Parallel Model Combination

PhD thesis of Mark John Francis Gales

Talk by Karthikesh Raju

Overview

- Basic Parallel Model Combination
- Cepstrum to Log-Spectral Domain
- The Mismatch Function
- Training HMMs on Statistical Data
- Non-Iterative PMC
- Iterative PMC
- Estimation of Convolutional Noise
- Conclusions

Basic Parallel Model Combination

- A technique model-based compensation scheme for training noisy data

- Techniques for compensating both static and dynamic coefficients

- Works in both additive and convolutive noise environments

- Model: Standard HMM with Gaussian emission probabilities

- Implies no modification to the existing HMM estimation formulae

$$\mathcal{L}(\mathbf{O}^c(\tau)|\mathcal{M}_s, \mathcal{M}_n) = \sum_{m=1}^{M} \hat{c}_m \mathcal{N}(\mathbf{O}^c(\tau); \hat{\mu^c_m}, \hat{\Sigma^c_m})$$

Basics - 2

- What is the best approach to estimate the new model parameters?

- Add noise to the clean training data to generate the new training database.

- Requires:
  - whole data base to be avbl online
  - availability of sufficient noise samples
  - a reasonable computation power
- Such a technique is impractical

- Use clean speech models that contain sufficient statistics of the training data, generate a model for background noise, and combine them accurately.

