“Nice” Model-Based Compensation Schemes for Robust Speech Recognition

Mark Gales

mjfg@eng.cam.ac.uk

Emmanuel College

and

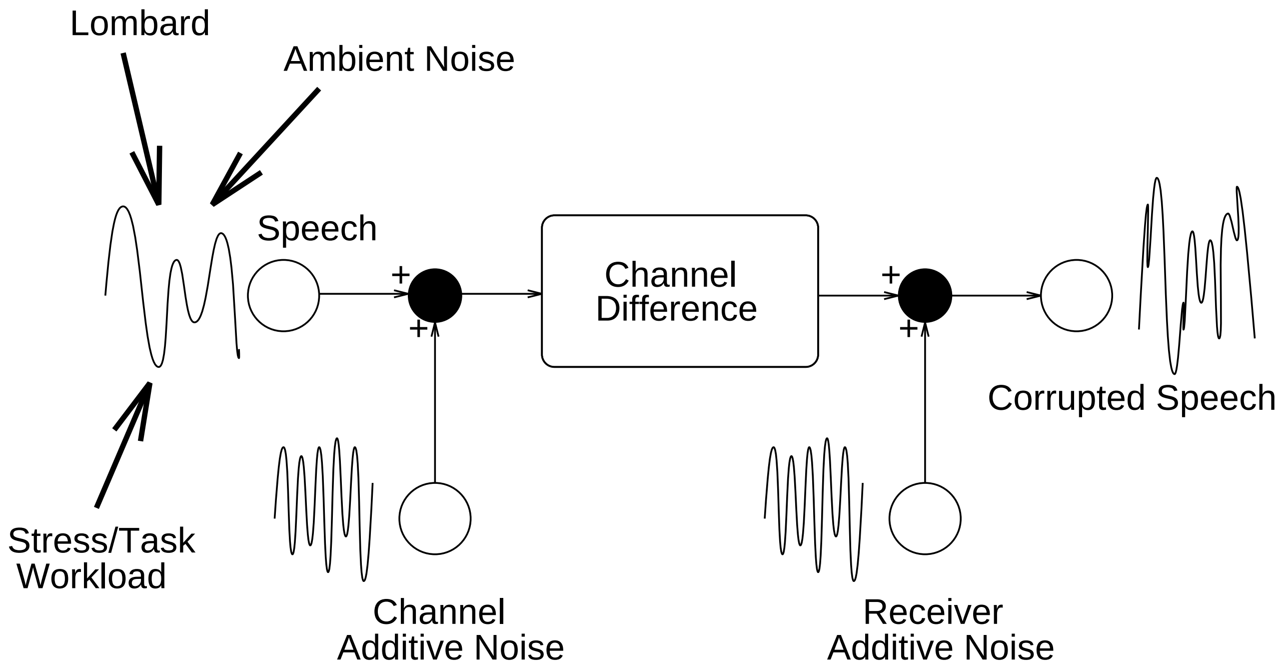
Department of Engineering

University of Cambridge

Talk Layout

- “Nice” Compensation Schemes
- Parallel Model Combination
- Advantages/Disadvantages
- Single-Pass Retraining
- PMC with Linear Transforms

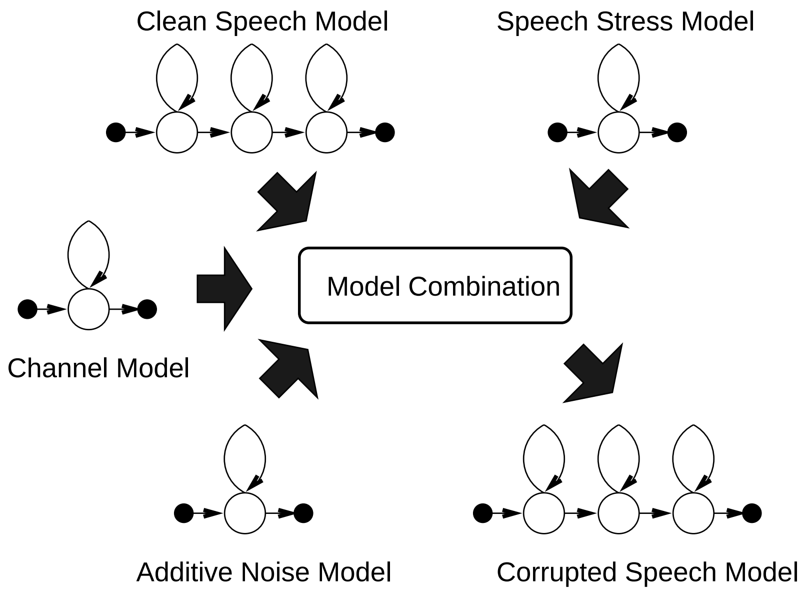
The Acoustic Environment



Typically stress ignored and all sources combined

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

"Nice" Compensation Framework



- What speech/noise models should be used?
- What domain should be used?
- How to combine the models?
- How to model the corrupted speech distribution?

