Intelligent Decision Technologies 8 (2014) 289–299
DOI 10.3233/IDT-140197
IOS Press

Statistical parametric speech synthesis with a novel codebook-based excitation model

Tamás Gábor Csapó and Géza Németh
Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

Abstract. Speech synthesis is an important modality in Cognitive Infocommunications, which is the intersection of informatics and cognitive sciences. Statistical parametric methods have gained importance in speech synthesis recently. The speech signal is decomposed to parameters and later restored from them. The decomposition is implemented by speech coders. We apply a novel codebook-based speech coding method to model the excitation of speech. In the analysis stage the speech signal is analyzed frame-by-frame and a codebook of pitch synchronous excitations is built from the voiced parts. Timing, gain and harmonic-to-noise ratio parameters are extracted and fed into the machine learning stage of Hidden Markov-model based speech synthesis. During the synthesis stage the codebook is searched for a suitable element in each voiced frame and these are concatenated to create the excitation signal, from which the final synthesized speech is created. Our initial experiments show that the model fits well in the statistical parametric speech synthesis framework and in most cases it can synthesize speech in a better quality than the traditional pulse-noise excitation. (This paper is an extended version of [10].)

Keywords: Text-to-speech synthesis, speech processing, excitation model, vocoding, parametric

1. Introduction

Speech is one of the main modalities of human-human communication and is important in human-computer communication as well. Cognitive Infocommunications (CogInfoCom) is the intersection of cognitive communication and informatics [4]. According to its definition, it investigates the link between the research areas of infocommunications and cognitive sciences [3]. This discipline evolved at the end of the last decade [31], when consistent terminology was proposed for the convergence [30]. Speech synthesis can have a major role in CogInfoCom by providing a natural inter-cognitive sensor-bridging communication mode [3]. Synthesized speech can be used in many applications where it is beneficial to extend the graphical user interface with speech interface [26]. Such applications include weather forecast in mobile phone or tablet, talking robot, car speech interface and telesurgery. In addition, speech synthesis is helpful for the visually impaired and blind people to access information.

State-of-the-art text-to-speech synthesis is often based on statistical parametric methods. Particular attention is paid to Hidden Markov-model (HMM) based text-to-speech (TTS) synthesis [44]. 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, typically simple vocoders 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 text-to-speech synthesis systems. The aim of this paper is to reduce the "buzziness" of HMM-TTS.

*Corresponding author: Tamás Gábor Csapó, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary. E-mail: csapot@tmit.bme.hu.

ISSN 1872-4981/14/$27.50

According to the source-filter theory, speech can be split into the source and filter [18]. The source signal (excitation) represents the glottal source that is created in the human glottis. The filter represents the vocal tract (including the mouth, tongue, lips, etc.). Traditionally linear prediction (LPC) analysis can be used for the source-filter decomposition, which results in a residual signal modeling the excitation source. Recently more complex and more accurate filtering methods have been used, including mel-generalized cepstrum (MGC) analysis [38]. The excitation source of speech can be obtained with inverse filtering.

1.1. Excitation models

In the baseline HMM-based speech synthesis system (HTS [44]), a very simple LPC vocoder is used for source-filter modeling: an impulse sequence is used as excitation in voiced parts, while unvoiced parts are modeled with white noise (see Fig. 1, left). However, this produces "buzzy" speech quality, for which HMM-based systems are often criticized. Several approaches have been proposed to overcome this problem. Figure 1 shows the difference between the oversimplified impulse train as source signal (as in the simple vocoder of baseline HTS) and the real excitation signal of speech that was obtained by MGC inverse filtering. The accurate modeling of the excitation signal of speech or the glottal source signal has proven to be very difficult.

Yoshimura and his colleagues were the first to introduce mixed excitation [43], meaning that the voiced parts include not only pulse components but noise excitation as well [45] continues this direction and introduces STRAIGHT-based vocoding which has been found to produce the best quality HMM-based synthesized speech until now. In [24], the impulse and noise parts of the excitation are modified with state-dependent filters to better model the excitation waveform. The procedures applied here resemble analysis-by-synthesis speech coding algorithms.

Cabral uses the Liljencrants-Fant (LF) acoustic model of the glottal source derivative [19] to construct the excitation signal [5]. A strong argument for using the LF model is that the LF waveform has a decaying spectrum at higher frequencies, which is more similar to the real glottal source excitation signal [8] than pulse or mixed excitation. In [7] Glottal Spectral Separation is introduced which consists of separating the glottal source effects from the spectral envelope of the speech [6] summarizes the latest results of HTS-LF which is claimed to have slightly better results than the STRAIGHT-based system. However, the model leaves room for improvement in terms of reproducing the original speaker characteristics.

