

Model-Based Speech Processing: Recognition, Enhancement and Separation

Prof. Hui Jiang

Department of Computer Science and Engineering
York University, Toronto, Ont. M3J 1P3, CANADA

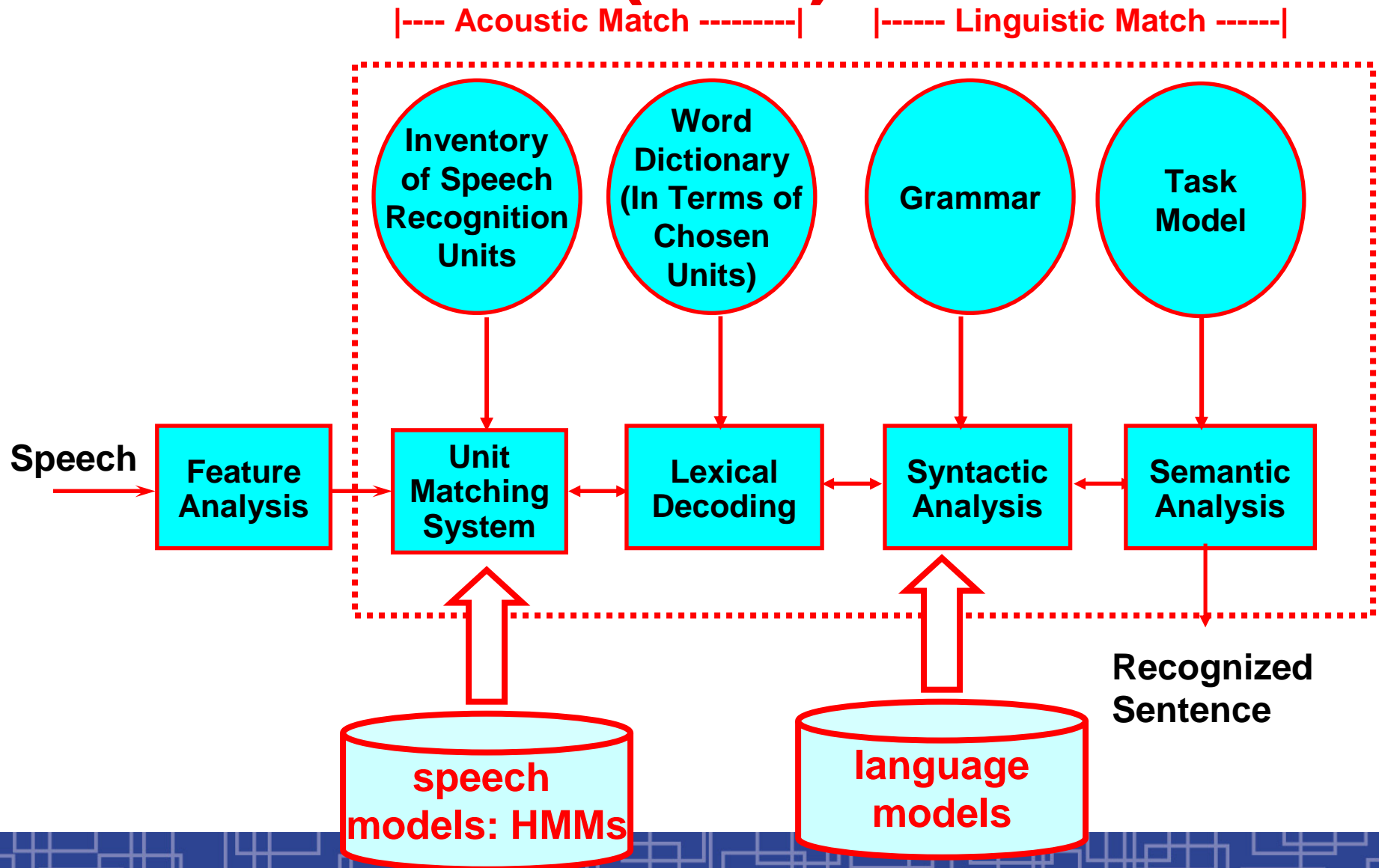
Email: hj@cs.yorku.ca

(Joint work with Guangji Shi, Qi Wang, Qian-Jie Fu)

Outline

- **Why Speech models?**
- **Single-microphone model-based noise removal**
 - **Two methods: nonlinear vs. piece-wise linear approximation**
 - **Experiments:**
 - **In-car hands-free speech recognition**
 - **Speech enhancement**
 - **Cochlear implant processing**
- **Dual-microphone model-based speech processing**
 - **Time-Frequency (T-F) phase-error filtering**
 - **Model-based adaptive phase-error filtering**
 - **Experiments:**
 - **noisy speech recognition**
 - **speech enhancement (ongoing)**
 - **speech separation (ongoing)**
- **Conclusions**

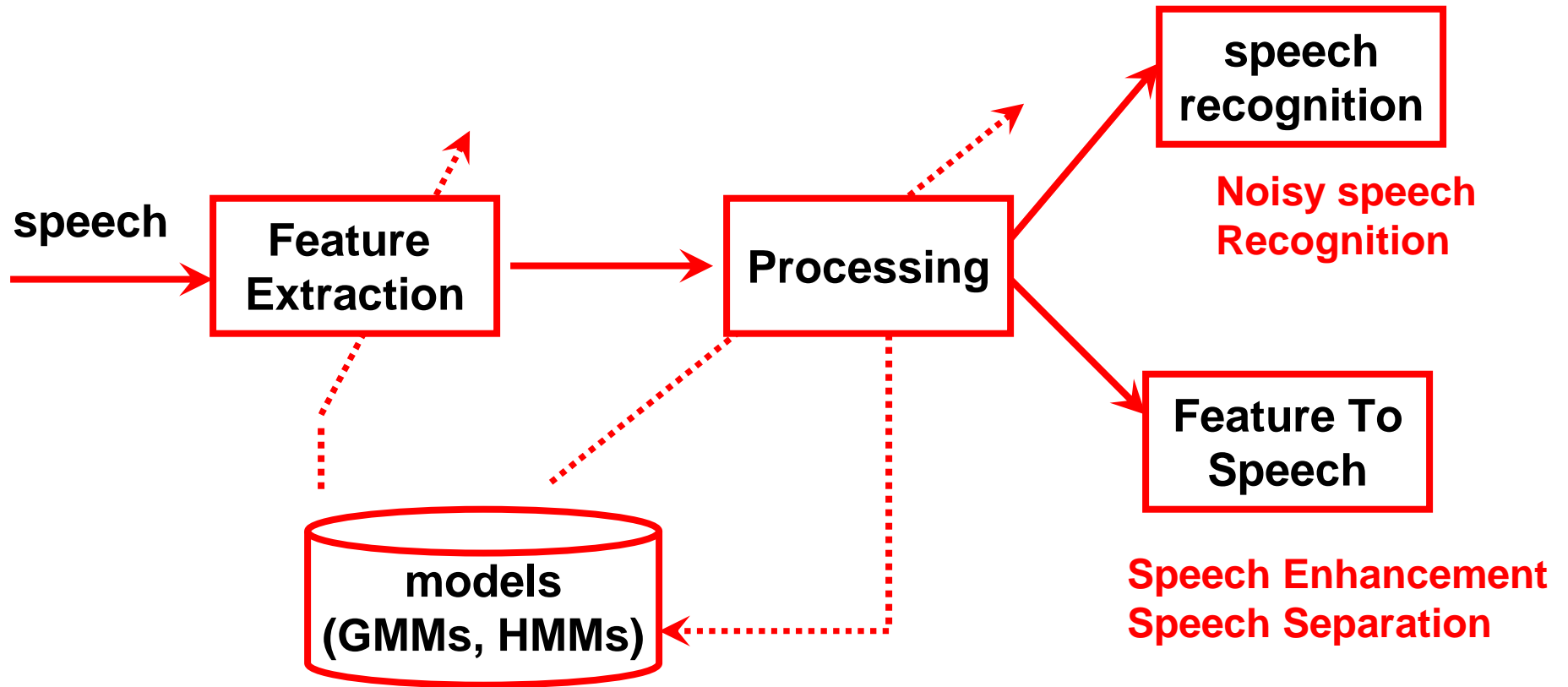
Automatic Speech Recognition (ASR)