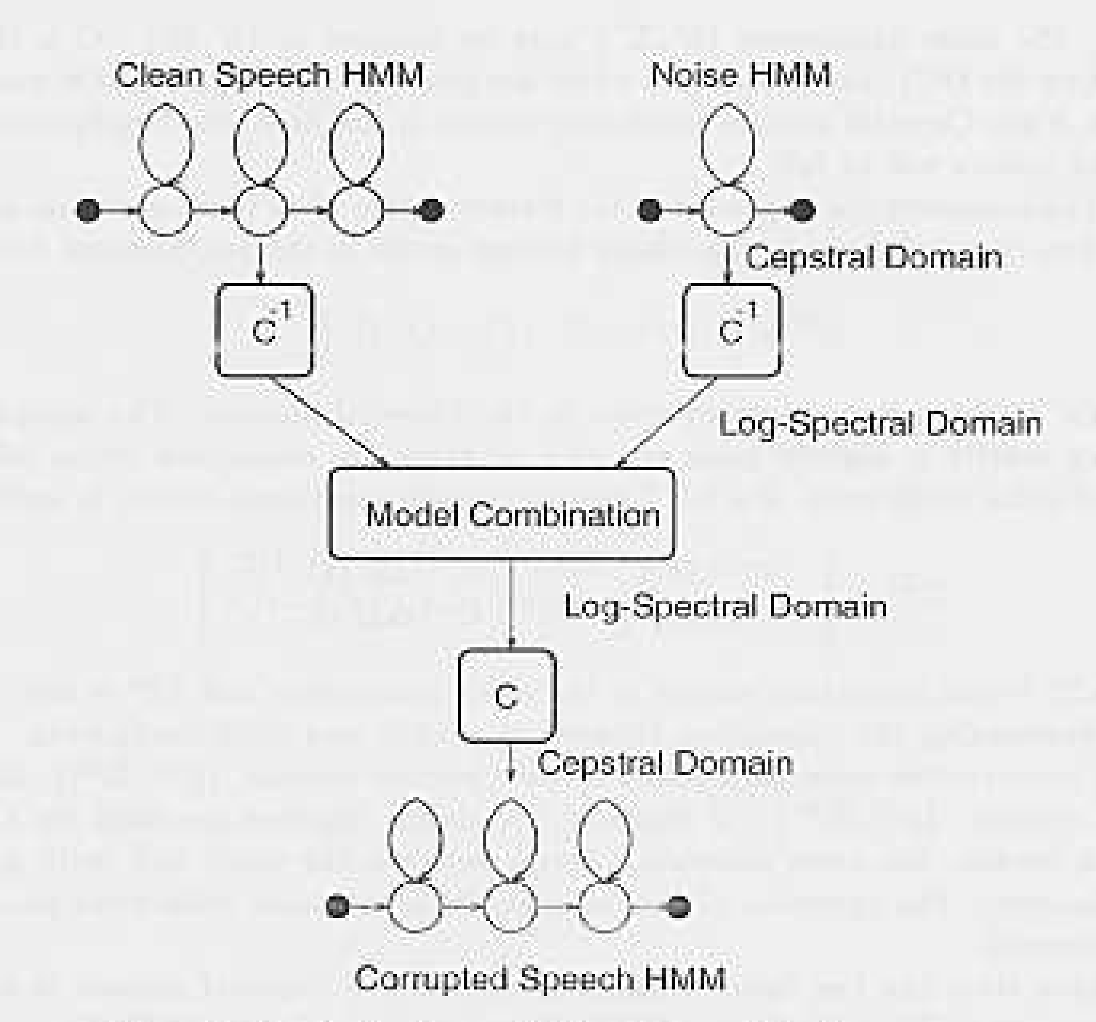

Figure 5.1: The basic parallel model combination process

- Inputs - clean speech and noise models

- Combination is expressed in linear spectral or log-spectral domains

- *Mismatch function*, approximates this combination

Cepstral to Log-Spectral Domain

- Required to map the cepstral parameters of the speech and noise models to the log-spectral domain

- Achieved using a DCT matrix ($\mathbf{C}$) as

$$\begin{aligned} \mu^l &= \mathbf{C}^{-1}\mu^c \\ \Sigma^l &= \mathbf{C}^{-1}\Sigma^c(\mathbf{C}^{-1})^T \end{aligned}$$

- Noise parameters can be similarly mapped

- Diagonal cepstral domain covariance matrices also results in full covariance matrices in the LS domain

- Mapping to Cepstral domain is a reverse operation. Replace inverse DCT $\mathbf{C}^{-1}$ by the DCT $\mathbf{C}$.

Mismatch Function

• Describes how noise affects each of the speech parameters

1. Speech and Noise are independent
2. Additive in time domain. Sufficient smoothing on the spectral domain also means they are additive at the psd level.
3. Single or multiple Gaussian mixture models sufficiently represents the observation vectors
4. Noise does not alter frame/state alignment

MF::Static Parameters

$$\begin{aligned} O_i^l(\tau) &= \mathcal{F}(S_i^l(\tau), N_i^l(\tau)) \\ &= \log(g \exp(S_i^l(\tau)) + \exp(N_i^l(\tau))) \end{aligned}$$

• g gain matching term, accounts for level differences between clean speech and the noisy speech.

MF:Dynamic Parameters

• Dymanic parameter coefficients attempt to model correlation between successive frames

• Observation feature vectors are functions of simple differences over a given window width w

• Statistics between successive frames $(\mathcal{E}(S^c(\tau + w)S^c(\tau - w)^T))$ known, delta parameters can be estimated using them

• But this increases model storage parameters

• Alternative is to re-express the delta parameters with terms of parameters whose statistics are known as

$$\begin{aligned}\Delta O_i^l(\tau) &= \mathcal{F}^{\Delta}(S_i^l(\tau - w), N_i^l(\tau - w), \Delta S_i^l(\tau), \Delta N_i^l(\tau)) \\ &= \log(\exp(\Delta S_i^l(\tau) + \Delta S_i^l(\tau - w) + g^l) + \\ &\quad \exp(\Delta N_i^l(\tau) + \Delta N_i^l(\tau - w))) - \\ &\quad \log(\exp(S_i^l(\tau - w) + g^l) + \exp(N_i^l(\tau - w)))\end{aligned}$$

• Similar mismatch function can be derived for delta-delta parameters.

MF:Additive and Convolutive Noise

- Observations under channel distortions and additive noise are

$$O_i(\tau) \quad = \quad H_i S_i(\tau) + N_i(\tau)$$

- Mismatch function can be expressed as

$$\begin{aligned} O_i^l(\tau) &= \mathcal{F}^H(S_i^l(\tau), N_i^l(\tau), H_i^l) \\ &= \log(\exp(S_i^l(\tau) + H_i^l) + \exp(N_i^l(\tau))) \end{aligned}$$

- So, the noise mismatch function for static case:

$$\mathcal{F}^H(S_i^l(\tau), N_i^l(\tau), H_i^l) \quad = \quad \mathcal{F}(S_i^l(\tau) + H_i, N_i^l(\tau))$$

- Dynamic parameters for clean speech are not affected as long as the noise in not time varying.

$$\mathcal{F}^\Delta(S_i^l(\tau - w) + H_i^l, N_i^l(\tau - w), \Delta S_i^l(\tau), \Delta N_i^l(\tau))$$

- Estimating H static speech parameters may be shifted and same estimation techniques can be used for additive and convolutive noise.

Training HMMs on Statistical Data

- HMMs here have only the statistics and not all the observations

- Requires modification of standard HMM rules

- HMMs have pdfs describing the observations for a particular states instead of training data.

