# Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis

Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics (BME TMIT), Budapest, Hungary

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh,

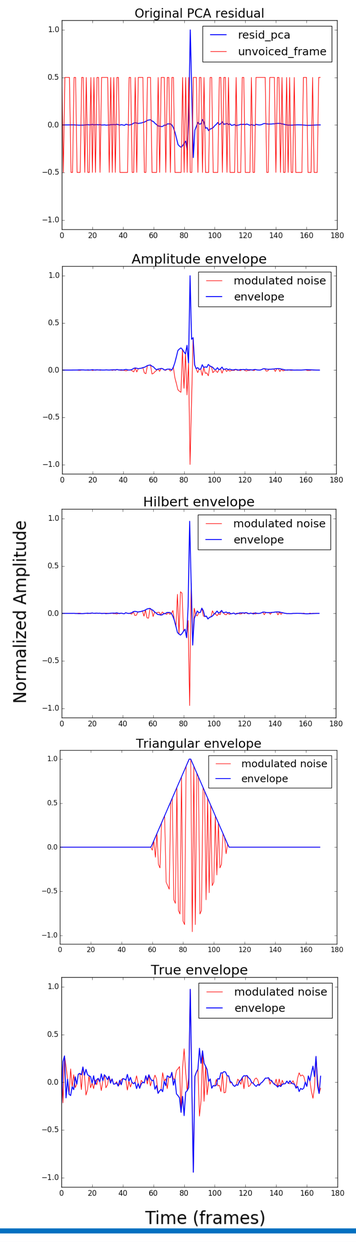
*{malradhi,csapot,nemeth}@tmit.bme.hu*

## 1. Introduction

- **vocoder problems**
  - buzziness
  - real-time processing
- **fundamental frequency (F0)**
  - continuous in voiced regions
  - discontinuous in unvoiced regions
  - hard to model boundaries between voiced and unvoiced segments
- **maximum voiced frequency (MVF)**
  - excitation parameter
  - separate the voiced and unvoiced components
- **standard Mel-Generalized Cepstral analysis (MGC)**
- **noise component**
  - according to [1], not accurately modeled even in the widely used STRAIGHT vocoder
- **goal of this paper**
  - extension of a continuous residual-based vocoder for statistical parametric speech synthesis [2] for advanced modeling of the noise excitation
  - **shaping the high-frequency component by adding envelope modulated noise to the voiced excitation**
  - evaluate four approaches for estimating the time envelope of the speech residual signal

## 2. Methods

- **Continuous vocoder (baseline [2])**
  - continuous F0 model [3] to decrease the disturbing effect of creaky voice
    - standard autocorrelation
    - no voiced/unvoiced decision
    - Kalman smoothing-based interpolation
  - MVF to model the voiced/unvoiced characteristics of sounds [4]
- **Time-domain envelope estimation (see Fig. 2)**
  - Amplitude envelope
    - filtering the absolute value of the voiced frame
  - Hilbert envelope
    - based on the Hilbert transform technique
    - taking the magnitude of the analytical signal
  - Triangular envelope
    - used three parameters as it assumes the triangular is symmetric
  - True envelope [5]
    - the original spectrum signal and the current cepstral representation is maximized
    - weighting factor makes the convergence more closely to the natural speech (see Fig. 3). In practice, the most successful weighting factor is 10



