Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation

Tamás Gábor Csapó*, *Student Member, IEEE*, Géza Németh

Abstract—**Statistical parametric text-to-speech synthesis is optimized for regular voices and may not create high quality output with speakers producing irregular phonation frequently. A number of excitation models have been proposed recently in the hidden Markov-model speech synthesis framework, but few of them deal with the occurrence of this phenomenon. The baseline system of this study is our previous residual codebook based excitation model, which uses frames of pitch-synchronous residuals. To model the irregular voice typically occurring in phrase boundaries or sentence endings, two alternative extensions are proposed. The first, rule-based method applies pitch halving, amplitude scaling of residual periods with random factors and spectral distortion. The second, data-driven approach uses a corpus of residuals extracted from irregularly phonated vowels and unit selection is applied during synthesis. In perception tests of short speech segments, both methods have been found to improve the baseline excitation in preference and similarity to the original speaker. An acoustic experiment has shown that both methods can synthesize irregular voice that is close to original irregular phonation in terms of open quotient. The proposed methods may contribute to building natural, expressive and personalized speech synthesis systems.**

Index Terms— **Creaky Voice, Excitation, Glottalization, HMM, Irregular Phonation, Parametric, Residual, Speech Processing, Speech Synthesis, Vocal Fry, Voice Quality.**

I. Introduction

State-of-the-art text-to-speech (TTS) synthesis is often based on statistical parametric methods. Particular attention is paid to hidden Markov-model (HMM) based text-to-speech synthesis [1] (HTS), which has gained much popularity due to its flexibility, smoothness and small footprint. In this speech synthesis technique, the speech signal is decomposed to parameters which are fed to a machine learning system. After the training data is learned, during synthesis, the parameter sequences are converted back to speech signal with reconstructing methods (e.g. speech coders, vocoders).

A number of excitation models have been proposed recently. Statistical parametric speech synthesis and most of these excitation models are optimized for regular, modal voices (with quasi-periodic vibration of the vocal folds in voiced regions) and may not produce high quality with voices having frequent non-modal sections. Irregular voice (definition in Section II.B) is such a non-modal phonation mode, which has not been extensively modeled yet in hidden Markov-model based text-to-speech synthesis.

In this paper we extend our previous residual codebook based excitation model of HMM-TTS with two alternative irregular voice models. Section II presents a survey of currently available excitation models and speech processing techniques dealing with irregular voice. The baseline residual analysis-synthesis method and its integration into HTS is introduced in Section III. A rule-based model of irregular voice is presented in Section IV, while in Section V another model is proposed which is data-driven and uses a unit selection corpus. In Section VI two perceptual tests, while in Section VII an acoustic experiment and their results are shown. Finally, Section VIII presents the advantages and drawbacks of our methods and concludes the paper.

II. Related work

There are three main factors in statistical parametric speech synthesis that are needed to deal with in order to achieve as high quality synthesized speech as unit selection: vocoder techniques, acoustic modeling accuracy and over-smoothing during parameter generation [2]. In this paper, we investigate the first factor. Most HMM-based TTS systems are based on the source-filter theory [3]. However, over-simplified vocoder techniques (e.g. pulse-noise excitation, [1]) make the quality of synthesized speech of HMM-TTS poor compared to high-quality unit selection based text-to-speech synthesis systems. To overcome this drawback, a large number of improved excitation models have been proposed recently.

A. *Excitation models*

Mixed excitation [4], two-band excitation [5] and STRAIGHT-based vocoding [6] have been found to produce high quality HMM-based synthesized speech. The extension of mixed excitation with state-dependent filtering which

This research is partially supported by the Paelife (Grant No AAL-08-1-2011-0001)), the EITKIC_12-1-2012-001 and the CESAR (Grant No 271022) projects.

T.G. Csapó and G. Németh are with the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary. Mailing address: Budapest, H-1117, Magyar tudósok krt. 2. Phone: +36-1-463-3512. E-mails: csapot@tmit.bme.hu, nemeth@tmit.bme.hu.

Parts of this manuscript have been presented at the 8th Speech Synthesis Workshop [49].

resembles analysis-by-synthesis speech coding methods can model the excitation even better [7]. [8] proposes the use of the complex cepstrum to model the mixed phase characteristics of speech in statistical parametric synthesis. 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 [9] (LF) acoustic model of the glottal source derivative to construct the excitation signal and introduces Glottal Spectral Separation [10] which is claimed to have slightly better results than the STRAIGHT-based system [11]. Raitio and his colleagues use glottal inverse filtering [12] within HMM-based speech synthesis for generating natural sounding synthetic speech [13], [14]. The single pulse technique [13] is extended with a glottal pulse library [15] and unit selection yielding a hybrid approach based on pulse concatenation [16]. 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 [17]. The LF model is also used in [18] with a new glottal source separation method controlling the breathiness of synthesized speech. The model is extended with Gaussian noise resulting in a mixed source model [19]. 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 (MVF) [20] or voicing cut-off frequency [21], [22], [23] to decompose speech into harmonic and stochastic parts. The advantage of using MVF 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 (therefore no recording of EGG or approximation of the glottal source signal is necessary). In [21], the residual is parameterized by the amplitude spectrum and zero-phase criterion is used to synthesize the excitation frame. [24] improves this approach with spectrum normalization and codebook construction, and shows that this excitation model is comparable with that of HTS-STRAIGHT. In [25], characteristic waveforms are extracted from the residual and Waveform Interpolation (WI) is used. [26] extends this model with the concept of slowly and rapidly evolving waveforms resulting lower spectral distortion, while [27] adds time and frequency domain zero padding techniques to the WI model in order to further reduce the spectral distortion. Drugman and his colleagues construct a codebook of pitch-synchronous residuals which is compressed with Principal Component Analysis (PCA) [28]. In [29], the Deterministic Plus Stochastic Model (DSM) of the residual signal is proposed and integrated into HTS [30]. The deterministic part of the excitation contains the low-frequency content, while the stochastic component is high-pass filtered white noise, similarly to the HNM model. The authors argue that the first PCA eigenvector of residuals ('eigenresidual') is usually dominating the deterministic component; therefore using eigenvectors of superior ranks is not necessary. This results in a very simple model in which excitation periods are only parameterized by the pitch, while providing high-quality speech synthesis. We have proposed a residual codebook based excitation model using peak locations of the windowed residual frames and Harmonics-To-Noise ratio to parameterize the residual periods [31]. This model has been integrated into the HMM-TTS (HTS-CDBK, [32]). It differs from DSM and WI in using a larger codebook containing several thousand frames. A perception test found HTS-CDBK of higher quality than simple pulse-noise excitation. An advantage of these 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 example for such a different phonation mode is irregular phonation (see next section).

B. *Irregular phonation*

During regular phonation (modal 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, which is a non-modal phonation type and is referred to as irregular phonation. It leads to abrupt changes in the fundamental frequency (F0), amplitude of the pitch periods or both. Irregular phonation is also called glottalization, creaky voice, vocal fry and laryngealization, and is a frequent phenomenon in both healthy speakers and people having voice disorders. It is often accompanied by extremely low pitch and the quick attenuation of glottal pulses. Glottalization is perceived as a creaky, rough voice [33], [34]. Fig. 1 shows an example for glottalization. The horizontal arrow denotes the section where the phonation is irregular. The occurrence of glottalization depends on the prosodic structure (it often coincides with prosodic boundaries like silences [35] and stressed syllables [36]) and carries information from the speaker identity, his/her dialect, mood, emotional state and vocal-fold health [37], [38]. Irregular phonation 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 [39], [40], [41], for transforming modal voice to irregular [34], [42], [43], and to statistical parametric speech synthesis with creaky voice [44], [45], [46].