The method presented in [2] allows high quality reconstruction of speech signals assuming a Harmonics plus Noise Model (HNM). The speech is decomposed to harmonic and stochastic parts. The harmonics are modeled with sinusoids, while the stochastic part is modeled as white Gaussian noise passing through a shaping filter. In [17] the method is extended with the modeling of Maximum Voiced Frequency, which is stated to have an even better synthesis performance.

In the excitation model of [41] the residual amplitude spectrum of only half of pitch period length is preserved in synthesis stage and zero-phase criterion is used to synthesize the excitation frame. In [42] the above model is extended with an adaptation of the Harmonic plus Noise Model, and the model is integrated into the HMM-based speech synthesis system. The Voicing Cut-Off Frequency is estimated and used for separating the harmonic and noise components to different frequency bands [40] continues this work and introduces an amplitude spectrum based excitation model which has comparable quality to that of STRAIGHT when integrated into HTS.

Waveform interpolation (WI) is introduced in [33] for excitation modeling in HMM-TTS. In this method, characteristic waveforms are extracted from LP residuals and they are compressed with Principal Component Analysis (PCA). It has been shown that using this model, the excitation signal evolves smoothly [22] extends this model with the concept of slowly evolving waveform (SEW) and rapidly evolving waveform (REW). In objective experiments it was found that the trainability of SEW and REW parameters is better than the parameters of HTS-STRAIGHT and the new method results lower spectral distortion [32] adds time domain and frequency domain zero padding techniques to the WI model in order to further reduce the spectral distortion. Furthermore, they apply non-negative matrix factorization to obtain a low-dimensional representation of the excitation signal.

Drugman was one of the first researchers to create a CELP (Code-Excited Linear Prediction) like excitation synthesis solution [16]. During analysis of speech, a codebook of pitch-synchronous residual frames (excitations) is constructed and similar techniques are used like the above HNM-based approaches. The codebook is applied in HMM-based speech synthesis: PCA is used for data compression and the result-

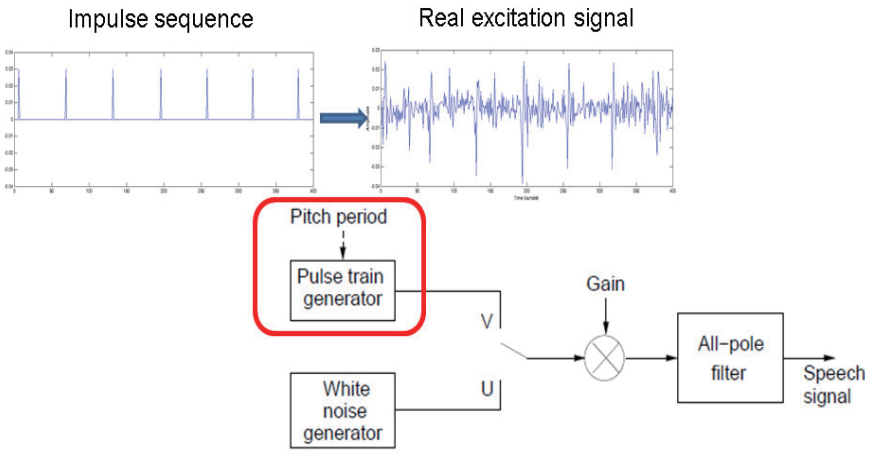

Fig. 1. Difference between source signals of speech within the HMM-TTS framework (extended from [46]). (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDT-140197)

ing 'eigenresiduals' are resampled to the suitable pitch and overlap-and-added together. The extended version of this method (Deterministic plus Stochastic Model, DSM) is introduced in [15] and several applications of it are shown in [11] and [12]. The deterministic part of the excitation contains the low-frequency contents, while the stochastic component is a high-pass filtered white noise. The authors argue that the first eigenvector of residuals usually dominates the deterministic component; therefore using eigenvectors of superior ranks is not necessary. This results in a very simple model, in which excitation is only parameterized by the pitch, while providing high-quality speech synthesis.

Raitio and his colleagues use glottal inverse filtering within HMM-based speech synthesis for generating natural sounding synthetic speech [28,29]. Glottal flow pulses are extracted from real speech via Iterative Adaptive Inverse Filtering (IAIF [1]), and these are used as voice source [36] introduces the GlottHMM system in which the glottal excitation is further modified to the desired voice source characteristics. During synthesis, one specific glottal source pulse is used for a whole sentence [27,34] extend this model with a glottal source pulse library. Here, a library of glottal source pulses is extracted from the estimated voice source signal and used during synthesis. The synthesized excitation is concatenated from the elements of the pulse library, retaining the dynamics of the voice source [35] introduces a hybrid approach, in which HMM-based speech synthesis is combined with unit selection glottal source concatenation. According to the subjective tests, the quality of the final GlottHMM system is high and clearly better than traditional excitation methods.

