IMPROVEMENT OF PROBABILISTIC ACOUSTIC TUBE MODEL FOR SPEECH DECOMPOSITION

Yang Zhang[1]
yzhan143@illinois.edu

Zhijian Ou[2]
ozj@tsinghua.edu.cn

Mark Hasegawa-Johnson[1]
jhasegaw@Illinois.edu

1 University of Illinois at Urbana Champaign, Urbana IL, USA
Department of Electrical and Computer Engineering

2 Tsinghua University, Beijing, China
Department of Electronic Engineering

Motivation

Drawbacks of Current Model-based Methods

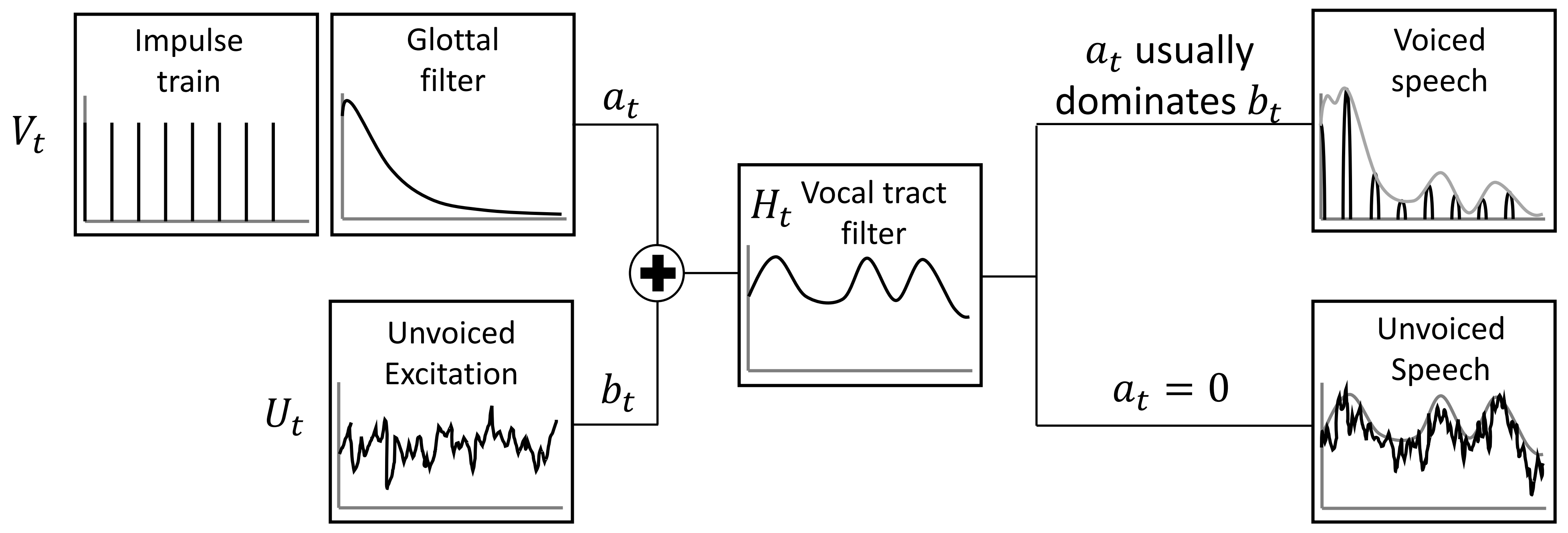
- **Incomplete** - tend to model only a part of parameters of interest, and disregard others that might also be important.
- Speech analysis may be **inaccurate** or even **incorrect**:
  - Chicken and egg effect;
  - LPC and MFCC corrupted by spectral tilt.

Highlights of PAT2

- A **probabilistic generative model** that **jointly** considers all speech parameters;
- Incorporates **breathiness** and **glottal vibration**;
- Incorporates **phase modeling** and so completely defines a probabilistic model for the complex spectrum of speech;
- **Makes U/V states a continuum** by introducing voiced amplitude and unvoiced amplitude, which is closer to the nature of speech.