- The statistics of the pdfs are represented by the models for clean speech and noise.

- $\mathbf{O}^c$ is the random variable associated with actual observations

- The state transition matrix contains temporal information, which implies the frame/state alignment is not altered

- Observation is alignment of a particular observation with a particular component M_m

of the current state.

• For state s_j, of the corrupted speech model $\hat{\mathcal{M}}$

$$p(M_m|\mathbf{O}^c, s_j, \hat{\mathcal{M}}) = \frac{c_m\mathcal{N}(\mathbf{O}^c; \mu_{jm}, \Sigma_{jm})}{\sum_{i=1}^{M} c_i\mathcal{N}(\mathbf{O}^c; \mu_{ji}, \Sigma_{ji})} = \mathcal{K}_m(\mathbf{O}^c)$$

• For multiple Gaussian mixture component HMMs

$$\hat{\mu}_{jm} = \frac{\int_{\mathcal{R}^n} \mathcal{K}_m(\mathbf{O}^c)\mathbf{O}^c\mathcal{L}(\mathbf{O}^c)d\mathbf{O}^c}{\int_{\mathcal{R}^n} \mathcal{K}_m(\mathbf{O}^c)\mathcal{L}(\mathbf{O}^c)d\mathbf{O}^c} \quad (1)$$

• Covariances can also be similarly re-estimated.

• The above scheme is iterative, ensuring frame/state alignment also allows single-pass, non-iterative, estimation.

• In single-pass case all mixtures are treated independently, as if each were in a distinct state

- This is **non-iterative PMC**

- If the frame/state alignment is allowed to vary, there are additional compensation problems

- Then **iterative PMC** estimation can be used.

Non-Iterative PMC

• Simplest and most computationally efficient

• Assumes frame/state alignment, implies re-estimation of STM or mixture components are not necessary

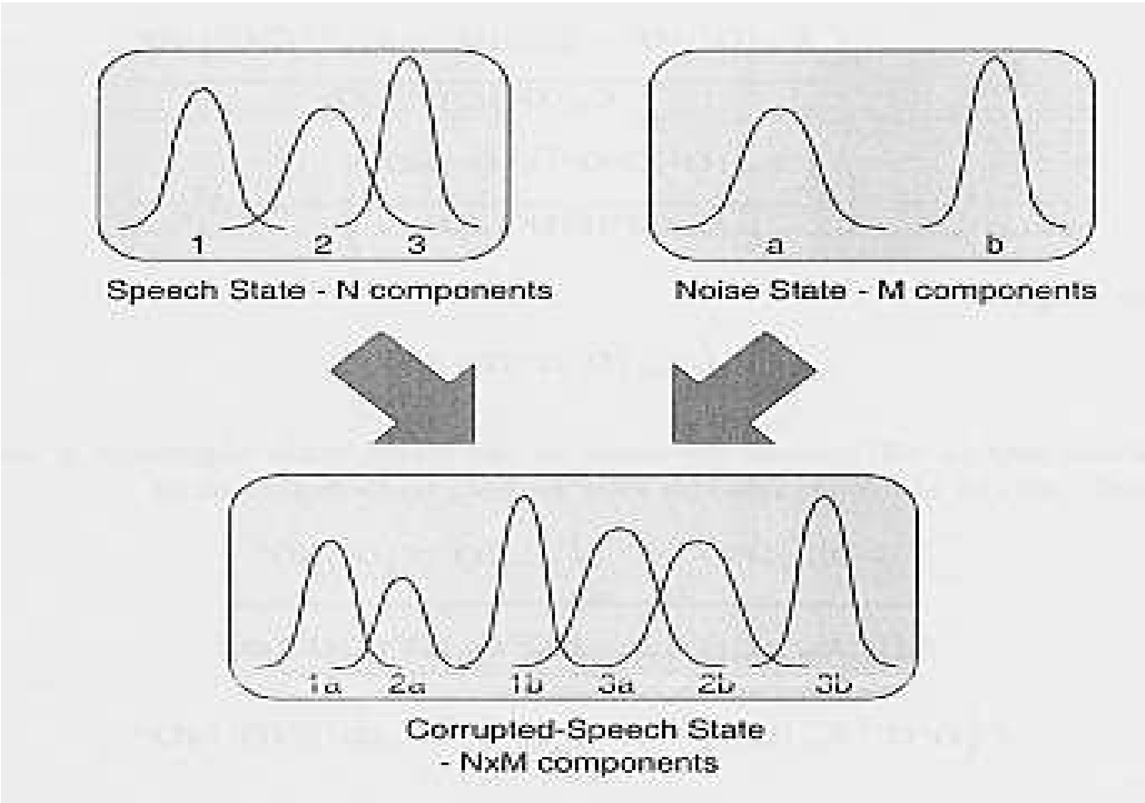

Figure 5.2: Non-iterative parallel model combination

• Most of the formulae here depend on the statistics at time $\tau - w$

• The formulae dont have a closed-form

expressions, hence it is necessary to make approximations about the distributions or use numerical integration techniques.

