T-61.184
Automatic Speech Recognition: From Theory to Practice

http://www.cis.hut.fi/Opinnot/T-61.184/
November 15, 2004

Prof. Bryan Pellom
Department of Computer Science
Center for Spoken Language Research
University of Colorado

pellom@cslr.colorado.edu

Announcements

- I still need 2 more volunteers to present their project topic (next week) on November 22nd
- The goal is to present to the class (and myself) your chosen topic area.
- Brief 10 minute presentation (project overview)
- Does not have to reflect your completed project (since that is due December 8th).

Today

- **How to reduce mismatch between training and test conditions to improve recognition reliability**
- **Noise Robustness Techniques**
- **Speaker Adaptation Techniques**

Classic References

- Boll, S. F. (1979) “Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, pp. 113-120, April.
- Ephraim, Y. and Malah, D. (1984) "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. ASSP, Vol. ASSP-32, No. 6, pp. 1109-1121.
- Hansen, J. H.L. and Clements, M. A. (1991) "Constrained Iterative Speech Enhancement With Application to Speech Recognition", IEEE Transactions on Signal Processing, Vol. 39, No. 4, pp. 785-805, April.
- Gong, Y. (1995) "Speech recognition in noisy environments: A survey", Speech Comm., Vol. 16, pp. 261-291.
- Gales, M. J. F. (1997) “Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition,” Technical Report CUED/F-INFENG/TR 291, Cambridge University, May.
- Huang, X., Acero, A., Hon, H.-W., (2001) *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development* , Prentice Hall, ISBN 0-13-022616-5

Training vs. Test Mismatch

- **Performance of speech recognition systems degrades whenever there is a mismatch between <u>training</u> and <u>test</u> conditions**

- **Mismatch can occur at various levels of processing and to various degrees,**
  - Acoustic Modeling
  - Language Modeling
  - Pronunciation Modeling

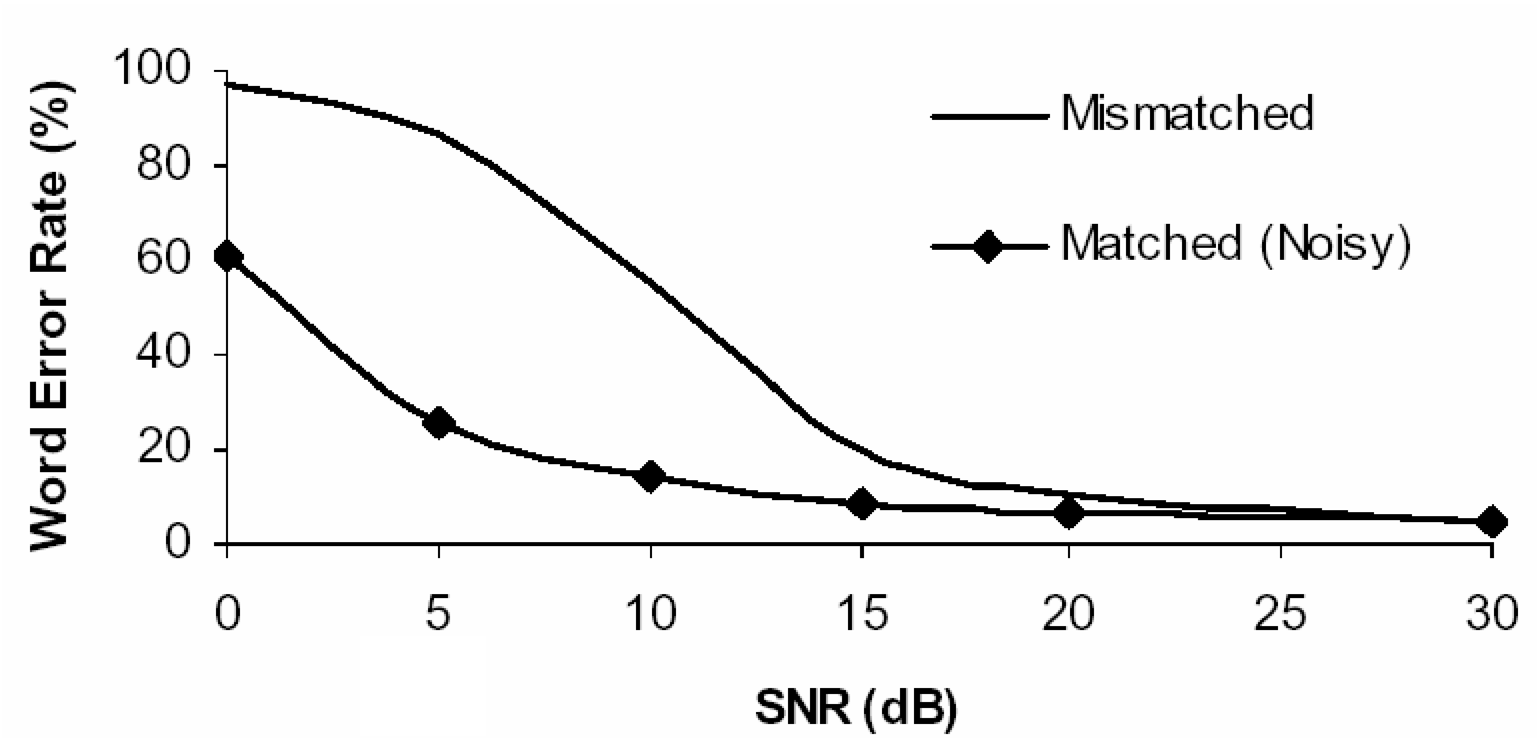
Acoustic Variability

- **Environmental**
  - Transducer — *Microphone frequency response*
  - Channel & Codec — *Telephone band-limiting, VoIP*
  - Noise — *Additive, Impulsive, etc.*
- **Speaker**
  - Accent — *Foreign Accent*
  - Dialect — *Regional Differences*
  - Vocal Tract Geometry
  - Lombard Effect — *Voice changes due to noise*
  - Age & Gender

Language Variability

- **Speaker-Specific**
  - Choice of words is highly speaker-specific
  - Vocabulary varies speaker-to-speaker
- **Task-Specific**
  - Topic shifts impact word choices, vocabulary, and statistical distributions of words
  - We can not expect a language model trained on financial news to work well when transcribing sports news.

Impact of Additive Noise on Word Error Rate



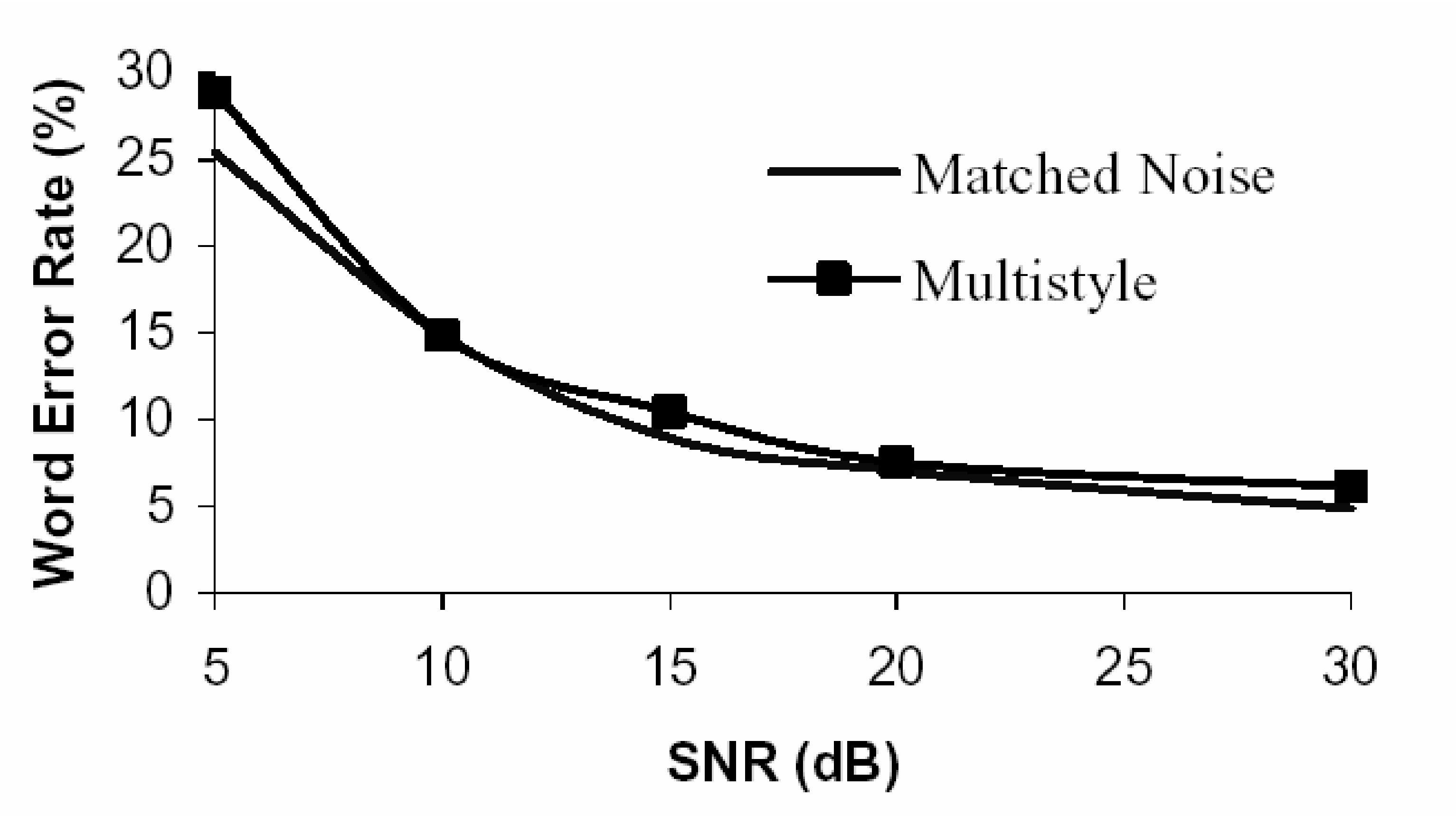
Train on noise-free data, test on noisy data…

Multi-Style Training

- Train on acoustic data that has been corrupted by (1) different noise types, and (2) different noise levels.
- Works well when all types of operating environments are known.
- Can't always predict the environment; Need methods to compensate for noise + channel

