An Excitation Model for HMM-based Speech Synthesis Based on Residual Modeling

Ranniery Maia[†,‡] Tomoki Toda[†,††] Heiga Zen[†‡]
Yoshihiko Nankaku[†‡] Keiichi Tokuda[†,†‡]

[†]National Institute of Information and Communications Technology (NiCT), Japan
[‡]ATR Spoken Language Communication Laboratories, Japan
[††]Nara Institute of Science and Technology, Japan
[†‡]Nagoya Institute of Technology, Japan

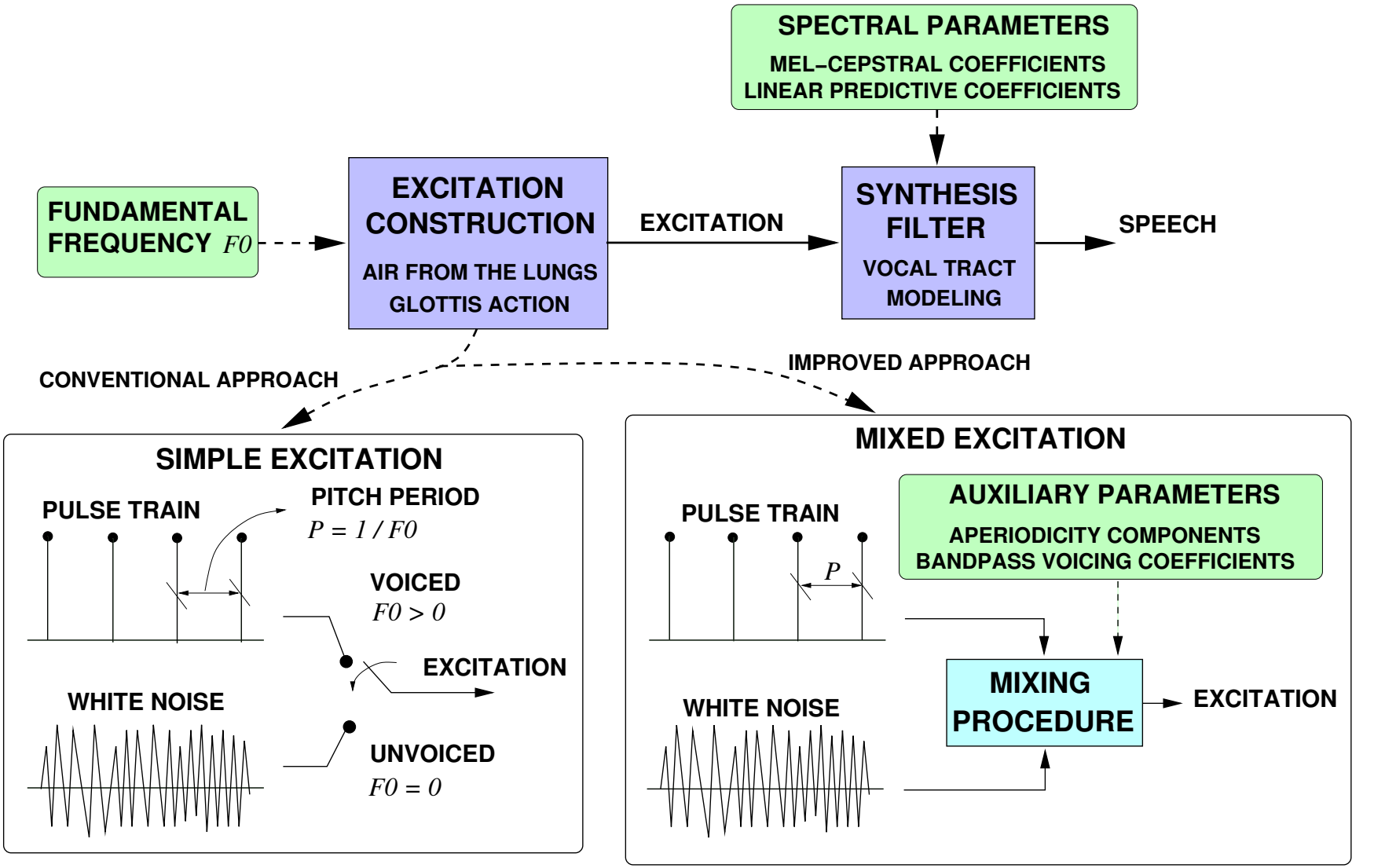
August 23, 2007

Contents

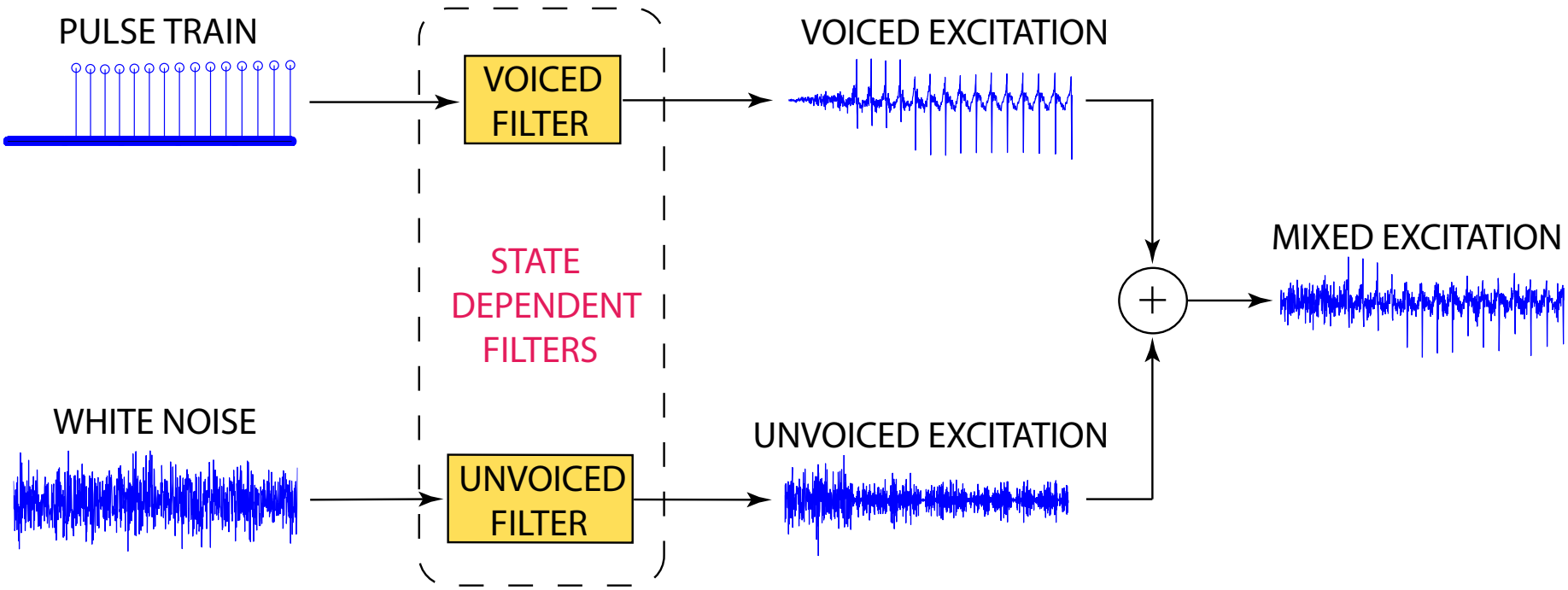
Introduction

- HMM-based speech synthesis: **flexibility vs. naturalness**
- Improved excitation models for HMM-based speech synthesis
  - Based on MELP speech coding: *Yoshimura (Eurospeech 2001)*
  - Utilization of STRAIGHT: *Zen (IEICE Trans. Inf. Sys. Jan. 2007)*
  - Approaches based on sinusoidal modeling: e.g., *S.J. Kim (IEICE Trans. Inf. Sys. Jan 2007)*
- Minimization of the error between synthesized and natural speech waveforms $\Rightarrow$ **analysis-by-synthesis concept** $\Rightarrow$ not yet proposed
- *Akamine (ICSLP 1998):* generation of speech units for concatenation through **closed-loop training**
- **Excitation model derived by closed-loop training?**

The HMM-based speech synthesis: utilization of the source-filter model for speech production