Log-Normal Approximations

- Sum of two log-normally distributed variables are also approx. log-normal

- Implies that mean and variance of the corrupted-speech model in the linear spectral domain is enough to derive the model in the cepstral domain

- Assuming Noise and speech are independent

$$\begin{aligned} \hat{\mu} &= g\mu + \tilde{\mu} \\ \hat{\Sigma} &= g^2\Sigma + \tilde{\Sigma} \end{aligned}$$

- The parameters for the clean speech in the linear spectral and log spectral domains are related by

$$
\begin{aligned}
\mu_i &= \exp(\mu_i^l + \sum_{ii}^{l} /2) \\
\Sigma_{ij} &= \mu_i \mu_j [\exp(\Sigma_{ij}^l) - 1]
\end{aligned}
$$

• Since, corrupted-speech is assumed to be log-normally distributed in the linear-spectral domain, the distribution in the log normal domain, $\{\hat{\mu}^l, \hat{\Sigma}^l\}$, are obtained by the inverses of the above equations.

• Usually the sum of two log-normally distributed variables are not log-normally distributed.

• This technique matches only the first two moments

• The log-normal approximation provides good approximation of the corrupted means

• Calculation of correlations is a complicated function and the log-normal approximation cannot be applied to compensate the delta parameter covariance matrix.

Log-Add Approximation

• Simplest approximation, where variances are assumed to be small, so that

$$\hat{\mu}_i^l \approx \mathcal{F}(\mu_i^l, \tilde{\mu}_i^l) \quad = \quad \log(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l))$$

• Analogous to Wiener filtering, the parameters can be written as

$$\hat{\mu}^c \approx \hat{\mu}^c + \mathbf{C}[\log(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)) - \mu^l]$$

• Hence, log-add approximation is a simple and efficient implementation of hypothesized Wiener filtering, with means obtained from the HMMs.

• Similar expressions exist for the delta parameters

• Covariance matrix not used: major reduction in computational load

Iterative PMC

- Assumption - for a particular speech component the corrupted-speech is Gaussian.

- Improved modeling can be achieved by allowing frame/state alignment to vary.

- The statistics are a function of all components of the feature vector

- Data-driven PMC (DPMC) is a reasonably complex PMC scheme.

- As, said, calculating statistics of all the n dimensions of the observation vectors is intractable.

- DPMC generates a speech "observation" and a noise "observation" for each pair of states and combines them using the mismatch function.

- HMM training is done assuming a multiple mixture component procedure using these "observations".

- Initial states can be obtained from non-iterative PMC

- By generating "observations", integration of Gaussian is avoided, all that is required is to accumulated appropriate statistics

- Computational overhead is dependent on the number of points generated

- Log-Add approx. - non-iterative, simplified and faster DPMC, which updates the means.

- Generates a single observation for each speech and noise component, at the means of the distributions.

- Combined with the mismatch function to yield the corrupted speech "Observation", which is an estimated of the mean.