The first attempts to model irregular phonation were either in the formant synthesis domain [42] or relied on increasing jitter and shimmer of the speech signal [43]. In [34], a simple semi-automatic transformation method was developed that introduces irregular pitch periods into a modal speech signal, yielding irregular speech that is as rough and as natural as original glottalized speech. In [47], we deal with automatic irregular-to-regular voice transformation by manipulating the residual. A perception test found the method to decrease the perceived roughness of glottalized speech samples.

To model vocal fry in statistical parametric speech synthesis, [44] introduces a robust F0 measure, improved voicing estimation and two-band voicing, which improves significantly the quality of HMM-based speech synthesis. However, it does not focus on the characteristics of creaky excitation and thus does not deal with producing correct timbre. [45] derives an extension of the DSM model [29] which can handle creaky excitation by integrating secondary pulses in the residual. The residual is obtained from the first 'eigenresidual' of a given speaker, of which only the closed period is resampled to the target pitch in order to preserve the sharp distribution in the open period. Copy-synthesis experiments with a subjective evaluation showed that this extension improves the standard DSM vocoder. [48] investigates the usefulness of contextual factors for creaky voice prediction and experiments with adding parameter streams describing irregular phonation into the HMM-TTS framework. 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 [46]. However, there is only a small difference compared to the baseline system in overall naturalness if the creaky excitation was included or not. In [49] we proposed an irregular voice model in HTS that applies pitch-synchronous residual modulation with periods multiplied by random scaling factors. In a perception experiment the method has been found to synthesize speech that is more similar to the original speaker and more pleasant than the baseline HTS-CDBK system.

As shown above, there have been only a few studies dealing with glottalization in statistical parametric speech synthesis. [40] found, that up to 15% of the vowels of several American English speakers are produced with irregular phonation. As the occurrence of glottalization is not negligible in normal speech, a proper irregular model in HMM-TTS may contribute to create personalized voices, especially for speakers producing frequent irregular phonation (e.g. elderly).

III. Baseline: HMM-TTS with a Residual Codebook Based Excitation Model (HTS-CDBK)

We have proposed a residual codebook based excitation model [31] and integrated it into HMM-TTS [32], that will be used as the baseline system. The methods used in this system are summarized here briefly.

A. Analysis

The input is a speech waveform low-pass filtered at 7.6 kHz with 16 kHz sampling rate and 16 bit linear PCM quantization. First, a codebook of pitch-synchronous residuals is built from a small database (see Section III.D). After that, residual analysis is performed. The fundamental frequency (F0) parameters are calculated by the publicly available Snack pitch tracker [50] with 25 ms frame size and 5 ms frame shift. In the next step 34-order Mel-Generalized Cepstral analysis (MGC) [51] is performed on the speech signal with α=0.42 and γ=-1/3. The residual signal (excitation) is obtained by MGLSA inverse filtering [52]. Next, the SEDREAMS Glottal Closure Instant (GCI) detection algorithm is used to find the glottal period boundaries in the voiced parts of the residual signal [53].

The further analysis steps are completed on the residual signal with the same frame shift values. For measuring the parameters in the voiced parts, pitch synchronous, two period long frames are used according to the GCI locations and they are Hann-windowed. A codebook is built from pitch-synchronous residual frames. Several parameters of these frames are used to fully describe the speech residuals:

- F0: fundamental frequency of the frame
- gain: energy of the windowed frame
- rt0 peak indices: the positions of prominent values (peaks or valleys) in the windowed frame (see Fig. 2)
- HNR: Harmonics-To-Noise ratio of the frame [54]

For each voiced frame, one codebook element is saved with the above parameters and the windowed signal is also stored without F0 normalization. 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. 2: the prominent values are determined by simple maximum / minimum peak picking in the windowed residuals. The position of the peaks is calculated as the distance from the main excitation in the middle (which corresponds to the instant of glottal closure). We found experimentally that it is advantageous to use four peaks i.e. one maximum and one minimum on both sides of the middle of the window (main excitation). Each peak should have a distance from the middle of the window exceeding 10% the length of the pitch period. In this case the consecutive rt0 parameters are slowly evolving enough and are suitable for machine learning in HTS. After the codebook has been built, during analysis of the speech corpus, the above parameters are extracted from each voiced frame. These parameters will be used for target cost calculations during synthesis (see III.C). In order to collect similar codebook elements, the Root Mean Squared Error (RMSE) is calculated between the pitch normalized versions of the codebook elements which will be used for concatenation cost. The normalization is done by resampling every frame to 40 samples. For unvoiced frames, only the gain parameter is calculated.

In the current approach the novelty compared to similar residual-based excitation models is the use of the rt0 parameter and the application of concatenation cost during residual unit selection.

B. Training

For training, the parameters of log(F0), log(gain), log(rt0) and log(HNR) of each frame are extracted to describe the residual and MGC is used for spectral representation. Logarithmic values are used as they were found to be more suitable in training experiments. F0 and rt0 are modeled with MSD-HMMs because these do not have values in unvoiced regions. MGC, HNR and gain are modeled as simple HMMs. The first and second derivatives of all of the parameters are also stored in the parameter files and used in the training phase. Altogether five streams of data are considered with the delta and delta-delta values resulting in 15 streams. Decision tree-based context clustering is used with context dependent labeling applied in the Hungarian version of HTS [55]. Independent decision trees are built for all the parameters and duration using a maximum likelihood criterion.

C. Synthesis

In the synthesis phase of HTS-CDBK the inputs are the parameters obtained from training (F0, gain, rt0 indices and HNR) generated by a maximum likelihood algorithm [56] and the codebook of pitch-synchronous residuals. 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 [57]. The target cost is the squared difference among the parameters (F0, rt0 and HNR) 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 (see Section III.A). 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. After that, the synthesized residual is lowpass filtered to 6 kHz and white noise is used in the frequency band above 6 kHz. 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.

Note that the computational cost of the residual unit selection during synthesis depends on the size of the codebook and the applied costs. In our experiments we found that using a small codebook the synthesis time might be suitable for real-time synthesis, therefore the method does not decrease the flexibility of the original HTS system.

D. Speech data

The speech data that was used for our experiments is a part of the PPBA database [58]. Ten Hungarian native speakers are included in this database, of which two males having F0 in the range of 60-250 Hz were chosen here for speaker dependent training (denoted FF3 and FF4). Speaker FF3 was 50 years old, while speaker FF4 was 66 years old at the time of the recordings. Both speakers produced irregular phonation frequently, mostly at the end of sentences. 1940-1940 phonetically balanced sentences (2-2 hours of speech) from the two speakers were used as training corpora. The sentences in the corpus are stored as waveform files (44.1 kHz sampling rate, 16 bit linear PCM quantization), which were resampled to 16 kHz. We created a residual codebook with 3394 elements for speaker FF3 and another one with 2218 elements for speaker FF4 extracted from about 10 minutes of speech from the first 150 sentences. It has been found that codebooks of similar scale are enough for high-quality speech synthesis, therefore we did not use the whole database for residual codebook construction [17].

