Speaker Verification Under Additive Noise Conditions With Non-stationary SNR Using PMC

Michael L P Wong

&

Martin J Russell

Digital Systems & Vision Processing

THE UNIVERSITY OF BIRMINGHAM

References

- M.J. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEE Transactions on Speech and Audio Processing, Vol. 4, No. 5, pp. 352-359, September 1996.

- T. Matsui, T. Kanno, S. Furui, "Speaker Recognition Using HMM Composition in Noisy Environments," Computer Speech and Language, Vol. 10, pp. 107- 116, 1996.
- O. Bellot, D. Matrouf, T. Merlin and Jean-Francois Bonastre, "Additive and Convolutive Noises Compensation for Speaker Recognition", Proceedings of the ICSLP 2000 Beijing, China, 2000.

Task Definition

- Clean verification speech : Good
- Noise-contaminated verification speech with non-stationary SNR : Bad

Preview of Results

- Clean speech models tested on non-stationary SNR phrases
  - Speech noise : 38.55% EER
  - Operations room noise : 34.78% EER

- Performance of compensated models
  - Speech noise : 19.92% EER
  - Operations room noise : 18.84% EER

Structure of Presentation

- Stage One
  - Evaluation of PMC on speaker verification tasks : stationary SNR conditions
- Stage Two
  - Recognition of unknown SNR conditions
- Stage Three
  - Modelling the dynamics of SNR in noise-contaminated verification phrases