PAT2 Signal Modeling

The Source Filter Model with Mixed Excitation



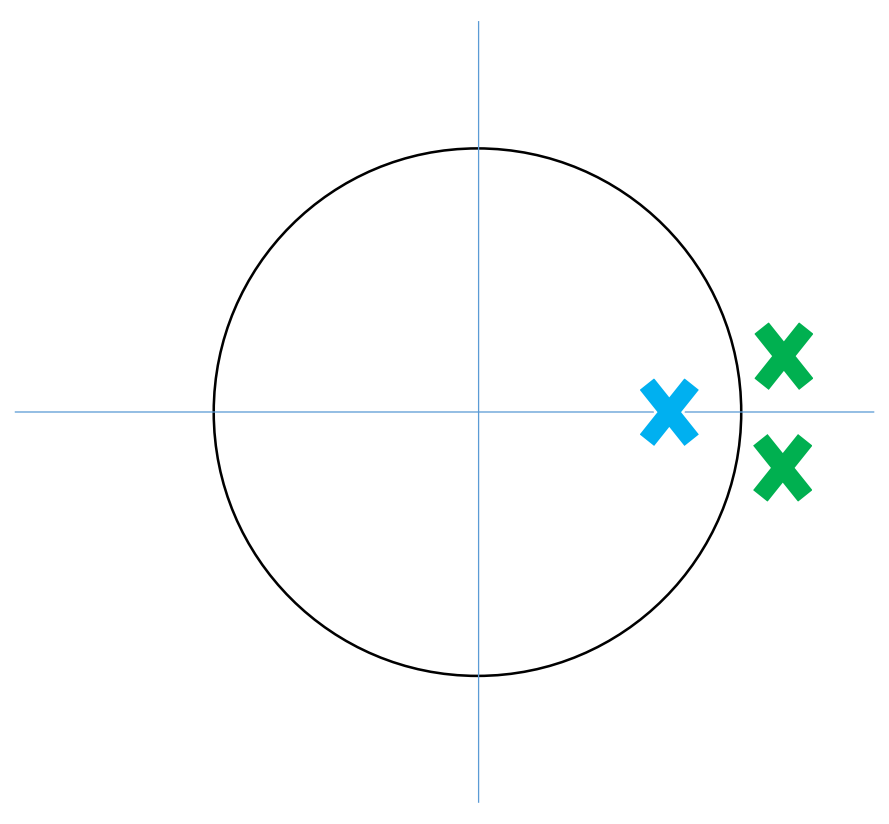
speech — voiced excitation — unvoiced excitation — vocal tract transfer function — windowing function

$$S_t(\omega) = [a_t V_t(\omega) + b_t U_t(\omega)] H_t(\omega) \circledast W_t(\omega)$$

$$V_t(\omega) = G_t(\omega) e^{-j\omega\tau_t} \sum_k \delta(\omega - k\omega_{0t})$$

glottal transfer function — group delay — dirac delta function — fundamental frequency

Glottal Filter



magnitude & phase of the anti-causal poles

$$G_t(\omega) = \left(1 + g_{1t}\cos\beta_t e^{-jw} + g_{1t}^2 e^{-2j\omega}\right)^{-1} \left(1 + g_{2t} e^{-j\omega}\right)^{-1}$$

magnitude of the causal pole

Vocal Tract Filter

When **negative sign** and **group delay** is removed, complex cepstrum decay at the rate of $1/\hat{n}$.

mel-frequency **complex** cepstral coefficients

$$H_t(\omega) = \exp\left[\sum_{\hat{n}=1}^{K} \hat{h}[\hat{n}] \exp(-jm(\omega)\hat{n})\right]$$

quefrency — mel-frequency

PAT2 Probabilistic Modeling

Convert to DFT and Vectorize

$$\boldsymbol{s}_t = a_t \boldsymbol{\xi}_t + b_t \boldsymbol{\eta}_t$$

$$\boldsymbol{\eta}_t \sim \mathcal{N}(0, \boldsymbol{H}_t)$$

Convert to Mel Frequency

$$\tilde{\boldsymbol{s}}_t = \boldsymbol{F}\boldsymbol{s}_t = a_t \boldsymbol{F}\boldsymbol{\xi}_t + b_t \boldsymbol{F}\boldsymbol{\eta}_t$$

$$\tilde{\boldsymbol{s}}_t \sim \mathcal{N}(a_t \boldsymbol{F}\boldsymbol{\xi}_t, b_t^2 \boldsymbol{F}\boldsymbol{H}_t\boldsymbol{F}^T)$$

