INCORPORATING AM-FM EFFECT IN VOICED SPEECH FOR PROBABILISTIC ACOUSTIC TUBE MODEL

Yang Zhang [*], Zhijian Ou [†], Mark Hasegawa-Johnson [*]
[*]University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering
[†]Tsinghua University, Department of Electronic Engineering
yzhan143@Illinois.edu, ozj@tsinghua.edu.cn, jhasegaw@Illinois.edu

Introduction: Probabilistic Acoustic Tube (PAT) Model

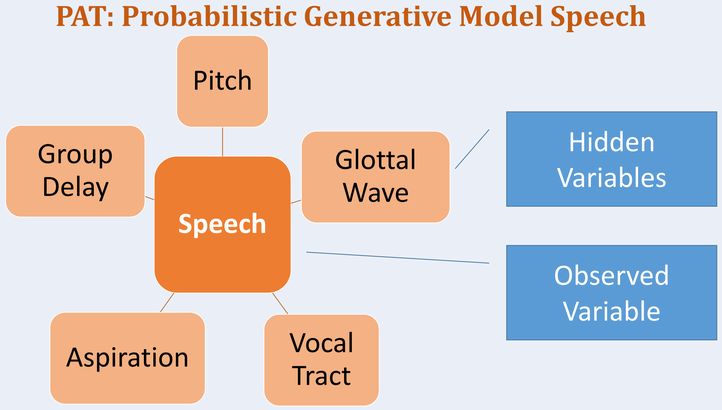
Joint Model of Main Speech Parameters

- Pitch and Spectral Envelope
- Glottal Source

Previous Joint Models

- STRAIGHT (Kawahara, et. al., ICASSP 2008):
- SVLN (Degottex, et. al., Speech Communication 2013)
- Still estimate parameters *separately*

PAT: Probabilistic Generative Model Speech



Problem with previous PAT

- Ignores AM/FM effect in voiced speech:
  Pitch Jitter & Amplitude Shimmer
- *Underestimates* voiced variations
- *Overestimate* unvoiced variations

The new PAT3 Model

- Incorporate AM/FM in voiced models

Signal & Probabilistic Modeling of PAT

The Voiced Excitation Model – Modulated Sinusoids

$$e_v[t] = \text{real}\left[\sum_d G(d\omega_0)\eta[t]\exp\big(jd(\omega_0(t-\tau)+\phi[t])\big)\right]$$

Glottal Wave Transfer Function | Fundamental Frequency | Amplitude Modulation | Group Delay | Phase Modulation

$G(\omega)$: Three-Pole Model

Speech

Voiced Excitation | Unvoiced Excitation

$$s[t] = (ae_v[t] + be_u[t]) * h[t]$$

Voiced Amplitude | Unvoiced Amplitude | Vocal tract impulse response

The Unvoiced Excitation Model – White Gaussian Noise

$$e_u[t] \sim \mathcal{N}(0,1)$$
$$\Rightarrow u[t] \triangleq be_u[t] * h[t] \sim \mathcal{N}(0,\Sigma)$$

Vocal Tract Transfer Function – Complex Cepstrum

$$h[t] \sim \text{IFFT}\left\{\exp\left[\hat{h}[n]\right]\right\}$$

Complex Cepstrum at low quefrency

Probabilistic Modeling for Voiced: AM/FM

Adapted Bayesian Spectral Estimation

$$v[t] \triangleq ae_v[t] * h[t]$$
$$= \text{real}\left[\sum_d \alpha_d\eta[t]\exp\big(jd(\omega_0(t-\tau)+\phi[t])\big)\right]$$
$$= x_d[t]\xi_d[t]$$

$aG(d\omega_0)H(d\omega_0)$

$$x_d[t] = \begin{bmatrix}|\alpha_d|\cos(d\omega_0(t-\tau)+\angle\alpha_d)\\ |\alpha_d|\sin(d\omega_0(t-\tau)+\angle\alpha_d)\end{bmatrix}$$
$$\xi_d[t] = \begin{bmatrix}\eta[t]\cos(d\phi[t])\\ \eta[t]\sin(d\phi[t])\end{bmatrix}$$

Deterministic Periodic Signal

Random AM-FM Variations

Autoregressive Model for $\xi_d[t]$

$$\xi_d[t] = \lambda_d\xi_d[t-1] + \varepsilon_d[t]$$

$$\varepsilon_d[t] \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\begin{bmatrix}1 & 0\\ 0 & \rho_d^2\end{bmatrix}\right)$$

$$\lambda_d = \exp(-d\delta)$$

Stationary Distribution

Assume the signal approaches stationary distribution:

$$\xi_d[t] \sim \mathcal{N}\left(0, \sigma_\xi^2\begin{bmatrix}1 & 0\\ 0 & \rho_d^2\end{bmatrix}\right)$$
$$\sigma_\xi^2 = \frac{\sigma_\varepsilon^2}{1-\lambda_d}$$
$$\rho_d = \tanh(d\gamma)$$

d small, $d\phi[t]$ small, $\rho_d \to 0$ d large, $d\phi[t]$ uniform, $\rho_d \to 1$

Experiments

Frame 32 Reconstruction without AM-FM

Frame 60 Reconstruction without AM-FM

Frame 32 Reconstruction with AM-FM

Frame 60 Reconstruction with AM-FM

Utterance 1, Male Speaker
Edinburgh Speech Corpus

Frame 32 Modeled Standard Deviation

Frame 60 Modeled Standard Deviation