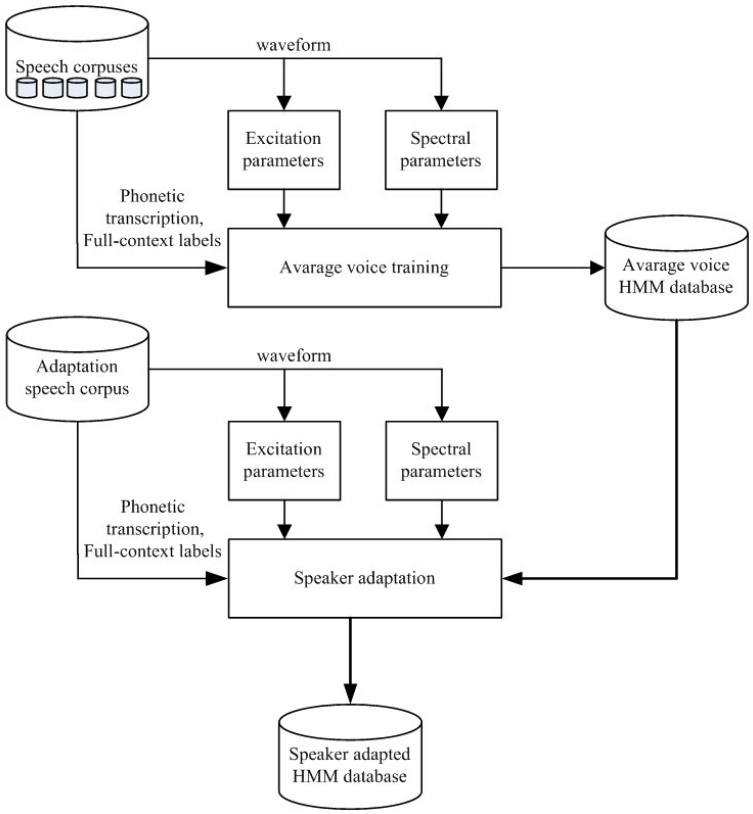

Figure 2: Block diagram of speaker adapted training.

2.1.2 Speaker adaptive training

Speaker adaptive training is divided into two parts (Figure 2). First the so called average voice must be constructed, then the average voice is adapted to the target speakers voice. For the average voice speech corpora (minimum 1-1.5 hours/speaker) from numerous speakers (minimum 4-5) is required. The excitation and spectral parameters and their first and second derivatives are extracted from the corpora. The average voice model is trained with this data and with the related phonetic transcriptions and context dependent labels. If an average voice is available, a much smaller speech corpus of 5-10 minutes is sufficient for adaptation. The same training data are extracted from the adaptation corpus for completeing the adaptation phase.

2.1.3 Context dependent labeling and decision trees

To describe the features of a phoneme in detail - to be able to select the most likely units in the synthesis phase - a number of phonetic features should be defined. These features are calculated for every sound. Labeling is done automatically, which may include errors (e.g. finding the accented syllables, defining the part of speech). This effect is likely not to influence the quality much, if the same algorithm is used in speech generation, thus the parameters are chosen by the HMMs consistently. In subsection 3.3. the features that were used in the Hungarian version of our HMM-based TTS system are described.

The combination of all possible context dependent features is a huge number. If only the possible variations of quintphones (this is a basic context dependent feature, see subsection 3.3.) are taken into account, that is over 160 million and this number increases exponentially if further context dependent features are included as well. Consequently it is impossible to design a speech corpus, which contains all combinations of context dependent features. To overcome this problem decision tree based clustering [13] is used. As different features influence the spectral parameters, the pitch values and the state durations, decision trees are separately handled for each. In subsection 3.3. the general questions used for building the decision trees in the Hungarian version of HMM-based synthesis are introduced.

2.2 Synthesis

The speech synthesis method is the same in the case of both training methods: the HMMs generate the most likely parameters (including pitch, state durations and spectral parameters) belonging to the text and then the speech is generated by a vocoder method. Depending on the type of the parameters, that were used during training, the vocoder may be a simple vocoder (e.g. LPC-10), although mixed excitation vocoders perform much better, as they significantly reduce the buzzyness of the speech. Certainly different vocoder techniques influence the choice of the parameters, that are to be extracted from the waveform, and they may also influence the training methods of the HMMs (e.g. pitch modeling requires MSD-HMMs).

