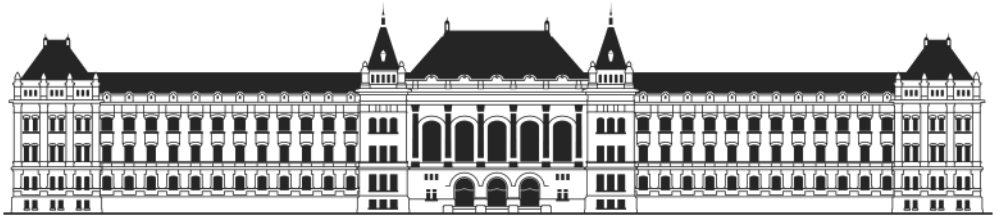

Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics

Hidden Markov-model based text-to-speech synthesis

Ph.D. thesis booklet
Doctoral School of Electrical Engineering

Bálint Pál Tóth, M.Sc.

Supervisors
Géza Németh, Ph.D.
Gábor Olaszy, D.Sc.

Budapest, Hungary
2013

1. Introduction

Speech production is a complex process: the brain precisely controls the articulatory system at high speed and the speaker gets audio feedback of his or her communication via the hearing organs. To be able to mimic speech production artificially, not only the articulatory system, but the mechanism of the brain should be understood. Since we are far from understanding the brain, models for speech production are built.

The general goal of speech synthesis is to create a natural sounding, highly intelligible synthetic voice. In addition the following basic engineering aspects must be kept in mind: available resources and target platforms. The dissertation and the current thesis booklet focus on the general goal and the engineering aspects as well.

2. Background

In general text-to-speech (TTS) synthesis systems consist of two main parts: text preprocessor and speech generation (see Figure 1.). The input text is converted into a feature matrix, which contains the phonemes of the input text and additional information (e.g. stresses, segmental features) generated by the text preprocessor. According to this feature matrix the synthesized voice waveform is created by the speech generation module.

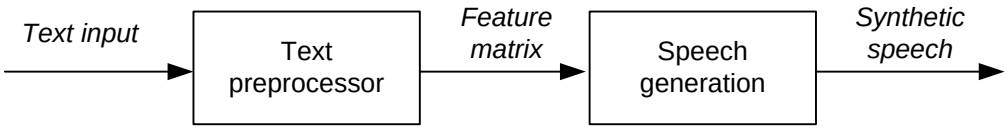

Figure 1. The general structure of TTS synthesis systems.

The artificial production of speech has a long history. The first mechanical speech production system was built by Farkas Kempelen back in 1791 [1]. In the last three decades computer based approaches have been preferred [2,3]. Articulatory [4] and formant [5] synthesis tries to model the mechanism of speech organs. Diphone a triphone based speech synthesis concatenates phoneme level waveforms [6,7]. Unit selection speech synthesis systems concatenate waveforms from a precisely labelled speech corpus based on concatenation and target costs [8,9].

Recently hidden Markov-model (HMM) based speech synthesis has become a focused-on research area [10]. HMM-based TTS systems produce high-quality human-like voices. Compared to other speech synthesis systems with similar speech quality, the HMM speech synthesis footprint is fairly small, but computational cost and playback latency are often high.

HMM-based TTS consists of two main parts: the training and the speech synthesis components. During the training process, HMM parameters are learned from a large precisely labelled speech corpus and generative models are built. The parameter approximation is based on the maximum likelihood (or similar) technique:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \{p(\mathbf{O} \mid W, \boldsymbol{\lambda})\} \tag{1}$$

where $\boldsymbol{\lambda}$ contains the model's parameters, $\boldsymbol{O}$ is the observation vector extracted from the speech corpus (training data) and W denotes the word sequence representing the speech corpus. As a result, a small HMM database is created, which includes the representative parameters of the speech corpus. At the synthesis stage, the best matching $\boldsymbol{o}$ output probabilities of $\hat{\boldsymbol{\lambda}}$ model to w textual word sequence are maximized:

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{p(\mathbf{o} \mid w, \hat{\boldsymbol{\lambda}})\} \tag{2}$$

From these parameters the synthetic voice is generated by a vocoder algorithm. Figure 2. shows the general architecture of a hidden Markov-model based text-to-speech (HMM-TTS) synthesis system. Excitation and spectral parameters are extracted from the waveform and based on the phonetic transcription context dependent labels are calculated. This information is passed to the training algorithm (Equation 1). Context dependent labels typically contain phonemes, phoneme boundaries, accents, segmental information (phoneme, syllable, word, phrase, sentence level), and they may contain several additional features as well. The possible combinations of context dependent labels are high. A representative speech corpus containing all possible variations cannot be created. To overcome this problem decision tree based context clustering is used. In the training phase separate generative models are built for excitation parameters, spectral parameters and state durations. Continuous parameter streams (e.g. spectral parameters) are modelled by Gaussian distributions, and discrete/continuous parameters (e.g. voiced/unvoiced regions in excitation) are modelled with multispace probability distributions (MSD). In order to model the timing properly state transition probabilities are modelled by Gaussian distributions. In the synthesis phase Equation 2 is maximized: the HMM generative models create the most likely parameter stream of the input text. The waveform is created from this parameter stream with a vocoder algorithm.

HMM-based speech synthesis 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), the voice characteristics can be changed by speaker adaptation and interpolation and emotions can be expressed as well.

3. Research objectives

My general research topic is hidden Markov-model based text-to-speech synthesis. I focus on three different research areas of HMM-TTS: one of them is Hungarian language specific principally; two of them are language independent.

The *first research* objective is **creating a Hungarian hidden Markov-model based speech synthesis system and improving its quality**. This part of the research includes designing speech corpora, introducing language specific features, distinctive features,

creating speaker dependent and speaker adaptive HMM-TTS systems and measuring the quality improvements of manual correction of automatic labelling.

The *second research* objective is **automatic speech recognizer transcription based unsupervised speaker adaptation of HMM-TTS systems with semi-spontaneous speech**. The possibility of speaker adaptation with semi-spontaneous speech is investigated and an unsupervised speaker adaptation method is introduced. Results of subjective evaluation show that the proposed method is not significantly different from the supervised case even though the phoneme error rate (PER) is about 50%. The unsupervised adaptation method is extended to higher PERs as well.

The *third research* topic is **optimizing HMM-TTS for low-resource devices**. The noise generation algorithm in the excitation modelling is modified, optimal spectral parameter settings are investigated, the number of nodes in decision trees is reduced and the segment size of parameter generation, the vocoder algorithm and waveform playback is optimized according to the performance of the device and the actual load of the CPU.

I have chosen hidden Markov-model based text-to-speech synthesis for my research topic because of its novelty and countless possibilities. Furthermore it was a challenge to pioneer HMM-TTS research in Hungary. In the current thesis booklet I summarize the novel outcomes of my research grouped in the three research objectives.

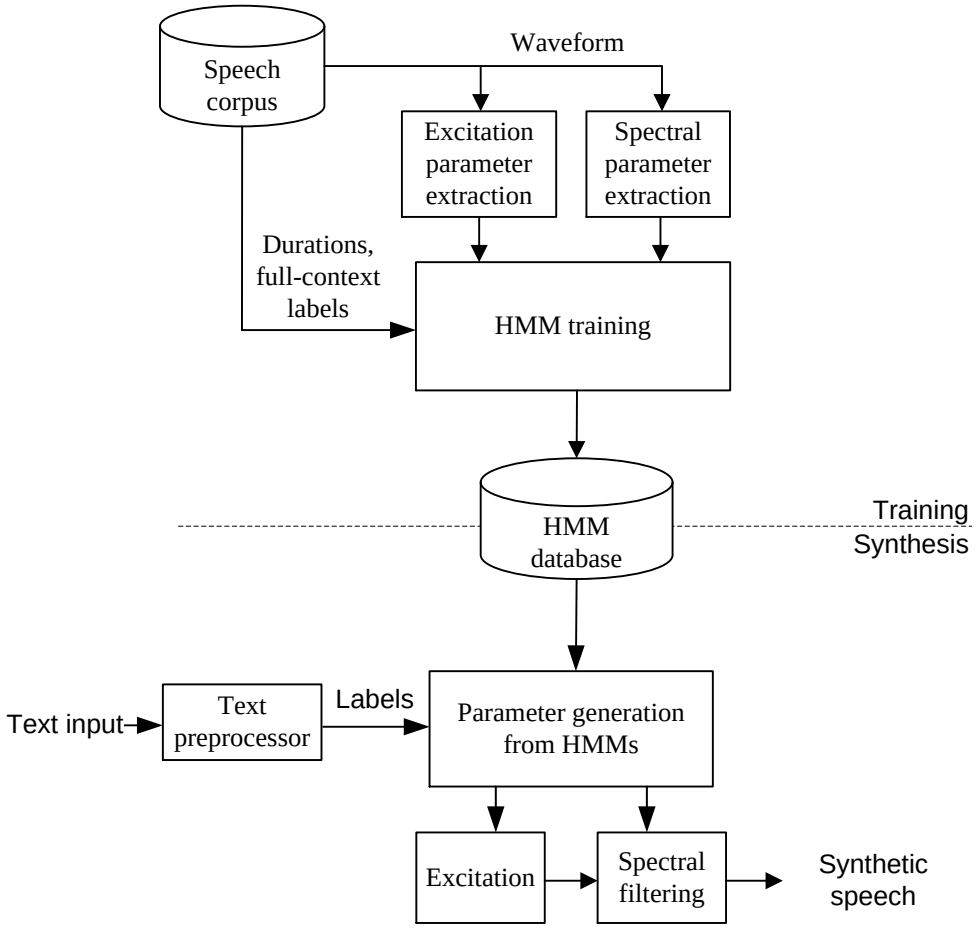

Figure 2. The general architecture of HMM-TTS; based on [10].

