MODELING-BY-GENERATION-STRUCTURED NOISE COMPENSATION ALGORITHM FOR GLOTTAL VOCODING SPEECH SYNTHESIS SYSTEM

Min-Jae Hwang[1], Eunwoo Song[1,2], Kyungguen Byun[1] and Hong-Goo Kang[1]

[1]DSP Lab., Yonsei University, Seoul, Korea

[2]NAVER Corp., Seongnam-si, Gyeonggi-do, Korea

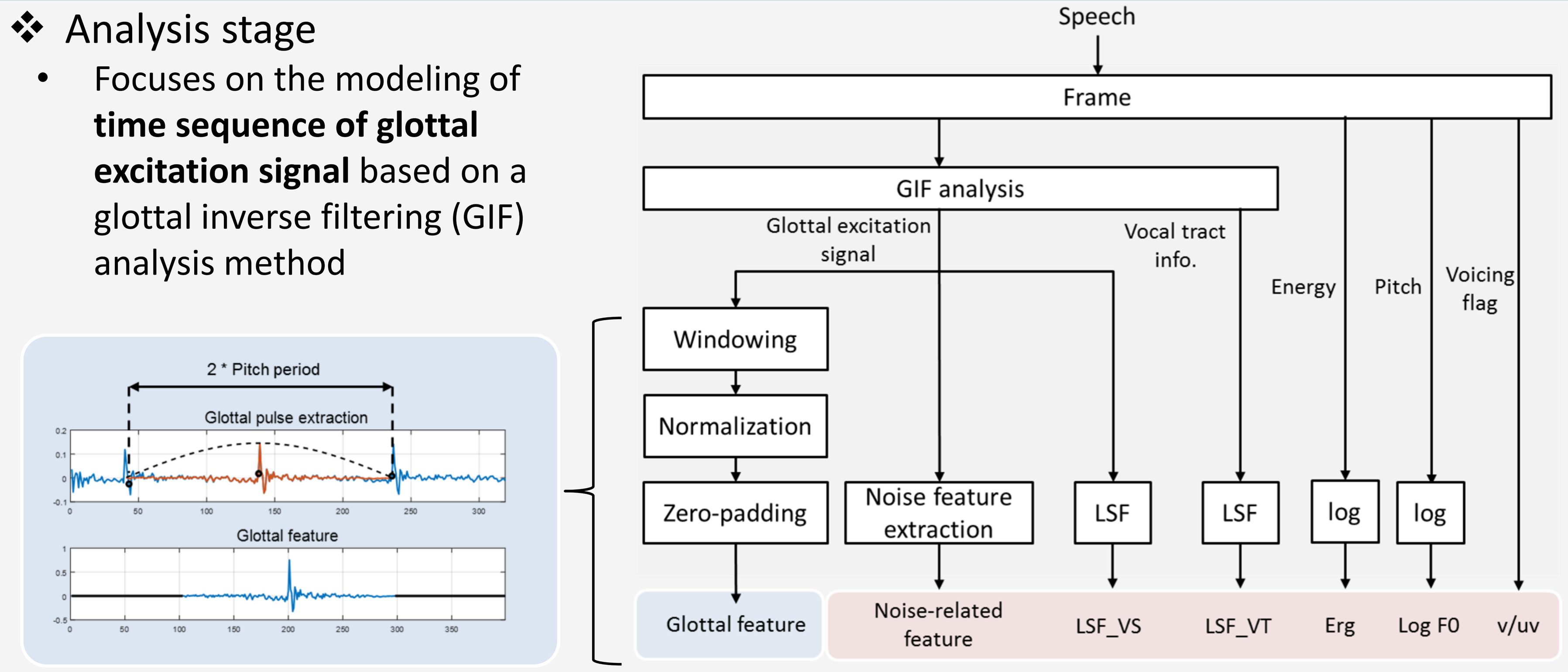
Introduction

- This paper proposes a **modeling-by-generation (MbG)-structured noise compensation method** for a glottal excitation model in a deep learning (DL)-based speech synthesis system.
- A generated glottal signal by the model training process does not faithfully represent noise component, thus the signal can be treated as **harmonic component**.
- In the proposed MbG approach, the **weighted subtraction** between original and generated glottal signals is parameterized into **noise features**, then they trained and generated by **additional DL network**.
- In the synthesis stage, the noise component is **phase-aligned** and **added** to the generated glottal signal.