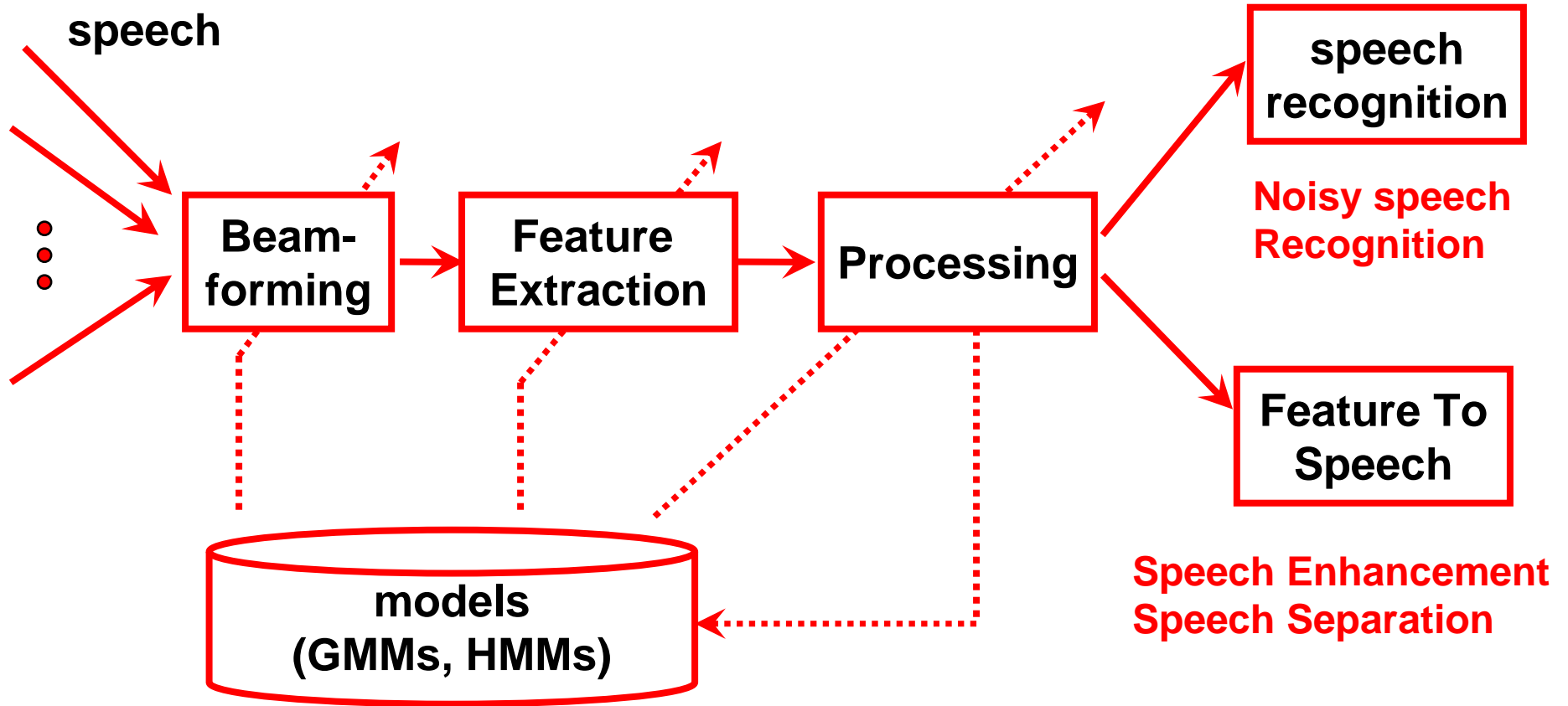
Why Speech Models?

- **Automatic learning from data**
- **Models: computationally usable knowledge**
 - **HMM: hidden Markov models**
 - **GMM: Gaussian mixture models**
- **Effective and efficient learning algorithms**
- **Huge success in speech recognition, language processing**
- **Extend to other speech processing tasks: noise removal, speech enhancement, speech separation, etc.**

Single-Microphone based



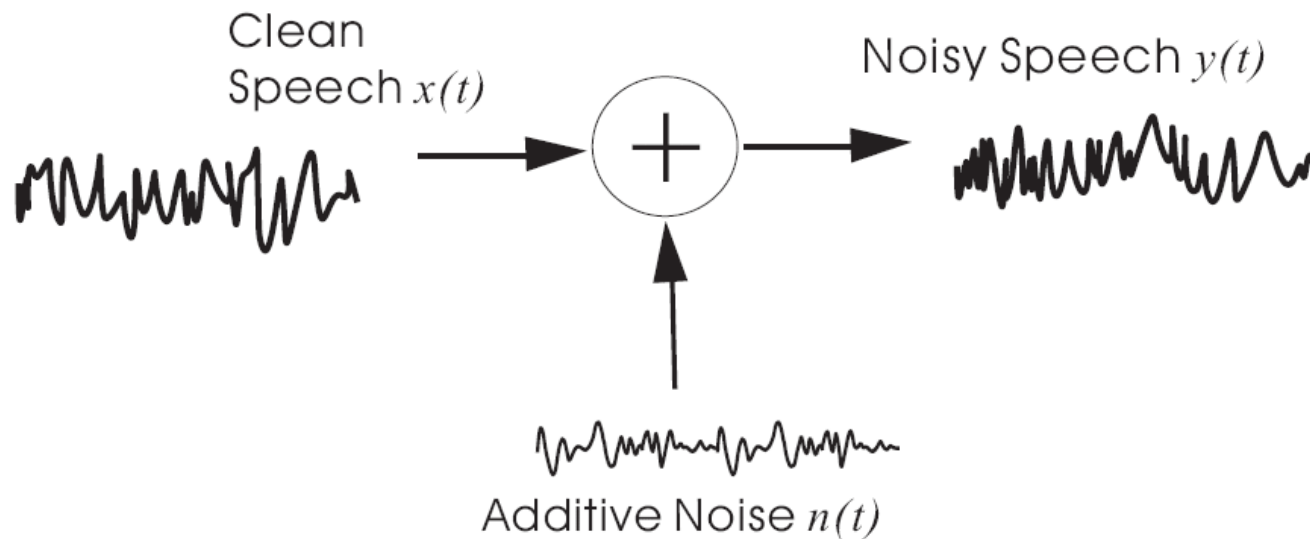
Multiple-Microphone based



Single-Microphone Model-Based Noise Removal

- Traditional approaches: spectral subtraction, Wiener Filtering, etc.
- Environmental Models for additive noises
- Models in the MFCC domain
- Noise Removal in Feature domain
 - Strict non-linear compensation
 - Piece-wise Linear Approximation with Vector Taylor Series
- Convert clean feature to clean speech
- Experiments:
 - In-car hands-free speech recognition
 - Speech enhancement
 - Cochlear implant processing

Environmental Model for additive noise



$$y(t) = x(t) + n(t)$$

Models in the MFCC domain

- **MFCC: Mel-Frequency Cepstral Coefficients**
- **MFCC: approx. log-spectrum + DCT**
- **The most effective speech models are in the MFCC domain**
 - **log → small spectrum dynamic range → more reliable spectrum estimation**
 - **DCT → diagonal covariance matrix**

What's MFCC?