E. Irregular voice handling in the baseline system

We have analyzed the training speech databases of the two speakers and conducted speaker dependent training. During the analysis, it was found that when glottalization occurs (typically in the vowels of the last syllables of the phrases), the Snack pitch tracker cannot measure F0 and sets the frame as being unvoiced. Note that the sound boundaries in the PPBA database were corrected manually. This way the most probable reason for unvoiced frames in a vowel is the sentence-final creaky voice. Therefore, this pattern is learned by the system and glottalization is modeled in HTS-CDBK similarly to unvoiced speech. During synthesis unvoiced excitation is often generated at the last vowels of the sentences. This produces a very unpleasant voice and is not a proper model of glottalization. Fig. 3 a) and b) show an example for the end of a sentence synthesized by the baseline system with the residual (a) and the final speech waveform (b). In the section denoted by a blue horizontal arrow unvoiced excitation was generated for some part of the vowel ‘á’ /a:/, and therefore there is only aperiodic noise at the end of the speech signal. This makes the synthesized speech very unnatural.

IV. Proposed #1: HTS-CDBK extended with a rule-based irregular voice model (HTS-CDBK+Irreg-Rule)

A rule-based irregular voice model is proposed which is hypothesized to better model glottalization than the baseline system. The basic idea of the method comes from [34] which is summarized first. The basic method is further improved to yield automatic transformation and integrated into HTS-CDBK. The novel system is denoted as HTS-CDBK+Irreg-Rule [49].

The analysis and training steps are the same as in the baseline system (Sections III.A and III.B, respectively), and the same speech data was used; only the synthesis step is different. During synthesis, this rule-based irregular voice model applies 1) pitch halving, 2) pitch-synchronous residual modulation with periods multiplied by randomized amplitudes and 3) spectral distortion.

A. Semi-automatic regular to irregular transformation

In [34], a regular-to-irregular voice transformation method was presented which uses amplitude scaling of individual glottal cycles. Here, the regular speech is pitch-synchronously windowed, the periods are multiplied by individual hand-selected scaling factors and finally speech is overlap-added from the modified signal. The scaling factors can either boost, attenuate, remove or leave unmodified the cycles. [34] extends this with stylized pulse pattern copying yielding in a semi-automatic transformation method.

In the present form, this method is not suitable to fit to HTS; partly because it is manual or semi-automatic and as it works on the speech signal itself and not on excitation. However, the concepts of this transformation method were re-used and further improved yielding in an automatic irregular voice model that was integrated into HTS-CDBK.

B. Detection of sections to synthesize irregularly

There is no explicit glottalization model (e.g. irregular phonation labels, questions for decision trees) in the HTS-CDBK+Irreg-Rule system, so sections with irregular phonation should be found from the generated F0 sequence. In our initial experiments the generated parameter and label files were checked automatically. Glottalization was applied if at least five consecutive frames (1*25 ms first frame + 4*5ms frame shifts, altogether 45 ms) were given zero F0 within a vowel. In these cases, the fundamental frequency was interpolated between the modal voiced parts to have a straight F0 line, or it was set to slightly decreasing if there were no voiced neighboring sounds.

By applying high precision F0 detection like in [44] it may be possible to obtain better F0 contours for training, but in this way we would lose the information where to apply glottalization in synthesis.

C. Synthesis

Fig. 4 shows the steps of the rule-based irregular voice model. 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 than modal speech (see Section II.A), and [34] argues that by removing every second or third cycle the perception of samples is similar to decreasing the open quotient. In the residual codebook, frames with extremely low F0 are rare. Therefore, during synthesis, residual frames are zero padded which results in a similar effect than removing every second cycle.

During residual synthesis, each pitch cycle is multiplied by a random amplitude scaling factor in the range of {0…1}. This amplitude scaling is similar to [34] but we do not boost any of the periods, only remove, attenuate or not modify them. Another important aspect is that using random scaling factors we do not need to apply manual scaling or semi-automatic pulse pattern copying. From the modified residual periods the whole residual signal is obtained by pitch-synchronously overlap-adding the frames as in Section III.C.

Finally, spectral distortion is applied. In [47] we found that the extracted MGC parameters of irregularly phonated speech are less smooth than those of regular speech. Therefore here we try to ‘distort’ the MGC parameters similarly by slightly modifying them: the parameter values are multiplied by random numbers having uniform distribution between {0.995…1.005}. This yields a less smooth parameter sequence for each dimension of MGC. MGC is a representation of the linear prediction spectrum which is suitable for interpolation or perturbation adding, therefore this step does not result in instabilities.

The residual sections that are unvoiced or should be synthesized with modal voiced phonation are created similarly as in the baseline system. Finally, synthesized speech is reconstructed by MGLSA filtering using the MGC parameters.

Fig. 3 shows an example for the results of the baseline (HTS-CDBK: a, b) and the proposed #1 systems (HTS-CDBK+Irreg-Rule: c, d). In a) and b) the blue horizontal arrow shows the section where the excitation is unvoiced within the vowel ‘á’ /a:/ in HTS-CDBK. As this section is longer than five frame shifts (45 ms), we apply glottalization for this vowel in the HTS-CDBK+Irreg-Rule system. In c) and d) the proposed residual and speech signal are shown and a red dashed horizontal line indicates the glottalized vowel ‘á’ /a:/. 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. 1) than the baseline system.

V. Proposed #2: HTS-CDBK extended with a data-driven irregular voice model (HTS-CDBK+Irreg-Data)

Another model of irregular phonation was created which is data-driven and based on the unit selection synthesis method. This system is denoted as HTS-CDBK+Irreg-Data. It is hypothesized that this system will produce synthesized irregular phonation that is closer to natural glottalization than that of the baseline system or HTS-CDBK+Irreg-Rule.

The analysis of the baseline system is extended with the creation of an irregular voice dataset: a corpus from residuals of irregularly phonated vowels is built. The training step is the same as in the baseline system and the same speech data was used; the difference is again in the synthesis step. During synthesis, we search for suitable vowel-length residuals from the corpus that fit in the residual signal and apply spectral distortion. Note that these vowel-length residuals differ from the pitch-synchronous residual frames of Section III.C in their length.

A. Analysis

First the five parameters are extracted from the speech database the same way as in the baseline system. After that we apply a recent high-precision creaky voice detection algorithm in the speech database [41].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.

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.

B. *Synthesis*

The sections to synthesize with irregular phonation are decided based on the F0 stream similarly to IV.A. Fig. 5 shows the steps of the data-driven irregular voice model.

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 the neighboring sounds will be modal. With this assumption we do not have to deal with concatenation among vowel-length residuals in the GLOTT corpus. 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 between the section to glottalize and the vowel-length residuals of the corpus 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. The context of the target residual is used to find vowel-length residuals which originate from suitable neighboring sounds. 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. This ensures that the synthesized speech will be as close to original irregular speech as possible.

MGC distortion is similarly applied here as in the HTS-CDBK+Irreg-Rule system. The residual sections that are unvoiced or should be synthesized with modal voiced phonation are created similarly as in the baseline system. Finally, synthesized speech is reconstructed by MGLSA filtering using the MGC parameters.

Fig. 3 e) and f) show an example for the residual and speech waveform of the proposed #2 system. Similarly to the baseline 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. This might be a better model than that of the proposed #1 system (c and d) where the whole vowel was synthesized with the rule-based irregular model. When comparing Fig. 3 e) and f) with Fig. 1, we can see that the irregular vowel of the proposed #2 method is close to the original irregular vowel, and the synthesized residual contains several secondary pulses similarly to the original residual of Fig. 1. However, it is a question whether the residual found by the unit selection fits to the overall residual of the sentence.

