Residual-based Excitation with Continuous F0 Modeling in HMM-based Speech Synthesis

Tamás Gábor Csapó[1], Géza Németh[1], and Milos Cernak[2]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
Magyar tudósok körútja 2., Budapest, Hungary
[2]Idiap Research Institute,
Rue Marconi 19, Martigny, Switzerland
{csapot,nemeth}@tmit.bme.hu
Milos.Cernak@idiap.ch

Abstract. In statistical parametric speech synthesis, creaky voice can cause disturbing artifacts. The reason is that standard pitch tracking algorithms tend to erroneously measure F0 in regions of creaky voice. This pattern is learned during training of hidden Markov-models (HMMs). In the synthesis phase, false voiced / unvoiced decision caused by creaky voice results in audible quality degradation. In order to eliminate this phenomena, we use a simple continuous F0 tracker which does not apply a strict voiced / unvoiced decision. In the proposed residual-based vocoder, Maximum Voiced Frequency is used for mixed voiced and unvoiced excitation. As all parameters of the vocoder are continuous, Multi-Space Distribution is not necessary during training the HMMs, which has been shown to be advantageous. Artifacts caused by creaky voice are eliminated with this speech synthesis system. A subjective listening test of English utterances has shown improvement over the traditional excitation.

Keywords: speech synthesis, HMM, creaky voice, vocoder, pitch tracking

1 Introduction

State-of-the-art text-to-speech synthesis is either based on unit selection or statistical parametric methods. Recently, particular attention has been paid to hidden Markov-model (HMM) based text-to-speech (TTS) synthesis [29], 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 representing excitation and spectrum of speech, and 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 vocoders, excitation models).

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 [31]. In this paper, we investigate the first factor. A large number of improved excitation models have been proposed recently (for a comparison, see [18]). 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 is such a non-modal phonation mode, which can cause disturbing artefacts in hidden Markov-model based text-to-speech synthesis. 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 referred to as irregular phonation, creaky voice, glottalization, vocal fry and laryngealization [4], [5]. It leads to abrupt changes in the fundamental frequency (F0), amplitude of the pitch periods or both. Irregular phonation 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 can cause problems for standard speech analysis methods (e.g. F0 tracking and spectral analysis) and it is often disturbing in speech technologies [8].

In this paper we propose an attempt to eliminate the phenomena of non-modal phonation in HMM-based speech synthesis. More specifically, we hypothesize that a continuous F0 tracker, which does not apply a strict voiced/unvoiced decision caused by creaky voice, can 'smooth' the voice irregularities that further improves modeling capabilities of the HMM-based training framework.

2 Related Work

In our earlier studies, we modeled the creaky voice in HMM-TTS explicitly using a rule-based [7] and a data-driven irregular voice model [8]. We used a residual codebook based excitation model [6], [9]. We also created an irregular-to-regular transformation method to smooth irregularities in speech databases [10]. Another alternative for overcoming the issues caused by creaky voice is to eliminate miscalculation of pitch tracking by using a more accurate fundamental frequency (F0) estimation method.

It has been shown recently that continuous F0 has advantages in statistical parametric speech synthesis [17]. For example, it was found that using a continuous F0, more expressive F0 contours can be generated [26–28] than using Multi-Space Distribution (MSD) [25] for handling discontinuous F0. Another important observation is that the voiced/unvoiced decision can be left up to the aperiodicity features in a mixed excitation vocoder [22]. This decision can also be modeled using a dynamic voiced frequency [13].

Accurate modeling of the residual has been shown to improve the synthesis quality [11, 18]. Using a Principal Component Analysis-based (PCA) 'Eigen-Residual' results in significantly more natural speech (in terms of artificiality, buzziness) than the traditional pulse-noise excitation [16].

In the following, we introduce a new combination of continuous F0 (based on pitch tracking with Kalman-smoother based interpolation, [17]), excitation modeling (PCA-based residual, [11]) and aperiodicity features (based on Maximum Voiced Frequency, MVF, [13]).

3 Methods

We trained three HMM-TTS systems with various parameter streams using two voices. In the following the used databases, the methods for analysis, training of HMMs and synthesis are presented for a baseline and two improved systems.

