ISCA Archive
http://www.isca-speech.org/archive
8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

A novel irregular voice model for HMM-based speech synthesis

Tamás Gábor Csapó, Géza Németh

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
{csapot,nemeth}@tmit.bme.hu

Abstract

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. HMM-TTS is optimized for ideal voices and may not produce high quality synthesized speech with voices having frequent non-ideal phonation. Such a voice quality is irregular phonation (also called as glottalization), which occurs frequently among healthy speakers. There are existing methods for transforming regular (also called as modal) to irregular voice, but only initial experiments have been conducted for statistical parametric speech synthesis with a glottalization model. In this paper we extend our previous residual codebook based excitation model with irregular voice modeling. The proposed model applies three heuristics, which were proven to be useful: 1) pitch halving, 2) pitch-synchronous residual modulation with periods multiplied by random scaling factors and 3) spectral distortion. In a perception test the extended HMM-TTS produced speech that is more similar to the original speaker than the baseline system. An acoustic experiment found the output of the model to be similar to original irregular speech in terms of several parameters. Applications of the model may include expressive statistical parametric speech synthesis and the creation of personalized voices.

Index Terms: irregular phonation, glottalization, voice quality, parametric, speech synthesis

1. 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). In this type of speech synthesis, the speech signal is decomposed to physical 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 speech coding methods. For this task, often simple vocoders (e.g. pulse-noise excitation) are used which make use of the source-filter model of speech. The advantages of HMM-TTS compared to other synthesis techniques include its flexibility and small footprint.

However, the over-simplified vocoder techniques make the quality of synthesized speech of HMM-TTS poor compared to high-quality unit selection based speech synthesis systems. To overcome this drawback, several improved excitation models have been proposed. STRAIGHT-based vocoding produces very good quality HMM-based synthesized speech [2]. Cabral uses the Liljencrants-Fant (LF) [3] acoustic model of the glottal source derivative to construct the excitation signal [4]. Drugman proposed the Deterministic Plus Stochastic Model (DSM) of the residual signal [5]. Raitio and his colleagues use glottal inverse filtering within HMM-based speech synthesis and unit selection of pulses for generating natural sounding synthetic speech [6], [7]. The latest excitation models introduce the voicing cut-off frequency [8] and waveform interpolation [9] to enhance the performance of HMM-TTS. We proposed a residual codebook based excitation model which also exceeds the quality of simple pulse-noise excitation [10], [11].

1.1. Irregular phonation

Statistical parametric speech synthesis and most of the above excitation models are optimized for ideal voices and may not produce high quality synthesized speech with voices having frequent non-ideal phonation. Such a non-ideal voice quality is irregular phonation.

During regular voiced phonation (ideal, modal voice) in human speech, the vocal cords are vibrating quasi-periodically. For shorter or longer periods of time this vibration may become irregular. Abrupt changes occur in the fundamental frequency (F0), amplitude of the pitch periods or both. This is called irregular phonation (or glottalization, vocal fry, creaky voice), which is a frequent phenomenon for 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 [12], [13]. Fig. 1 shows an example for glottalization (LP residual on the top and speech signal on the bottom). The horizontal arrow denotes the section where the phonation is irregular. Amplitude attenuations in the waveform and missing impulses in the residual are clearly visible.

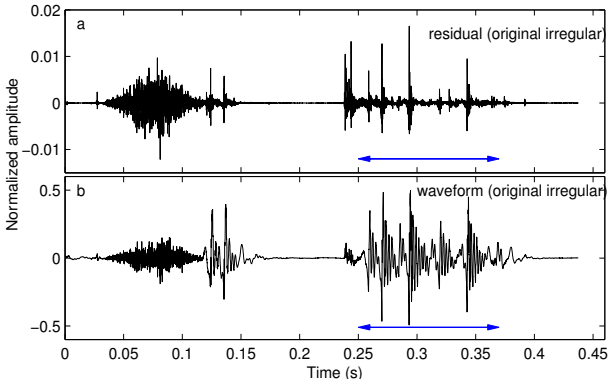

Figure 1: *A speech recording of the word 'cipő' having irregular phonation at the section denoted by an arrow. a) residual signal and b) speech signal.*