VI. Perceptual experiments

In order to evaluate the quality that can be achieved by the proposed HTS-CDBK+Irreg-Rule and HTS-CDBK+Irreg-Data methods, we have conducted two listening tests. A major factor that determines the usefulness of these methods is if human listeners accept the synthesized speech extended with an irregular voice model.

Therefore, our aim was to measure the perceived 'pleasantness' and the similarity to the original speaker. We compared speech synthesis samples of the baseline system with samples of the proposed systems. The other purpose of the evaluation was to find the better representation of irregular voice from the two solutions proposed.

A. *Methods of subjective experiment #1*

To obtain the speech stimuli, we created four voice models with the baseline and proposed #1 systems and the two speakers [49]. 130-130 sentences not included in the training database were synthesized with all four voice models and 10-10 sentences having at least one irregularly synthesized vowel at the end were selected for the subjective test. The last word (with at least two syllables) of each sentence was cut and used as stimuli as we wanted the listeners to focus only on the sentence endings. An example for a word that was included in the test can be seen on Fig. 3 a-d.

In the test, the two versions of each word were included, resulting altogether 40 utterances (2 speakers * 10 words * 2 versions). We created a web-based paired comparison test with two CMOS-like questions. Before the test, listeners were asked to listen to an example from speaker FF3. In the first part of the test, the listeners had to rate their preference ('Which version do you think is more pleasant?', '1 – first is much more pleasant' … '5 – second is much more pleasant'). In the second part of the test, they were asked which version is more similar to the original speaker. For this, a reference creaky speech sample was shown first and the two stimuli after that ('Which version is more similar to the original speaker?', '1 – first is more similar', '2 – equal', '3 – second is more similar'). The utterances were presented in a randomized order (different for each participant).

B. *Results of subjective experiment #1*

Altogether 11 listeners participated in the test. They were all students or computer science professionals, between 19-31 years (mean: 24 years). All of them were native speakers of Hungarian and none of them reported any hearing loss. On average the whole test took 9 minutes to complete.

The results of this listening test are presented in Fig. 6 for the two speakers. The figure provides a comparison between the baseline HTS-CDBK system (left part, blue color) and the proposed HTS-CDBK+Irreg-Rule system (right part, red color). Green color in the middle shows the percentage of equal answers. The answers of the listeners for the first question were pooled together for the visualization: the levels 1 and 2 are included in the left blue bar, the level 3 is shown in the middle green bar, whereas the levels 4 and 5 are included in the right red bar. It can be seen that for the preference question, for both speakers the results are higher than the equal answer of 50% (CMOS score=3.0) meaning that the proposed system was more preferred (mean CMOS for the speakers together: 3.36). Similarity scores are higher than the equal 50% (CMOS=2.0) for both speakers FF3 and FF4 (mean altogether: 2.38). The ratings of the listeners were compared by t-tests as well. The statistical analysis showed that the proposed method was significantly preferred in terms of pleasantness ($p<0.0005$) and was significantly more similar to the original speaker ($p<0.0005$) than the baseline system. By investigating the scores for the stimuli one by one, we found that all of the utterances of the proposed system ranked higher in the similarity test, while in 18 out of 20 sample pairs the extended model was preferred. Fig. 6 does not show any speaker dependency in the results, and the differences are similarly significant when we analyze the answers for the two speakers separately.

From this subjective experiment, we can conclude that the HTS-CDBK+Irreg-Rule system improves the perceived naturalness of synthesized speech using a rule-based irregular voice model and the proposed method can generate speech that is more similar to the original speaker.

C. *Methods of subjective experiment #2*

Another listening test was conducted for measuring the acceptability of the HTS-CDBK+Irreg-Data system compared to the baseline and the HTS-CDBK+Irreg-Rule systems. Two more voice models were created with the proposed #2 system and the two speakers. The methods applied here were similar to that of subjective experiment #1. An example for a word that was included in the test can be seen on Fig. 3 a-f.

In this test, three versions of each of the 10 words were included in a paired comparison, resulting altogether 80 utterance pairs (versions of the pairs: baseline vs. proposed #2 and proposed #1 vs. proposed #2). A similar test was created as in Section VI.A with the same questions.

D. *Results of subjective experiment #2*

Altogether 17 listeners participated in the test (partly different from test #1). All of them were Hungarian students or speech technology professionals, between 19-65 years (mean: 32 years). One of the listeners reported hearing loss, therefore she was excluded from the evaluation, and the results of the remaining 16 subjects were analyzed. On average the whole test took 17 minutes to complete.

Some of the listeners of subjective test #2 reported that it might have been useful to add an answer to the second question that ‘none of them is similar’ to the original speaker. In these cases, they evaluated the utterance pair as equal.

Fig. 7 shows the evaluation results of the baseline (left part, blue color) vs. HTS-CDBK+Irreg-Data (right part, red color) systems, while Fig. 8 presents the differences between the rule-based (left part, blue color) and the data-driven (right part, red color) irregular voice models. In both figures, green color in the middle shows the percentage of equal answers. The answers of the listeners for the first question were pooled together for the visualization, similarly as in Fig. 6.

The result of comparing the utterances of the baseline HTS-CDBK system with the data-driven HTS-CDBK+Irreg-Data system is shown in Fig. 7 with the two speakers and two questions separately. The CMOS scores of pleasantness and similarity show that the proposed #2 system was preferred: for the first question mean CMOS=3.36 which is significantly different from 3.0 ($p<0.0005$) and for the second question mean CMOS=2.28 which is significantly different from 2.0 ($p<0.0005$). The results are similarly significant when we investigate the two speakers separately.

The evaluation results of the two alternative irregular voice models are presented in Fig. 8. In terms of preference, there is no significant difference between the utterances synthesized by the models (mean CMOS=3.07; no significant difference from 3.0; p=0.16). The listeners also did not perceive significant differences in the similarity to the original speaker between the two methods (mean CMOS=1.95, no significant difference from 2.0; p=0.23). When we investigate the two speakers separately, listeners found that for speaker FF3 the rule-based irregular model was a little closer to the original speaker; whereas for speaker FF4 the data-driven method was found to be slightly more similar to original creaky utterances.

One of the listeners reported that in certain utterances he found the stimuli to be extremely creaky which does not occur in natural speech. After investigating the stimuli, we found that this observation might be the result of the too sharp amplitude changes in the rule-based irregular samples. However, other subjects did not perceive this as disturbing.

From this second subjective experiment, we can draw the conclusions that 1) the irregular voice of HTS-CDBK+Irreg-Data system was preferred over the baseline in terms of pleasantness and similarity to the original version for both speakers 2) the results of the HTS-CDBK+Irreg-Rule and HTS-CDBK+Irreg-Data models are not significantly different in terms of preference and similarity.

VII. ACOUSTIC EXPERIMENT

The perception tests showed that both proposed irregular voice models are preferred over the baseline system, and they can generate speech that is more similar to the original speaker. However, from the listening test results it is not known whether the proposed systems model irregular voice properly or it was just preferred to use other excitation instead of white noise in the investigated vowels. The basis of our acoustic experiment is the one presented in [34].