4. Methodology

I will introduce in this chapter the speech corpora, the tools and the evaluation method used during my research.

4.1. Speech corpora

A speech corpus contains the following: waveform (studio recordings are preferred), phonetic transcription and segmentation labels (phoneme boundaries). At the beginning of the research there was no available Hungarian speech corpus suitable for HMM-TTS. It is important for HMM-TTS that the speech corpus contains phonetically balanced sentences with regular phoneme distribution based on Hungarian language characteristics. The MTBA speech corpus includes about 6-7 minute long telephone conversations from 500 speakers, basically for speech recognition purposes [11]. I investigated the utterances of MTBA database and I found them suitable for HMM-TTS purposes, although at least one hour of studio quality recordings (min. 44 kHz, 16 bit) is required from each speaker, and a smaller number of speakers is enough (5-10 speakers). Based on the MTBA sentences, therefore, and with the help of BME-TMIT Speech Laboratory colleagues, we recorded and segmented speech corpora from seven speakers (approx. 20 hours altogether). We took into account the experiences from the creation of the MTBA database [12]. I used these speech corpora in the first thesis group.

In the second thesis group, in addition to the speech corpora of the first thesis group, I used semi-spontaneous parliamentary speeches. The semi-spontaneous speech corpora were collected from four speakers (approx. 4 hours), from which I selected 10 minutes per speaker for unsupervised adaptation with different methods.

In the third thesis group I worked with an English speech corpus. I used the SLT speaker of the ARCTIC database from the Speech Technology Laboratory of the Carnegie Mellon University [13].

The speech corpora used in my research are summarized in Table 1. In my first thesis group I used only a part of the whole speech corpus of a given speaker for speaker adaptation (10-15 minutes). In the second thesis group for the average voice I used the databases from the first thesis group. For speaker adaptation I created several adaptation corpora which are described briefly in this booklet and in detail in my dissertation (these are not shown in the table).

4.2. Synthesized sentences for listening tests

My goal is to create a general solution. Consequently I synthesized generic declarative sentences (not domain specific ones) with the systems I created. These synthesized utterances were used in the listening tests. The language of the test sentences was defined by the language of the given HMM-TTS system.

Table 1. The speech corpora used for HMM-TTS research.

Thesis group	Symbol	Length	Sex	Language	Purpose
I.	M1	190 min	male	Hungarian	Speaker dependent training, average voice training, (supervised) speaker adaptation
	M2	137 min			
	M3	170 min			
	M4	214 min			
	M5	198 min			
	F1	128 min	female		
	F2	193 min			
II.	M6	11.4 min	male		Supervised and unsupervised speaker adaptation
	M7	9.6 min			
	M8	8.9-10.2 min			
	M9	9.7 min			
III.	CMU-ARCTIC-SLT	47 min	female	English	HMM speech synthesis on low-resource devices (speaker dependent)

4.3. Experimental environment

I used open source tools and previous solutions of BME-TMIT. The main toolkits and applications used were as follows (the complete list can be found in my dissertation):

- **HTS** (HMM-based Text-To-Speech System): Training of speaker dependent HMMs, training of average voice HMMs, speaker adaptation. [14]
- **SPTK** (Speech Processing Toolkit): Parameter extraction and pulse-noise excitation based vocoder algorithm. [15]
- **STRAIGHT**: Parameter extraction and mixed excitation based vocoder algorithm. [16]
- **hts_engine**: Parameter generation from HMMs and waveform generation with pulse-noise excitation based vocoder algorithm. [14]
- **ProfiVox**: phonetic transcription and accents determination. [7]
- Hungarian large vocabulary automatic speech recognizer. [17]
- Forced alignment to determine phoneme boundaries automatically. [17]

4.4. Subjective evaluation

I used mean opinion score (MOS) and comparison mean opinion score (CMOS) listening tests for subjective evaluation. Test subjects had to score each utterance from 1 (worst) to 5 (best, integers) in MOS tests. In the case of CMOS tests subjects had to decide from two utterances which one fulfils the given criterion better in a 5 point scale (e.g. quality, naturalness, intelligibility). In some cases test subjects had to decide what they thought about the meaning of "quality". This way I received a general feedback about how the subjects consider the "overall quality" of the TTS system. Several features are included in the "overall quality", e.g., naturalness, sympathy, emotions triggered by the voice characteristics. In other cases test subjects were asked to score a specific feature, e.g.

naturalness of the synthetic voice. The precise settings of the listening tests are described in my dissertation in detail. The MOS and CMOS figures show the average value and confidence interval of 95%. I have checked significance in each case. If two results had to be compared I used two sample t-test (MOS tests) or one sample t-test (in the case of CMOS tests). If more than two results had to be compared, I used ANalysi Of VAriance (ANOVA) significance test. If the ANOVA showed significant difference, then for post hoc comparison Tukey's test was used. I tested significance at 95% confidence level (α=0.05).

In some cases MOS tests resulted in rather low values (~3), in other cases similar HMM-TTS systems scored better (~3.5-4). The reason for this difference can be explained by the involvement of natural speakers in the former case, with only synthetic voices involved in the latter. If a natural speaker is present, synthetic voices are considered worse, than in the case when there are only artificial voices.

5. New results

5.1. Thesis Group I. Hidden Markov-model based text-to-speech synthesis applied to Hungarian and quality enhancements.

First I created a Hungarian HMM-TTS system and compared it to previous Hungarian TTS systems. The synthetic voice quality of the system introduced in Thesis I.1 is enhanced by distinctive features in Thesis I.2. I applied speaker adaptation and I showed that with speaker adaptation it is possible to create a synthetic voice with significantly better quality than in the case with speaker dependent training (Thesis I.3). At the end of Thesis Group I the effects of manual correction of labelling (segmentation, phonetic transcription) in speaker dependent and speaker adaptive cases are investigated. The evaluation of the results includes subjective listening tests in each thesis, and investigation of decision trees in Thesis I.2.

Thesis I.1. [J2, J3, J4, B2a, B3, C6, C7] *I designed and implemented hidden Markov-model based text-to-speech synthesis in Hungarian and I showed that with significantly smaller database size the quality of the HMM-TTS system is not significantly worse than the quality of the state-of-the-art, domain specific corpus based Hungarian text-to-speech system.*

At the beginning of the research there was no Hungarian HMM-TTS solution, consequently I could only use international publications as guidelines [18,19]. Because of the difference between languages and due to the structure of Hungarian it wasn't a trivial task to create a Hungarian HMM-TTS system. As a first step of the research suitable speech corpora were created (see Chapter 4.1). Based on features of the Hungarian language I defined the possible phonemes, context-dependent labels and questions for decision trees [20]. These are described in detail in the dissertation and in related publications. I created an HMM-TTS voice with *M1* speech corpus (see Table 1) and synthesized sentences with a mixed-excitation vocoder.

Evaluation: I measured the quality of the resulting HMM-TTS system with listening tests. I compared the novel solution with two previous TTS systems developed at BME-TMIT: a triphone based [7] and a corpus based TTS system [9]. Figure 3. shows the results of the listing tests (left) and the runtime database sizes (right). According to the results the quality of HMM-TTS is not significantly worse than the state-of-the-art corpus based system, and it has a significantly smaller runtime database. Furthermore the quality of HMM-TTS is significantly better than the quality of triphone based synthesis, and the database size of these systems is not significantly different.

Conclusion: the runtime database size of the corpus based TTS system is about 850 MBytes (12 hours of recordings), whereas the runtime database size of the HMM-TTS system is about 10 MBytes (with 2 hours of recordings in the training database). The corpus based system produces constant quality in fixed domains (e.g. weather forecast); the HMM-TTS system gives constant quality in general domains. (The sub-phoneme level parametric model of HMM-TTS is a general speech synthesis technique. Furthermore in Thesis I.2, I.3, I.4, II.1 and II.2 listening tests were carried out with synthesized sentences from different domains and there was no perceptual difference between the quality of the synthetic speech from different domains.) According to the listening test of the current thesis there was no significant perceptual difference between the HMM-TTS and the corpus-based TTS even in a fixed domain. These results prompted me to make deeper investigations with HMM-TTS.