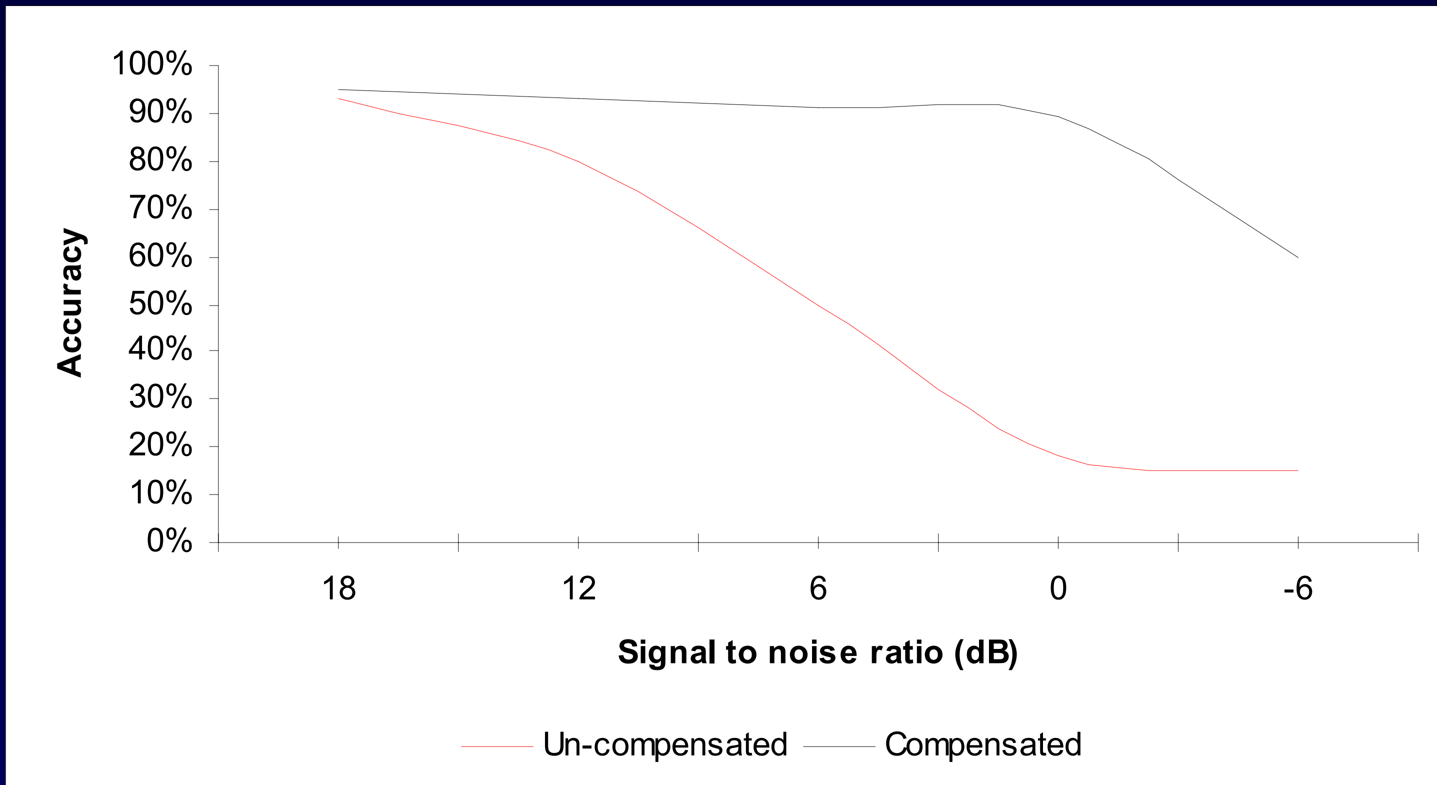
Problem Formulation

- Text-dependent speaker verification
- Deployment in dynamic real world environments
- Model based approach
- Ultimately multi noise multi SNR scenario

Evaluation Using PMC

- Successful in improving the performance of ASR systems
- Based on work by Mark Gales
- Evaluate use of PMC in text-dependent speaker verification tasks

Performance of PMC in ASR Experiments



Reference : Gales

Design Criteria

- Additive noises considered
- Scaling to be performed on noises

$$\mu^{l}_{S \otimes N} = \log(\exp(\mu^{l}_{S}) + g\exp(\mu^{l}_{N}))$$

$$\Sigma^{l}_{S \otimes N} = \log(\exp(\Sigma^{l}_{S}) + g\exp(\Sigma^{l}_{N}))$$

- Compensate only for static parameters

Implementation

- Selection of databases
- Preparation of data
- System Structure
- Scoring Procedures

Selection of Databases

- Yoho speaker verification database
  - Standard database used, performance comparison available
- Timit database
  - Used for the initialisation of isolated phone models prior to Yoho training
- Noisex-92 noise database
  - Selection of repetitive noise sources. Two noise sources reported in this paper. Speech noise and operations room noise

Preparation of Data

- Scaling of both enrolment and verification data
- Measurement of verification speech power
  - Silence periods ignored [ref 7, ITU-T Rec.]
- Mixing of speech and noise from –18dB to +18dB at 6dB intervals. Retain multiplication factor, g, and take an average

System Structure

- Front-end
  - 25ms, Hamming windowed, MEL scale warped
  - 12 cepstral coefficients with 0th energy appended, 1st and 2nd order derivatives included
- HTK Software for both training and recognition
- 3 state 4 component tied-triphone speaker dependent models, 1 state 4 component noise models

System Structure

- Training
  - 96 phrases per speaker
  - 118 authorised
  - 20 for General Speaker model
- Recognition
  - 40 phrases used for both FR and FA experiments