3.1 Data

Two English speakers were chosen from the CMU-ARCTIC database [21], denoted EN-M-AWB (Scottish English, male) and EN-F-SLT (American English, female). Both of them produced irregular phonation frequently, mostly at the end of sentences. For speaker EN-M-AWB, the ratio of the voiced frames produced with irregular phonation vs. all voiced frames is 2.25%, whereas for speaker EN-F-SLT, this ratio is 1.88%, measured on the full database using automatic creaky voice detection [12].

3.2 Analysis

A. HTS-F0std In the baseline system, the input is a speech waveform low-pass filtered at 7.6 kHz with 16 kHz sampling rate and 16 bit linear PCM quantization. The fundamental frequency (F0) parameters are calculated by a standard pitch tracker, the RAPT algorithm [23] implemented in Snack [2]. We denote this as 'F0std'. 25 ms frame size and 5 ms frame shift are used. In the next step 34-order Mel-Generalized Cepstral analysis (MGC) [24] is performed on the speech signal with $\alpha = 0.42$ and $\gamma = -1/3$. The results are the F0std and the MGC parameter streams.

As PCA-based residual has been shown to produce better speech quality than pulse-noise excitation [16], we perform residual analysis in the baseline system. The residual signal (excitation) is obtained by MGLSA inverse filtering [19]. The SEDREAMS Glottal Closure Instant (GCI) detection algorithm [15] is used to find the glottal period boundaries in the voiced parts of the residual signal. Pitch synchronous, two period long frames are used according to the GCI locations and they are Hann-windowed. The pitch-synchronous residual frames are resampled to twice the average pitch period of the speaker (e.g. for EN-M-AWB, $Fs = 16kHz$, $F0_{avg} = 123Hz$, $T0_{avg} = 130samples$, $framelen_{resampled} = 260samples$). Finally, Principal Component Analysis is applied on these frames, and the first principal component is calculated. Fig. 1 shows examples for the PCA residual. Instead of impulses, this PCA residual will be used for the synthesis of the voiced frames.

For text processing (e.g. phonetic transcription, labeling, etc.), the Festival TTS front-end is used [1].

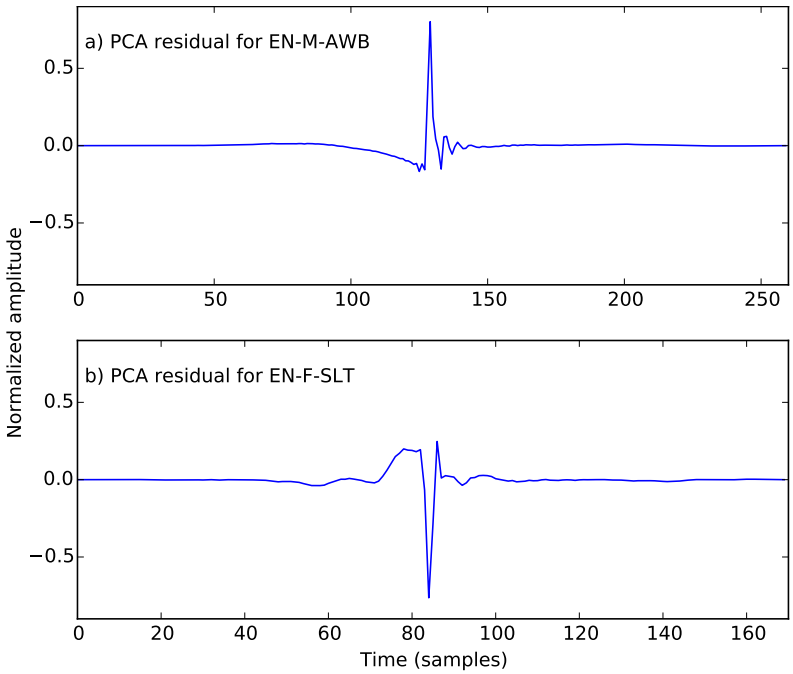

Fig. 1. PCA residuals obtained from two period long pitch-synchronous residuals.

B. HTS-F0std+MVF In the 2nd system, the analysis of the speech waveform and residual is similar to HTS-F0std, resulting in the MGC and F0std parameters. After these steps, Maximum Voiced Frequency is calculated from the speech signal using the MVF_Toolkit [14] with 5 ms frame shift, resulting in the MVF parameter.