A. *Acoustic properties of irregular phonation*

Voice quality has a number of acoustic correlates consistently reported in the literature (e.g. [33]). In natural speech, irregular phonation can be distinguished from regular phonation by several properties [34], [42]:

- the time that is elapsed between successive glottal pulses is longer and more irregular, resulting in lower F0 and higher jitter
- the overall intensity level is lower
- abrupt changes occur in the amplitude of the pitch periods
- the open quotient (proportion of the glottal cycle where the glottis is open) is lower
- first formant bandwidth is increased because of more acoustic losses at the glottis
- the closure of the vocal folds is more abrupt, i.e. spectral tilt is lower

Some of these properties are observable in both the speech signal and in the residual signal. An example for this can be seen in Fig. 1. The bottom (b) shows a section of speech, while the top (a) is the residual of this speech signal. In the irregularly phonated interval (denoted by an arrow) the pitch is lower and the periods have clearly abrupt changes in amplitude.

In the acoustic experiment the three most important acoustic cues [34], [42] are used: open quotient (OQ), first formant bandwidth (B1) and spectral tilt (TL). OQ and TL are expected to be lower for irregular phonation, while B1 is increased compared to regular voice. If the synthesized utterances match these correlates, that might provide an explanation for their perceptual acceptability.

B. Methods of the acoustic experiment

The above parameters are more convenient to consider in the frequency domain; therefore the changes in H1-H2 (the difference of the amplitudes of the first two harmonics), H1-A1 (H1 relative to the first formant amplitude) and H1-A3 (H1 relative to the third formant amplitude) were measured which are correlated with OQ, B1 and TL, respectively [59]. These parameter values can be biased by the effects of the first three formants (F1, F2 and F3). To compensate this, we used the equations suggested by [60] and implemented in VoiceSauce [61]: the value of H1 and H2 was corrected for F1 and F2 frequencies (H1* and H2*), and the value of A3 was corrected for F1, F2, and F3 (A3*). Altogether, the measurement of the following parameters was necessary: H1, H2, F1, F2, F3, A1, A3 and frequencies of H1 and H2.

The measurements were conducted partly on the stimuli used in the perceptual evaluation (10-10 words synthesized by the baseline system, proposed #1 and proposed #2 models). The other part of the investigated speech material consisted of 10-10 original regular and original irregular vowels selected from the PPBA database from both speakers (we selected words which were available in both versions). Altogether the parameters of 100 vowels were measured. First the wave files were resampled to 8 kHz (this ensured that only the spectrum of 0-3.8 kHz was visible). Then a glottalized vowel from the original irregular version was selected and three points within the vowel (roughly equally spaced and aligned with the pitch marks) were chosen and the same vowel was measured in the original regular version. In the synthesized versions, the vowels created by the baseline system and the irregular voice models were measured. In Wavesurfer [62], the 512-point FFT spectrum, calculated using a Hamming window, was displayed at the chosen locations and the parameters were graphically measured. In the irregular versions often strong subharmonics appeared; here we measured H1 and H2 as the lowest two of the spectral peaks. We approximated the formant frequencies and amplitudes by the frequency and amplitude of the strongest harmonic in the formant peak. In the utterances of the baseline system sections of the vowels contained unvoiced excitation. Here we measured H1 and H2 as the two lowest peaks in the spectrum similarly to the voiced cases.

C. Results of the acoustic experiment

The mean values of H1*-H2* (proportional to OQ), H1*-A1 (proportional to 1/B1) and H1*-A3* (proportional to TL) are shown in Fig. 9 for the five utterance versions separately. In one-way ANOVAs, stimulus type had significant effects on all three acoustic parameters (F(4,295)=11.89, 7.70, 4.49, respectively; p<0.005). Tukey-HSD post hoc test was used to compare the mean parameter values of each stimulus type.

H1*-H2* was similar for the baseline utterances and for the original regular speech (p=0.37, n.s.). It was almost the same for the original irregular and for the synthesized irregular recordings (difference not significant; p=0.99). However, the regular versions were significantly different from the irregular versions (p<0.0005). This means that in terms of open quotient, the synthesized versions are close to the original irregular versions. In terms of H1*-A1, the homogeneous subsets are the original regular plus synthesized baseline which differ significantly from the original irregular plus synthesized rule-based irregular (p<0.05). In the figure we can see the trends that the irregular voice models have created. In terms of the H1*-A1 and first formant bandwidth the rule-based synthesized irregular utterances are very similar to the original irregular recordings. The data-driven model has resulted in B1 parameters that are between original regular and irregular vowels. In this experiment, H1*-A3* was not helpful to differentiate between the regular and irregular utterances. Only the synthesized baseline versions are significantly different from other versions (p<0.05). This might be caused by the wrong measurement of H1 in unvoiced segments of the baseline system. According to Fig. 9 and the statistical analysis, the spectral tilt parameters measured on the recordings do not show a tendency.

From the acoustic experiment the conclusion is that the proposed irregular models can reconstruct two of the three investigated acoustical correlates of irregular speech. An explanation for the higher difference in the acoustic parameters between original irregular speech and utterances synthesized by the data-driven model can be that in the second version only smaller sections of the vowels have irregular properties (see Fig. 3 e-f) and thus the measurements in the middle and end of the vowel did not catch the acoustic correlates of glottalization.

VIII. Discussion and conclusions

This paper presented two alternative methods to synthesize irregular voice within the HTS framework. The first, rule-based method uses pitch halving, amplitude scaling of the pitch periods of the residual signal and spectral distortion. The second, data-driven model builds a corpus of irregular vowel residuals and searches for suitable vowel-length residuals from this corpus with unit selection methods.

The first method is fully automatic because amplitude scales are determined randomly and no manual scaling is necessary. By adding jitter and shimmer, or applying predefined stylized pulse patterns as in [34] instead of random scaling factors, the naturalness of synthesized glottalization might be further improved.

The second method is data-driven in a sense that residuals extracted from irregular sections of speech are re-used during synthesis. The method might be further improved by adding concatenation cost between irregular residuals to process longer irregular sections in synthesized speech.

By applying an irregular vs. regular classification algorithm (e.g. [40], [41]), glottalization could be modeled explicitly in both models, and a more reliable decision could be made where to use irregular voice instead of the decision based on the generated F0 sequence of the current models. A recent study in this topic found that the use of posterior probability of a detection algorithm [41] gives the best performance for creaky voice prediction [48].

Compared to the first experiments of statistical speech synthesis with creaky voice our models differ in several properties. [44] deals with the parameterization of vocal fry and tries to remove irregular sections from synthesized sentences, while we keep the ratio of the occurrence of glottalized vowels of training corpora in the synthesized speech. [45] extends the Deterministic plus Stochastic Model and concentrates on using one period of 'eigenresidual' obtained from a dataset of creaky utterances, but does not deal with acoustic correlates of irregular voice. It was found that creaky voice typically occurs in the last syllable of phrases [48], therefore the prediction of irregular phonation in HTS results in mainly the last and second last syllable. This is similar to our observations. Recently, the HTS-DSM vocoder has been extended with creaky voice prediction and creaky synthesis [46]. In subjective tests on English and Finnish data, an improvement has been found compared to the baseline DSM model in terms of creakiness, however there was no difference in naturalness. In the future, we plan to compare the HTS-CDBK+Irreg-Rule and HTS-CDBK+Irreg-Data models with the above creaky speech synthesis systems.

