# A novel codebook-based excitation model for use in speech synthesis

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

***Abstract* —Speech synthesis is an important modality in Cognitive Infocommunications. 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 propose a novel speech coding method with codebook-based excitation. In the analysis stage the speech signal is analyzed frame-by-frame and a codebook of phoneme-dependent, pitch synchronous residuals is built from the voiced parts. During the synthesis stage the codebook is searched for a suitable element in each voiced frame and these are concatenated. Our initial experiments show that in most cases the method can re-synthesize speech to a similar quality than the original. This new excitation model fits well in the machine learning part of the statistical parametric speech synthesis framework.[1]**

## I. INTRODUCTION

Speech is one of the main modalities of human-human communication and is important in human-computer communication as well. Speech synthesis can have a major role in Cognitive Infocommunications [1] by providing a natural inter-cognitive sensor-bridging communication mode. Synthesized speech can be used in applications like talking robot, car speech interface and telesurgery. In addition, it is helpful for the vision 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 synthesis [2]. 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.

According to the source-filter theory, speech can be split into the source and filter [3]. 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, but recently more complex and more accurate filtering methods have been used, including mel-spectrum and mel-generalized cepstrum (MGC) analysis [4]. The excitation signal (source) of speech can be obtained with inverse filtering.

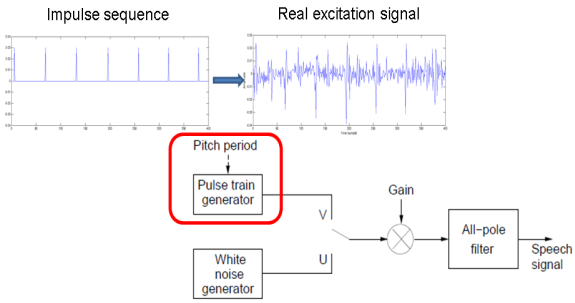


Figure 1. Difference between source signals of speech within the HMM TTS framework.

In the baseline HMM-based speech synthesis system (HTS, [2]), a very simple LPC vocoder is used for source-filter modelling and 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: Cabral et al. use the Liljencrants-Fant acoustic model of the glottal source derivative to construct the excitation signal [5,6], Erro et al. apply the Harmonic Plus Noise Model [7], Maia et al. experiment with MELP (Mixed Excitation LP) [8], while Wen and Tao modify the spectrum of the residual [9]. CELP (Codebook Excited LP) based methods offer the highest quality solutions to alleviate the "buzzyness" [, 10, 11, 12, 13]. Such speech coders have two main steps: encoding (analysis) when parameters are obtained from the speech signal, and decoding (synthesis) in which the speech signal is reconstructed from the parameters.

Fig. 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 goal of our initial research is to synthesize excitation signals that resemble the properties of real excitation more properly than the impulse sequence.

Drugman was one of the first researchers to create such an excitation synthesis solution [10, 11]. He constructs a codebook of residual frames (excitations) obtained from natural speech and uses it in HMM synthesis: "eigenresiduals" are resampled to the suitable pitch. Raitio et al. use unit selection methods for the synthesis of excitation, where glottal periods obtained from real speech are concatenated resulting in a smooth excitation signal [12, 13].

[1] This research is partially supported by the Paelife (Grant No AAL-08-1-2011-0001) and the CESAR (Grant No 271022) projects.

In our initial approach, we aim to create a phoneme-dependent codebook-based excitation model that uses unit selection. During the analysis part, the excitation signal is obtained from natural speech with MGC-based inverse filtering. Starting from this signal, a codebook is built from phoneme-dependent pitch-synchronous excitation frames. Several parameters (e.g. period, peak indices and gain) of these frames are used to fully describe the modeled signal. During synthesis, 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 goal of our work is to improve the source-filter model, particularly the modeling of the excitation signal in statistical parametric speech synthesis. Phoneme-dependent excitation codebooks have not been widely studied previously for this task, because according to the source-filter theory it is assumed that excitations are phoneme-independent. However, in practice it was found that the residual of LPC analysis has phoneme specific characteristics. 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: Section II describes the methods we used for speech processing. In Section III the results of evaluations of our algorithms are shown. Section IV concludes the paper.