In this study we have tested the two most commonly used vocoder techniques in HMM-based speech synthesis, the pulse-noise and mixed excitation vocoders.

2.2.1 Pulse-noise excitation vocoder

The pitch (voiced regions) or a binary flag (unvoiced regions), the spectral parameters and the state durations should be extracted from the speech corpus and trained for the HMMs in the pulse-noise excitation model. To be able to model voiced and unvoiced regions, MSD-HMMs are used. In the synthesis phase the excitation is modeled as periodic pulse trains at the rate of the pitch that was generated by the HMMs (voiced phonemes) or as white noise (unvoiced phonemes). This excitation signal is filtered by a Mel-Log Spectral Approximation (MLSA) filter [14] and the synthesized speech is generated (see Figure 3).

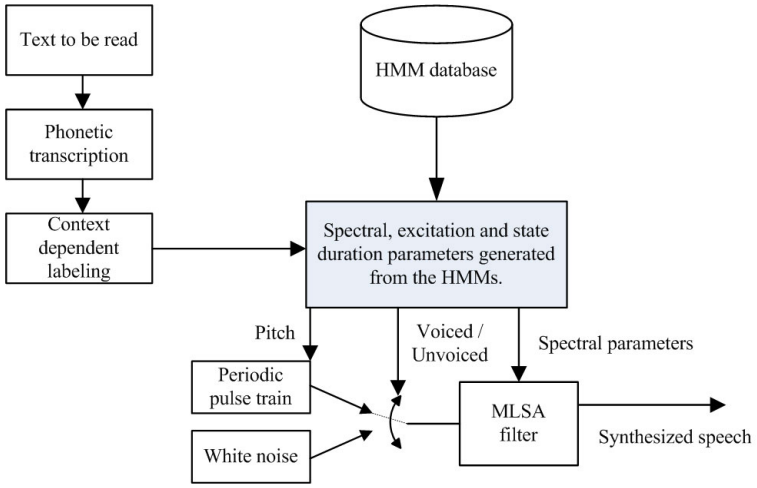

Figure 3: Block diagram of HMM TTS with pulse-noise excitation.

The advantage of pulse-noise excitation is the simplicity, furthermore a small footprint runtime database is enough and the computational cost mainly depends on the order of the MLSA filter. The main disadvantage is the buzzyness of the synthesized voice.

2.2.2 Mixed excitation vocoder

To make the synthesized voice more natural and to eliminate the buzzyness mixed-excitation vocoders were introduced [15]. In the mixed excitation model (see Figure 4) the pitch, the bandpass voicing strengths and spectral parameters are extracted and trained for the HMMs. In the synthesis phase the parameters of the bandpass filters for the periodic pulse train and for the white noise excitation are generated by the HMMs (bandpass voicing strengths). After the excitation signals passed through the bandpass filters, the results are summed and filtered by an MLSA filter. As a result the synthesized voice is generated.

The main advantage of using mixed excitation is the good, natural sounding quality, although more computational cost is required as the number and the order of filters increases. Further improvements in quality can be achieved by post filtering the synthesized voice [16].

3 Improvements of Hungarian HMM-based TTS

Several language specific steps are necessary to create a Hungarian HMM-based text-to-speech engine. The basics of a Hungarian HMM-based speaker dependent text-to-speech engine are described in [17]. In this chapter the most significant