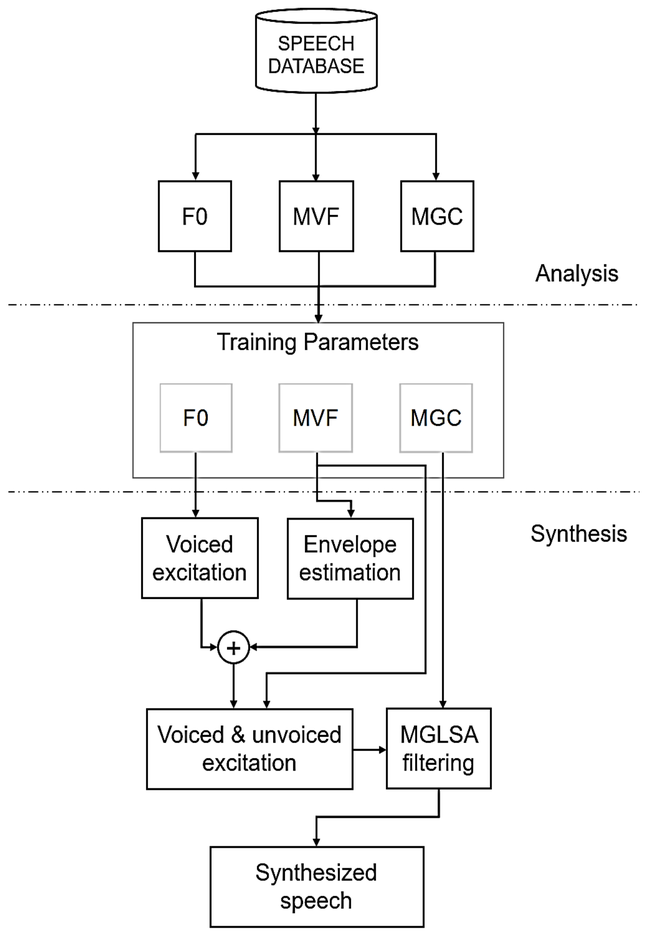




**Figure 1.** *Workflow of the proposed method.*

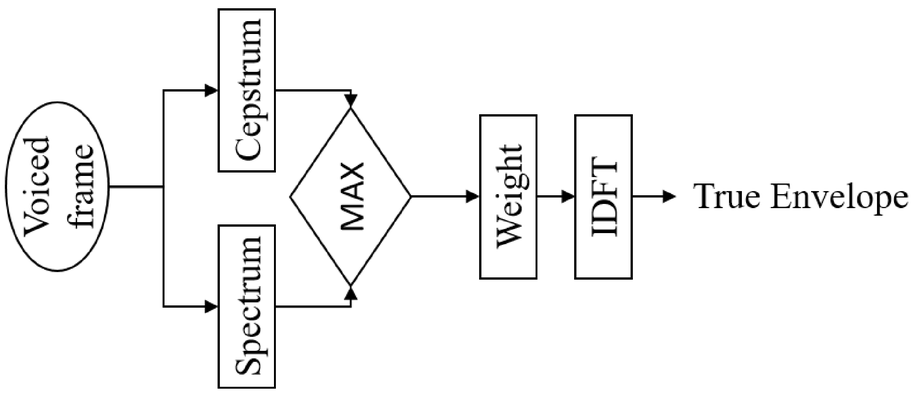



**Figure 3.** *Procedures for estimating the true envelope.*

**Figure 2.** *Illustration of the time envelopes performance*

## 3. Objective evaluation

- Data: from CMU-ARCTIC
  - AWB (Scottish English, male) and SLT (American English, female)
  - 100 sentences analyzed and re-synthesized with all vocoder variants
- Phase Distortion Deviation (PDD)
  - good measure of noisiness, and a strong correlate of the maximum-phase component of the voice source
  - we zeroed out the PDD values below the MVF contour to quantify the noisiness in the higher frequency bands
  - in Fig. 5 the proposed systems have PDD values closer to the natural speech; especially for 'Hilbert' and 'True' envelopes




**Figure 5.** *Mean PDD values by sentence type.*

**Figure 4.** *Phase Distortion Deviation of a natural and vocoded speech samples above the Maximum Voiced Frequency region (sentence: "He made sure that the magazine was loaded, and resumed his paddling.", from speaker AWB). The warmer the color, the bigger the PDD value and the noisier the corresponding time-frequency region.*

## 4. Perceptual evaluation

- Multi-Stimulus test with Hidden Reference and Anchor listening test
- aim: compare natural vs. vocoded sentences (unvoiced component)
- 12 participants (mean age: 38 years) with engineering background
- rate from 0 (highly unnatural) to 100 (highly natural)
- for the male speaker, the vocoder using the Hilbert envelope is slightly better than the baseline system (see Fig. 6)
- samples: http://smartlab.tmit.bme.hu/interspeech2017_vocoder_envelope

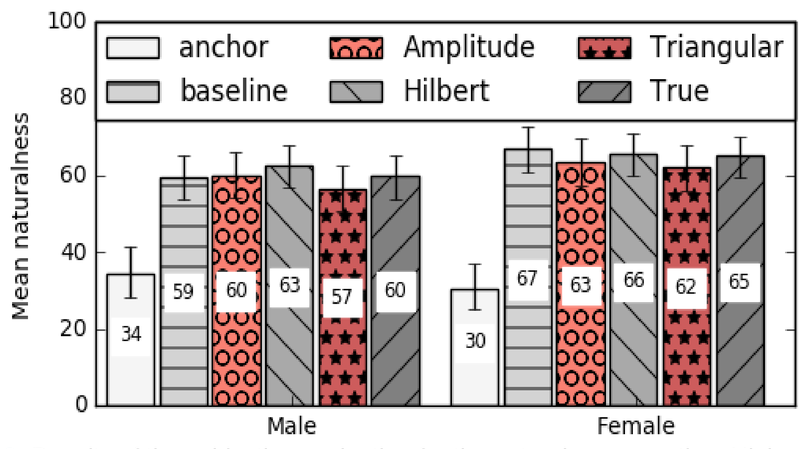



**Figure 6.** *Results of the subjective evaluation for the naturalness question. Higher value means larger naturalness. Error bars show the bootstrapped 95% confidence intervals.*

## 5. Discussion and Conclusion

- this work aims to further control the time structure of the high-frequency noise component in continuous vocoder
- it can be concluded that the Hilbert and True envelopes are the best when combined with the continuous vocoder
- plans of future research involve adding a Harmonics-to-Noise Ratio parameter to the analysis, statistical learning and synthesis steps in order to further reduce the buzziness caused by vocoding

## Key references

[1] G. Degottex, P. Lanchantin, and M. Gales, "A Pulse Model in Log-domain for a Uniform Synthesizer," in Proc. ISCA SSW9, p. 230–236, 2016.

[2] Tamás Gábor Csapó, Géza Németh, Milos Cernak, and Philip N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in EUSIPCO, Budapest, 2016.

[3] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102-105, 2013.

[4] T. Drugman and Y. Stylianou, "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," IEEE Signal Processing Letters, vol. 21, no. 10, p. pp. 1230–1234, 2014.

[5] A.Röbel and X.Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in International Conference on Digital Audio Effects, Madrid, 2005

## Acknowledgements

The research was partly supported by the VUK (AAL-2014-1-183) and the EUREKA / DANSPLAT projects. We would like to thank the listeners participating in the test.