In our approach, we aim to create a codebook-based excitation model that uses unit selection. We have presented the initial version of this excitation model in [10] for speech analysis and synthesis, which is further improved here. During the encoding part of the model, the excitation signal is obtained from natural speech with MGC-based inverse filtering. Starting from this signal, a codebook is built from pitch-synchronous excitation frames. Several parameters (e.g. period, peak indices, harmonic-to-noise ratio and gain) of these frames are used to fully describe the modeled signal. During decoding, excitation frames are selected from the codebook with unit selection, and concatenated to each other. The final synthesized speech is obtained with MGC-based filtering, the parameters of which are set by the HMM-based speech synthesis framework.

1.2. Structure of the paper

The goal of our work is to further improve the way the source-filter model is used in statistical parametric speech synthesis and to introduce the improved version of our novel excitation model [10]. A great advantage of the model is that the residual can be obtained directly from the inverse filtered speech signal (therefore no approximation of the glottal source signal is necessary). The residual extracted from real speech can be used as the excitation for the synthetic speech signal, which provides more natural synthesis quality compared to the pulse train excitation. The method is flexible and scalable enough to optimize it for a mobile phone based text-to-speech system, as we use only a

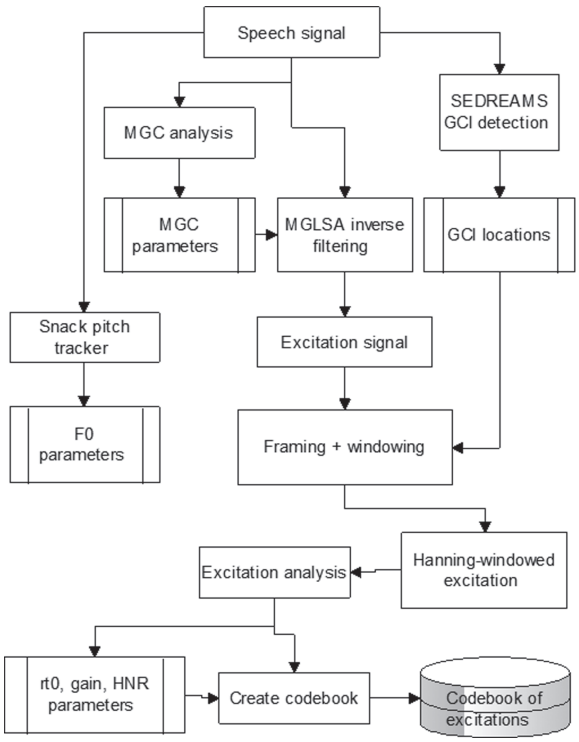

Fig. 2. Encoding of the speech signal.

few parameters to describe speech. It is not straightforward to create a model which suits to the requirements of the machine learning part of the statistical parametric speech synthesis framework. According to our preliminary experiments, our method works well with the HTS system.

The next sections are organized as follows: in Section 2 the details of our novel excitation model are presented. Section 3 introduces the Hidden Markov model based text-to-speech synthesis framework and the baseline system. Section 4 shows how we have integrated the excitation model to the HTS system. In Section 5 a subjective evaluation of the method and its results are presented. Finally, Section 6 summarizes the paper and shows the advantages of our novel excitation model applied in speech synthesis.

2. Novel excitation model

In our approach, the aim is to create a codebook-based excitation model for use in text-to-speech synthesis. Similarly to other speech coding methods, it consists of two main steps: encoding speech to parameters and decoding speech from parameters. In the encoding part, speech residual is obtained, divided into frames and several parameters describing these frames are saved. A codebook of residuals is built from voiced frames. Unvoiced frames are modeled by white noise. The residual signal is reconstructed from the parameters on a frame-by-frame basis using the previously built codebook with pitch synchronous overlap-and-add.

2.1. Encoding of excitation

Figure 2 shows the details of the analysis (speech encoding) stage. 16 kHz, 16 bit speech stored in a waveform is the input of the method. First, the fundamental frequency (F0) parameters are calculated by the publicly available Snack ESPS pitch tracker [48] with 25 ms frame size and 5 ms frame shift [37]. After that, Mel-Generalized Cepstrum (MGC) analysis [38] is performed on the same frames with the SPTK toolkit [47]. MGC is used here similarly as in HTS, as these features capture the spectral envelope efficiently. For the MGC parameters, we use alpha $=$ 0.42 and gamma $= -1/3$ instead of the default HTS parameters, as recommended in [16]. The residual signal (excitation) is obtained by inverse filtering with a MGLSA (Mel-Generalized Log Spectral Approximation) digital filter [21]. Next, the SEDREAMS Glottal Closure Instant (GCI) detection algorithm is used to find the glottal period boundaries (GCI locations) in the voiced parts of the speech signal [14]. We chose SEDREAMS because it has been shown that among the available GCI detection algorithms this method has the highest identification rate and accuracy of finding the GCI peaks in the excitation signal, and it is robust to additive noise and reverberation [14].