Step 1: Mel-Scale Filter Bank Processing

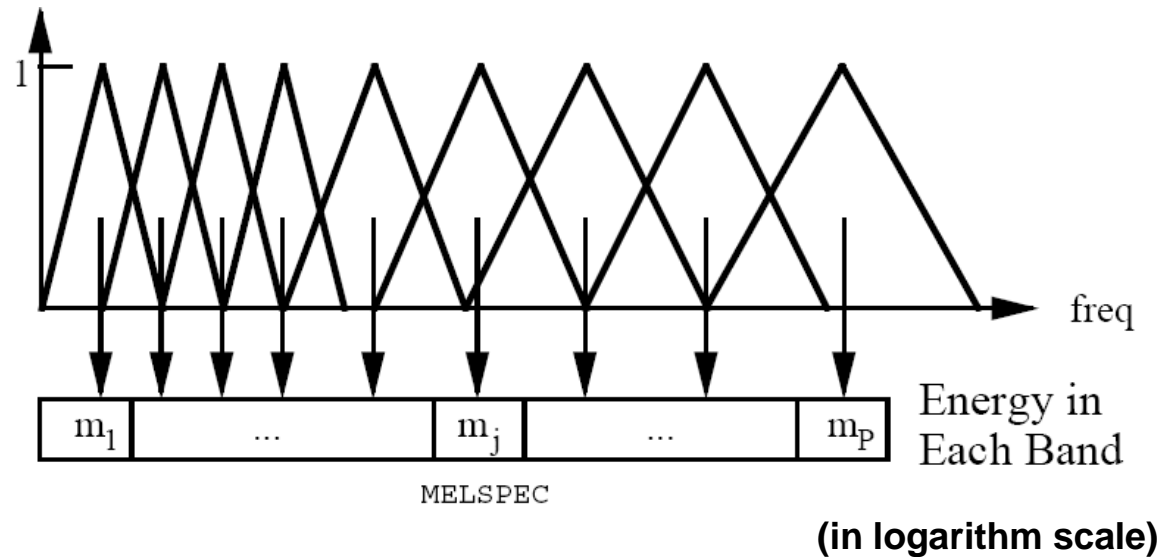


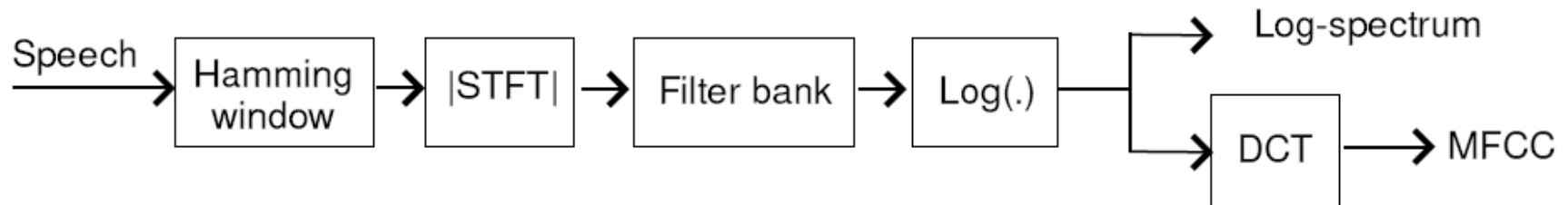
Fig. 5.3 Mel-Scale Filter Bank

Step 2: DCT (Discrete Cosine Transform) to de-correlate

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos \left(\frac{\pi i}{N} (j - 0.5) \right)$$

MFCC: Mel-Frequency Cepstral Coefficients

Environmental model in the Log-Spectrum domain



$$y(t) = x(t) + n(t)$$

$$Y(\omega) = X(\omega) + N(\omega)$$

$$\log Y(\omega) = \log[X(\omega) + N(\omega)]$$

$$Y = X + \log(1 + e^{(N-X)})$$

Statistical Models for Noise Compensation

- Assume clean speech \mathbf{x} is modeled by a Gaussian mixture model (GMM), $p(\mathbf{x})$, and noise is modeled a single Gaussian, $p(\mathbf{n})$:

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_{xk}, \sigma_{xk}^2) = \sum_{k=1}^K w_k \cdot \prod_{d=1}^D \frac{1}{\sqrt{2\pi\sigma_{xkd}^2}} \cdot e^{-\frac{(x_d - \mu_{xkd})^2}{2\sigma_{xkd}^2}}$$

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n} \mid \mu_n, \sigma_n^2) = \prod_{d=1}^D \frac{1}{\sqrt{2\pi\sigma_{nd}^2}} \cdot e^{-\frac{(n_d - \mu_{nd})^2}{2\sigma_{nd}^2}}$$

- Clean speech model $p(\mathbf{x})$ is estimated from some clean speech data beforehand.
- Noise speech model $p(\mathbf{n})$ is estimated for each noisy speech utterance.

Noisy Compensation (I)

- Problem: given a noisy feature $Y_0 \rightarrow$ clean estimate X
- Minimum Mean Square Error (MMSE) estimation:

$$\hat{X} = E[X | Y_0] = \int X \cdot p(X | Y_0) dX$$

$$\begin{aligned}\hat{\mathbf{x}} &= E_{\mathbf{x}}[\mathbf{x} | \mathbf{y}_0] = \int \int \mathbf{x} \cdot p(\mathbf{x} | \mathbf{y}_0) d\mathbf{x} \\ &= \int \int \frac{\mathbf{x} \cdot p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x})}{p(\mathbf{y}_0)} d\mathbf{x} = \frac{\int \int \mathbf{x} \cdot p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x}) d\mathbf{x}}{\int \int p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x}) d\mathbf{x}} \\ &= \frac{\sum_{k=1}^K w_k \int \int \mathbf{x} \cdot \mathcal{N}(\mathbf{x} | \mu_{xk}, \sigma_{xk}^2) \cdot p(\mathbf{y}_0|\mathbf{x}) d\mathbf{x}}{\sum_{k=1}^K w_k \int \int \mathcal{N}(\mathbf{x} | \mu_{xk}, \sigma_{xk}^2) \cdot p(\mathbf{y}_0|\mathbf{x}) d\mathbf{x}}\end{aligned}$$

Noisy Compensation (II)

- How to estimate $p(Y_0|X)$?

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &\equiv \left| \frac{d\mathbf{n}}{d\mathbf{y}} \right| \cdot p(\mathbf{n}) \Big|_{\mathbf{n}=\mathbf{x}+\ln(e^{\mathbf{y}-\mathbf{x}}-1)} \\ &= \prod_{d=1}^D \left| \frac{dn_d}{dy_d} \right| \cdot p(n_d) \Big|_{n_d=x_d+\ln(e^{y_d-x_d}-1)} \\ &= \prod_{d=1}^D \frac{1}{\sqrt{2\pi\sigma_{nd}^2}} \cdot \frac{\psi(x_d, y_d)}{\psi(x_d, y_d) - 1} \cdot e^{-\frac{[x_d - \mu_{nd} + \ln(\psi(x_d, y_d) - 1)]^2}{2\sigma_{nd}^2}} \end{aligned}$$

where $\psi(x, y) = e^{y-x}$

- Involve integral of complex non-linear functions

Method I: Numeral Method

- Uniformly partition each feature dimension into many intervals

$$l_{kd} = x_{kd0} < x_{kd1} < x_{kd2} < \dots < x_{kdJ-1} < x_{kdJ} = u_{kd}$$

- Numerical integral:

$$\hat{x}_e = \frac{\sum_{k=1}^K w_k \cdot \mathcal{M}_{ke} \cdot \prod_{d=1, d \neq e}^D \mathcal{N}_{kd}}{\sum_{k=1}^K w_k \cdot \prod_{d=1}^D \mathcal{N}_{kd}}$$

$$\mathcal{M}_{ke} = \Delta_{ke} \left[x_{ke0} \mathcal{U}_k(x_{ke0} | y_{0e}) + x_{keJ} \mathcal{U}_k(x_{keJ} | y_{0e}) + 2 \sum_{j=2}^{J-1} x_{kej} \mathcal{U}_k(x_{kej} | y_{0e}) \right]$$

$$\mathcal{N}_{kd} = \Delta_{kd} \left[\mathcal{U}_k(x_{kd0} | y_{0d}) + \mathcal{U}_k(x_{kdJ} | y_{0d}) + 2 \sum_{j=2}^{J-1} \mathcal{U}_k(x_{kdj} | y_{0d}) \right]$$

Method II: Piece-wise Linear Approximation

- **Nonlinear environmental model:**

$$Y = X + \log(1 + e^{(N-X)})$$

- **Expand log in a point (X_0, N_0) with zero-th order VTS**

$$Y = X + \log(1 + e^{(N_0 - X_0)})$$

- **Piece-wise linear approximation: given clean speech model and noise model, expand it around noise mean μ_n and all clean speech means μ_{xk} :**

$$\mathbf{x} = \mathbf{y} - \ln(1 + e^{\mu_n - \mu_{xk}}) \quad \text{for } k = 1, 2, \dots, K$$