T-61.184

Multi-Style Training

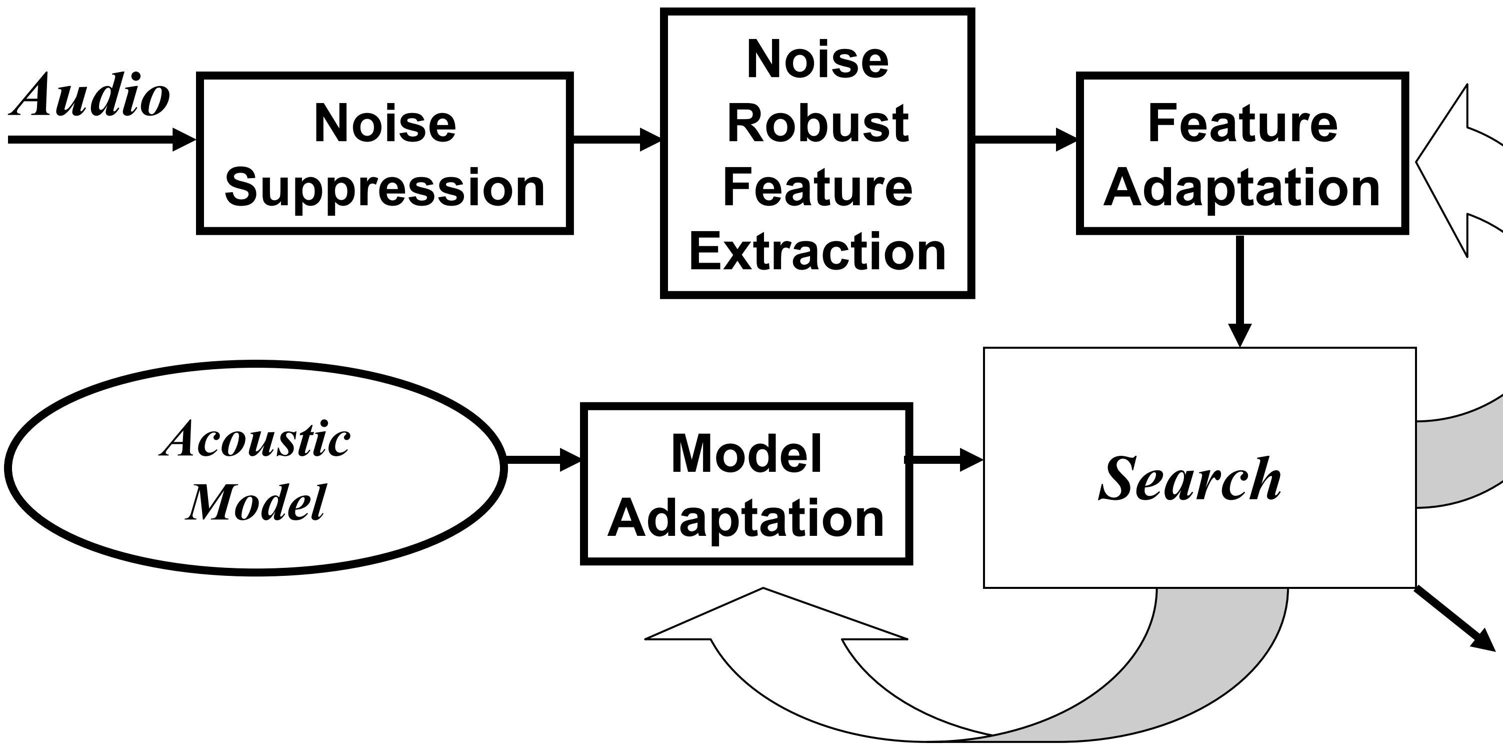


Train on data corrupted by noise of various levels.

Robust Acoustic Modeling

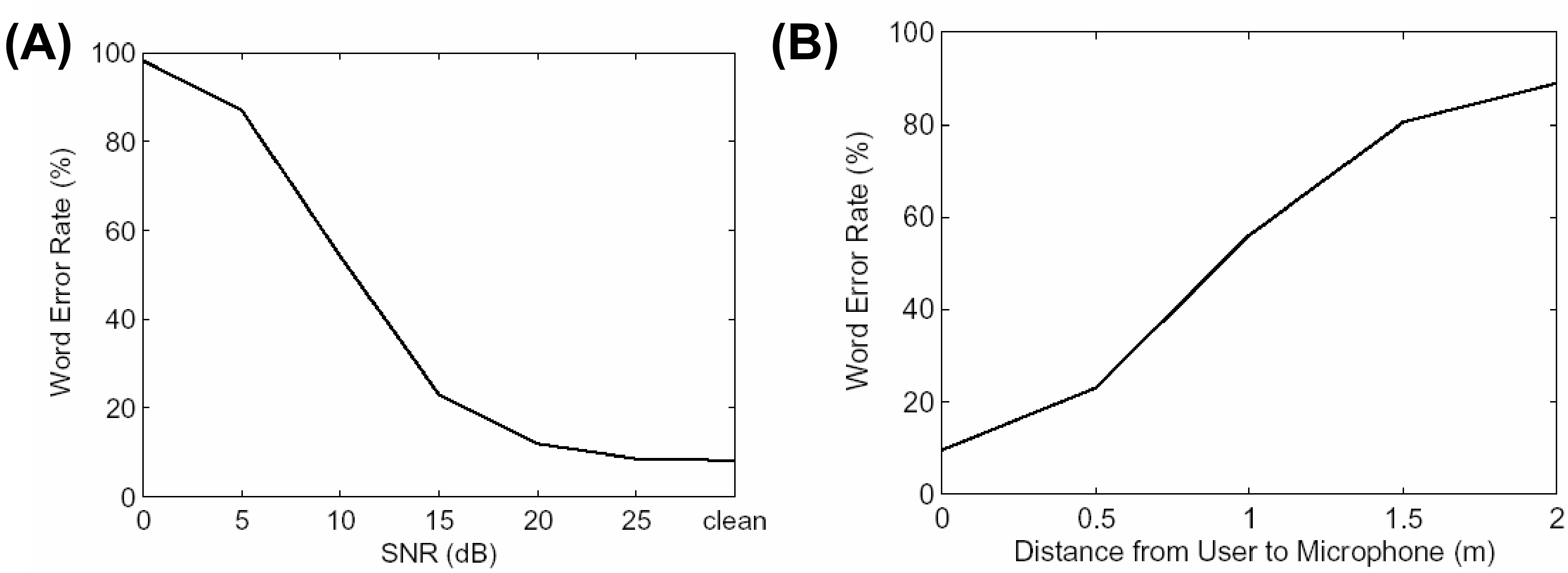
- **Robust Front-end Processing**
  - Remove noise from speech
  - Design features to be as noise, channel robust as possible
- **Feature Compensation**
  - Reduce the observed mismatch between the extracted features and estimated model parameters
- **Acoustic Model Compensation**
  - Modify acoustic model parameters to closer match the observed test environment

Robust Acoustic Modeling



Front-End Noise Suppression

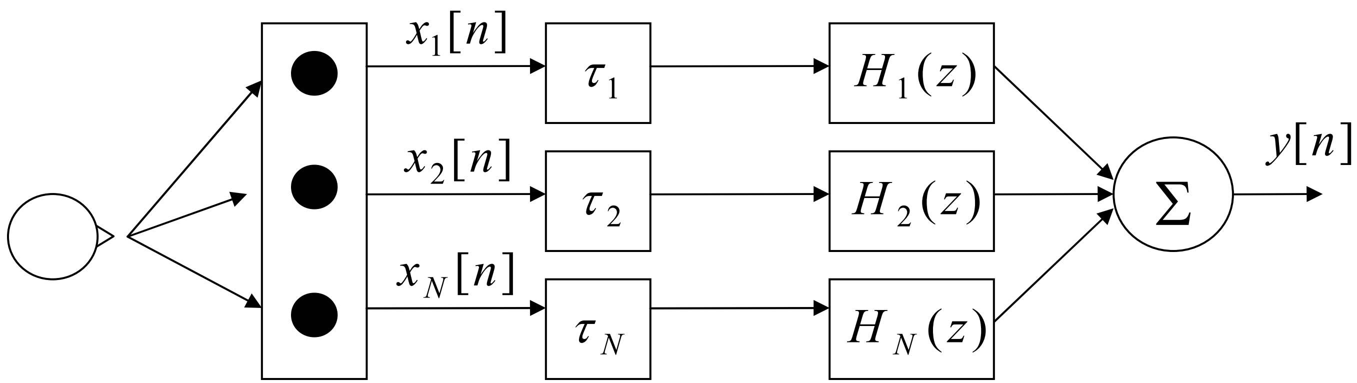
Scope of the Problem



- Error Rate for WSJ Dictation system trained on clean speech (Paul & Baker, 1992)
- (A) Word Error Rate as a function of SNR for White Noise
- (B) Word Error Rate as a function of Microphone Distance

Microphone Arrays

- **Microphone arrays provide spatial selectivity**
- **Spatial selectivity varies as a function of frequency**
- **Example: Filter-and-sum beam former,**



$$y[n] = \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} h_m[p] x_m[n - p - \tau_m]$$

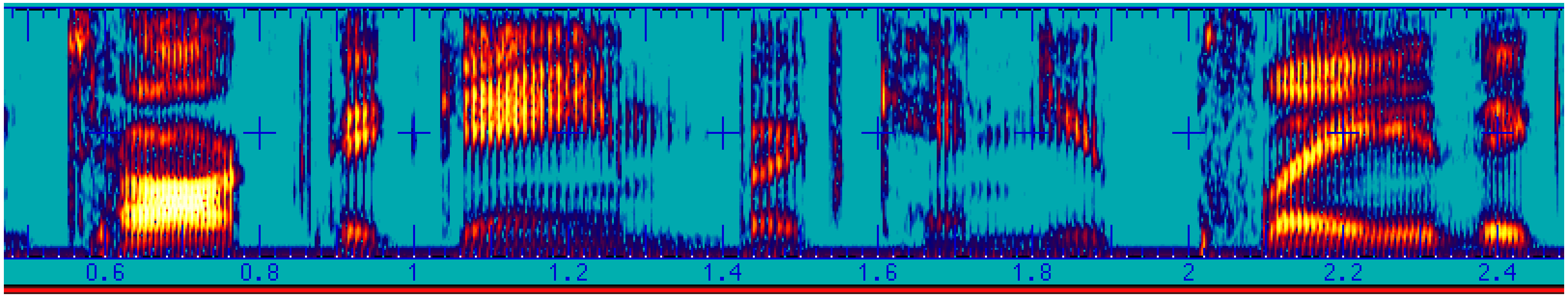
Noise Suppression

- Goal is to reduce the impact of noise on speech
- Typical approaches estimate the clean-speech magnitude spectrum from the noisy-speech magnitude spectrum
- Phase of the noisy signal used as the estimate of the phase of the clean-speech signal