Scoring Procedures

- Likelihood ratio test employed

$$\boxed{\frac{P(X \mid S)}{P(X \mid GSM)} \geq t}$$

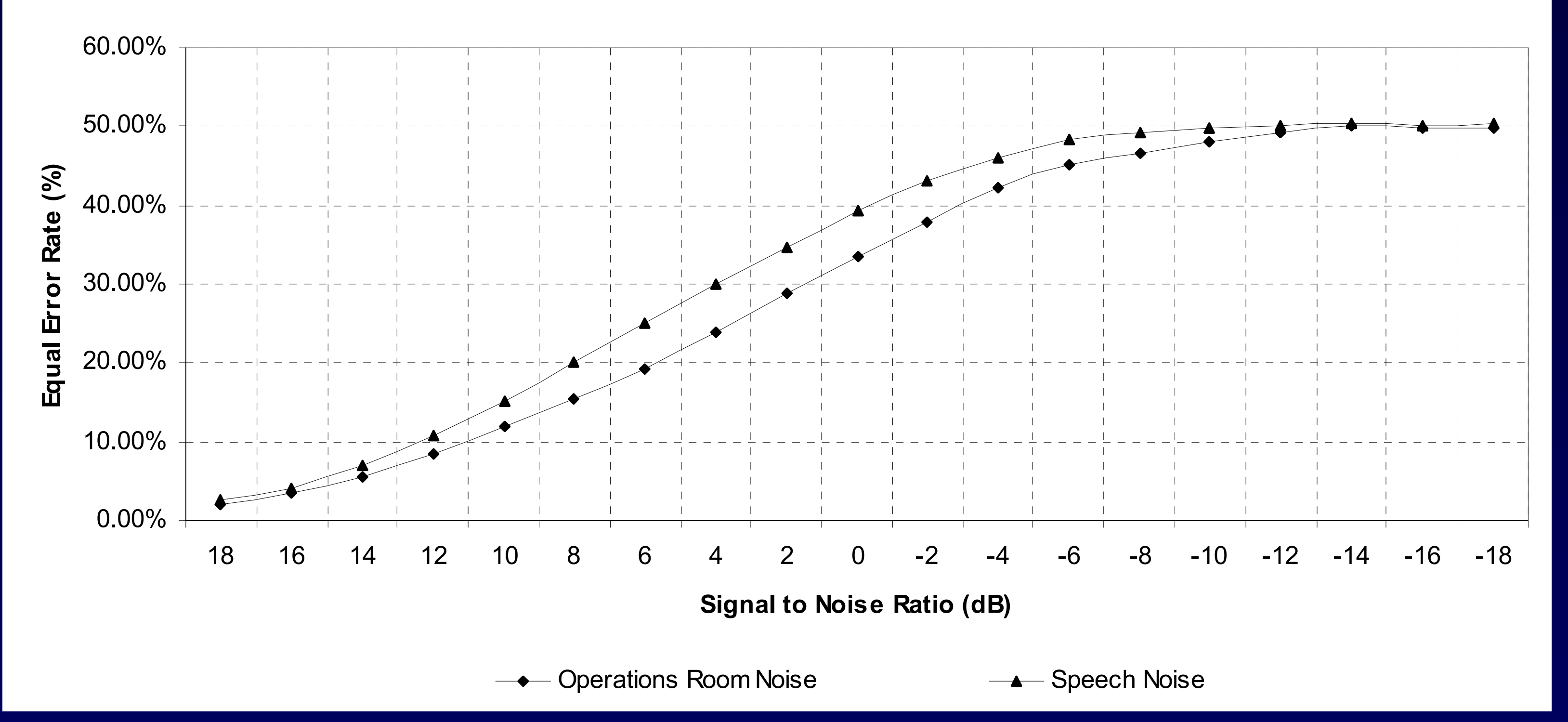
- Performance quoted in % EER

Experiment Methodology

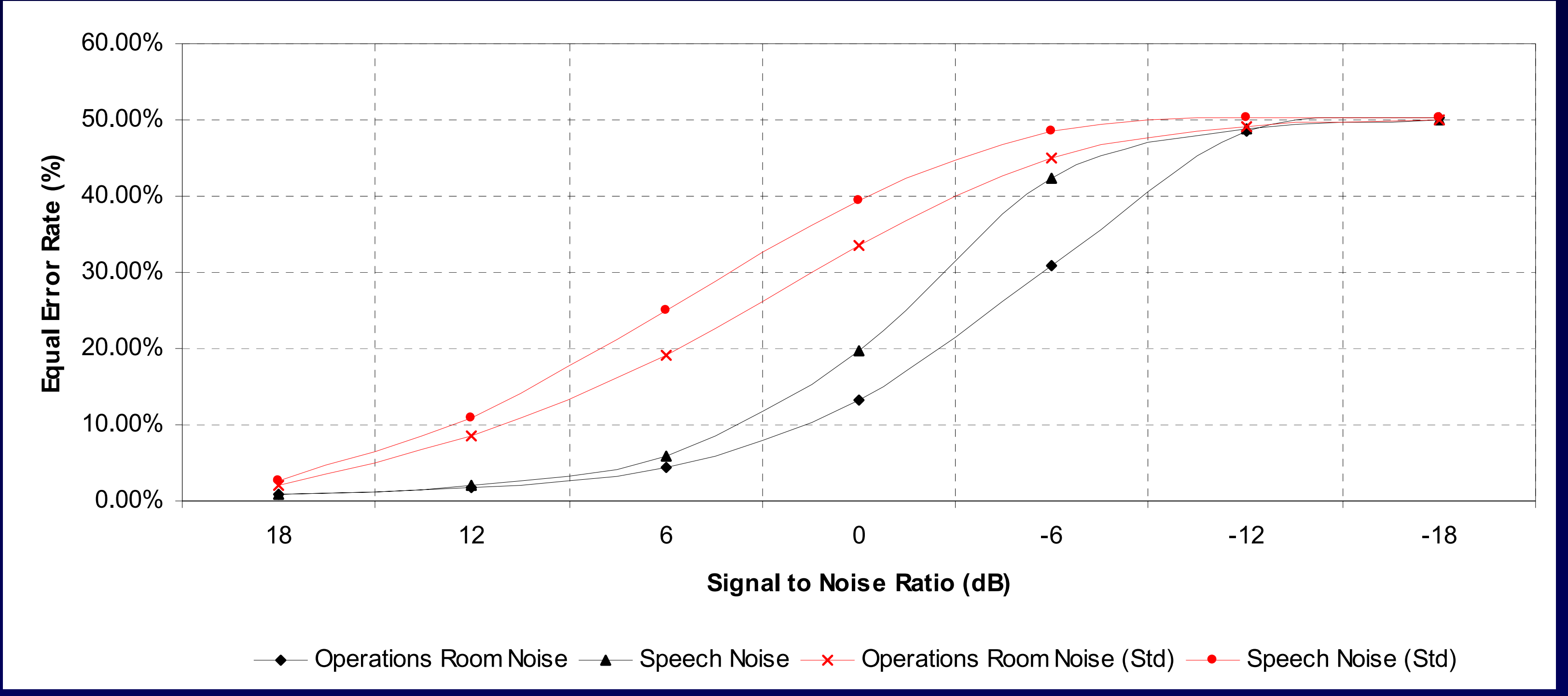
- Establish baseline performance using clean speaker models and clean verification data
- Evaluate performance of clean speaker models under multi SNR verification data
- Evaluate performance of PMC compensated speaker models under multi SNR verification data

Un-compensated Models

Clean speech and models performance = 0.57%



Compensated Models



Digital Systems & Vision Processing

THE UNIVERSITY OF BIRMINGHAM

Stage One Summary

- Text-dependent SV task
- HTK Software used with modifications for PMC
- Yoho, Timit and Noisex-92 databases used
- 7 SNR scenarios considered (-18dB to +18dB)

Stage One Summary

- PMC improves SV performance
- 2 additive noises considered
- Static parameters compensated
- Baseline used : clean models, clean/contaminated speech