C. HTS-F0cont+MVF In the third system, we use the same MGC parameter stream as in the baseline. For the calculation of the fundamental frequency, the open-source implementation [3] of a simple continuous pitch tracker [17], denoted as 'F0cont', is used. In regions of creaky voice, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. Similarly to the 2nd system, MVF is also calculated here. That is, in the 3rd system we use the F0cont, MGC and MVF parameter streams.

Fig. 2 (above the dashed line) shows all the steps applied in the analysis part of the HTS-F0cont+MVF system.

3.3 HMM Training

For training, the various parameters are calculated from each frame to describe the residual (F0std / F0cont / MVF) and the spectrum (MGC). During training, logarithmic values are used as they were found to be more suitable in our

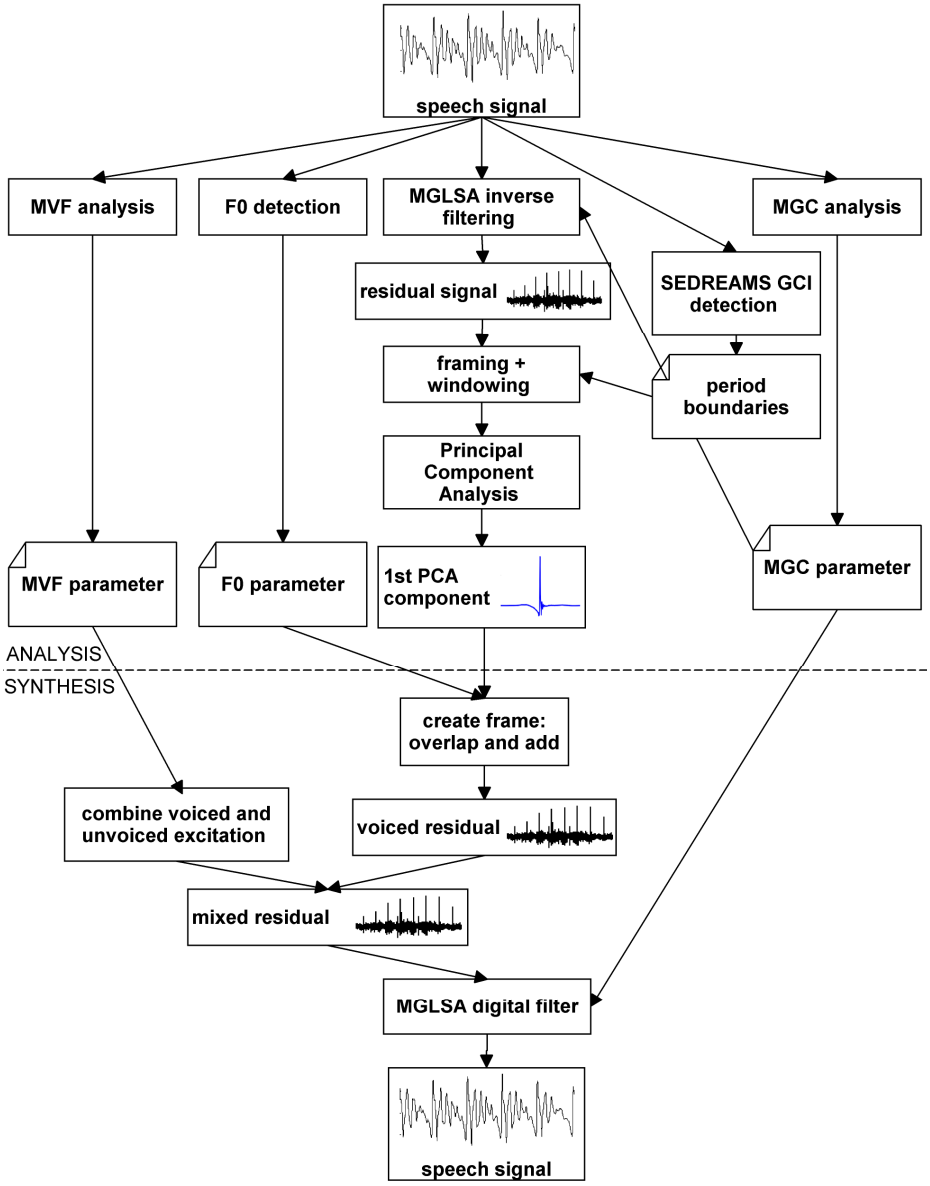

Fig. 2. Analysis (above the dashed line) and synthesis (below the dashed line) with the HTS-F0cont+MVF system.