It was found that up to 15% of the vowels of healthy American English speakers may be produced with irregular phonation [14]; therefore it is not negligible in normal speech. The occurrence of glottalization depends on the prosodic structure (it often coincides with prosodic boundaries and stressed syllables [15]) and carries information from the speaker, his/her dialect, mood and emotional state [16]. Irregular phonation can cause problems for standard speech analysis methods (e.g. F0 tracking and spectral analysis). Proper modeling of irregularly phonated speech may

contribute to building natural, emotional and personalized speech synthesis systems. Irregular phonation is frequently adopted in lively story-telling, natural interactive conversation [17] and can signal sadness [18] or boredom [19]. Therefore an irregular phonation model improves expressive speech synthesis systems. Such a model allows speaker adaptation for deep elderly voices (e.g. radio announcers) having frequent glottalization.

First attempts to model irregular phonation were either in the formant synthesis domain [20] or relied on increasing jitter and shimmer of the speech signal [21]. In [13], a simple semi-automatic transformation method is developed which introduces irregular pitch periods into a modal speech signal, based on amplitude scaling of the individual periods. In perception and acoustic experiments, this method was shown to yield irregular speech that is as rough and as natural as original glottalized speech. To model vocal fry in statistical parametric speech synthesis, [22] introduces a robust F0 measure and two-band voicing, which improves significantly the quality of HMM-based speech synthesis. However, they do not focus on the characteristics of creaky excitation. Drugman and his colleagues derive an extension of the DSM model [5] which can handle creaky excitation by integrating secondary pulses in the residual, and investigate this in copy-synthesis experiments [23]. After that they investigate the usefulness of contextual factors for creaky voice prediction and experiment with adding parameter streams describing irregular phonation into the HMM-TTS framework [17]. To the best of our knowledge this extended analysis-synthesis method with the creaky voice model has not been integrated into HTS yet.

In this paper we extend our previous residual codebook based excitation model (HTS-CDBK) with irregular voice modeling. The baseline residual analysis-synthesis framework and the model of irregular voice are introduced in Sections 2 and 3, respectively. In Section 4 a perceptual test, while in Section 5 an acoustic experiment and their results are shown. In Section 6, we present the advantages and drawbacks of our method and conclude the paper.

2. HMM-TTS with a residual codebook based excitation model

We have proposed a residual codebook based excitation model [10] and integrated it into HMM-TTS ([11], HTS-CDBK), that will be used here as the baseline system.

2.1. Analysis

The input is a speech waveform with 16 kHz sampling rate and 16 bit linear PCM quantization. First, the F0 parameters are calculated by the publicly available Snack pitch tracker with 25 ms frame size and 5 ms frame shift. In the next step 34-dimensional MGC analysis is performed on the speech signal with the SPTK tool. The residual signal (excitation) is obtained by MGLSA inverse filtering. Next, a Glottal Closure Instant (GCI) detection algorithm is used to find the pitch boundaries in the voiced parts of the modal speech signal [24]. Finally, a codebook of pitch-synchronous residuals is built, obtained from a small speech database (see Section 2.4) and residual analysis is performed.

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 Hanning-windowed (see Fig. 2). 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: RMS energy of the windowed frame
- rt0 peak indices: the locations of prominent values (peaks or valleys) in the windowed frame (see Fig. 2)
- HNR: Harmonic-To-Noise ratio of the frame [25]

For each voiced frame, one codebook element is saved with the above parameters and the windowed signal is also stored. The rt0 parameter is a 4-dimensional vector, which is a new idea for describing the residual frames. We found that the consecutive rt0 parameters are slowly evolving enough and are suitable for machine learning in HTS. In the used parameters our model is different from similar excitation models, like DSM [5]. These parameters will be used for target cost calculations during synthesis. In order to collect similar codebook elements, the RMSE distance 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.