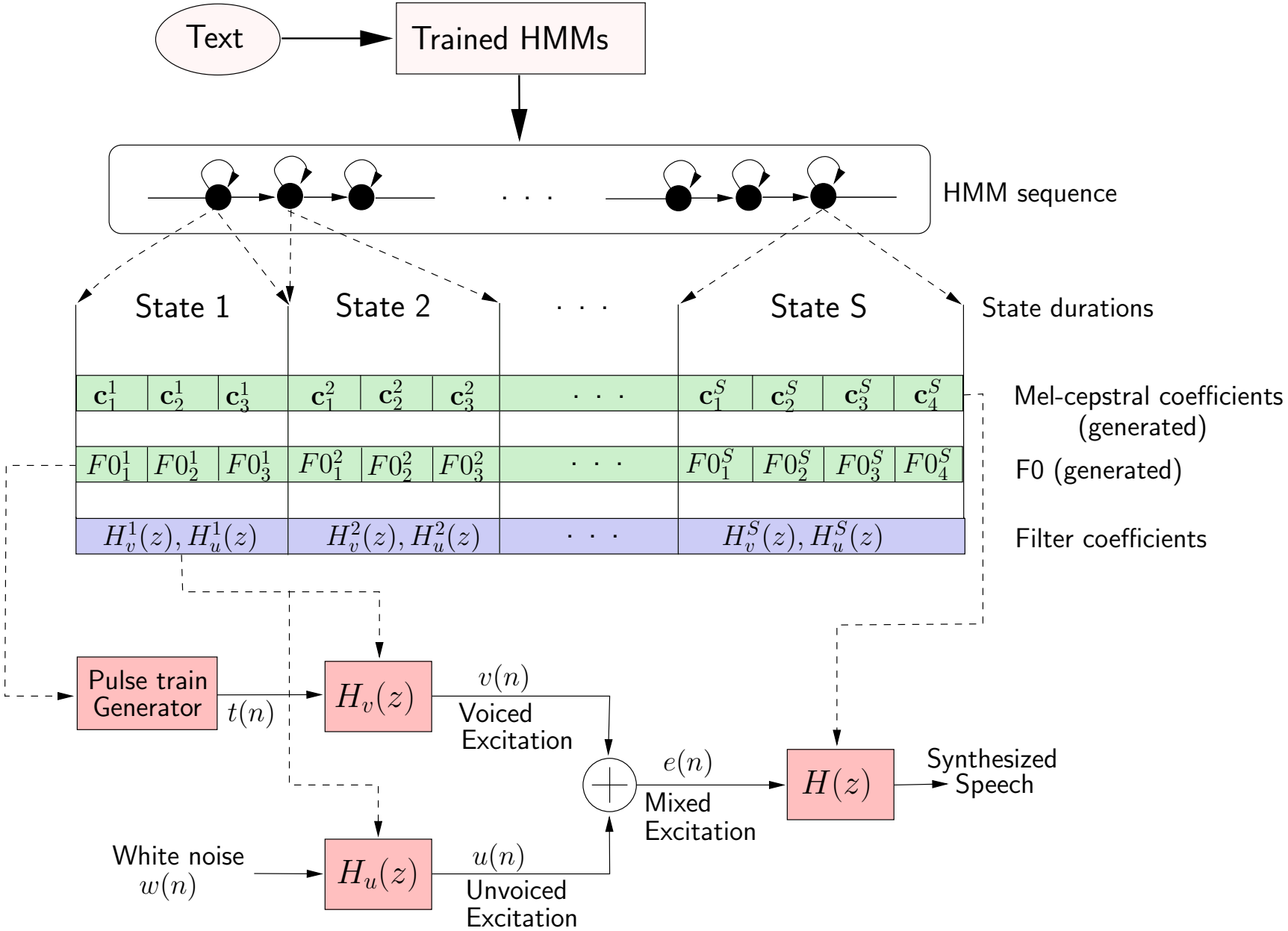


The proposed excitation model

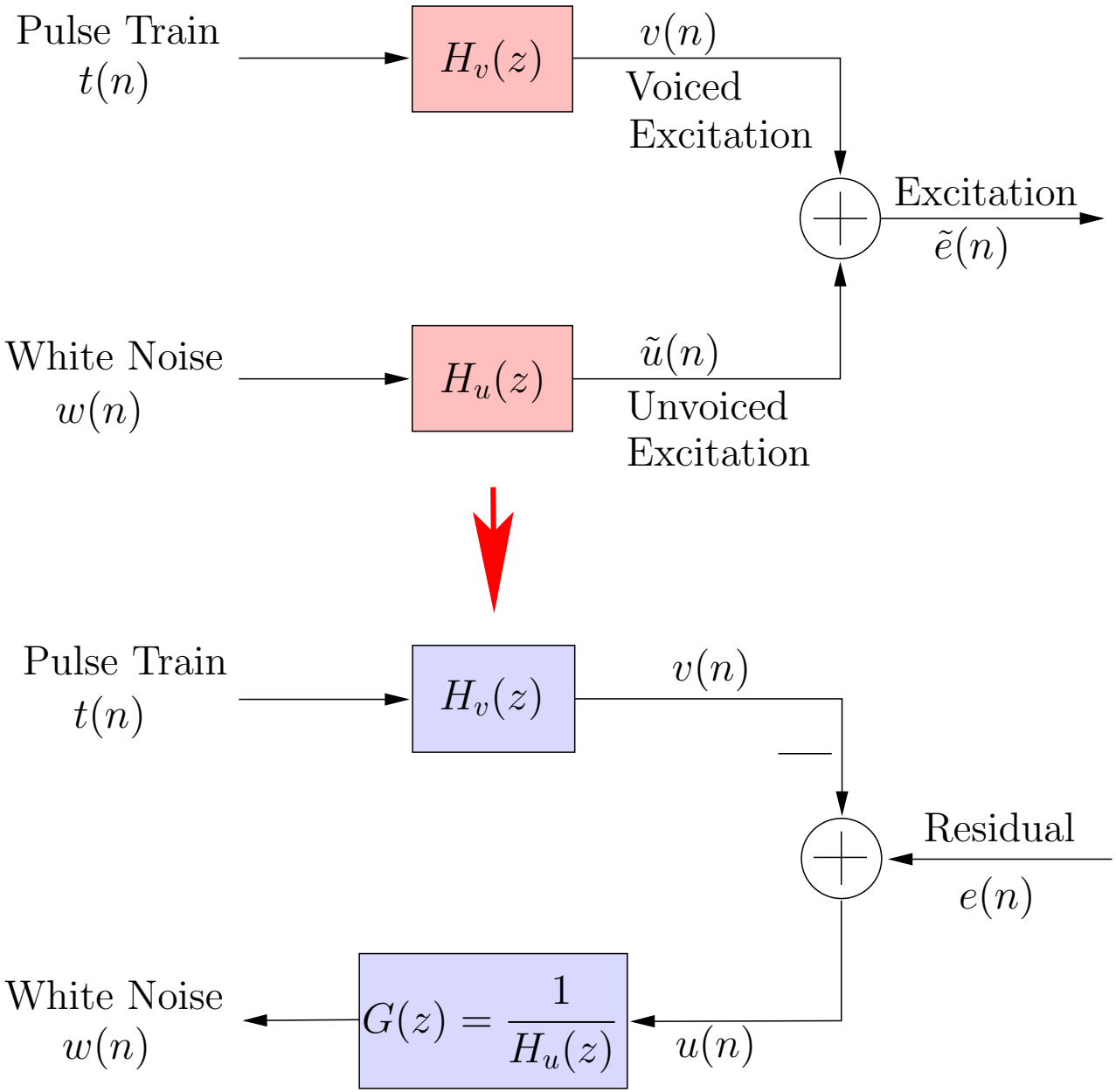


- **Voiced filter:** $H_v(z) = \sum_{l=-\frac{M}{2}}^{\frac{M}{2}} h(l) z^{-l}$
  - Process pulse train to generate close-to-residual voiced excitation
- **Unvoiced filter:** $H_u(z) = \frac{K}{1 - \sum_{l=1}^{L} g(l) z^{-l}}$
  - Noise shaping