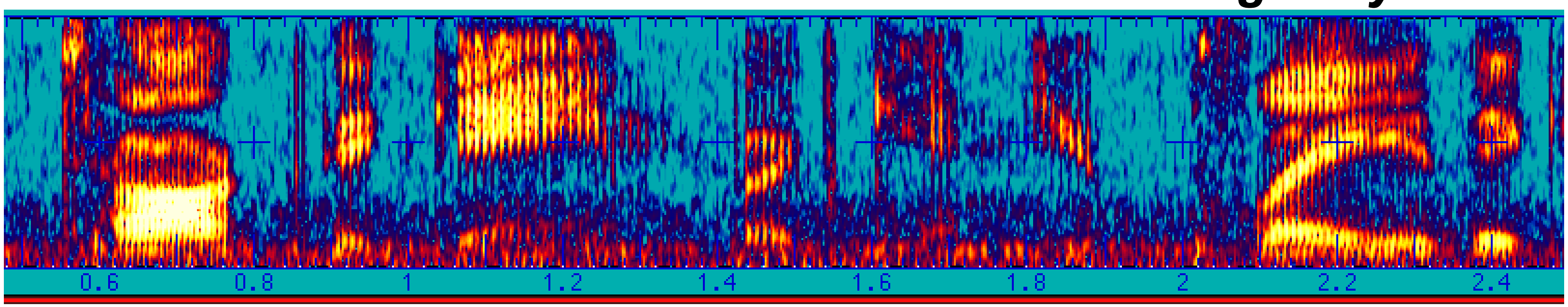
Stationary Additive Noise

- **Relatively constant spectral shape across time**
- **Has a non-uniform impact on speech,**
  - Speech sounds have different spectral shapes across time
  - Speech sounds have different energy levels across time
- **Signal-to-noise ratio therefore is a function of time even when the noise is additive and stationary.**

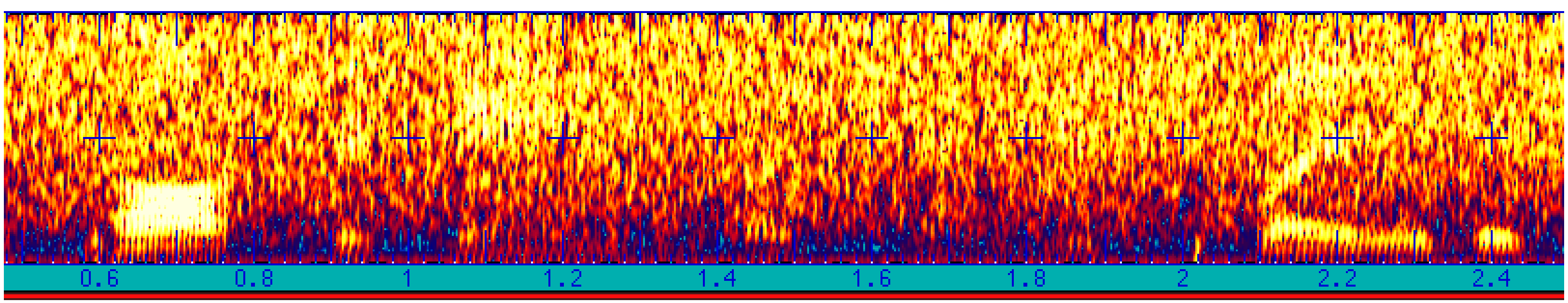
Original



5dB SNR Car Highway Noise



5dB SNR White Noise



Additive Noise Model

$$y_t(n) = s_t(n) + d_t(n)$$

$$Y_t(\omega) = F\{s_t(n) + d_t(n)\} = S_t(\omega) + D_t(\omega)$$

$$|Y_t(\omega)|^2 = |S_t(\omega) + D_t(\omega)|^2$$

$$= |S_t(\omega)|^2 + |D_t(\omega)|^2 + 2\,\mathrm{Re}\left\{ S_t(\omega) D_t^*(\omega) \right\}$$

$$\boxed{|Y_t(\omega)|^2 \approx |S_t(\omega)|^2 + |D_t(\omega)|^2}$$

Spectral Subtraction

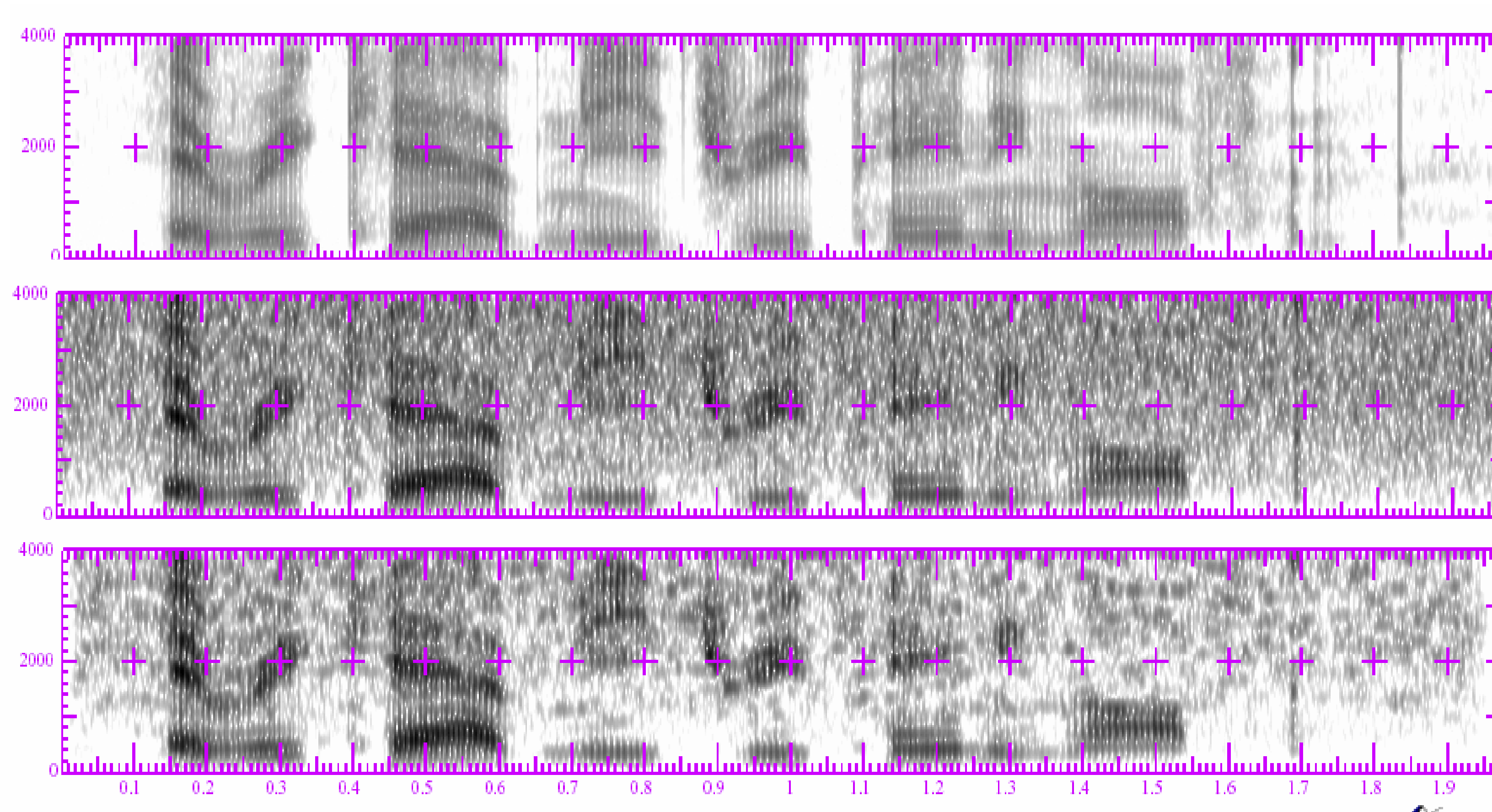
- **Subtract estimate of noise power spectrum from observed (speech+noise) power spectrum**

$$\left|\hat{S}_t(\omega)\right|^2 = \left|Y_t(\omega)\right|^2 - \mathrm{E}\left\{\left|D_t(\omega)\right|^2\right\}$$

- **Must ensure that spectral estimate is positively valued.**
- **Compute features from estimated clean speech power spectrum**

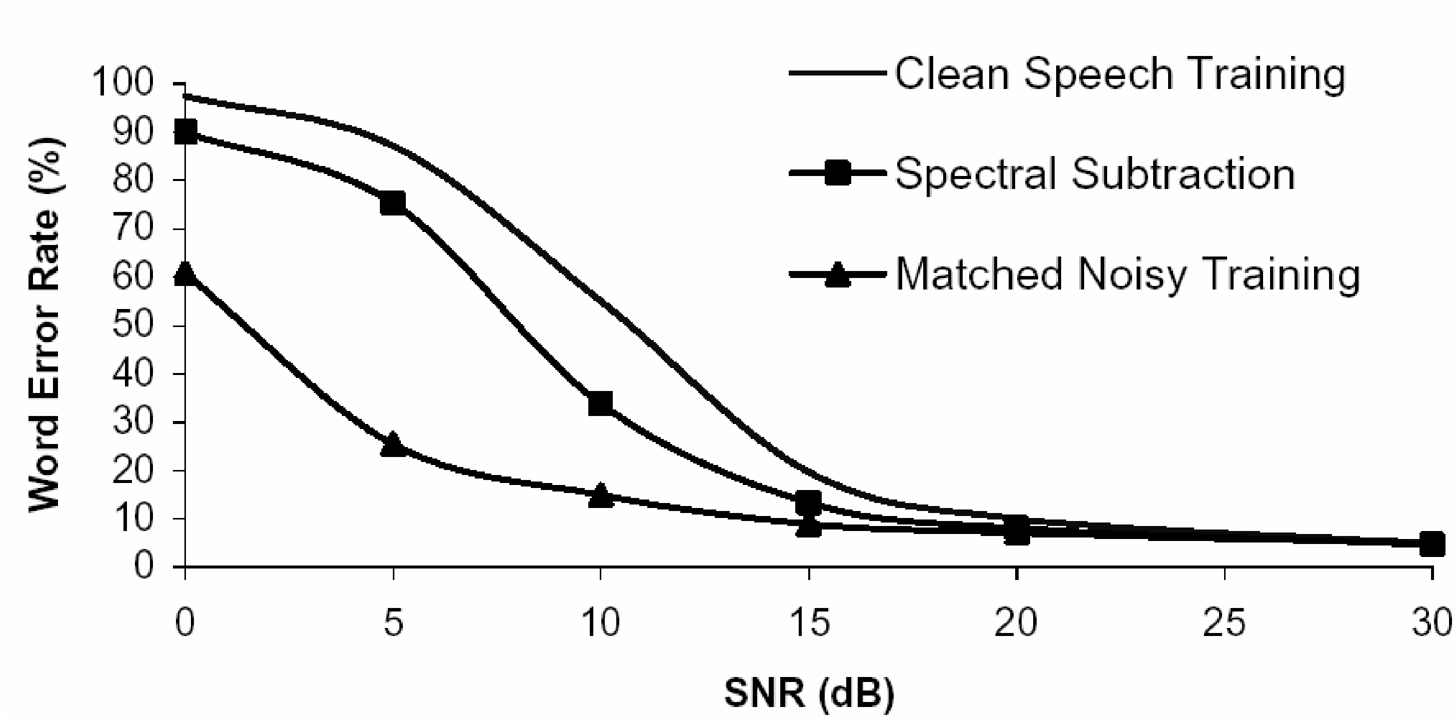
Spectral Subtraction Example

(a)
(b)
(c)