## II. METHODS

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: analysis and synthesis. In the analysis part, speech excitation (called as residual) is obtained, divided into frames and several parameters of these frames are saved. A codebook of residuals is built from voiced frames. Unvoiced frames are modeled with white noise. The residual signal is reconstructed from the parameters on a frame-by-frame basis using the previously built codebook.

### A. Analysis

Fig. 2 shows the details of the analysis (speech encoding) part. 16 kHz, 16 bit speech is the input of the method. First, the F0 parameters are obtained by the publicly available Snack ESPS pitch tracker [14] with 25 ms frame size and 5 ms frame shift. After that, Mel-Generalized Cepstrum (MGC) analysis is performed [4] on the same frames. 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 found in [11]. The residual signal (excitation) is obtained by inverse filtering with a MGLSA (Mel-Generalized Log Spectral Approximation) digital filter [4]. 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 [54]. We chose SEDREAMS because it has been shown that among the available GCI detection algorithms it has the highest identification rate and accuracy, and it is robust to additive noise and reverberation [15].

The further analysis steps are completed on the residual signal with the same frame and shift values. First, the gain (energy) of the frame is measured. As second step voiced/unvoiced decision follows. If the frame is unvoiced, we do not apply further processing. If the frame is voiced, a two pitch period long signal part is cut according to the GCI locations and it is Hanning-windowed. Fig. 3 shows 1) a frame of speech 2) a frame of the corresponding residual signal with the GCI locations indicated, and 3) the windowed signal.

Starting from this two pitch period long signal, a codebook is built from phoneme-dependent pitch-synchronous excitation frames. Several parameters of these frames are used to fully describe the speech excitation:

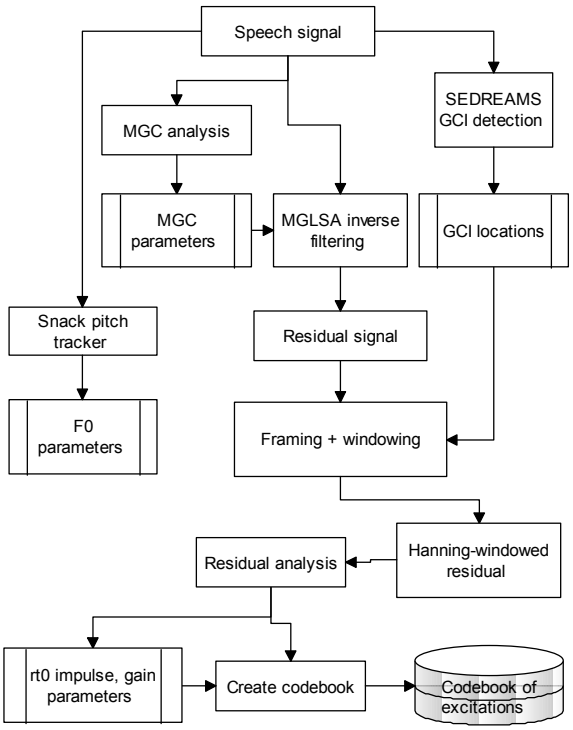


Figure 3. Analysis of the speech signal.

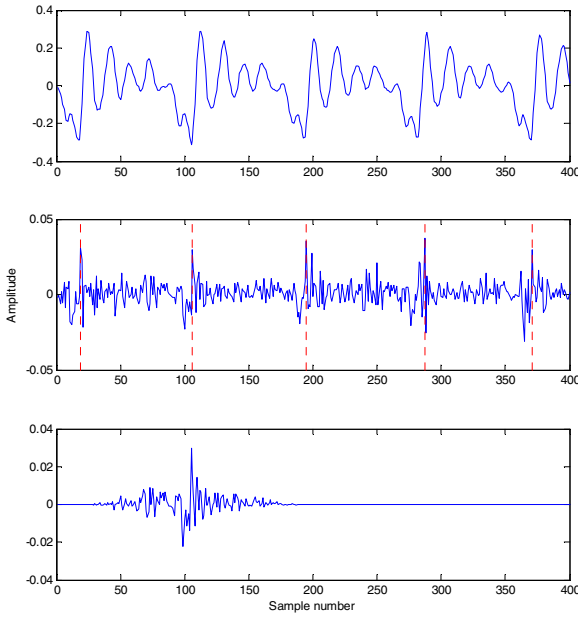


Figure 2. Result of the speech analysis.
1) Top: 400 samples of a 25 ms voiced speech frame (16 kHz sampling rate).
2) Middle: residual frame. Red vertical lines show the GCI locations.
3) Bottom: a 2 period long, Hanning-windowed part of the signal.