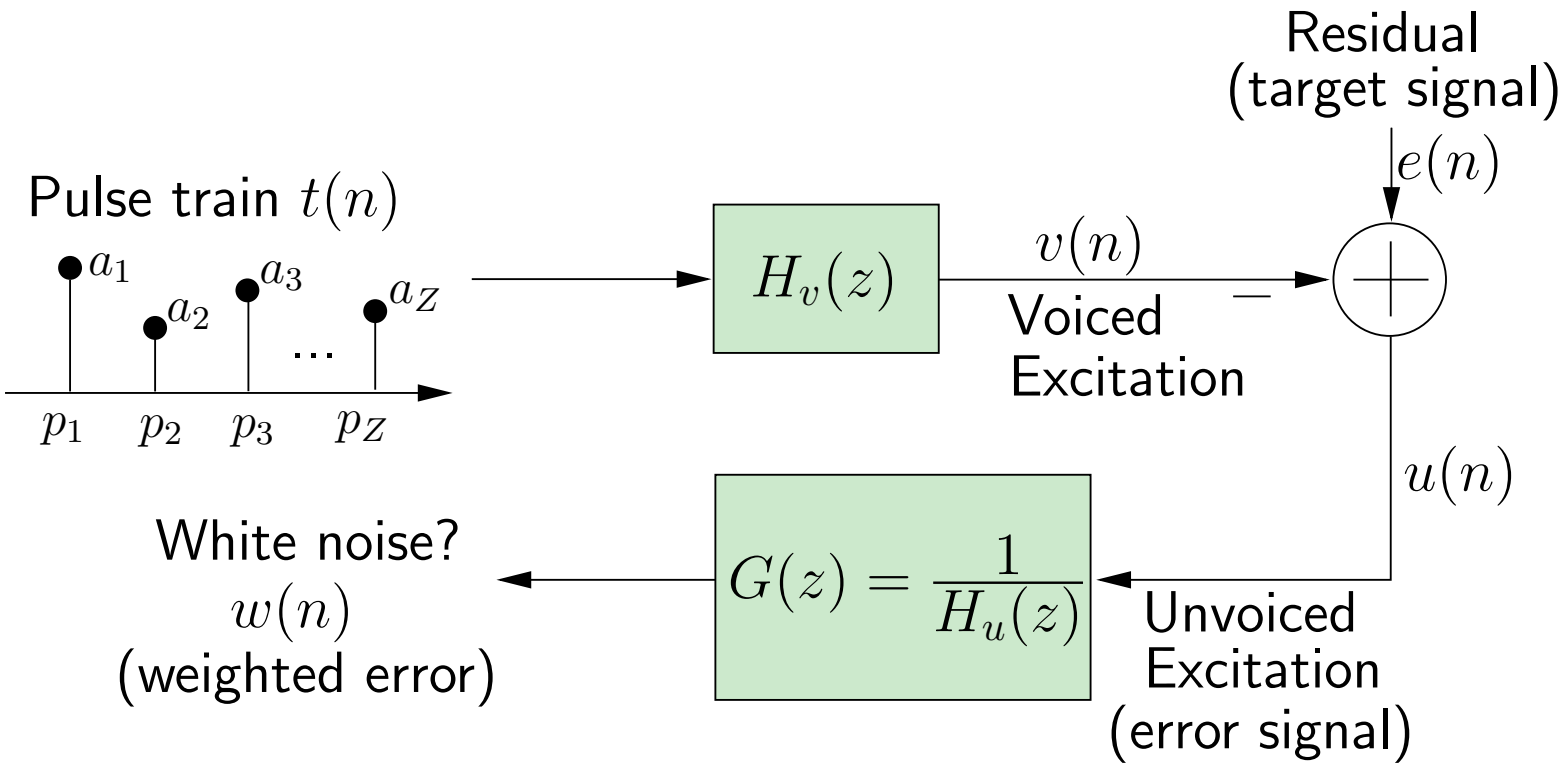
Overall picture: synthesis of an utterance



Excitation training



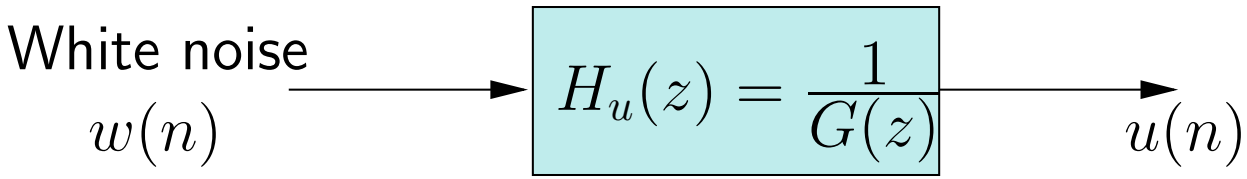
Excitation training: analogy with AbS coding



- Target signal: $e(n)$
- Terms to be optimized: $t(n)$, $H_v(z)$ and $H_u(z)$
- Error to be minimized: $\varepsilon = E\{w^2(n)\}$ (MSE)

Filter determination: maximum likelihood criterion

- Likelihood of $u(n)$ given $H_u(z)$



$$P\left[\mathbf{u}|H_u(z)\right] = \frac{1}{\sqrt{(2\pi)^N|\mathbf{R}|}} e^{-\frac{1}{2}\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u}}$$

- Where

$$\begin{cases} \mathbf{u} & = \mathbf{e} - \mathbf{v} = \mathbf{e} - \mathbf{H}_v\mathbf{t} \\ \mathbf{R}^{-1} & = \mathbf{G}^T\mathbf{G} \end{cases}$$

Likelihood maximization vs. MSE minimization

- Likelihood of $e(n)$ given $H_v(z)$, $H_u(z)$ and $t(n)$

$$\log P\left[\mathbf{e}|H_v(z), H_u(z), t(n)\right] = -\frac{N}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{G}^T\mathbf{G}| - \\ -\frac{1}{2}\left[\mathbf{e}-\mathbf{H}_v\mathbf{t}\right]^T\mathbf{G}^T\mathbf{G}\left[\mathbf{e}-\mathbf{H}_v\mathbf{t}\right]$$

- Mean squared error

$$\varepsilon = \mathbf{w}^T\mathbf{w} = \left[\mathbf{e}-\mathbf{H}_v\mathbf{t}\right]^T\mathbf{G}^T\mathbf{G}\left[\mathbf{e}-\mathbf{H}_v\mathbf{t}\right]$$