Further analysis steps are completed on the excitation signal with the same frame shift values. The first step is voiced/unvoiced decision. For measuring the parameters in the voiced sections, pitch synchronous, two period long frames are used according to the GCI locations and they are Hanning-windowed. In the unvoiced parts, a fixed 25 ms frame length is used. First, the gain (energy) of the frame is measured. If the frame is unvoiced, we do not apply further processing. If the frame is voiced, a codebook is built from pitch-synchronous excitation frames. Several parameters of these frames are used to fully describe the speech excitation:

- F0: fundamental frequency of the frame
- gain: energy of the frame
- rt0 peak indices: the locations of prominent values (peaks or valleys) in the windowed frame
- HNR: Harmonic-To-Noise ratio of the frame [23]

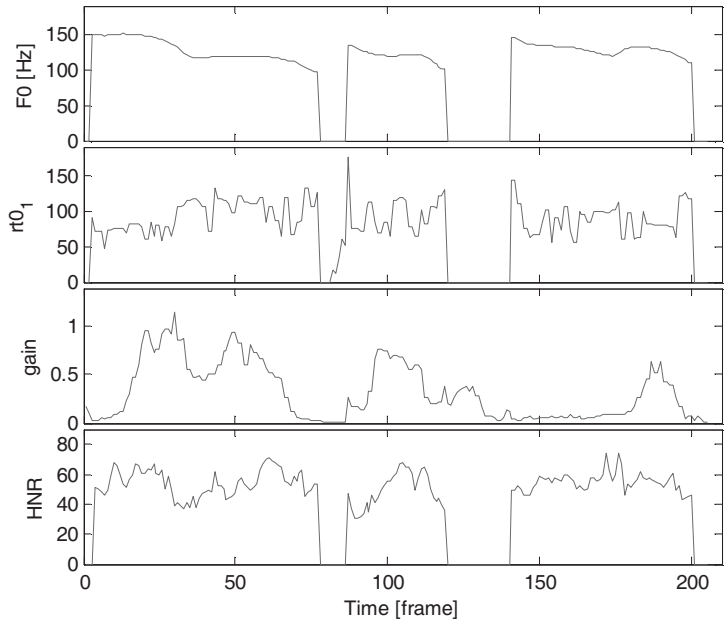

Fig. 3. Parameter types for describing excitation.

Figure 3 shows an example for these parameters extracted from a short sentence. The F0 curve (1st row) is smooth and shows the voiced and unvoiced sections of the signal (in unvoiced parts F0 is zero). The 2nd and 4th rows (rt0_1 and HNR) are calculated only in voiced regions, and these parameters are quite unstable. Gain (3rd row) is calculated both in voiced and unvoiced regions. 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 synthesis. In order to collect similar codebook elements, the RMSE (Root Mean Squared Error) distance is calculated between the pitch normalized versions of the codebook elements. The normalization is performed by resampling the codebook element to 40 samples. This distance will be used as concatenation cost during encoding. For excitation codebook building, more sophisticated methods are presented in Section 4.

2.2. Decoding of excitation

Figure 4 shows the steps of the synthesis (speech decoding) stage. The input parameters are obtained during encoding (F0, gain, rt0 indices, HNR and the codebook of pitch-synchronous excitations), or they are generated by HMMs during text-to-speech synthesis (see Section 4). For each parameter set, a 25 ms frame is built with 5 ms shift.

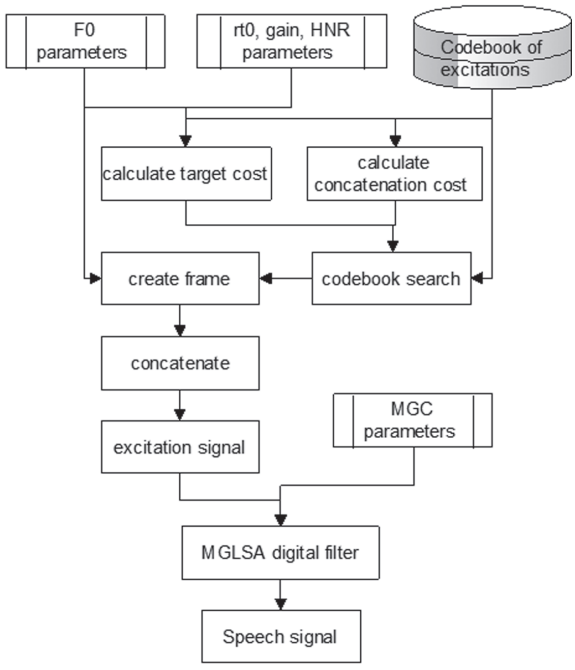

Fig. 4. Decoding of the speech signal.

If the frame is unvoiced, random noise is generated with the gain as energy. 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, sim-

ilarly to unit selection speech synthesis [20]. 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 excitation is created by pitch synchronously overlap-adding the Hanning-windowed excitation periods. Finally, the energy of the frame is set using the gain parameter in both voiced and unvoiced regions.