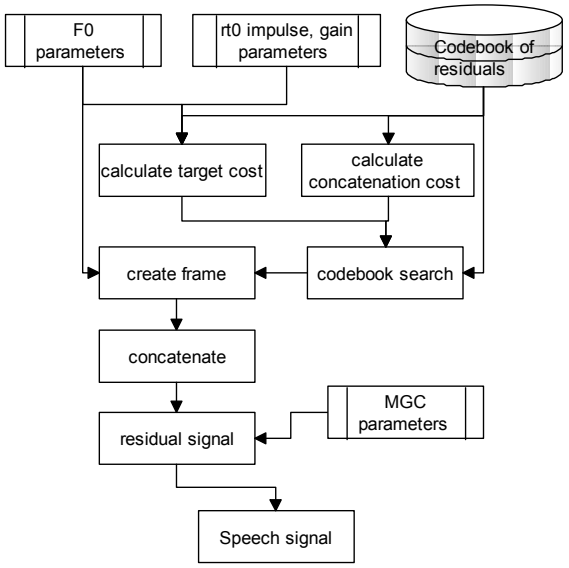


Figure 4. Synthesis of the speech signal.

- period: fundamental period of the signal
- gain: energy of the signal
- rt0 peak indices: the locations of prominent values (peaks or valleys)
- phoneme: the phoneme of the frame

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 version of the codebook elements belonging to the same phoneme. The normalization is performed by resampling the codebook element to 40 samples. This distance will be used as concatenation cost during synthesis.

### *B. Synthesis*

Fig. 4 shows the steps of the synthesis (speech decoding) stage. The input are the parameters obtained during encoding (F0, gain, rt0 indices and phoneme code) and the codebook of pitch-synchronous residuals. For each parameter set, a 25 ms frame is built with 5 ms shift.

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 pitch is searched from the codebook. We apply target cost and concatenation cost with hand-crafted weights, similarly to unit selection speech synthesis [16]. The target cost is the squared difference among the parameters of the current frame and the parameters of those elements in the codebook, which belong to the same phoneme. The concatenation cost shows the similarity of codebook elements of the same phoneme to each other. When a suitable codebook element is found, its fundamental period is set to the target pitch by either zero padding or deletion. Next, a 25 ms residual frame (400 samples at 16 kHz) is created by overlap-adding the Hanning-windowed residual period. The initial delay is set according to the fixed frame size to preserve the fundamental periods. Finally, the energy of the frame is set using the gain parameter.

The whole residual signal is built by concatenating the frames. Synthesized speech is obtained from the residual signal with MGC-based filtering using the MGLSA digital filter.

## III. EVALUATION

During synthesis, 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.

The total target cost can be written as:

$C_{total} = w_1 C_{concatenation}^2 + w_2 C_{target}^2$ or similarly

$C_{total'} = w_1 / w_2 C_{concatenation}^2 + C_{target}^2$

We experimented with the $w_1 / w_2$ setting. 5-5 short sentences from four Hungarian speakers (two female and two male, denoted F1, F2, M1 and M2) were selected. They represented the typical pitch range of adult speech. First these signals were resampled to 16 kHz, mono, 16 bit in order for the experiment. The sentences were resynthesized with five weight values:

$w_1 / w_2 = \{0.01, 0.1, 1, 10, 100\}$

In this way, 5 x 4 x 5 = 100 utterances were obtained, which were compared to the original speech sentence: the similarity was judged by the first author in terms of naturalness, creakiness, timbre, intelligibility. If the sentence was found to be similar enough to the original, a '+' judgment was given, while the other case was indicated by a '-'.