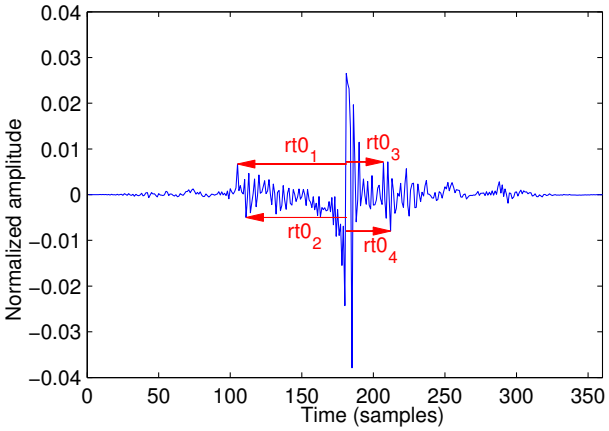

Figure 2: *Calculation of the rt0 parameter for a windowed residual segment. $rt0_i$ is the distance of prominent peaks from the main impulse, in samples.*

2.2. Training

For training, the parameters of MGC, log(F0), log(gain), log(rt0) and log(HNR) of each frame are extracted. 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.

2.3. Synthesis

In the synthesis phase of HTS-CDBK the inputs are the parameters obtained during training (F0, gain, rt0 indices and HNR) and the codebook of pitch-synchronous residuals. If the frame is voiced, a suitable codebook 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 [26]. 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 is calculated as the RMSE difference 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 Hanning-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 and synthesized speech is reconstructed by MGLSA filtering.

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.

2.4. Speech data

The speech data that was used for our experiments is a part of the PPBA database [27]. Two Hungarian males were chosen for speaker dependent training (denoted FF3 and FF4). Both speakers produced irregular phonation frequently, mostly at the end of sentences. 1940-1940 phonetically balanced sentences (2-2 hours of speech) from them 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. Other excitation models use codebooks of similar scale [7].

2.5. 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 sentences), the Snack pitch tracker cannot measure 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. During synthesis unvoiced excitation is often generated at the last vowels of the sentences. This produces a very unpleasant voice and it 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 showing 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', and therefore there is only aperiodic noise in the end of the speech signal.

3. HTS-CDBK extended with an irregular voice model

First, several acoustic properties of glottalization are introduced. Then an available semi-automatic regular-to-irregular transformation method is described. Finally, this method is further improved and integrated into HTS-CDBK. The novel system is denoted as HTS-CDBK+Irreg-Rule.

3.1. Acoustic properties of irregular phonation

In natural speech, irregular phonation can be distinguished from regular phonation by several properties ([13], [20]):

- the overall intensity level is lower
- the time that is elapsed between successive glottal pulses is longer and more irregular, resulting in lower F0 and higher jitter
- abrupt changes occur in the amplitude of the 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

Some of these properties are observable in both the speech and in the residual signal. An example for this can be seen in Fig. 1. In the irregularly phonated interval the pitch is lower and the periods have clearly abrupt changes in amplitude.