Glottal vocoding speech synthesis system

❖ Analysis stage

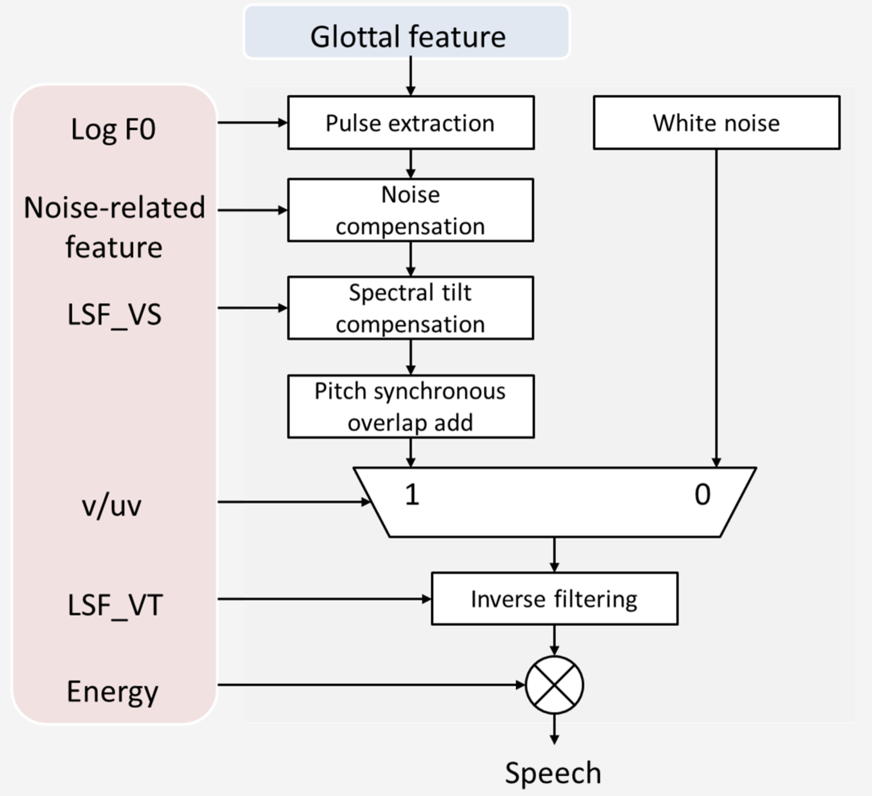
- Focuses on the modeling of **time sequence of glottal excitation signal** based on a glottal inverse filtering (GIF) analysis method



Glottal features (GFs)   Acoustic features (AFs)

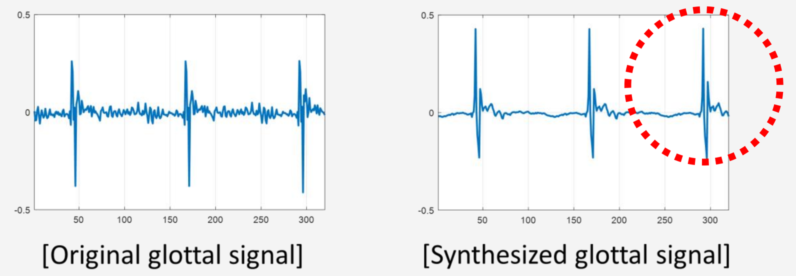
❖ Synthesis stage

- Glottal pulse extraction
- Noise compensation
- Spectral tilt compensation
- Pitch synchronous overlap-add
- Inverse filtering