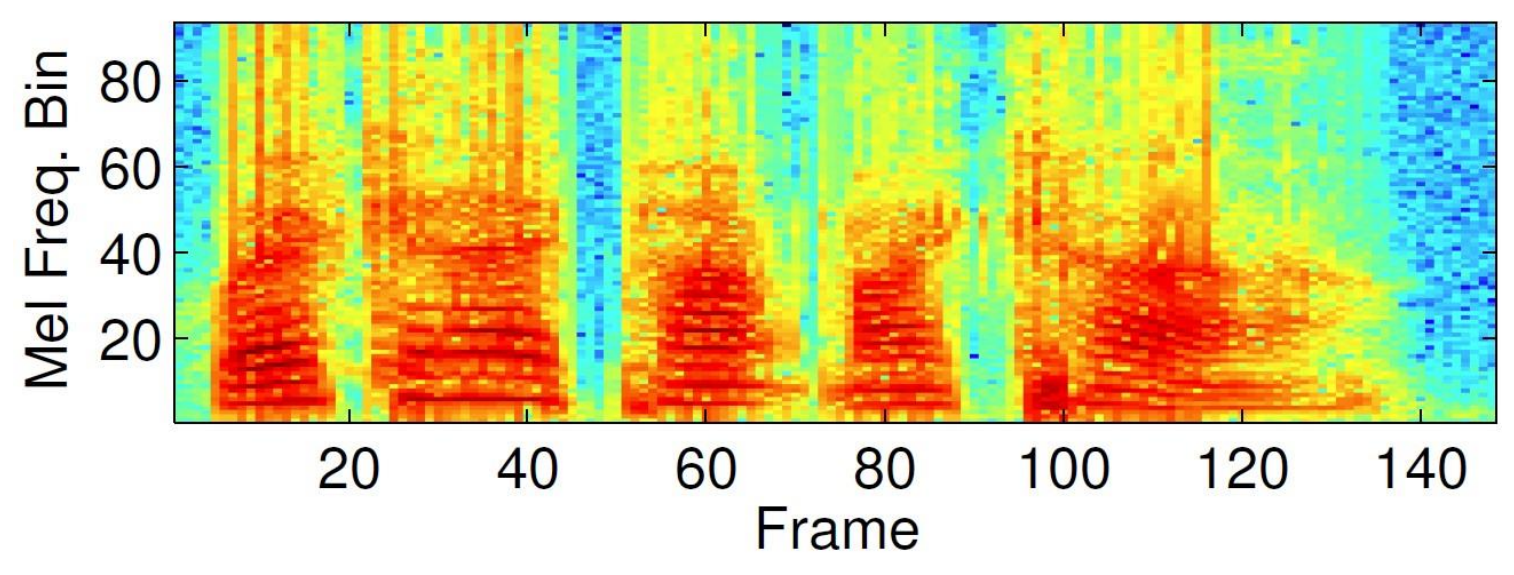
Add Prior

$$P_{\theta_t|\theta_{t-1}}(u|v) \propto -\frac{(u-v)^2}{\sigma_\theta^2}$$

