Use of Particle Filtering and MCMC for Inference in Probabilistic Acoustic Tube Model

Ruobai Wang[1] wangruobai11@tsinghua.org.cn
Yang Zhang[2] yzhan143@illinois.edu
Zhijian Ou[1] ozj@tsinghua.edu.cn
Mark Hasegawa-Johnson[2] jhasegaw@illinois.edu

[1] Tsinghua University, Department of Electronic Engineering
[2] University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering

The Signal Model of PAT3

$$x(t) = \Big[\big(v(t) + \tilde{b}\,u(t)\big) * h(t) + \tilde{e}\,\varepsilon(t)\Big]w(t)$$

voiced source; unvoiced amplitude; Gaussian unvoiced signal; vocal tract response; noise; window

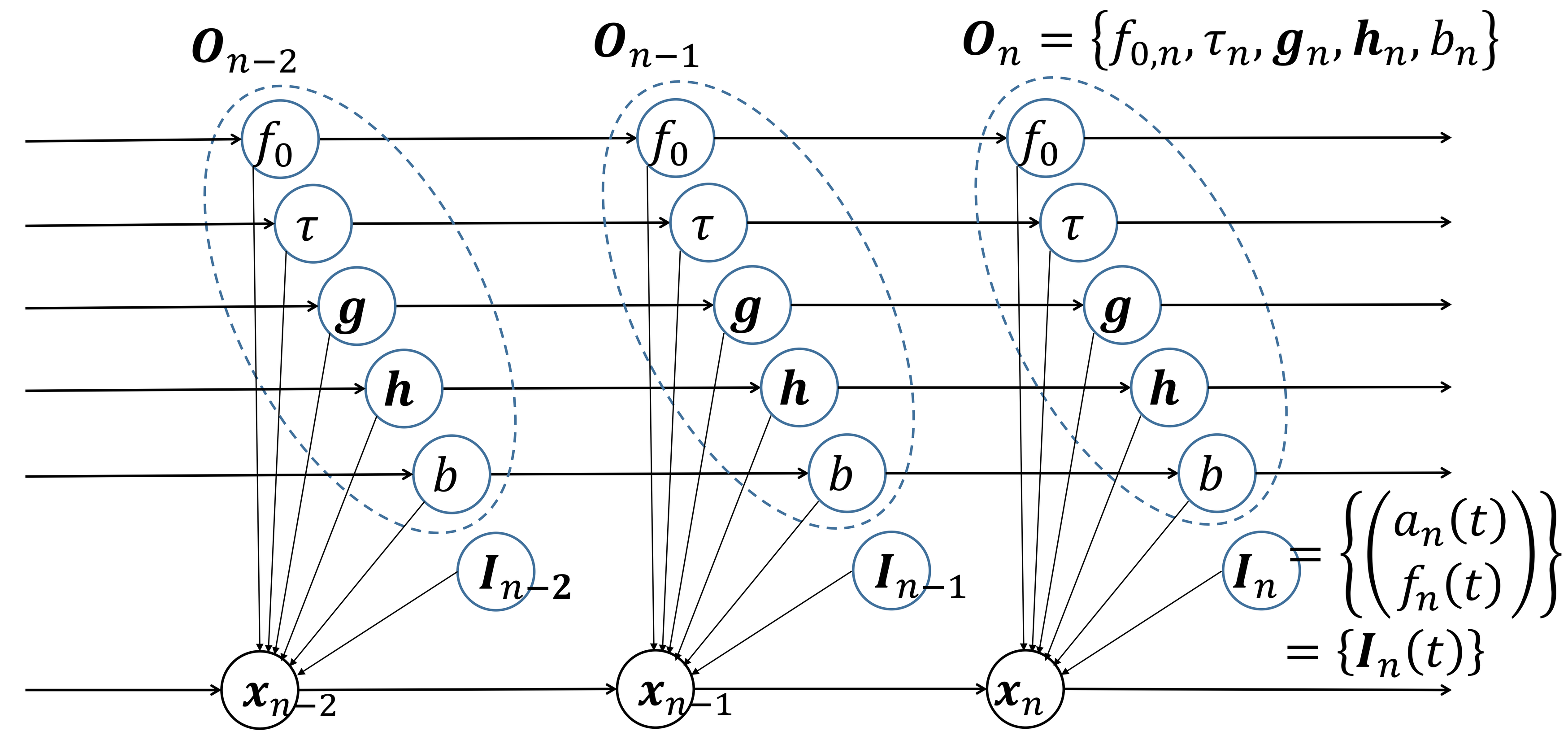
- $X_n(\omega) = \Big[\big(V_n(\omega) + \tilde{b}_n U_n(\omega)\big) H_n(\omega) + \tilde{e}_n E_n(\omega)\Big] \otimes W_n(\omega)$
- Stack the real and imaginary parts of $X_n\left(\frac{2\pi k}{L}\right)$, $k = 0$ to $L/2$, as $\boldsymbol{x}_n$.
- Voiced part $S_n(\omega) = [V_n(\omega)H_n(\omega)] \otimes W_n(\omega)$, vectorized as $\boldsymbol{s}_n$.
- Unvoiced part implies variance: $\sigma_{n,k}^2 = \tilde{b}_n^2 \left|H_n\left(\frac{2\pi k}{L}\right)\right|^2 + \tilde{e}_n^2$