Noise compensation methods and their limitations

❖ Necessity for introducing the noise compensation process

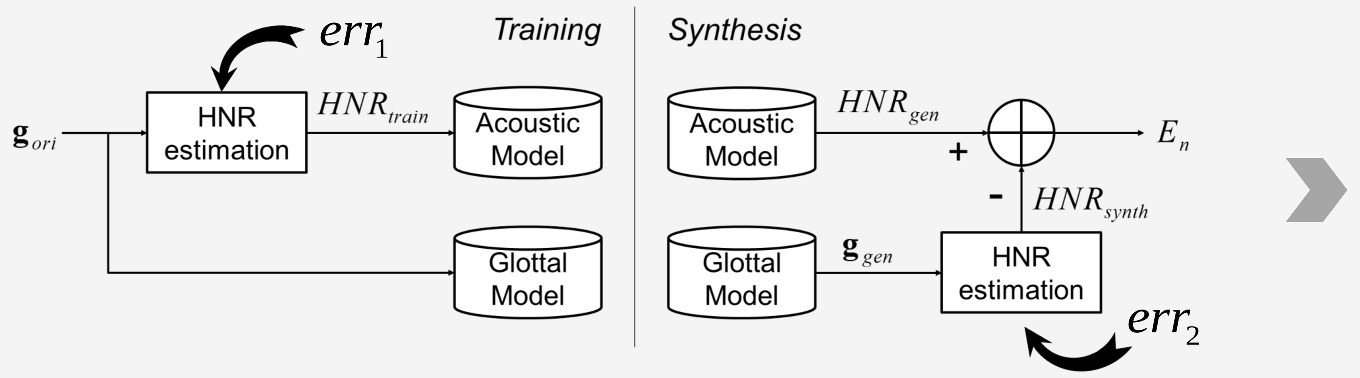


[Original glottal signal]   [Synthesized glottal signal]

Generated glottal signal output with the trained model becomes impulsive

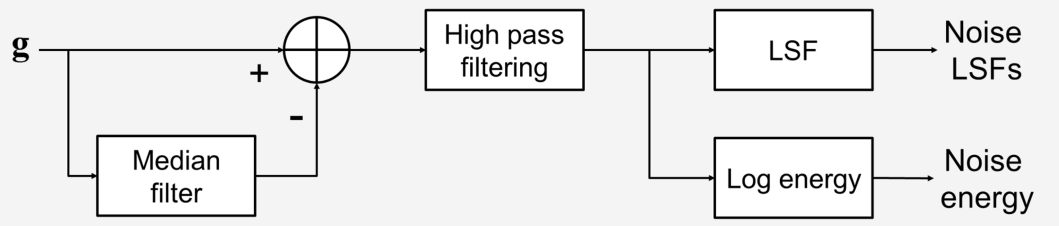
→Unnatural buzzy sound

❖ **HNR-NC**: Harmonic-to-noise ratio (HNR)-based noise compensation [1]

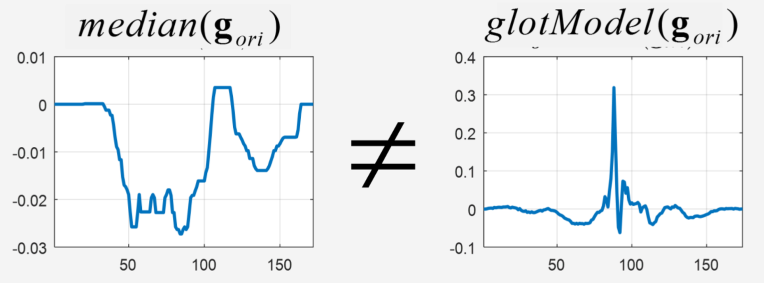


$$E_n = E_{n,ideal} + err_1 + err_2$$

❖ **MF-NC**: Median-filter (MF)-based noise compensation [2]



Bring noisy artifact in the synthesis stage



$median(\mathbf{g}_{ori}) \neq glotModel(\mathbf{g}_{ori})$ ➤ Mismatch in $\mathbf{n}$ ➤ $\mathbf{n}_{synth} = \mathbf{n}_{ideal} + \mathbf{n}_{unwanted}$

Glottal vocoder with MbG-structured noise compensation algorithm

❖ Concept

- Directly estimate the missing noise component with the optimized glottal model

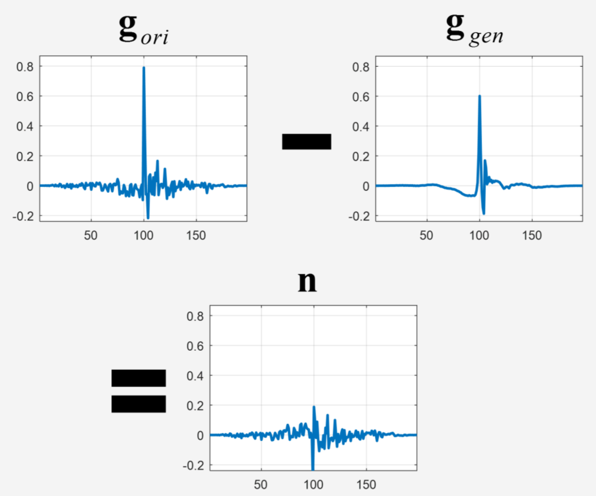
❖ Noise analysis in the training stage

- **Weighted subtraction**

$$\mathbf{g}_{ori} = \mathbf{h} + \mathbf{n} \text{ and } \mathbf{g}_{gen} = \alpha \cdot \mathbf{h}$$

$$\alpha = \frac{\mathbf{g}_{gen}^T \mathbf{g}_{gen}}{\mathbf{g}_{ori}^T \mathbf{g}_{gen}}, \text{ where } \mathbf{n}^T \mathbf{g}_{gen} = 0$$

$$\mathbf{n} = \mathbf{g}_{ori} - \frac{1}{\alpha}\mathbf{g}_{gen}$$



[Noise extraction process]

