Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Telecommunications and Media Informatics

Increasing the naturalness of synthesized speech in hidden Markov-model based text-to-speech synthesis

Ph.D. thesis booklet
BME-VIK Doctoral School of Informatics

Tamás Gábor Csapó
M.Sc. in Technical Informatics

Supervisor:
Géza Németh, Ph.D.

Budapest, Hungary
2013

1. Introduction

Research on human-computer interaction is important in the information society. Speech technology research (e.g. text-to-speech synthesis) fits into this process. The speech-driven communication between the user and the system is essential if the hands and vision of the user are busy (e.g. during car driving), or cannot be used because of disabilities (e.g. visually impaired), furthermore, if the used service is telephone-based (e.g. intelligent directory assistance, news reading on mobile devices). Expressive, emotion-imitating speech synthesis can be beneficial if the goal is to decrease the monotony during the reading of longer texts (e.g. audiobooks). The personalized text-to-speech that has the voice of a given person can be useful for those users, who have lost their speaking ability because of injuries or illnesses.

Numerous simplified models have been created for the investigation of speech production, which are mostly based on source-filter separation of speech [1]. The larynx, the voice source organ can be roughly modeled with a simple impulse sequence in voiced regions and with white noise in unvoiced parts. For modeling the vocal tract (oral and nasal cavities, etc.), which is the filter in the model, many types of methods have been developed. Statistical parametric speech synthesis, which is a state-of-the-art technique in the domain, also applies the source-filter model [2]. The modeling of the vocal tract has reached a mature level and thus the further quality improvement is not related to this part [3]. However, for modeling the source signal there is still lack of techniques which would ensure that the quality of statistical parametric speech synthesis can reach the naturalness provided by unit selection[1] [4]. The modeling of the voice source is nowadays still an active research field.

Most speech technology methods were developed for the processing of idealized speech. During ideal voiced phonation the vocal folds are vibrating quasi-periodically, meaning that only small differences are observable in the consecutive pitch periods. In natural speech speakers use other voice qualities from time to time (like irregular phonation), and significantly different pitch periods may also occur (e.g. with extremely high or low amplitude). Although there are methods for the analysis, detection and transformation of this phenomenon [5], few research has been done in the modeling of non-ideal speech in text-to-speech synthesis and voice transformation domains.

The above source-filter based models assume that the source and the filter are perfectly separable during human speech production. However, this does not hold in all cases and non-linear coupling may occur because of the interaction between

[1] The main concept of unit selection text-to-speech is to choose and concatenate segments of natural speech (e.g. words, word-groups) from a larger speech corpus.

the source and the filter. In recent years it has been shown that besides the larynx and the organs above it, the area below it (lower airways, e.g. lungs, trachea, bronchi) also influences the speech [6]. According to this concept the lower airway (subglottal, i.e. below the glottis) system contributes to the separation of vowel groups in terms of distinctive features. The speech technology relations of the subglottal resonances have only been investigated in initial experiments till now.

I decomposed my research into three thesis groups according to the above topics. In the first thesis group I introduce a novel excitation model and present an irregular-regular voice transformation method based on this. In the second thesis group I show the speech synthesis aspects of the excitation model and introduce two new irregular voice models that can be used in statistical parametric speech synthesis. In the third thesis group I deal with the interaction between the vocal tract and the lower airways: I present my research on the investigation of subglottal resonances in Hungarian speech.

2. Background

One of the state-of-the-art techniques of speech synthesis is the hidden Markov-model based text-to-speech synthesis (*HMM-TTS*), which is a sub-category of the statistical parametric speech synthesis systems [3]. One research tool in this context is the open source HTS system [2]. In recent years HMM-TTS has gained much popularity due to its advantages: flexibility, low footprint and smoothness [3].

During statistical parametric speech synthesis the waveforms of the speech database are not processed directly, but the speech signal is decomposed into parameters which are fed to a machine learning system. From the training database, which is typically a few hours long, the excitation and spectral parameters are extracted first. Using these parameters the HMMs can be trained with the help of the phonetic transcription of the speech corpus and the contextual labels of the speech sounds. The result of the training is the small HMM database, which can be used in synthesis. During text-to-speech synthesis, first the phonetic transcription and context dependent labeling is performed on the input text. After that the parameters belonging to the labeled text are generated using the HMM database. From the generated excitation and spectral parameters a vocoder can synthesize the speech waveform [7].

2.1 Excitation models in statistical parametric speech synthesis

Most HMM-TTS systems are based on the source-filter decomposition of speech [1]. The simplest, pulse-noise method is shown in Fig. 1: voiced segments are modeled with a pitch-dependent ($F0$) impulse train, while unvoiced parts are modeled with white noise. After concatenating the frames and gain scaling the all-pole spectral filtering results in the speech signal. However, the simple pulse-train excitation in the baseline HTS makes the quality of the system 'buzzy', robotic compared to unit selection systems. In buzzy speech we mean the metallic, robotic, machine-like voice that is the result of vocoding. In order to overcome this drawback, numerous improved excitation models have been proposed in the literature. They can be categorized according to the properties of the excitation signal and the type of spectral filtering.

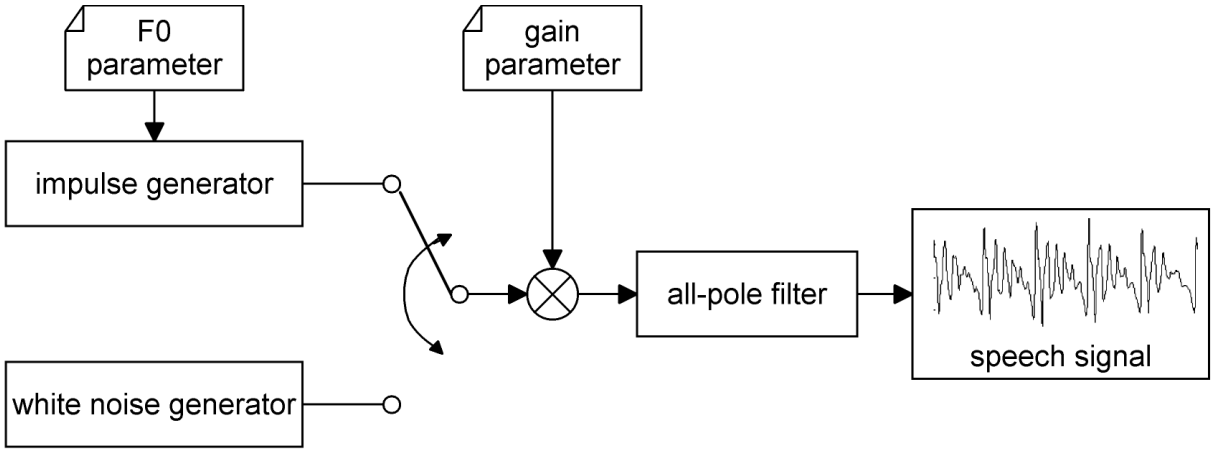

Fig. 1: Pulse-train excitation in the baseline HTS system. Source: [7], with modifications.

Mixed excitation and STRAIGHT-based vocoding [8] have been found to produce high quality HMM-based synthesized speech, but these are hard to use in real-time applications because of their high computational requirements. Mixed excitation is particularly useful for modeling sounds which do not have clearly voiced or unvoiced characteristics, but are produced as a mix of these (e.g. voiced fricatives).

Glottal source parameters are expected to be a suitable framework for describing the glottal excitation mechanism of speech. Cabral uses the Liljencrants-Fant (LF) acoustic model of the glottal source derivative to construct the excitation signal and introduces Glottal Spectral Separation [9]. Raitio and his colleagues use glottal inverse filtering within HMM-based speech synthesis for generating natural sounding synthetic speech (GlottHMM, [10]). The single pulse technique is

extended with a glottal pulse library and unit selection, but in the latest experiments it is shown that a mean-based excitation scheme is of similar quality than the complex unit selection of glottal pulses [11]. Overall the methods applying glottal source can synthesize high quality speech, but there are stability problems between voiced and unvoiced segments.

Several solutions have proposed the application of the Harmonic Plus Noise model (*HNM*) within the HTS framework and use maximum voiced frequency [12] or voicing cut-off frequency [13]. The advantage of these systems is that by applying stochastic noise in the higher spectral bands the buzziness of synthesized speech can be decreased.

Numerous approaches make use of the residual signal of speech. A great advantage of these models is that the residual can be obtained directly from the speech signal with inverse filtering. In [13], the residual is parameterized by the amplitude spectrum and zero-phase criterion is used to synthesize the excitation frame. In another approach, Waveform Interpolation (*WI*) is used and extended with time and frequency domain zero padding techniques in order to reduce the spectral distortion [14]. Drugman and his colleagues construct a codebook of pitch-synchronous residuals which is compressed with Principal Component Analysis (*PCA*) [15]. In [16], the Deterministic Plus Stochastic Model (*DSM*) of the residual signal is proposed and integrated into HTS. Excitation periods are the result of resampling an 'eigenresidual' to the target $F0$. An advantage of these residual-based models is that they may be applied for synthesizing different voice qualities with the proper manipulation of the residual signal.

Statistical parametric speech synthesis and most of the above excitation models are optimized for modal voices and may not produce high quality with voices having frequent non-modal sections. A specific reason for such a different phonation mode is irregular voice.

2.2 Occurrence and modeling of irregular voice

During ideal voiced phonation (modal or regular voice) in human speech, the vocal cords are vibrating quasi-periodically. For shorter or longer periods of time instability may occur in the larynx causing irregular vibration of the vocal folds. This is a non-modal phonation type and is referred to as irregular phonation, glottalization, creaky voice, vocal fry or laryngealization. It arises from abrupt changes in the length and/or amplitude of the pitch periods or both. Irregular phonation is a frequent phenomenon in both healthy speakers and people having voice disorders, usually in phrase boundaries (e.g. end of sentence) or in vowel-vowel transitions. It is often accompanied by extremely low pitch and the quick

attenuation of glottal pulses [17]. Glottalization is perceived as a creaky, rough voice [5]. Bőhm and his colleagues found, that up to 15% of the vowels of several American English speakers are produced with irregular phonation, therefore this phenomenon is not negligible in normal speech [18].

Fig. 4 shows an example for irregular (a) and regular (e) phonation (horizontal arrow denotes the section where the phonation is irregular). Glottalization can cause problems for standard speech analysis methods (e.g. $F0$ tracking and spectral analysis). Proper modeling of irregular phonation may contribute to building natural, expressive and personalized speech synthesis systems.

There are existing methods for classification of regular vs. irregular phonation [18, 19], for transforming modal voice to irregular [5], and for initial experiments with statistical parametric speech synthesis modeling creaky voice [20, 21, 22, 23]. We did not find any methods for irregular-regular transformation in the literature.

To model vocal fry in statistical parametric speech synthesis, Silén et al. [20] introduce a robust $F0$ measure, improved voicing estimation and two-band voicing, which removes glottalized sections from synthesized speech. However, it does not focus on the characteristics of creaky excitation of the original speaker. Drugman et al. derive an extension of the DSM model which can handle creaky excitation by integrating secondary pulses in the residual, and test this in analysis-synthesis experiments [21]. They investigate the usefulness of contextual factors for creaky voice prediction and experiments with adding parameter streams describing irregular phonation into the HMM-TTS framework [22]. This extended analysis-synthesis method with the creaky voice model and the new contextual factors have been recently integrated into the HTS-DSM vocoder combined with GlottHMM $F0$ estimation [23]. However, there is only a small difference compared to the baseline system in overall naturalness in case of including the modeling of creaky excitation.