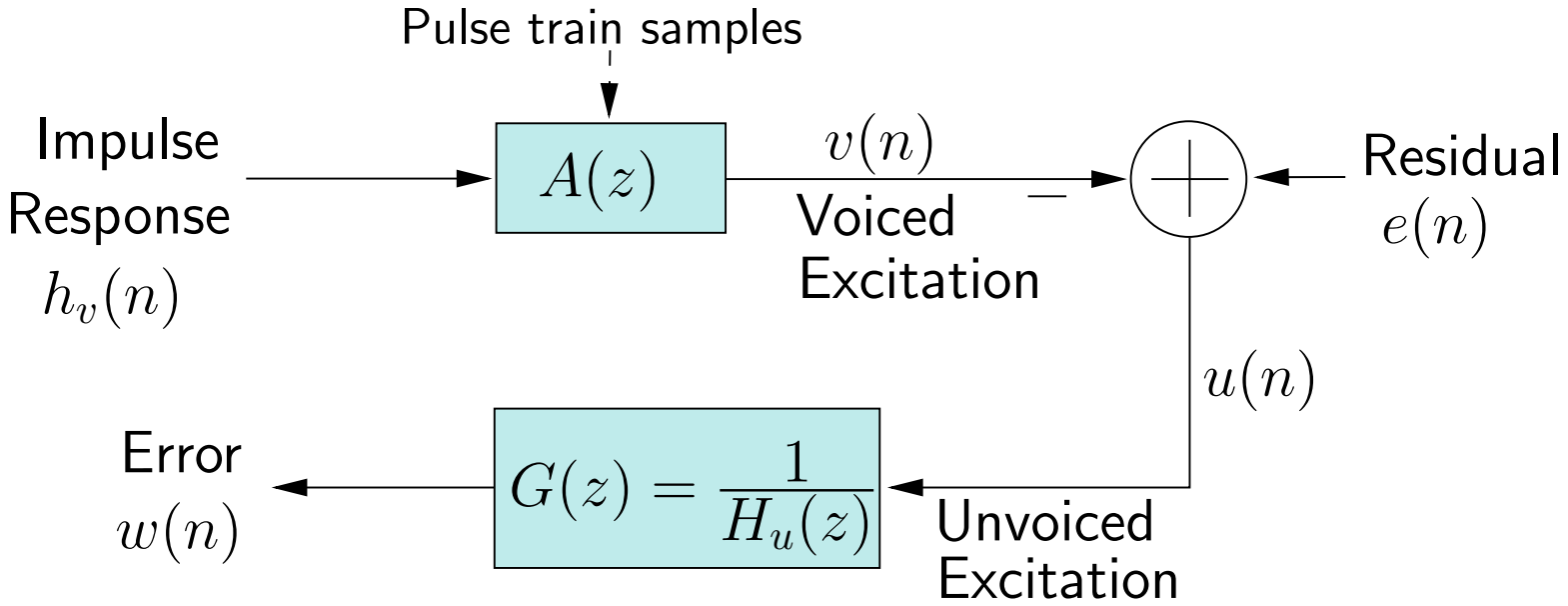
- Since $G(z)$ is minimum-phase

$$\log|\mathbf{G}^T\mathbf{G}| = 0$$

$$\Downarrow$$

- **Likelihood maximization $\Longleftrightarrow$ MSE minimization**

Voiced filter determination

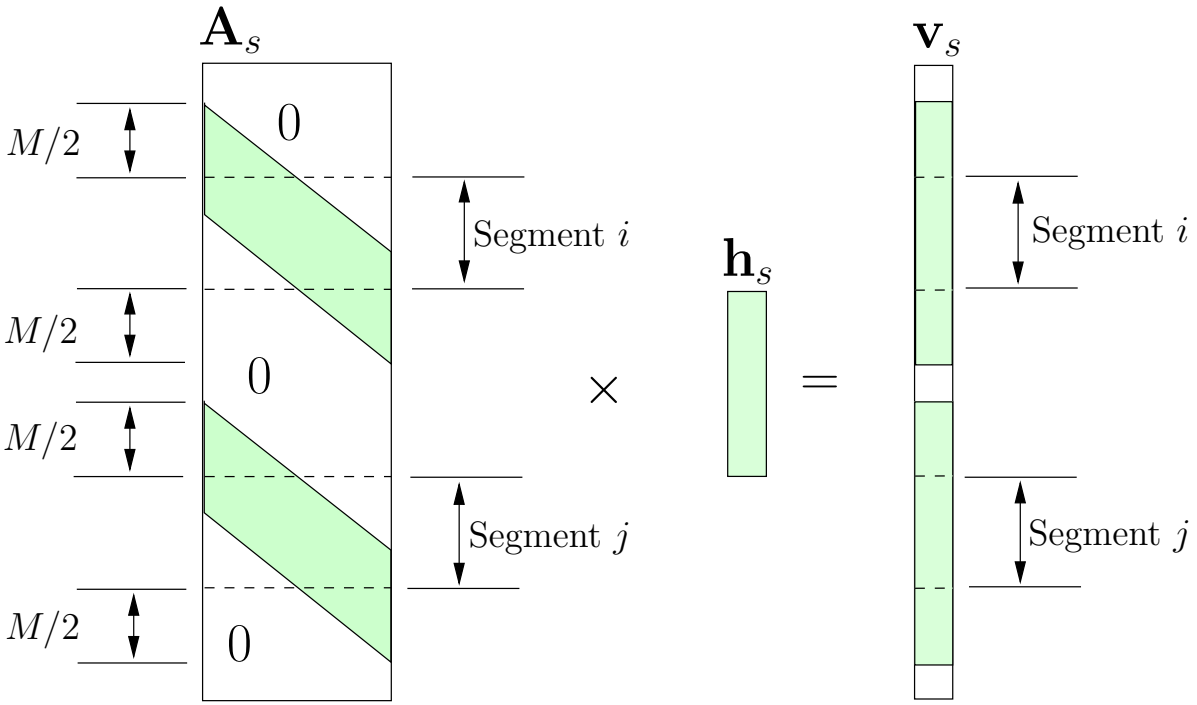


- **Voiced excitation vector**

$$\mathbf{v} = \mathbf{A}_1\mathbf{h}_1 + \ldots + \mathbf{A}_S\mathbf{h}_S = \sum_{s=1}^{S} \mathbf{A}_s\mathbf{h}_s$$