Method II: Piece-wise Linear Approximation

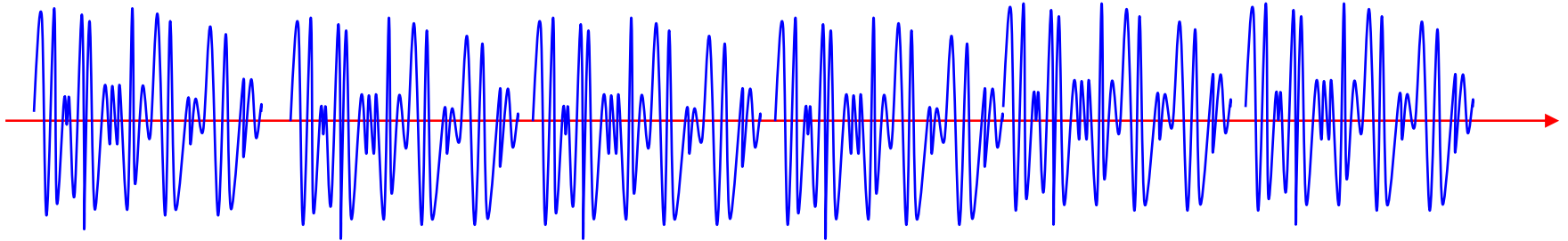
- MMSE estimation of clean speech feature:

$$\hat{\mathbf{x}} = \mathbb{E}_{\mathbf{x}}[\mathbf{x} | \mathbf{y}_0] = \mathbf{y}_0 - \sum_{k=1}^K \text{Pr}(k | \mathbf{y}_0) \cdot \ln(1 + e^{\mu_n - \mu_{xk}})$$

$$\text{Pr}(k | \mathbf{y}_0) = \frac{w_k \cdot \mathcal{N}(\mathbf{y}_0 | \mu_{xk} + \ln(1 + e^{\mu_n - \mu_{xk}}), \sigma_{xk}^2)}{\sum_{k=1}^K w_k \cdot \mathcal{N}(\mathbf{y}_0 | \mu_{xk} + \ln(1 + e^{\mu_n - \mu_{xk}}), \sigma_{xk}^2)}$$

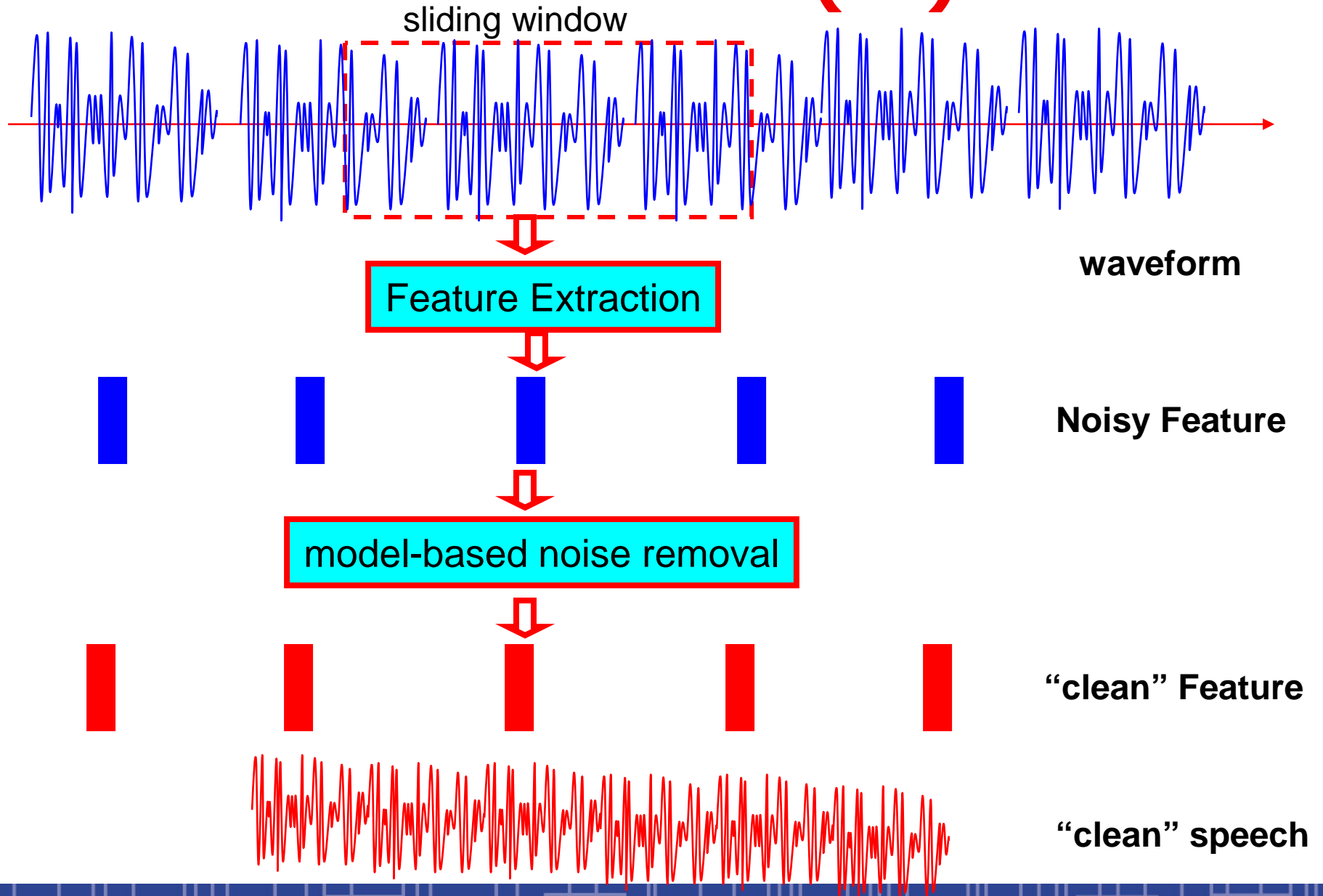
Noise Removal (I)

- Estimate clean speech model (GMM) offline
- For each noisy speech utterance



- Estimate noise model on-line
 - Use beginning silence part
 - Refine with the whole utterance based on EM algorithm

Noise Removal (II)

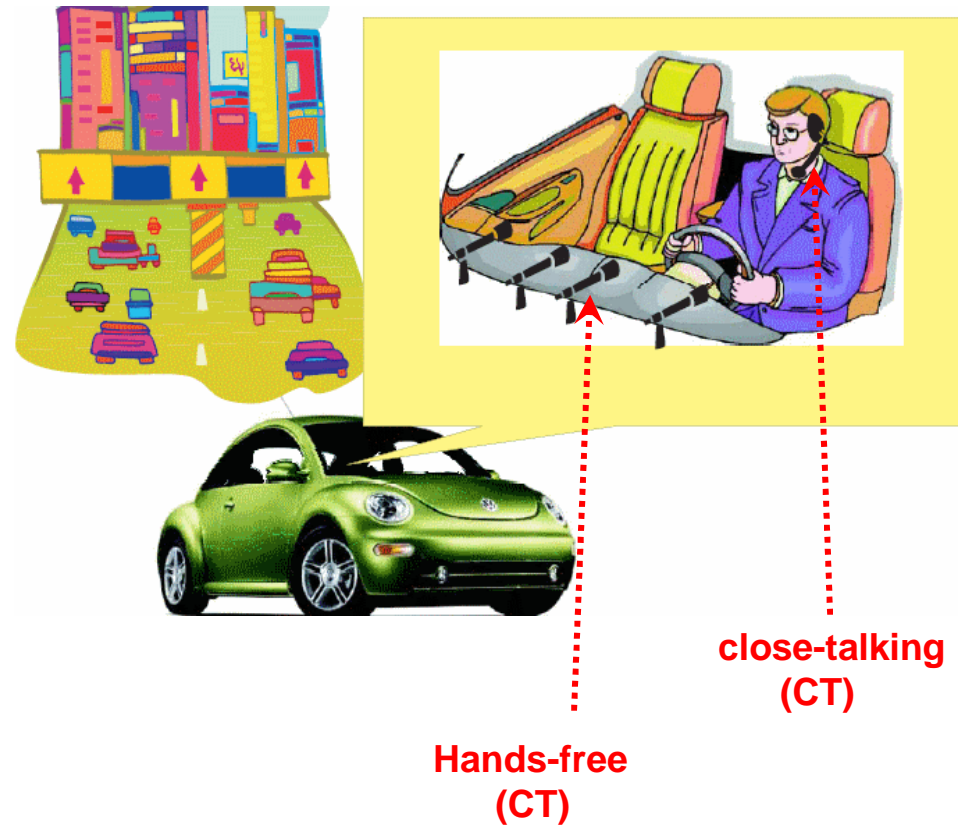


Convert Feature to Speech

- How to convert X back to time domain $x(t)$?
 - DFT of each frame
 - Modify DFT spectrum
 - IDFT (inverse Fourier Transform)
 - Overlapped Adding
- Filter-bank Smoothing: reduce music noises
- Smoothing across consecutive frames: reduce pre-echo