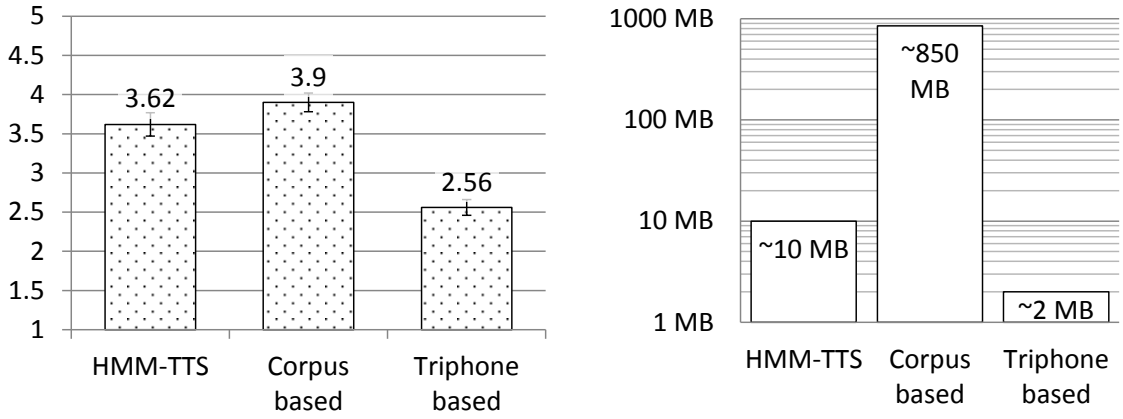

Figure 3. Subjective evaluation with listening test (left) and runtime database sizes (right) of HMM, corpus and triphone based TTS systems.

Thesis I.2. [C2] *I designed and implemented a method to apply distinctive features to a hidden Markov-model based text-to-speech system, and I showed that it is possible to increase the quality of synthetic speech by applying them.*

The same organs are used for speech production, and sound generation is independent from language [20]. The possibility of speech production is universal, although there are many language dependent differences. Distinctive features describe phonemes with binary and unary values language independently [21]. In Thesis I.1 I introduced a classification of Hungarian phonemes. With the help of distinctive features a more general description of phonemes is possible. I defined a set of distinctive features suitable for HMM-TTS for engineering purposes considering general linguistic principles and concepts. In the

elaborated hierarchy 18 distinctive features of three groups (articulator-free, articulator-bound, larynx) were used. I add distinctive features to the HMM-TTS system of Thesis I.1. I extended the questions used for decision tree building according to these distinctive features. I assigned two questions for binary and one question for unary features. My expectation was that distinctive features create more general clusters than the conventional notation.

Evaluation: I investigated the effects of distinctive features by analysing the changes in decision trees (compared to the system of Thesis I.1). The results are shown in Table 2 and Table 3. The decision trees for the five states are summarized in the figure. This way each value represents 5 states × 5 speakers = 25 decision trees. The header of both tables shows the parameter streams of mixed excitation. The results of Table 2 show how parameter streams were influenced by distinctive features. The biggest influence was in the case of spectral parameters, although all other streams are affected as well. Table 3. shows the ten most frequent distinctive features; articulatory-free features occur in more than 50% of the decision trees.

I measured the perceptual effect of distinctive features by MOS and CMOS listening tests. The results of the CMOS test are shown in Figure 4. *M1* denotes the experimental system of Thesis I.1 and *M1-DF* denotes the HMM-TTS with distinctive features. The figure shows that distinctive features increased the quality (*M1-DF* is preferred to *M1*). The results of the MOS test show an increment in quality as well. These results are described in the dissertation and in the related publication in detail.

Table 2. The ratio of distinctive features in the decision trees of Hungarian HMM-TTS (mixed excitation).

	F0	**Spectral parameters**	**Duration**	**Voicing strength**	**Σ**
Number of nodes	13821	3272	1153	4486	22732
Distinctive features	2664	1411	314	1018	5407
Ratio	19.3%	43.1%	27.2%	22.7%	23.8%

Table 3. The ten most frequent distinctive features in decision trees (articulatory-free distinctive features are bold and italic).

	F0	**Spectral parameters**	**Duration**	**Voicing strength**
1.	***sonorant***	back	***lateral***	***sonorant***
2.	low	***sonorant***	***sonorant***	***continuant***
3.	***continuant***	round	***continuant***	***nasal***
4.	***lateral***	***nasal***	round	high
5.	***nasal***	coronal	voiced	round
6.	round	low	***nasal***	***consonantal***
7.	voiced	high	low	***lateral***
8.	high	***lateral***	***strident***	voiced
9.	***strident***	***continuant***	***consonantal***	***strident***
10.	back	labial	low	low

Conclusion: distinctive features increased the speech quality of HMM-TTS and the structure of decision trees was also remarkably affected. Apart from the practical aspects (better speech quality), distinctive features bring HMM-TTS closer to the nature of speech production.

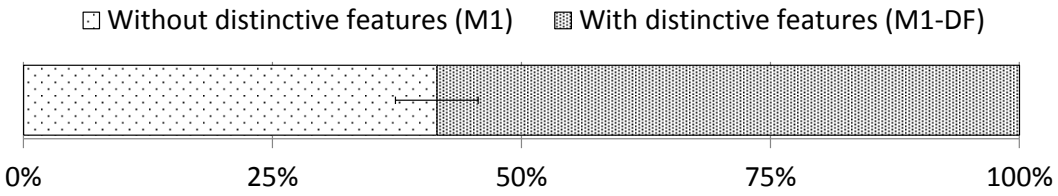

Figure 4. Subjective evaluation of effects of distinctive features with CMOS listening test.

Thesis I.3. [J2, B1, B2a, C6, C7] *I designed a supervised speaker adaptation method for hidden Markov-model based text-to-speech synthesis in Hungarian, which requires less than 10% of the speech corpus of the speaker dependent case to create new voices from the average voice model. I showed that it is possible to produce a synthetic voice with significantly better quality than in the speaker dependent case.*

In the current thesis I examined one of the most important features of hidden Markov-model based speech synthesis: speaker adaptation. I created the average voice model with *M2, M3, M4, M5* and *F2* speech corpora considering distinctive features (Thesis I.2). Next I modified the HMMs by an MLLR (Maximum Likelihood Linear Regression) procedure according to the parameters extracted from *M1* and *F1* speech corpora [22]. I denote speaker dependent cases with *SD* and speaker adapted cases with *SA*.

Evaluation: after speaker adaptation a male (*SA-M1*) and a female (*SA-F1*) voice were created. I compared the perceptual difference of these systems to the speaker dependent models of *M1* and *F1* (denoted by *SD-M1* and *SD-F1*) with MOS and CMOS listening tests. Figure 5. shows the results of the CMOS test: speaker adapted systems were preferred. The results of the MOS test are described in the dissertation in detail. In both cases (CMOS, MOS) the quality of speaker adapted systems was significantly better than in the speaker dependent case.

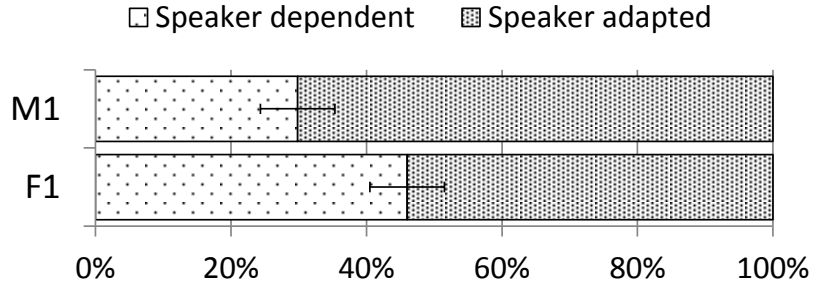

Figure 5. Subjective evaluation of speaker dependent and speaker adapted HMM-TTS with CMOS listening test.

Conclusion: based on the results of this thesis 10-15 minute recordings are enough for creating new HMM-TTS voice characteristics in contrast to the speaker dependent case in Thesis I.1 (2-3 hour recordings). The quality of speaker adapted voice can even be significantly better.

Thesis I.4. [B1] *I showed experimentally that manual correction of automatic labelling of training speech corpus may be substituted by automatic methods only, because manual correction does not always cause significant improvement in quality of synthetic speech in speaker dependent and speaker adapted HMM-TTS systems.*

After creating a Hungarian speaker dependent and speaker adapted HMM-TTS system my goal was to investigate the correlation between the quality of synthetic speech and the precision of the speech corpus. Manual correction of automatic labels requires deep knowledge and high precision; it is a time consuming work. Speaker dependent (*SD*) and speaker adapted (*SA*) HMM-TTS voices with automatically labelled (*auto*) and manually corrected (*manual*) male (*M1*) and female (*F1*) speech corpora were created. The average voice was built by automatically labelled speech corpora (the same as used in Thesis I.3).

Phoneme error related data are shown in Table 4, segmentation error related results are shown in Table 5. Table 4 shows that only a small percentage of the database was affected by phoneme errors in the speaker dependent speech corpora (0.83%, 0.52%). In the adaptation speech corpora a higher phoneme error ratio was measured (15.5%, 6%). The header of Table 5 denotes the difference in time between the automatic and manually corrected phoneme boundaries, and the values refer to the number of corrections in the given speech corpus. Comparing the number of phonemes with Table 4 it can be concluded that about 17 to 31 percent of phoneme boundaries were manually corrected.

Table 4. Features of automatic and manually corrected phonemes in speech corpora (speaker dependent and speaker adapted systems).

	SD-M1-manual	**SD-M1-auto**	**SD-F1-manual**	**SD-F1-auto**	**SA-M1-manual**	**SA-M1-auto**	**SA-F1-manual**	**SA-F1-auto**
No. of sentences	1936	1936	1937	1937	104	104	164	164
Duration [minutes]	190	190	128	128	10	10	11	11
Number of phonemes	80964	81053	80893	81058	4281	4370	6934	7099
Correct phonemes	80964	80380	80893	80663	4281	3697	6934	6674
Deletions	-	32	-	51	-	32	-	51
Substitutions	-	57	-	114	-	57	-	114
Insertions	-	584	-	260	-	584	-	260
Number of corrections	-	673	-	425	-	673	-	425
PER	0%	0.83%	0%	0.52%	0%	15.5%	0%	6%

Table 5. Precision of automatic segmentation (phoneme boundary).