Table 1. Parameter streams of the three systems.

	HTS-F0std	HTS-F0std+MVF	HTS-F0cont+MVF
$F0std$	MSD-HMM	MSD-HMM	-
$F0cont$	-	-	HMM
MGC	HMM	HMM	HMM
MVF	-	HMM	HMM

experiments. F0std is modeled with MSD-HMMs because this stream does not have values in unvoiced regions. All the other parameter streams (F0cont, MVF, and MGC) 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. Decision tree-based context clustering is used with context dependent labeling applied in the English version [29] of HTS 2.2. Independent decision trees are built for all the parameters and duration using a maximum likelihood criterion. The parameter streams for the systems are summarized in Table 1.

3.4 Synthesis

A. HTS-F0std In the baseline system, unvoiced excitation (if F0std = 0) is modeled as white noise. Voiced excitation (if F0std > 0) is generated using pitch-synchronously overlap-adding the first PCA component obtained during the analysis. This is lowpass filtered at 6 kHz (similarly to the HTS 2.2 demo system), and unvoiced excitation (white noise) is used in the higher frequency bands. For an example for the result of the synthesis with the HTS-F0std system, see Fig. 3 a) and b).

B. HTS-F0std+MVF In the 2nd system, PCA residuals are overlap-added similarly to the baseline system, depending on F0std. After that, this voiced residual is lowpass filtered at the frequency given by the MVF parameter. In the frequencies higher than the actual value of MVF, white noise is used. For an example for the result of the synthesis with the HTS-F0std+MVF system, see Fig. 3 c) and d).

C. HTS-F0cont+MVF Fig. 2 shows all the steps applied in the synthesis part of the HTS-F0cont+MVF system (below the dashed line). In this 3rd system, PCA residuals are overlap-added, but now the density of the residual frames is dependent on the F0cont parameter. As there is no strict voiced / unvoiced decision in the F0cont stream, the MVF parameter models the voicing information: for unvoiced sounds, the MVF is low (around 1 kHz), for voiced sounds, the MVF is high (typically above 4 kHz), whereas for mixed excitation sounds, the MVF is in between (e.g. for voiced fricatives, MVF is around 2–3 kHz). Voiced and unvoiced excitation is added together similarly to the 2nd system, depending on the MVF parameter stream (see 'mixed residual signal' in Fig. 2).