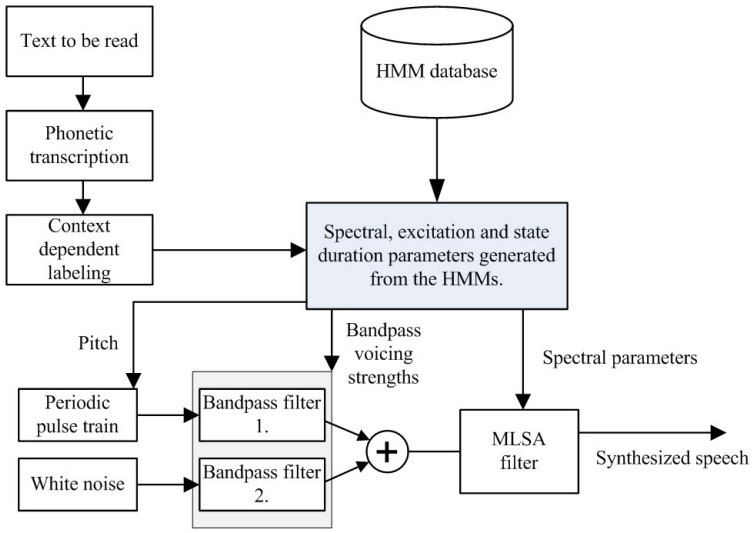

Figure 4: Block diagram of HMM TTS with mixed excitation.

issues of creating a Hungarian HMM-based speaker adapted text-to-speech system are investigated.

3.1 Speech databases

Five speech databases were recorded: four males and one female, for the average voice. The utterances are well designed, phonetically balanced sentences . The content of the utterances was manually verified. Phoneme boundaries were determined by forced alignment with a wide beam. The properties of the speech databases are shown in Table 1.

Table 1: Speech corpora for the average voice (44 kHz, 16 bit, mono format)

Speaker	Number of sentences	Duration	Size
1. male speaker	1941	170 minutes	857 MB
2. male speaker	1938	137 minutes	694 MB
3. male speaker	1944	191 minutes	966 MB
4. male speaker	1938	214 minutes	1082 MB
5. female speaker	1940	129 minutes	652 MB

For adaptation we used several different databases, including semi-spontaneous political speeches, weather forecasts, price list utterances (planned speech), and general, phonetically balanced utterances. The length of the adaptation speech

databases was between 5-15 minutes. In the current paper adaptation with a general, phonetically balanced database is investigated (see section 4.) with the properties shown in Table 2.

Table 2: Speech corpus for adaptation (44 kHz, 16 bit, mono format)

Speaker	Number of sentences	Duration	Size
Female speaker	117	8 minutes	40 MB

The speech databases were resampled at a rate of 16 kHz on 16 bits and windowed by a 25 ms Hanning-window with 5 ms shift. The feature vectors consisted of 39 mel-cepstral coefficients (including the 0th coefficient), logF0, aperiodicity measures, and their dynamic and acceleration coefficients.

3.2 Adaptation technique

There are two main techniques of speaker adaptation in the HMM paradigm: maximum likelihood linear regression (MLLR) [7] and maximum a posteriori (MAP) estimation [8]. MLLR is applied when the amount of adaptation data is small, for MAP more data is required as the Gaussian distributions are updated individually.

The Hungarian version uses the MLLR adaptation method. MLLR modifies the parameters of the average voice to the target voice by linear transforms. In this case the state outputs are:

$$b_j(o_t) = \mathbb{N}(o_t; \hat{\mu}_j, \hat{\Sigma}_j) \tag{1}$$

$$\hat{\mu}_j = A_{\gamma(j)}\mu_j + b_{\gamma(j)} \tag{2}$$

$$\hat{\Sigma}_j = H^T_{r(j)}\Sigma_j H_{r(j)} \tag{3}$$

b_j corresponds to the output probability function, o_t is the observation vector, μ_j and Σ_j are the original mean vector and covariance matrix. $\hat{\mu}_j$ is the linearly transformed mean vector of the j-th state output distribution and $\hat{\Sigma}_j$ is the linearly transformed covariance matrix of the j-th state output distribution. The covariance matrix adaptation is performed after the mean vector adaptation. $A_{\gamma(j)}$, $b_{\gamma(j)}$ and $H_{r(j)}$correspond to the mean linear transformation matrix, to the bias vector and to the covariance linear transformation matrix for the r_j-th regression class.

Generally there are two types of MLLR adaptation. If A and H linear transformation matrices are the same, than we talk about constrained MLLR (CMLLR), otherwise it is unconstrained MLLR. We used CMLLR for adaptation. The state output distributions are clustered by regression class trees; in a given class we use the same transformation matrices and bias vectors. The linear transform is derived from the labeled adaptation data. In order to perform adaptation with less data, the context-dependent models with regression or decision trees are used. The

complexity and generalization abilities of the adaptation can be controlled by adjusting the size of the regression-class / decision tree to the size of the adaptation data. CMLLR is the most commonly used adaptation technique, but other, more sophisticated schemes are available as well [18].

