Automatic transformation of irregular to regular voice by residual analysis and 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

This paper presents an automatic speech transformation method of non-ideal phonation of speech (irregular or creaky voice). The irregular-to-regular transformation is performed by analyzing and resynthesizing the residual. A recent continuous pitch estimation algorithm is used for interpolating F0 in regions of irregular voice. The linear prediction residual of irregular sections of speech is replaced by overlap-added frames from a codebook of pitch-synchronous residuals. Finally, speech is reconstructed from the residual. A listening experiment showed that by transforming natural speech samples containing irregular voice, the perceived roughness of the transformed speech is decreased.

Index Terms: irregular phonation, voice transformation, modal voice, creaky voice, glottalization, voice quality

1. Introduction

During ideal voiced phonation in human speech, the vocal cords are vibrating quasi-periodically, producing regular (also called as modal) speech. For shorter or longer periods of time instability may occur in the larynx causing irregular vibration of the vocal folds. It results in abrupt changes in the length and/or amplitude of the pitch periods. This is called glottalization (or irregular phonation, creaky voice), which is a frequent phenomenon in the speech of 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 [1], [2]. Figure 1 shows an example for the difference between irregular (1a) and regular (1c) speech. Amplitude attenuations in the (1a) waveform are clearly visible in the section denoted by the horizontal arrow.

It was found that up to 15% of the vowels of healthy American English speakers are produced with irregular phonation [3]; therefore it is not negligible in natural speech. The occurrence of glottalization depends on the prosodic structure (it often coincides with prosodic boundaries and stressed syllables [4] or occurs at vowel-vowel transitions) and carries information from the speaker, his/her dialect, mood and emotional state [5]. Irregular phonation can cause problems for standard speech analysis methods (e.g. F0 tracking and spectral analysis) and it is often disturbing in speech technologies – e.g. glottalized sections of speech in statistical parametric speech synthesis might cause false voiced/unvoiced decision which decreases the overall quality [6]. Acoustic modeling of glottalization can improve automatic speech recognition [7]. In such cases the transformation of irregular sections of speech to modal voice might be useful. Proper manipulation of irregularly phonated speech may contribute to building natural, emotional and personalized speech synthesis systems.

There have been a couple of studies dealing with glottalization in speech analysis, voice conversion and speech synthesis. There are existing methods for the automatic detection of irregular phonation [3], [8]–[11], for transforming modal voice to irregular [2], [12], [13], for analysis and synthesis of glottalization [7], and for speech synthesis with creaky voice [6], [14], [15]. However, there is a lack of methods to transform glottalized speech to modal speech by preserving the characteristics of the original speaker. In [16], we introduced a semi-automatic irregular-to-regular transformation algorithm which is based on a vocoder [17]. The study relied on manual irregular voice labels and simple linear interpolation of the F0 contour. In cases when the linear interpolation did not give suitable results, manual correction of the F0 curve was also necessary before resynthesis.

In this paper we show our work on the automatic transformation of irregular phonation to modal voice. We extend our semi-automatic method [16] with the use of a recent continuous pitch estimation algorithm of Garner et al. [18] and automatic creaky voice detection of Drugman et al. [11]. The residual analysis-synthesis framework and the method of the voice transformation are introduced in Section 2. In Section 3, a listening test and its results are shown. In Section 4, we present the advantages and drawbacks of our method and conclude the paper.