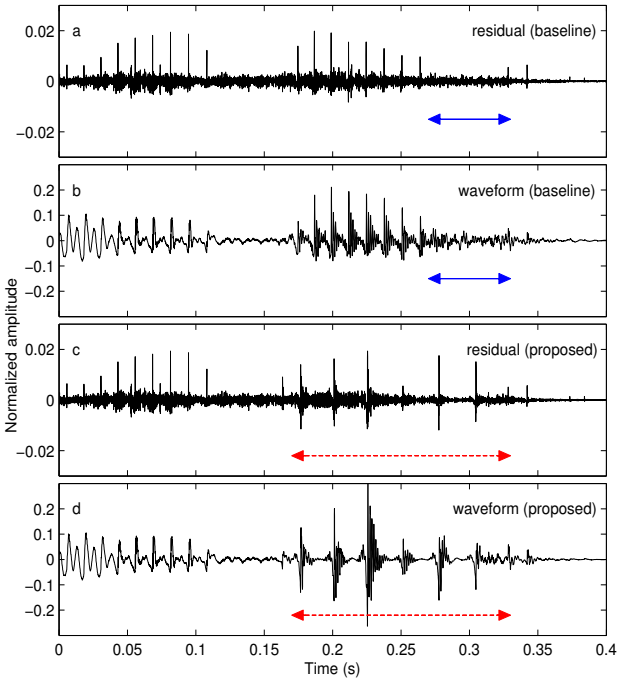

Figure 3: *Synthesized version of the word 'mihái' extracted from the end of a longer sentence with a) and b) from the baseline system and c) and d) from the proposed system.*

3.2. Regular to irregular transformation method

In [13], a regular-to-irregular voice transformation method was proposed which uses amplitude scaling of individual glottal cycles. Here, the modal 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. [13] extends this with stylized pulse pattern copying yielding in a semi-automatic transformation method.

In the present form, this method is not suitable to be integrated into 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 model that was integrated into HTS-CDBK.

3.3. The proposed model

The proposed model differs from the baseline only in the synthesis phase. The analysis, training and the training speech

database are the same as in the baseline system (see Sections 2.1, 2.2 and 2.4, respectively).

The proposed model applies three heuristics similarly to [13]: 1) pitch halving, 2) pitch-synchronous residual modulation with periods multiplied by random amplitudes and 3) spectral distortion. Although the theoretical correctness of these heuristics cannot be proven because irregular phonation does not have a strict definition and each occurrence is different, in our preliminary experiments these ideas were useful and improved the baseline system. All of the heuristics are motivated by acoustic properties of irregular phonation, which are described in detail here:

1) In the sections that should be synthesized with irregular phonation, the half of the F0 of the generated parameter sequence is used. If there is F0=0 in the parameters of the glottalized section as in the baseline system, than before the halving the F0 is first interpolated according the neighboring frames. We applied the pitch halving because glottalization has often significantly lower F0 than modal speech (see Section 3.1), and [13] 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.

2) During residual synthesis, each pitch cycle is multiplied by a random scaling factor in the range of {0..1}. This is similar to [13] but we do not boost any of the periods, only attenuate or leave them unchanged. This heuristic is motivated by the property of glottalization that is visible in Fig. 1: irregular phonation has often strong amplitude attenuations during the consecutive pitch cycles. From the modified residual periods the residual signal is obtained by overlap-adding the frames.

3) Finally, spectral distortion is applied. In [28] we found that the frame-by-frame 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 between {0.995…1.005}. This yields a less smooth parameter sequence for each dimension of MGC. Note that one might argue that by adding random numbers to the residual or waveform samples itself the speech signal could be directly distorted. However, there is only a small chance that such a distortion would lead to a speech signal that is similar to original irregular utterances.

As there is no explicit glottalization model (e.g. irregular phonation labels, questions for decision trees) in the HTS-CDBK system, sections with irregular phonation have to be found from the generated F0 sequence. In our experiments the generated parameter and label files were checked automatically. Glottalization was applied if at least five consecutive frames were given zero F0 within a vowel. In these cases, fundamental frequency was interpolated between the voiced parts to have a straight F0 line, or was set to slightly decreasing if there were no voiced neighboring sounds.

Fig. 3 shows an example for the results of the baseline (HTS-CDBK: a, b) and the extended 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' in HTS-CDBK. As this section is longer than five frame shifts (25 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 red dashed horizontal line indicates the glottalized vowel 'á'. It is clearly visible on both the residual and speech signals that the extended model is closer to the original irregular signal (Fig. 1) than the baseline system.

4. Perceptual evaluation

In order to evaluate the quality that can be achieved by the proposed HTS-CDBK+Irreg-Rule method, a listening test was conducted according to the guidelines of [29]. A major factor that determines the usefulness of this method is if human listeners accept the synthesized speech. Therefore, our aim was to measure the perceived 'pleasantness' and the similarity to the original speaker. Synthesized samples of the baseline system were compared to those of the proposed solution.