The whole excitation signal is built by concatenating the pitch synchronous frames and white noise parts. Synthesized speech is obtained from the excitation signal with MGC-based filtering using the MGLSA digital filter [21].

2.3. Weight setting

During decoding of the excitation, we use a cost for the codebook search that consists of concatenation cost and target cost. In order to calibrate the suitable weight of these costs, a simple evaluation procedure was established in [10]. The ratio of target cost and concatenation cost was varied between $C_{ratio} = \{0.01, 0.1, 1, 10, 100\}$ and five short sentences from four Hungarian speakers (two male and two female) were selected. The sentences were encoded and decoded with the proposed excitation model with all five cost settings. After listening to the recoded sentences, usually the equal weight for concatenation and target cost were preferred ($C_{ratio} = 1$). When the concatenation cost was stronger ($C_{ratio} = 0.01$ or 0.1) the utterances sounded buzzy because of the repeated excitation frames. In the other extreme, when the target cost was stronger ($C_{ratio} = 10$ or 100), the synthesized sentences often contained abrupt discontinuities caused by the sudden change of the excitation periods. Target cost is made of several subcosts (including a separate subcost for F0, rt0 and HNR), whose weights were set by hand crafting.

3. Hidden Markov model based speech synthesis baseline system

Hidden Markov model is a machine learning algorithm which has been successfully applied in both speech recognition and in speech synthesis, as this can simulate properly the behavior of physical processes based on observations [39]. HMM-based text-to-speech synthesis contains two main steps: training and speech synthesis. During the training stage, the parameters extracted from a large, precisely labeled speech corpus are trained by HMMs. As a result of the training, a small HMM database is created that includes the representative parameters. During the speech synthesis stage, the best matching parameters to the text to be read are selected from the database and a synthesized sentence is generated by a suitable vocoder.

3.1. Baseline system

As a baseline system, we used the Hungarian version of HTS with the simple pulse-noise excitation model [39], referred as HTS-PN. During our experiments, we applied speaker dependent training. The speech of only one speaker was used for training, analysis and synthesis. 1940 phonetically balanced sentences (2 hours of speech) from a male native Hungarian speaker were used as training corpus [25]. The sentences in the corpus are stored as 44.1 kHz, 16 bit waveforms, which were resampled to 16 kHz. The F0 range of the speaker is 50–220 Hz.

The training of the baseline system is shown in Fig. 5. First, pitch and spectral analysis is performed similarly as in Section 2.1. Log(F0), MGC and their first and second derivatives are stored in the parameter files. After that, phonetic transcriptions are extended to context dependent labels. During the training phase, the HMMs learn the parameters according to the context dependent labels. Parameters with varying dimensions are modeled by multi-space distribution HMMs (MSD-HMM). For example, log(F0) has a real number value in voiced regions and is undefined in unvoiced regions. For rhythm modeling, speech state duration densities are calculated for each phoneme. Phoneme-dependent state durations are modeled by multi-dimensional Gaussian distributions. Context-dependent labeling and decision trees are applied to reduce the combination of all context dependent features, using quintphones. Spectral, excitation and duration parameters are handled with separate decision trees [39].

Figure 6 shows the steps of the synthesis stage. The most likely parameters (pitch, state durations and spectral parameters) belonging to the text are generated by the HMMs and then speech is synthesized by the pulse-noise vocoder. Here, the excitation is modeled as a pe-

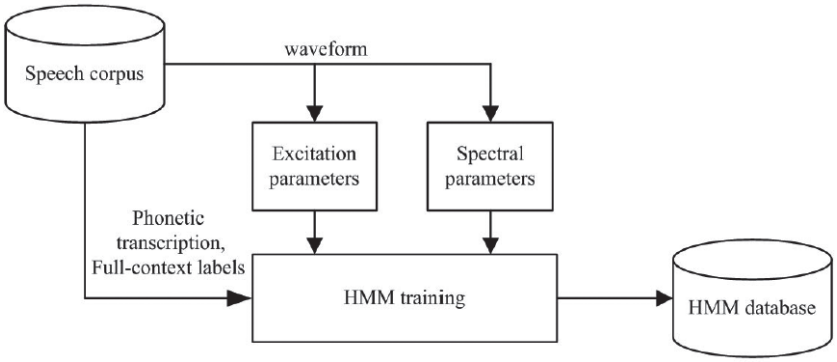

Fig. 5. HTS training with the baseline system [39], adapted from [46].