Experiments (I): in-car hands-free speech recognition

- The CARVUI database:
 - collected in a running car
 - Recorded with multiple-microphones
 - Only two microphones channels are used in this experiment: CT + HF
 - 56 speakers (hundreds utterances each speaker)
 - Digit strings + control commands + sentences



Experiments (I): in-car hands-free speech recognition

- Tested with CT + various levels of white noises

SNR	Baseline	Method I: nonlinear	Method II: linear
∞ dB	3.7	4.1	4.0
15dB	30.8	19.6	20.2
10dB	54.2	31.2	32.7
5dB	77.3	50.9	57.3
0dB	87.6	69.4	74.2

(string recognition error rate in %)

Experiments (I): in-car hands-free speech recognition

- Tested with real noisy data: HF data

	Baseline	Method I: nonlinear	Method II: linear
HF	26.6	19.0	19.9

(string recognition error rate in %)

Experiments(II): Speech Enhancement in White Noises

- SNR 25dB
 - Example 1: Noisy → Cleaned
 - Example 2: Noisy → Cleaned
- SNR 15dB
 - Example 1: Noisy → Cleaned
 - Example 2: Noisy → Cleaned
- SNR 9dB
 - Example 1: Noisy → Cleaned
 - Example 2: Noisy → Cleaned

Experiments(II): Speech Enhancement in Speech Babble Noise

- SNR 25dB
 - Example 1: Noisy → Cleaned
 - Example 2: Noisy → Cleaned
- SNR 14dB
 - Example 1: Noisy → Cleaned
 - Example 2: Noisy → Cleaned

More Speech Enhancement Examples: 14dB Babble noise

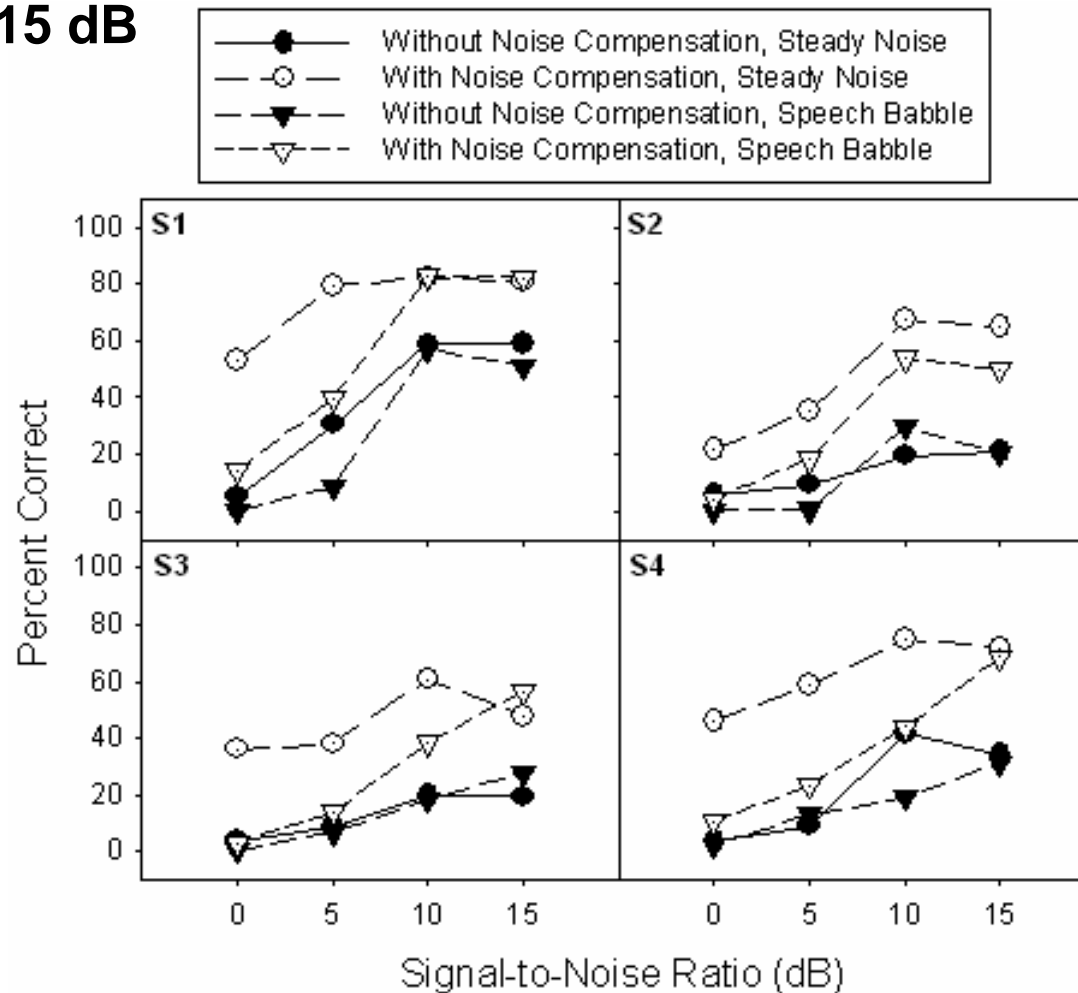
- Example 1: noisy → cleaned
- Example 2: noisy → cleaned
- Example 3: noisy → cleaned
- Example 4: noisy → cleaned
- Example 5: noisy → cleaned
- Example 6: noisy → cleaned

More Speech Enhancement Examples: 9dB White noise

- Example 1: noisy → cleaned
- Example 2: noisy → cleaned
- Example 3: noisy → cleaned
- Example 4: noisy → cleaned
- Example 5: noisy → cleaned
- Example 6: noisy → cleaned
- Example 7: noisy → cleaned

Experiments(III): Cochlear Implant Processing

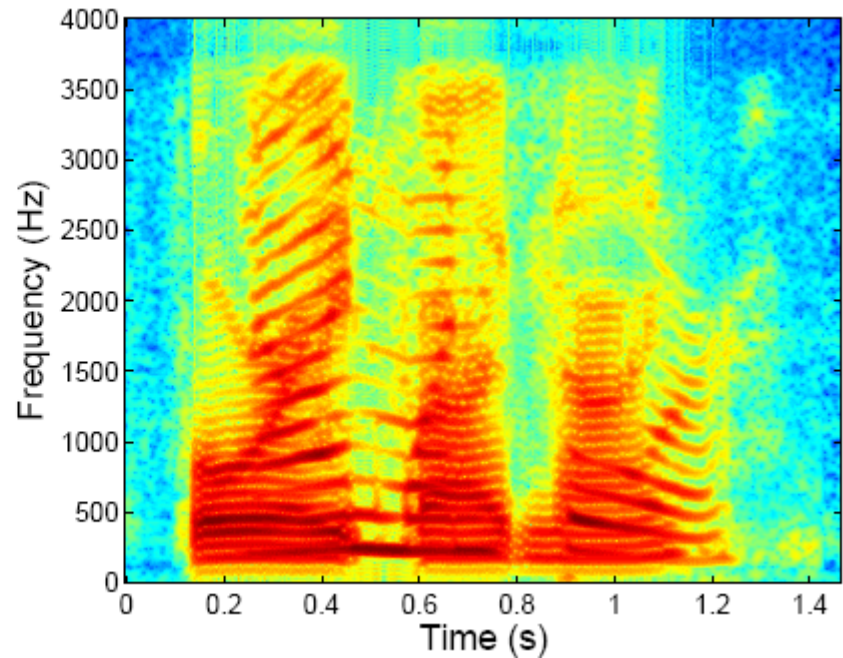
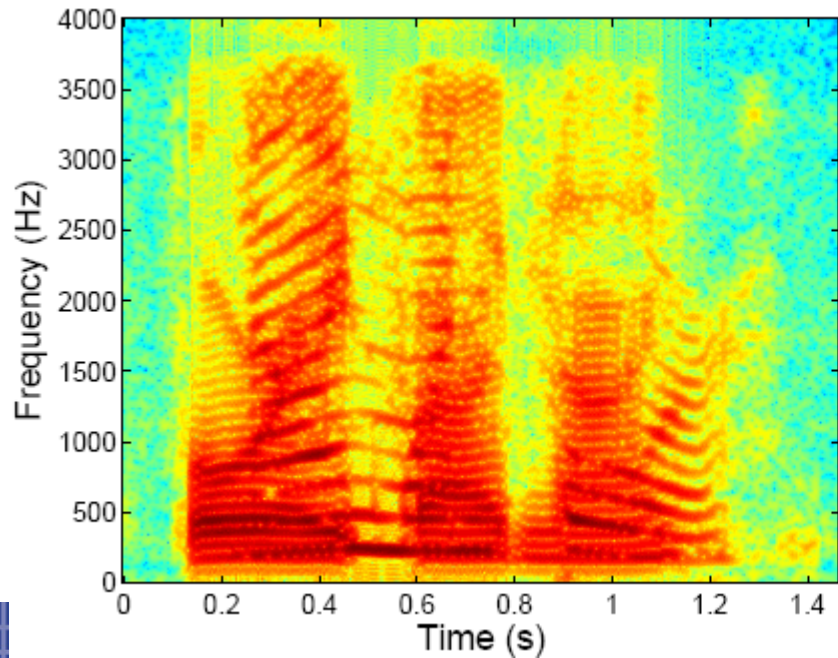
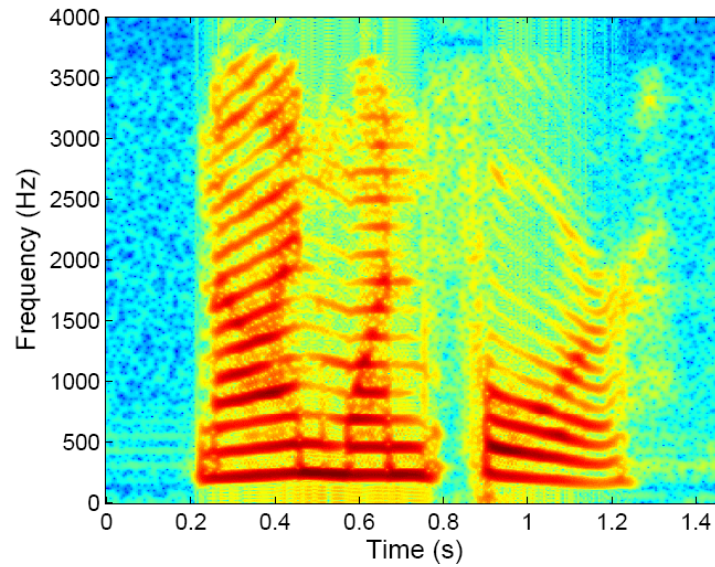
- Four cochlear implant patients are tested with 720 English utterances corrupted with stable and babble noises under SNR=0,5,10,15 dB