2.3 Role of subglottal resonances in speech

The acoustic quality of speech sounds is determined not only by the larynx and the organs above it, but several properties of the respiratory system (e.g. volume of lungs, length of trachea) also influence them. The source-filter model does not deal with the nonlinear interaction between the source and the filter. The resonances of the lower airways (subglottal resonances, SGR) shape the spectrum of voiced sounds similarly to formants, but the formants amplify the nearby harmonics, whereas the frequency environment of the SGRs is acoustically disadvantageous. For this reason, it is hypothesized that during speech production we try to

avoid those articulatory positions when there might be interaction between the formants and the subglottal resonances. This way, the formants try to avoid the SGR frequencies. As the physiological dimensions of the trachea and the bronchi do not change significantly during speech, the subglottal resonance frequencies are nearly constant for a given speaker. The typical values of the first three SGRs are near 600, 1500 and 2300 Hz [6].

Recently it has been shown for several languages (American English [24], Spanish, German and Korean) that the resonances of the lower airways divide the vowels and consonants into distinct groups based on the frequency structure. These are in relation with specific categories (phonological distinctive features, [6]). Numerous theoretical approaches tried to explain these categories, of which one of the most successful is Quantal Theory (QT) [25]. Quantal Theory is based on the claim that in some regions of articulatory-acoustic space, small articulatory movements lead to large acoustic changes, while in other regions large movements lead to small acoustic changes [6]. A good example for this is the case of back and front vowels: the two vowel groups are differentiated by the horizontal tongue movement. QT has been recently extended with the SGRs [24], according to which the second SGR ($Sg2$) is a natural division line between front and back vowels in American English. The effect of the first subglottal resonance ($Sg1$) is less strong, but it was found that the $Sg1$ plays a role in dividing low and non-low vowels in terms of $F1$. The third subglottal resonance ($Sg3$) is often between the tense and lax vowels in American English [24].

Subglottal resonances can cause disfluencies of the formant track, are observable for human perception and can be useful for automatic speaker normalization [26]. However, the relation of vowel formants and SGRs have been investigated only for a few languages. There has previously been no research regarding the role of subglottal resonances in speech sounds for the Hungarian language.

3. Research objectives

With my research I wish to contribute to advance hidden Markov-model based speech synthesis to be more natural and help to understand the source-filter interaction. During the research, my specific aims are:

1) improve the naturalness of statistical parametric speech synthesis,
2) analysis and correction of irregular phonation to improve creaky, rough voices,
3) construction of irregular voice models for speech synthesis that can be used for expressive and personalized speech synthesis,
4) better understanding of source-filter interaction in human speech production, with special regard to the role of the subglottal system.

I chose these research aims because they contain numerous challenges. With my research I can contribute to more natural human-computer interaction. In my work I conducted the experiments on Hungarian speech, but most results are easily applicable for other languages as well. In the first thesis group I deal with the 1) and 2) research goals, while in the second thesis group 1) and 3) goals are dealt with. In the third thesis group I accomplish the 4) goal.

4. Methodology

During my research the success of the created methods was investigated using experimental methodology.

The experiments regarding speech analysis, synthesis and investigation of irregular voice (Thesis groups I and II) were conducted on the data of 5 native Hungarian speakers of the PPBA database [27]. I used the recordings of four males (FF1, FF2, FF3 and FF4) and one female (NO3), which were recorded in a professional studio and contain approximately 2 hours of speech for each speaker. I tested and validated my methods on Hungarian samples, but these models are language independent and it is expected that they can easily be used for other languages.

For investigating the subglottal resonances (Thesis group III), the speech and subglottal recordings were recorded during my research with Hungarian speakers[2]. The analyses were conducted partly on the logatom readings of 4 speakers [C4] (two males: Log_FF1, Log_FF2 and two females: Log_NO1, Log_NO2), in other part on the spontaneous speech and logatom readings [J4] of 6 speakers (five males: Spo_FF1 – Spo_FF5 and one female: Spo_NO1) from the BEA database [28]. The recordings were made in a quiet chamber, the subglottal signal was recorded with an accelerometer device attached to the skin of the neck below the thyroid cartilage. The textual and phonetic transcriptions and the labeling of the sound boundaries were prepared during the research with automatic tools and manual corrections.

During my research I used to following tools and software:

BME-TMIT forced alignment: automatic labeling of sound boundaries,
GLOAT / SEDREAMS: decompose speech signal to pitch-synchronous periods, http://tcts.fpms.ac.be/~drugman/Toolbox/
HTS: training of parameters with HMMs [2], http://hts.sp.nitech.ac.jp/
HTS-HUN: Hungarian version of HTS [7],
Matlab: speech analysis, synthesis, ROC analysis, t-test,

[2] Plural refers to the other persons involved in the research: Zsuzsanna Bárkányi, Tekla Etelka Gráczi, Tamás Bőhm, András Beke and Steven M. Lulich. The recordings, the manual measurements and corrections were performed collectively.

Praat: pitch detection; formant measurement; visual analysis of speech, http://www.fon.hum.uva.nl/praat/
SoX: lowpass filtering and resampling, http://sox.sourceforge.net/
SPTK: spectral analysis, inverse filtering and digital filtering, http://sp-tk.sourceforge.net/
Voice_Analysis_Toolkit / creak_detect: irregular voice detection [19], https://github.com/jckane/Voice_Analysis_Toolkit
VoiceSauce: correction of acoustic parameters; calculation of Harmonics-to-Noise Ratio, http://www.ee.ucla.edu/~spapl/voicesauce/
Wavesurfer: visual analysis of speech and accelerometer signal; measurement of acoustic parameters, http://www.speech.kth.se/wavesurfer/
Weka: decision tree implementation, http://www.cs.waikato.ac.nz/ml/weka/.

I investigated the results of the transformation and synthesis methods with perception (listening) tests. During the preparation of the experiments I used the proposed test types of the literature. In the listening tests the listeners answered either 1–5 scale MOS (*Mean Opinion Score*) questions, or in case of sample pairs they answered 1–3 or 1–5 scale CMOS (*Comparative Mean Opinion Score*) questions. The details of the perception tests can be found in Section 3.4 of the dissertation.

In the statistical analyses I applied t-tests, paired t-tests and one-way ANOVA analysis with Tukey-HSD post-hoc test in the Matlab and SPSS programs. During the analyses I reject the null hypothesis below two-sided $p < 0.05$ significance level (above 95% confidence level).

The abbreviations and notations are listed at the end of this thesis booklet.

5. New results

Thesis group I: A novel excitation model and its application for irregular voice correction

A number of speech analysis-synthesis methods exist in the literature. These have the original aim to decompose speech to parameters having as small bandwidth in the telecommunications channel as possible while the coded speech remains still fairly intelligible. Besides, in the field of speech processing nowadays it is more and more important to find such a parametric representation of speech which can be used for transformations and can be applied in machine learning systems. According to our initial experiments[3], the state-of-the-art speech coding methods (e.g. CELP, *Code-Excited Linear Prediction* vocoders) are not suitable

[3] In the following, I will use plural in favor of the easier reading. My own results are summarized in the theses.

for direct integration into machine learning systems (e.g. the codebook index of the CELP vocoder contains abruptly changing values which cannot be modeled easily with HMMs). The simpler excitation models of Section 2.1 result buzzy speech. With the more complicated models it is possible to synthesize speech with better quality, but often it is difficult to use them in real-time applications because of their larger computing requirements. Our goal was between these two extremes to create such an excitation model which has good quality and can be used in devices with limited resources as well.

This thesis group introduces a language independent residual-based excitation model which can decompose speech to parameters with analysis-synthesis methods. Its parameters are suitable for direct integration into the hidden Markov-model based machine learning. There are several similar excitation models in earlier solutions. The DSM method is residual based as well, but it does not use concatenation cost during the unit selection [15]. In the GlottHMM system both target cost and concatenation cost are used, but they are computed on the cycles of the glottal source signal [11]. According to this, the proposed model of thesis I.1 differs in essential parts from earlier systems. Besides, I introduce novel parameters for the representation of the residual signal which have not yet been used previously. In the additional thesis points I introduce the application of the model for a transformation that can decrease the perceived roughness of natural speech produced with irregular voice.

Thesis I.1: [C3] I created a novel residual codebook unit selection based language independent excitation model which is suitable for decomposition of speech to parameters (analysis) and restore it back (synthesis). During the analysis the method creates a codebook of pitch-synchronous residual periods derived from natural speech. During synthesis, the method applies automatic unit selection for selecting the units to be concatenated using both target cost and concatenation cost.

Fig. 2 shows the details of the analysis part. The input of the analysis method is speech waveform, which is lowpass filtered at 7.6 kHz and stored with 16 kHz sampling and 16 bit linear PCM quantization. The method first builds a codebook consisting of pitch-synchronous residual periods, then the analysis of the residual is performed. The $F0$ parameters are obtained by the publicly available Snack pitch tracker with 25 ms frame size and 5 ms frame shift. After that, MGC (*Mel-Generalized Cepstrum*) analysis is performed on the same frames. For the MGC parameters, we use $\alpha = 0.42$ and $\gamma = -1/3$. The residual signal (excitation) is obtained by MGLSA (*Mel-Generalized Log Spectral Approximation*) in-

verse filtering. Next, the SEDREAMS (*Speech Event Detection using the Residual Excitation And a Mean-based Signal*) algorithm is used to find the glottal period boundaries (GCI locations) in the voiced parts of the speech signal.

The further analysis steps are completed on the residual signal with 50 ms frame and 5 ms frame shift values. From the voiced parts two pitch period long signal parts are cut and are Hann-windowed. From these, a codebook is built with the following parameters for each frame:

F0: fundamental frequency of the signal,
gain: energy of the signal:
$gain_i = \sqrt{\sum_{j=0}^{N} r_j^2}$, where r_j is the jth sample of the ith windowed element,
rt0: the locations of peaks in the windowed element (example: Fig. 3),
HNR: Harmonics-to-Noise Ratio of the signal, based on cepstral harmonics [29].

For each voiced frame, one codebook element is saved with the given parameters and the windowed signal is also stored. The rt0 parameter is a 4-dimensional vector, which is a novel idea for describing the residual frames. The calculation of the parameter is shown in Fig. 3. The earlier solutions did not use such a parameter. The calculation of the parameter can be found in Section 4.1.1 of the dissertation. In order to collect similar codebook elements, the RMSE (*Root Mean Squared Error*) distance is calculated between the pitch normalized version of the codebook elements belonging to the same phoneme. During analysis, the above parameters are calculated for each voiced frame (if $F0 > 0$). For unvoiced frames ($F0 = 0$), only the gain parameter is calculated.

Fig. 2 shows the steps of the synthesis phase. The input of the synthesis are the parameters obtained during analysis ($F0$, $gain$, $rt0$, HNR and MGC) and the codebook of pitch-synchronous residuals. During the reconstruction, the residual is created frame-by-frame. If the frame is voiced, a suitable element with the target $F0$, $rt0$ and HNR is searched from the codebook. We apply target cost and concatenation cost with hand-crafted weights, similarly to unit selection speech synthesis [4]. The target cost is the squared difference among the parameters of the current frame and the parameters of those elements in the codebook. The concatenation cost shows the similarity of codebook elements to each other and it is calculated as the RMSE distance of the pitch normalized frames. When a suitable codebook element is found, its fundamental period is set to the target $F0$ by either zero padding or deletion. If the frame is unvoiced, white noise is used as excitation. Next, the residual is created by pitch synchronously overlap-adding the Hann-windowed residual periods and by concatenating the unvoiced parts. Finally, the energy of the frames is set using the $gain$ parameter sequence and synthesized speech is reconstructed by MGLSA filtering using the MGC parameters.