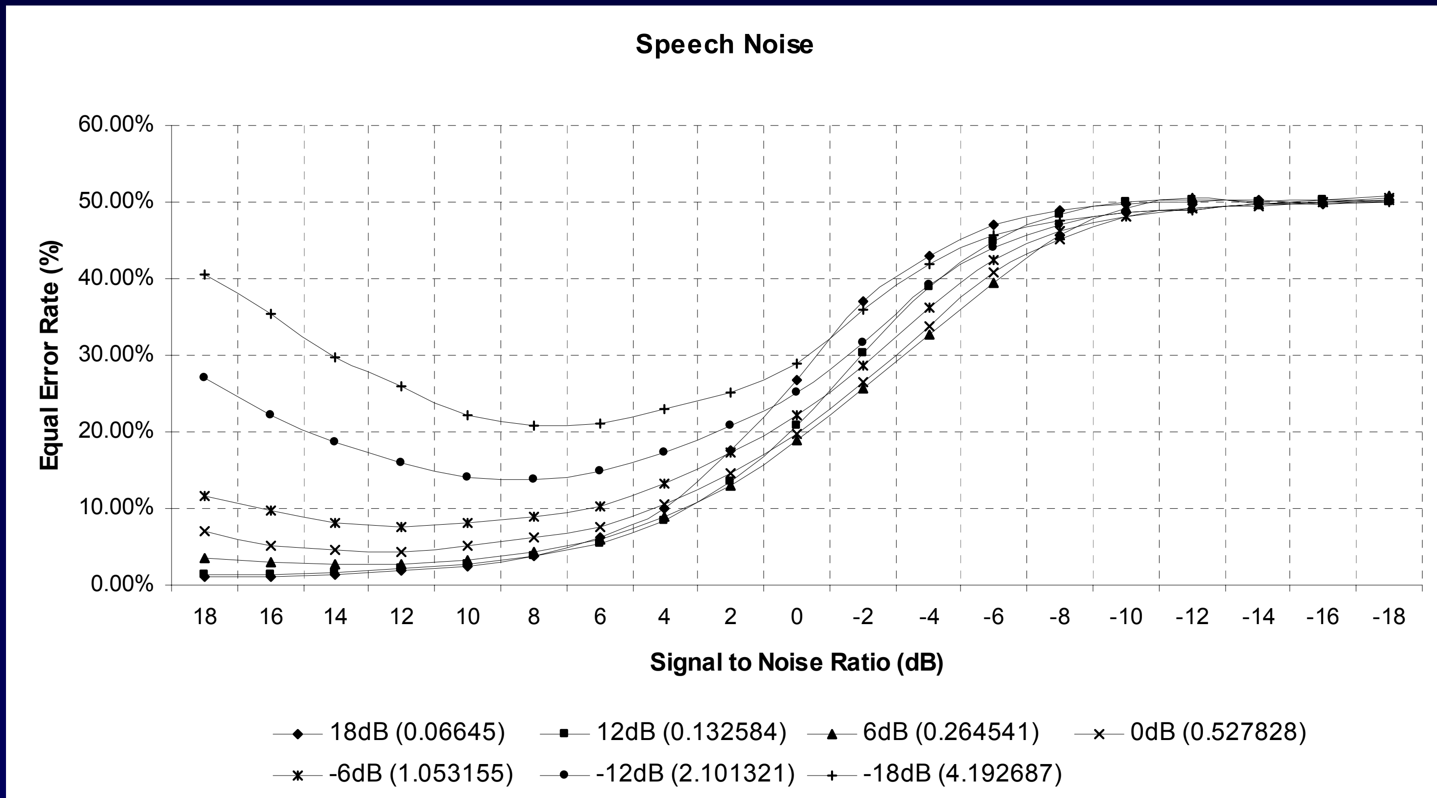
Experimental Extension

- We now have 7 SNR specific PMC models
- Can SNR specific PMC models be used for other SNRs? How sensitive are they?
- If yes, how well do they perform?