	10-19ms	**20-29ms**	**30-39ms**	**40-49ms**	**50-59ms**	**>60ms**
SD-M1-auto	17238	5355	1664	555	188	169
SD-F1-auto	13854	2317	656	227	91	92
SA-M1-auto	884	264	86	25	8	6
SA-F1-auto	1037	148	36	15	7	4

Evaluation: to determine if manual correction of the labels leads to an increment in speech quality CMOS and MOS listening tests were carried out. Figure 6 shows the results of the CMOS test. There was no significant difference between automatic labelling and manual correction of automatic labels in the case of *SD-F1*, *SA-M1*, *SA-F1* voices. Manual correction caused a significant improvement in speech quality in the case of *SD-M1*, although results of MOS tests did not show significant difference in either case. The results of the MOS test are introduced in the dissertation in detail.

Conclusion: according to the results there are cases, when it is possible to create consistently good speech quality without manual correction of automatic labels, thus a remarkable amount of work can be saved in HMM-TTS systems. In the speaker adapted cases CMOS and MOS tests did not show significant difference. This result makes it reasonable to investigate the error ratio, which still does not influence speech quality if only automatic methods are used. If generative models can produce similar quality even with higher phoneme error ratios, than automatic speech recognizer (ASR) transcription based speaker adaptation may be possible. I investigate this topic in Thesis Group 2.

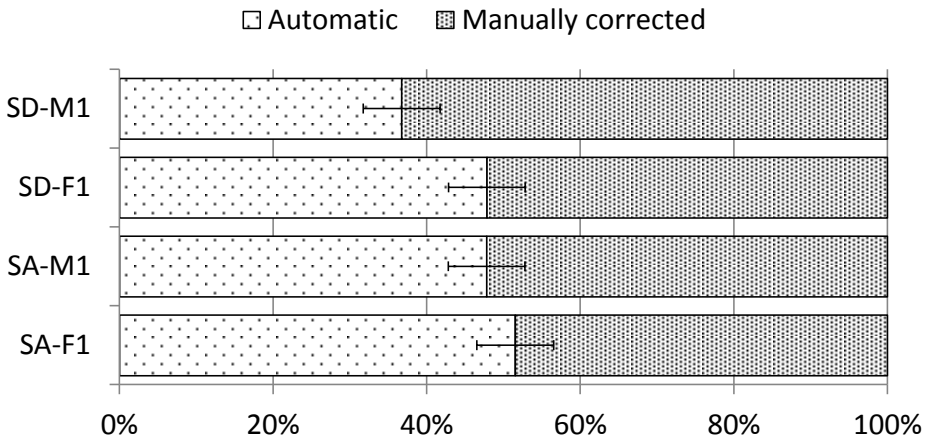

Figure 6. Subjective evaluation of automatic labels and manually corrected automatic labels with CMOS listening test.

5.2. Thesis Group II. Unsupervised speaker adaptation of hidden Markov-model based text-to-speech synthesis with semi-spontaneous speech.

The results of Thesis I.4 prepared the vision of completely automatic creation of new HMM-TTS voices; thus waveforms would be enough for speaker adaptation. Automatic creation of HMM-TTS voices makes sense in the case of spontaneous and semi-spontaneous

speech corpora, because planned speech usually has a phonetic transcription; consequently there is no reason for unsupervised speaker adaptation. I conducted research with semi-spontaneous speech.[1]

Based on the results of related research I suggested a novel solution: the transcription of automatic speech recognizer (ASR) is used as the basis of the adaptation database. Phoneme boundaries are determined by forced alignment with an automatically controlled beam.[2] Thus the method can be applied to ASRs even if confidence is not available. (I investigate previous works of unsupervised speaker adaptation in my related papers and in my dissertation in detail.)

First I developed a segmentation and selection algorithm suitable for semi-spontaneous speech. The goal of segmentation is to determine virtual sentences of semi-spontaneous speech. The goal of selection is to select an optimal subset of adaptation data. I performed subjective evaluation with different automatically created adaptation databases. These databases had 0%, 17%, 21%, 42%, 52%, 55%, 68%, 70%, 88% and 89% of phoneme error rate (PER), respectively. In practice it is likely that high quality recordings are not available and recognition accuracy varies. Consequently it is beneficial to test the solution with wide restrictions.

The procedure which is described in Thesis Group II contains language specific components, although the applied methodology is language independent.

Thesis II.1. [C1, C5, C6] *I designed and implemented an unsupervised procedure for speaker adaptation with semi-spontaneous speech based on the transcription of an automatic speech recognizer. I showed that it is possible to create not significantly different quality with the proposed method from supervised speaker adaptation.*

I created a method for the segmentation of semi-spontaneous speech and I had it recognized with a Hungarian ASR system. Furthermore I determined the phoneme boundaries with forced alignment. The ASR gave word level output, so forced alignment had to be performed in a separate step. From the resulting speech corpus I dismissed with an automatic method the utterances that are not favourable for HMM-TTS. I selected 10 minutes of the speech corpus randomly, and I created the manual transcription of these 10 minutes for reference. The average voice model was trained with the same corpora that were introduced in Thesis Group I.

In the first phase I created HMM-TTS voices with semi-spontaneous speech from four speakers. The PER varied between 10…42% (see Table 6). In the second phase I made further experiments with a male speaker's speech corpus (*M8*); the PER in this case was between 17…89%. The features of these corpora are described in the dissertation and related publications in detail.

Evaluation: for subjective evaluation CMOS and MOS listening tests were carried out. According to the results the quality of synthetic speech increases as PER decreases. When

[1] Semi-spontaneous (or semi-reproductive) is speech that has the features of live speech, although the speaker has previously planned it, usually in written form.
[2] Beam is a parameter of forced alignment.

PER was lower than 55% there was no significant difference between unsupervised and supervised cases. The results of CMOS tests show (Figure 7.) that supervised and unsupervised systems were considered similar in *M8-RND* and *M9-RND* cases, and even in *M6-RND* and *M7-RND* cases the difference was not significant. In case of higher PER significant difference was measured. I investigate higher PERs in Thesis II.2.

Conclusion: the results are quite surprising, because they state that the ASR transcription based speaker adaptation is not significantly different from the supervised case. This is the extension of Thesis I.4, because not only the phonetic transcription and segmentation is done automatically, but the textual transcription and utterance selection is determined automatically as well.

Table 6. Semi-spontaneous adaptation speech corpora for unsupervised speaker adaptation.

Symbol	Speaker	Method	Selection	Duration	PER[3]	WER[4]
M6-S-RND	Male 6.	Supervised	Random	11.4 min	"error free"	
M6-U-RND	Male 6.	Unsupervised	Random	11.4 min	42%	87%
M7-S-RND	Male 7.	Supervised	Random	9.6 min	"error free"	
M7-U-RND	Male 7.	Unsupervised	Random	9.6 min	21%	74%
M8-S-RND	Male 8.	Supervised	Random	10.2 min	"error free"	
M8-U-RND	Male 8.	Unsupervised	Random	10.2 min	17%	57%
M9-S-RND	Male 9.	Supervised	Random	9.7 min	"error free"	
M9-U-RND	Male 9.	Unsupervised	Random	9.7 min	10%	44%

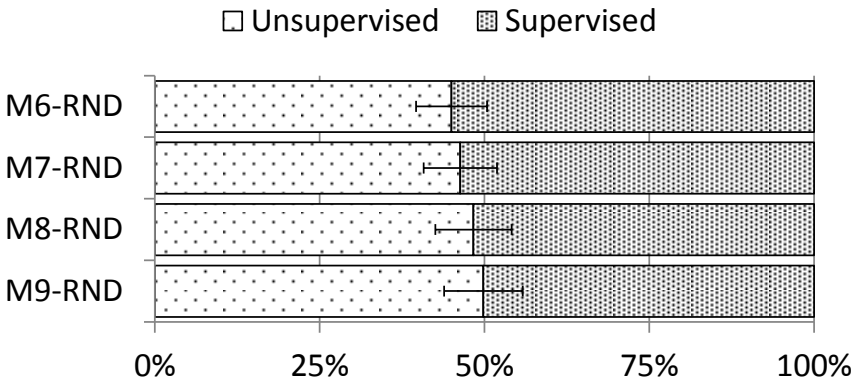

Figure 7. Quality evaluation of unsupervised speaker adaptation with semi-spontaneous speeches under 50% PER by CMOS pair comparison.