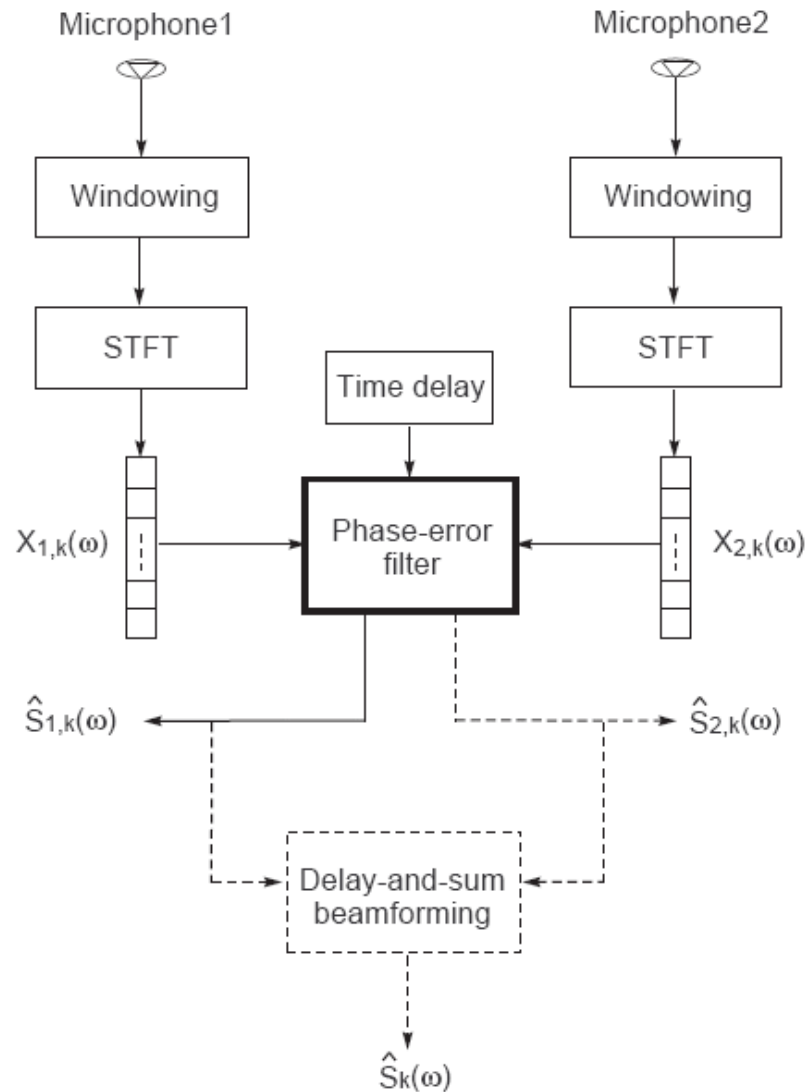
Dual-Microphone Based Speech Processing

- **Traditional approaches:**
 - **Beam-forming for recognition and enhancement**
 - **Independent Component Analysis for separation**
- **Time-Frequency (T-F) masking based on phase error**
- **Model-based adaptive T-F masking**
- **Experiments**
 - **noisy speech recognition**
 - **speech enhancement (ongoing)**
 - **speech separation (ongoing)**

Time-Frequency Masking



Time-Freq Filtering based on phase-error



Phase-Error vs. input SNR

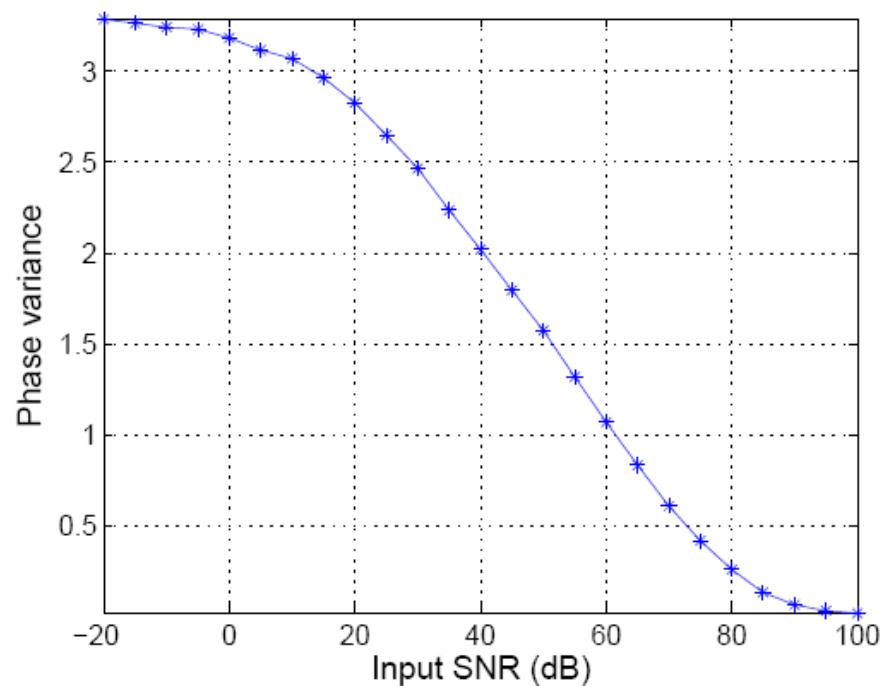


Figure 5.4: Phase variance versus input SNR (Gaussian noise).