(a) Original clean, (b) degraded 10dB WGN, (c) Enhanced

Improvements from Spectral Subtraction



***** Many times, these gains are not as significant in real-world operating environments. Sometimes spectral subtraction can degrade ASR performance!***

T-61.184

Wiener Filtering

- **Estimate optimal Weiner filter and apply to the noisy speech spectral magnitudes,**

$$\hat{S}_t(\omega_k) = Y_t(\omega_k) H_t(\omega_k)$$

$$H_t(\omega_k) = \frac{|S_t(\omega_k)|^2}{|S_t(\omega_k)|^2 + |D_t(\omega_k)|^2}$$

- **Numerator for filter (H) is unknown, must be estimated. Sometimes this is done iteratively using LPC-based models for speech as a constraint in the estimation process.**

- ***J. H. L. Hansen, M. A. Clements, "Constrained Iterative Speech Enhancement With Application to Speech Recognition", IEEE Transactions on Signal Processing, Vol. 39, No. 4, pp. 785-805, April, 1991.***

Minimum Mean-Square Error (MMSE) Spectral Amplitude Estimation

- **Estimate the clean speech spectral amplitude from corrupted speech using MMSE methods**

$$\hat{S}_t(\omega_k) = Y_t(\omega_k) H_t(\omega_k)$$

- ***Y. Ephraim and D. Malah, "Speech Enhancement Using Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109-1121, 1984***

- **Derived MMSE Estimator:**

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k \qquad \begin{Bmatrix} \xi_k & \text{: a priori SNR} \\ \gamma_k & \text{: a posteriori SNR} \end{Bmatrix}$$

$$H_t(\omega_k) = \left(\frac{\sqrt{\pi}}{2}\right)\left(\frac{\sqrt{\upsilon_k}}{\gamma_k}\right)\exp\left(-\frac{\upsilon_k}{2}\right)\cdot\left[(1+\upsilon_k)I_o\left(\frac{\upsilon_k}{2}\right)+\upsilon_k I_1\left(\frac{\nu_k}{2}\right)\right]$$

T-61.184

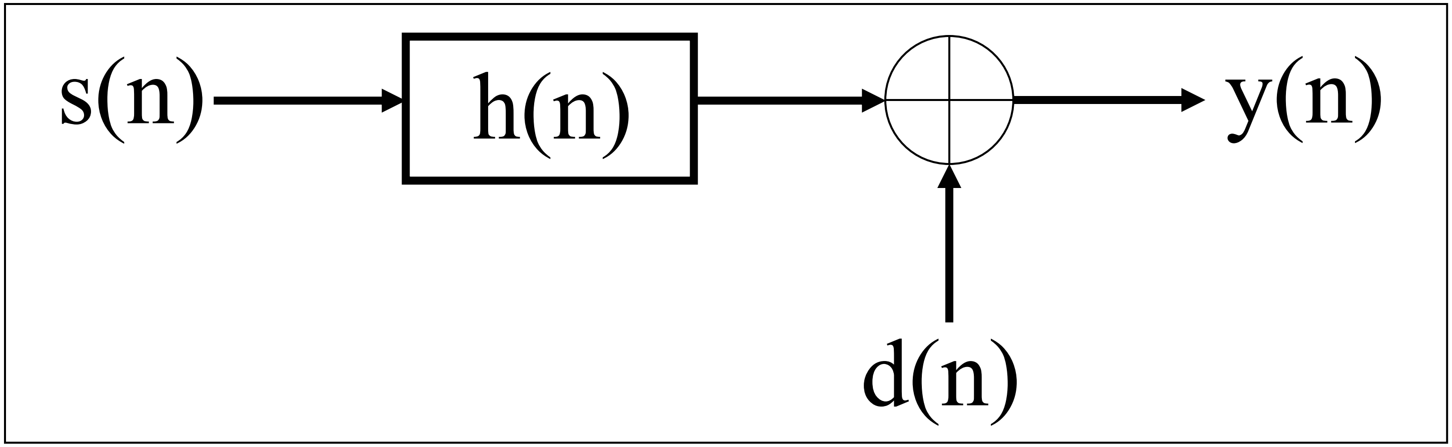
Several Issues with Speech Enhancement Approaches

- **Requires estimate of noise to be updated during periods of silence (noisy-only regions). How to do this well?**
- **Most algorithms subtract variable amounts of the noise estimate to obtain trade-offs in distortion vs. noise attenuation. How does this impact the ASR engine?**
- **Algorithms developed to improve intelligibility *may not necessarily improve ASR accuracy.* Sometimes worse!**
- **Many speech enhancement algorithms have not been formulated to improve ASR accuracy. Be careful!**

Considering the Communication Channel

- **Communication channels generally act to filter the input signal (multiplicative distortion in the frequency domain)**
- **Analog telephone networks band-limit the signal to a range of approximately 200Hz to 3400Hz.**
- **Spectral shape of channel can vary from call-to-call, sometimes with <u>echo</u>.**
- **Other types of channels,**
  - Voice over IP
  - Variability due to Microphones
  - Telephone handset variability (also wireless phones)
  - Cellular telephony

Noisy-Channel Model



$$y(n) = s(n) * h(n) + d(n)$$

s(n): clean speech signal

h(n): channel (telephone, microphone)

d(n): additive noise

Spectral Domain Noisy-Channel Model

$$y_t(n) = s_t(n) * h_t(n) + d_t(n)$$

$$\downarrow$$

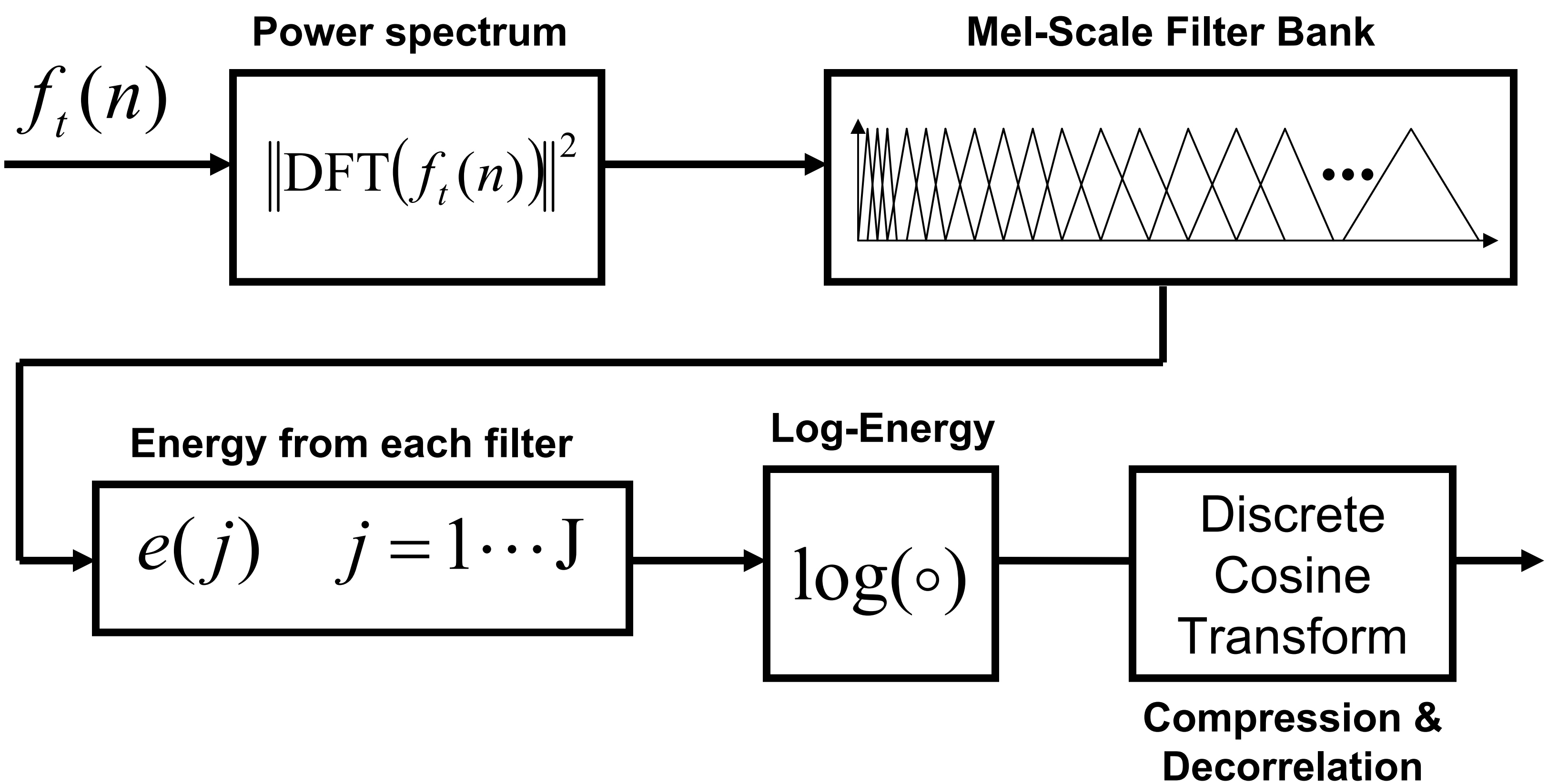
$$|Y_t(\omega)|^2 = |S_t(\omega)|^2 |H_t(\omega)|^2 + |D_t(\omega)|^2 + 2\,\mathrm{Re}\left\{ S_t(\omega) H_t(\omega) D_t^*(\omega) \right\}$$