Thesis II.2. [C1, C3, C5] *I designed and implemented an unsupervised method to select a favourable subset of a speech corpus for speaker adaptation, and I showed that it is possible to create better synthetic speech quality with the proposed method than with random selection of the adaptation speech corpus.*

The method, which was described in Thesis II.1, resulted in not significantly worse synthetic speech quality in an automatic way. The PER of the adaptation corpus was smaller

[3] PER: Phoneme Error Rate
[4] WER: Word Error Rate

than 50% in Thesis II.1. In the current thesis I investigate how the method can be enhanced where PER of the speaker adaptation corpus is larger than 50%. Based on the results of previous research (which are introduced in the dissertation), of Thesis I.4 and II.1 I designed and implemented the following method: segmentation, automatic speech recognition and phoneme boundary detection was performed as in Thesis II.1, although the selection method was modified. The optimal value of the beam is defined by the quality of waveform and the errors in ASR transcription. Furthermore my goal was to select about 10 minutes of adaptation data from any speech corpus. In unsupervised speaker adaptation varying quality of utterances are probable, so the quality of ASR transcription is not predictable, thus in empirical terms an exact beam value cannot be determined. Therefore the width of the beam is set in an iterative way to find its optimal value, when the length of successfully forced aligned wave files is closest to 10 minutes (*t_limit*).

Each virtual sentence of semi-spontaneous speech is represented by one wave file. I run forced alignment on these files with the set beam width, and I investigate the length of the successfully forced aligned wave files (denoted by *t_adaptation_corpus*). I search for the optimal beam width with *bisection method*. The core of the method may be written in pseudocode as follows:

```
 1. i=0
 2. beam_max=beam[0]=maximum beam width
 3. beam_min=0
 4. t_limit=10 minutes
 5. DO
 6.    CALL forced alignment WITH beam[i] on each wave file
       RETURNING t_adaptation_corpus[i]
 7.    IF t_adaptation_corpus[i]>t_limit THEN
 8.       beam_max=beam[i]
 9.       beam[i+1]=beam[i]-floor((beam[i]-beam_min)/2)
10.    ELSE
11.       beam_min=beam[i]
12.       beam[i+1]=beam[i]+floor((beam_max-beam[i])/2)
13.    END IF
14.    i++
15. WHILE beam[i] != beam[i-1]
```

The method stops, when the beam value is the same in two consecutive steps. Next full-context labelling of phonetic transcription is done and the result is a speaker adaptation corpus. At the beginning of the research the quality of ASR was high due to high-quality, domain specific recordings. In order to investigate the deeper effects of phoneme errors I simulated worse recognition results with 0-gram language models[5] and additive noise. With these settings the method can be practically tested, because varying quality and domain-free utterances are likely in general.

The adaptation corpora with higher than 50% PER are summarized in Table 7. All of these corpora were generated in an unsupervised way (denoted by *U*), and the method, which was introduced above, is denoted by *BBS* (Beam-based selection). To be able to measure the effectiveness of the BBS method, I also created adaptation speech corpora with the random selection method, and these are denoted by *RND* (random). *0G* means 0-gram

[5] 0-gram means each morpheme occurs once, with the same probability in the language model.

language model, *NOISE* and *NOISE2* mean -50 dB and -25 dB additive white noise compared to the maximum level. The maximum level of the original recordings was normalized to 0 dB per sentence. With the speech corpora in Table 7 I created speaker adapted HMM-TTS voices.

Evaluation: a CMOS and a MOS listening test were carried out to determine the efficiency of the beam-based selection method. The naturalness and the similarity to the target speaker were measured by MOS tests, the preference score by CMOS tests. The results of the CMOS tests are shown in Figure 8. In the case of higher phoneme error rates (*M8-U-0G-NOISE, M8-U-0G-NOISE2*) the proposed method (*BBS*) resulted in significantly better speech quality than the random selection method. The results of the MOS tests (which are described in the dissertation) show significant difference in the case of NOISE2.

Conclusion: in the listening tests the proposed method gave significantly better results, even in the case of high phoneme error rates (i.e. bad recognition results with additive noise), than with random selection of adaptation data.

Table 7. Semi-spontaneous speech corpora with simulated bad recognition results for unsupervised speaker adaptation.

Symbol	Speaker	Language model	Noise	Duration	PER	WER
M8-U-0G-RND	Male 8.	0-gram	-	9.5 min	55%	100%
M8-U-0G-BBS	Male 8.	0-gram	-	10 min	52%	100%
M8-U-0G-RND-NOISE	Male 8.	0-gram	-50 dB	8.9 min	70%	100%
M8-U-0G-BBS-NOISE	Male 8.	0-gram	-50 dB	9.4 min	68%	100%
M8-U-0G-RND-NOISE2	Male 8.	0-gram	-25 dB	9.7 min	89%	100%
M8-U-0G-BBS-NOISE2	Male 8.	0-gram	-25 dB	10.2 min	88%	100%

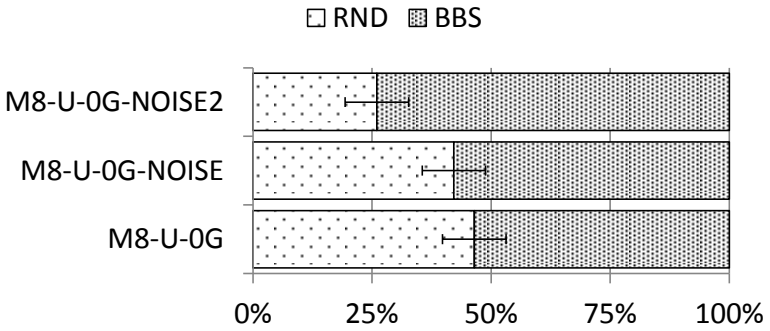

Figure 8. Subjective evaluation of RND and BBS methods with CMOS listening tests.

5.3. Thesis Group III. Optimizing Hidden Markov-model based text-to-speech synthesis for low-resource devices.

HMM-TTS speech generation runs faster than real-time on modern desktop computers. On low-resource devices, i.e. smartphones, the calculations must still be optimized to achieve low response times with near real-time functionality. Optimizing a speech synthesis system on mobile devices is a challenging task because both the storage capacity and the computing power are limited. The latest high-end mobile devices possess large storage size

and high performance CPU. Speech synthesis still needs to compete with other applications for precious storage space and the computing power is also shared among system and third party processes. A further disadvantage of resource demanding computations is that they cause higher power consumption and shorter battery life.

My research includes the introduction of codebook based noise excitation; investigation of the relationship between line spectral pairs, parameter streams and perceived quality; furthermore, the parallelization of parameter generation, vocoding algorithm and waveform playback. In this thesis group the most computational power demanding parts of synthesis are determined and I design and implement methods for decreasing the computation and response times. I investigate the synthetic speech quality after each incremental step with listening tests. I measure the required time of loading the database, parameter generation algorithm and vocoder algorithm until the response of the system. In the following I refer to these three stages as (1), (2) and (3). I carried out the measurements on three different smartphones, which are shown in Table 8. In this thesis group I carried out the research with an English speech corpus (CMU ARCTIC / SLT) [13].

Table 8. The devices used in the experiment of optimizing HMM-TTS.

Device	**CPU type**	**Speed [MHz]**
Mob1 (iPhone)	Samsung ARM 11	412
Mob2 (Spica)	Samsung S3C6410	800
Mob3 (Desire)	Qualcomm QSD8250	1000

Thesis III.1. [J1, C4] *I designed a low-resource model for hidden Markov-model based speech synthesis and I showed experimentally that without significant loss in quality the proposed model significantly improves the performance.*

The excitation of unvoiced sounds is modelled with Gaussian noise in the case of impulse-noise excitation based vocoders. The Box-Muller procedure generates Gaussian noise [23]. This method is widely used in HMM-TTS systems as well. Codebook based Gaussian noise generation which uses only integer operations achieved a significant increment in performance (~ten times faster) compared to floating-point arithmetic [24]. Codebook based noise generation and integer operations mean loss in precision compared to the Box-Muller procedure, although significant loss in perceived quality on mobile phones was not expected. Thus I introduced this method on low-resource devices.

Next I modified the modelling of spectral parameters. Generally HMM-TTS systems use MGC (Mel-Generalized Cepstrum) and MGC-LSP (Mel-Generalized Cepstrum-Line Spectral Pairs) [25]. MGC is the generalized logarithm of the spectrum modified by the perception based Mel-scale. The spectral filtering with MGC and MGC-LSP parameters is usually completed with MLSA (Mel Log Spectrum Approximation) filters. The ideal transfer characteristics of MLSA filters cannot be realized, thus it is approximated by a 20^{th} order Padé approximation in practice. So the complexity of speech synthesis depends on the order of spectral analysis and the order of Padé approximation. If we change MGC and MGC-LSP spectral modelling to LSP, then the spectral filtering can be performed by LPC;

thus the complexity of the system will depend only on the order of spectral analysis. Further enhancement in speed can be achieved by reducing the order of LSP analysis, although it affects the quality of synthetic speech as well. I created HMM-TTS systems with 24th, 22nd, 20th, 18th, 14th, 12th and 10th order LSP analysis. (In the case of 24th, 22nd and 20th order filters the listening test were not carried out with all the test subjects, because preliminary listening tests by speech experts did not show significant differences to the 18th order LSP case.) The depth of decision trees influences both computational cost and footprint size. Fewer leaves and leaf nodes decrease computational cost and footprint size. Smaller decision trees result in degradation in speech quality, because larger sets of parameters are clustered in leaves. The number of leaves in decision trees used during further optimization is shown in Table 9.

Evaluation: all these steps causes loss in speech quality, consequently it is important to investigate, if this loss is significant. The calculation time measurements were performed incrementally, and a listening test was carried out with all the resulting systems. The results of calculation time measurements are shown in Figure 9 and the results of listening tests are shown in Figure 10. Figure 9 shows (1), (2) and (3) parts of the calculation time measurements in one column so the response time of the system can be easily seen (the sum of the three parts).