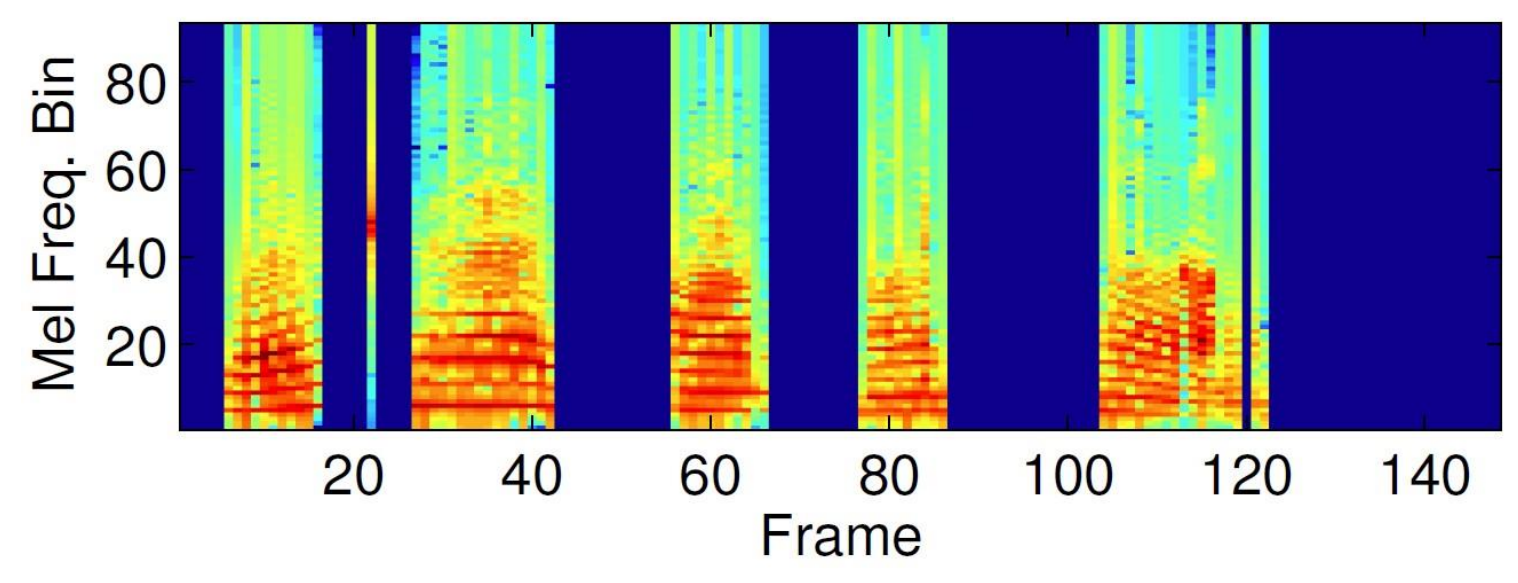
Analysis with **MAP**

Experimental Results

Voiced Reconstruction