"Mismatch Functions"

Additive and Convolutional Noise

$$o_i^l(\tau) = \log\left(\exp(s_i^l(\tau) + h_i^l) + \exp(n_i^l(\tau))\right)$$

Bandwidth Limited Channel:

$$\mathbf{o}^{bc}(\tau) = \tilde{\mathbf{W}}\mathbf{o}^c(\tau)$$

Delta parameters and delta-delta parameters may also be compensated

Parallel Model Combination

No simple closed-form solutions to mismatch functions when speech and noise are assumed to be Gaussian distributed. So

Numerical integration

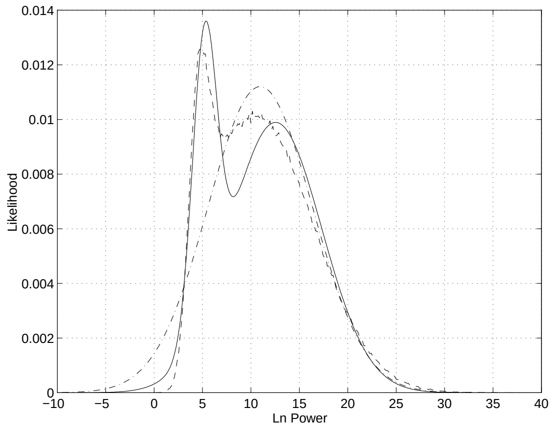
"Monte-Carlo" simulation

Log-normal approximation

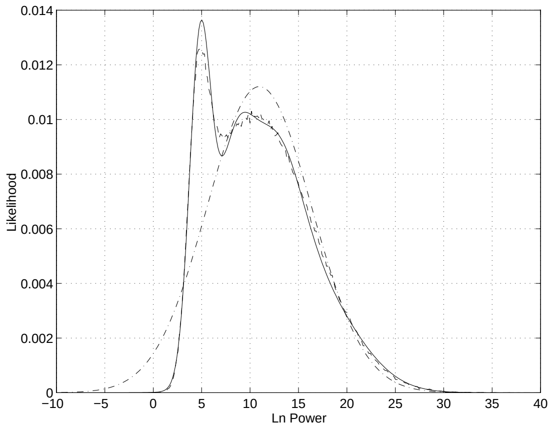
Log-add approximation

Iterative PMC

Use multiple-components to model corrupted speech distribution



(e) Two components



(f) Four components

Slight performance improvements, but computationally expensive.

Advantages

Parameters individually compensated

Meaningful models

Handles complex noise environments

No speech data required

Can handle non-Gaussian distributions

Disadvantages

Mismatch functions

Parameterisations

Noise models

Speaker characteristics

Accurate models

Speed

Single-Pass Retraining

Add "noise" to the actual speech data.

Assumes

Frame/state alignments not altered by noise.

Eases

Mismatch functions

Parameterisations

Accurate models

Problems

- Requires complete database on-line;
- Computational overhead to re-code the data;
- Computational overhead to re-train models.

PMC with Linear Transforms

$$\hat{\mathcal{M}} = \mathcal{G}(\mathcal{F}(\mathcal{M}_s, \mathbf{o}_T) \otimes \mathcal{M}_n, \mathbf{o}_T)$$

1. $\mathcal{F}(\mathcal{M}, \mathbf{o}_T)$ models to a certain extent the convolutional noise, speaker stress and speaker characteristics. Usual form

$$\hat{\mu}^c = \mu^c + \mathbf{h}^c$$

2. $\mathcal{G}(\mathcal{M}, \mathbf{o}_T)$ takes into account inaccuracies in the mismatch functions and limitations in $\mathcal{F}(\mathcal{M}, \mathbf{o}_T)$. *Any* standard form may be used.