Classic speaker adaptation uses precise phonetic transcriptions, manually transcribed or automatically annotated segmentation and linguistic labels - this is called supervised speaker adaptation. In the unsupervised case the adaptation process does not require any manual interaction. The advantages of unsupervised adaptation are quite appealing: the creation of target voices becomes automatic which is favorable if several voices are required or if no pre-processing of the speech data is possible. There are some solutions for unsupervised speaker adaptation, which are introduced in [19], [20] and [21]. We also conducted some experiments of ASR transcription based unsupervised adaptation in Hungarian with promising results [22].

The gender of the average voice database speakers is an important question. If large speech corpora are available then creating gender dependent average voices is ideal. In practice only some speech corpora are available from both males and females, thus a mixed gender average voice is used often. [23] introduces a method, which causes minimal quality degradation in case of adapting a mixed gender average voice to male or to female voices, compared to the gender dependent case. As shown in Table 1 four male and one female speakers were used in our experiments for the average voice. According to some inner tests in our laboratory, there was no significant difference between adapting to male or to female voices from the average voice.

3.3 Context dependent labeling and decision trees

In 2.1.3. context dependent labels and decision trees were introduced in general. In this subsection we investigate their language specific features. Table 3 shows the context dependent labels, which were used in the Hungarian HMM-based TTS system. An example for a context dependent label looks like the following:

```
a^l-a1+bb=i@2_1/A:2_1/B:0-2@2-1&6-6$2-0;0-...
```

The questions for the decision tree building algorithm have been defined according to these features. Depending on the modeled parameter (spectral, pitch, duration) the most significant question varies, although generally the questions regarding to phonemes are dominant. These questions are determined by the behavior of the Hungarian phonemes [24]. Table 4 shows some important features that are used for the creation of the decision trees.

Figure 5 shows an example for decision trees in the case of spectral features. C_, L_ and R_ denote the central, left and right neighbouring phonemes that are under examination. The figure shows that the "Is the central phoneme in the quintphone a vowel?" was the most significant question in this case (it is on the "top" of the decision tree). On the next level there are the "Is the center phoneme a low

Table 3: The main features used by Hungarian context dependent labeling.

Sounds	The current and the two previous and the two following sounds/phonemes (quintphones). Pauses are also marked.
Syllables	Mark if the current / previous / next syllable is accented. The number of phonemes in the current / previous / next syllable. The number of syllables from / to the previous / next accented syllable. The vowel of the current syllable.
Word	The number of syllables in the current / previous / next word. The position of the current word in the current phrase (forward and backward).
Phrase	The number of syllables in the current / previous / next phrase. The position of the current phrase in the sentence (forward and backward).
Sentence	The number of syllables in the current sentence. The number of words in the current sentence. The number of phrases in the current sentence.

Table 4: The most important features used for building the decision tree.

Phonemes	Is it vowel or consonant? Is it short or long? Is it stop / fricative / affricative / liquid / nasal phoneme? Is it front / central / back vowel? Is it high / medium / low vowel? Is it rounded / unrounded vowel?
Syllable	Is it a stressed or a not stressed syllable? Numeric parameters (see Table 3).
Word	Numeric parameters (see Table 3).
Phrase	Numeric parameters (see Table 3).
Sentence	Numeric parameters (see Table 3).

vowel?" and the "Is the center phoneme unvoiced stop?" questions. The same idea is followed at lower levels.

4 Results

A modified version of the HTS framework with STRAIGHT [6] was applied for training and for generation. The speech corpora shown in Table 1 was processed to

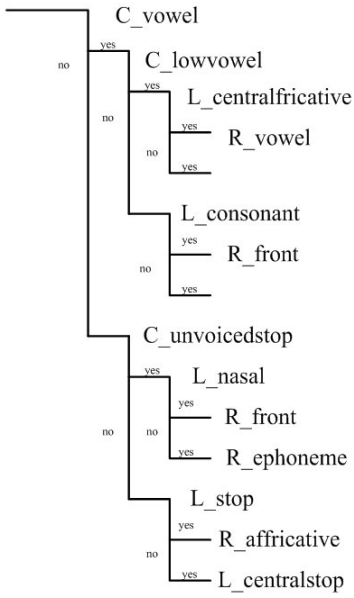

Figure 5: Example for decision trees (spectral features).

create the models of the average voice and the speech corpus in Table 2 was used for adaptation. A listening test was carried out to evaluate the quality of comparable Hungarian TTS systems.