Evaluation of Non-ideal PMC Models

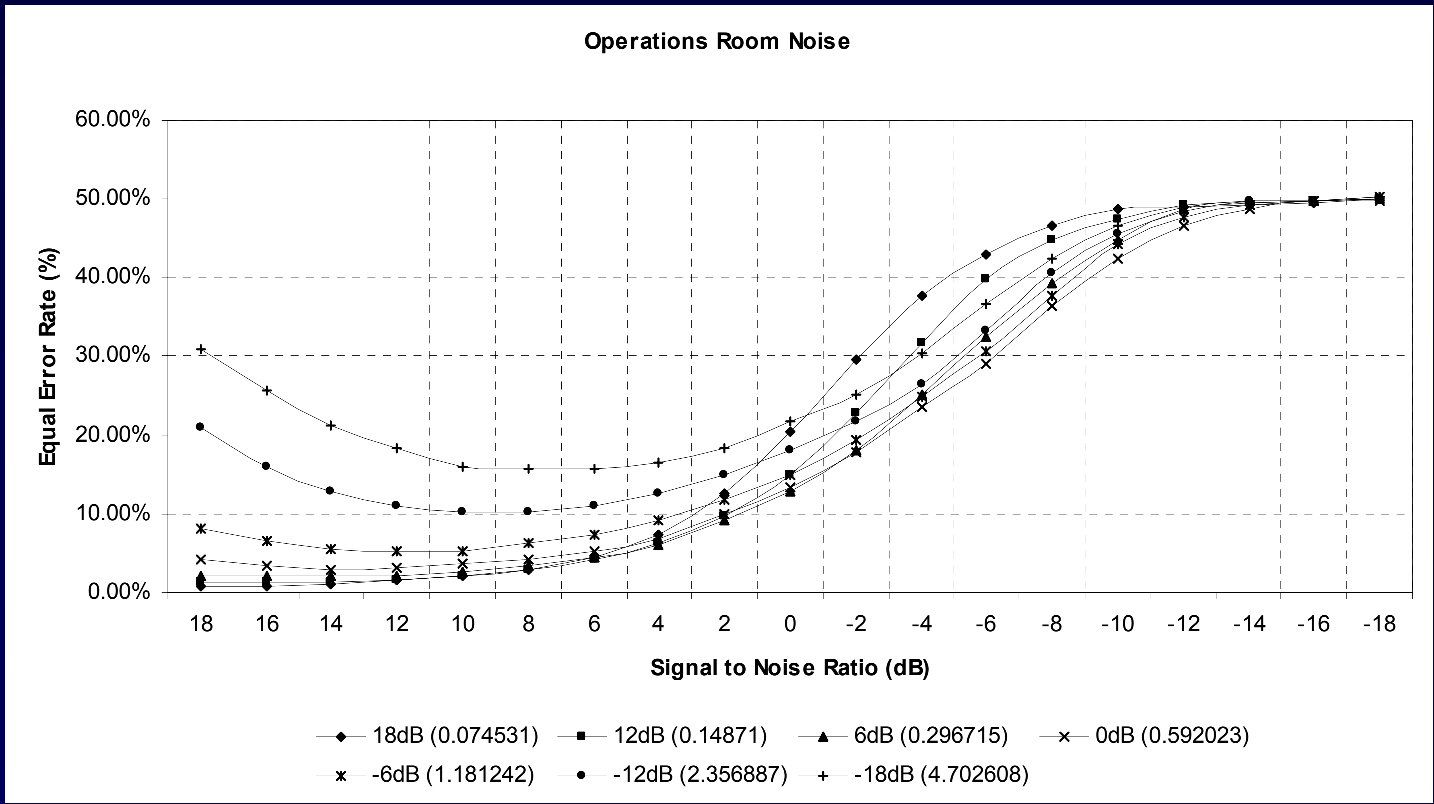
- For each SNR specific PMC model, perform SV task on noise contaminated verification phrases from –18dB to +18dB at 2dB intervals
- Observe any degradation in performance from using non-ideal models