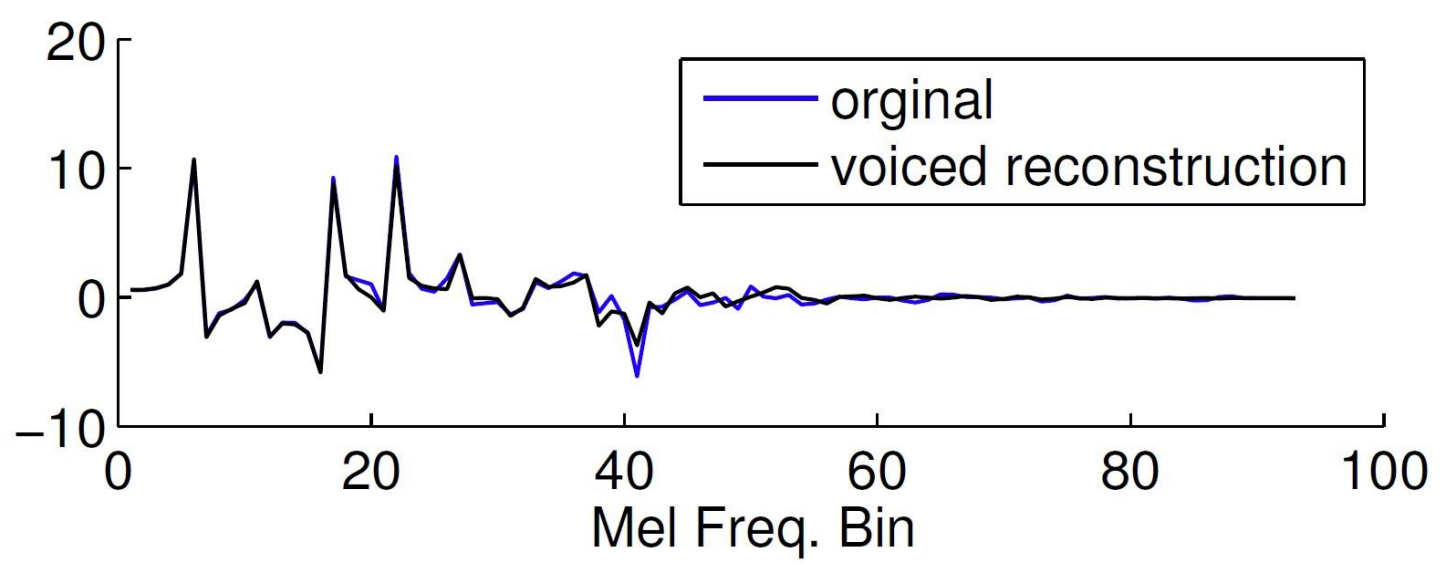


Original Spectrogram

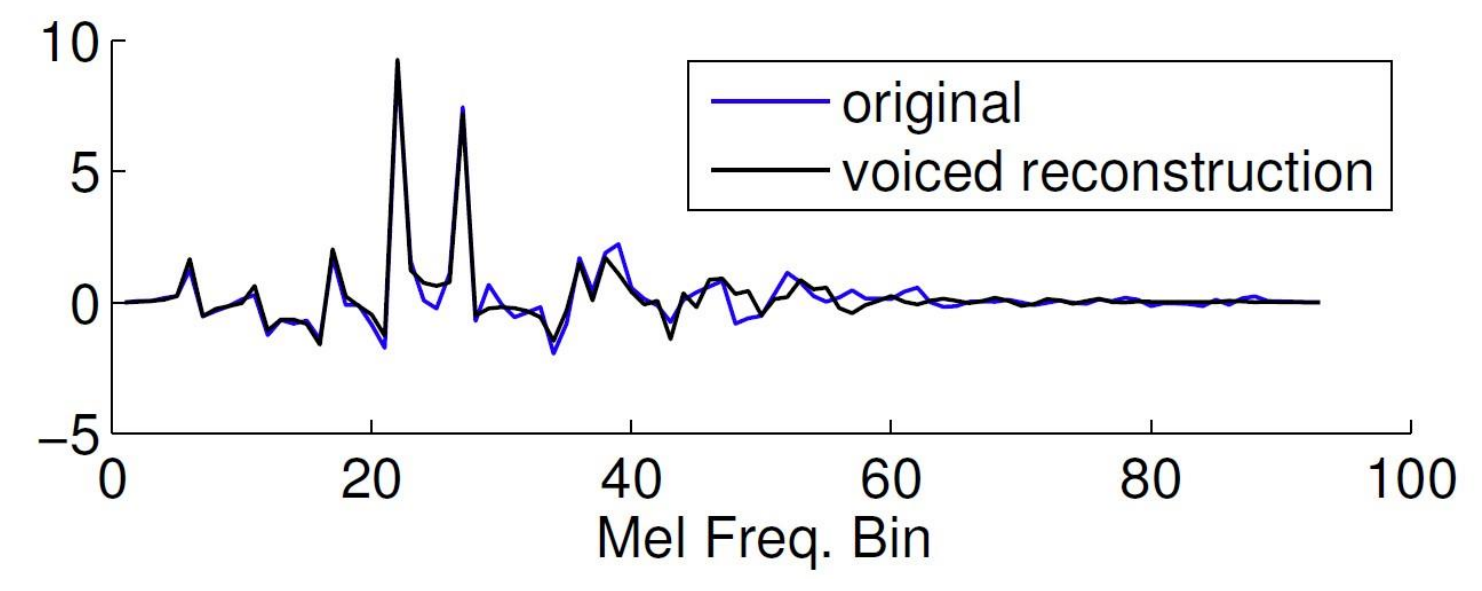


Voiced Reconstruct