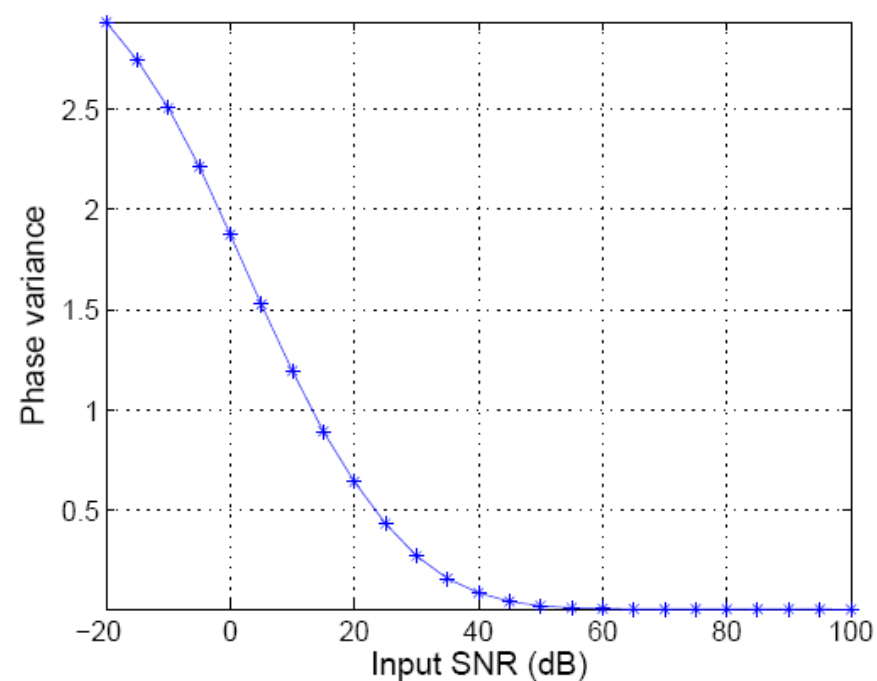
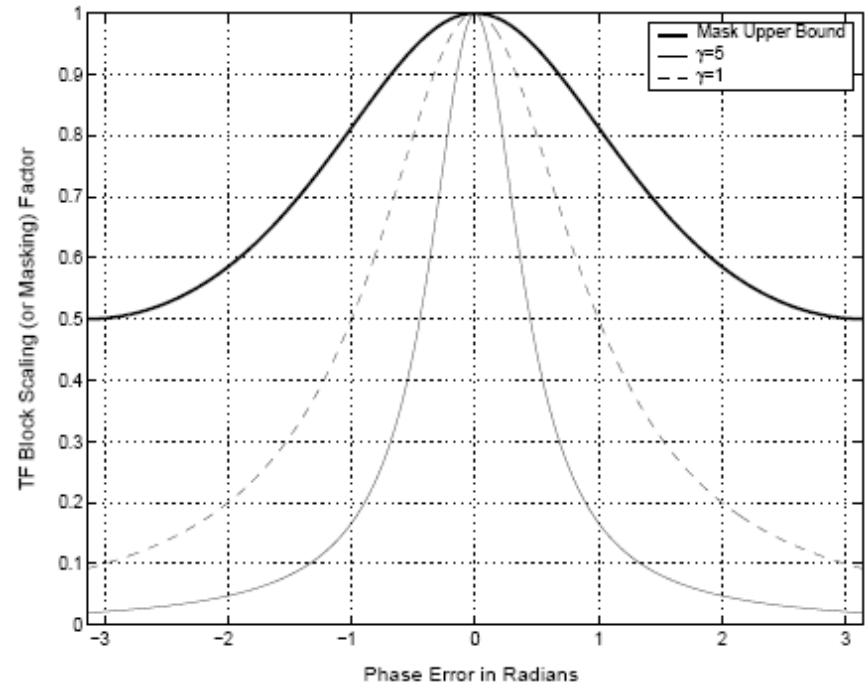


Figure 5.5: Phase variance versus input SNR (speech noise).

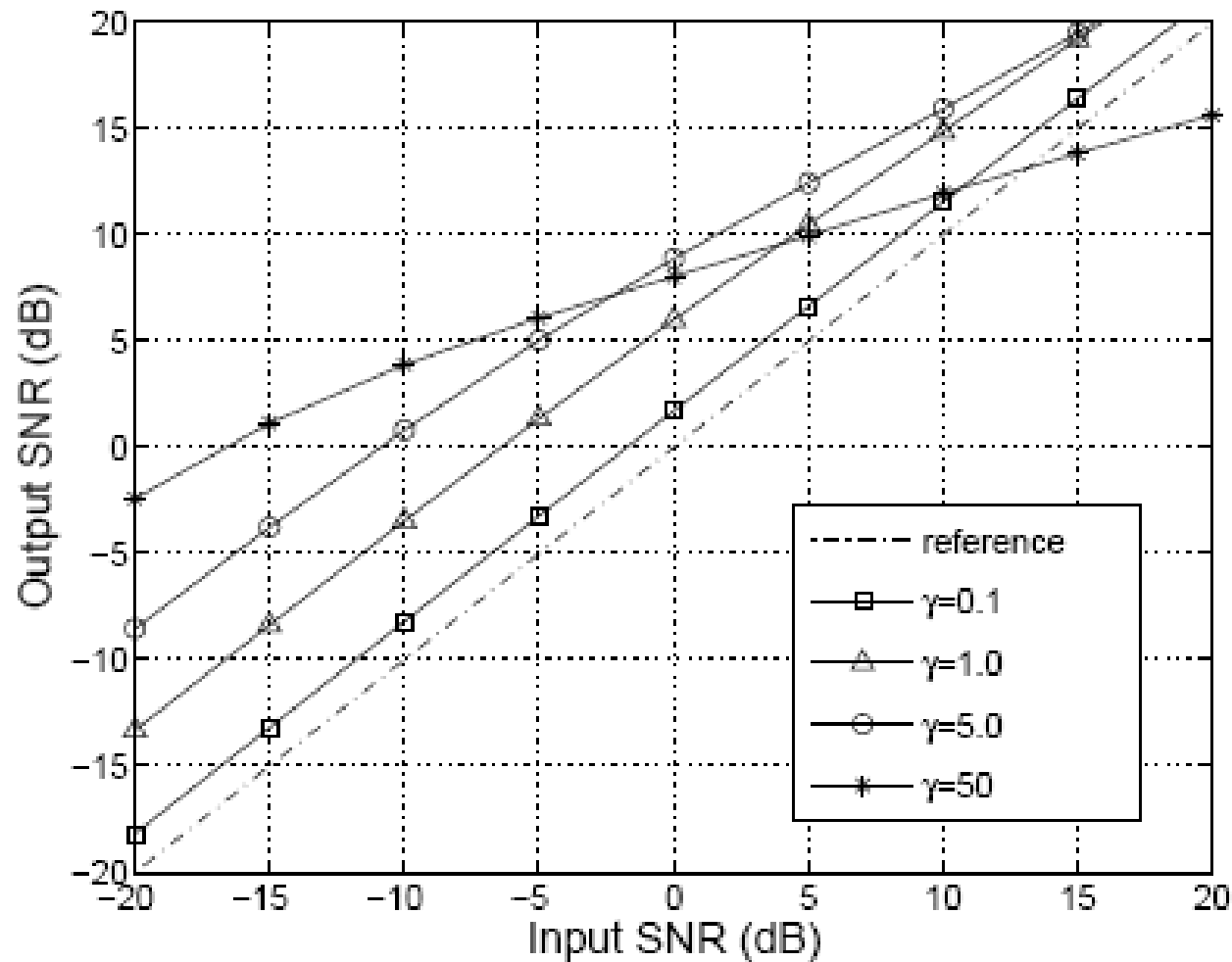
Phase-Error Filter

- Phase error in each T-F block reflects noise level.
- Each T-F block is filtered by