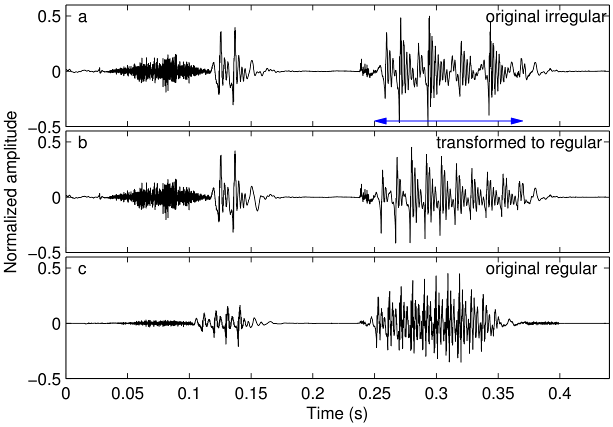

Figure 1. *Speech waveforms of the word /tsipøː/ with a) original irregular ending (the horizontal arrow shows the section where phonation is irregular), b) its transformed version to regular, c) another realization of the same word with original regular ending.*

2. Transformation method

For the transformation method, we use the extended version of the analysis-synthesis excitation model (vocoder) introduced in [17]. This model has been created for the purposes of statistical parametric speech synthesis, but it is suitable for speech parametrization and reconstruction as well. Similar methods have been successfully used for creaky speech synthesis and pitch transposition [14], [19].

2.1. Residual analysis and codebook of residuals

The input of the analysis part of the transformation method is a speech waveform with 16 kHz sampling rate and 16 bit linear PCM quantization. First, the F0 parameter is calculated by a recent continuous pitch tracker [18]. In this paper we refer to this method as CONT_F0. The CONT_F0 algorithm has the advantage that it results in a pitch estimate for all frames and it does not give a strict voiced/unvoiced decision. This property is especially useful in sections of speech phonated with irregular voice, as standard F0 estimation algorithms (e.g. RAPT, [20]) tend not to measure F0 in regions of creaky voice. An example for the effect of irregular voice is shown in Figure 2. The figure shows a) a short speech segment with red dashed line showing the regions of irregular voice; b) its estimated pitch track by Snack using the RAPT algorithm; and c) its estimated pitch track by the CONT_F0 pitch tracker. In some regions of creaky voice (see the blue horizontal arrows of b) the standard RAPT pitch tracking cannot estimate F0 because of the abrupt changes of the pulse amplitudes, whereas the CONT_F0 pitch tracking interpolates the F0 contour in the case of irregular voice as well.

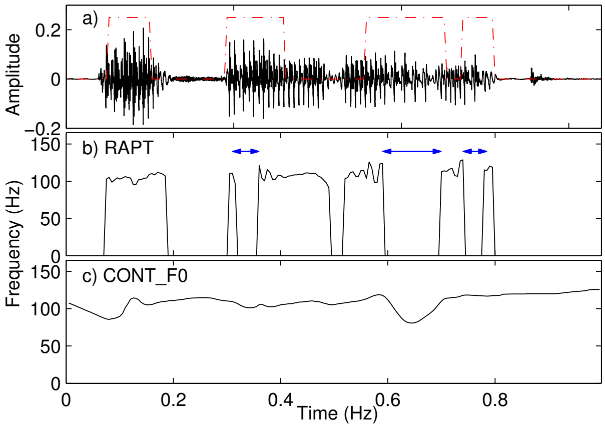

Figure 2. *Effect of creaky voice on pitch estimation. a) a sample speech waveform and regions of irregular voice (red dashed line); b) pitch estimated by RAPT* [20] *(blue arrows indicate inaccurate pitch estimation); and c) pitch estimated by the CONT_F0 pitch tracker* [18].

For pitch tracking, 25 ms frame size and 5 ms frame shift are used. After that, 34-dimensional MGC (Mel-Generalized Cepstral) analysis is performed on the speech signal with the SPTK toolkit [21]. The residual signal (excitation) is obtained by inverse filtering with the MGLSA (Mel-Generalized Log Spectral Approximation) method [22]. Next, a Glottal Closure Instant (GCI) detection algorithm is used to find the glottal period boundaries in the voiced parts of the modal speech signal [23]. Finally, a codebook of pitch-synchronous residuals is built from modal speech, obtained from a small speech database.