Table 1 shows the results of this simple evaluation summarizing the four speakers and five sentences. For most of the sentences, the *1* or *10* values of the $w_1 / w_2$

TABLE I.
SIMILARITY OF RESYNTHESIZED SENTENCES TO THE ORIGINAL.

| | $w_1 / w_2$ | | | | |
|---|---|---|---|---|---|
| **SentID** | **0.01** | **0.1** | **1** | **10** | **100** |
| **F1_1** | + | + | + | + | + |
| **F1_2** | - | - | + | - | - |
| **F1_3** | - | - | - | + | + |
| **F1_4** | + | + | + | - | - |
| **F1_5** | - | - | + | + | - |
| **F2_1** | - | - | + | + | + |
| **F2_2** | - | - | + | + | - |
| **F2_3** | - | - | + | + | + |
| **F2_4** | - | + | + | + | - |
| **F2_5** | - | + | + | - | - |
| **M1_1** | + | + | + | + | + |
| **M1_2** | - | + | - | - | - |
| **M1_3** | - | - | + | + | - |
| **M1_4** | + | + | + | + | + |
| **M1_5** | - | - | + | + | - |
| **M2_1** | - | - | - | + | + |
| **M2_2** | - | - | - | - | - |
| **M2_3** | + | + | - | - | - |
| **M2_4** | + | + | + | + | - |
| **M2_5** | + | + | + | + | + |
| **Sum+** | **7** | **10** | **15** | **14** | **8** |

weight were preferred, therefore we chose an equal weight for the target and concatenation cost ($w_1 / w_2 = 1$).

During the evaluation we found that the method is more suitable for female voices to create resynthesized speech than for male voices. This might be caused by the higher pitch of female voices. In the specific case of sentence M2_2 none of the resynthesized versions were found to be similar enough to the original sentence.

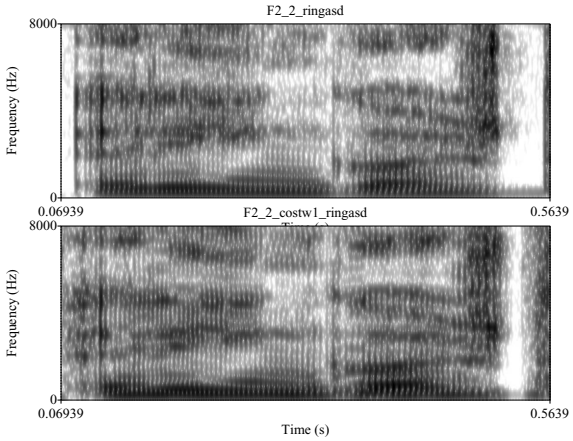


Figure 5. Resynthesis of the speech signal 'ringasd' by speaker F2. Sample of good quality resynthesis.

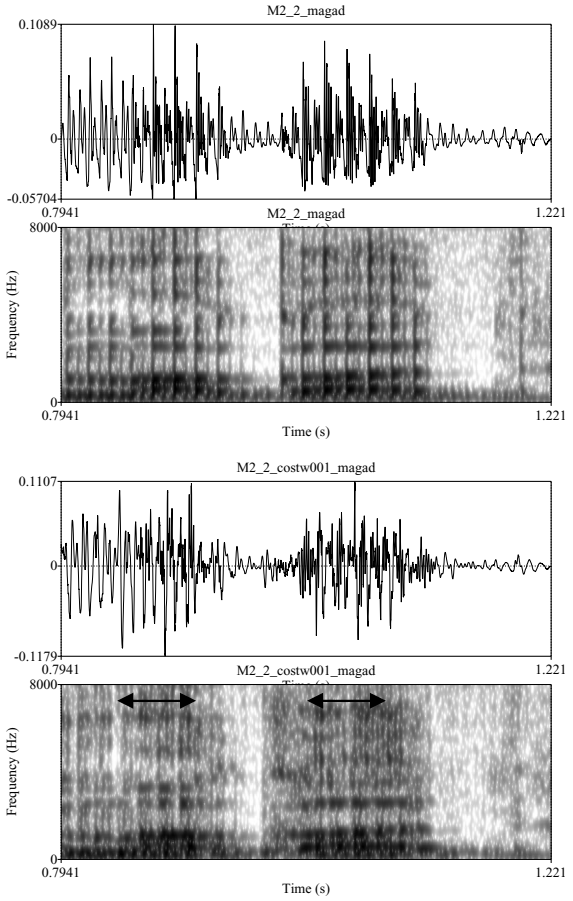


Figure 6. Resynthesis of the speech signal 'magad' by speaker M2. Samples of the irregularities found in resynthesis are marked with horizontal arrows.

## IV. Discussion and Conclusions

In most cases of the evaluation in Section III, the analysis and synthesis resulted in good quality speech.