With the proposed methods we extend previous speech processing techniques dealing with irregular phonation. Experiments on the synthesized speech of two speakers have proven the appropriateness of our methods to synthesize irregular-like speech. Perception tests of short speech segments found the proposed models to be suitable to synthesize glottalized speech that is closer to the original speaker while increasing naturalness. According to an acoustic experiment, both irregular voice models are able to reconstruct open quotient values that are close to original irregular voice, while the rule-based system was similar to natural glottalized speech in terms of the first formant bandwidth.

Our methods may contribute to building natural, expressive and personalized speech synthesis systems. Irregular phonation is frequently adopted in lively story-telling and natural interactive conversation [48]. During the analysis of Hungarian expressive voices, [63] found that glottalization is one of the cues of sadness; while in Japanese the creaky voice may convey that the speaker is displaying an attitude of being under high pressure [64] and in British English it tends to signal boredom [65]. Therefore the proposed algorithms might be suitable to extend expressive speech synthesis systems. Assistive communication systems are helpful for individuals with speech impairment, but usually the text-to-speech voice does not reflect the user's vocal quality or personality [66]. A specific example for the application of the irregular phonation models in personalized systems is to create 'elderly voices' where glottalization occurs very frequently.

ACKNOWLEDGMENT

The authors thank the listeners for participating in the subjective tests. They would like to thank John Kane and Thomas Drugman for kindly providing the creaky voice detection algorithm. Last, but not least, they would also like to thank the anonymous reviewers for their valuable comments and questions.

REFERENCES

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

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[3] G. Fant, Acoustic theory of speech production. The Hague: Mouton, 1960, pp. 15–20.

[4] T. Yoshimura and K. Tokuda, "Mixed excitation for HMM-based speech synthesis," in Proc. Eurospeech, 2001, pp. 2263–2266.

[5] S. Kim and M. Hahn, "Two-Band Excitation for HMM-Based Speech Synthesis," IEICE Transactions on Information and Systems, vol. E90-D, no. 1, pp. 378–381, Jan. 2007.

[6] 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.

[7] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in Proc. Interspeech, 2007, pp. 1909–1912.

[8] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum for statistical parametric speech synthesis," Speech Communication, vol. 55, no. 5, pp. 606–618, Feb. 2013.

[9] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR, vol. 4, pp. 1–13, 1985.

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

[11] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in Proc. ICASSP, 2011, pp. 4704–4707.

[12] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," Speech Communication, vol. 11, no. 2–3, pp. 109–118, Jun. 1992.

[13] 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, 2008, pp. 1881–1884.

[14] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[15] T. Raitio, A. Suni, and H. Pulakka, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in Proc. ICASSP, 2011, pp. 4564–4567.

[16] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach," in Blizzard Challenge 2012, http://festvox.org/blizzard/bc2012/HELSINKI_Blizzard2012.pdf, accessed Oct 8, 2012.

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

[18] P. Lanchantin, G. Degottex, and X. Rodet, “A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method,” in Proc. ICASSP, 2010, pp. 4630–4633.

[19] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, “Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis,” Speech Communication, vol. 55, no. 2, pp. 278–294, Feb. 2013.

[20] D. Erro, I. Sainz, E. Navas, and I. Hernáez, “Improved HNM-based Vocoder for Statistical Synthesizers,” in Proc. Interspeech, 2011, pp. 1809–1812.

[21] Z. Wen and J. Tao, “An excitation model based on inverse filtering for speech analysis and synthesis,” in IEEE MLSP, 2011.

[22] Z. Wen and J. Tao, “Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis,” in Proc. Interspeech, 2011, pp. 1805–1808.

[23] Z. Wen, H. Kawahara, and J. Tao, “Pitch-Scaled Analysis based Residual Reconstruction for Speech Analysis and Synthesis,” in Proc. Interspeech, 2012, pp. 374–377.

[24] Z. Wen and J. Tao, “Amplitude spectrum based Excitation model for HMM-based Speech Synthesis,” in Proc. Interspeech, 2012, pp. 1428–1431.

[25] J. S. Sung, D. H. Hong, K. Oh, and N. Kim, “Excitation modeling based on waveform interpolation for HMM-based speech synthesis,” in Proc. Interspeech, 2010, pp. 813–816.

[26] C. Jung, Y. Joo, and H. Kang, “Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems,” IEEE Signal Processing Letters, vol. 19, no. 12, pp. 809–812, Dec. 2012.

[27] 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.

[28] 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, 2009, pp. 3793 – 3796.

[29] T. Drugman, G. Wilfart, and T. Dutoit, “A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis,” in Proc. Interspeech, 2009, pp. 1779–1782.

[30] T. Drugman and T. Dutoit, “The Deterministic Plus Stochastic Model of the Residual Signal and its Applications,” IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 3, pp. 968–981, 2012.

[31] T. G. Csapó and G. Németh, “A novel codebook-based excitation model for use in speech synthesis,” in IEEE CogInfoCom, 2012, pp. 661–665.

[32] T. G. Csapó and G. Németh, “Statistical parametric speech synthesis with a novel codebook-based excitation model,” Intelligent Decision Technologies, 2013, accepted.

[33] 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, no. 5, pp. 2649–2658, May 1998.

[34] 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, 2008, pp. 6141–6146.

[35] J. Slifka, “Irregular phonation and its preferred role as a cue to silence in phonological systems,” in ICPhS, 2007, pp. 229–232.

[36] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, “Glottalization of word-initial vowels as a function of prosodic structure,” Journal of Phonetics, vol. 24, no. 4, pp. 423–444, Oct. 1996.

[37] C. Gobl and A. N. Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” Speech Communication, vol. 40, no. 1–2, pp. 189–212, Apr. 2003.

[38] N. Malyska, “Analysis of nonmodal glottal event patterns with application to automatic speaker recognition,” PhD thesis, MIT, 2008, http://dspace.mit.edu/handle/1721.1/43804, accessed Mar 20, 2013.

[39] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, “A Method for Automatic Detection of Vocal Fry,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 47–56, Jan. 2008.

[40] T. Bőhm, Z. Both, and G. Németh, “Automatic Classification of Regular vs. Irregular Phonation Types,” in NOLISP, 2009, pp. 43–50.

[41] J. Kane, T. Drugman, and C. Gobl, “Improved automatic detection of creak,” Computer Speech & Language, vol. 27, no. 4, pp. 1028–1047, Jun. 2013.

[42] 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, no. 2, pp. 820–857, Feb. 1990.

[43] A. V. McCree and T. P. Barnwell, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[44] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, “Parameterization of vocal fry in HMM-based speech synthesis,” in Proc. Interspeech, 2009, pp. 1775–1778.

[45] T. Drugman, J. Kane, and C. Gobl, “Modeling the Creaky Excitation for Parametric Speech Synthesis,” in Proc. Interspeech, 2012, pp. 1424–1427.

[46] T. Raitio, J. Kane, T. Drugman, and C. Gobl, “HMM-based synthesis of creaky voice,” in Proc. Interspeech, , 2013, pp. 2316–2320.

[47] T. G. Csapó and G. Németh, “Transformation of irregular voice to regular voice by residual analysis and synthesis,” IEEE Signal Processing Letters, 2013, in preparation.

[48] T. Drugman, J. Kane, T. Raitio, and C. Gobl, “Prediction of Creaky Voice from Contextual Factors,” in Proc. ICASSP, 2013, pp. 7967–7971.