The further analysis steps are completed on the residual signal with the same frame shift values. For measuring the parameters in the modal voiced parts, pitch synchronous, two period long frames are used according to the GCI locations and they are Hanning-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: RMS energy of the frame,
- rt0 peak indices: the locations of prominent values in the windowed frame (see Figure 3),
- HNR: Harmonics-To-Noise ratio of the frame [24].

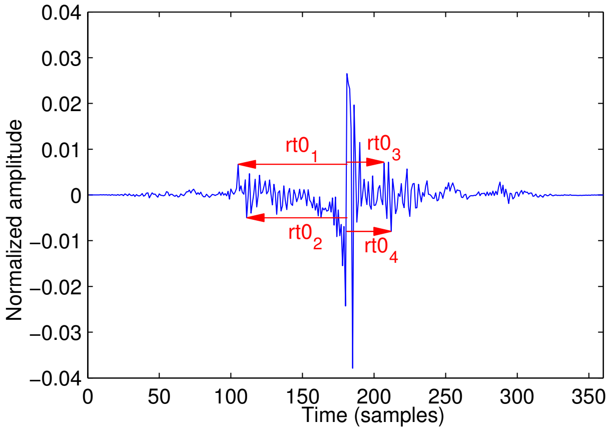

Figure 3. *Calculation of the rt0 parameter for a windowed residual segment. $rt0_i$ is the distance of prominent peaks from the main impulse, in samples.*

The rt0 parameter is a 4-dimensional vector, which is helpful for parameterizing the residual frames. The calculation of the parameter is shown in Figure 3: 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 main excitation exceeding 10% the length of the pitch period.

For each voiced frame, one codebook element is saved with the given parameters and the windowed signal is also stored. These parameters will be used for target cost calculations during the reconstruction of the transformed speech regions. In order to collect similar codebook elements, the RMSE (Root Mean Squared Error) distance is calculated between the pitch normalized versions of the elements. The normalization is done by resampling every frame to 40 samples. Previously we concluded that a codebook consisting of about 6500 residual frames is enough for high quality speech synthesis [17]. We used the same codebooks here.

During the analysis step of the irregular-to-regular transformation, the same processing is done as during the codebook building, but the parameters are only used in the parts labeled as irregular. When analyzing the speech to be transformed, an automatic creaky voice detection algorithm [11] is used to label the irregular voice sections.

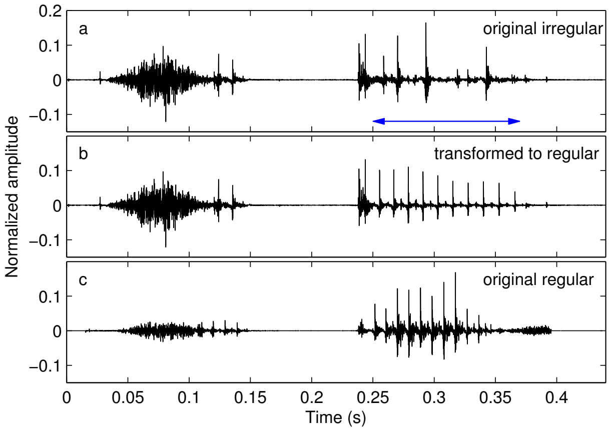

Figure 4. *Residuals of speech recordings of Figure 1: a) residual with original irregular ending (horizontal arrow shows the section where phonation is irregular), b) its transformed version to regular, c) is the residual of another realization of the same word with original regular ending.*

2.2. Residual correction and synthesis

In the regions labeled as irregular, the residual signal is replaced by a periodic shape similar to modal regions of the residual. Figure 4a) and c) show the residuals of speech samples from Figure 1a) and c) obtained by MGLSA inverse filtering. In the section denoted by an arrow in Figure 4a) the pulses are clearly irregular, while on Figure 4c) the pulses are more regular.

In regions of creaky voice, the CONT_F0 pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing [18]. This way the transformation method is fully automatic and there is no need for semi-automatic F0 interpolation or manual correction of the F0 contour as in our previous study [16].