4.1 Experimental conditions

Five TTS systems were involved in the listening test: a triphone based system (System A); a general domain corpus based unit selection system (System B); a domain specific corpus based unit selection system (System C); a HMM-based speaker adapted system with pulse-noise excitation (System D) and a HMM-based speaker adapted system with mixed excitation (System E). The original speaker (from whom the speech corpora were recorded) was the same in case of all TTS systems. The corpus based unit selection system had all the waveforms from one speaker in the runtime database. The HMM-based speech synthesis system had the waveforms in the training database from five (average voice) plus one speaker (adaptation), as it is described in 4.1. The language of the test was Hungarian. The properties of the different systems are shown in Table 5.

The listening test consisted of two parts:

- the first part was a Comparison Mean Opinion Score (CMOS) test;
- the second part was Mean Opinion Score (MOS) test.

In the first part test subjects had to decide on a five point scale from two synthesized samples which one sounds more natural. On the scale 3 meant the quality of the two samples are considered same, higher values meant that the second sample was considered more natural (4 more natural, 5 much more natural), lower values meant that the first sample was considered more natural (2 more natural, 5 much more natural). The text of the utterance in a pair was the same. In a pair different speech synthesis systems were used. Altogether 9 pairs were played; each pair was played twice (normal and inverted order). CMOS pair comparison as the first part of the test is favorable, as subjects get used to the synthetic voice and they will give consistent answers for the MOS tests in the next part. In the second section the test subjects had to mark on a five point scale the naturalness of 20 samples, 4 samples from each system. Lower values meant worse naturalness, higher values meant better naturalness. In the second section the text of the utterances was different.

We have chosen this order of the two main parts to minimize the chance that the test subjects memorize the different systems. The samples were selected from a larger set of sentences in order to get the desired information about the systems and not about the speech samples. Furthermore the samples were sorted in different pseudo-random orders for every test subject to avoid memory effects. The distribution of the samples and the systems was kept even.

The authors carried out a pre-test with five subjects to verify the effectiveness of the test design. The results of the pre-test were adequate, so the same design was kept and the results of the pre-test were also included in computing the final results.

Altogether 24 test subjects (7 female, 17 male) were involved in the test. All the test subjects were native Hungarian speakers with no known hearing loss. The test was internet-based, the average age was 32, and the youngest subject was 22, the oldest 67 years old. 7 test subjects were speech experts.

4.2 Analysis of the results

The results of the listening text are shown in Table 6 and Table 7. Table 6 contains the general preference scores of the CMOS test and the results of the MOS test. In Table 7 the particular values related to the HMM-based speech synthesis systems are shown. The results are represented in Figure 6 on boxplot diagrams according to the guidelines of [25]. On boxplot diagram systems can easily be compared by the median (black thick line), by the 1st and 3rd quartiles (bar), by the whiskers and outliers. The most significant information are the median, the 1st and the 3rd quartiles. As it was expected System A scored the worst in both parts of the test. Although the naturalness of System A is much worse then the naturalness of other systems, it has got a small footprint and its computational costs are very low, so it can be applied in low resource systems. The naturalness of System B was considered also quite low and its runtime database is large and the computational costs are also high.

System D achieved the third position. Its global preference score is almost

Table 5: Speech synthesis systems involved in the listening test

System	Technique	Training database	Runtime database
A	Triphone based unit selection	-	285 MB (triphones)
B	Corpus based unit selection (general domain)	-	4634 MB (one speaker, 44 kHz, 16 bit, mono waveforms)
C	Corpus based unit selection (domain specific)	-	3113 MB (one speaker, 44 kHz, 16 bit, mono waveforms)
D	HMM-based speech synthesis (speaker adapted, pulse-noise excitation)	4251 + 40 MB (five + one speakers, 44 kHz, 16 bit, mono waveforms)	2 MB (HMM parameters, decision trees)
E	HMM-based speech synthesis (speaker adapted, mixed excitation)	4251 + 40 MB (five + one speakers, 44 kHz, 16 bit, mono waveforms)	11 MB (HMM parameters, decision trees)

the same as the score of System B, but its general naturalness and CMOS score compared to System B are higher. In addition System B has a small runtime database.