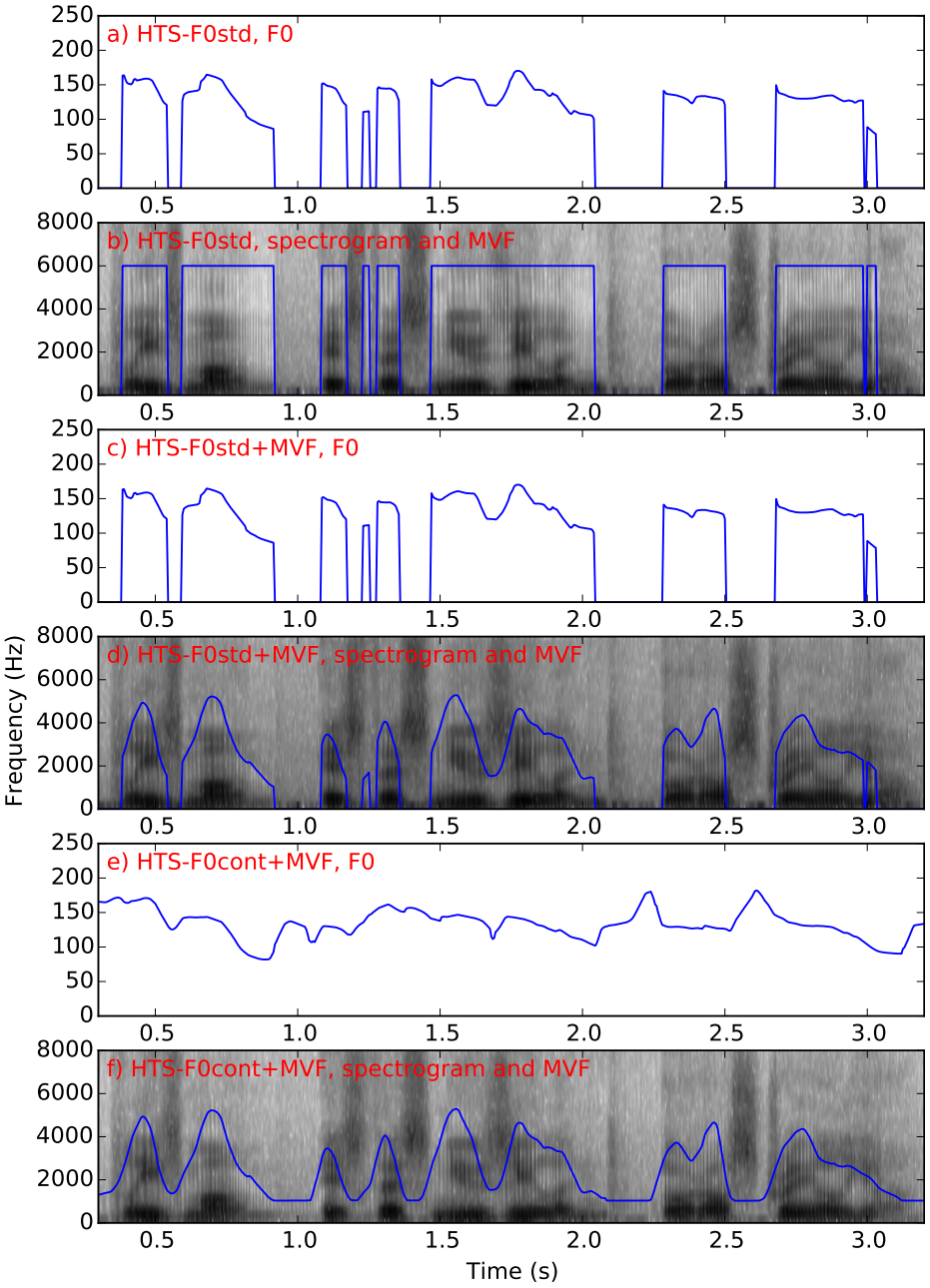

Fig. 3. Synthesis examples for the sentence: *'Please Mom, is this New Zealand, or Australia?'*, by speaker EN-M-AWB. Subfigures a), c) and e) show the generated F0std / F0cont parameters; subfigure b) shows a spectrogram and a fixed 6 kHz MVF; subfigures d) and e) show the spectrograms and the generated MVF parameter.

Fig. 3 e) and f) shows an example for the result of the synthesis with the HTS-F0cont+MVF system. By comparing all three sentence variants, it can be seen that in the baseline and 2nd systems (subfigures a) and c), F0std is modeled erroneously at the regions of creaky voice (between 1.23–1.36 s and 2.98–3.13 s) as a result of miscalculated F0 during the analysis. In the 3rd system (subfigure e), F0cont models this well and there is no unvoiced excitation at the final vowel of the sentence. In the baseline system, the voiced fricative sound 'z' (subfigure b), between 1.65–1.74 s) is extremely buzzy because of the fixed 6 kHz frequency between the voiced and unvoiced excitation. This is modeled better by adding the MVF parameter in the 2nd and 3rd systems: the Maximum Voiced Frequency in the 'z' sound is around 2.2 kHz (subfigures d) and f), between 1.65–1.74 s).

4 Evaluation

In order to evaluate the quality of the proposed HTS-F0std+MVF and HTS-F0cont+MVF methods, we have conducted a subjective listening test. A major factor that determines the usefulness of these methods is if human listeners accept the synthesized speech with no strict voiced / unvoiced decision and a dynamic Maximum Voiced Frequency.

Therefore, our aim was to measure the perceived 'naturalness'. We compared speech synthesis samples of the HTS-F0std baseline system with samples of the proposed systems (HTS-F0std+MVF and HTS-F0cont+MVF).