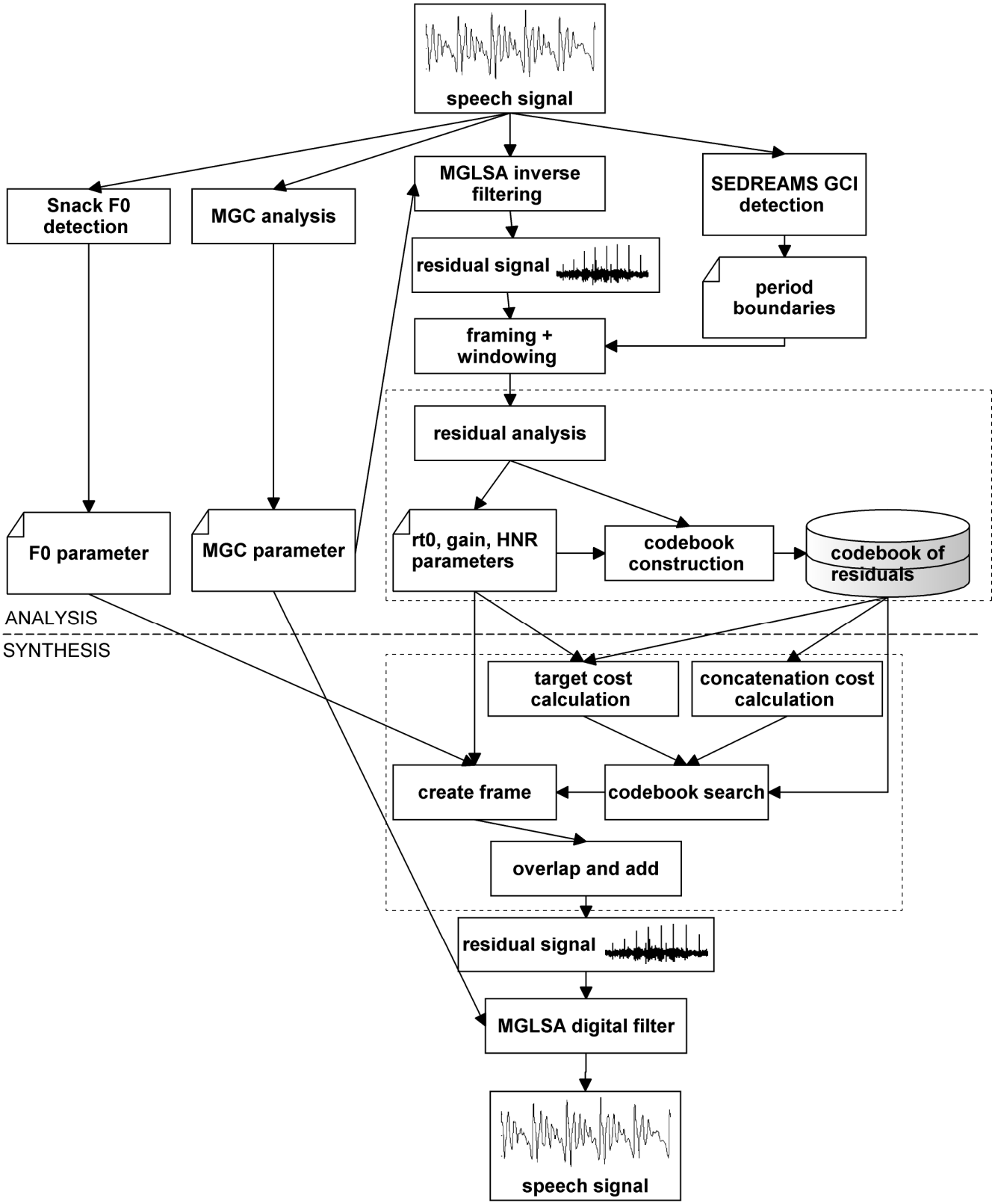

Fig. 2: Analysis (above the dashed line) and synthesis (below the dashed line) of speech signal with the method of Thesis I.1. Processes and waveforms are denoted by boxes; parameters are denoted by dog-eared boxes. The boxes with dashed line show my own methods.

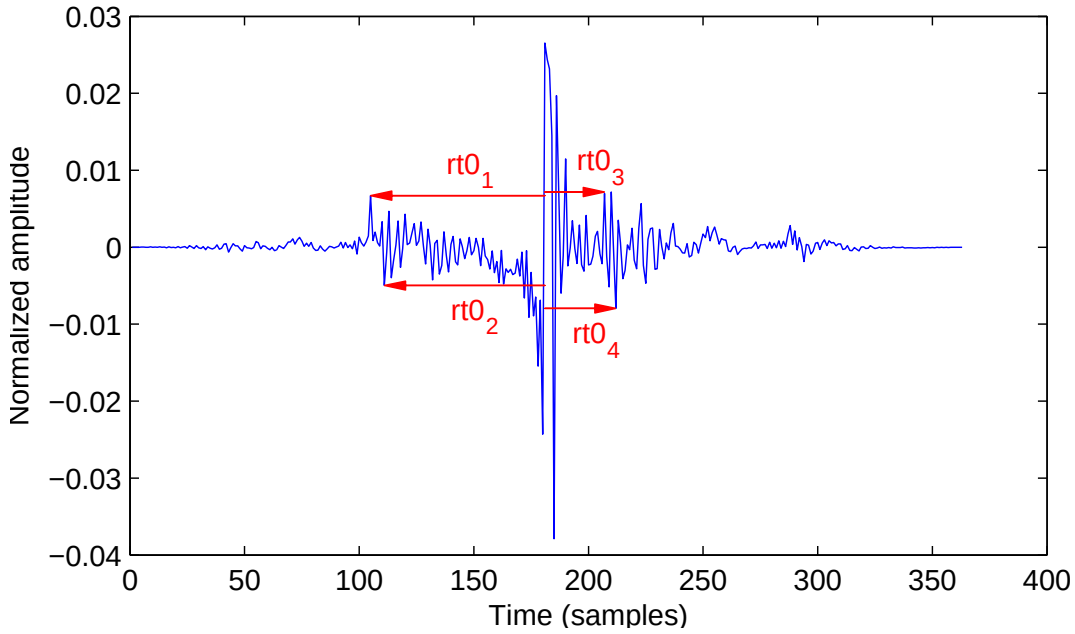

Fig. 3: Calculation of the $rt0$ parameter for a windowed codebook frame. The $rt0_i$ values give the distance in samples between the impulse ($T = 181$) and peaks. The values in the figure: $rt0_3 < rt0_4 < rt0_2 < rt0_1$.

Thesis I.2: [C1] I created a language independent method to transform speech produced with irregular phonation to regular speech using the model of Thesis I.1. In a perception test based on Hungarian samples I showed that the transformed speech is significantly less rough than original irregular speech.

The method is based on the analysis-synthesis method of Thesis I.1. The analysis is performed similarly as in Thesis I.1, with the difference that the codebook is built from modal residual parts, skipping the parts produced by irregular voice.

During the transformation, those parts of the residual are modified which are labeled as having irregular voice, while the regular voiced and unvoiced parts are left unmodified. The parameters obtained during synthesis are modified: $F0$ is interpolated, while $gain$ and MGC values are smoothed in the irregular parts. Irregular phonation may cause errors in $F0$ tracking: it may happen that the algorithm cannot calculate a pitch value, or half of the original $F0$ is measured. To correct this, we interpolate $F0$ with a straight line in sections having zero $F0$. We investigated all the stimuli used in the experiments and hand-corrected the interpolated $F0$ values if necessary. Therefore, the algorithm is semi-automatic. Irregular phonation causes small perturbations in the frame-by-frame $gain$ and MGC values as well, partly because of the abrupt changes in the amplitude of the pitch periods. Therefore, we used a 5-point moving average on these parameters which was found to be suitable for removing the perturbations. The further steps of the synthesis are the same as those of Thesis I.1.

An example for the result of the irregular-regular transformation is shown in Fig. 4. In the figure it can be seen that the ‘transformed to regular’ (c and d) and the ‘original regular’ (e and f) signals have similar regular pitch periods, while ’original irregular’ (a and b) is very different and has period-by-period amplitude attenuations.

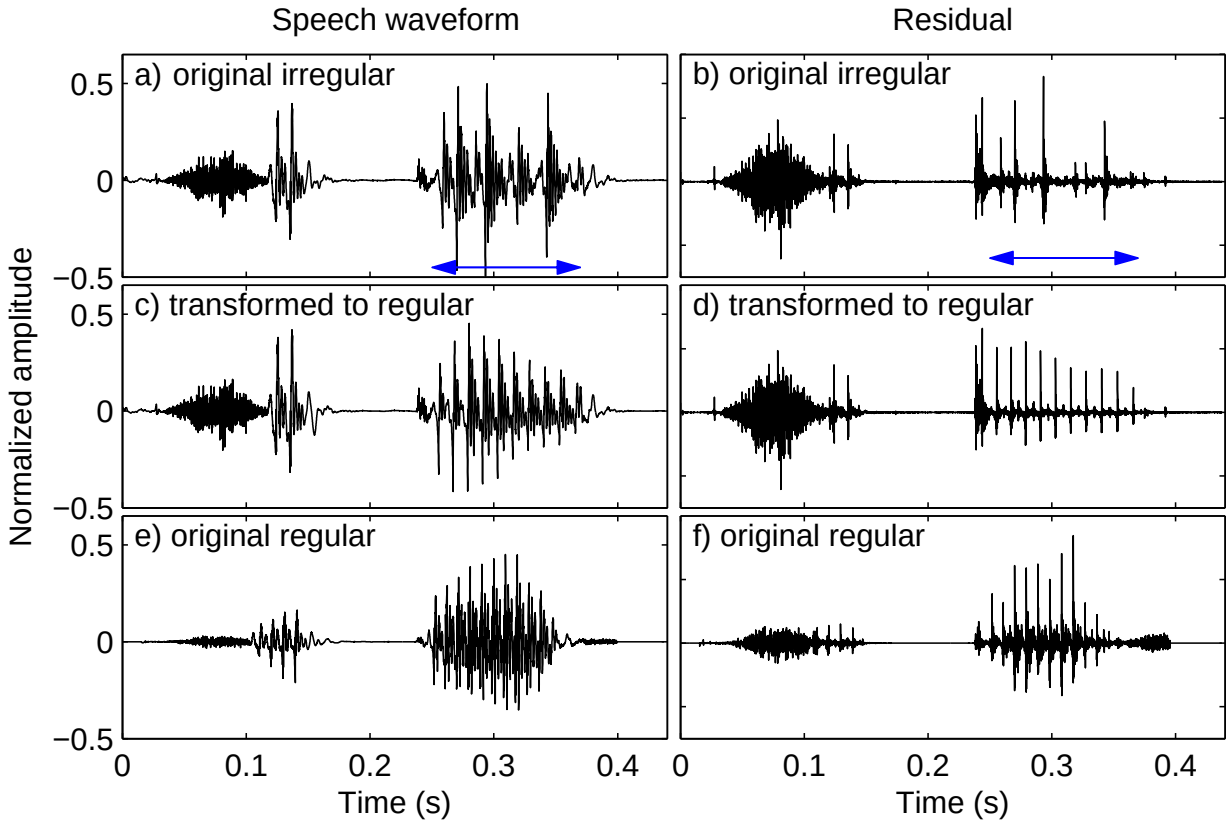

Fig. 4: Original and transformed speech waveforms and residuals of the word ’cipő’ from speaker FF3: a) speech signal b) residual signal with original irregular closing vowel (arrow denotes the irregular phonation).
c) speech signal d) residual signal with transformed closing vowel.
e) speech signal f) residual signal with original regular closing vowel.

The results of the irregular-regular transformation method were tested on the speech data of four Hungarian speakers (3 males: FF1, FF3 and FF4 and one female: NO3) from the PPBA database [27]. Four-four words having both regular and irregular versions were selected from the database for each speaker. The utterance versions having irregular sections were transformed to modal voice by the proposed method. The three versions of the words (original irregular, transformed to regular and original regular) were compared in a perception test, performed by 9 listeners. We compared the results with paired samples t-test and found that the transformation significantly ($p < 0.05$) decreased the perceived roughness compared to the original irregular samples.