System C and System E performed the best in the listening test. System C was considered better than System E in the pair comparison part (on Figure 6 they have the same median, but System C has higher 3rd quartile), in the general naturalness part System E was considered better. These differences are mostly not significant and the reason, why the two systems performed different in the two parts is that their naturalness is quite close to each other. The only significant difference is the median of the systems in the MOS test, where System E performs better (see Figure 6). In case of more test subjects the scores of systems C and E may get closer. However System C performed well only in a given domain with a large runtime database, System E performed the same quality on general sentences with a small runtime database.

5 Conclusions

In the current paper the basics of HMM-based speech synthesis are introduced, including speaker dependent and speaker adaptive training, furthermore two different speech generation techniques, the pulse-noise and mixed excitation based

Table 6: Results (mean ± variance) of the listening test. Higher numbers mean better naturalness.

	CMOS (Global preference score) Compared naturalness of speech synthesis systems	MOS General naturalness of the systems
System A	2.3 ± 1.14	2.1 ± 0.9
System B	2.8 ± 1.27	2.6 ± 1.1
System C	3.6 ± 1.3	3.2 ± 1.1
System D	2.9 ± 1.27	3.1 ± 1.2
System E	3.4 ± 1.22	3.5 ± 1.0

Table 7: CMOS pair comparison values for System D and System E (3 means identical naturalness, higher values mean that the system in the row was considered more natural, lower values mean that the system in the column was considered more natural)

System	A	B	C	D	E
D	3.3 ± 1.15	3.3 ± 1.3	2.7 ± 1.4	N/A	2.5 ± 1.0
E	3.9 ± 1.1	3.5 ± 1.3	2.7 ± 1.3	3.5 ± 1.0	N/A

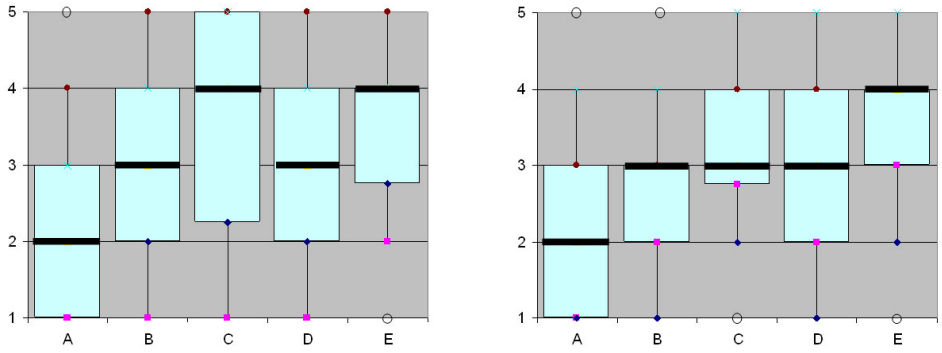

Figure 6: Boxplot showing compared naturalness of the speech synthesis systems (left) and general naturalness of the speech synthesis systems (right).

vocoders are described. The Hungarian version of a speaker adapted HMM-based speech synthesis engine was investigated, and the most important language specific features are shown. To measure the quality of the system a listening test was carried out with some Hungarian speech synthesis engines. The results showed that

HMM-based speech synthesis with mixed excitation performs with a small runtime database on general sentences like the state-of-the-art corpus-based unit selection system with a large runtime database on domain specific sentences.

In the future we plan further error corrections and more precise labeling of the training data, as it is likely to increase the quality of the synthesized voice. Additionally the solution will be optimized for embedded environments. Other voice coding algorithms will also be applied.

References

[1] P. Mermelstein. Articulatory model for the study of speech production. Journal of the Acoustical Society of America, Volume 53, 1973, pp. 1070-1082.