  - $\{1, \ldots, S\}$: state set
  - $\{\mathbf{h}_1, \ldots, \mathbf{h}_S\}$: corresponding voiced filter coefficients

Voiced filter determination



- MSE

$$\varepsilon = \left[\mathbf{e} - \sum_{s=1}^{S} \mathbf{A}_s \mathbf{h}_s\right]^T \mathbf{G}^T\mathbf{G} \left[\mathbf{e} - \sum_{s=1}^{S} \mathbf{A}_s \mathbf{h}_s\right]$$

- Voiced filter for state s

$$\frac{\partial \varepsilon}{\partial \mathbf{h}_i} = 0 \implies \mathbf{h}_s = \left(\mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \mathbf{A}_s\right)^{-1} \mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \left[\mathbf{e} - \sum_{r \neq s} \mathbf{A}_r \mathbf{h}_r\right]$$

Unvoiced filter determination

- MSE again

$$\varepsilon = \mathbf{w}^T\mathbf{w} = [\mathbf{e}-\mathbf{v}]^T\,\mathbf{G}^T\mathbf{G}\,[\mathbf{e}-\mathbf{v}] = \frac{1}{K^2}\sum_{n=0}^{N-1}\left[u(n) - \sum_{l=1}^{L} g(l)u(n-l)\right]^2$$

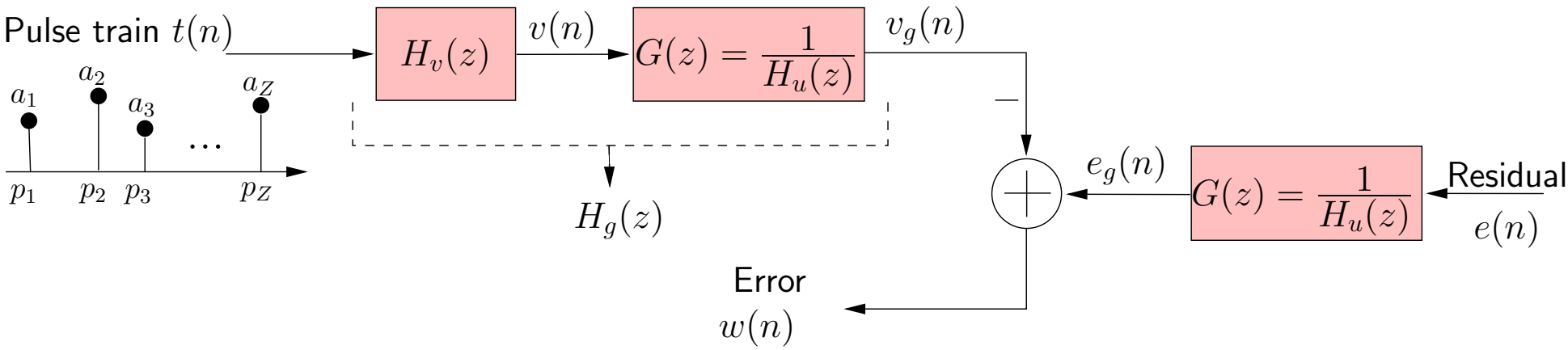
- By making $\dfrac{\partial\varepsilon}{\partial K} = 0$

$$\begin{cases} K = & \sqrt{\varepsilon_m} \\ \varepsilon_m = & \min\limits_{g(1),\ldots,g(l)} \left\{ \dfrac{1}{N}\sum\limits_{n=0}^{N-1}\left[u(n) - \sum\limits_{l=1}^{L} g(l)u(n-l)\right]^2 \right\} \end{cases}$$

