http://dx.doi.org/10.35596/1729-7648-2020-18-2-23-29

*Оригинальная статья*
*Original paper*

UDC 004.934.5

# AN EFFICIENT SPEECH GENERATIVE MODEL BASED ON DETERMINISTIC/STOCHASTIC SEPARATION OF SPECTRAL ENVELOPES

MOSTAFA TAHA, ELIAS S. AZAROV, DENIS S. LIKHACHOV, ALEXANDER A. PETROVSKY

*Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)*

*Submitted 03 June 2019*



**Abstract.** The paper presents a speech generative model that provides an efficient way of generating speech waveform from its amplitude spectral envelopes. The model is based on hybrid speech representation that includes deterministic (harmonic) and stochastic (noise) components. The main idea behind the approach originates from the fact that speech signal has a determined spectral structure that is statistically bound with deterministic/stochastic energy distribution in the spectrum. The performance of the model is evaluated using an experimental low-bitrate wide-band speech coder. The quality of reconstructed speech is evaluated using objective and subjective methods. Two objective quality characteristics were calculated: Modified Bark Spectral Distortion (MBSD) and Perceptual Evaluation of Speech Quality (PESQ). Narrow-band and wide-band versions of the proposed solution were compared with MELP (Mixed Excitation Linear Prediction) speech coder and AMR (Adaptive Multi-Rate) speech coder, respectively. The speech base of two female and two male speakers were used for testing. The performed tests show that overall performance of the proposed approach is speaker-dependent and it is better for male voices. Supposedly, this difference indicates the influence of pitch highness on separation accuracy. In that way, using the proposed approach in experimental speech compression system provides decent MBSD values and comparable PESQ values with AMR speech coder at 6,6 kbit/s. Additional subjective listening testsdemonstrate that the implemented coding system retains phonetic content and speaker's identity. It proves consistency of the proposed approach.

**Keywords:** speech generative model, harmonic plus noise model, speech analysis, speech coding.

**Conflict of interests.** The authors declare no conflict of interests.

**For citation.** Taha M., Azarov E.S., Likhachov D.S., Petrovsky A.A. An efficient speech generative model based on deterministic/stochastic separation of spectral envelopes. Doklady BGUIR. 2020; 18(2): 23-29.

## Introduction

Contemporary speech synthesis algorithms have made a great leap forward due to developing of artificial neural networks. Now it is possible to synthesize high-quality speech using WaveNet algorithm [1] or one of the wide range of similar solutions [2, 3]. One of the drawbacks, however, is high computational complexity of these algorithms that can reach tens of billions of floating-point operations per second (GFLOPS) which requires using high-end GPUs. A much more efficient solution LPCNet has been reported recently [4], however, it requires around 2.8 GFLOPS which is still very high compared to conventional parametric methods. The crucial part of the synthesis is the problem of transforming amplitude spectrum or amplitude spectral envelope into waveform. The classical solution to the problem was proposed by Griffin and Lim known as Griffin/Lim

algorithm [5]. The algorithm has less computational requirement, however, it is very sensitive to hop size, requires amplitude spectrum and does not work with amplitude spectral envelopes.

In the present paper an efficient algorithm is proposed for speech waveform generation from its amplitude spectral envelopes. The algorithm utilizes Harmonic plus Noise Model (HNM) and statistical deterministic/stochastic separation of the envelopes. The main idea behind the approach originates from the fact that speech signal has a determined spectral structure that is statistically bound with deterministic/stochastic energy distribution in the spectrum. The separation function is estimated through a training procedure that involves fitting of data obtained through instantaneous harmonic analysis and short time spectrum.

The flexible HNM synthesis where the deterministic part accounts for the periodic (harmonic) structure of the signal and the stochastic models its noise part was presented in [6, 7]. The model has been successfully applied to a number of different speech applications: speech coding, text to speech synthesis, voice conversion and other. The main benefits of the model can be shortly listed as follows:

– explicit control over prosodic features of the speech that is a benefit in text to speech synthesis and voice conversion;

– efficiency of the representation;

– high quality speech reconstruction.

However, harmonic parameters estimation is pitch-based. It means that the method is extremely sensitive to pitch estimation errors. Pitch estimation itself is a fundamental problem of speech analysis that does not have an ultimate solution yet. The estimation is prone to errors especially for transitional (partially voiced) speech sounds. Inaccuracy of harmonic parameters values causes audible artifacts in reconstructed speech. Majority of the mentioned speech processing applications require estimation of harmonic spectral envelopes rather than parameters of individual harmonics. This is also true for stochastic part of the signal that is usually estimated as difference between source and harmonic signals in time domain and then represented by spectral envelopes (e. g. using all-pole filter).