Table 9. Decision trees with different sizes used for optimization.

Symbol	Number of leaves in decision tree			Size [KByte]
	LSP	$LogF_0$	Duration	
Baseline	2883	3545	555	666
#1	2282	2104	376	463
#2	1227	1344	172	214
#3	651	543	79	140

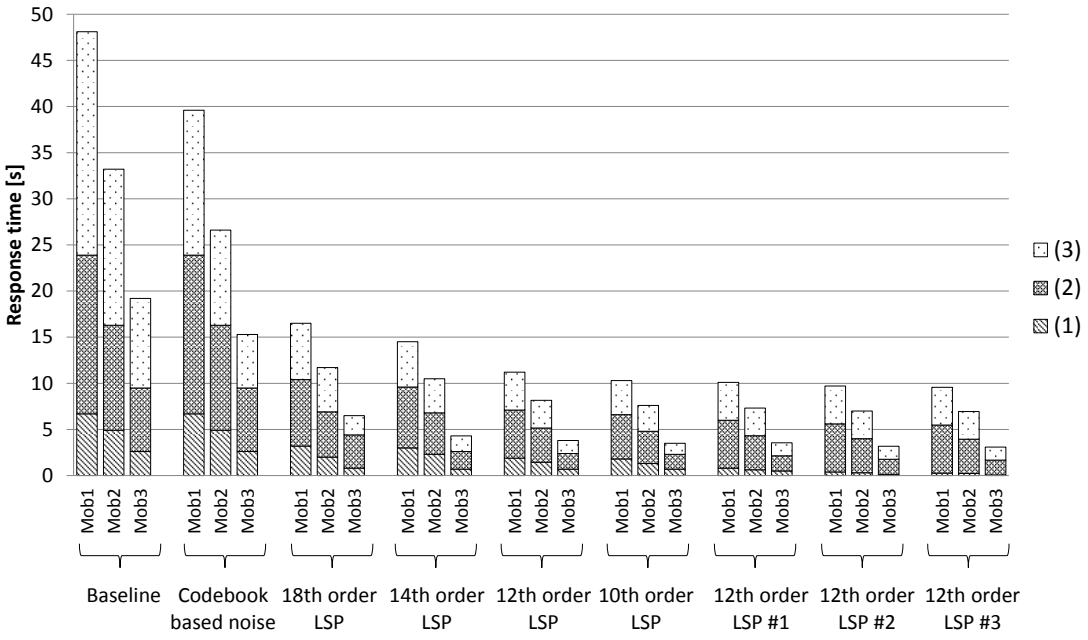

Figure 9. Calculation time measurements of HMM-TTS system on low-resource devices.

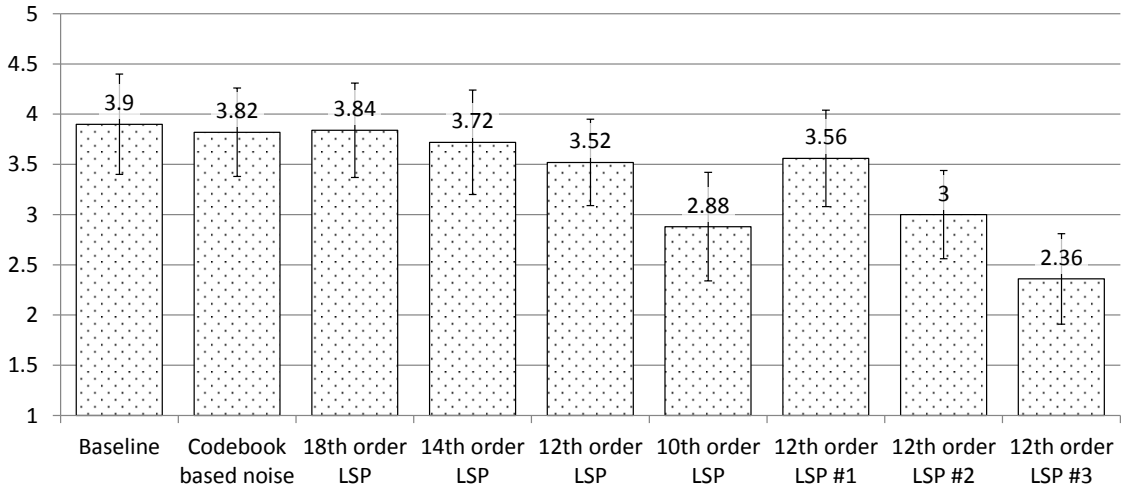

Figure 10. Subjective evaluation of Thesis III.1 optimization steps with MOS listening tests.

Conclusion: in the case of 12th order LSP with codebook based noise generation, and about 30% reduction of the size of the decision tree there was no significant loss in quality, while the calculation times became about five times faster. In the other cases there was either less improvement in performance or quality decreased significantly.

Thesis III.2. [J1, C4] *I designed and implemented a parallel method for resource demanding processes of HMM-TTS synthesis (parameter generation, vocoder algorithm) taking into account the actual load, and I showed experimentally that without loss in quality the proposed method significantly improves the response time.*

After the optimization steps of Thesis III.1 I designed a method to reduce the response time of HMM-TTS system. This method does not affect the quality of the synthetic speech. I extended the time-recursive algorithm of parameter generation as follows [26]: the vocoder algorithm and waveform playback is done in segments, and segment size is set according to the performance and actual load of the system. In general in text-to-speech engines waveform playback is not realized in order to remain platform independent. Introducing waveform playback the response time of the system can be reduced by the parallelization of parameter generation, vocoder algorithm and waveform playback, although platform specific steps must be taken. The parallelization can be realized in the following way (I define segment as a parameter stream of *k* frames):

1. The time-recursive parameter generation algorithm is calculated for the given segment (*k* frames), and the parameter stream is passed to the vocoder algorithm. The computation continues with the next segment.
2. The vocoder algorithm generates waveform from the parameter stream of the segment.
3. The segment's waveform added to the playback queue.

I determine the segment's length in runtime on the analogy of audio playback in computer networks. Ramjee et al. designed a network audio playback method [27], which I will introduce briefly in the following. Let n^i be the delay of the *i*-th audio packet. Delay estimate (d^i) and variation (v^i) of every incoming packet is calculated in the following way:

$$\hat{d}^i = A * \hat{d}^{i-1} + (1 - A) * n^i \tag{3}$$

$$\hat{v}^i = A * \hat{v}^{i-1} + (1 - A) * \left|\hat{d}^i - n^i\right| \tag{4}$$

Equations (3) and (4) are calculated for each packet, but they are used after pauses only. The time at which packet *i* is played out at the receiving host after pauses is given by:

$$p^i = \hat{d}^i + B * \hat{v}^i \tag{5}$$

The A constant in equations (3) and (4) determines the memory of the approximation, the delay / packet loss ratio is defined by the B constant in equation (5). In practice *A=0.998002* and *B=4* values are often used.

I tailored this method to the speech synthesis in HMM-TTS systems. Let's denote the time required for parameter generation and vocoder algorithm of the *i*-th segment with n^i. The values of d^i, v^i and p^i are calculated according to equation (3)-(5) and the initial values are $d^1=n^1$, $k^1=30$, $v^0=0$ and the constants are *A=0.99, B=4* (*i>0*). The number of frames in the *i+1*-th segment (k^{i+1}) is calculated after every 60 frames (T_{frame} is the length of the frame, 25 ms respectively in the experimental system):

$$k^{i+1} = \left\lceil \frac{p^i}{T_{frame}} \right\rceil \tag{6}$$

The schematic block diagram of the method can be seen in Figure 11. and it is described in the dissertation in detail.

Evaluation: the method, which is described above does not influence the speech quality; consequently listening tests were not necessary. The values of the calculation time measurements are shown in Figure 12.

Conclusion: the results of Thesis III.2 show about five times improvement in response time compared to the system of Thesis III.1. Compared to the baseline system the response time is about twenty times faster according to the results of Thesis Group III.