4.1 Methods of the Subjective Experiment

To obtain the speech stimuli, we created six models with the baseline and the two proposed systems and the two English speakers (EN-M-AWB and EN-F-SLT). 50 sentences not included in the training database were synthesized with all models and 10 sentences having at least one irregularly synthesized vowel in the baseline system were selected for the subjective test.

In the test, the three versions of each sentence were included in pairs, resulting altogether 60 utterances (2 speakers · 10 sentences · 3 versions). We created a web-based paired comparison test with one CMOS-like question (Comparative Mean Opinion Score). Before the test, listeners were asked to listen to an example from speaker EN-F-SLT to adjust the volume. In the test, the listeners had to rate the naturalness ('Which of the sentences is more natural?', '1 – 1st is much more natural' ... '5 – 2nd is much more natural') as a question for overall quality. The utterances were presented in a randomized order (different for each participant). The listening test can be found at `http://leszped.tmit.bme.hu/slsp2015_en/`.

4.2 Results of the Subjective Experiment

Altogether 8 listeners participated in the test (3 females, 5 males). They were all students or speech experts, between 24-45 years (mean: 37 year). They were not

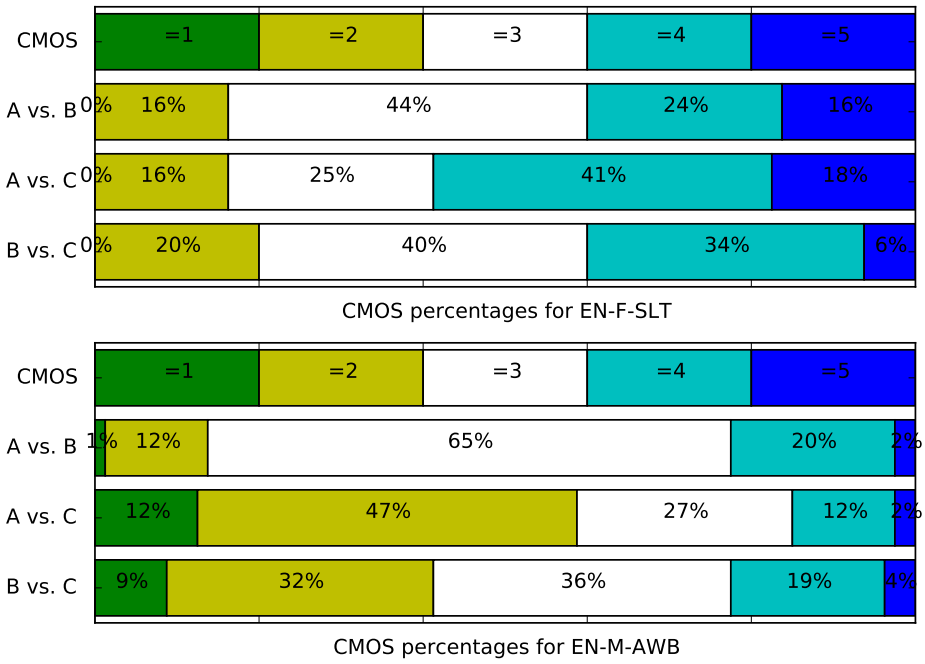

Fig. 4. Results of the listening test for the 'naturalness' question (top: speaker EN-F-SLT, bottom: speaker EN–M-AWB). A: HTS-F0std, B: HTS-F0std+MVF, C: HTS-F0cont+MVF. Average CMOS values can be found in the text of Section 4.2.

native speakers of English. On average the whole test took 18 minutes to complete. Two listeners noted that some of the sentences were too long to evaluate properly.

The results of the listening test are presented in Fig. 4 for the two speakers and three systems. The figure provides a comparison between the baseline A: HTS-F0std system and the two proposed systems (B: HTS-F0std+MVF, C: HTS-F0cont+MVF) pair by pair. The answers of the listeners for the 'naturalness' question were pooled together for the visualization (levels 1-2-3-4-5).

The ratings of the listeners were compared by t-tests as well, with a 95 % confidence level. For speaker EN-F-SLT, HTS-F0std+MVF was significantly preferred over HTS-F0std (average CMOS for A vs. B: 3.40) and HTS-F0cont+MVF was significantly preferred over both HTS-F0std and HTS-F0std+MVF (average CMOS for A vs. C: 3.60 and for B vs. C: 3.26). This result means that for the female voice, listeners evaluated the proposed systems as being significantly more natural than the baseline system. Fig. 4 also shows that for speaker EN-F-SLT (top subfigure), there was no '=1' answer from the listeners.