Thesis I.3: [C1, C5] I validated with experimental methods on Hungarian samples that the transformation method of Thesis II.2 changes a number of relevant acoustic parameters of speech (open quotient, first formant bandwidth, spectral tilt) during irregular-regular transformation towards the values characteristic of regular phonation.

We also performed acoustic analysis on the samples selected for the perception test of Thesis I.2 (original irregular, transformed to regular, original regular). Three acoustic cues were chosen which were found previously to differentiate irregular and regular speech [30, 5]. According to this, in irregular voice the open quotient (OQ) is lower; the first formant bandwidth ($B1$) is higher; the spectral tilt (TL) is steeper than in regular speech.

The effect of the transformation on the OQ, $B1$, TL acoustic parameters were investigated with measurements. The measurements were performed in frequency domain [31] in the Wavesurfer program by visual inspection, with the correction of the parameters [32]: OQ was approximated by the first harmonic amplitude relative to the second harmonic amplitude (i.e. by $H1^* - H2^*$ in dBs); TL by the first harmonic amplitude relative to the third formant amplitude ($H1^* - A3^*$) and $B1$ by the first harmonic amplitude relative to the first formant amplitude ($H1^* - A1$).

The three acoustic parameters measured on the three speech sample types are shown in Fig. 5. ANOVA analysis and Tukey-HSD post-hoc test were performed. We found that the a $H1^* - H2^*$ was approximately the same for the original regular and for the transformed recordings ($p = 0.938$, n.s. difference), while it was significantly different for original irregular recordings ($p < 0.0005$). This means that in terms of open quotient, the transformed versions are close to the original modal versions. $H1^* - A1$ and $H1^* - A3^*$ are both significantly different for original irregular recordings from the others ($p < 0.0005$ and $p < 0.05$), but they are not significantly different for original regular and transformed utterances ($p = 0.336$ és $p = 0.321$, n.s. difference). The transformed utterances are close to the original regular utterances in $B1$ and TL. We can conclude from the acoustic experiments that the proposed transformation method reconstructs the above acoustical correlates of regular speech.

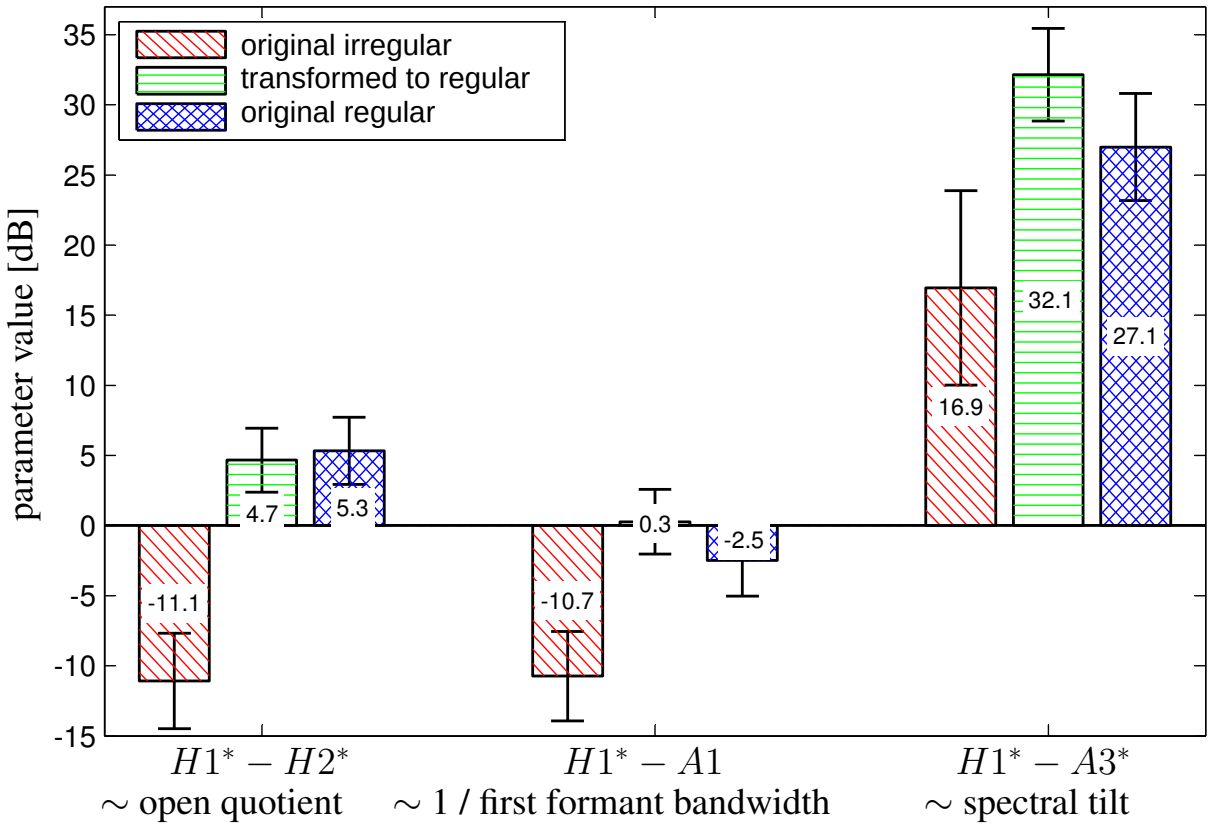

Fig. 5: Results of the acoustic analysis on the original irregular, irregular-regular transformed and original regular samples. Vertical black lines show the 95% confidence intervals.

Thesis group II: Integration of the new excitation model into a text-to-speech system and its application for irregular voice synthesis

The literature overview has shown numerous excitation models that are used in statistical parametric speech synthesis. Some of them use mixed excitation, other methods aim to model the glottal source signal, in some experiments the harmonics-noise model is further developed, while in quite a few cases speech residual based models are used.

In this thesis group I integrate the excitation model of Thesis I.1 into statistical parametric speech synthesis. The proposed system is compared to the freely accessible version of HTS using impulse-noise excitation. After that I extend the proposed system with two alternative irregular voice models.

Thesis II.1: [J2] I integrated the language independent excitation model of Thesis I.1 into hidden Markov-model based text-to-speech. In a perception test based on Hungarian samples I showed that the speech created by the new model has significantly better quality than the system with impulse-noise excitation.

The parameters used in the analysis part of Thesis I.1 ($F0$, $gain$, $rt0$, HNR and MGC) are calculated on each frame of every sentence of the training database with 50 ms frame size and 5 ms frame shift. The first and second derivatives of the parameters are also stored. After that the parameters were fitted to the training part of the HTS-HUN system [7]. Parameters with varying dimensions ($log(F0)$, $log(rt0)$ and $log(HNR)$) are modeled by multi-space distribution HMMs (MSD-HMM). The logarithm values were found to have better results during the experiments. The other parameters ($log(gain)$ and MGC) are modeled by traditional HMMs. For rhythm modeling, speech state duration densities are calculated for each phoneme. Phoneme-dependent state durations are modeled by Gaussian distributions. Context-dependent labeling and decision trees are applied to reduce the combination of all context dependent features. The parameter sequences are handled by separate decision trees.

The synthesis step is similar to that of Thesis I.1 with several extensions. The residual signal is synthesized from the $F0$, $gain$, $rt0$ and HNR parameters obtained as the result of machine learning and using the residual codebook. After that 6 kHz lowpass filter is applied and white noise is used in the frequency region above 6 kHz similar to HNM-based systems. This step ensures that the buzziness of voiced sounds is significantly decreased. Finally the speech signal is synthesized using the MGC parameters by a MGLSA filter. The new system is denoted HTS-CDBK.

We performed speech synthesis experiments with the speech data of speaker FF2 from the PPBA database. For this, the whole 137 minutes (1938 sentences) speech and the labels were used in speaker dependent training. The original 44.1 kHz signals were resampled to 16 kHz after 7.6 kHz lowpass filtering. As a baseline system the simple impulse-noise version of HTS was used (HTS-PN). Based on the residuals of speaker FF2, a codebook of 6 500 elements was created in the HTS-CDBK system. We synthesized 130-130 sentences which did not occur in the training database with both systems. 20-20 sentences were selected for a perception test, where the quality of the samples was judged by 15 listeners in paired comparison. According to the statistical analysis the HTS-CDBK was found to have significantly ($p < 0.0005$) better quality than the HTS-PN system.

Statistical parametric speech synthesis and most of the above excitation models (including the method of Thesis II.1) are optimized for ideal voices and may not produce high quality with voices having frequent non-modal sections like irregular phonation. When glottalization occurs (typically in the vowels of the last syllables of the phrases), usually the pitch tracker cannot measure proper $F0$ and sets the frame as being unvoiced. Therefore, this pattern is learned by the system and glottalization is modeled in HTS-CDBK similarly to unvoiced speech. This

produces a very unpleasant voice and is not a proper model of glottalization. In the following the HTS-CDBK system of Thesis II.1 will be used as baseline and it will be extended with irregular voice models.

Thesis II.2: [C2, J1] I created a language independent rule based irregular voice model and fitted this to the text-to-speech system of Thesis II.1. The model uses pitch halving, residual period amplitude scaling and spectral distortion. In a perception test based on Hungarian samples I showed that the synthesized speech extended with the new model is significantly more preferred and is more similar to the original speaker than the system of Thesis II.1.

The analysis and training steps are the same as in the system of Thesis II.1, only the synthesis step is different. The extended system is denoted as HTS-CDBK+Irreg-Rule. There is no explicit glottalization model for predicting the location of irregular voice, so sections with irregular phonation should be found from the generated $F0$ sequence. Glottalization was applied if at least five consecutive frames were given zero F0 within a vowel. In these cases, the $F0$ was interpolated between the modal voiced parts, or it was set to slightly decreasing if there were no voiced neighboring sounds.

The HTS-CDBK+Irreg-Rule model applies three heuristics for modeling irregular voice: 1) $F0$ halving, 2) pitch-synchronous residual modulation with periods multiplied by randomized amplitudes and 3) spectral distortion. During synthesis, in the modal voiced and unvoiced parts the residual generated by the HTS-CDBK system is used. In the sections that should be synthesized with irregular phonation, the half of the $F0$ of the generated and interpolated parameter sequence is used. Glottalization has often significantly lower $F0$, but in the residual codebook frames with extremely low $F0$ are rare. Therefore, during synthesis, residual frames are zero padded before the overlap-add step. The $F0$ halving and zero padding have a similar effect than removing every second cycle, which is perceptually similar to lowered open quotient [5]. During residual synthesis, each selected pitch cycle is multiplied by a random amplitude scaling factor in the range of $\{0 \dots 1\}$. This heuristic was motivated by the property of irregular voice that the consecutive pitch periods have often strong amplitude differences. In an earlier study we found that the extracted MGC parameters of irregularly phonated speech are less smooth than those of regular speech (Thesis I.2, [C1]). Therefore here we try to 'distort' the MGC parameters similarly by slightly modifying them: the parameter values are multiplied by random numbers between $\{0.995 \dots 1.005\}$, which is expected to have a similar effect than irregular voice. The synthesized speech is obtained from the residual similarly as earlier by MGLSA filtering, using the MGC parameters.

Fig. 6 shows an example for the results of the HTS-CDBK (a and b) and the HTS-CDBK+Irreg-Rule (c and d) systems. The irregular voice model was applied in the 'á' sound of the word 'Mihály'. The effect of zero padding the residual frames is that the waveform has separated pitch cycles. The gain scaling with random factors resulted in the strong amplitude attenuation of the fourth cycle. It is clearly visible on both the residual and the speech signals that the extended model is closer to the original irregular signal (Fig. 4. a and 4. b) than the baseline system.