Estimating $\mathcal{F}(\mathcal{M}, \mathbf{o}_T)$

Various schemes (Gales and Young '93, Gauvain et al '96).
Nice scheme (Minami and Furui '96)

1. Make initial estimate of $\mathbf{h}^c$, $\mathbf{h}^{c(0)}$, $k = 0$.

2. Calculate

$$\delta h_i^{c(k+1)} = \frac{\sum_{\tau=1}^{T} \sum_{m=1}^{M} \frac{\gamma_m(\tau)(o_i^c(\tau) - \hat{\mu}_i^{c(k)})}{\hat{\sigma}_i^{c(k)2}}}{\sum_{\tau=1}^{T} \sum_{m=1}^{M} \frac{\gamma_m(\tau)}{\hat{\sigma}_i^{c(k)2}}}$$

   where

$$\hat{\mathcal{M}}^{(k)} = \mathcal{F}(\mathcal{M}_s, \mathbf{h}^{c(k)}) \otimes \mathcal{M}_n$$

3. Set $\mathbf{h}^{l(k+1)} = \mathbf{h}^{l(k)} + \delta \mathbf{h}^{l(k+1)}$ and unless "converged" goto step 2 with $k = k + 1$.

Maximum Likelihood Linear Regression

General mean transform solved (Leggetter and Woodland '95)

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{LHL}^T \quad \text{or} \quad \hat{\mathbf{\Sigma}} = \mathbf{H\Sigma H}^T$$

Usually block-diagonal or full mean transforms and diagonal variance transforms are used.

Experiments

System description:

- Cross-word, state-clustered, triphone system (SI-284 trained);
- Trigram language model, 5K vocabulary;
- 12 MFCC + C0 + deltas + delta-deltas
  (12 MFCC + log-energy + deltas + delta-deltas + CMN);

Tasks (ARPA 1994 CSRNAB data):

1. **Additive Noise (S10)**: Car noise was artificially added onto the clean speech database at various SNRs.
2. **Unknown Microphone (S5)**: The test set was recorded with an unknown microphone.

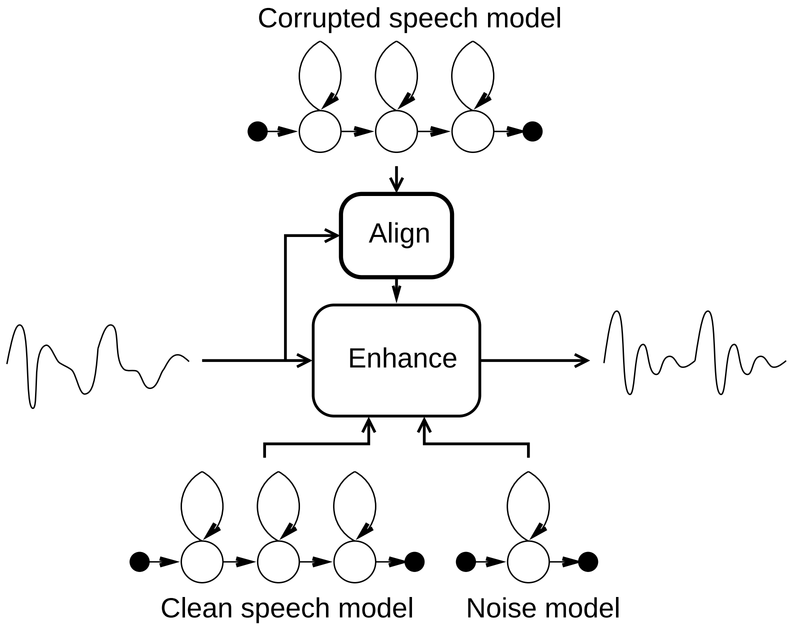
Results

Task	Model Set	Noise Condition			
		Clean	Level 1	Level 2	Level 3
S10	Baseline (+CMN)	5.8	30.0	14.7	54.3
	Baseline (-CMN)	6.7	35.0	20.4	59.3
S10	Diagonal (-$\mathcal{F}$(.))	6.7	10.3	9.4	15.3
S10	Diagonal	6.7	7.8	7.2	10.3
	Diagonal (Mean)	—	7.7	6.9	10.7
	Extended	—	7.5	7.2	10.1

Results

Task	Model Set	None	Means	Means +Vars
S5	Baseline (+CMN)	17.4	12.1	—
	Log-Add	10.3	8.6	8.0
S10	Baseline (+CMN)	54.3	—	—
	Log-Add	10.7	9.3	8.9

Enhancement



Various domains examined:

- Waveform (Ephraim '89)
- Linear-spectral (Gagnon '92)
- Cepstral (Seymour '96)

Conclusions

Parallel Model Combination - Useful!