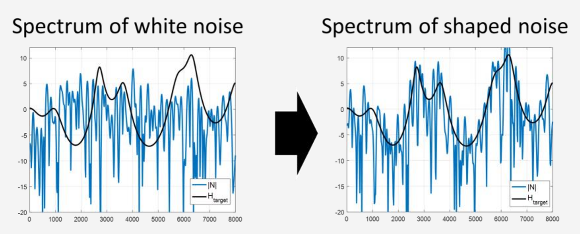
- **Noise features (NFs) extraction**
  - **Noise LSFs** for spectral information
  - **Pulse-wise HNR** for gain information

$$HNR = 10\log_{10}\left[E\left[\mathbf{h}^2\right] / E\left[\mathbf{n}^2\right]\right]$$

❖ Noise compensation in the synthesis stage

- **Spectral shaping**

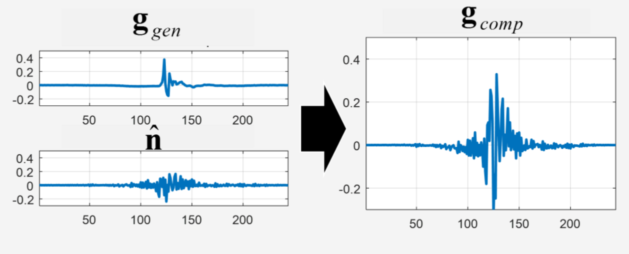
$$\hat{N}(\omega) = \frac{H_{target}(\omega)}{H_{noise}(\omega)} \cdot N(\omega)$$



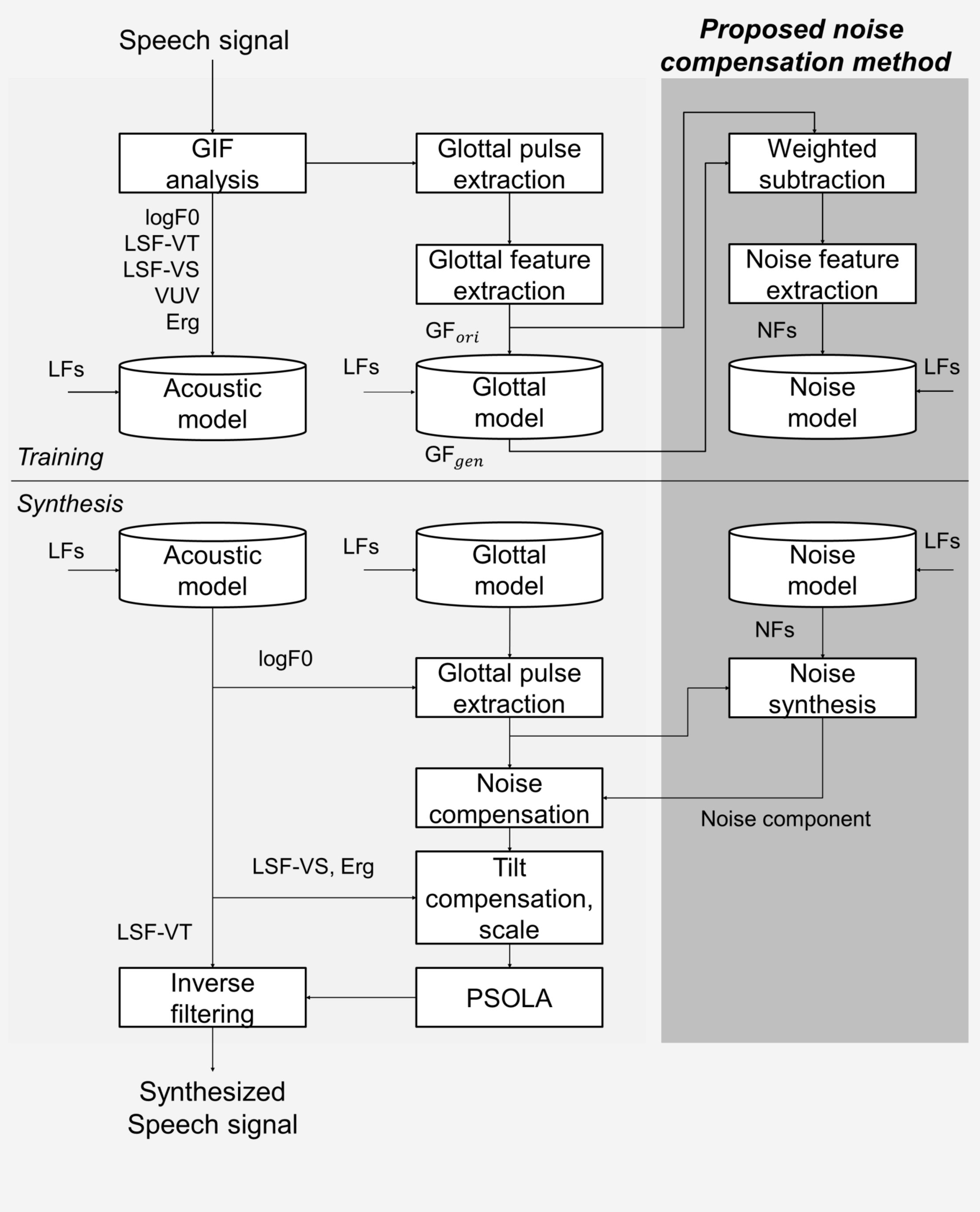
[Spectral shaping process]