After the residual has been analyzed, the frame-by-frame gain parameters and MGC coefficients are smoothed. Irregular phonation causes small perturbations in the frame-by-frame gain and MGC values, partly because of the abrupt changes in the amplitude of the pitch periods (see Figure 4a). By using a 5-point moving average, the smoothing was found to be good enough during resynthesis. Spectral parameters are represented as MGC values which are suitable for such smoothing.

In the synthesis stage of the transformation the inputs are the interpolated / smoothed parameters obtained during analysis and the codebook of residuals. If the frame is voiced, a suitable codebook element with the target F0, rt0 and HNR is searched from the codebook. For the selection of the residual frames, we apply target cost and concatenation cost with hand-crafted weights, similarly to unit selection speech synthesis. 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 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. Next, the residual is created by pitch synchronously overlap-adding the Hann-windowed residual frames. Finally, the energy of the frame is set using the smoothed gain parameter.

In the regions labeled as regular, the original residual is preserved. Original and transformed regions of the residual signal are concatenated. Resynthesized speech is obtained from the residual with MGLSA filtering. Figure 4b) shows an example for the transformed residual, while Figure 1b) shows the speech signal that is the result of the irregular-to-regular transformation method. It can be seen in both Figure 1 and Figure 4 that the b) ‘transformed to regular’ and c) ‘original regular’ signals have similar regular pitch periods, while a) ’original irregular’ is very different and has amplitude attenuations.

3. Perceptual evaluation

In order to evaluate the quality that can be achieved by the proposed transformation method, we have conducted a listening test. A major factor that determines the usefulness of this method is if human listeners accept the transformed speech. Therefore, our aim was to measure the perceived ‘roughness’ (whether phonation is irregular or not) and the naturalness of several utterances that were transformed by the new method. Naturalness was also investigated here because the method might produce unwanted ‘buzzy’ regions which is the result of using a vocoder. We compared the results of the irregular-to-modal transformed speech samples to unmanipulated natural sentences.

3.1. Methods

For the evaluation, four Hungarian speakers (three males: FF1, FF3, FF4 and one female: NO3) were selected from the PPSD (Parallel Precision Speech Database) database [25]. All of them produced irregular phonation frequently, mostly at the end of sentences. Table 1 shows the ratio of the voiced frames produced with irregular phonation vs. all voiced frames for these speakers. The ratio was measured on the same 180 sentences from each speaker, based on automatic creaky voice detection [11].

Table 1. *Ratio of irregular voice usage for the four speakers of the perceptual evaluation.*

speaker	no. of voiced frames	ratio of irregular frames
FF1	58 912	11.83%
FF3	59 496	5.48%
FF4	69 051	8.38%
NO3	68 109	6.34%

To create the speech stimuli of the perceptual test, five sentences were selected from each speaker, containing irregular voice in at least 15% of the voiced frames. The utterance versions having irregular sections were transformed to modal voice by the proposed method. In some of the cases only one word was produced with irregular phonation, while in other realizations a longer section of the utterance was glottalized.

In the test, both versions of each sentence (original irregular and transformed to regular) were included, resulting in altogether 40 utterances (4 speakers * 5 sentences * 2 versions). These utterances are attached to the submission as multimedia files. We created a web-based test with two 5 point MOS-like questions. Before the test, the subjects were asked to listen to an example of glottalized speech to clarify the meaning of the term ‘rough’. After listening to each utterance, the listeners had to rate the roughness (‘1 – very rough’ … ‘5 – not rough at all’) and naturalness (‘1 – very unnatural’ … ‘5 – very natural’) of the stimuli. The sentences and the versions in the pairs were presented in randomized order (different for each participant).

Altogether 13 listeners participated in the test (11 males, 2 females). They were university students or speech technology experts with a mean age of 25 years. All of them were native speakers of Hungarian and none of them reported any hearing loss. On average the whole test took 8 minutes to complete.

