Effects of adding a Harmonics-to-Noise Ratio parameter to a Continuous vocoder

Budapest University of Technology and Economics,
Department of Telecommunication and Media Informatics,
Budapest, Hungary
Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh
{malradhi,csapot,nemeth}@tmit.bme.hu

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)**

MOTIVATION

- **related work**
  - in [1], we proposed a vocoder using continuous F0 in combination with MVF, which was successfully used with hidden Markov models based text-to-speech (TTS).
  - Our previous work has shown the advantages of adding envelope modulated noise to the voiced excitation [2].
- **goal of this paper**
  - propose to add a Harmonics-to-Noise Ratio parameter to the analysis and synthesis steps in order to further reduce the buzziness caused by vocoding.

BASELINE SYSTEM

- **Continuous vocoder (see Fig. 1)**
  - continuous F0 model [3] to decrease the disturbing effect of creaky voice, and easier to handle mixed excitation
  - standard autocorrelation
  - no voiced/unvoiced decision
  - Kalman smoothing-based interpolation
- **MVF to model the voiced/unvoiced characteristics of sounds [4]**
- **Noise component**
  - shaping the high-frequency component by adding envelope modulated noise to the voiced excitation [2]
- **Spectral envelope refinement**
  - using CheapTrick algorithm

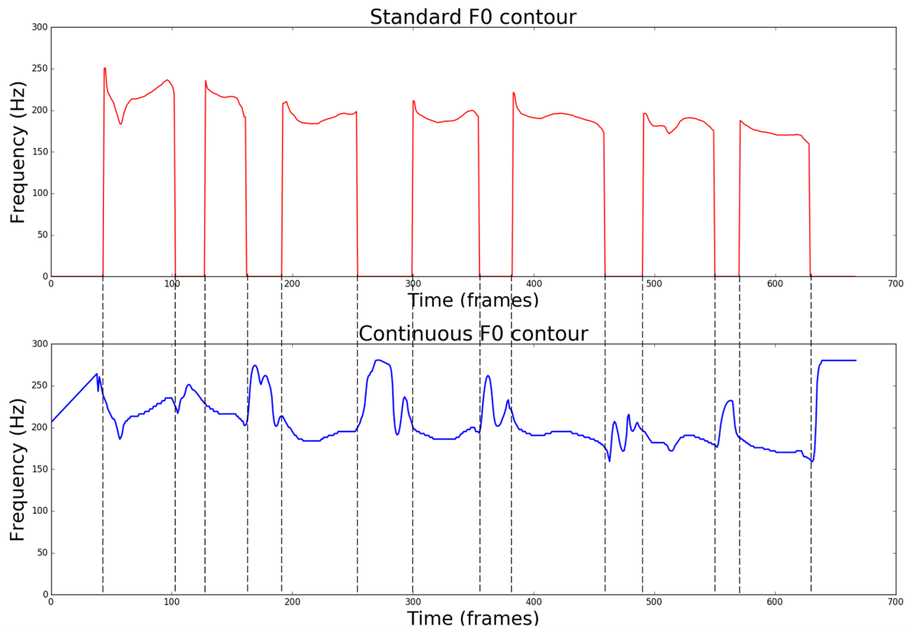

Figure 1. F0 trajectories for the synthesized speech signal using the standard (red), and Continuous algorithms (blue). sentence: "The girl faced him, her eyes shining with sudden fear" from speaker SLT.

PROPOSED SYSTEM: Definition

- show the effect of adding Harmonics-to-Noise Ratio (HNR) as a new excitation parameter to the Continuous vocoder (see Fig. 2)
- according to [5],

$$HNR = 10log_{10}\frac{r'(\tau_{max})}{1-r'(\tau_{max})}$$

- from relative height of the maximum autocorrelation function, the HNR of the voice can be found
- For perfectly periodic sounds, the HNR is positive infinite; while for the noise, it is very low.

PROPOSED SYSTEM: Architecture

- For each frame, a F0 estimate and HNR are calculated
- The voiced and unvoiced signal are added in the ratio suggested by the HNR

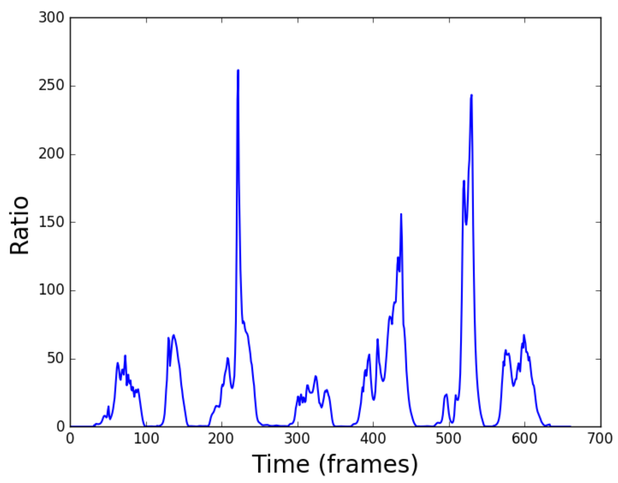

Figure 3. Example of a HNR parameter for the synthesized speech. sentence: "The girl faced him, her eyes shining with sudden fear" from speaker SLT.