The performance of the proposed speech generation algorithm is evaluated using an experimental low bitrate wide-band speech coder. The quality of reconstructed speech is rated using objective and subjective methods.

## Overview of the method

Deterministic/stochastic spectrum separation of a speech signal is carried out using separation function that is determined through a training procedure using a speech data corpus. The training process illustrated in Fig. 1 involves the following steps:

1. Speech data is analyzed using instantaneous harmonic analyzer [8] and separated into deterministic and stochastic parts.

2. Harmonic spectral envelopes are calculated using interpolation from instantaneous harmonic parameters; noise spectral envelopes are calculated from stochastic part using short-time Fourier transform (STFT).

3. Short-time spectra are calculated from the source signal and transformed into spectral envelopes.

4. The separation function is estimated that minimizes quadratic error of separated spectra.

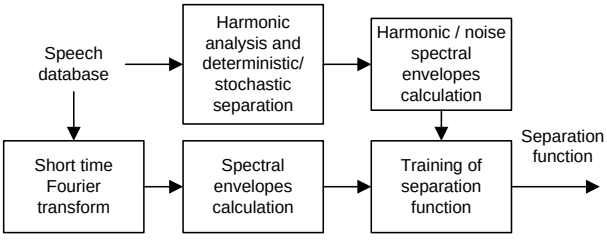


**Fig. 1.** Training of separation function

During harmonic analysis speech frames are classified either as voiced or unvoiced. Unvoiced frames are modeled as pure stochastic signals.

The spectrum separation process illustrated in Fig. 2 consists of the following steps:

1. Short-time Fourier transform of the frames are calculated.
2. Short-time spectrum envelopes are calculated from short-time spectra.
3. Spectrum envelopes are separated into harmonic and noise envelopes using separation function.

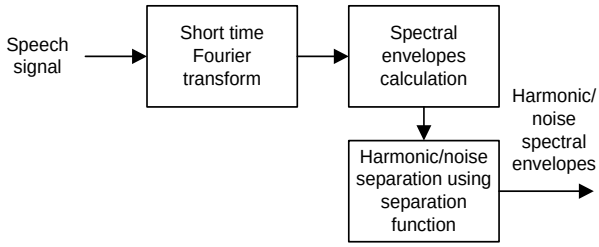


**Fig. 2.** Stochastic/deterministic separation of speech spectrum

The spectrum separation process is quite simple while the training process requires implementation of complex algorithms: pitch estimation, harmonic analysis and training.

### Estimation of instantaneous harmonic parameters

The hybrid deterministic/stochastic model assumes that the signal $s(n)$ can be expressed as the sum of its periodic and noise parts:

$$s(n) = \sum_{k=1}^{K} MAG_k(n)\cos\varphi_k(n) + r(n),$$

where $MAG_k(n)$ – the instantaneous magnitude of the $k$-th sinusoidal component, $K$ is the number of components, $\varphi_k(n)$ is the instantaneous phase of the $k$-th component and $r(n)$ is the stochastic part of the signal. Instantaneous phase $\varphi_k(n)$ and instantaneous frequency $f_k(n)$ are related as follows:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi \cdot f_k(i)}{F_s} + \varphi_k(0),$$

where $F_s$ is sampling frequency and $\varphi_k(0)$ is the initial phase of the $k$-th component. The harmonic model states that frequencies $f_k(n)$ are integer multiples of the fundamental frequency $f_0(n)$ and can be calculated as: $f_k(n) = kf_0(n)$.

The harmonic model is often used in speech coding since the instantaneous harmonic parameters $MAG_k(n)$, $f_k(n)$ and $\varphi_k(0)$ represent voiced speech in a highly efficient way.

Instantaneous harmonic parameters are calculated using the technique based on analysis filters [3]. Filter bands are recalculated for each frame of the signal using estimated pitch values. Given a speech frame, multiplied by a window function $s(n)$, $0 \le n \le N-1$ and a filter passband specified by center frequency contour $F_c(n)$ and bandwidth $2F_\Delta$, instantaneous magnitude $MAG(n)$, phase $\varphi(n)$ and frequency $f(n)$ are calculated as:

$$MAG(n) = \sqrt{A^2(n) + B^2(n)},$$

$$\varphi(n) = \arctan\left(\frac{-B(n)}{A(n)}\right),$$

$$f(n) = \frac{\varphi(n+1) - \varphi(n)}{2\pi} F_s,$$

where

$$A(n)=\sum_{i=0}^{N-1}\frac{s(i)F_S}{2\pi(n-i)F_\Delta}\sin\left(\frac{2\pi(n-i)}{F_S}F_\Delta\right)\cos\left(\frac{2\pi}{F_S}\varphi_C(n,i)\right),$$

$$B(n)=\sum_{i=0}^{N-1}\frac{-s(i)F_S}{2\pi(n-i)F_\Delta}\sin\left(\frac{2\pi(n-i)}{F_S}F_\Delta\right)\sin\left(\frac{2\pi}{F_S}\varphi_C(n,i)\right),$$

$$\varphi_c(n,i)=\begin{cases}\sum_{j=n}^{i}F_c(j), & n<i\\ -\sum_{j=i}^{n}F_c(j), & n>i\\ 0, & n=i.\end{cases}$$

Central frequencies of the filter bands are calculated as instantaneous pitch multiplied by number of the respective harmonic $F_c^k(n)=kf_0(n)$:

$$f_0(n)=\sum_{i=0}^{K}\frac{f_i(n)MAG_i(n)}{(i+1)\sum_{j=0}^{k}MAG_j(n)}.$$

The procedure goes from the first harmonic to the last, adjusting fundamental frequency at every step. The fundamental frequency recalculation formula can be written as follows:

The fundamental frequency values become more precise while moving up the frequency range. It allows making proper analysis of high order harmonics with significant frequency modulations.

Harmonic envelopes are calculated from instantaneous harmonic parameters using linear interpolation. The deterministic part of the signal is synthesized using estimated harmonic parameters and subtracted from the source signal frame in order to obtain residual. The residual (stochastic) part of the signal $r(n)$ is parameterized as a bark-band noise. The noise envelopes are calculated as energies of the signal in bark subbands. After applying the parameterization technique, the speech signal is represented as a set of instantaneous harmonic envelopes, short-time noise envelopes and a pitch contour.

## Estimation of energy separation function

Let us denote spectral envelope vector estimated through STFT as $E(e_1,e_2,...,e_m)$ and harmonic/noise envelopes estimated using harmonic analysis technique as $H(h_1,h_2,...,h_m)$ and $V(v_1,v_2,...,v_m)$, respectively. Assuming that the separation function can be expressed in terms of linear regression, it can be estimated by use of least-squares method minimizing the following functions:

$$Er_h=\sum_{l=1}^{L}(E_l^T\alpha-H_l)^2,\ \ Er_n=\sum_{l=1}^{L}(E_l^T\beta-V_l)^2,$$

where $Er_h$ and $Er_n$ – harmonic and noise estimation errors, respectively, $L$ – training sequence length, $\alpha$ and $\beta$ – $m\times m$ matrices of regression coefficients for harmonic and noise parts, respectively; $T$ stands for transposition. Evaluation of $\alpha$ and $\beta$ matrices is carried out independently.

The envelopes $E_l$ are estimated as decimal logarithm of the energy values. The length of the vector $E$ corresponds to the number of spectrum bands where the energy of the signal is calculated. The experiments show that the best result is achieved when these bands are not uniform. A relatively good result was obtained for bark scale – the average squared error was 0.07 per harmonic/noise vector value for multi-speaker speech database. Fig. 3 illustrates an example of bark-band spectrum envelopes separation into deterministic and stochastic parts. The possibility of speech reconstruction from its bark-band energy envelopes and pitch contour can be especially useful in speech coding.