Assuming the speech and noise are statistically independent,

$$\boxed{|Y_t(\omega)|^2 \approx |S_t(\omega)|^2 |H_t(\omega)|^2 + |D_t(\omega)|^2}$$

Robust Feature Extraction

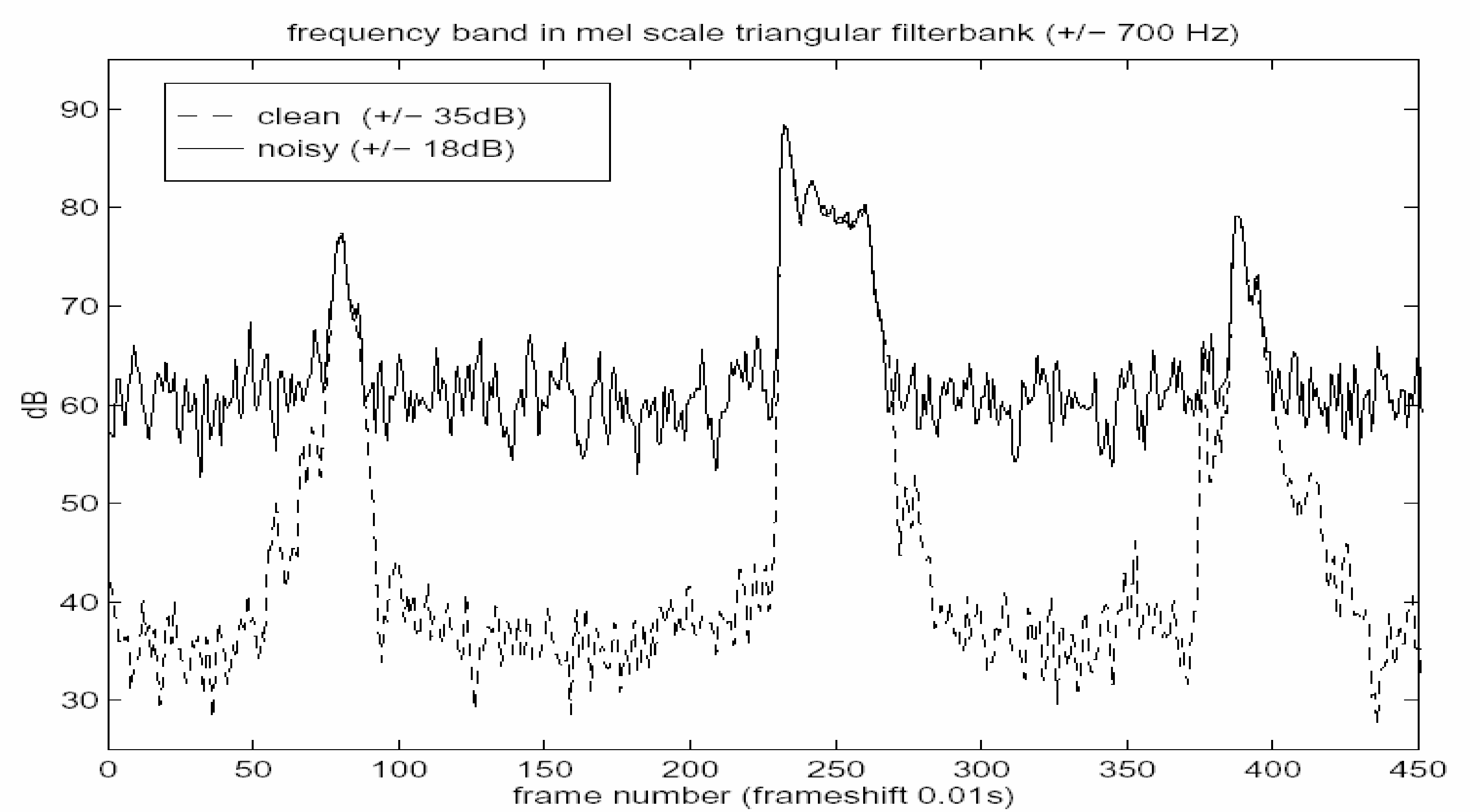
MFCC Block Diagram



Robust Feature Extraction

- **Attempts to compensate for channel and noise during feature extraction process**
- **Recall that the typical cepstral parameter representation is based on the log-scaled outputs of a series of non-linear spaced filters…**
- **Additive Noise will impact the distributions of the filterbank values, but how?**

Impact of Noise on Mel-Scale Filter banks



Classic Feature Normalization Methods

- **Cepstral Mean Normalization (CMN)**
  - ❑ Mainly compensates for channel (and some noise)
- **Cepstral Variance Normalization**
  - ❑ Mainly compensates for noise
- **Vocal Tract Length Normalization (VTLN)**
  - ❑ Mainly compensates for speaker-differences

Impact of Noise on (Simulated) Cepstral Parameters

- **Assume noise and clean speech are Gaussianly distributed in log-spectral domain,**

$$y_t = \log\left(\exp\left(s_t\right) + \exp(d_t)\right)$$

- **Assume speech with mean=10, variance=5.**
- **Simulate the adding of noise at various levels**

Impact of Noise on (Simulated) Cepstral Parameters

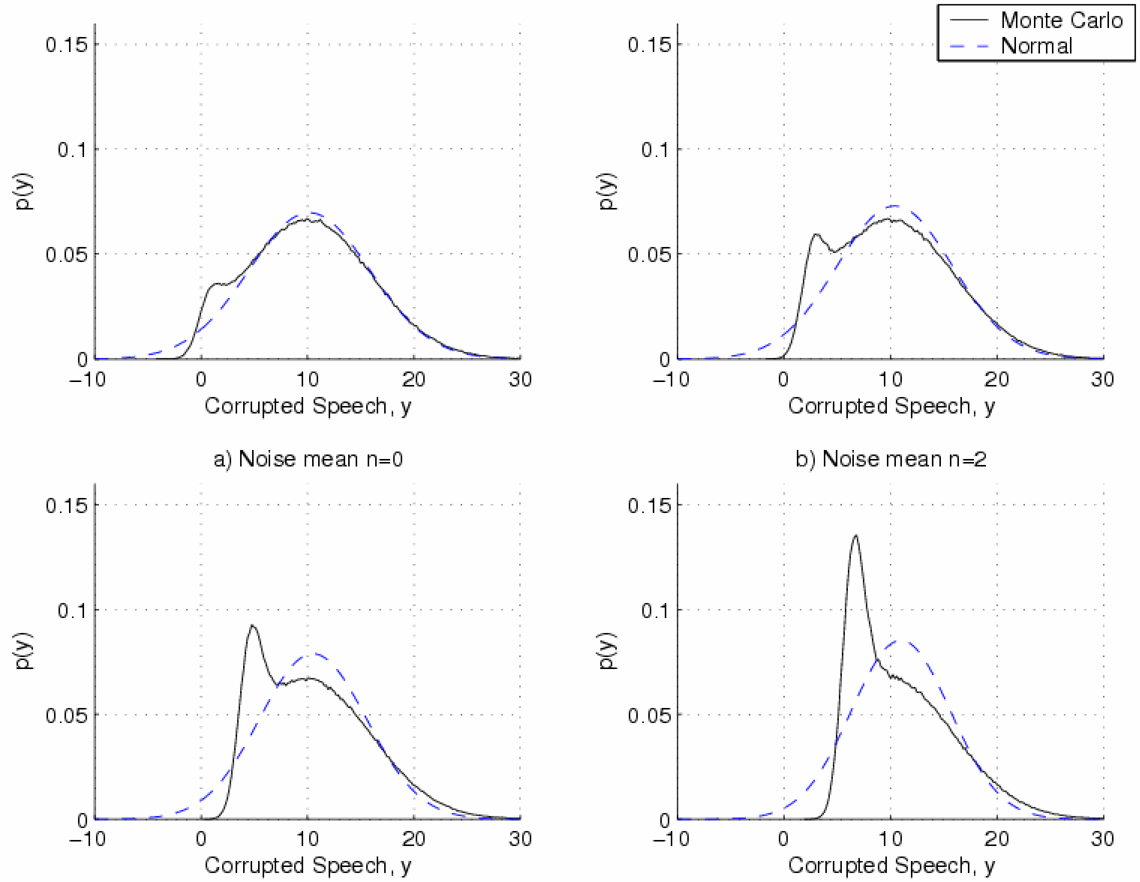


Image from Liao and Gales (2004)

- **Resulting distribution initially becomes bimodal, then skewed unimodal**
- **Mean shifts**
- **Variance decreases with increasing noise level**

Noisy-Channel Model in Log-Spectral Domain

- **Our channel model (k refers to filter bank bin),**

$$\left|Y_t(\omega_k)\right|^2 = \left|S_t(\omega_k)\right|^2 \left|H_t(\omega_k)\right|^2 + \left|D_t(\omega_k)\right|^2$$

- **In the log-domain,**

$$\log\left(\left|Y_t(\omega_k)\right|^2\right) = \log\left(\left|S_t(\omega_k)\right|^2\right) + \log\left(\left|H_t(\omega_k)\right|^2\right) + \\ \ldots\log\left(1+\exp\left(\log\left(\left|D_t(\omega_k)\right|^2\right)\right) - \log\left(\left|S_t(\omega_k)\right|^2\right) - \log\left(\left|H_t(\omega_k)\right|^2\right)\right)$$

T-61.184

Noisy-Channel Model in Log-Spectral Domain

- **Define,** $$g(a) = C \cdot \log\left(1 + \exp\left(C^{-1} a\right)\right)$$
- **Where C and C^{-1} are the discrete cosine transform (DCT) and inverse DCT.**
- **The cepstral parameters can then be described by the following non-linear model,**

$$y^{(c)} = s^{(c)} + h^{(c)} + g\left(d^{(c)} - s^{(c)} - h^{(c)}\right)$$

Cepstral Mean Normalization

- **Channel distortion (h) in linear spectral domain result in additive distortions in the log-spectral (cepstral) domain**

$$y^{(c)} = s^{(c)} + h^{(c)} + g\left(s^{(c)} - h^{(c)}\right)$$

- **Telephone channel differences compensated by subtracting the long-term mean from the cepstral features**

$$\hat{s}_t^{(c)} = y_t^{(c)} - \bar{y}_t^{(c)} = y_t^{(c)} - \frac{1}{T}\sum_{n=1}^{T} y_n^{(c)}$$