4.1. Methods

To create the speech stimuli, four voice models with the two systems and the two speakers were created. Note that HTS-CDBK and HTS-CDBK+Irreg-Rule only differ in the synthesis part, therefore the analysis, training and speech data was the same here. 130-130 sentences 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 (containing 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 an utterance from the test can be seen in Fig. 3.

In the test, the two versions of each word were included, resulting altogether 40 utterances (2 speakers · 10 words · 2 versions). A web-based paired comparison test with two CMOS-like questions was created. 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, they were asked which version is more similar to the original speaker. For this, a reference 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.

4.2. Results

Altogether 11 listeners participated in the test. They were all university students or computer science professionals, between ages of 19-30 years. All of them were native speakers of Hungarian and none of them reported any hearing loss. On the average the whole test took 9 minutes to complete.

The results of the listening test are presented in Fig. 4 for the two speakers. The figure provides a comparison between the baseline system (left part, blue color) and the proposed system (right part, red color). 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 altogether: 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 ranked higher in the similarity test, while in 18 out of 20 sample pairs the extended model was preferred.

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

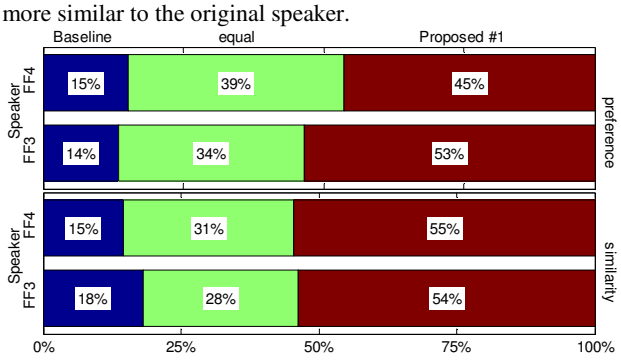

Figure 4: *Results of the subjective evaluation showing percentages of Comparative MOS scores between baseline and proposed systems.*

5. Acoustic evaluation

The perception test showed the preference of the proposed model. However, from the listening test results it is not known whether the proposed system models irregular voice properly or it was just preferred to use other excitation instead of white noise in the investigated vowels. Therefore we investigated several acoustic cues which were found previously to distinguish original irregular and regular speech [13].

5.1. Methods

The acoustic properties of glottalization were introduced in Section 3.1. In the acoustic experiment the three most important acoustic cues [20] 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.

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 [31, 32]. These parameter values can be biased by the effects of the first three formants. To compensate this, we used the equations suggested by [30] and implemented in VoiceSauce: the value of H1 and H2 was corrected for F1 and F2 (H1* and H2*), and the value of A3 was corrected for F1, F2, and F3 (A3*).

The measurements were conducted partly on the stimuli used in the perceptual evaluation (10-10 words synthesized by the proposed model). 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. Altogether the parameters of 80 vowels were measured. First the wave files were resampled to 8 kHz. Then a glottalized vowel from the original irregular version was selected and the middle of the vowel (roughly aligned with the pitch marks) was chosen and the same vowel was measured in the original regular version. In the synthesized versions, the vowels created by the irregular voice models were measured. In Wavesurfer, 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.

5.2. Results

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. 5 for the three utterance versions separately. In one-way ANOVAs, stimulus type had a significant effect on the difference between the first two harmonics ($p<0.0005$), while the other two calculated parameters were not significantly different. Tukey's post hoc test was used to compare the mean parameter values of each stimulus type.

H1*-H2* was significantly different for the original regular speech ($p<0.05$) whereas it was approximately the same for the original irregular and for the synthesized irregular recordings ($p=0.97$, n.s.). This means that in terms of open quotient, the synthesized versions are close to the original irregular versions. H1*-A1 and H1*-A3* are not significantly different for any of the groups, but in the figure we can see the trends that the irregular voice model have created. In terms of the H1*-A1 and first formant bandwidth the synthesized irregular utterances are close to the original irregular recordings. 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 model can reconstruct two of the three investigated acoustical correlates of irregular speech.