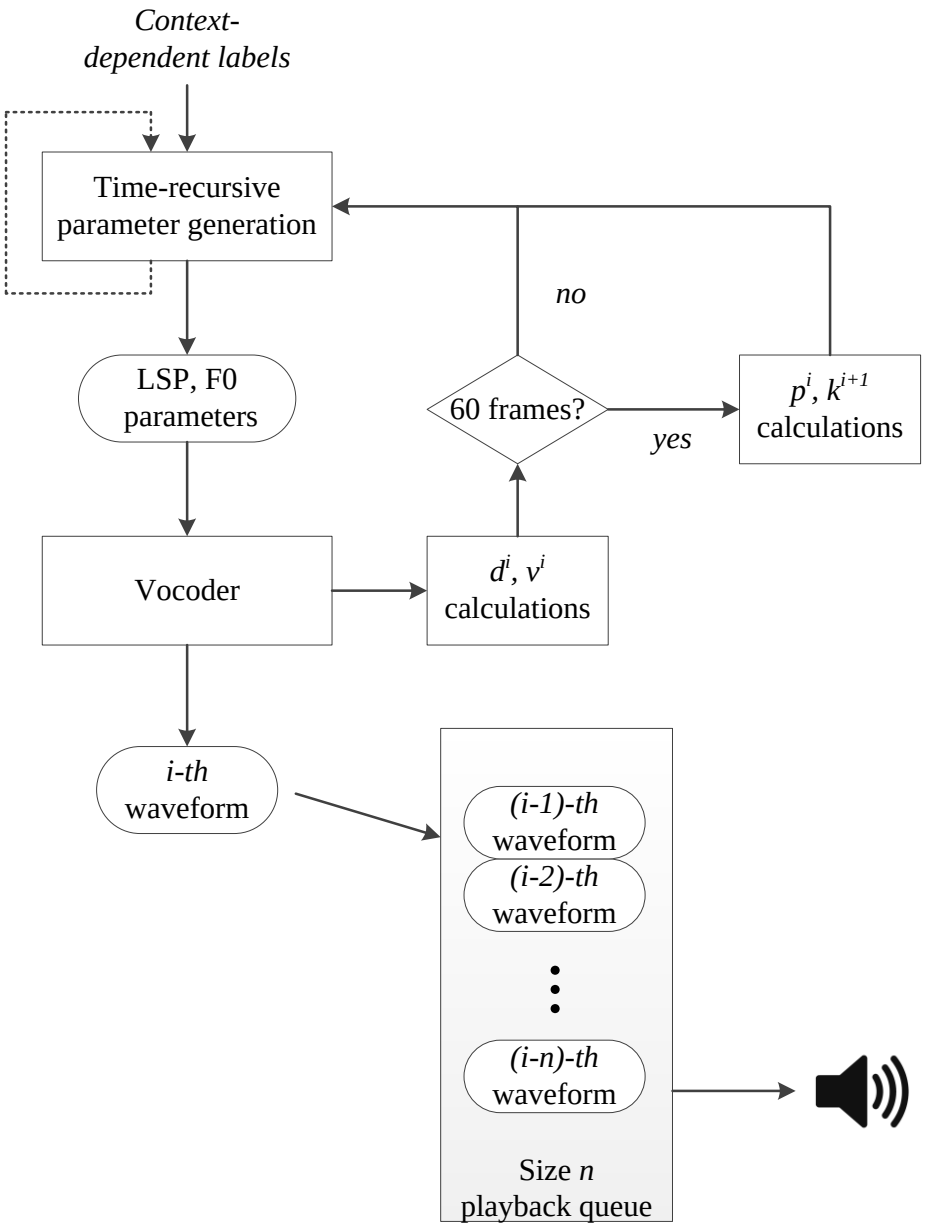

Figure 11. Schematic block diagram of parallelization of parameter generation, vocoder algorithm and waveform playback in HMM-TTS systems based on the actual load.

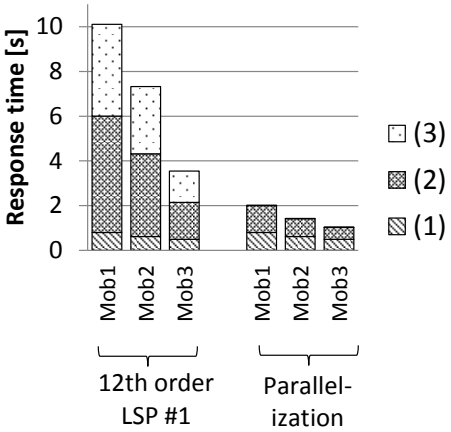

Figure 12. Enhancement of response times after parallelization.

6. Practical application of scientific results

Besides the theoretical outcomes of the thesis groups, as introduced in this booklet, their practical application is also an important factor.

A high-quality, domain free Hungarian text-to-speech engine is created based on the results of theses I.1. and I.2. The general application of this TTS engine makes several speech enabled systems possible, e.g. screen readers for blind users, interactive voice response systems, prompt generators and additional systems with speech user interface. The results of Thesis I.2 could be extended to other languages. Based on the results of Thesis I.3 10-15 minutes of utterances are enough to create new voice characteristics. Thesis I.4 suggests that manual correction of the adaptation database is not absolutely necessary.

The novel outcome of Thesis Group II is the possibility of creating new HMM-TTS voices without manual work. With the unsupervised adaptation method, which is introduced in this thesis group, it is possible to create thousands of voices automatically, e.g., from a speech corpus of telephone conversations. The results also make it possible to tailor the voice of a particular system to a target speaker's voice characteristics. Thus a smartphone can learn the voice characteristics of its owner automatically. The results of Thesis II.1 give a method for unsupervised speaker adaptation in the case of better recognition results, while Thesis II.2 also makes it possible to extend the method to worse recognition results. These methods were tested with Hungarian speech corpora, although the method does not contain any language specific parts.

The systems from Thesis Group I and II could be realized on low-resource devices, i.e. smartphones, based on the results of Thesis Group III. The experimental system of this thesis group takes into account the performance and actual load of the system during synthesis on Google Android 2.x and 4.x smartphones. The application programming interface (API) level realization of the HMM-TTS makes a wider usage possible. The resulting system can be used in any speech enabled application on the device, including message readers (SMS, e-mail, social network messages, etc.), e-book readers, screen readers, navigation systems. The research of Thesis Group III was carried out with English HMM-TTS, although the solution does not contain any language specific element: it can be applied to other languages as well. Furthermore the voice characteristics of the mobile HMM-TTS system can be modified based on the results of Thesis Groups I and II.

The results of all thesis groups were implemented in experimental systems.

Acknowledgement

I am thankful to my supervisors, Dr. Géza Németh and Dr. Gábor Olaszy, for their patient support and encouragement during my research. They taught me a lot about speech synthesis and they guided me throughout my Ph.D. studies. I conducted my research in a positive, humane atmosphere with their leadership in the BME-TMIT Speech Technology Laboratory.

I would like also thank the colleagues of BME-TMIT Speech Technology Laboratory, including Mátyás Bartalis, Dr. Tamás Böhm, Tamás Csapó and Dr. Csaba Zainkó for their theoretical and practical help.

I wish also to acknowledge the help and support of Tibor Fegyó, Dr. Péter Mihajlik and Balázs Tarján in automatic speech recognition related questions.

I would like to express my appreciation to Dr. Péter Siptár for his great help in phonology, and to Dr. Alexandra Markó for her advice in linguistics.

Special thanks to Dr. Tamás Henk, the Head of the Department, for his help and supervision in the process of dissertation.

I thank Dr. Géza Gordos for all his encouraging advice and support; it is highly appreciated.

My sincere thanks go to Dr. György Takács and Dr. László Tóth for their encouraging comments and advice that helped me to improve the dissertation.

Last but not least, I would like to express my thanks to my family. I thank my father, Dr. Pál Péter Tóth for his inspiring comments and remarks during my research, and my mother, Dr. Klára Gyires for her general advice and guidance in carrying out research. I also thank my sister, Dr. Veronika Tóth for her selfless help every day. I wish to acknowledge the patience and support of my beloved Roberta Deák.

I dedicate my doctoral dissertation to the memory of my grandfather, Dr. Béla Gyires, academician.

My research was partly supported by NAP (OMFB-00736/2005), Teleauto (OM-00102/2007), BelAmi (ALAP2-00004/2005), ETOCOM (TÁMOP-4.2.2-08/1/KMR-2008-0007), TÁMOP-4.2.1/B-09/1/KMR-2010-0002, CESAR (No271022), EITKIC_12-1-2012-0001 and Paelife (Grant No AAL-08-1-2011-0001) projects.

References

[1] Kempelen, F.: Az emberi beszéd mechanizmusa, valamint a szerző beszélőgépének leírása. Szépirodalmi Könyvkiadó, Budapest (1989)

[2] Gordos, G., Takács, Gy.: Digitális beszédfeldolgozás. Műszaki Könyvkiadó, Budapest (1983)

[3] Németh, G., Olaszy, G., eds.: A magyar beszéd. Akadémiai Kiadó, Budapest (2010)

[4] Mermelstein, P.: Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53 (4), 1070-1082 (1973)

[5] Klatt, D. H., Klatt, L. C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America vol. 87., issue 2, 820-857 (1990)

[6] Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communications 9., 453–467 (1990)

[7] Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., Gordos, G.: Profivox – a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4., 201-215 (2000)

[8] Möbius, B.: Corpus-based speech synthesis: methods and challenges. Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition, 79–96 (2000)

[9] Németh, G., Olaszy, G., Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás 2006, 183-196 (2006)

[10] Zen, H., Tokuda, K., Black, A. W.: Statistical parametric speech synthesis. Speech Communication vol. 51, 1039-1064 (2009)

[11] Vicsi, K., Tóth, L., Kocsor, A., Gordos, G., Csirik, J.: MTBA - magyar nyelvű telefonbeszéd-adatbázis. Hiradastechnika Vol. LVII, NO.8, 35-43 (2002)

[12] Tóth, L., Kocsor, A.: Az MTBA magyar telefonbeszéd-adatbázis kézi feldolgozásának tapasztalatai. Beszédkutatás, 134-146 (2003)

[13] Kominek, J., Black, A. W.: The CMU Arctic speech databases. Proc. of 5th ISCA Speech SynthesisWorkshop, 223-224 (2004)

[14] Zen, H., Oura, K., Nose, T., Yamagishi, Y., Sako, S., Toda, T., Masuko, T., Black, A. W., Tokuda, K.: Recent development of the HMM-based speech synthesis system (HTS). Proc. of Asia-Pacific Signal and Information Processing Association, 121-130 (2009)