Fig. 5 shows and example of the spectrogram of the word 'ringasd' uttered by speaker F2 (top) and its resynthesized version (bottom). Visually, the two spectrograms are similar, and perceptually the whole original sentence and its resynthesized version are close to each other.

However, the evaluation revealed some imperfection as well: in several cases large irregularities were found in the resynthesized speech signal. Fig. 6 shows the waveform and spectrogram of the word 'magad' uttered by speaker M2 (top) and its resynthesized version (bottom). This figure shows that the resynthesized signal contains unwanted low-frequency amplitude modulation, which is audible and has a strong 'creaky'-like voice. The original sound has very low pitch and it is heard as 'hoarse' when listening to it. The resynthesized signal does not have clear pitch periods, the waveform looks like unvoiced speech. It is possible that our method is not suitable to resynthesize speech with such properties. Note that creaky voice synthesis is a new topic and includes several challenges [17].

During synthesis, our method modifies the pitch of the excitations by zero padding or deletion instead of resampling. In [11], Drugman et al. use resampling 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.

The proposed algorithm uses unit selection for excitation selection similarly to [12] and [13]. However, we build a phoneme-specific codebook, in which the codebook elements of different phonemes are handled separately. Theoretically (according to the source-filter separation) the excitation is independent from the final sound, but in practice it was found that the residual of LPC analysis has phoneme specific characteristics, that is why we chose to use a phoneme-dependent codebook.

We have not performed yet a more detailed perceptual evaluation of the resynthesized sentences. An evaluation with naïve listeners is planned as future work.

This novel analysis / synthesis excitation model can be used in hidden Markov-model based speech synthesis by replacing the simple vocoder of HTS. Initial experiments show that the model is suitable for machine learning in HTS. This way, synthesized speech will become better in terms of voice quality. By further improving the excitation model, we will be able to synthesize different voice qualities (e.g. breathy, whispered) as well.

Cognitive Infocommunications can gain from better speech-driven human-machine interfaces, as they provide a natural communication modality between infocommunication systems and users.

## References

[1] P. Baranyi and A. Csapó, "Cognitive infocommunications: Coginfocom," Computational Intelligence and Informatics, 11th Int. Symposium on Computational Intelligence and Informatics, pp. 141-146, 2010.

[2] Zen, H., Nose, T., Yamagishi, J., Sako, S. Masuko, T., Black, A.W., Tokuda, K. 2007. The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6.

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

[4] SPTK working group, 2011. Reference Manual for Speech Signal Processing Toolkit Ver. 3.5, December 25, 2011.

[5] Cabral, J. P., 2010. HMM-based Speech Synthesis using an Acoustic Glottal Source Model, Ph.D. Thesis, CSTR, University of Edinburgh, UK.

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

[7] Erro, D.,Sainz, I., Navas, E., Hernaez, I. Improved HNM-based Vocoder for Statistical Synthesizers, Proc. Interspeech, pp. 1809-1812, 2011.

[8] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, An excitation model for HMM-based speech synthesis based on residual modeling, Proc. ISCA SSW6, Aug. 2007.

[9] Zhengqi Wen, Jianhua Tao, Amplitude Spectrum based Excitation Model for HMM-based Speech Synthesis, Interspeech 2012, in press.

[10] Drugman, T., 2011. Advances in Glottal Analysis and its Applications, PhD Thesis, University of Mons, Belgium.

[11] T.Drugman, T.Dutoit, The Deterministic plus Stochastic Model of the Residual Signal and its Applications, IEEE Transactions on Audio, Speech and Language Processing, vol. 20, Issue 3, pp. 968-981, March 2012.

[12] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. Alku, P. 2011.: HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Transactions on Audio, Speech & Language Processing 19(1): 153-165.

[13] Tuomo Raitio, Antti Suni, Hannu Pulakka, Martti Vainio, Paavo Alku: Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. ICASSP 2011: 4564-4567.

[14] The Snack Sound Toolkit [online], http://www.speech.kth.se/snack/.

[15] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit, Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review, IEEE Transactions on Audio, Speech and Language Processing, vol. 20, Issue 3, pp. 994-1006, March 2012.

[16] Hunt, A.J. and Black, A.W., Unit selection in a concatenative speech synthesis system using a large speech database, ICASSP 1996, pp. 373 – 376.

[17] T.Drugman, J.Kane, C.Gobl, Modeling the Creaky Excitation for Parametric Speech Synthesis, Interspeech 2012, in press.