$$\Downarrow$$

Linear prediction of $u(n)$

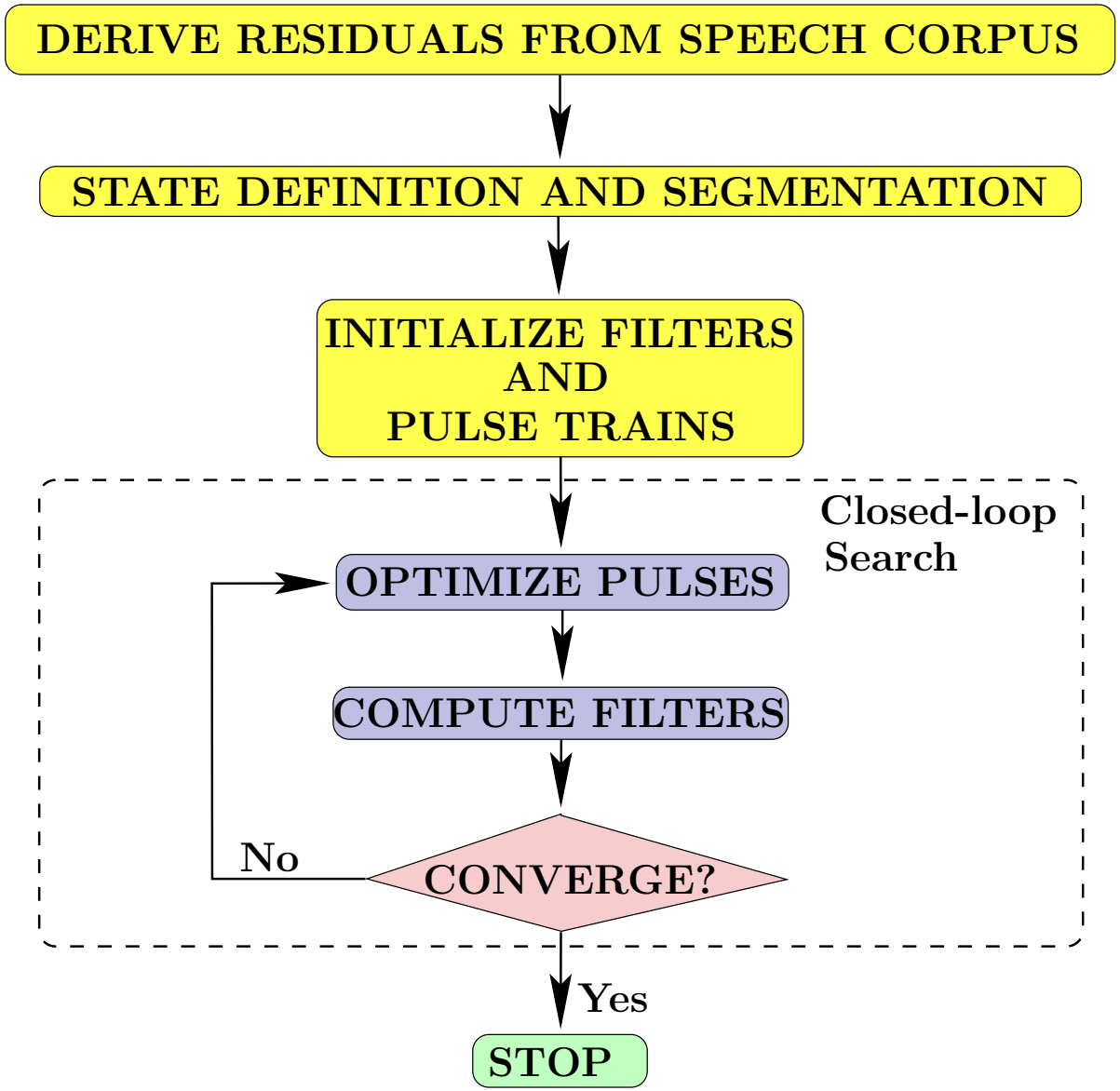
Pulse optimization



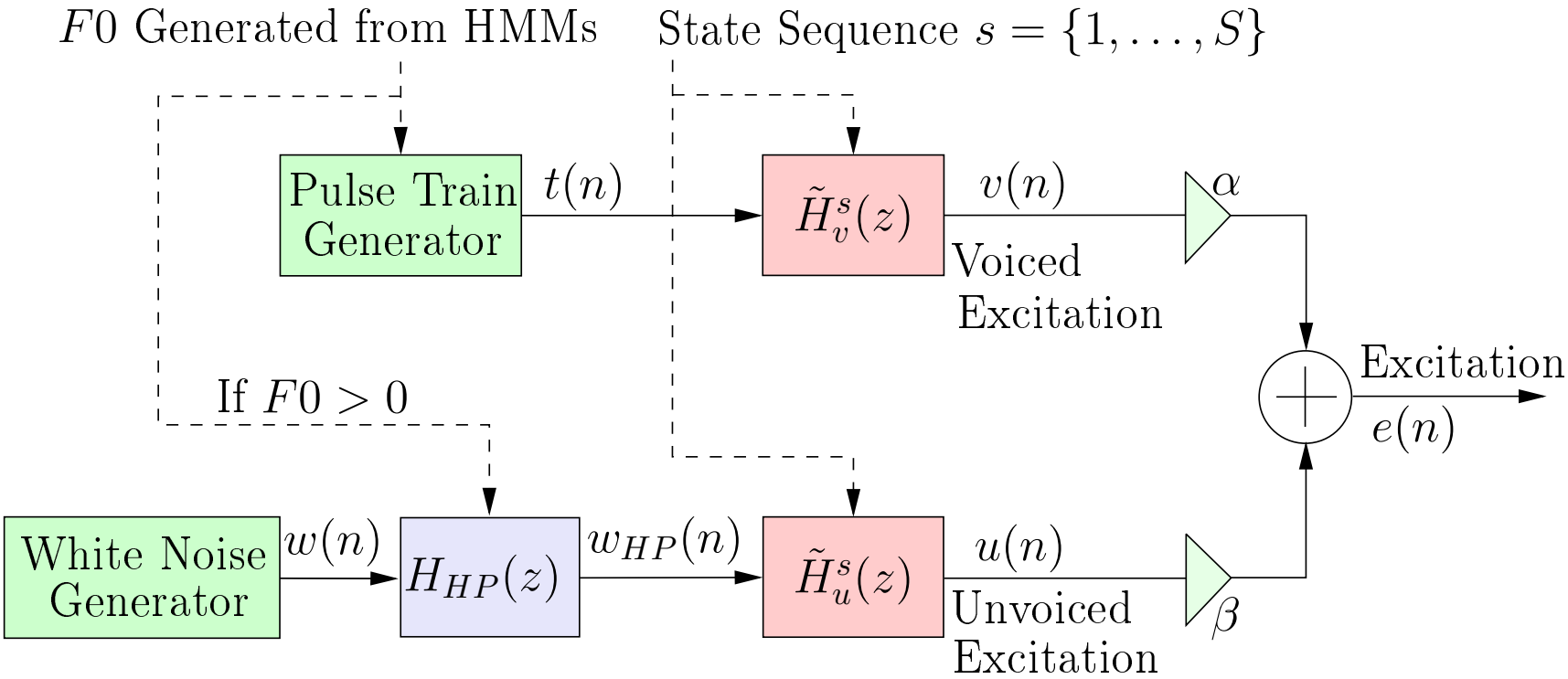
Analogy: Multipulse Excitation Linear Prediction