[15] Oura, K., Tamamori, A., Sako, S., Zen, H., Nose, T., Takahashi, T., Yamagishi, J., Nankaku, Y.: Speech Signal Processing Toolkit (SPTK), Version 3.5. (Accessed 2013) Available at: http://sp-tk.sourceforge.net/

[16] Kawahara, H., Masuda-Katsuse, I., Cheveign´e, 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 vol. 27, 187-207 (1999)

[17] Mihajlik, P., Fegyó, T., Tüske, Z., Ircing, P.: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages - like Hungarian. Proc. of Interspeech, 1497-1500 (2007)

[18] Krstulovic, S., Hunecke, A., Schröder, M.: An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. Proc. of Interspeech, 1897-1900 (2007)

[19] Tokuda, K., Zen, H., Black, A. W.: An HMM-based speech synthesis system applied to English. Proc. of IEEE SSW, 227-230 (2002)

[20] Gósy, M.: Fonetika, a beszéd tudománya. Osiris kiadó (2004)

[21] Durand, J., Siptár, P.: Bevezetés a fonológiába. Osiris Kiadó, Budapest (1997)

[22] Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. Proc of ICASSP, 805-808 (2001)

[23] Jeruchim, M. C., Balaban, P., Shanmugan , K. S.: Simulation of Communication Systems: Modeling, Methodology and Techniques., 383-384 (2000)

[24] Chu, P. L.: Fast Gaussian Noise Generator. IEEE Transactions on Acoustics, Speech and Signal Processing 37(10), 1593-1596 (1989)

[25] Zen, H., Toda, T., Tokuda, K.: The Nitech-NAIST HMM-Based Speech Synthesis System for the Blizzard Challenge 2006. Journal IEICE - Transactions on Information and Systems E91-D(6), 1764-1773 (2008)

[26] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis. Proc. of ICASSP vol. 3, 1315-1318 (2000)

[27] Ramjee, R., Kurose, J., Towsley, D., Schulzrinne, H.: Adaptive playout mechanisms for packetized audio applications in wide-area networks. Proc. of IEEE Infocomm, 680-688 (1994)

Publications

Publication related to Ph.D. Thesis

Journal papers

[J1] Tóth, B., Németh, G.: Optimizing HMM Speech Synthesis for Low Resource Devices, Journal of Advanced Computational Intelligence & Intelligent Informatics, Vol. 16, No. 2., 327-334 (2012)

[J2] Tóth, B., Németh, G.: Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis, Acta Cybernetica, 19(4), 715-731 (2010)

[J3] Tóth, B., Németh, G.: Hidden Markov Model Based Speech Synthesis System in Hungarian, Infocommunications Journal, Volume LXIII, 2008/7, 30-34 (2008)

[J4] Tóth, B., Németh, G.: Rejtett Markov-Modell Alapú Mesterséges Beszédkeltés Magyar Nyelven, Híradástechnika, Volume LXIII., 2-6 (2008) (in Hungarian)

Chapters in edited book

[B1] Tóth, B., Németh, G., Olaszy G.: Beszédkorpusz tervezése magyar nyelvű, rejtett Markov-modell alapú szövegfelolvasóhoz, Gósy M.: Beszédkutatás 2012, MTA Nyelvtudományi Intézet, 278-295 (2012) (in Hungarian)

[B2a] Tóth, B., Németh, G.: A rejtett Markov-modellen alapuló gépi szövegfelolvasás, Németh, G., Olaszy, G. (eds.), A magyar beszéd, 512-518 (2010) (in Hungarian)

[B3] Tóth, B., Németh, G.: Rejtett Markov-modell alkalmazása magyar nyelvű gépi szövegfelolvasóhoz, Gósy, M.: Beszédkutatás 2008, MTA Nyelvtudományi Intézet, 182-193 (2008) (in Hungarian)

Conference paper

[C1] Székely, É., Csapó, T-G., Tóth, B., Mihajlik, P., Carson-Berndsen J.: Synthesizing Expressive Speech from Amateur Audiobook Recordings, Proc. of IEEE Workshop on Spoken Language Technology, Miami, USA, 297-302 (2012)

[C2] Tóth, B., Berki, S., Németh, G.: Distinctive Features in a Hungarian Hidden Markov Model Based TTS System, Proc. of 53rd International Symposium ELMAR-2011, Zadar, Croatia, 213-216 (2011)

[C3] Tóth, B., Fegyó, T., Németh, G.: The Effects of Phoneme Errors in Speaker Adaptation for HMM Speech Synthesis, Proc. of 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, 2805-2808 (2011)

[C4] Tóth, B., Németh, G.: Some Aspects of HMM Speech Synthesis Optimization on Mobile Devices, Proc. of 2nd International Conference on Cognitive Infocommunications, Budapest, Hungary, 1-5 (2011)

[C5] Tóth, B., Fegyó, T., Németh, G.: Some Aspects of ASR Transcription based Unsupervised Speaker Adaptation for HMM Speech Synthesis, Proc. of 13th International Conference on Text, Speech and Dialogue (TSD), Brno, Czech Republic, 408-415 (2010)

[C6] Tóth, B., Németh, G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel, Proc of. VI. Magyar Számítógépes és Nyelvészeti Konferencia (MSZNY), Szeged, Hungary, 246-256 (2009) (in Hungarian)

Conference presentation summary

[C7] Tóth, B., Németh, G.: Hidden Markov Model Based Speaker Dependent and Adaptive Training of Hungarian Text-to-Speech System, Proc. of International Conference Probability and Statistics with Applications, Debrecen, Hungary, abstract (2009)

Additional publications

Chapters in edited book

[B2b] Tóth, B., Németh, G., Kiss, G.: Mobiltelefonba épített SMS felolvasó, Németh G., Olaszy G.: A magyar beszéd, 560-561 (2010) (in Hungarian)

[B2c] Viktóriusz, Á., Németh, G., Tóth, B.: NaviSpeech – beszélő navigátor látássérült gyalogosoknak, Németh G., Olaszy G.: A magyar beszéd, 591-595 (2010) (in Hungarian)

[B2d] Tóth, B., Németh, G.: Beszédkommunikátor beszédsérültek segítésére, Németh G., Olaszy G.: A magyar beszéd, 620-623 (2010) (in Hungarian)

[B4] Németh, G., Kiss, G., Zainkó, Cs., Olaszy, G., Tóth, B.: Speech Generation in Mobile Phones. In: Gardner-Bonneau, D., Blanchard, H. (eds.), Human Factors and Interactive Voice Response Systems, New York: Springer, 163-191 (2008)

[B5] Németh, G., Kiss, G., Tóth, B.: Cross Platform Solution of Communication and Voice/Graphical User Interface for Mobile Devices in Vehicles, In: Abut, H., Hansen, J. H. L., Takeda, K. (eds.), Advances for In-Vehicle and Mobile Systems: Challenges for International Standards, Springer, 237-250 (2007)

Conference papers

[C8] Tóth, B., Nagy, P., Németh, G.: New Features in the VoxAid Communication Aid for Speech Impaired People, Proc of. Computers Helping People with Special Needs: Lecture Notes in Computer Science. Linz, Ausztria, 295-302 (2012)

[C9] Németh, G., Csapó, T., Tóth, B.: Improving the Quality of Unit Selection and HMM based Speech Synthesis, Proc of. FuturICT, Budapest, Hungary (2009)
[C10] Tóth, B., Németh, G.: XML Based Multimodal Interfaces on Mobile Devices in an Ambient Assisted Living Scenario, Proc of. Workshop on Intelligent User Interfaces for Ambient Assisted Living, International Conference on Intelligent User Interfaces, Maspalomas, Gran Canaria, January 13-16 (2008)
[C11] Tóth, B., Németh, G.: Speech Enabled GPS Based Navigation System for Blind People on Symbian Based Mobile devices in Hungarian, Proc. of Regional Conference on Embedded and Ambient Systems, Budapest, Hungary, 69-74 (2007)
[C12] Tóth, B., Németh, G.: Challenges of Creating Multimodal Interfaces on Mobile Devices, Proc. of 49th International Symposium ELMAR-2007 focused on Mobile Multimedia, Zadar, Croatia, 171-174 (2007)
[C13] Tóth, B., Németh, G.: Creating XML Based Scalable Multimodal Interfaces for Mobile Devices, Proc. of 16th IST Mobile and Wireless Communications Summit, Budapest, Hungary, CD-ROM Proceedings (2007)
[C14] Németh, G., Kiss, G., Tóth, B.: Proposals for Extending the Speech Synthesis Markup Language (SSML) 1.0 from the Point-of-View of Hungarian TTS Developers, Proc. of W3C Second Workshop on Internationalizing SSML, Crete, Greece, (2006)
[C15] Tóth, B., Németh, G.: VoxAid 2006: Telephone Communication for Hearing and/or Vocally Impaired People, Proc. of 10th International Conference on Computer Helping People with Special Needs, Springer, Linz, Austria (2006)
[C16] Németh, G., Kiss, G., Tóth, B.: Cross Platform Solution of Communication and Voice / Graphical User Interface for Mobile Devices in Vehicles, Proc. of Biennial on DSP for in-Vehicle and Mobile Systems, Sesimbra, Portugal, CD-ROM Proceedings (2005)