[49] T. G. Csapó and G. Németh, “A novel irregular voice model for HMM-based speech synthesis,” in Proc. ISCA SSW8, 2013, pp. 229–234.

[50] “The Snack Sound Toolkit [Computer program], Version 2.2.10.” http://www.speech.kth.se/snack/, accessed Sep 15, 2012.

[51] “Reference Manual for Speech Signal Processing Toolkit, Ver. 3.5.” http://sp-tk.sourceforge.net/, accessed Dec 25, 2011.

[52] S. Imai, K. Sumita, and C. Furuichi, “Mel Log Spectrum Approximation (MLSA) filter for speech synthesis,” Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, pp. 10–18, 1983.

[53] T. Drugman and M. Thomas, “Detection of glottal closure instants from speech signals: a quantitative review,” IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 3, pp. 994–1006, 2012.

[54] 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, no. 2, pp. 254–266, Apr. 1993.

[55] 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.

[56] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, “An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features,” in Proc. Eurospeech, 1995, pp. 757–760.

[57] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. ICASSP, 1996, vol. 1, pp. 373–376.

[58] 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], 2013.

[59] 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, no. 6, pp. 1212–1223, Dec. 1995.

[60] 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, 2004, pp. 669–672.

[61] “VoiceSauce - A program for voice analysis [Computer program], Version 1.12.” http://www.ee.ucla.edu/~spapl/voicesauce/index.html, accessed Nov 8, 2012.

[62] K. Sjölander and J. Beskow, “Wavesurfer [Computer program], Version 1.8.5.”, http://www.speech.kth.se/wavesurfer/, accessed Apr 3, 2009.

[63] Cs. Zainkó, M. Fék, and G. Németh, “Expressive Speech Synthesis Using Emotion-Specific Speech Inventories,” Lecture Notes in Computer Science, no. 5042, pp. 225–234, 2008.

[64] T. Sadanobu, “A natural history of Japanese pressed voice,” Journal of the Phonetic Society of Japan, vol. 8, no. 1, pp. 29–44, 2004.

[65] J. Laver, The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press, 1980.

[66] C. Jreige, R. Patel, and H. T. Bunnell, “VocaliD: personalizing text-to-speech synthesis for individuals with severe speech impairment,” in ASSETS, 2009, pp. 259–260.

Tamás Gábor Csapó (S’10) was born in Mosonmagyaróvár, Hungary, in 1985. He obtained his MSc in computer science from the Budapest University of Technology and Economics (BME),

Budapest, Hungary in 2008. Since 2008, he has been a PhD student at the Speech Technology Laboratory of BME.
In 2007, Mr. Csapó was awarded with 1st prize of the National Conference of Scientific Student's Associations, Hungary. He received a CIRE student grant of the Acoustical Society of America in 2010 and a Fulbright scholarship in 2013. His research interests include speech synthesis, speech analysis, excitation models and subglottal resonances in speech acoustics.

Géza Németh was born in 1959. He obtained his MSc in electrical engineering, major in Telecommunications at the Faculty of Electrical Engineering of BME in 1983. Also at BME: dr. univ, 1987, PhD 1997.
He is an associate professor at BME. He is the author or co-author of more than 140 scientific publications and 4 patents. His research fields include speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications.
He is the Head of the Speech Technology Laboratory of BME.

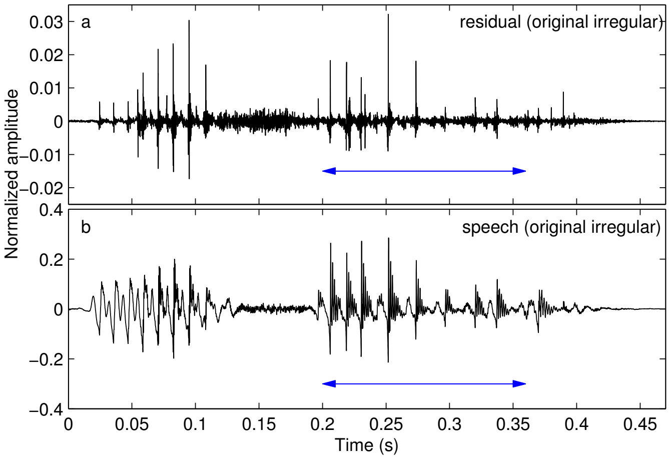

Fig. 1. *A speech recording of the word 'Mihály' /miha:j/ having irregular phonation at the section denoted by an arrow. a) inverse filtered residual signal b) speech signal.*

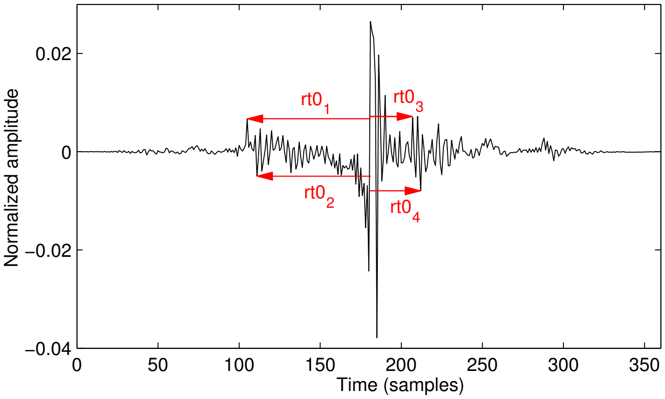

Fig. 2. *Calculation of the rt0 parameter for a windowed residual segment in the baseline system. rt0$_i$ is the distance of prominent peaks from the main impulse, in samples.*

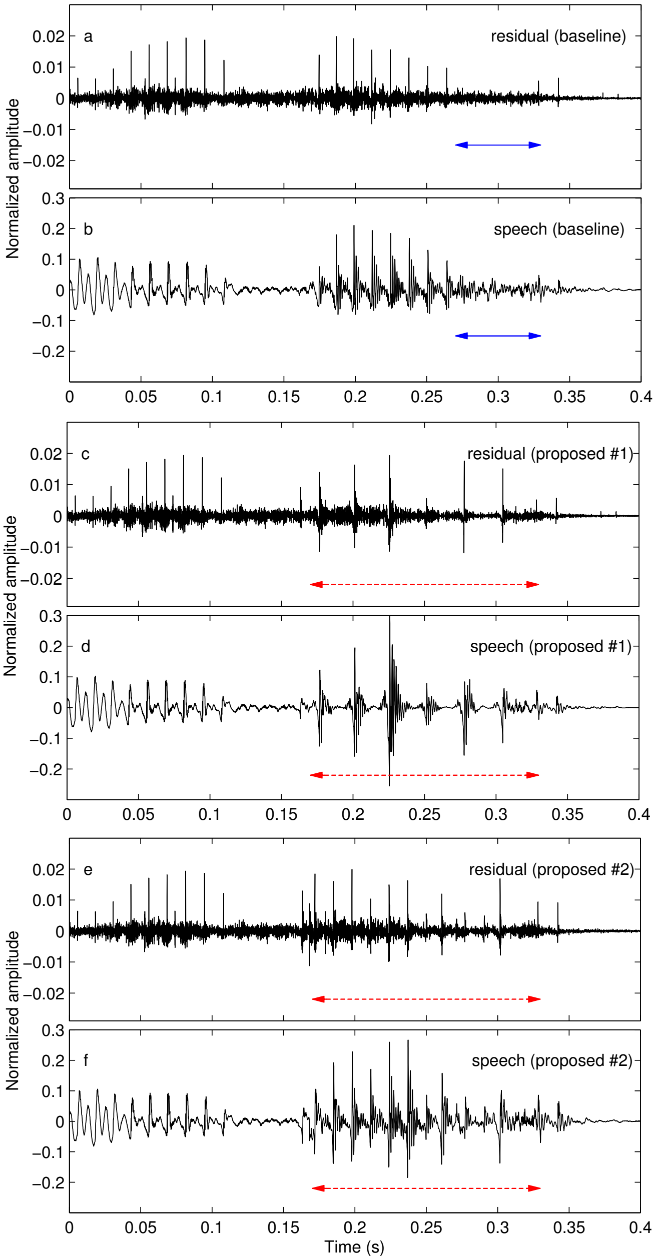

Fig. 3. *Synthesized versions of the word 'Mihály' /miha:j/ extracted from the end of a longer sentence with a) and b) from the baseline system; c) and d) from the proposed system #1(HTS-CDBK+Irreg-Rule) and e) and f) from the proposed system #2 (HTS-CDBK+Irreg-Data).*