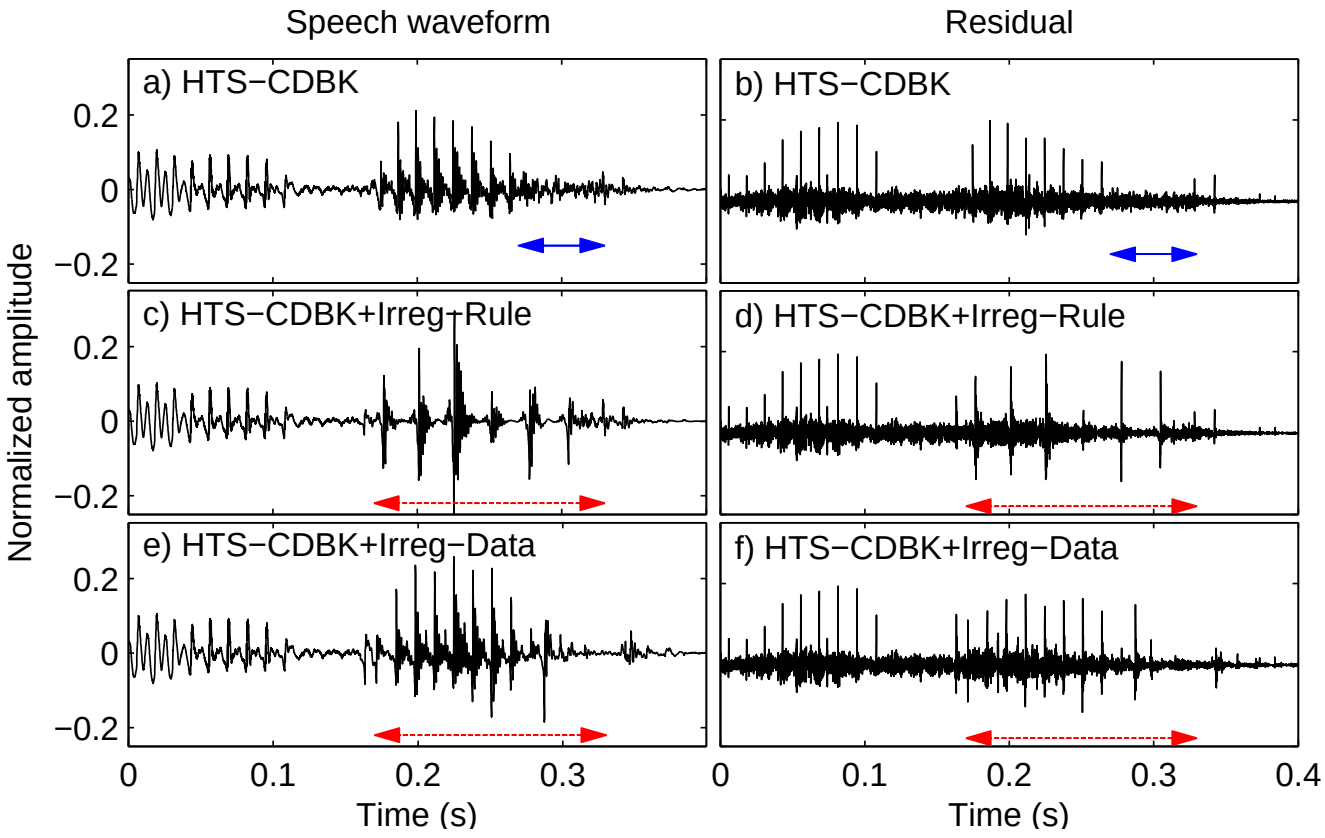

Fig. 6: Synthesized versions of the word 'Mihály' (extracted from a longer sentence) a) speech signal b) residual with the baseline system
c) speech signal d) residual with the model of Thesis II.2
e) speech signal f) residual with the model of Thesis II.3
Horizontal arrows denote the sections with irregular phonation.

In order to evaluate the quality that can be achieved by the rule-based irregular voice model, we conducted a perception test. To obtain the speech stimuli, we created four voice models with the HTS-CBK baseline and HTS-CDBK+Irreg-Rule systems and used two male speakers of the PPBA database (FF3 and FF4). 130-130 sentences were synthesized with all four voice models and 10-10 sentences having at least one irregularly synthesized vowel were selected. The last word (containing irregular phonation) of each sentence was cut and used in the

perception experiment. According to the evaluation of 11 listeners, the system of Thesis II.2 is significantly more preferred ($p < 0.0005$) and significantly more similar ($p < 0.0005$) to the original speaker than the baseline system.

Thesis II.3: [J1] I created a language independent data-driven irregular voice model and fitted this to the text-to-speech system of Thesis II.1. The model uses a corpus built from residuals of irregular voice sections and applies unit selection during synthesis. In a perception test based on Hungarian samples I showed that the synthesized speech extended with the new model is significantly more preferred and is more similar to the original speaker than the system of Thesis II.1.

Another model of irregular phonation was created for synthesis purposes which is data-driven and based on residual unit selection. This extended system is denoted as HTS-CDBK+Irreg-Data. The analysis and the training steps are the same as in the system of Thesis II.1; the difference is in the synthesis step.

After the analysis step a corpus from residuals of irregularly phonated vowels of the speech database is built („GLOTT" corpus). For this, we apply a recent high-precision creaky voice detection algorithm („creak_detect", [19]). We include the residuals of the vowels in the GLOTT corpus which have the creaky binary decision in more than half of the frames of the vowel. In the data-driven method the full, vowel-length residuals are stored in the corpus contrary to the pitch-synchronous residual frames.

During synthesis, the modal residual sections are synthesized with the HTS-CDBK method. Similarly to the HTS-CDBK system, there is no explicit glottalization model for predicting the location of irregular voice, so sections with irregular phonation should be found from the generated $F0$ sequence. The residual for the sections that should be synthesized with irregular phonation is searched from the GLOTT corpus. In this initial version of the method we hypothesize that only one vowel should have irregular phonation and we do not deal with concatenation among vowel-length residuals. For selecting a target residual from the corpus only target cost is used which is composed of several sub-costs: 1) mean $F0$ difference 2) mean length difference 3) context of the residuals. During unit selection we constrain that the target residual should be at least as long as the section to produce irregularly. After the target residual is found by minimizing the target cost, the residual section is resized to the target length by removing the last samples. Its gain is normalized to fit to the overall intensity curve of the residual signal, but other properties are not modified. MGC distortion is similarly applied here as in the HTS-CDBK+Irreg-Rule system. Finally, synthesized speech is reconstructed by MGLSA filtering using the MGC parameters.

Fig. 6 shows an example for the residual and speech waveform of the data-driven irregular voice model (e and f). Similarly to the baseline HTS-CDBK system (a and b), the last vowel of the HTS-CDBK+Irreg-Data residual contains irregular-like voice (amplitude attenuations) only in the last part of the vowel. When comparing this to the original irregular sample (Fig. 4. a and 4. b), we can see that the synthesized residual contains several secondary pulses similarly to the original residual of the irregular sample.

Another listening test was conducted on the samples of speakers FF3 and FF4 for measuring the acceptability of the HTS-CDBK+Irreg-Data system compared to the baseline HTC-CDBK system. For speaker FF3, a glottalization corpus consisting of 1116 vowel-length residuals, for speaker FF4, a corpus consisting of 1822 vowel-length residuals was built from the speech residuals of the whole speech database. 130-130 sentences were synthesized with the HTS-CDBK and HTS-CDBK+Irreg-Data systems and 10-10 sentences having at least one irregularly synthesized vowel were selected. The last word (containing irregular phonation) of each sentence was cut and used in the perception experiment. According to the evaluation of 16 listeners, the system of Thesis II.3 is significantly more preferred ($p < 0.0005$) and significantly more similar ($p < 0.0005$) to the original speaker than the baseline system.

Thesis II.4: [J1] I validated with experimental methods on Hungarian samples that the methods of Theses II.2 and II.3 model a number of relevant acoustic parameters of speech (open quotient II.2 and II.3, first formant bandwidth: II.2) during synthesis similarly to the characteristic of irregular phonation.

We also performed acoustic analysis on the samples selected for the perception tests of Theses II.2 and II.3. Three acoustic cues were chosen which were found previously to differentiate irregular and regular speech [30, 5]. According to this, in irregular voice the open quotient (OQ) is lower; the first formant bandwidth ($B1$) is higher; the spectral tilt (TL) is steeper than in regular speech.

The measurements were conducted on the speech data of the two speakers (10-10 words synthesized by the baseline system, by the rule-based and data-driven models; 10-10 original regular and original irregular recordings). Instead of OQ we measured $H1^* - H2^*$, $1/B1$ was investigated by the analysis of $H1^* - A1$, whereas the TL acoustic cue was measured based on $H1^* - A3^*$. The comparison of the three parameters and the five speech sample types are shown in Fig. 7. According to the ANOVA analysis and Tukey-HSD post-hoc test the difference of the first two harmonics is significantly different between the original regular, synthesized baseline and other samples ($p < 0.05$), whereas it does not differ significantly between the original and synthesized irregular samples ($p = 0.99$,

n.s.). This means that in terms of open quotient, the synthesized versions are close to the original irregular versions. According to Fig. 7, in terms of $H1^* - A1$ and thus the first formant bandwidth the rule based irregular voice model is close to original irregular speech, while the result of the data-driven model is between the regular and irregular samples. In this experiment, $H1^* - A3^*$ was not helpful to differentiate between the regular and irregular utterances. From the acoustic experiment the conclusion is that the proposed irregular models can reconstruct two of the three investigated acoustical correlates of irregular speech.

In summary we can conclude according to the perception tests and acoustic analysis that both irregular voice models are suitable for synthesis of glottalized speech, and the quality of the synthesized samples with the two systems is similar.

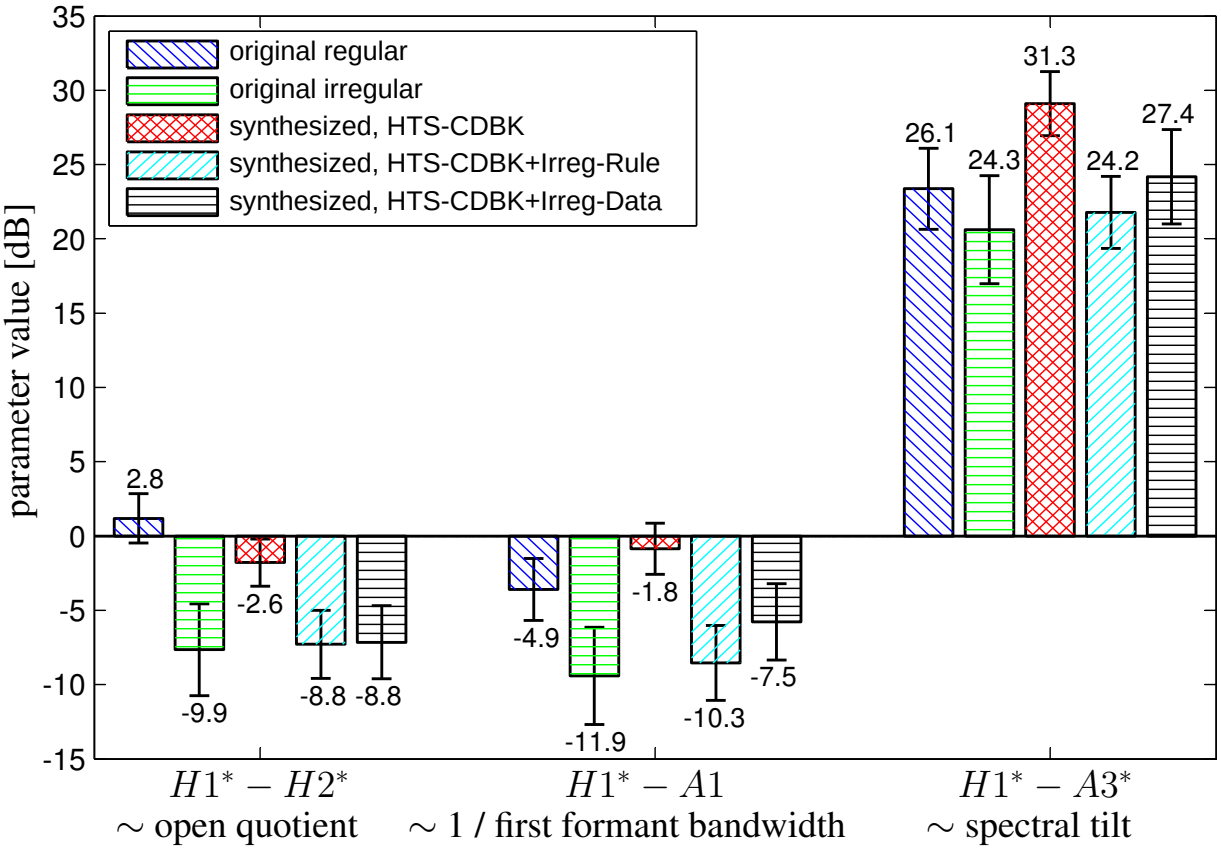

Fig. 7: Results of the acoustic analysis on the synthesized samples with the baseline system and irregular voice models. Vertical black lines show the 95% confidence intervals.