$$v(t) = \tilde{a}(t) \sum_{d=1}^{D} \mathrm{real}\big[\, G(d\omega_0)\ \exp(\, -d\omega_0\tau + d\phi(t)\,)\big]$$

amplitude envelope; harmonic order; glottal transfer function; pitch frequency; group delay; instant phase for d=1

The Probabilistic Model of PAT3

Speech frames are modeled as a non-linear state-space model.



Observation likelihood :

$$\log p(\boldsymbol{x}_n|\boldsymbol{s}_n) = \sum_{k=0}^{L-1}\left[-\frac{1}{2}\log(2\pi\sigma_{n,k}) - \frac{\left|X_n\left(\frac{2\pi k}{L}\right) - S_n\left(\frac{2\pi k}{L}\right)\right|^2}{2\sigma_{n,k}^2}\right]$$

State transition :

$$p(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}) = \mathcal{N}(\boldsymbol{O}_{n-1}, \Sigma_O)$$

$$p(\boldsymbol{I}_n(t)|\boldsymbol{I}_n(t-2), \boldsymbol{I}_n(t-1)) = \mathcal{N}(\lambda\boldsymbol{I}_n(t-1) + (1-\lambda)\boldsymbol{I}_n(t-2), \Sigma_I)$$

Outer Loop using APF

Frame-level MAP inference : $\widehat{\boldsymbol{O}}_{1:N}^{MAP} = \underset{\boldsymbol{O}_{1:N}}{\operatorname{argmax}}\, p(\boldsymbol{O}_{1:N}|\boldsymbol{x}_{1:N})$.

Suppose that $p(\boldsymbol{O}_{n-1}|\boldsymbol{x}_{1:n-1})$ is approximated by : $\left\{\boldsymbol{O}_{n-1}^{(i)}, w_{n-1}^{(i)}\right\}_{i=1:M}$

Then, APF aims to sample from

$$p(\boldsymbol{O}_n|\boldsymbol{x}_{1:n}) \propto p(\boldsymbol{x}_n|\boldsymbol{O}_{1:n}) \sum_{i=1}^{M} p\left(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}^{(i)}\right) w_{n-1}^{(i)}$$

with proposal

$$q\left(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}^{(i)}, \boldsymbol{x}_n\right)\beta_n^{(i)}.$$

For $n = 1, \dots, N$:

1. **Resample** $j_1, \dots, j_M$ from $\{1, \dots, M\}$ according to $\left\{\beta_n^{(i)}\right\}_{i=1,\dots,M}$
2. **Propagate**: Sample $\boldsymbol{O}_n^{(i)}$ from $q\left(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}^{(j_i)}, \boldsymbol{x}_n\right)$
3. **Reweight** each particle $\boldsymbol{O}_n^{(i)}$ as $w_n^{(i)} \propto \dfrac{p\left(\boldsymbol{x}_n|\boldsymbol{O}_n^{(i)}\right)p\left(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}^{(j_i)}\right)w_{n-1}^{(j_i)}}{q\left(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}^{(j_i)}, \boldsymbol{x}_n\right)\beta_n^{(j_i)}}$

Inner Loop using Taylor Expansion Assisted MCMC

Viterbi approximation :

$$p(\boldsymbol{x}_n|\boldsymbol{O}_n) \approx \max_{\boldsymbol{I}_n} p(\boldsymbol{x}_n|\boldsymbol{O}_n, \boldsymbol{I}_n) = p(\boldsymbol{x}_n|\boldsymbol{O}_n, \widehat{\boldsymbol{I}}_n)p(\widehat{\boldsymbol{I}}_n)$$

$\widehat{\boldsymbol{I}}_n$ is picked from the samples $\left\{\widehat{\boldsymbol{I}}_n^{(r)}\right\}_{r=1:R}$ drawn from $p(\boldsymbol{I}_n|\boldsymbol{O}_n, \boldsymbol{x}_n)$.