- **High-pass filtering**
- **Noise compensation** after gain adjustment

$$\mathbf{g}_{comp} = \mathbf{g}_{gen} + \sqrt{\frac{HNR_{noise}}{HNR_{target}}} \cdot \hat{\mathbf{n}}$$



[Noise compensation process]



[Proposed speech synthesis system]

Experiments

❖ Settings

Database	Korean male speaker, 16kHz		
Training / validation / test	2,500 (~3h) / 200/ 200		
GIF analysis method	Quasi-closed phase analysis		
Network type	Acoustic model	Glottal model	Noise model
Input layer	210-dim. LFs	210-dim. LFs	210-dim. LFs
Hidden feed forward layer	1024 x 2	512 x 3	512 x 2
Hidden LSTM layer	512 x 2	256 x 1	256 x 1
Output layer	142-dim. AFs	400-dim. GFs	48-dim. NFs
Initialization	Xavier initialization		
Optimizer	Adam optimizer		
Activation function	tangent hyperbolic for hidden / linear for output		
Post-processing	Maximum likelihood parameter generation		
	LSF-sharpening processing on all LSFs		
	Formant enhancement on LSF_VT		

❖ Objective evaluation

- Log spectral distance (dB) between **original** and **noise compensated** glottal pulses

HNR-NC	MF-NC	MbG
8.55	7.93	7.61

The proposed MbG approach shows smaller errors than those with conventional approaches.

❖ Subjective evaluation with baseline STRAIGHT-based synthesis system

- A/B/X preference test result

STR.	HNR-NC	MF-NC	MbG	No pref.	p-vale
-	3.3	**87.5**	-	9.2	$< 10^{-79}$
-	-	5.4	**47.1**	47.5	$< 10^{-21}$
25.8	-	**64.6**	-	9.6	$< 10^{-10}$
17.9	-	-	**75.4**	6.7	$< 10^{-23}$
72.1	10	-	-	17.9	$< 10^{-34}$
-	1.7	0	**93.3**	5.9	$< 10^{-113}$

- MOS test result with 95% confidence interval

STR.	HNR-NC	MF-NC	MbG
2.90±0.10	2.16±0.13	3.20±0.13	3.72±0.11

In both ABX and MOS tests, the proposed system presented significantly better perceptual performances than the conventional systems.

Conclusion

- A **modeling-by-generation (MbG)-structured noise compensation algorithm** for glottal vocoding speech synthesis system was proposed.
- By directly modeling the missing noise component from the **difference between the glottal pulses of the original and generated ones** and by including it in the entire training process, we were able to construct a **glottal model-adaptive noise compensation method**.
- The experimental results verified **that the proposed system was superior to conventional glottal vocoding systems, both objectively and subjectively**.

[1] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.

[2] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN-a full-band glottal vocoder for statistical parametric speech synthesis." *in Proc. INTERSPEECH*, 2016, pp. 2473–2477.