Thesis group III: Role of subglottal resonances in Hungarian speech

The source-filter model of speech production [1], which was used in the excitation model of Thesis I.1 applies the simplification that the source and the filter are perfectly separable. In reality nonlinear interaction might occur between the source (glottis) and filter (vocal tract), which is caused in part by the subglottal system. There is no direct cause and effect correlation between the formants (the resonance frequencies of the vocal tract) and the subglottal resonances (the resonance frequencies of the lower airways). However, the indirect relation between them contribute to separated vowel groups which can be explained by Quantal Theory. In principle, Quantal Theory [25] forms a system for the classification of speech sounds universally and independent of the language used. However, in practice it is not known whether subglottal resonances contribute to the separation of speech sounds in all of the languages. The relation between subglottal resonances and vowel formants has been investigated in detail from speech production aspects for American English [24], Spanish, German and Korean; but there have been no results for Hungarian till now.

In this thesis group I present my investigations regarding the role of subglottal resonances on Hungarian speech and introduce a new, SGR-based vowel classification method.

Thesis III.1: [C4, J4] I created a model for the effect of the resonances of the lower airway (subglottal) system on Hungarian speech. I showed that the subglottal resonances (the first three resonance frequencies of the lower airway system) can be used for formant-based separation of Hungarian vowels groups using the indirect relationship between the subglottal resonances and formants.

In the first experiment we investigated the role of the subglottal system in Hungarian vowels. For this, new recordings were made and we analyzed them speaker by speaker and by pooling them together via normalization.

During the research we analyzed the speech and accelerometer signal of four Hungarian speakers in logatom reading. The first three formants ($F1$, $F2$ and $F3$) were automatically measured from the speech signal with Praat, and manually corrected. Subglottal resonances were measured manually with Wavesurfer from the accelerometer signal as the peaks of the spectral envelope, in 25 points for each speaker and SGR. The median values of the SGRs were used, which are in the expected range. According to the measurements we created a new acoustic model for the effect of subglottal resonances on Hungarian speech. We applied the model developed for American English [24] for Hungarian and we found that

1) the first subglottal resonance ($Sg1$) is between the low and non-low vowels in the range of the first formant ($F1$),
2) the second subglottal resonance ($Sg2$) is between the front and back vowels in the range of the second formant ($F2$),
3) the third subglottal resonance ($Sg3$) divides the front, unrounded, non-low vowels from other front vowels in the range of the second formant ($F2$).

We verified the above model by engineering methods. The vowel formants were normalized: the values of the formants were divided by the corresponding subglottal resonance of the speaker ($F1/Sg1$, $F2/Sg2$ and $F2/Sg3$), and the data were pooled together for all speakers. Fig. 8 shows the SGR-normalized formant histograms: e.g. in figure b) it can be seen that the $Sg2$ (vertical line) separates the front and back vowels nearly optimally.

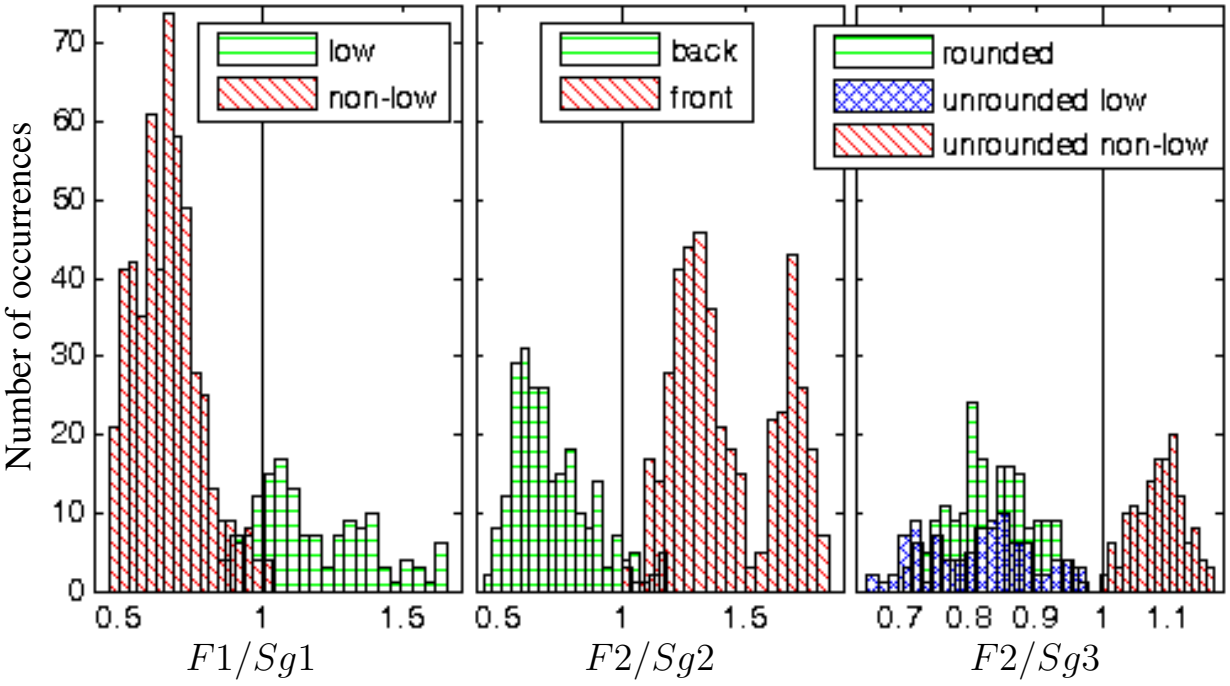

Fig. 8: Normalized formant histograms based on logatom speech: the $F1/Sg1$, $F2/Sg2$, $F2/Sg3$ valued pooled together for all of the speakers. The vertical lines denote $Sg1$, $Sg2$ and $Sg3$, respectively.

According to the detailed investigations the above claims do not hold for each speaker and every category. To investigate the optimal separation between the categories, we conducted ROC (*Receiver Operating Characteristics*) analysis separately for every SGR and speaker. Fig. 9 shows that from 6 cases out of 12 the median SGR is in the optimal separating range (** in the figure), in 4 more cases it is within one standard deviation (*), whereas in the remaining 2 cases it is farther.

As a summary, the subglottal resonances divide nearly optimally the low vs. non-low, front vs. back, front unrounded non-low vs. other front vowels in the Hungarian language.

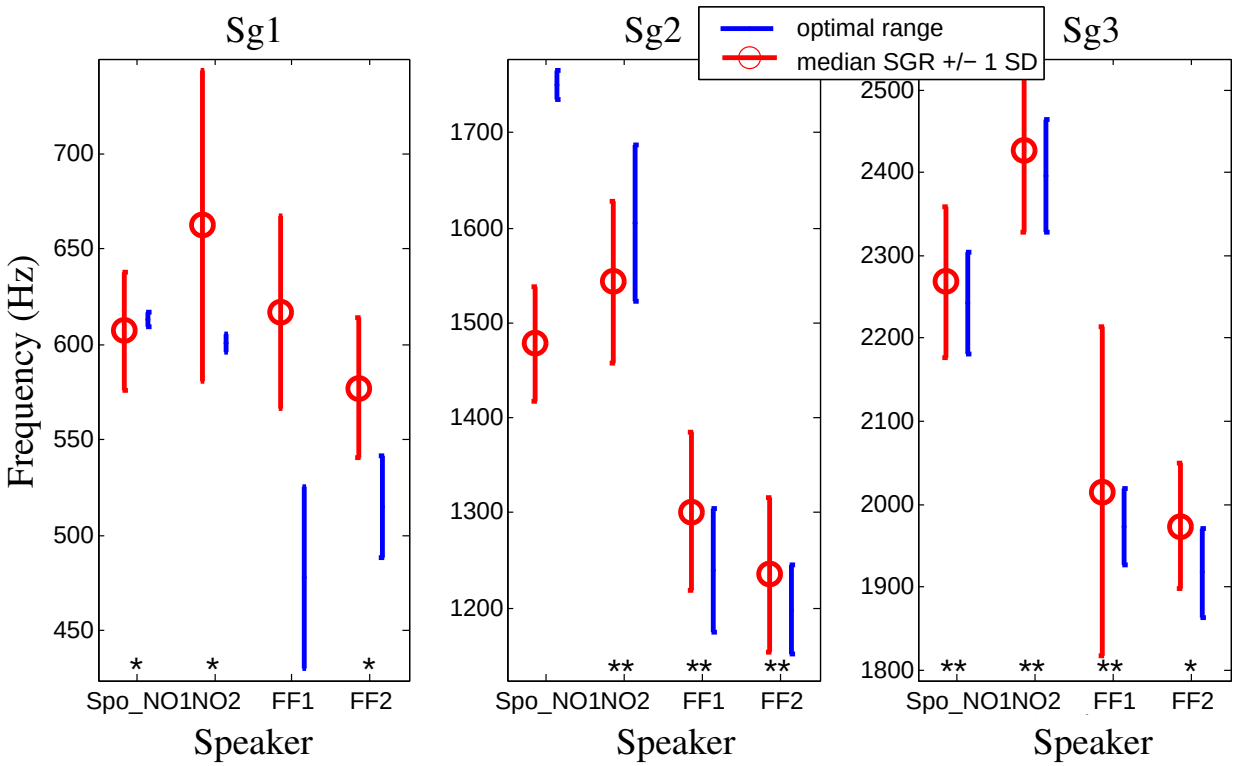

Fig. 9: Results of the ROC analysis for the investigation of SGR-separation in vowel classes. Light lines denote SGR values and one standard deviation; dark lines denote the optimal separating range.

The formal, quantitative model that was introduced in Thesis III.1 is suitable for computer implementation. In order to test whether the knowledge of subglottal resonances can help speech processing, we planned a new experiment. In this, we investigate whether it is possible to improve vowel classification using the SGRs.

Thesis III.2.: [J4] I created automatic classifiers based on the indirect relationship between the subglottal resonances and formants. The method uses subglottal resonance based normalization for classifying vowels into the categories of Thesis III.1. I showed that on the examined Hungarian samples the Sg2 based method has always higher accuracy, the Sg3 based method has higher accuracy when using small training data, while the Sg1 based method has lower accuracy than a decision tree based reference classifier using raw formants.

As training and test data the spontaneous speech recordings of six Hungarian speakers (5 males and 1 female) [28] were used, of which 5948 vowels were extracted. The $Sg1$, $Sg2$ and $Sg3$ values were measured separately from read speech recordings in Wavesurfer in 20 points for each speaker and SGR, and the median values were used later.