Variations on CMN

- **CMN-2**
  - Compute a running mean for the speech and silence separately
  - Detect speech vs. silence and use the appropriate mean
- **Real-Time Implementations**
  - Running average (typically 5 seconds):

$$\bar{y}_t^{(c)} = \alpha \cdot y_t^{(c)} + (1-\alpha) \cdot \bar{y}_{t-1}^{(c)}$$

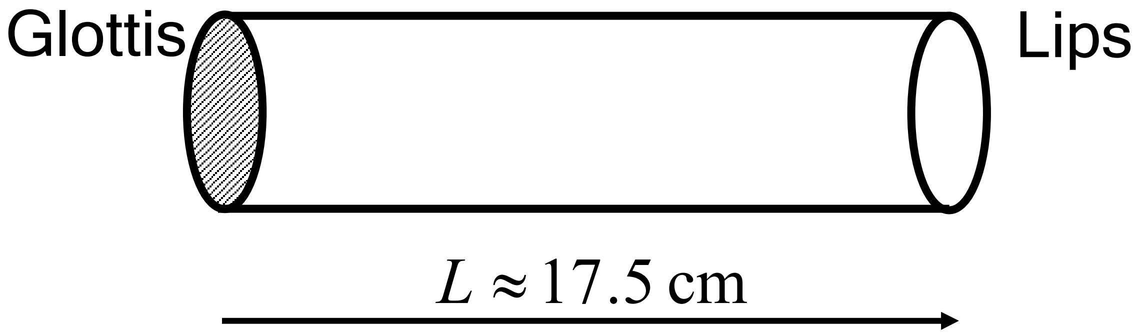
Cepstral Variance Normalization

- **Additive noise reduces the variance of cepstral features**
- **Compensate by normalizing all feature components to have variance of 1.0**
- **Typically, compute standard deviation of features over large block of adaptation data. Divide features by their standard deviation**

Vocal Tract Length Normalization

- **Vocal tract lengths vary from speaker to speaker**
  - Influence formant frequency locations
  - Source of intra-speaker variability
- **VTLN methods generally warp the speech spectrum as to remove variabilities due to vocal tract length**
- ***Welling, L., Ney, H., and Kantahak, S. (2002) "Speaker Adaptive Modeling by Vocal Tract Normalization," IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 6, pp. 415—426, September.***

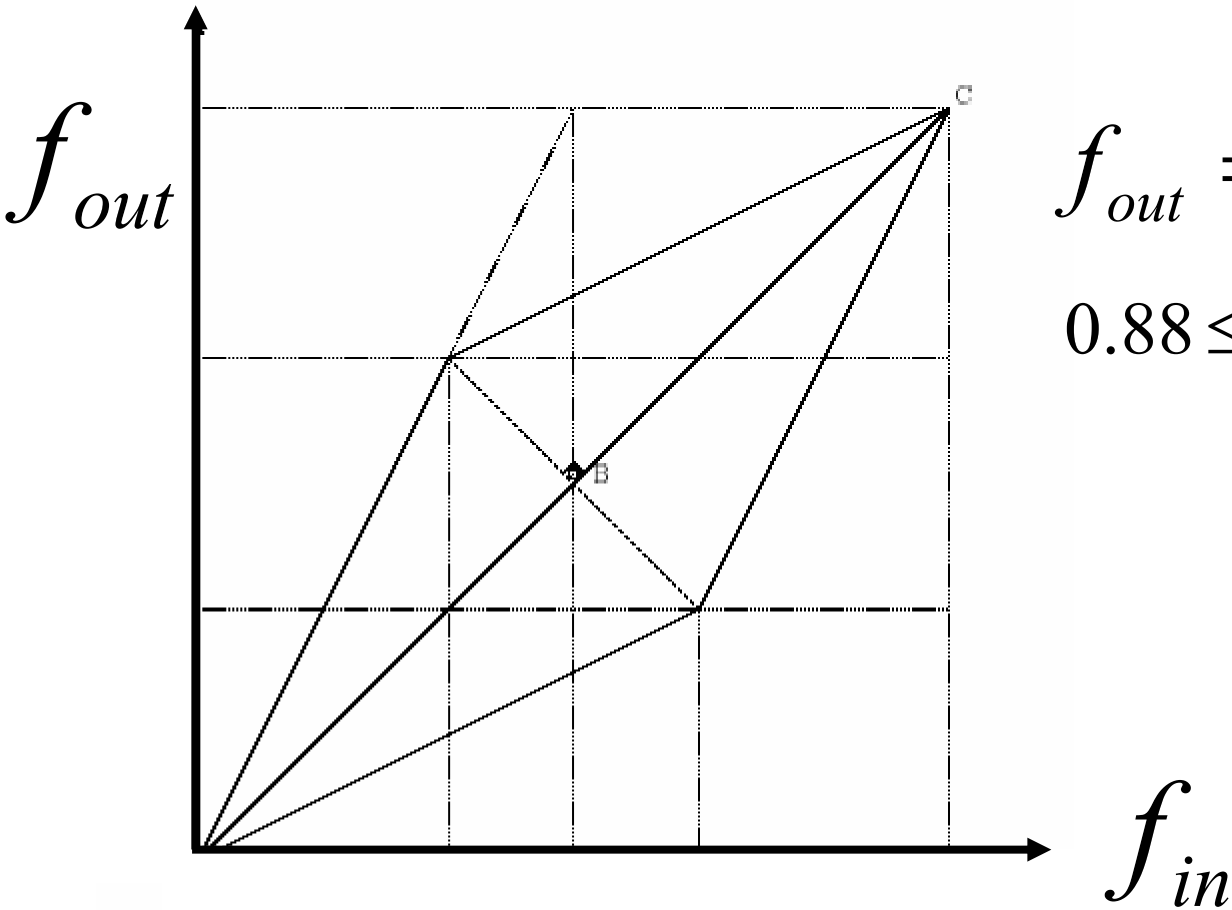
Uniform Lossless Tube Model of Vocal Tract



- **Formant locations linearly related to tube length,**

$$F_i = \frac{c}{4L}[2i-1] \qquad c = 350\ \mathrm{m/sec} \qquad i = 1,2,3,4,\ldots$$

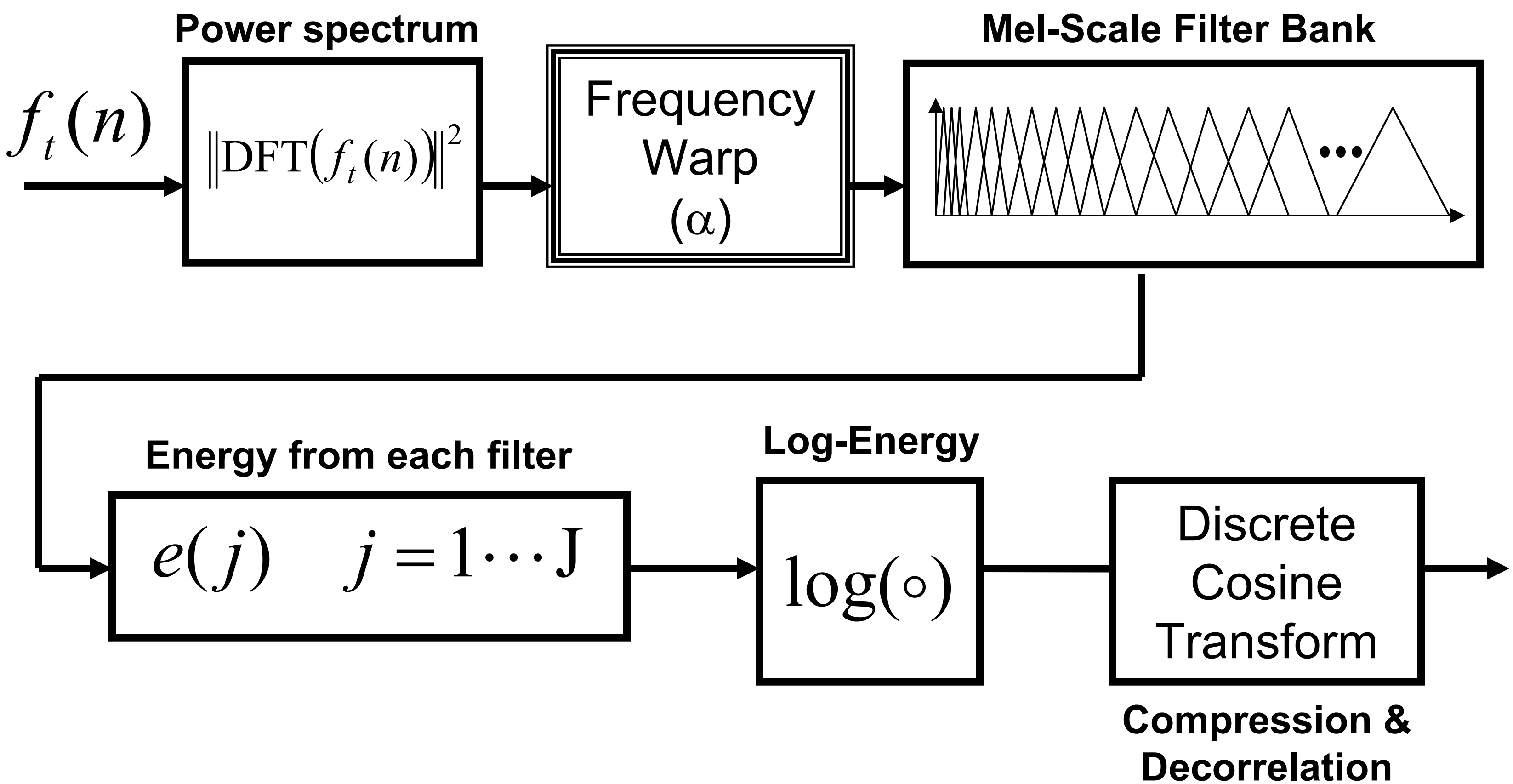
VTLN via Linear Frequency Warping



$$f_{out} = \alpha \cdot f_{in}$$

$$0.88 \leq \alpha \leq 1.12$$

MFCC with VTLN Block Diagram



T-61.184

VTLN Implementation

- **Must determine frequency warp factor (alpha) for each speaker in training set**
- **Apply warping during feature extraction, train acoustic models**
- **During recognition, must determine optimal frequency warp factor for each test speaker**