- $e(n)$: target signal
- $H_g(z) = H_v(z)G(z)$: fixed
- $\{a_1, \ldots, a_Z\}$: amplitudes to be optimized
- $\{p_1, \ldots, p_Z\}$: positions to be optimized
- Error to be minimized: $\varepsilon = \frac{1}{N} \sum_{n=0}^{N-1} w^2(n)$

Joint filter computation and pulse optimization



Synthesis part

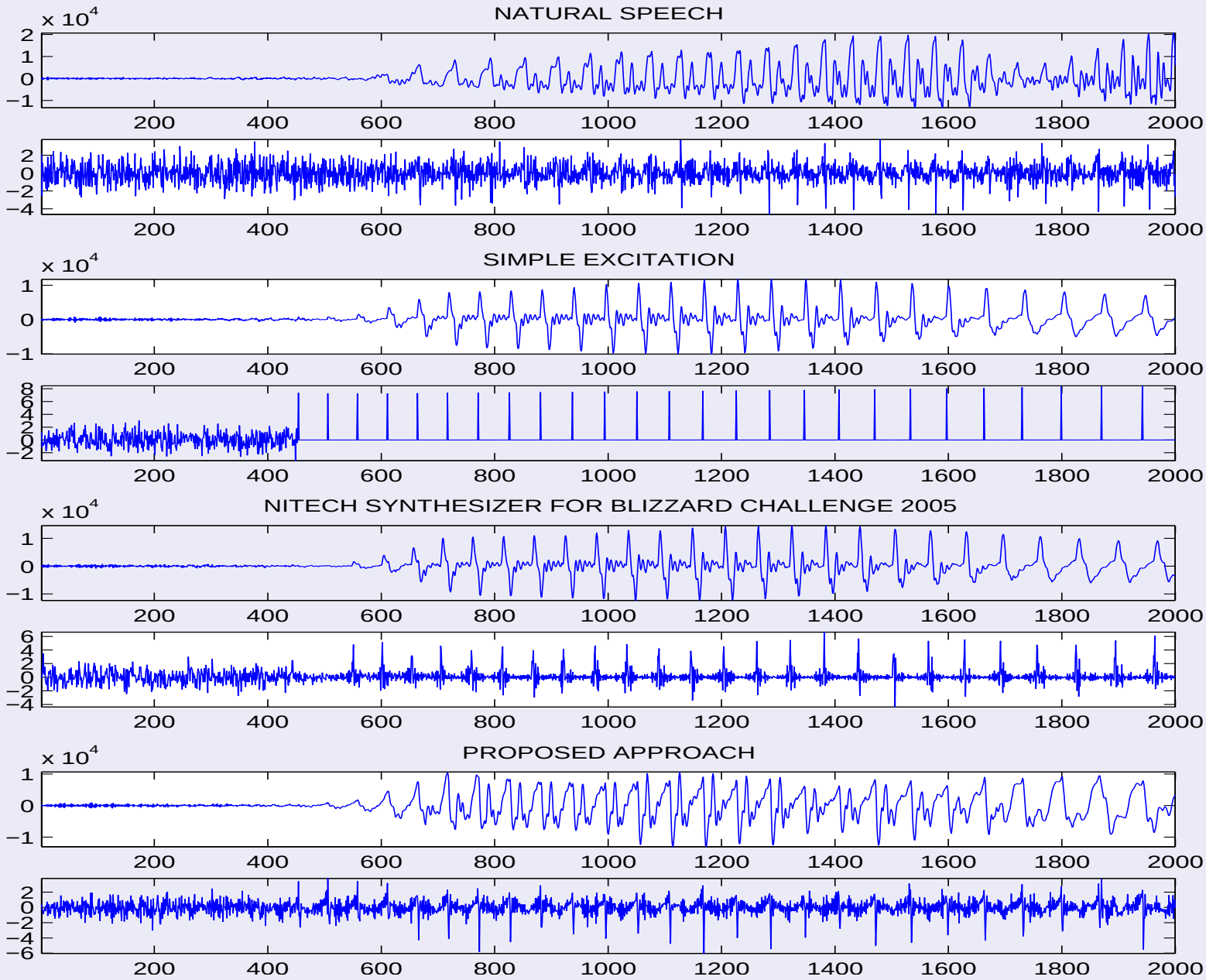


- $t(n)$: power-normalized pulse train
- α, β: gain control
- $\tilde{h}_s(n) = \frac{h_s(n)}{\sqrt{\sum_{l=-M/2}^{M/2} h_s^2(l)}}$: energy-normalized $h_s(n)$
- $\tilde{H}_u^s(z) = \frac{1}{1-\sum_{l=1}^{L} g_s(l) z^{-l}}$
- $H_{HP}(z)$: highpass filter ($f_c = 4$kHz)

Experiment

- Corpus
  - **ATR503 F009** $\Rightarrow$ Japanese female
  - Approximately 30 minutes
- Filter orders: $M = 512$ (voiced) and $L = 64$ (unvoiced)
- **The states**
  - Small decision tree for mel-cepstral coefficients
    - ★ High correlation between spectrum and residual
    - ★ Questions regarding the central phone
    - ★ $\lambda = 12$ (MDL factor)
    - ★ 75 clusters
  - Segmentation procedure
    1. Alignment of the entire database using the usual HMMs
    2. Mapping of the aligned contextual labels onto the states of the small decision tree

Example: *"Ippyou no kakusa wa sarani hirogaru darou."*



Conclusions

- The proposed approach reduces the "buzziness" of HMM-based synthesis
  - Better than simple excitation
  - Comparable in quality to one of the best excitation approaches so far proposed for HMM-based synthesis
- Minimization of the distortion between natural and synthesized waveforms
  - **Concept of analysis-by-synthesis speech coders**
- Synthesized speech sounds slightly harsh
  - Similar issues have been reported for CELP speech coders
- Future steps
  - State clustering
  - Pulse train modeling for the synthesis part