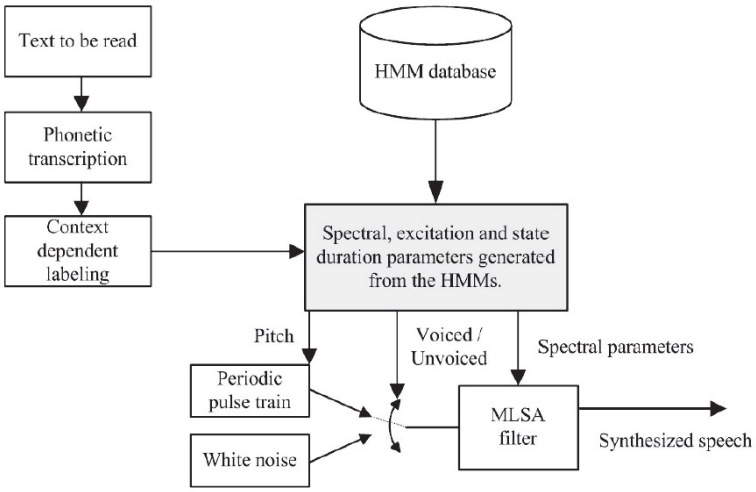

Fig. 6. HTS synthesis with the baseline system [39], adapted from [46].

riodic pulse train at the rate of the pitch that was generated by the HMMs in voiced frames, and as white noise in unvoiced frames. The excitation signal is filtered by a Mel-Generalized Log Spectral Approximation (MGLSA) filter [21] for the generation of synthesized speech.

4. HMM-TTS with novel excitation model

In this section we show how the proposed excitation model (Section 2) was integrated to the HTS system. The new excitation parameters were incorporated into the baseline Hungarian HTS system. We tested whether the parameters are suitable for machine learning. The new system is denoted by HTS-CDBK.

4.1. Parameter extraction, training and synthesis

Similarly to the HTS-PN system, HTS-CDBK consists of training and synthesis stages. During training, the same steps are applied as in the baseline system. The parameters for each sentence in the learning database are extended with those calculated in the encoding step of the novel excitation model (gain, rt0 and HNR) as described in Section 2.1. The logarithm of each of the parameters is calculated as these have more Gaussian distributions. Log(rt0) and log(HNR) are modeled by MSD-HMMs similarly to F0 because these all are undefined in unvoiced regions. Log(gain) has values in both voiced and unvoiced frames, therefore it is 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. During training, altogether five streams of data are con-

sidered: MGC coefficients, pitch, gain, rt0 and HNR. Several other parameters were tested as well, but those were not suitable for machine learning as the distribution of parameters did not satisfy the requirements of the HTS training procedures and the training stage was unsuccessful. At synthesis time, parameters generated from a constrained maximum likelihood algorithm are fed into the decoding part of the novel excitation model to produce the synthetic speech.

4.2. Codebook of excitations

In the decoding part of the excitation model, a codebook of pitch-synchronous excitation frames is used. We have experimented with the size (the number of frames stored) and the structure of the codebook, which is presented here in detail.

In [35], a pulse library consisting of about 20 000 pulses is used. However, it is suggested that a library with a much smaller size (e.g. 1000 pulses) might also be enough to achieve similar quality. In [16] PCA compression is applied to reduce the size of the codebook, but in [35] such reduction is not used.

While experimenting with the codebook size, our aim was to find a suitable size with which the synthesis is quick enough but the quality of speech is not degraded. The reason for using smaller codebooks is that the calculation of concatenation and target costs can be high when using large codebooks. First we used codebooks with about 30 000 frames, and the size could be reduced to 6 500 excitation frames without noticeable quality degradation. While creating codebooks, randomly selected sentences were chosen from the speech database and they were encoded and decoded with the excitation analysis-synthesis technique. During the codebook reduction, we paid attention to include frames with F0 values having a similar distribution than the original codebook.

We have experimented with phoneme-dependent codebooks as well. A separate codebook was built for each voiced phoneme type. This is motivated by the fact that the source-filter separation is never perfect, and in the excitation signal some phoneme-dependent information might remain. This approach has not yielded significant quality improvement, but by this phoneme-based clustering the calculation time of the concatenation cost could be reduced.

A simple optimization of the codebook size was also conducted. We synthesized 130 sentences with the HTS-CDBK system using the codebook consisting of 6500 frames. After that, we kept only those frames in the codebook which were used in the unit selection process of the decoding part of the excitation model. This way we could further reduce the codebook size to 1900 frames without any change in the quality of synthesized speech and this size of the codebook is suitable for real-time speech synthesis.

5. Subjective evaluation of the novel HMM-TTS

In order to evaluate the quality that can be achieved by our proposed HTS-CDBK system, we have conducted a listening test. 130 sentences were selected and synthesized with both HTS-PN and HTS-CDBK systems using a male voice, and 20 of them were included in the test. We created a web-based CMOS-like (Comparative Mean Opinion Score [9]) 5 point scale paired comparison test. After listening to each sentence pair, the listeners had to answer the question 'Which of the sentences has better quality?' with one of '1 – the first is much better, 2 – the first is better, 3 – equal, 4 – the second is better, 5 – the second is much better'. The sentences were presented in a randomized order (different for each participant) and the systems in the pairs were also randomized.