Voiced Reconstruction – Single Frame

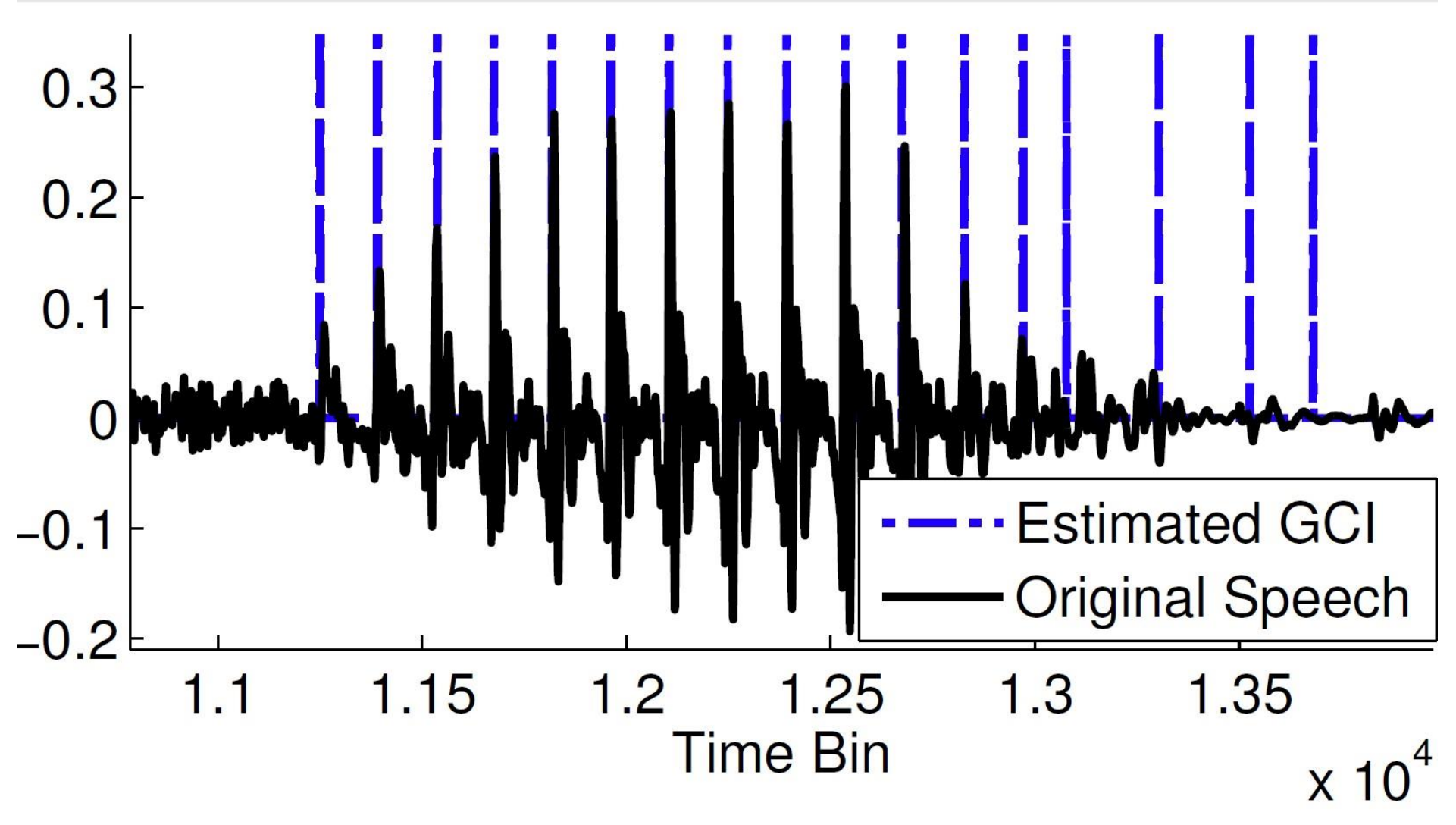


Real Spectrum

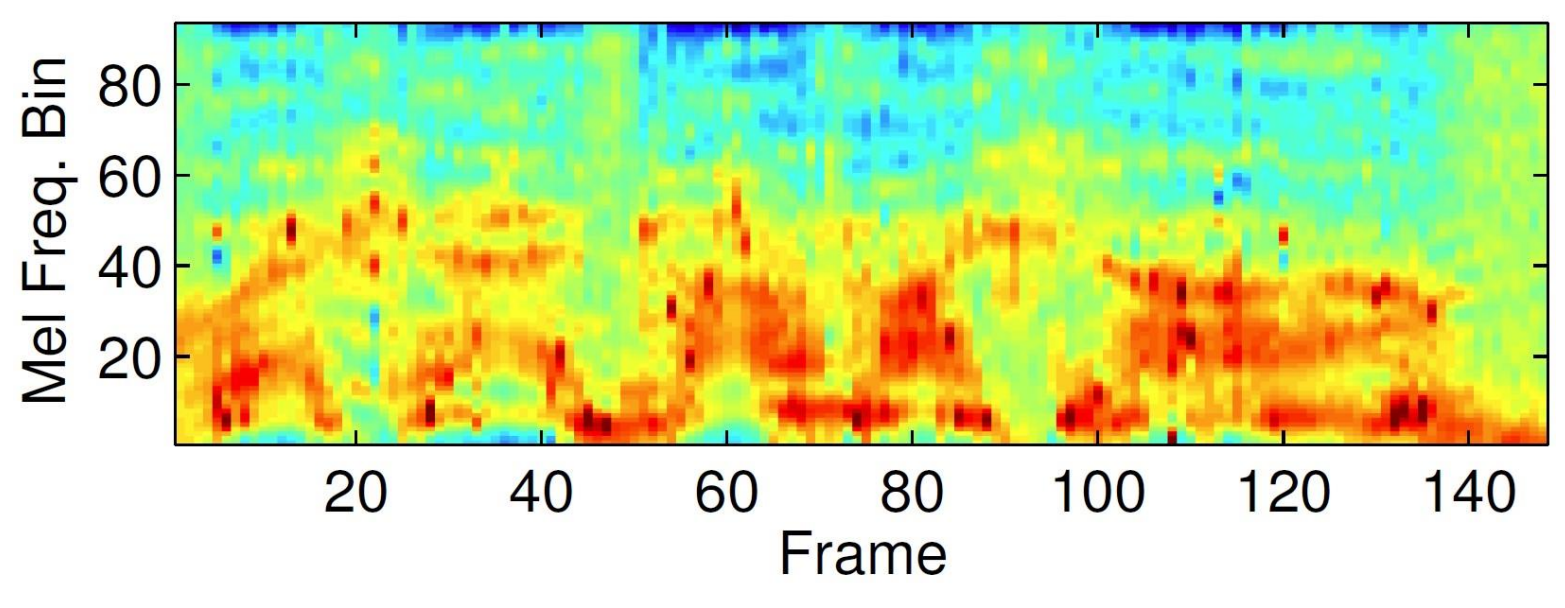