During the experiment we used J4.8 decision trees in Weka as baseline classifiers. We chose this decision tree because it is the improved version of the widely used C4.5 classifier and in most cases it results in nearly optimal classification. We created three classifiers according to the model of Thesis III.1. As input they have the raw formant values, whereas the outputs are the categories of the model:

a) input: $F1$, output: low – non-low
b) input: $F2$, output: front – back
c) input: $F2$, output: front, unrounded, non-low – other front

We created three new SGR-based classification methods using formant normalization according to the three subglottal resonances and the three categories of Thesis III.1. The input of the classifiers is the SGR-normalized value of the $F1$ or $F2$ formant, i.e. the formant frequency divided by the corresponding subglottal resonance value ($Sg1$, $Sg2$ or $Sg3$). The output of the classifier is the vowel categories used in the model:

a) input: $Fn1 = F1/Sg1$, output: low – non-low
b) input: $Fn2 = F2/Sg2$, output: front – back
c) input: $Fn3 = F2/Sg3$, output: front, unrounded, non-low – other front

The classifier performs a simple threshold decision, e.g. in the case b) : if the $Fn2 >= 1.0$ holds for the input vowel, the method decides for the 'front' output, whereas in case of $Fn2 < 1.0$ the classifier has the 'back' output.

The baseline decision tree-based classifiers (using raw formants as input without the knowledge of SGRs) were compared to the SGR-normalization based classifiers. In the experiment we investigated which classifier performs better as a function of the amount of training data used. In case of the decision tree the amount of training data was changed between 0.2...90 % of the whole dataset (12 – 5353 data points), and the remaining part was used for testing. The accuracy of the SGR-based classifier does not depend on the amount of training data, if the subglottal resonance values have been measured. In case of the SGR-based classifier the tests were made on 50 % of the whole dataset. All measurements were repeated on 100 random groups and the results were averaged.

Fig. 10 shows the results of the experiment. In case a) the knowledge of $Sg1$ did not help the classification. The reason for this can be that the measurement of $Sg1$ is often difficult from the accelerometer signal, because the strong lower

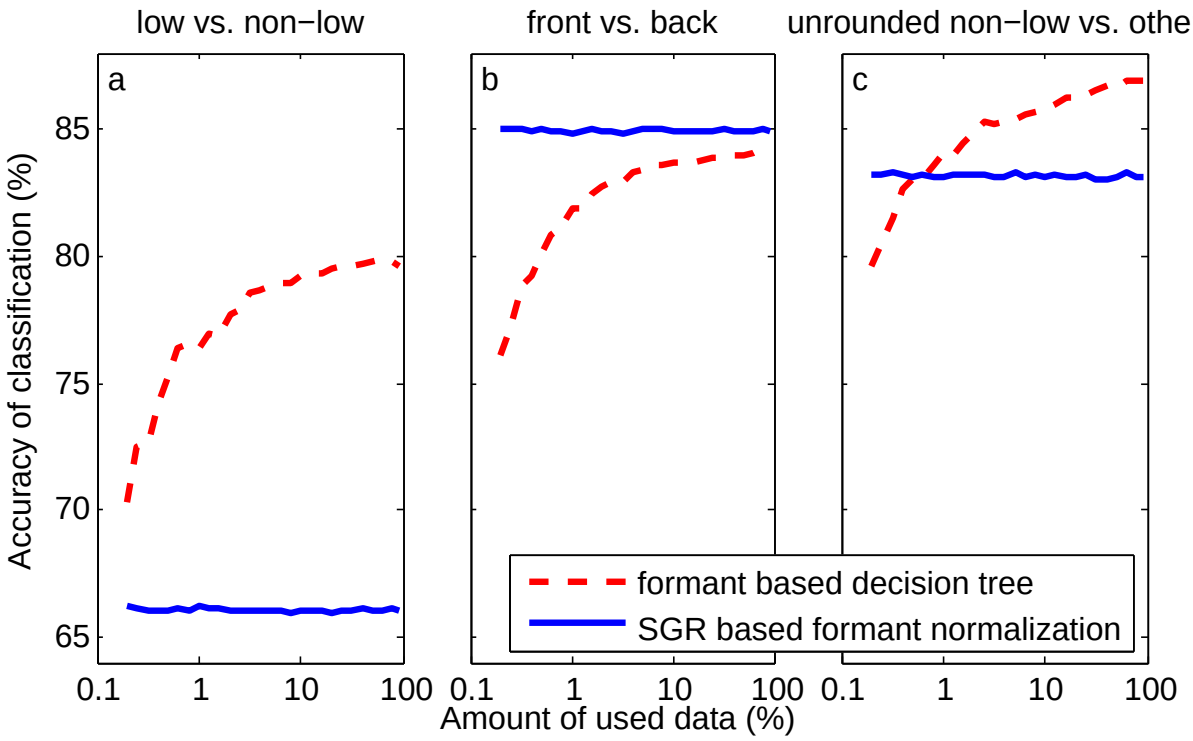

Fig. 10: Comparison of the accuracies of the formant based decision trees and the SGR-normalized formant-based classifications as a function of the amount of used data: a) $Sg1$, b) $Sg2$, c) $Sg3$.

harmonics can distort the measurements. In figure b) the result of the separation of front and back vowels is shown. In this case the knowledge of $Sg2$ clearly improved the classification. With little data the SGR-based classification has 20 % higher accuracy than the raw formant based classifier, while using most of the data the improvement is still 1 %. This meets the results of the literature as usually the $Sg2$ has the strongest effect in dividing the vowels into specific categories [24]. In case c) the $Sg3$-based classification has higher accuracy, if less than 1 % (50 vowels) are used in the training.

According to this experiment the $Sg2$-based classification is always, while the $Sg3$-based classification is in limited data conditions (less than 50 vowels) better than the baseline decision tree. For the SGR-normalization based classification it is enough to have 10–20 vowels as training data, which are needed for the measurement of the subglottal resonances. Thus the SGR-based method adapts quickly to the speaker, and it is theoretically established based on the model of Thesis III.1. However, the raw formant based classifier is sensitive for the type and amount of the training data.

In the above research we investigated the relation of vowel formants and subglottal resonances in speech production and automatic classification, for the Hungarian language. According to the analyses and the experiments the SGRs

help the separation of vowels to phonological distinctive categories in Hungarian as well. This contributes to the extension of Quantal Theory [25] with subglottal resonances [24].

6. Applicability of the results

The results of my research can be used in numerous speech technology applications. On the one hand they contribute to more natural human-computer interaction. On the other hand they help to understand human speech production. I validated my methods on Hungarian samples. The methods in Thesis groups I and II are language independent, therefore they can be easily used for other languages as well. In the following I show several potential applications.

The residual-based analysis-synthesis excitation model of Thesis I.1 is suitable for automatic conversion of voice qualities. According to my preliminary experiments the method might be used for breathy-to-modal speech transformation. The glottalization correction method of Thesis I.2 could be extended for longer speech regions as well, for making creaky, pathologic voices more pleasant (e.g. the voice of speakers or announcers). The automatic version of the irregular-regular transformation method could be used for improving speech databases by eliminating the unwanted irregular sections, which would make the speech more ideal for further processing purposes.

The speech synthesis system introduced in Thesis II.1 could improve the quality of synthesized speech in low resource devices (e.g. smartphones). Because of the limited resources the more complicated excitation models are usually difficult to use. On the other hand the model of the thesis might be suitable for real-time speech synthesis on specific devices. The irregular voice models of Theses II.2 and II.3 can contribute to making speech synthesis more natural, expressive and personalized. By more natural and personalized systems I mean that the synthesized voice can have similar ratio of irregular sections as it is characteristic of the original speaker. It was found earlier that the speakers use different voice qualities for certain emotions (e.g. sad and irritated); thus the irregular voice models can improve expressive speech synthesis as well.

The investigation of subglottal resonances (Thesis III.1) contributes to understanding the function of phonological distinctive features in Quantal Theory. According to the assumptions the vowel formants are partly normalized to the SGRs during human speech perception. There are large differences in the acoustic product of single speakers and SGRs might help understanding the speech of each other. This property can be used in speech technology as well: subglottal resonances have been successfully used in automatic speaker normalization, improving the accuracy of child speech recognition [26]. In a preliminary perception

test we found that the relation of vowels and subglottal resonances can be connected to the quality of the perceived vowel [J4]. According to the results, if the ratio of $F2$ and $Sg2$ was not in accordance with Thesis III.1 (it did not hold for front vowels that $F2 > Sg2$ and back vowels that $F2 < Sg2$), then it was more difficult for the listeners to recognize the original vowel. We expect that the wrong $F2 - Sg2$ ratio is unfavorable from perception aspects and makes the understanding of speech more difficult. It might be possible to create a method based on the above expectation to clear the unsuitable speech sounds from training databases, thus contributing to make synthesized speech more intelligible. The classifier introduced in Thesis III.2 can be extended for the classification of longer sound sequences (e.g. CV or VC transition) to articulatory groups, which has been investigated previously for American English. If the source-filter model used in hidden Markov-model based speech synthesis could be extended with the modeling of subglottal resonances, that might further improve the naturalness of synthesized speech.

7. Acknowledgement

Hereby I thank my advisor, Dr. Géza Németh for his supervising, for the continuous help and support during my work, and for the useful advices and remarks. I thank him that with his work he laid the foundations of my scientific way of thinking.

I have to thank the current and previous colleagues of the Speech Technology Laboratory. Mátyás Bartalis helped with friendly talks, Dr. Tamás Bőhm with research and methodology, Dr. Márk Fék with his speech coding knowledge, Géza Kiss with his programming skills, Dr. Gábor Olaszy with his extensive experience, Bálint Tóth with introducing to me statistical parametric speech synthesis, Dr. Csaba Zainkó with his signal processing knowledge - they all helped me during my work and contributed to the writing of my dissertation. Besides, I thank the help of Tibor Fegyó, Gábor Kiss, Dr. Péter Mihajlik, Péter Nagy, Dr. György Szaszák, Dávid Sztahó, Balázs Tarján and Dr. Klára Vicsi.

I thank Dr. Steven M. Lulich (Indiana University, Bloomington, USA) that he introduced to me his research on subglottal resonances and supported my experiments in this topic. I have to thank Dr. Tekla Etelka Gráczi (Research Institute for Linguistics, Hungarian Academy of Sciences), Zsuzsanna Bárkányi (Research Institute for Linguistics, Hungarian Academy of Sciences), and András Beke (Research Institute for Linguistics, Hungarian Academy of Sciences) for the research co-operation and for widening my horizon. I thank all of my co-authors for the possibility of writing papers together and for the joys of working as a group.

Furthermore I thank Dr. Tamás Henk and Dr. Gábor Magyar department heads for they help in the process of the dissertation.

I thank the speakers of the PPBA, BEA databases and Thesis group III that I could use their voice for my experiments. I thank the listeners of the perception experiments that they listened and evaluated the samples and helped the research directions with their comments.

I thank Dr. Mária Gósy and Dr. Péter Olaszi that they helped to improve the dissertation with their valuable remarks and useful suggestions.

I specially thank my family: my wife Berni, my daughter Lili, my son Ábel, my mother Édi, my father István and my brother Krisztián that they continuously supported me during the years of the graduate school and created a calm atmosphere that is needed for research.

The research was supported by the following projects: NAP (OMFB-00736/2005), Enhances (NKFP 2/034/2004), Teleauto (OM-00102/2007), BelAmi (ALAP2-00004/2005), ETOCOM (TÁMOP-4.2.2-08/1/KMR-2008-0007), Research university (TÁMOP-4.2.1/B-09/1/KMR-2010-0002), CESAR (Grant No. 271022), Paelife (Grant No. AAL-08-1-2011-0001) and EITKIC (EITKIC_12-1-2012-001).