For speaker EN-M-AWB, system B was slightly preferred over system A, although this difference is not significant (average CMOS for A vs. B: 3.10). However, both system A and system B reached significantly higher CMOS scores than system C (average CMOS for A vs. C: 2.46 and for B vs. C: 2.76).

From this result we can conclude that adding the MVF parameter increased the naturalness, but combined with F0cont, this introduced audible vocoding artifacts. By investigating the synthesis samples of speaker EN-M-AWB we found that the HTS-F0cont+MVF system often resulted in too strong voiced component in the lower frequency bands for the unvoiced sounds, which might have been disturbing for the listeners. The original recordings of speaker EN-M-AWB contain significant background noise, and the vocoder introduced unwanted buzzy components because of this.

During the listening test we noted that subjective ratings can hardly be focused on buzziness or voiced/unvoiced transitions, but are mostly influenced by overall speech quality. Hence, it was difficult to evaluate changes in segmental level separately.

5 Discussion and Conclusions

It was found earlier that using continuous F0 has advantages in HMM-TTS [17, 26–28]. Our experiments further support this, because the disturbing artifacts caused by creaky voice were eliminated by the proposed vocoder. During training the HMMs, Multi-Stream Distribution modeling was not necessary, because all paramaters of the HTS-F0cont+MVF system are continouos. In this system, the voiced/unvoiced decision was left up to the Maximum Voiced Frequency parameter. This kind of aperiodicity modeling is similar to other mixed excitation vocoders [22, 13], but our system is simpler, i.e. uses less parameters compared to STRAIGHT-based mixed excitation [20, 30]. However, MVF does not always work well for voiced/unvoiced decision (e.g. in case of unvoiced stops there is a disturbing component in the samples of HTS-F0cont+MVF). In future work we will decrease the periodic component of unvoiced sounds.

In this paper we introduced a new vocoder, using 1) Principal Component Analysis-based residual frames, 2) continuous pitch tracking, and 3) Maximum Voiced Frequency. In a listening test of English speech synthesis samples, the proposed system with a female voice was evaluated as significantly more natural than a baseline system using only PCA-based residual excitation. The listening test results of the male voice have shown that there is room for improvement in modeling the unvoiced sounds with this continuous F0 model. MVF-based mixed voiced and unvoiced excitation was found to be extremely useful for modeling the voiced fricatives (e.g. 'z' in Fig. 3). However, in case of unvoiced sounds, the lowest MVF value was 1 kHz, which was disturbing for the male voice, but acceptable for the female voice. It is a question whether the buzziness caused by the combined F0cont and MVF modeling can be reduced. In the future, we plan to add a Harmonics-to-Noise Ratio parameter to both the analysis and synthesis steps in order to investigate this and to further reduce the buzziness caused by vocoding.

With the proposed methods we extend previous speech processing techniques dealing with irregular phonation. The above models and results might be useful for more natural, expressive, personalized speech synthesis.

Acknowledgments.

We would like to thank the listeners for participating in the subjective test. We thank Philip N. Garner for providing the continuous pitch tracker open source and Bálint Pál Tóth for useful comments on this manuscript. 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 no IZ73Z0_152495-1) and by the EITKIC project (EITKIC_12-1-2012-001).