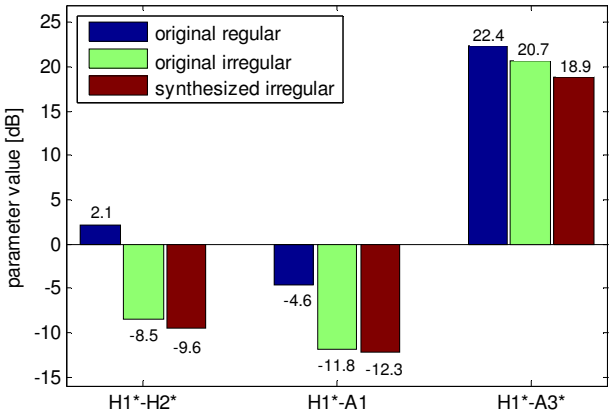

Figure 5: *Results of the acoustic experiment.*

6. Discussion and Conclusions

This paper presented a method to synthesize irregular voice within the HTS framework. The proposed method uses pitch halving, amplitude scaling of the pitch periods of the residual signal and spectral distortion. Although the theoretical correctness of these heuristics cannot not be proven because irregular phonation does not have a strict definition and every occurrence is different, in our experiments these ideas were useful and improved the baseline system. The proposed method was supported by perception and acoustic tests. A perception experiment found the proposed method to synthesize glottalized speech that is closer to the original speaker while increasing naturalness. An acoustic experiment

found the output of the model to be similar to original irregular speech in terms of open quotient and first formant bandwidth.

The new method is fully automatic because amplitude scales are determined randomly and no manual scaling is necessary. By applying predefined stylized pulse patterns as in [13] instead of random scaling factors, the naturalness of synthesized glottalization might be further improved. With the application of an irregular vs. regular classification algorithm (e.g. [14]), glottalization could be modeled explicitly in HTS. To create a full speech synthesis system that is able to synthesize irregular speech, it will be necessary to include new contextual factors or additional parameter streams like in [17]. In [33] we extend this model and show another data-driven approach for irregular voice synthesis.

With the new method we extend previous speech processing techniques dealing with irregular phonation: it may contribute to building natural, emotional and personalized speech synthesis. Irregular phonation is frequently adopted in lively story-telling, natural interactive conversation [17] and can signal sadness [18] or boredom [19]. Therefore an irregular voice model improves expressive speech synthesis systems. For example it is possible to create speaker adaptation for deep elderly voices (e.g. those of famous radio announcers) having frequent glottalization.

7. Acknowledgements

We would like to thank the listeners for participating in the subjective test. This research was 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.

8. 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, 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.

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

[4] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis," in Proc. ISCA SSW6, 2007, pp. 113–118.

[5] 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.

[6] 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.

[7] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach," in Blizzard Challenge 2012, 2012.

[8] 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.

[9] 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.

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

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

[12] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," JASA, vol. 103, pp. 2649–2658, 1998.

[13] 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.

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

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

[16] 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.

[17] T. Drugman, J. Kane, T. Raitio, and C. Gobl, "Prediction of Creaky Voice from Contextual Factors," in Proc. ICASSP, 2013.

[18] 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.

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

[20] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers.," JASA, vol. 87, no. 2, pp. 820–857, 1990.

[21] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech and Audio Processing, vol. 3, no. 4, pp. 242–250, 1995.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[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], 2013.

[28] 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.

[29] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical Analysis of the Blizzard Challenge 2007 Listening Test Results," in Blizzard Challenge 2007, 2007.

[30] 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.

[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," J.Speech Hear.Res, vol. 38, no. 6, pp. 1212–1223, 1995.

[32] N. Henrich, C. d'Alessandro, B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data", Proc. Eurospeech 2001, pp. 47–50.

[33] T. G. Csapó, G. Németh, „Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation," IEEE Journal on Selected Topics in Signal Processing, submitted, 2013.