VTLN During HMM Training

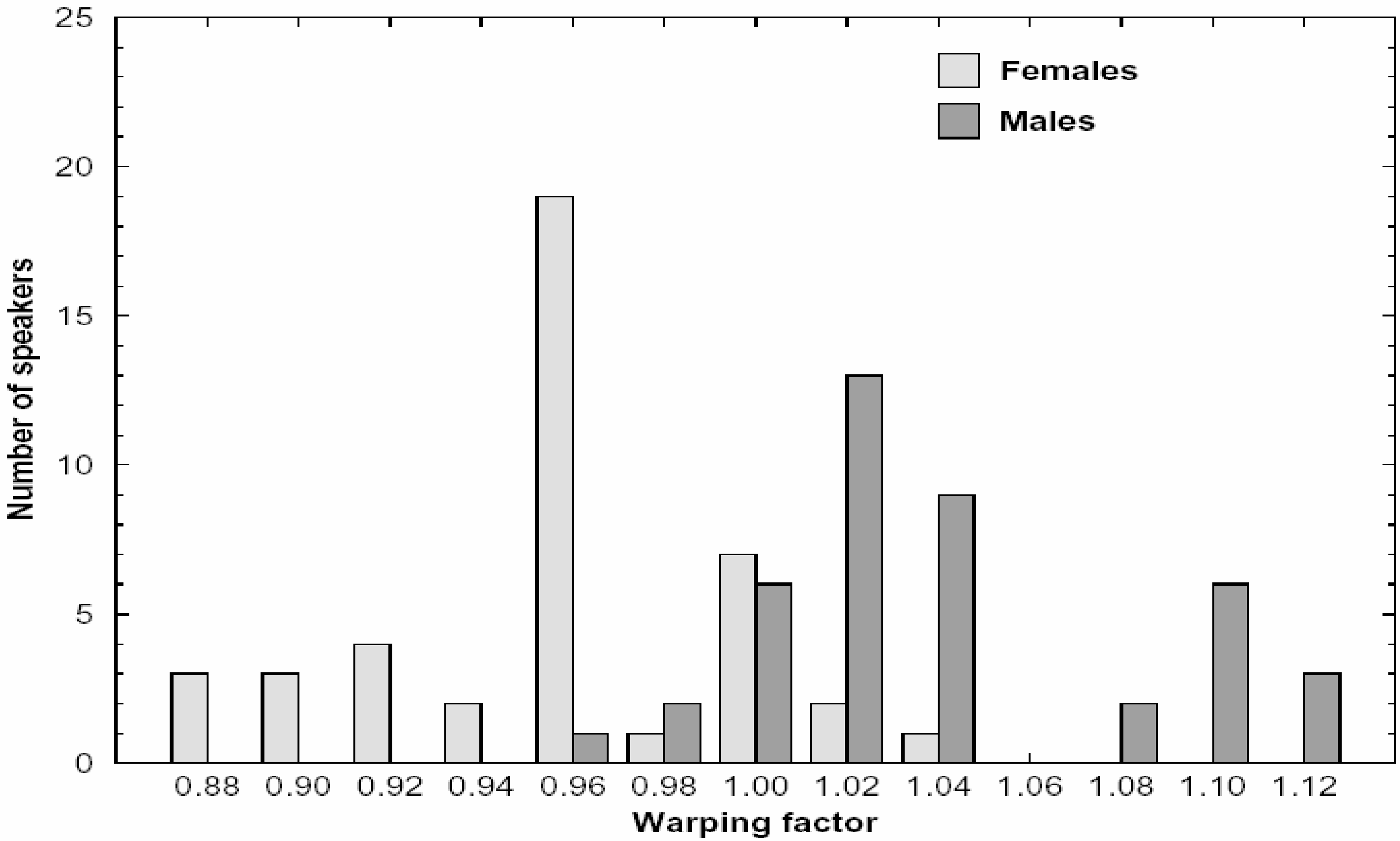
- **Construct an initial model λ using a single Gaussian mixture component for each clustered triphone state**

- **For each training speaker, select a frequency warping factor that maximizes the likelihood of the training data given the reference transcription**

$$\alpha_s = \arg\max_{\alpha} P(\mathbf{O}^{(\alpha)} \mid \mathrm{W}, \lambda)$$

- **Estimate normalized acoustic models by extracting features using speaker-dependent warp factors. Retrain HMMs using standard decision tree method**

Speaker-Dependent Frequency Warp Factors Estimated During Training



VTLN During Recognition

- **Perform a first pass recognition using standard acoustic models (trained with alpha=1.0). This provides an initial estimate of the word strings.**
- **Select a frequency warping factor that maximizes the likelihood of the speaker's frequency warped features given the hypothesized transcription**

$$\alpha_s = \arg\max_{\alpha} P(\mathbf{O}^{(\alpha)} \mid \hat{\mathrm{W}}, \lambda)$$

- **Perform a second recognition pass using VTLN normalized acoustic models with features extracted using the speaker-dependent frequency warp factor**

Linear Discriminant Analysis (LDA)

- **Improve Discrimination via a linear transform on features**
- **Maps a set of input feature vectors (x) of dimension D to a set of output feature vectors (y) of dimension D by means of a linear transformation θ,**

$$y_t^{(c)} = \theta^T x_t^{(c)}$$

- **θ chosen to maximize the ratio of between-class scatter to within-class scatter given a set of class-labels on the training set (x).**

T-61.184

Estimation of LDA Transform

- **Given a set of vectors assigned to a set of C classes, define <u>Scatter-Matrix</u> as,**

$$S_i = \sum_{x \in c(i)} \left(x_t^{(c)} - m_i\right)\left(x_t^{(c)} - m_i\right)^T$$

Where m_i is the mean of the *i*th class.

- **Define <u>Within-class Scatter,</u>**

$$S_W = \sum_{i=1}^{c} S_i$$

T-61.184

Estimation of LDA Transform

- Define <u>Between-Class Scatter,</u>

$$S_B = \sum_{i=1}^{c} (m_i - m)(m_i - m)^T$$

- Can be shown that the solution for θ is the eigenvectors of the matrix,

$$S_W^{-1} S_B$$

LDA as used in Speech Recognition

- Viterbi Align training data using non-LDA system (associate feature vectors to phone or subphone units)
- For 50 phones with 3 states per phone, we might have 50x3 = 150 classes
- Take each training vector (just static 13-D cepstrum) and augment it with between 4-8 surrounding vectors.
- Estimate an LDA transform for this extended vector representation.
- Keep the top N dimensions (based on eigen values) for recognition (40-60).
- Retrain the acoustic models with this new feature representation.

Acoustic Model Adaptation

Acoustic Model Adaptation

- **For HMM states modeled with Gaussian distributions, Model Adaptation Methods attempt to shift the means and variances of Gaussians to better match the input feature distributions**

- **Example Techniques,**
  - ❑ Parallel Model Combination (PMC)
  - ❑ Maximum Likelihood Linear Regression (MLLR)
  - ❑ Maximum A Posteriori (MAP) Adaptation

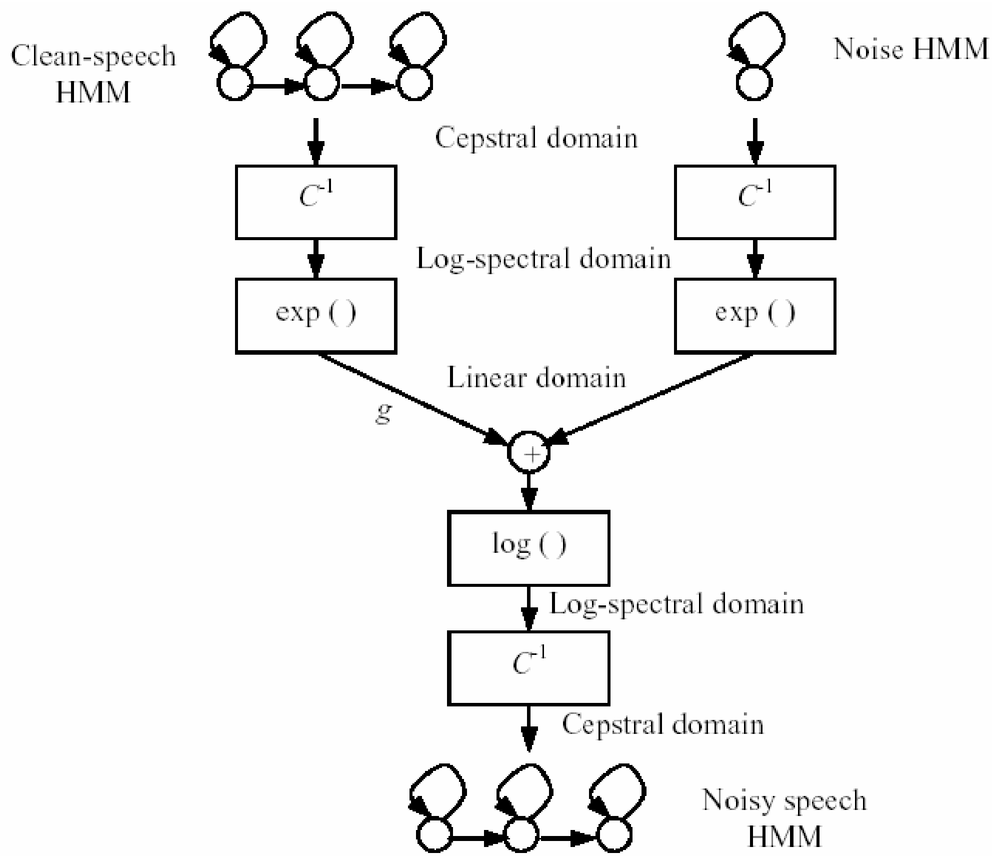
Parallel Model Combination (PMC)

- **Acoustic model compensation for additive and convolutional noise sources**
- **Assumes**
  - HMMs trained on clean speech
  - Typically assumes noise is stationary and modeled using 1 HMM state
- **A clever approach that modifies the (static) cepstral means of modeled Gaussians**

Parallel Model Combination (PMC)

- **Basic Idea,**
  - ❑ Convert each HMM cepstral mean vectors from log-spectral domain to linear spectral domain.
  - ❑ Add noise spectrum estimate to clean speech spectrum from the acoustic model
  - ❑ Convert new noisy-speech spectrum back to cepstral domain to get compensated HMM model mean vectors