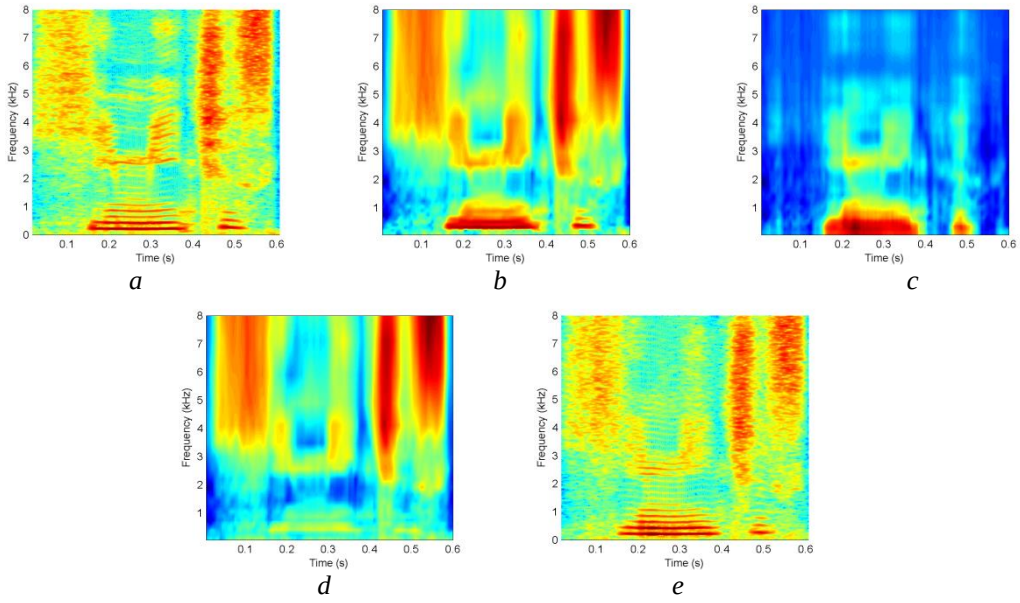


**Fig. 3.** An example of speech signal reconstruction from its bark-band energy envelopes using spectrum separation: source speech signal (*a*); bark-band energy envelopes (*b*); estimated harmonic envelopes (*c*); estimated noise envelopes (*d*); reconstructed signal (*e*)

## Experimental results

In order to evaluate applicability of the method to speech compression experimental speech coding systems for narrow- and wide-band speech were implemented. The encoding/decoding processes are illustrated in Fig. 4. The distinguishing feature of the coding scheme is that harmonic/noise separation is done at the decoding phase. The input of the decoder consists of quantized bark-band energy values and pitch contour. The energy values are obtained using 1024-point short-time Fourier transform (512-point for narrow-band version) and combined in bark-band envelope vectors. The envelopes are calculated with 10ms time offset. Pitch values are estimated using analysis filters as was reported in [3].

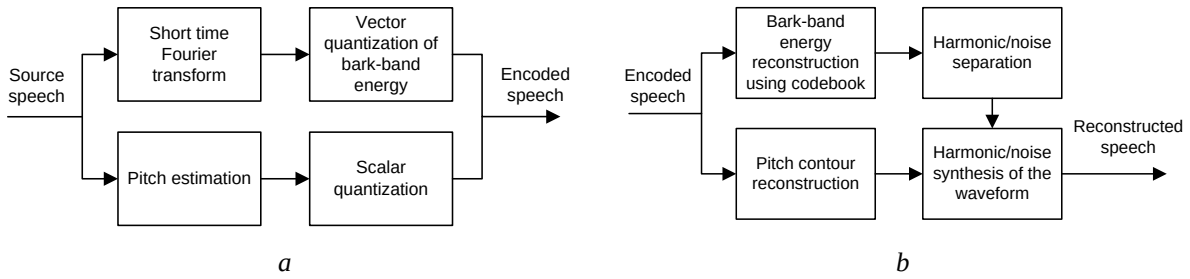


**Fig. 4.** Speech compressing scheme: encoding (*a*); decoding (*b*)

Quantization of energy values is made using common vector quantization technique. Each wide-band energy vector $E(e_1,e_2,...,e_{21})$ is split into three separate vectors $E_1(e_1,...,e_9)$, $E_2(e_{10},...,e_{17})$ and $E_3(e_{18},...,e_{21})$ that are quantized using different codebooks. The codebooks were trained on multi-speaker speech material (with duration about 10 minutes) through standard *K*-means algorithm. The sequence of energy envelopes is reconstructed in the decoder and their harmonic/noise separation is carried out using separation function. The function was trained using the same training speech material. The coding scheme is very efficient and can be implemented using a non-uniform filter bank [4].

Performance of the proposed speech coder was evaluated using objective measures of speech quality. The two following quality characteristics were calculated: Modified Bark Spectral Distortion (MBSD) and Perceptual Evaluation of Speech Quality (PESQ). Proposed narrow-band solution was

compared with MELP (Mixed Excitation Linear Prediction) speech coder and wide-band version was compared with AMR (Adaptive Multi-Rate) speech codec. The speech base used for testing contained sentences pronounced by four different speakers (two male and two female speakers whose speech was not used during training). The average obtained values are presented in Tables 1, 2 (proposed speech coding system is labeled as 'joint coding').