3.2. Results

The ratings of the subjects were compared by paired t-tests. The analysis showed that the difference between the scores of ‘original irregular’ and ‘transformed to regular’ samples is significant for roughness ($p < 0.05$) when analyzing all of the data together. By investigating the MOS scores we can conclude that the roughness of the original samples was significantly decreased. Altogether, the method significantly ($p < 0.05$) decreased the naturalness during transformation.

Table 2. *Speaker by speaker means and standard deviations (in parenthesis) for the roughness and naturalness questions of the subjective evaluation.*

	roughness		naturalness	
speaker	original	transf.	original	transf.
FF1	2.77 (0.93)	2.92 (1.24)	3.71 (1.04)	2.49 (1.03)
FF3	2.80 (1.28)	2.89 (1.24)	3.94 (1.03)	3.02 (1.08)
FF4	2.89 (1.05)	3.26 (1.03)	3.94 (0.88)	3.26 (1.11)
NO3	3.71 (1.13)	3.69 (1.18)	3.88 (0.93)	2.80 (1.20)

The speaker by speaker results of the listening test are shown in Table 2. For speakers FF1, FF3 and FF4, the method was able to decrease the perceived roughness (higher number in the table means less rough voice, according to the possible answers for the first question of the test). However, this difference is only significant for speaker FF4. For speaker NO3, the transformation slightly increased the roughness. According to Table 1, speakers FF3 and NO3 use relatively few creaky voice, therefore it might have happened that the listeners did not perceive strong differences in the roughness of these sentences. When listening to the original samples of speaker FF4, one can observe strong and clear creaky voice, resulting from extremely low fundamental frequency.

In Table 2, we can observe differences among speakers in the success of the transformation method in terms of keeping the naturalness. For speaker FF4, the method has not significantly modified it. For the remaining three speakers (FF1, FF3 and NO3), the naturalness scores of the utterances were significantly decreased. This can be explained by three possible reasons. 1) The CONT_F0 method interpolates F0 for all frames. If there is no modal voice around the section of creaky voice, this interpolation is often inaccurate, e.g. relatively high F0 values are predicted for the end of the sentence. This contradicts to the natural F0 contour of the declarative sentences where F0 decreases and has the lowest F0 values at the end of the sentence. 2) Using a vocoder for resynthesis might cause ‘buzzy’ voiced quality. This can be an audible artefact depending on the length of the resynthesized section. 3) The 34th order of the MGC analysis may be too high for the female speaker.

From this subjective experiment we can conclude that the method can decrease the perceived roughness of irregular utterances. This is especially true for speaker FF4, whose glottalization is clearly audible in natural speech.

4. Discussion and Conclusions

We presented a fully automatic method to transform irregular voice to regular voice by analyzing and resynthesizing the residual. For this, we used a codebook-based residual analysis-synthesis method. Using such an analysis method is more suitable than direct waveform manipulation like PSOLA with hand-crafted weights, as the residual could be corrected automatically. During the transformation the original irregular residual sections were replaced by overlap-added frames from the codebook, yielding a regular residual. Using residual frames originating from different sounds in the concatenation is appropriate, as similar solutions have been successfully applied in other models [14].

The current version of the new method is fully automatic because of the automatic creaky voice detection [11] and the CONT_F0 pitch tracker which involves interpolation of the F0 contour [18]. The Kalman smoothing of CONT_F0 was found to be more suitable than the simple linear F0 interpolation we used in [16]. However, one drawback of this interpolation is that in some cases it causes high F0 at the end of the sentence, which is unnatural when the utterance is resynthesized. A solution might be to combine the F0 interpolation with a rule-based intonation model [26].