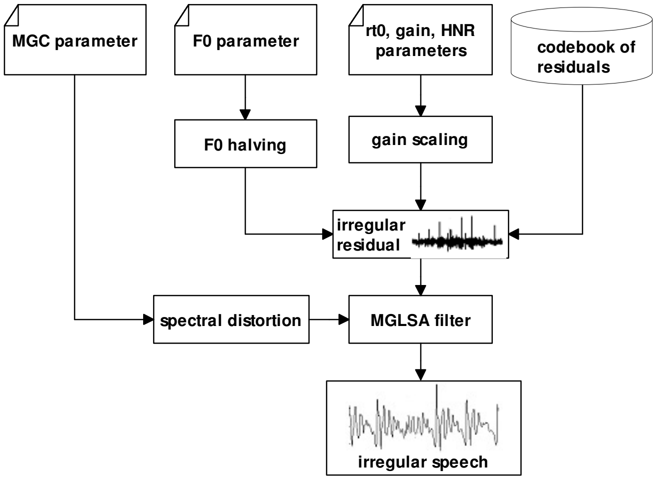

Fig. 4. *Irregular voice synthesis in the proposed #1 (rule-based) system.*

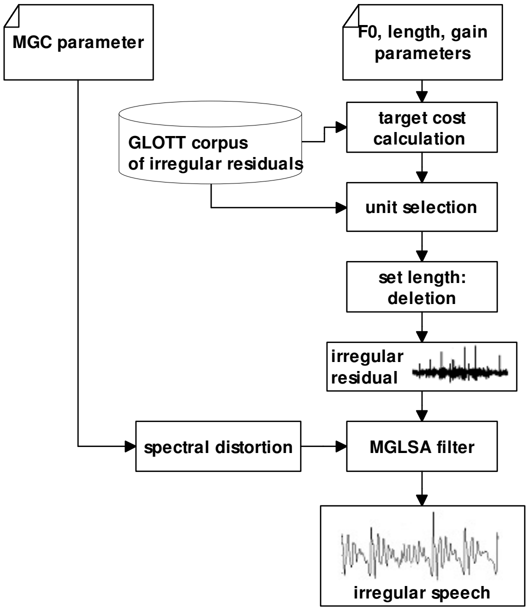

Fig. 5. *Irregular voice synthesis in the proposed #2 (data-driven) system.*

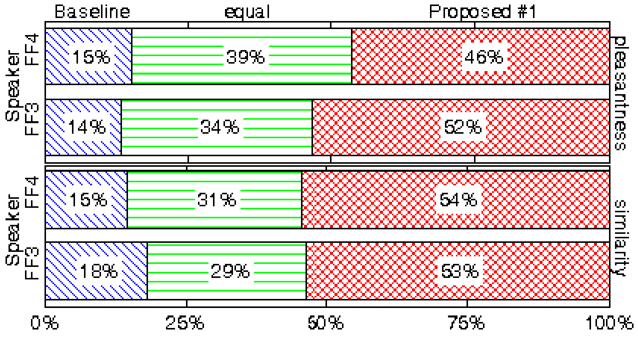

Fig. 6. *Results of the 1st subjective evaluation showing preference percentages between baseline and Proposed #1 systems. The Proposed #1 system was preferred over the baseline system for both speakers in both questions.*

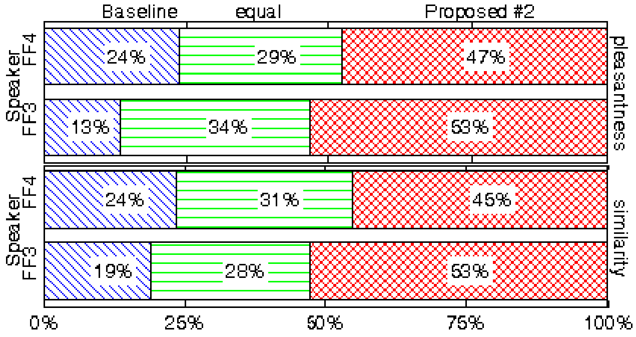

Fig. 7. *Results of the 2nd subjective evaluation showing preference percentages between baseline and Proposed #2 systems. The Proposed #2 system was preferred over the baseline system for both speakers in both questions.*

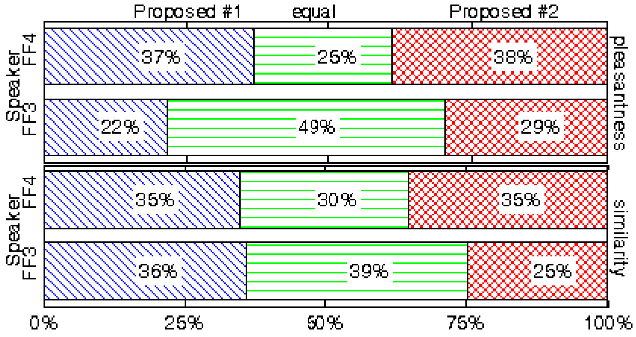

Fig. 8. *Results of the 2nd subjective evaluation showing preference percentages between Proposed #1 and Proposed #2 systems. The systems were found to produce irregular voice having roughly equal quality.*

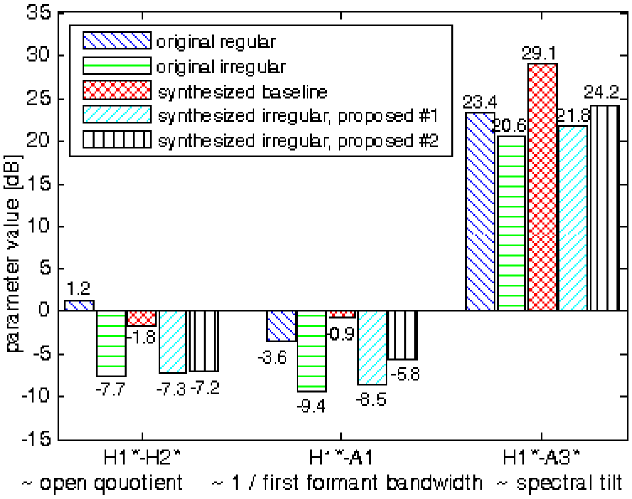

Fig. 9. *Results of the acoustic experiment. Synthesized irregular versions of the words are close to original irregular utterances in terms of the first two acoustic cues.*