References

1. The Festival Speech Synthesis System [Computer program], Version 2.1 (2010), http://www.cstr.ed.ac.uk/projects/festival/
2. The Snack Sound Toolkit [Computer program], Version 2.2.10 (2012), http://www.speech.kth.se/snack/
3. Speech Signal Processing - a small collection of routines in Python to do signal processing [Computer program] (2015), https://github.com/idiap/ssp
4. Blomgren, M., Chen, Y., Ng, M.L., Gilbert, H.R.: Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. The Journal of the Acoustical Society of America 103(5), 2649–2658 (May 1998)
5. Bőhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G., Aubergé, V.: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In: Acoustics'08. pp. 6141–6146. Paris, France (2008)
6. Csapó, T.G., Németh, G.: A novel codebook-based excitation model for use in speech synthesis. In: IEEE CogInfoCom. pp. 661–665. Kosice, Slovakia (Dec 2012)
7. Csapó, T.G., Németh, G.: A novel irregular voice model for HMM-based speech synthesis. In: Proc. ISCA SSW8. pp. 229–234. Barcelona, Spain (2013)
8. Csapó, T.G., Németh, G.: Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE Journal of Selected Topics in Signal Processing 8(2), 209–220 (2014)
9. Csapó, T.G., Németh, G.: Statistical parametric speech synthesis with a novel codebook-based excitation model. Intelligent Decision Technologies 8(4), 289–299 (2014)
10. Csapó, T.G., Németh, G.: Automatic transformation of irregular to regular voice by residual analysis and synthesis. In: Proc. Interspeech (2015), accepted.
11. Drugman, T., Dutoit, T.: The Deterministic Plus Stochastic Model of the Residual Signal and its Applications. IEEE Transactions on Audio, Speech and Language Processing 20(3), 968–981 (Mar 2012)
12. Drugman, T., Kane, J., Gobl, C.: Data-driven detection and analysis of the patterns of creaky voice. Computer Speech and Language 28(5), 1233–1253 (2014)
13. Drugman, T., Raitio, T.: Excitation Modeling for HMM-based Speech Synthesis: Breaking Down the Impact of Periodic and Aperiodic Components. In: Proc. ICASSP. pp. 260–264. Florence, Italy (2014)
14. Drugman, T., Stylianou, Y.: Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra. IEEE Signal Processing Letters 21(10), 1230–1234 (2014)
15. Drugman, T., Thomas, M.: Detection of glottal closure instants from speech signals: a quantitative review. IEEE Transactions on Audio, Speech and Language Processing 20(3), 994–1006 (Mar 2012)

16. Drugman, T., Wilfart, G., Dutoit, T.: Eigenresiduals for Improved Parametric Speech Synthesis. In: EUSIPCO09 (2009)
17. Garner, P.N., Cernak, M., Motlicek, P.: A simple continuous pitch estimation algorithm. IEEE Signal Processing Letters 20(1), 102–105 (2013)
18. Hu, Q., Richmond, K., Yamagishi, J., Latorre, J.: An experimental comparison of multiple vocoder types. In: Proc. ISCA SSW8. pp. 155–160 (2013)
19. Imai, S., Sumita, K., Furuichi, C.: Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications) 66(2), 10–18 (1983)
20. Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication 27(3), 187–207 (1999)
21. Kominek, J., Black, A.W.: CMU ARCTIC databases for speech synthesis. Tech. rep., Language Technologies Institute (2003)
22. Latorre, J., Gales, M.J.F., Buchholz, S., Knil, K., Tamura, M., Ohtani, Y., Akamine, M.: Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification? In: Proc. ICASSP. pp. 4724–4727. Prague, Czech Republic (2011)
23. Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (eds.) Speech Coding and Synthesis, pp. 495–518. Elsevier (1995)
24. Tokuda, K., Kobayashi, T., Masuko, T., Imai, S.: Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In: Proc. ICSLP. pp. 1043–1046. Yokohama, Japan (1994)
25. Tokuda, K., Mausko, T., Miyazaki, N., Kobayashi, T.: Multi-space probability distribution HMM. IEICE Transactions on Information and Systems E85-D(3), 455–464 (2002)
26. Yu, K., Thomson, B., Young, S., Street, T.: From Discontinuous To Continuous F0 Modelling In HMM-based Speech Synthesis. In: Proc. ISCA SSW7. pp. 94–99. Kyoto, Japan (2010)
27. Yu, K., Young, S.: Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis. IEEE Transactions on Audio, Speech, and Language Processing 19(5), 1071–1079 (2011)
28. Yu, K., Young, S.: Joint modelling of voicing label and continuous F0 for HMM based speech synthesis. In: Proc. ICASSP. pp. 4572–4575. Prague, Czech Republic (2011)
29. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.: The HMM-based speech synthesis system version 2.0. In: Proc. ISCA SSW6. pp. 294–299. Bonn, Germany (2007)
30. Zen, H., Toda, T., Nakamura, M., Tokuda, K.: Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Transactions on Information and Systems E90-D(1), 325–333 (2007)
31. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51(11), 1039–1064 (Nov 2009)