For each frame n, we run Metropolis-Hasting algorithm :

$$\min\left\{1, \frac{p(\boldsymbol{I}_n|\boldsymbol{O}_n, \boldsymbol{x}_n)\,q\left(\boldsymbol{I}_n^{(r)}|\boldsymbol{I}_n, \boldsymbol{O}_n, \boldsymbol{x}_n\right)}{p\left(\boldsymbol{I}_n^{(r)}|\boldsymbol{O}_n, \boldsymbol{x}_n\right)q\left(\boldsymbol{I}_n|\boldsymbol{I}_n^{(r)}, \boldsymbol{O}_n, \boldsymbol{x}_n\right)}\right\}$$

with proposal

$$q\left(\boldsymbol{I}_n|\boldsymbol{I}_n^{(r)}, \boldsymbol{O}_n, \boldsymbol{x}_n\right) \triangleq \prod_t q\left(\boldsymbol{I}_n(t)|\boldsymbol{I}_n^{(r)}(1{:}t-1), \boldsymbol{I}_n^{(r)}(t+1{:}T), \boldsymbol{O}_n, \boldsymbol{x}_n\right)$$

We exploit 2nd-order Taylor expansion of the single-site conditional distribution

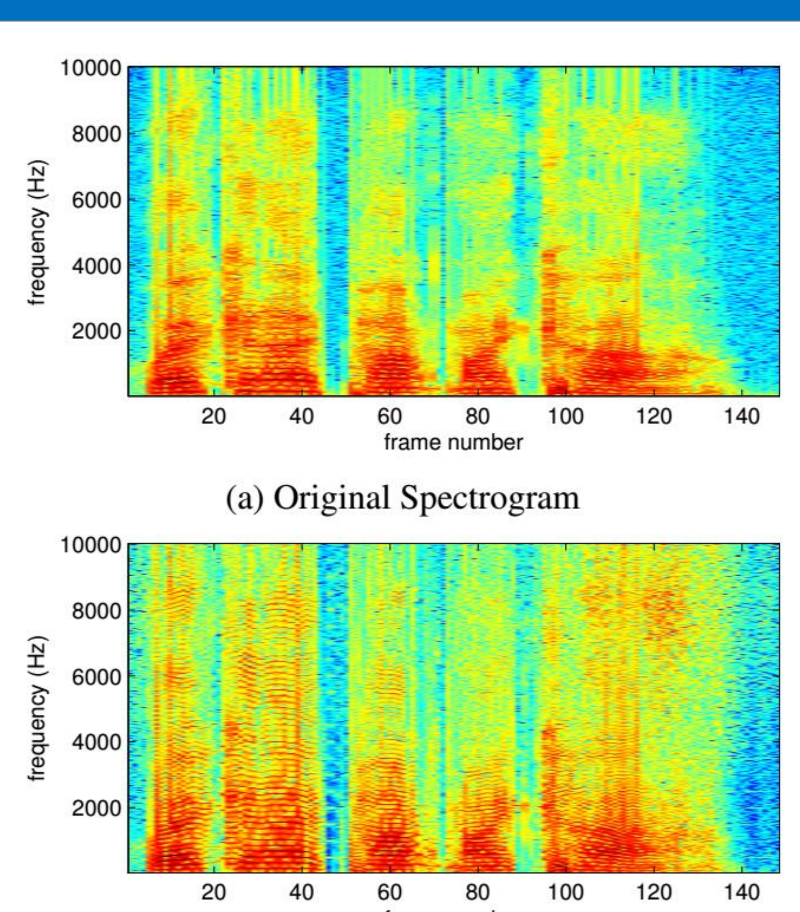
$$\ln p\left(\boldsymbol{I}_n(t)|\boldsymbol{I}_n^{(r)}(1{:}t-1), \boldsymbol{I}_n^{(r)}(t+1{:}T), \boldsymbol{O}_n, \boldsymbol{x}_n\right)$$

at the a prior conditional mean of $\boldsymbol{I}_n(t)$

$$\mathrm{E}\left(\boldsymbol{I}_n(t)|\boldsymbol{I}_n^{(r)}(t-2{:}t-1), \boldsymbol{I}_n^{(r)}(t+1{:}t+2)\right)$$

to define the single-site proposal distribution.