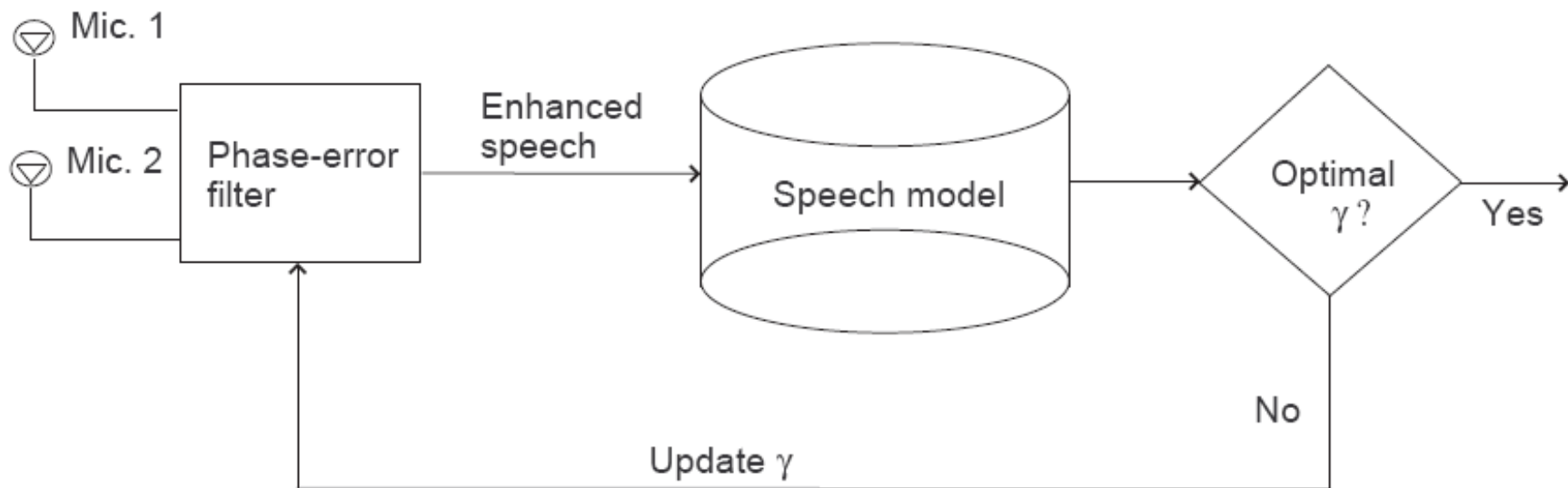
$$\eta(\omega) = \frac{1}{1 + \gamma \cdot \theta_k(\omega)}$$



How to determine γ ?



Adaptive Phase-Error Filtering: Estimating γ with speech model



$$\begin{aligned}\gamma^* &= \arg \max_{\gamma} \mathcal{L}(f_{\gamma}(\mathbf{X})|\gamma, \Lambda_{\tilde{s}}) \\ &= \arg \max_{\gamma} \log \prod_{t=1}^T p(f_{\gamma}(\mathbf{x}_t)|\gamma, \Lambda_{\tilde{s}}) \\ &= \arg \max_{\gamma} \sum_{t=1}^T \log p(f_{\gamma}(\mathbf{x}_t)|\gamma, \Lambda_{\tilde{s}})\end{aligned}$$

Estimate γ with Generalized EM

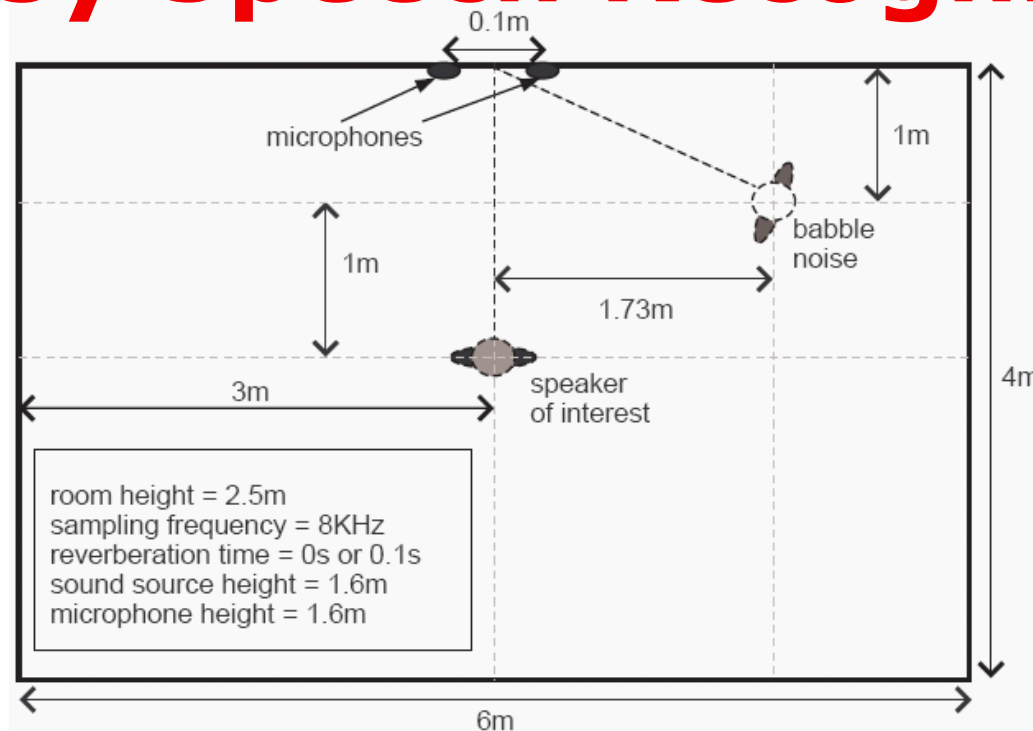
- **E-Step:**
$$Q(\gamma|\gamma^{(k)}) = E [\log p(\mathbf{C}(\gamma), M|\gamma, \Lambda_{\tilde{S}})|\mathbf{C}, \gamma^{(k)}, \Lambda_{\tilde{S}}]$$
$$= \sum_t \sum_m \log p(\mathbf{c}_t(\gamma), m|\gamma, \Lambda_{\tilde{S}})$$
$$\times p(m|\mathbf{c}_t, \gamma^{(k)}, \Lambda_{\tilde{S}})$$

- **M-Step:**

$$\gamma^{(k+1)} = \arg \max_{\gamma} Q(\gamma|\gamma^{(k)})$$

$$\gamma^{(k+1)} = \gamma^{(k)} + \eta Q'(\gamma|\gamma^{(k)})|_{\gamma=\gamma^{(k)}}$$

Experiments: Noisy Speech Recognition



- Adding noise at SNR=20, 15, 10, 5, 0, -5 dB.
- Speaker-independent digit string and command recognition.
- Speech models are trained with clean speech data.

Experiments: Noisy Speech Recognition

SNR	baseline	delay-sum beam-forming	GEM phase-err filtering
-5dB	99.2	85.9	51.5
0dB	78.6	52.8	30.2
5dB	31.8	19.4	15.4
10dB	7.9	6.9	6.4
15 dB	2.7	2.9	3.0
20 dB	2.3	2.0	1.7
clean	n/a	1.4	1.4

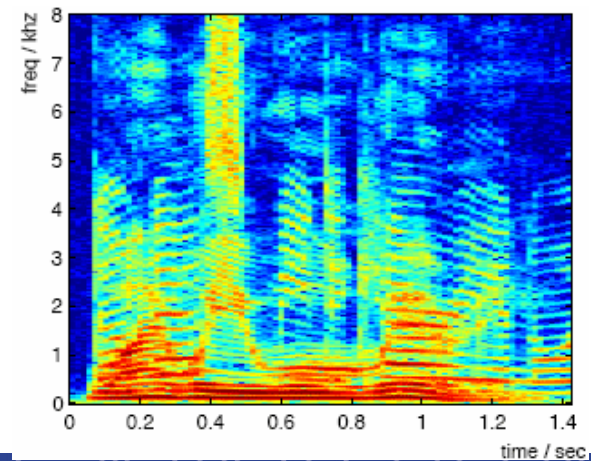
Recognition Word Error Rate in %

Experiments: Speech Enhancement (ongoing)

- **baseline: beam-forming from two microphone**
 - **The proposed:**
 - **delay estimation**
 - **Process both channels with phase-error filtering**
 - **beam-forming**
 - **Expect better SNR improvement.**
- 
- A stylized illustration of a person with a beard, wearing a red shirt, sitting in a light blue armchair at a desk. They are using a laptop. The scene is set in a room with a window showing a view of trees and a framed picture on the wall. The entire illustration is enclosed within a circular frame with a yellow and orange gradient.

Experiments: Speech Separation (ongoing)

- **Dual-microphone:**
 - Model-based adaptive phase-error filtering works.
 - Models are training by target speaker's speech
- **Single-microphone:**
 - Models are training by target speaker's speech
 - Time-Frequency masking
 - Decide how to suppress each T-F block to maximize the likelihood function of the model.



Summary

- **Introduce two speech processing techniques based on speech models:**
 - **single-microphone noise removal.**
 - **dual-microphone phase-error filtering.**
- **Both techniques can be applied to**
 - **speech recognition.**
 - **speech enhancement.**
 - **speech separation.**
- **Model-based speech processing shows promising results.**


```
ERROR: undefined
OFFENDING COMMAND:

STACK:
```