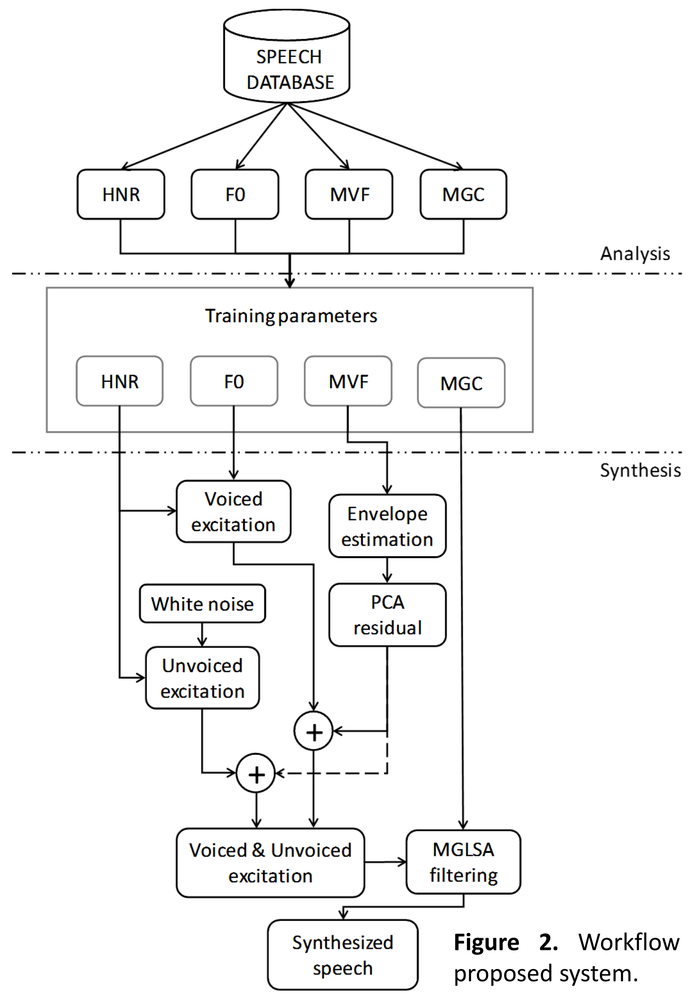

Figure 2. Workflow of the proposed system.

RESULTS AND 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 [6])**
  - good measure of noisiness, and a strong correlate of the maximum-phase component of the voice source
  - in figure 4, the proposed system has PDD values closer to the natural speech
- **Empirical measures [7] (see Table 1)**
  - Perceptual Evaluation of Speech Quality
  - Weighted Spectral Slope
  - Composite Objective Speech Quality
  - Normalized Covariance Metric
- For all empirical metrics, a calculation is done frame-by-frame and a lower value indicates better performance except for the NCM measure that is +1 in natural speech signal.

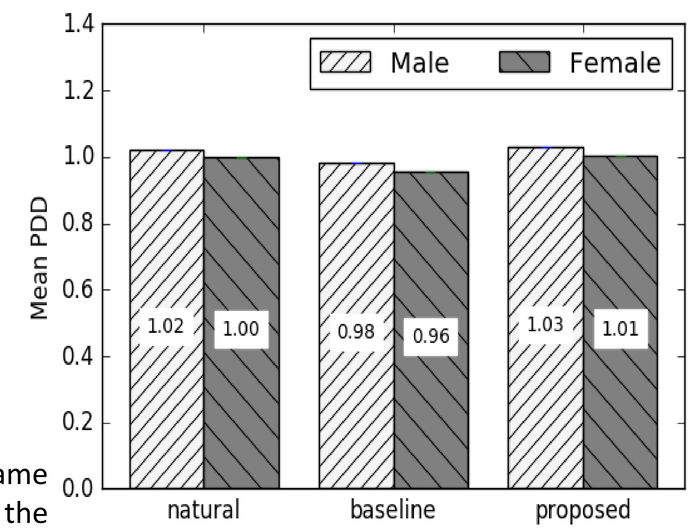

Figure 4. Mean PDD values by sentence type.

Systems	PESQ		WSS		Composite		NCM	
	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB
baseline	2.27	2.16	4.98	4.84	2.70	2.60	0.667	0.701
proposed	1.72	1.73	3.62	3.04	2.33	2.25	0.842	0.885

Table 1. Objective measures for the proposed systems.

DISCUSSION AND CONCLUSION

- this work aims to further reduce the buzziness caused by vocoding
- Experimental results based on objective evaluation have proven that the voice built with the proposed framework gives state-of-the-art speech synthesis performance while outperforming the previous baseline
- plans of future research involve adding the HNR parameter besides the F0, MVF, and MGC to the statistical deep learning based neural networks

REFERENCES

[1] T. G. Csapó, G. Németh, M. Cernak, and P. N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in EUSIPCO, Budapest, pp. 1338-1342, 2016.
[2] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh. Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis. In Proc. Interspeech 2017, Stockholm.
[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, pp. 1230–1234, 2014.
[5] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in proceedings of the Institute of Phonetic Sciences, Netherlands: University of Amsterdam, 1993.
[6] 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.
[7] P. C. Loizou, "Speech Enhancement: Theory and Practice", 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.