Imaginary Spectrum

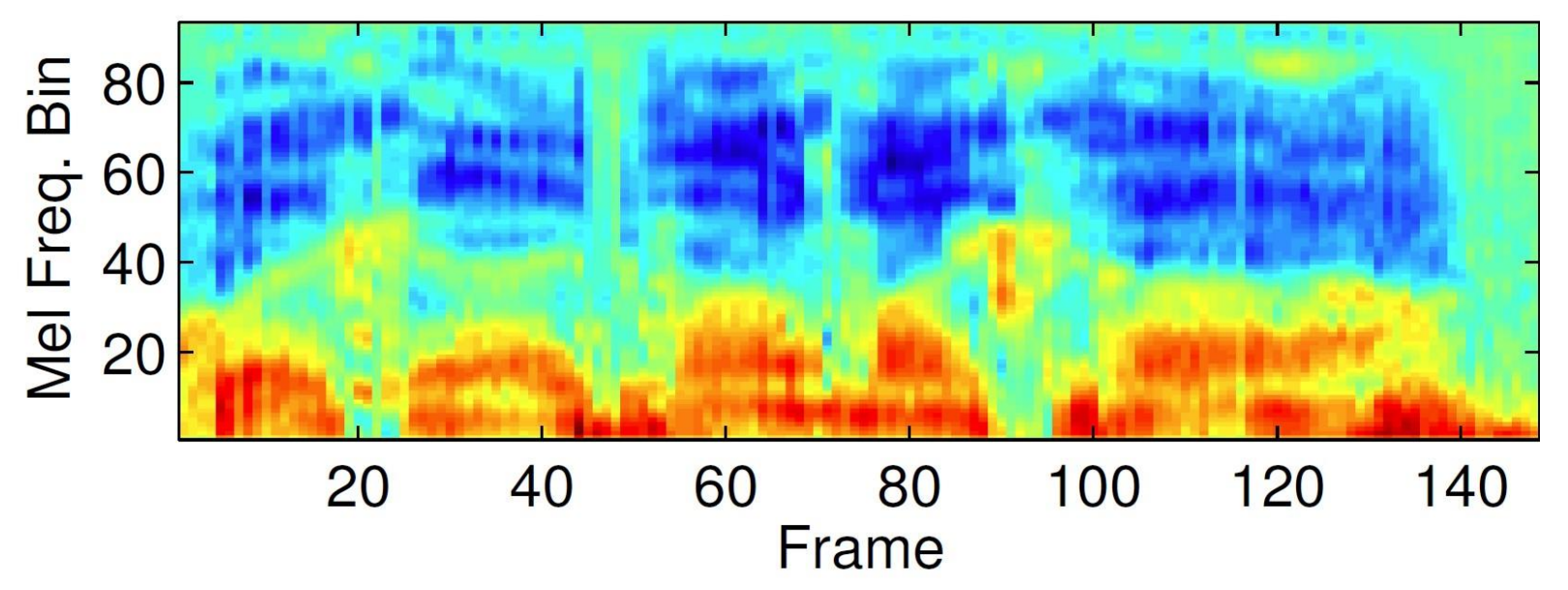
GCI Location Estimation



Vocal Tract Filter Estimation