In the listening test we found that the transformation method could significantly decrease the roughness in one of the four speakers, whereas the difference for the remaining three speakers was not significant. It is known that several types of glottalization can be differentiated and the occurrence of these types is speaker dependent [11], [27]. We found that our method was suitable for transforming the creaky voice type used by speaker FF4. The naturalness of the transformed samples can be improved by using a different vocoder, e.g. using Maximum Voiced Frequency [28] to decrease buzziness.

With this new voice transformation method we fill a gap in speech processing techniques dealing with irregular phonation. Applications of the model may include the correction of voices where unwanted irregular phonation occurs frequently (e.g. those of radio announcers or voice actors). The method may be used to transform glottalized parts of large speech databases in order to help further automatic speech processing and to create better synthetic voices from those databases. The transformation method might be useful for voice conversion, i.e. to transform the speech of one individual to sound similar to someone else.

5. Acknowledgements

We would like to thank the listeners for participating in the subjective test and to Philip N. Garner for providing the continuous pitch tracker open source. This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF n° IZ73Z0_152495-1).

6. References

[1] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, “Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers,” *J. Acoust. Soc. Am.*, vol. 103, no. 5, pp. 2649–2658, May 1998.

[2] 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.

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

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

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

[6] T. G. Csapó and G. Németh, “Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation,” *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 2, pp. 209–220, 2014.

[7] I. Kraljevski, M. P. Bissiri, and G. Strecha, “Analysis and Synthesis of Glottalization Phenomena in German-Accented English,” in *SPECOM 2014, Lecture Notes in Artificial Intelligence 8773*, A. Ronzhin et al., Ed. Springer International Publishing Switzerland, 2014, pp. 97–104.

[8] K. Surana, “Classification of vocal fold vibration as regular or irregular in normal voiced speech,” MIT, USA, 2006.

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

[10] J. Kane, T. Drugman, and C. Gobl, “Improved automatic detection of creak,” *Comput. Speech Lang.*, vol. 27, no. 4, pp. 1028–1047, Jun. 2013.

[11] T. Drugman, J. Kane, and C. Gobl, “Data-driven detection and analysis of the patterns of creaky voice,” *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1233–1253, 2014.

[12] D. H. Klatt and L. C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers.,” *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

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

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

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

[16] T. G. Csapó and G. Németh, “Irreguláris beszéd regulárissá alakítása beszédkódoláson alapuló módszerrel [Transforming irregular speech to regular speech based on voice coding] (in Hungarian),” *Beszédkutatás 2014 [Speech Res. 2014]*, pp. 193–204, 2014.

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

[18] P. N. Garner, M. Cernak, and P. Motlicek, “A simple continuous pitch estimation algorithm,” *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 102–105, 2013.

[19] T. Drugman and T. Dutoit, “A comparative evaluation of pitch modification techniques,” in *European Signal Processing Conference*, 2010, pp. 756–760.

[20] D. Talkin, “A Robust Algorithm for Pitch Tracking (RAPT),” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.

[21] “Reference Manual for Speech Signal Processing Toolkit, Ver. 3.5.” 2011.

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

[23] T. Drugman and T. Dutoit, “Glottal closure and opening instant detection from speech signals,” in *Proc. Interspeech*, 2009, pp. 2891–2894.

[24] G. de Krom, “A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals,” *J. Speech Hear. Res.*, vol. 36, no. 2, pp. 254–266, Apr. 1993.

[25] 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 Res. 2013]*, pp. 261–270, 2013.

[26] G. Olaszy, G. Németh, and P. Olaszi, “Automatic prosody generation-a model for hungarian.,” in *Proc. Eurospeech*, 2001, pp. 525–528.

[27] L. Redi and S. Shattuck-Hufnagel, “Variation in the realization of glottalization in normal speakers,” *J. Phon.*, vol. 29, no. 4, pp. 407–429, 2001.

[28] T. Drugman and T. Raitio, “Excitation Modeling for HMM-based Speech Synthesis: Breaking Down the Impact of Periodic and Aperiodic Components,” in *Proc. ICASSP*, 2014, pp. 260–264.