**Table 1.** Objective quality of narrow-band speech coding, $F_S$ = 8 kHz

| Coder / bitrate (kbit/s) | MBSD | | PESQ | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| MELP / 2.4 | 0.19 | 0.25 | 3.04 | 2.98 |
| Joint coding / 2.4 | 0.81 | 0.93 | 2.40 | 2.31 |

**Table 2.** Objective quality of wide-band speech coding, $F_S$ = 16 kHz

| Coder / bitrate (kbit/s) | MBSD | | PESQ | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| AMR / 6,6 | 0.51 | 0.58 | 3.42 | 3.47 |
| Joint coding / 6.6 | 0.26 | 0.49 | 3.02 | 2.75 |
| Joint coding / 4.4 | 0.32 | 0.59 | 2.93 | 2.69 |
| Joint coding / 2.4 | 0.81 | 0.97 | 2.31 | 2.25 |

Considering that accuracy of deterministic/stochastic spectrum decomposition might be speaker-depended average results are calculated for male and female voices separately.

The objective quality tests show that overall performance of the proposed approach is speaker-dependent. It can be seen from the presented results demonstrating that the quality of reconstructed speech is better for male voices. Supposedly, this difference indicates influence of pitch highness on separation accuracy. The experimental speech compression system provides decent MBSD values and comparable PESQ values with AMR codec at 6,6 kbit/s (it is the lowest bitrate possible for a free AMR encoder used in the experiments).

Additional subjective listening tests were carried out as well. The quality of signal reconstruction was compared in the following pairs: AMR 6.6 – joint coding 6.6 and MELP 2.4 – joint coding 2.4. Twenty different sentences were chosen and played back in random order. Five listeners were asked to rate which sentence from the pair sounded more natural. The proposed encoder was chosen in about 40 percent of cases for wide-band speech and in 35 percent of cases for narrow-band speech. All listeners approved that the implemented coding system retains phonetic content and speaker's identity at every bitrate that proves consistency of the proposed approach.

## Conclusion

A model for speech generation from its spectral amplitude envelopes has been proposed. The model involves deterministic/stochastic decomposition that is carried out using separation function without conventional harmonic analysis. The separation function is represented as a matrix of linear regression coefficients and evaluated using least-squares method. Training sequence contains harmonic and noise envelopes estimated via instantaneous harmonic analysis. The method has been experimentally applied to speech coding. The quality of reconstructed signal has been rated using objective and subjective methods. The obtained results show high potential of the presented approach.

## Authors contribution

Taha M. realized the speech modeling.
Azarov E.S. designed the structure of the proposed approach.
Likhachov D.S. developed the experimental speech coding systems.
Petrovsky A.A. performed the statement of the problem and coordination.

## References

1. A. van den Oord, Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A., Kavukcuoglu K. WaveNet: A generative model for raw audio, arXiv:1609.03499, 2016.
2. Shen J., Pang R., Weiss R. J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhang Y., Wang Y., Skerrv-Ryan R. "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018:4779-4783.
3. Arik S., Diamos G., Gibiansky A., Miller J., Peng K., Ping W., Raiman J., and Zhou Y. Deep voice 2: Multi-speaker neural text-to-speech. arXiv:1705.08947, 2017.
4. Valin J.-V., Skoglund J. LPCNet: Improving neural speech synthesis through linear prediction, arXiv:1810.11846
5. Griffin D., Lim J. A new model-based speech analysis/synthesis system. In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1985;10:513-516.
6. Laroche J., Stylianou Y., Moulines E.. HNS: Speech modification based on a harmonic+noise model. Proceedings of the ICASSP-93 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993;2:550-553.
7. Serra X. *Musical sound modeling with sinusoids plus noise. Musical Signal Processing* (C. Roads, S. Popea, A. Picialli, G. De Poli Eds.). Swets & Zeitlinger Publishers; 1997.
8. Azarov E., Petrovsky A. Instantaneous harmonic analysis for vocal processing. Proceedings of DAFx-09. Como, Italy, September 14; 2009.

## Information about the authors

Taha M., Master of Sciences, PhD student of Computer Engineering Department of Belarussian State University of Informatics and Radioelectronics.
Azarov E.S., D.Sci., Professor of Computer Engineering Department of Belarussian State University of Informatics and Radioelectronics.
Likhachov D.S., PhD, Associate Professor of Computer Engineering Department of Belarussian State University of Informatics and Radioelectronics.
Petriovsky A.A., D.Sci., Professor of Computer Engineering Department of Belarussian State University of Informatics and Radioelectronics.

## Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki str., 6,
Belarussian State University of Informatics and Radioelectronics
tel. +375172938805;
e-mail: likhachov@bsuir.by
Likhachov Denis Sergeevich