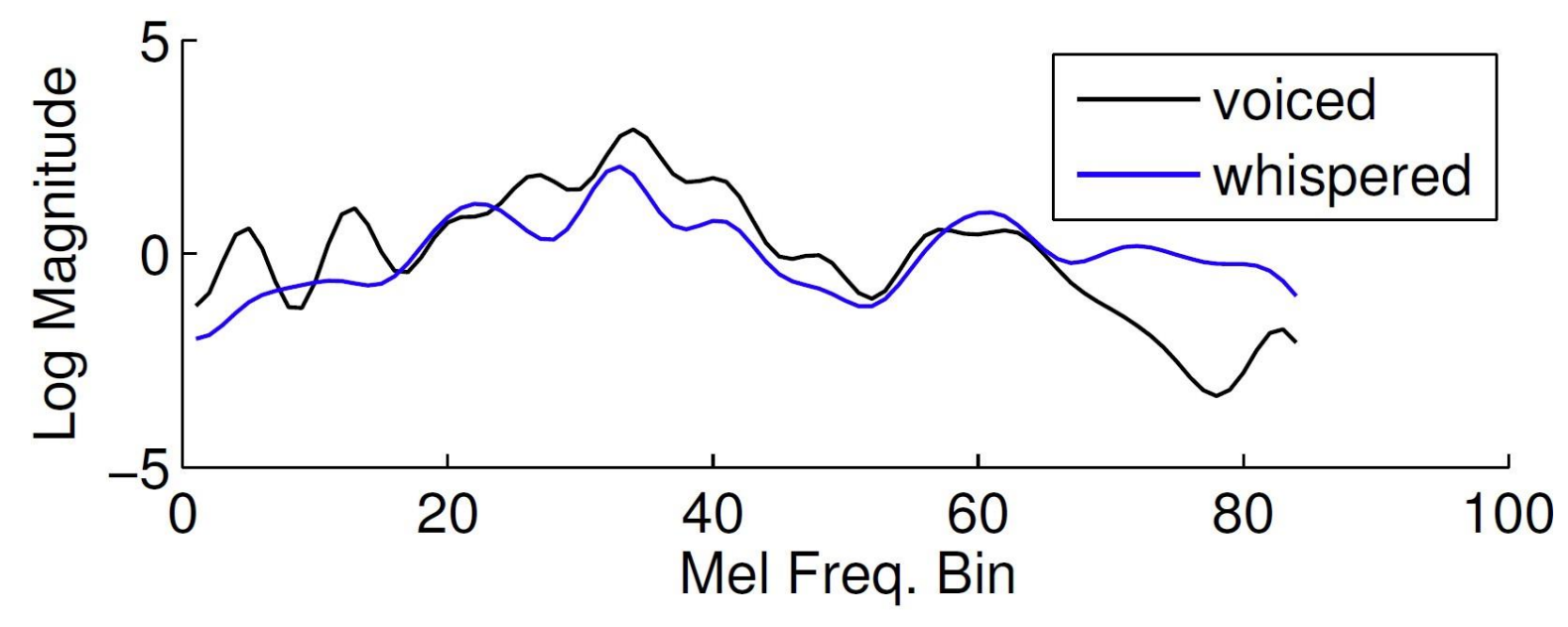


PAT2

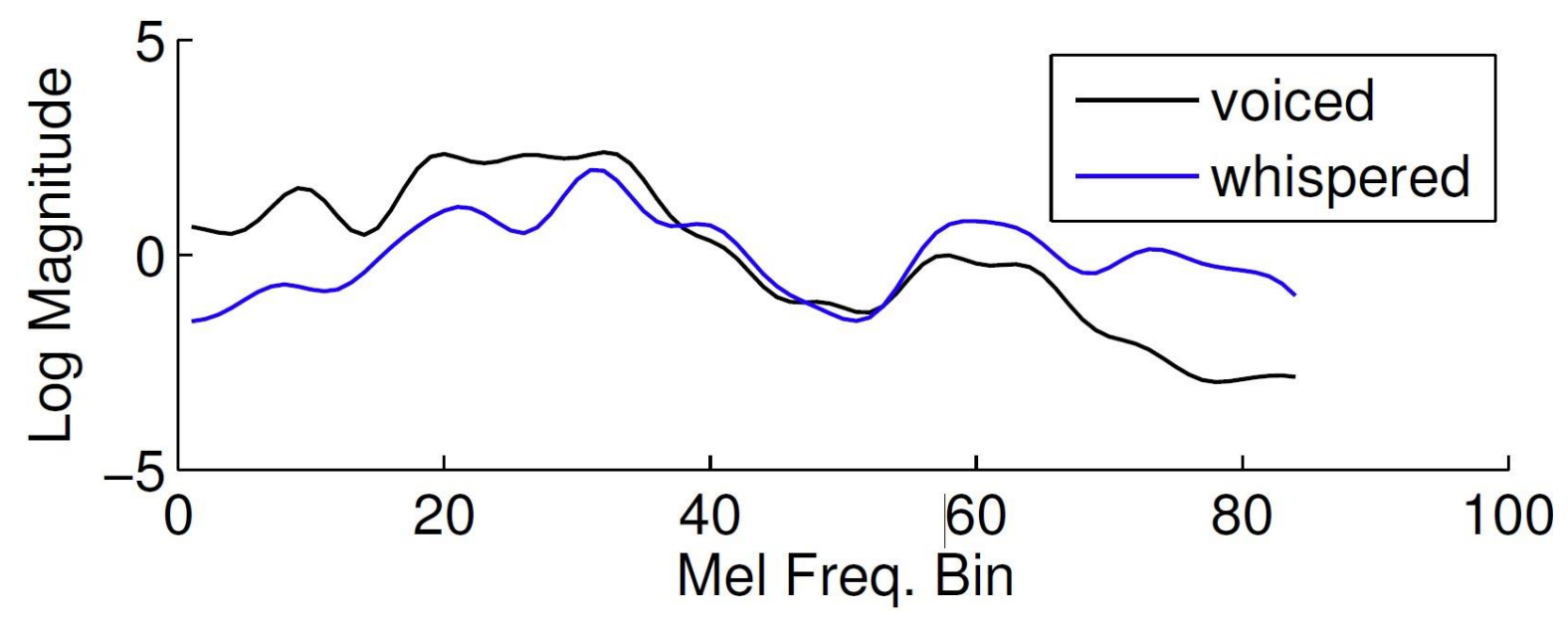


MFCC

Voiced vs Whispered



PAT2



MFCC