Parallel Model Combination (PMC)



- **M.J.F. Gales and S.J. Young (1995). "Robust Speech Recognition in Additive and Convolutional Noise using Parallel Model Combination."** ***Computer Speech and Language*** **Volume 9.**

Maximum Likelihood Linear Regression (MLLR)

- **Model adaptation technique which estimates new values for the Gaussian mean vectors via a linear transform**

- **The linear transform is estimated from labeled training data (features and their state alignments)**

- ***M.J.F. Gales and P.C. Woodland (1996), “Mean and Variance Adaptation within the MLLR Framework,” Computer Speech and Language Volume 10.***

Maximum Likelihood Linear Regression (MLLR)

- Estimates a linear transform matrix (A) and bias vector (b) to transform HMM model means:

$$\mu_{new} = A_r \mu_{old} + b_r$$

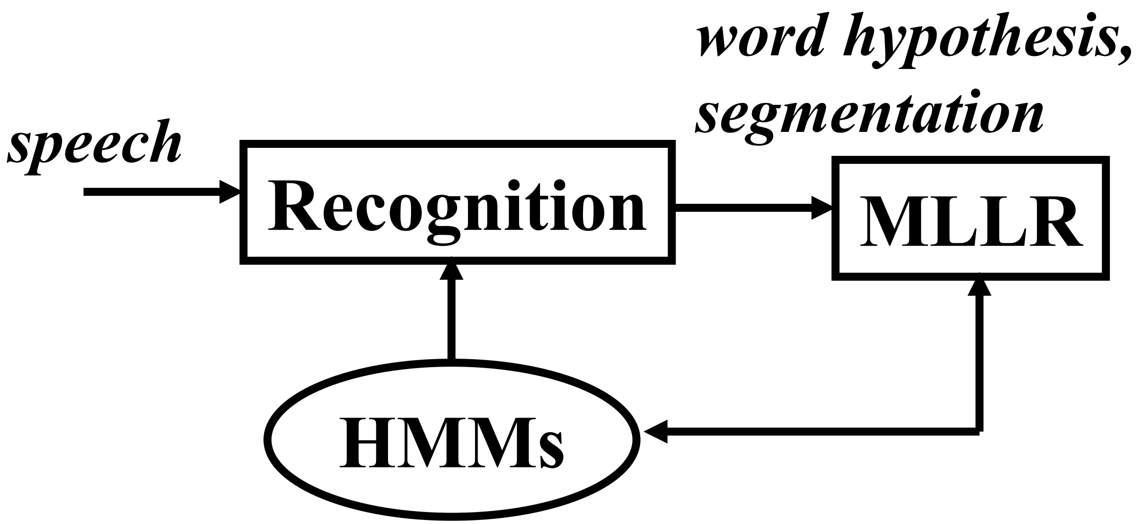
- Transform estimated to maximize the likelihood of the adaptation data
- Speech sounds can be clustered into a set of regression classes.
- Transform applied to all Gaussians within the same regression class.

MLLR Variance Adaptation

- **Can also adapt the Gaussian variances under the MLLR framework (Gales & Woodland, 1997)**
- **Variance adaptation provides only minor benefit as a technique for speaker adaptation**
  - 2-7%% relative reduction in WER
- **For speech recognition in noisy environments, adaptation of variances provides more benefit**
  - Assists in compensating for variance reduction due to noise

Iterative Unsupervised MLLR

- **Typically MLLR is applied in an iterative, and unsupervised manner,**



- **Typical error reductions on the order of 15-20% relative**

Word Error Rate vs. MLLR Iteration

- **Convergence is generally reached after 2-5 iterations of decoding followed by MLLR adaptation.**
- **Phoneme Recognition (Italian Children's Speech):**
  - First-Pass (WER) 18.1%
  - MLLR Iter #1 16.0%
  - MLLR Iter #2 15.3%
  - MLLR Iter #3 14.9%
  - MLLR Iter #4 14.7%
  - MLLR Iter #5 14.6%

19.3% Relative Error reduction

Maximum A Posteriori Adaptation (MAP)

- **MAP Adaptation can only be applied Gaussians that are "seen" in the test data,**

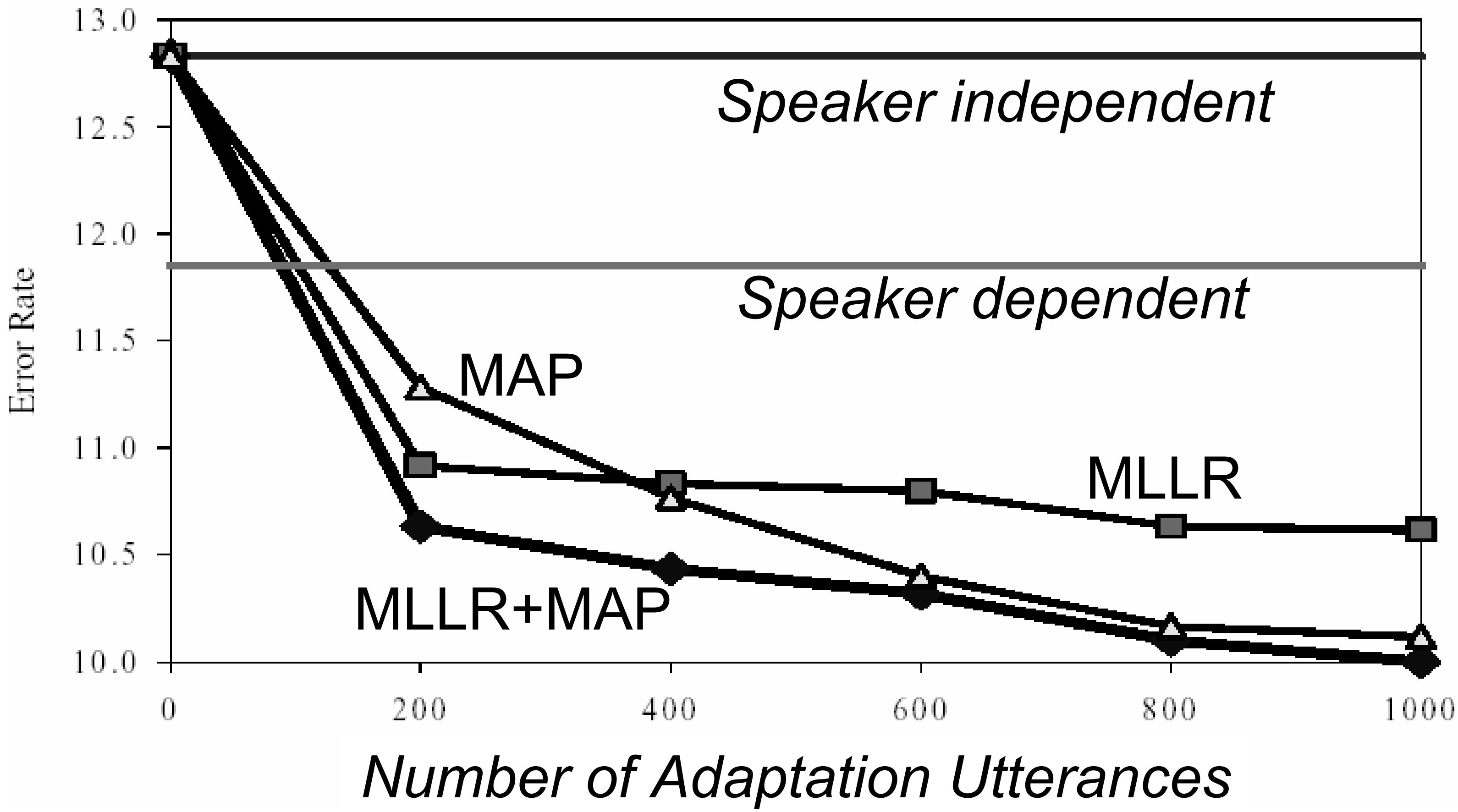
$$\mu_{new} = \frac{\hat{N}}{\hat{N}+\alpha}\hat{m}_{obs} + \frac{\alpha}{\hat{N}+\alpha}\mu_{old}$$

$\hat{N}$	Number of frames of adaptation data
α	Weight for prior estimate of old mean
$\hat{m}_{obs}$	Mean vector of adaptation data assigned to Gauss.

MAP vs. MLLR

- **MLLR Adaptation preferred over MAP for sparse adaptation sets**
  - MAP only re-estimates "seen" data components
  - Can estimate Block-Diagonal MLLR transforms
- **As adaptation data increases,**
  - MAP generally outperforms MLLR
  - Why? MLLR based on a fixed set of regression classes
- **Can combine MAP+MLLR**
  - Apply MLLR first,
  - Adapt "seen" Gaussians with MAP approach given sufficient data

Performance of MLLR and MAP



Constrained MLLR (CMLLR)

- **A variation on MLLR**
- **Transform can be applied to the <u>features</u> rather than to the model Gaussians**
- **Typically both CMLLR and MLLR are applied together for added benefit**

Speaker Adaptive Training (SAT)

- **Attempts to remove speaker-specific characteristics from the training data to build robust speaker-independent models**
- **SAT using Feature-Space CMLLR Transforms,**
  - Train standard acoustic models
  - Estimate CMLLR feature transform for each training speaker (use models from step #1)
  - Transform each speaker's features using speaker-dependent CMLLR transform
  - Retrain acoustic models

Benchmarking "Robust" Systems

- **DARPA SPINE**
  - "Speech Processing in Noisy Environments (SPINE)"
  - University and Commercial Several participants in 2001-2002
  - Human-to-Human Dialogs in various noisy environments
  - 3,000 word vocabulary
  - Word Error Rates ranging from 20-50%

- **AURORA**
  - Popular in several of the last speech conferences
  - End Goal: A standard for Distributed Speech Recognition
  - Attempts to compare systems using same database and recognizer, but allow researchers to propose new front-ends

What to Expect from Adaptation

- **Most techniques are used in combination**
- **Typical Research System might use,**
  - CMN+CVN+LDA(HLDA)+VTLN+MLLR+CMLLR
  - Speech Enhancement front-ends seem less common
- **Typical Relative Error Reductions**

Cepstral Mean Normalization		~ 5 %
Cepstral Variance Normalization		~ 5 %
LDA:		10 → 15%
VTLN:		5 → 10%
MLLR:	(unsupervised)	15 → 20%
	(supervised)	25 → 35%

Next Week

- **A few words about hypothesis combination**
- **ASR Course Review for 30 minutes**
- **Initial (Short) Project Presentations**
  - 10 minutes presentation about your project