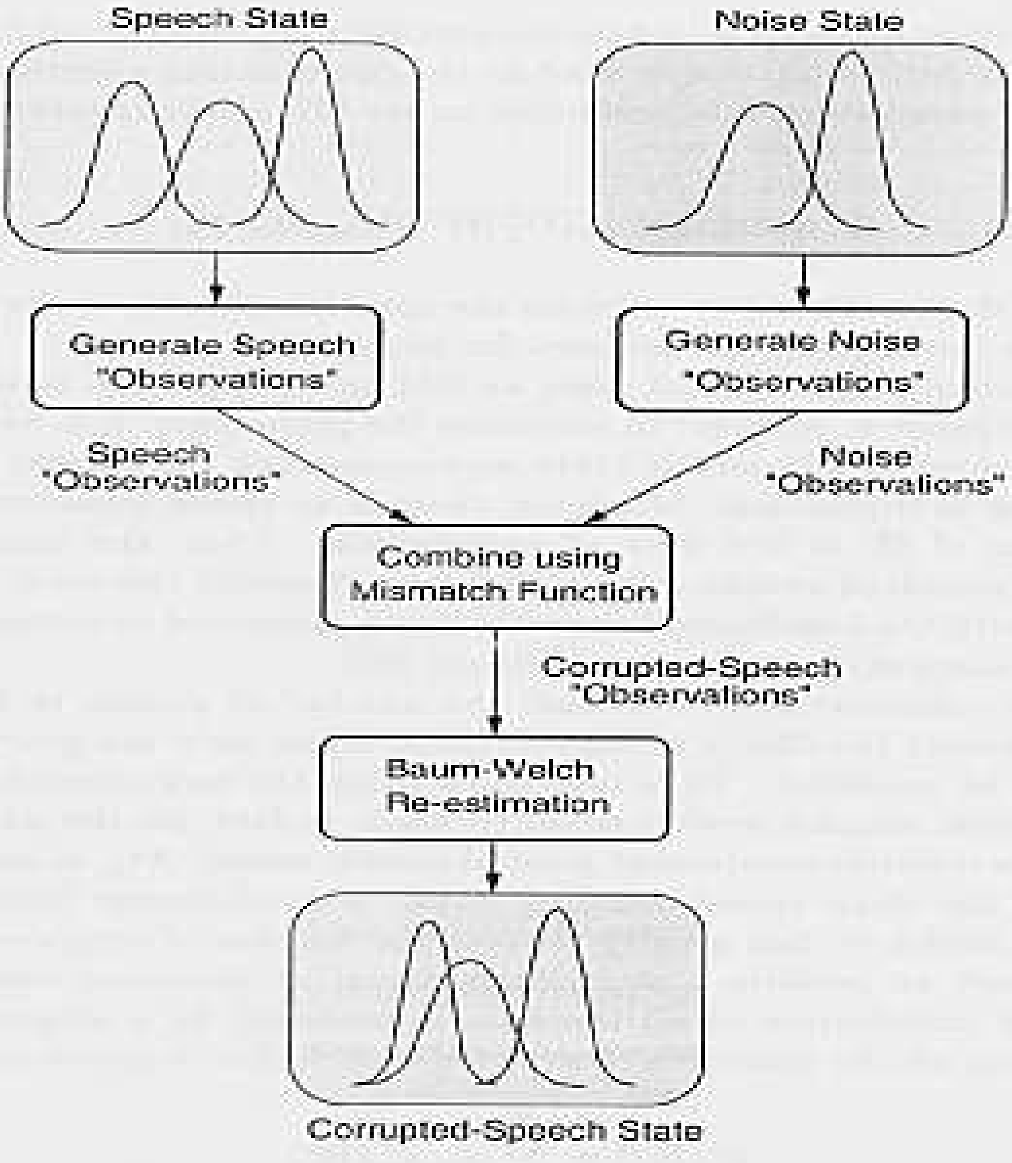

Figure 5.3: Data-driven parallel model combination

Estimation of Convolutional Noise Component

- ML estimation for obtaining statistics for the channel difference $\mathbf{H}^c$

$$\hat{\mathbf{H}}^c = \arg\max_{\mathbf{H}^c}[\mathcal{L}(\mathbf{O}_T^c|\mathbf{H}^c, \mathcal{M}_s, \mathcal{M}_n, \nu)] \qquad (2)$$

- $\mathbf{O}_T^c$ is the data available to estimate spectral tilt

- EM based approach is computationally infeasible

- Reduction of the number of models reduces computational load.

- Modeling the corrupted speech distribution by a single-component model, and combining all individual corrupted-speech distributions, we have

$$
\begin{aligned}
\hat{\mu}^c(\mathbf{H}^c) &= \sum_{m=1}^{M} c_m \hat{\mu}_m^c(\mathbf{H}^c) \\
\hat{\Sigma}^c(\mathbf{H}^c) &= \sum_{m=1}^{M} c_m (\hat{\Sigma}_m^c(\mathbf{H}^c) + \\
&\quad \hat{\mu}_m^c(\mathbf{H}^c)\hat{\mu}_m^c(\mathbf{H}^c)^T) - \hat{\mu}^c(\mathbf{H}^c)\hat{\mu}^c(\mathbf{H}^c)^T
\end{aligned}
$$

