# Probabilistic Acoustic Tube:
# a probabilistic generative model of speech for speech analysis/synthesis

**Zhijian Ou**
ozj@tsinghua.edu.cn

**Yang Zhang**
zhangyangbill@gmail.com

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

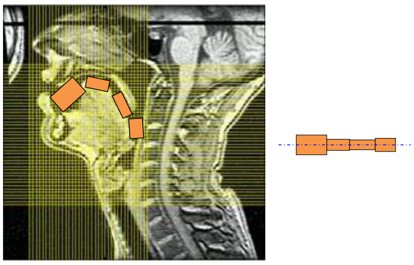
THE FIRST AUTHOR GRATEFULLY ACKNOWLEDGES THE SUPPORT OF K. C. WONG EDUCATION FOUNDATION, HONG KONG.

## Motivation

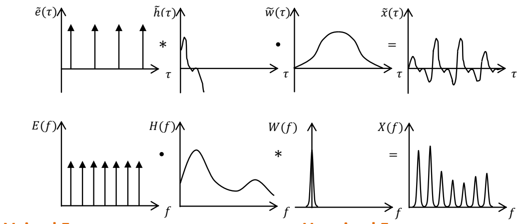
| Current Speech Analysis Methods | Probabilistic Acoustic Tube (PAT) |
|---|---|
| Signal processing front-end + probabilistic back-end.<br>Most features are nonlinear operators of speech (autocorrelation, cepstrum). | Directly model the spectrogram.<br>Preserve additivity. |
| Different tasks of speech analysis are carried out separately.<br>Ignore the chicken-and-egg relationship. | A unified probabilistic model to integrate the pitch, energy and spectral envelope. |

## Model Formulation

### Physical Acoustic Tube



### Signal Processing Modeling



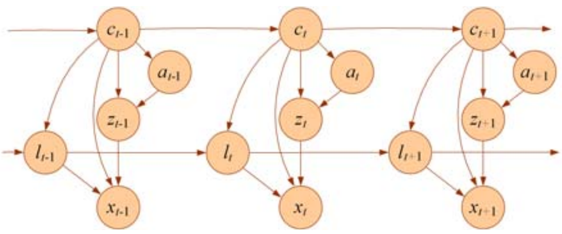

**Voiced Frame:**

$$|X(\omega)| = |[E_t(\omega)H(\omega)] * W(\omega)|$$
$$\approx \sum_n |H(n\eta_t)||W(\omega - n\eta_t)|$$

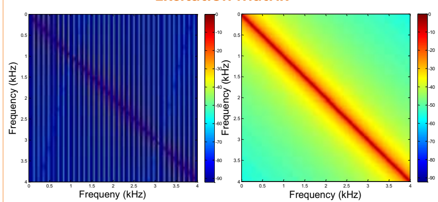
**Unvoiced Frame:**

$$|X(\omega)| = \left|\sum_\xi H(\xi)W(\omega-\xi)\right|$$
$$\approx \sum_\xi |H(\xi)||W(\omega-\xi)|$$

### Probabilistic Modeling




Excitation Matrix




| $c_t$ | $l_t$ | $a_t$ | $z_t$ | $x_t$ |
|---|---|---|---|---|
| Phoneme Cluster | Discretized Frequency | Excitation Gain | DCT Coef. of Spectral Envelope | Observed Spectrogram |
| $p(c_t \mid c_{t-1})$<br>Constrain UV transition | $p(l_t \mid l_{t-1}, c_t)$<br>Constrain abrupt changes in pitch | $p(a_t \mid c_t = c)$<br>$= \mathcal{N}(a_t; m_c, \sigma_c^2)$ | $p(z_t \mid c_t = c, a_t)$<br>$= \mathcal{N}(z_t; a_t\mu_c, \Phi_c)$<br>$\mu_c'\mu_c = 1$ | $x_t = E_{l_t} C z_t + n_t$<br>$p(n_t \mid c_t = c)$<br>$= \mathcal{N}(n_t; 0, m_c^2\Psi)$ |

**The parameters of PAT - $\Theta \triangleq \{\mu_c, \Phi_c, \Psi, m_c, \sigma_c^2\}$ can be solved using the EM algorithm**

## Experimental Results

**The capability of PAT is demonstrated for a number of speech analysis/synthesis tasks.**

### Pitch Tracking

Pitch tracking result

| | PAT | Get_f0 |
|---|---|---|
| UE (%) | 5.38 | 8.84 |
| VE (%) | 4.83 | 4.29 |
| GPE (%) | 0.91 | 2.86 |
| RMS (Hz) | 5.46 | 5.83 |

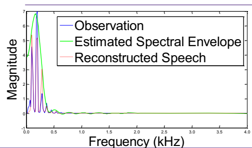
Pitch tracking result with UV labeling

| | PAT | Get_f0 |
|---|---|---|
| GPE (%) | 1.51 | 2.07 |
| RMS (Hz) | 5.4556 | 5.7792 |

### Speech Synthesis

MOS grading result

| | PAT | others |
|---|---|---|
| Z_SYNTHESIS vs. LPC | 4.33 | 2.21 |
| Z_SYNTHESIS vs. original | 4.37 | 4.69 |
| MU_SYNTHESIS vs. LPC | 3.24 | 2.31 |
| MU_SYNTHESIS vs. original | 3.34 | 4.98 |

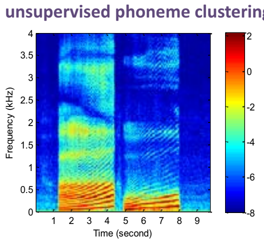



$$x_t^{Z\_SYNTHESIS} = \Gamma_{\hat{l}_t} \cdot E[z_t \mid x_{1:T}]$$

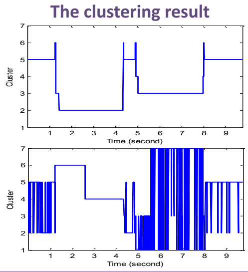
$$x_t^{MU\_SYNTHESIS} = E[a_t \mid x_{1:T}] \cdot \Gamma_{\hat{l}_t} \mu_{\hat{c}_t}$$

### Phoneme Clustering

Speech spectrogram for unsupervised phoneme clustering




The clustering result




### Speech Enhancement

Speech enhancement result for the vowel /ɔ:/