5.1. Subjective test environment

Altogether 16 listeners participated in the test. One subject was found to produce the answers randomly, therefore his results were not included. The data of 15 listeners was used in the statistical analysis. 12 males and 3 females were involved, between ages of 25–59 years. All of them were native speakers of Hungarian and none of them reported any hearing loss. On the average the whole test took 4.6 minutes to complete.

5.2. Test results

The distribution of CMOS values of the sentences can be seen in Fig. 7 and the average and standard deviation values are shown in Table 1. From the 20 sentences, in 13 cases the HTS-CDBK system was preferred, in five sentence pairs the systems were ranked as equal and in the remaining one case the HTS-PN system was preferred. We conducted significance tests as well (one-tailed t-test, $p < 0.0005$) and the results of the sentences altogether are significantly different from the average of 3.0 (mean CMOS $= 3.23$). In most of the cases the proposed HTS-CDBK system was preferred over the baseline HTS-PN system, or the systems were ranked as equal.

Table 1
Sentence by sentence mean and standard deviation results of the subjective listening test

Sentence #	1	2	3	4	5	6	7	8	9	10
CMOS mean	3.00	2.86	3.20	3.00	3.27	3.47	2.93	3.60	3.47	3.33
CMOS stddev	1.13	0.99	1.01	0.85	0.80	1.07	0.70	0.99	1.07	1.23
Sentence #	11	12	13	14	15	16	17	18	19	20
CMOS mean	3.27	3.53	3.13	3.53	3.47	3.40	2.93	3.40	2.73	3.07
CMOS stddev	1.34	0.92	1.25	1.12	0.99	1.24	0.96	1.06	1.03	1.39

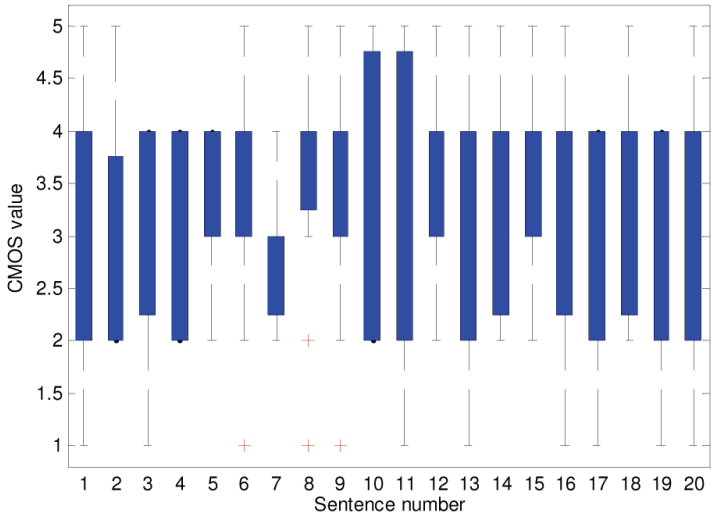

Fig. 7. Sentence by sentence results of the subjective listening test. Crosses show the outliers. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDT-140197)

6. Discussion and conclusions

In most cases of the evaluation in Section 5, HTS-CDBK synthesis resulted in good quality speech. However, the evaluation revealed some imperfections as well: in several cases large intensity perturbations were found in the synthesized speech signal. This might be caused by improper machine learning or by the concatenation of excitation frames that are too different from each other. In some other sentences 'creaky'-like voice was observed at the final word of the synthesized sentence. After investigating the training database we found that at the end of sentences, utterances are frequently characterized by irregular phonation, which can prevent the F0 detection algorithm from obtaining good results. In the future this can be solved by using other F0 detection algorithms or manually measuring the F0 in the critical sections. Note, that creaky voice synthesis is a new topic and includes several challenges [13].

During synthesis, our method modifies the period of the excitation frames by zero padding or deletion. In [12] resampling is used for this task, but [5] argues that resampling the residual results in unwanted spectral distortion. Therefore we tried to avoid such distortion when adjusting the pitch. We also investigated if the novel excitation model is suitable for pitch modification. According to a preliminary test it can achieve similar quality than [12] when increasing or decreasing the F0 of speech. Compared to the DSM [15] and GlottHMM [34], our approach is assumed to have a similar speech synthesis quality. However, subjective tests were not conducted yet to compare these systems.

In this paper we have shown that the novel excitation model is suitable for machine learning in HTS. This way, synthesized speech became more natural compared to the pulse-noise excitation. A great advantage of the model is that it uses MGC residual which can be obtained automatically with inverse filtering. Another advantage is that a few parameters are enough to describe the speech signal. By further improving the excitation model, we plan to synthesize different voice qualities (e.g. breathy, whispered or creaky) as well. As the model is flexible and scalable enough, we plan to implement and optimize it for mobile phone usage.