- $\hat{\mathbf{H}}^c$ is now estimated as

$$
\hat{\mathbf{H}}^c = \arg\min_{\mathbf{H}^c} [|\hat{\mu}^c(\mathbf{H}^c) - \frac{1}{T}\sum_{t=1}^{T} \mathbf{O}^c(t)|]
$$

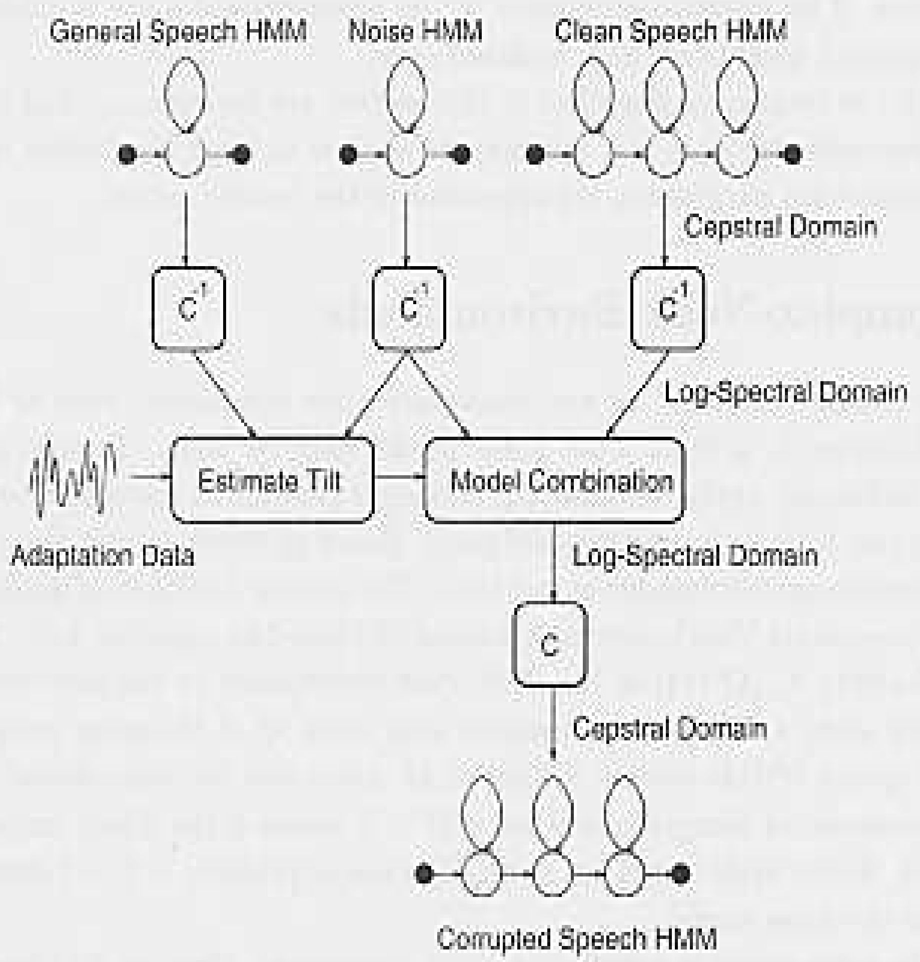

Figure 5.4: Parallel model combination for additive and convolutional noise

- Inputs: general speech model, adaptation data to estimate convolutional noise

- Is known to all the PMC methods

Limitations

- Gaussian distribution is not a good model for corrupted-speech data

- Higher order moments are ignored

- Implicit assumption that adaptation data is a representative of training data

- Size of adaptation data can affect results

- Little data: two-pass approach - estimate tilt data and then the weights for the noise model

- Requires a "good" representation of clean speech training data.

- Current estimates of $\hat{\mathbf{H}}^c$ can be used to obtain new alignment for the speech model

- Optimization is not a true maximization, convergence is not guaranteed

- Although they dont necessarily give a true ML estimate, they are fast and can be applied to large vocabulary systems, where delta and delta-delta parameters are appended to the feature vector

What i did not cover

- Variations of mismatch functions
- Handling complex and non-Gaussian noise
- Training models on noise-corrupted data
- Certain Performance Criteria
- Please refer:

$http : //svr -$
$www.eng.cam.ac.uk/\ mjfg/publications.html$

Conclusions

- How to train with noisy data?

- PMC is a model based compensation scheme.

- Depends on mismatch function, that approximates how this combination is

- Two schemes: Non-Iterative PMC (requires frame alignment) and Iterative PMC (does not require alignment)

- Adaptation data can be used to estimate and compensate the tilt between test and training data due to convolutional noise.