8. Abbreviations

ANOVA	ANalysis Of VAriance
BEA	BEszélt nyelvi Adatbázis (Spoken Language Database)
CELP	Code-Excited Linear Prediction
CMOS	Comparative Mean Opinion Score
DSM	Deterministic plus Stochastic Model
GCI	Glottal Closure Instant
HNM	Harmonic plus Noise Model
HMM	Hidden Markov-model
HNR	Harmonics-To-Noise Ratio
HTS	HMM-based Speech Synthesis System (H-Triple-S)
IPA	International Phonetic Alphabet
LF	Liljencrants-Fant
MGC	Mel-Generalized Ceptstrum
MGLSA	Mel-Generalized Log Spectral Approximation
MOS	Mean Opinion Score
MSD	Multi-Space Distribution
OQ	Open Quotient
PCA	Principal Component Analysis
PN	Pulse-Noise
PPBA	Preciziós, Párhuzamos magyar Beszédadatbázis (Hungarian precisely labeled and segmented, parallel speech database)
PSOLA	Pitch Synchronous Overlap and Add
QT	Quantal Theory
RMS	Root Mean Square
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristics
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SGR	Subglottal Resonance
TL	Spectral Tilt
TTS	Text-To-Speech

9. Notations

A1	Amplitude of first formant
A3, A3*	Amplitude of third formant (*: corrected value)
B1	Bandwidth of first formant
F0	Fundamental frequency
F1	Frequency of the first formant
F2	Frequency of the second formant
FF1, FF2, FF3, FF4	Four male speakers of PPBA database
Fn1	Sg1-normalized first formant
Fn2	Sg2-normalized second formant
Fn3	Sg3-normalized second formant
H1, H1*	First harmonic (*: corrected value)
H2, H2*	Second harmonic (*: corrected value)
HTS-CDBK	HTS with residual codebook excitation
HTS-CDBK+Irreg-Rule	Rule-based irregular voice model in HTS
HTS-CDBK+Irreg-Data	Data-driven irregular voice model in HTS
HTS-HUN	Hungarian version of the HTS system
HTS-PN	HTS with pulse-noise excitation
gain	Energy of residual frame
Log_FF1, Log_FF2	Two male speakers of logatom recordings
Log_NO1, Log_NO2	Two female speakers of logatom recordings
NO3	A female speaker of PPBA database
rt0	Parameter describing residual frame peaks
Sg1	First subglottal resonance
Sg2	Second subglottal resonance
Sg3	Third subglottal resonance
Spo_FF1 ... Spo_FF5	Five male speakers of spontaneous speech recordings
Spo_NO1	A female speaker of spontaneous speech recordings

10. References

[1] G. Fant, *Acoustic theory of speech production*. The Hague: Mouton, 1960.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. Black, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 294–299, 2007.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, Nov. 2009.

[4] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, vol. 1, (Atlanta, Georgia, USA), pp. 373–376, 1996.

[5] T. Bőhm, N. Audibert, S. Shattuck-Hufnagel, G. Németh, and V. Aubergé, "Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles," in *Acoustics'08*, (Paris, France), pp. 6141–6146, 2008.

[6] K. N. Stevens, *Acoustic Phonetics*. Cambridge: Cambridge University Press, 1998.

[7] B. Tóth and G. Németh, "Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis," *Acta Cybernetica*, vol. 19, no. 4, pp. 715–731, 2010.

[8] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[9] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1829–1832, 2008.

[10] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1881–1884, 2008.

[11] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7830–7834, 2013.

[12] D. Erro and I. n. Sainz, "HNM-based MFCC+ F0 extractor applied to statistical speech synthesis," in *Proc. ICASSP*, (Prague, Czech Republic), pp. 4728–4731, 2011.

[13] Z. Wen and J. Tao, "Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis," in *Proc. Interspeech*, (Florence, Italy), pp. 1805–1808, 2011.

[14] J. S. Sung, D. H. Hong, H. W. Koo, and N. S. Kim, "Statistical Approaches to Excitation Modeling in HMM-Based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 2, pp. 379–382, 2013.

[15] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis," in *Proc. ICASSP*, (Taipei, Taiwan), pp. 3793 – 3796, 2009.

[16] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1779–1782, 2009.

[17] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, pp. 2649–2658, May 1998.

[18] T. Bőhm, Z. Both, and G. Németh, "Automatic Classification of Regular vs. Irregular Phonation Types," in *NOLISP*, (Vic, Spain), pp. 43–50, 2009.

[19] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, pp. 1028–1047, June 2013.

[20] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Parameterization of vocal fry in HMM-based speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1775–1778, 2009.

[21] T. Drugman, J. Kane, and C. Gobl, "Modeling the Creaky Excitation for Parametric Speech Synthesis," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 1424–1427, 2012.

[22] T. Drugman, J. Kane, T. Raitio, and C. Gobl, "Prediction of Creaky Voice from Contextual Factors," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7967–7971, 2013.
[23] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Proc. Interspeech*, pp. 2316–2320, 2013.
[24] S. M. Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.
[25] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.
[26] S. Wang, S. M. Lulich, and A. Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *The Journal of the Acoustical Society of America*, vol. 126, pp. 3268–3277, Dec. 2009.
[27] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian)," *Beszédkutatás 2013 [Speech Research 2013]*, pp. 261–270, 2013.
[28] M. Gósy, "Magyar spontánbeszéd-adatbázis - BEA [Hungarian spontaneous speech database] (in Hungarian)," *Beszédkutatás 2008 [Speech Research 2008]*, pp. 194–207, 2008.
[29] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, Apr. 1993.
[30] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers.," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, Feb. 1990.
[31] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, pp. 1212–1223, Dec. 1995.
[32] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. ICASSP*, (Montreal, Quebec, Canada), pp. 669–672, 2004.

11. Publications

Publications related to Ph.D. Thesis

Journal papers

[J1] Tamás Gábor Csapó, Géza Németh, „Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation," *IEEE Journal on Selected Topics in Signal Processing*, accepted, 2013.
(BME-PA points: 100% · 6p = 6p) Scopus / Web of Science, IF: 3.297.

[J2] Tamás Gábor Csapó, Géza Németh, „Statistical parametric speech synthesis with a novel codebook-based excitation model," *Intelligent Decision Technologies*, accepted, 2013.
(BME-PA points: 100% · 6p = 6p) Scopus.

[J3] Tamás Gábor Csapó, „Increasing the naturalness of synthesized speech (PhD summary)," *The Phonetician*, No. 104–105, pp. 88–97, 2012.
(BME-PA points: 100% · 0p = 0p) (summary paper)

[J4] Tamás Gábor Csapó, Tekla Etelka Gráczi, Zsuzsanna Bárkányi, András Beke, Steven M. Lulich, „Patterns of Hungarian vowel production and perception with regard to subglottal resonances," *The Phonetician*, No. 99–100, pp. 7–28, 2011.
(BME-PA points: 50% · 6p = 3p)

Conference papers

[C1] Tamás Gábor Csapó, Géza Németh, „Transformation of irregular voice to modal voice by residual analysis and synthesis," *IEEE Signal Processing Letters*, in preparation, 2013.
(BME-PA points: 100% · 0p = 0p)

[C2] Tamás Gábor Csapó, Géza Németh, „A novel irregular voice model for HMM-based speech synthesis," *ISCA 8th Speech Synthesis Workshop (SSW8)*, (Barcelona, Spain), pp. 229–234, 2013.
(BME-PA points: 100% · 3p = 3p)

[C3] Tamás Gábor Csapó, Géza Németh, „A novel codebook-based excitation model for use in speech synthesis," *IEEE CogInfoCom 2012*, (Kosice, Slovakia), pp. 661–665, 2012.
(BME-PA points: 100% · 3p = 3p)

[C4] Tamás Gábor Csapó, Zsuzsanna Bárkányi, Tekla Etelka Gráczi, Tamás Bőhm, Steven M. Lulich, „Relation of formants and subglottal resonances in Hungarian vowels," *Proc. Interspeech 2009*, (Brighton, United Kingdom), pp. 484–487, 2009.
(BME-PA points: 50% · 3p = 1.5p)

Conference presentation abstracts

[C5] Csapó Tamás Gábor, Németh Géza, „Irreguláris beszéd regulárissá alakítása beszédkódoláson alapuló módszerrel," *Beszédkutatás*, (Budapest, Hungary), Nov 14–15, 2013. (in Hungarian)
(BME-PA points: 100% · 0p = 0p)

[C6] Csapó Tamás Gábor, Bárkányi Zsuzsanna, Gráczi Tekla Etelka, Beke András, Bőhm Tamás, „A magánhangzó-formánsok és a szubglottális rezonanciák összefüggése a spontán beszédben," *Beszédkutatás*, (Budapest, Hungary), Oct 16–17, 2009. (in Hungarian)
(BME-PA points: 20% · 0p = 0p)

Additional publications

Journal papers

[J5] Tamás Gábor Csapó, Csaba Zainkó, Géza Németh, „A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System," *Infocommunications Journal*, Vol. LXV., No. I., pp. 32–37, 2010.
(BME-PA points: 50% · 4p = 2p)

[J6] Csapó Tamás Gábor, ,,Változatos prozódia megvalósítása szövegfelolvasó rendszerekben,” *Akusztikai Szemle*, Vol. IX., No. 3., pp. 16–18, 2009. (in Hungarian)
(BME-PA points: 100% · 2p = 2p)

[J7] Csapó Tamás Gábor, Németh Géza, Fék Márk, ,,Szövegfelolvasó természetességének növelése,” *Híradástechnika*, Vol. LXIII., No. 5., pp. 21–30, 2008. (in Hungarian)
(BME-PA points: 50% · 1p = 1p)

Conference papers

[C7] Éva Székely, Tamás Gábor Csapó, Bálint Tóth, Péter Mihajlik, Julie Carson-Berndsen ,,Synthesizing Expressive Speech from Amateur Audiobook Recordings,” *SLT 2012*, (Miami, Florida, USA), pp. 297–302, 2012.
(BME-PA points: 20% · 3p = 0.6p)

[C8] Csapó Tamás Gábor, Németh Géza, ,,Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval,” *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged, Hungary), pp. 167–177, 2011. (in Hungarian)
(BME-PA points: 100% · 1p = 1p)

[C9] Tekla Etelka Gráczi, Steven M Lulich, Tamás Gábor Csapó, András Beke, ,,Context and speaker dependency in the relation of vowel formants and subglottal resonances - Evidence from Hungarian,” *Proc. Interspeech 2011*, (Florence, Italy), pp. 1901–1904, 2011.
(BME-PA points: 25% · 3p = 0.75p)

[C10] Géza Németh, Gábor Olaszy, Tamás Gábor Csapó, ,,Spemoticons: Text-To-Speech based emotional auditory cues,” *ICAD 2011*, (Budapest, Hungary), 2011.
(BME-PA points: 50% · 2p = 1p)

[C11] Csaba Zainkó, Tamás Gábor Csapó, Géza Németh, ,,Special Speech Synthesis for Social Network Websites,” *Lecture Notes In Computer Science*, 6231: pp. 455–463, Paper 58, 2010.
(BME-PA points: 50% · 6p = 3p)

[C12] Csapó Tamás Gábor, Németh Géza, ,,Mássalhangzó-magánhangzó kapcsolatok automatikus osztályozása szubglottális rezonanciák alapján,” *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged, Hungary), pp. 226-237, 2009. (in Hungarian)
(BME-PA points: 100% · 1p = 1p)

[C13] Géza Németh, Márk Fék, Tamás Gábor Csapó, ,,Increasing Prosodic Variability of Text-To-Speech Synthesizers,” *Proc. Interspeech 2007*, (Antwerp, Belgium), pp. 474–477, 2007.
(BME-PA points: 100% · 3p = 1.5p)