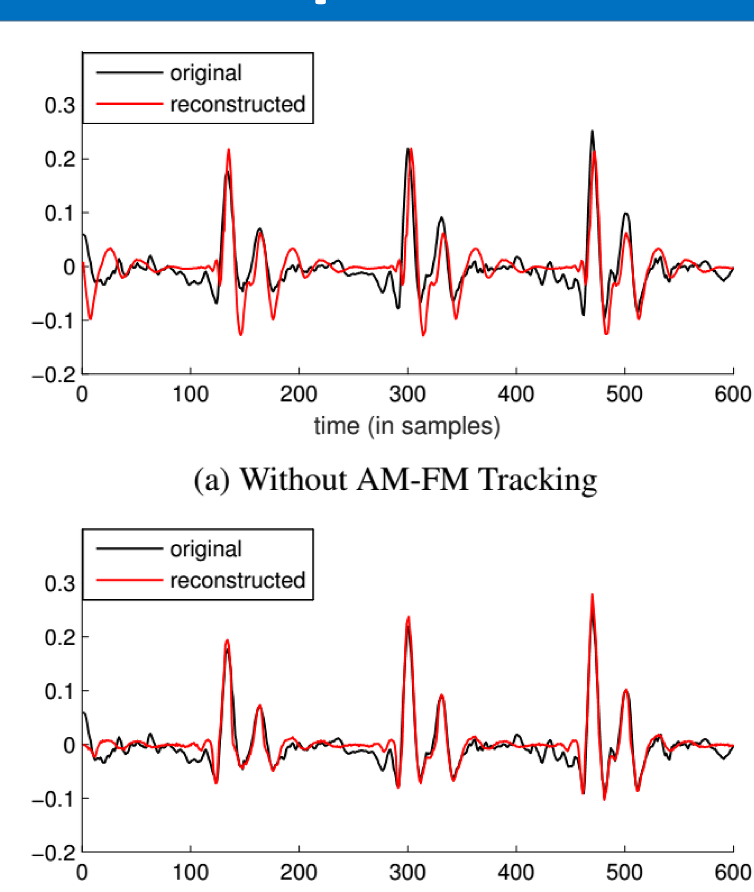
Experiments



(a) Original Spectrogram

(b) Reconstructed Spectrogram

Fig. 1: Comparison of the reconstructed and original spectrogram.



(a) Without AM-FM Tracking

(b) With AM-FM Trakcing

Fig. 2: The reconstructed voiced waveform (red line) with and without AM-FM tracking, compared with the original waveform (black).

PAT3	STRAIGHT
8.79	-2.57

Table 1: SNR of speech reconstruction

	PAT3	GetF0
GPE (%)	2.10	2.07
RMS (Hz)	5.052	5.780

Table 2: Pitch tracking results on Edinburgh dataset

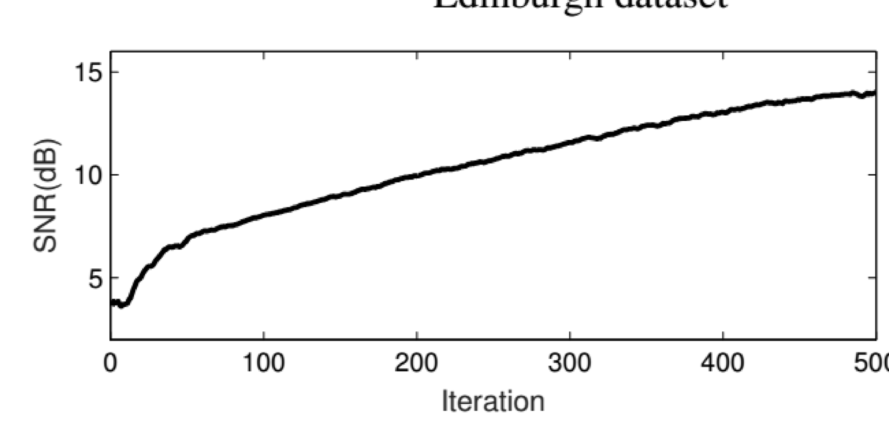


Fig. 3: SNR of reconstruction as a function of MCMC iterations.