[2] D. H. Klatt, L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America, Volume 87, Issue 2, February 1990, pp. 820-857

[3] D. O'Shaughnessy, L. Barbeau, D. Bernardi and D. Archambault. Diphone speech synthesis. Speech Communications, Volume 7, Issue 1, March 1988, pp. 55-65

[4] B. Möbius. Corpus-based speech synthesis: methods and challenges. Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition, 2000, pp. 7996

[5] A.W. Black, H. Zen, K. Tokuda. Statistical parametric speech synthesis. Proceedings of ICASSP, Apr. 2007, pp. 1229-1232

[6] J. Yamagishi, T. Nose, H. Zen, T. Toda, K. Tokuda. Performance evaluation of the speaker-independent HMM-based speech synthesis system "HTS-2007" for the Blizzard Challenge 2007. Proceedings of ICASSP 2008, Las Vegas, U.S.A, April 2008, pp. 3957-3960

[7] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi. Speaker adaptation for HMM-based speech synthesis system using MLLR. Proceedings of ESCA/COCOSDA Workshop on Speech Synthesis, November 1998, pp. 273-276

[8] K. Ogata, M. Tachibana, J. Yamagishi, T. Kobayashi. Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis. Proceedings of ICSLP 2006, September 2006, pp. 13281331.

[9] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura. Speaker interpolation for HMM-based speech synthesis system. Journal of the Acoustical Society of Japan (E), Volume 21, Issue 4, 2000, pp. 199-206

[10] T. Nose, M. Tachibana, T. Kobayashi. HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation. IEICE Trans. Inf. & Syst., Volume E92-D, Issue 3, Mar. 2009, pp. 489-497

[11] Lawrence R., Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, pp. 257286

[12] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi. Hidden markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of ICASSP-99, March 1999, pp. 229232

[13] Young, S., Ollason, D., Valtchev, V. and Woodland, P. The HTK book. (for HTK version 3.4), March 2009. http://htk.eng.cam.ac.uk

[14] S. Imai, K. Sumita, C. Furuichi. Mel log spectral approximation filter for speech synthesis. Trans. IECE, Volume J66-A, February 1983, pp. l22-129

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. Mixed excitation for HMM-based speech Synthesis. Proceedings of Eurospeech, Sept. 2001, pp.2259-2262

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. Incorporation of mixed excitation model and postfilter into HMM-based text-to-speechsynthesis. Systems and Computers in Japan, Volume 36, Issue 12, September 2005, pp. 4350

[17] B. Tóth, G. Németh. Hidden Markov model based speech synthesis system in Hungarian. Infocommunications Journal, Volume LXIII, no. 2008/7, 2008, pp. 3034

[18] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai. Analysis of Speaker Adaptation Algorihms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Audio, Speech, and Language Processing, Volume 17 Issue 1, January 2009, pp. 66-83

[19] S. King, K. Tokuda, H. Zen, and J. Yamagishi. Unsupervised adaptation for HMM-based speech synthesis. Proceedings of Interspeech 2008, 2008, pp. 18691872

[20] M. Gibson, T. Hirshimaki, R. Karhila, M. Kurimo and W. Byrne. Unsupervised Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis Using Two-Pass Decision Tree Construction. Proceedings of ICASSP 2010, Dallas, USA

[21] K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester. Unsupervised Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis. Proceedings of ICASSP 2010, Dallas, USA

[22] B. Tóth, T. Fegyó, G. Németh. Some aspects of ASR transcription based unsupervised speaker adaptation for HMM speech synthesis. 13th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, September 2010

[23] J. Yamagishi, T. Kobayashi, S. Renals, S. King, H. Zen, T. Toda, K. Tokuda. Improved Average-Voice-based Speech Synthesis using Gender-Mixed Modeling and A Parameter Generation Algorithm considering GV. Proceedings of ISCA SSW6, Bonn, Germany, August 2007, pp. 125-130

[24] M. Gósy. Phonetics, the Science of Speech (in Hungarian). Budapest, Osiris, 2004, p. 350

[25] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King. Statistical analysis of the Blizzard Challenge 2007 listening test results. Proceedings of Blizzard 2007 (in Proceedings of Sixth ISCA Workshop on Speech Synthesis), Bonn, Germany, August 2007, pp. 1-6