Acknowledgements

We would like to thank the listeners for participating in the subjective test. We thank the two anonymous reviewers for the helpful comments and suggestions. This research was partially supported by the Paelife (Grant No AAL-08-1-2011-0001), the CESAR (Grant No 271022) and the EITKIC_12-1-2012-001 projects.

References

[1] P. Alku, Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering, *Speech Communication* **11**(2–3) (Jun. 1992), 109–118.

[2] E. Banos, D. Erro, A. Bonafonte and A. Moreno, Flexible harmonic/stochastic modeling for HMM-based speech synthesis, in Proc. V Jornadas en Tecnologias del Habla, 2008, pp. 145–148.

[3] P. Baranyi and Á. Csapó, Definition and Synergies of Cognitive Infocommunications, *Acta Polytechnica Hungarica* **9**(1) (2012), 67–83.

[4] P. Baranyi, G. Németh and P. Korondi, 3D internet for Cognitive Infocommunication, *Informatika – A Gábor Dénes Föiskola Közleményei* **12**(2) (Dec. 2010), 3–6.

[5] J. Cabral, HMM-based Speech Synthesis Using an Acoustic Glottal Source Model, University of Edinburgh, United Kingdom, 2010, http://www.era.lib.ed.ac.uk/handle/1842/4877, accessed Feb 29, 2012.

[6] J. 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.

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

[8] 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.

[9] 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, http://festvox.org/blizzard/bc2007/blizzard_2007/full_papers/blz3_003.pdf, accessed Feb 29, 2012.

[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. Drugman, Advances in Glottal Analysis and its Applications, University of Mons, Belgium, 2011, http://tcts.fpms.ac.be/∼drugman/files/DrugmanPhDThesis.pdf, accessed Feb 20, 2012.

[12] 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* **20**(3) (2012), 968–981.

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

[14] T. Drugman and M. Thomas, Detection of glottal closure instants from speech signals: A quantitative review, *IEEE Transactions on Audio, Speech and Language Processing* **20**(3) (2012), 994–1006.

[15] 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.

[16] 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.

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

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

[19] G. Fant, J. Liljencrants and Q. Lin, A four-parameter model of glottal flow, *STL-QPSR* **4** (1985), 1–13.

[20] 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.

[21] S. Imai, K. Sumita and C. Furuichi, Mel Log Spectrum Approximation (MLSA) filter for speech synthesis, *Electronics and Communications in Japan (Part I: Communications)* **66**(2) (1983), 10–18.

[22] C. Jung, Y. Joo and H. Kang, Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems, *IEEE Signal Processing Letters* **19**(12) (Dec. 2012), 809–812.

[23] G. de Krom, A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals, *Journal of Speech and Hearing Research* **36**(2) (Apr. 1993), 254–266.

[24] R. Maia, T. Toda, H. Zen, Y. Nankaku and K. Tokuda, An excitation model for HMM-based speech synthesis based on residual modeling, in Proc. ISCA SSW6, 2007, pp. 131–136.

[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 Research 2013], 2013, pp. 261–270.

[26] M. Pleva, S. Ondas, J. Juhar, A. Cizmar, J. Papaj and L. Dobos, Speech and mobile technologies for cognitive communication and information systems, in IEEE CogInfoCom, 2011, pp. 1–5.

[27] 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.

[28] 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.

[29] 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* **19**(1) (Jan. 2011), 153–165.

[30] Gy. Sallai, Defining Infocommunications and Related Terms, *Acta Polytechnica Hungarica* **9**(6) (2012), 5–15.

[31] Gy. Sallai, The Cradle of the Cognitive Infocommunications, *Acta Polytechnica Hungarica* **9**(1) (2012), 171–178.

[32] 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* **E96-D**(2) (2013), 379–382.

[33] 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.

[34] A. Suni, T. Raitio, M. Vainio and P. Alku, The GlottHMM

entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation, in Blizzard Challenge 2011, http://festvox.org/blizzard/bc2011/HELSINKI_Blizzard2011.pdf, accessed Feb 22, 2012.

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

[36] A. Suni, T. Raitio, M. Vainio and P. Alku, The GlottHMM speech synthesis entry for Blizzard Challenge 2010, in Blizzard Challenge 2010, 2010, http://festvox.org/blizzard/bc2010/HELSINKI_Blizzard2010.pdf, accessed Feb 20, 2012.

[37] 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.

[38] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, Mel-generalized cepstral analysis – a unified approach to speech spectral estimation, in Proc. ICSLP, 1994, pp. 1043–1046.

[39] B. Tóth and G. Németh, Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis, *Acta Cybernetica* **19**(4) (2010), 715–731.

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

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

[42] 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.

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

[44] 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.

[45] 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* **E90-D**(1) (2007), 325–333.

[46] H. Zen, K. Tokuda and A.W. Black, Statistical parametric speech synthesis, *Speech Communication* **51**(11) (Nov. 2009), 1039–1064.

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

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