Speech Noise Result



Digital Systems & Vision Processing
THE UNIVERSITY OF BIRMINGHAM

Operations Room Noise Result



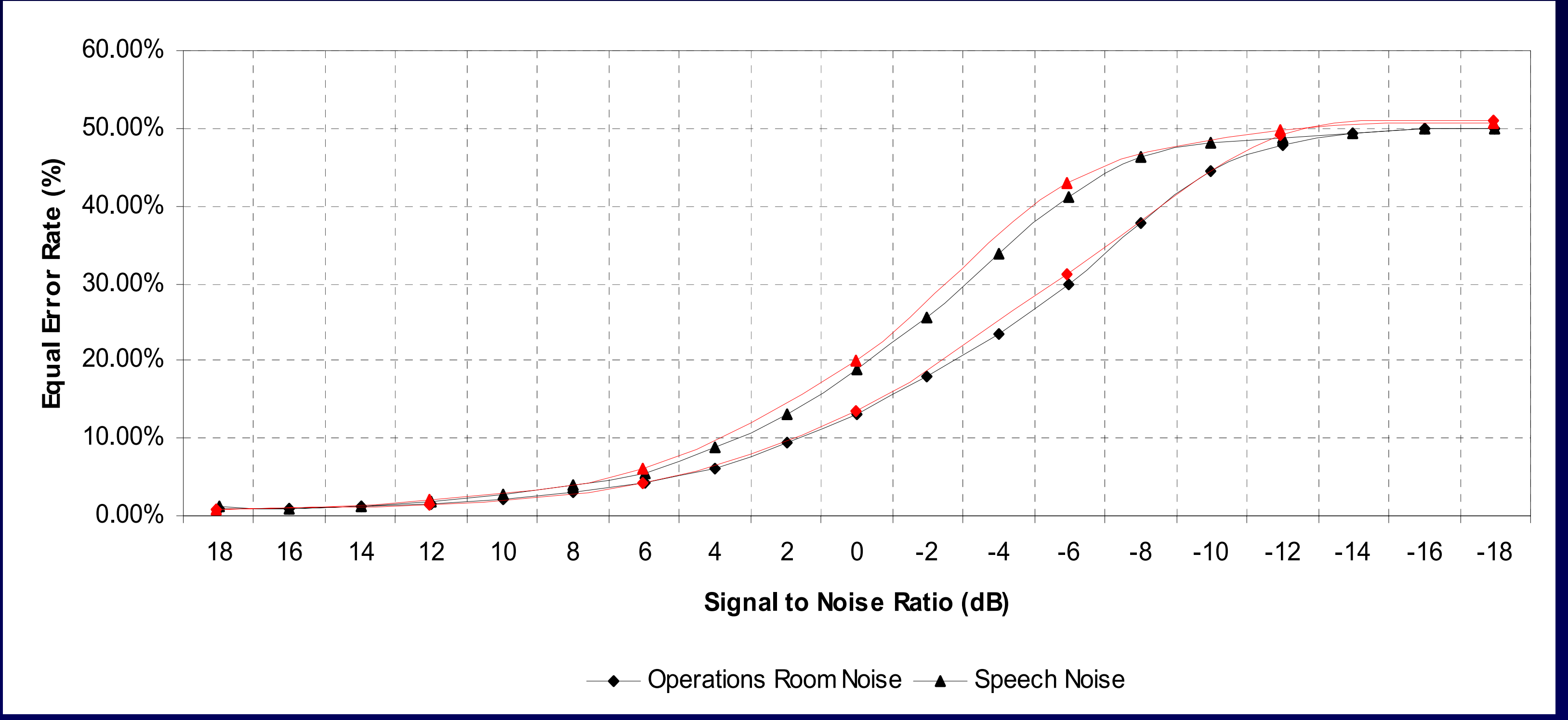
Digital Systems & Vision Processing

THE UNIVERSITY OF BIRMINGHAM

Discussion

- Allow the selection of SNR specific PMC models based on which has the highest probability for a given observation

Automatic Model Selection



Digital Systems & Vision Processing

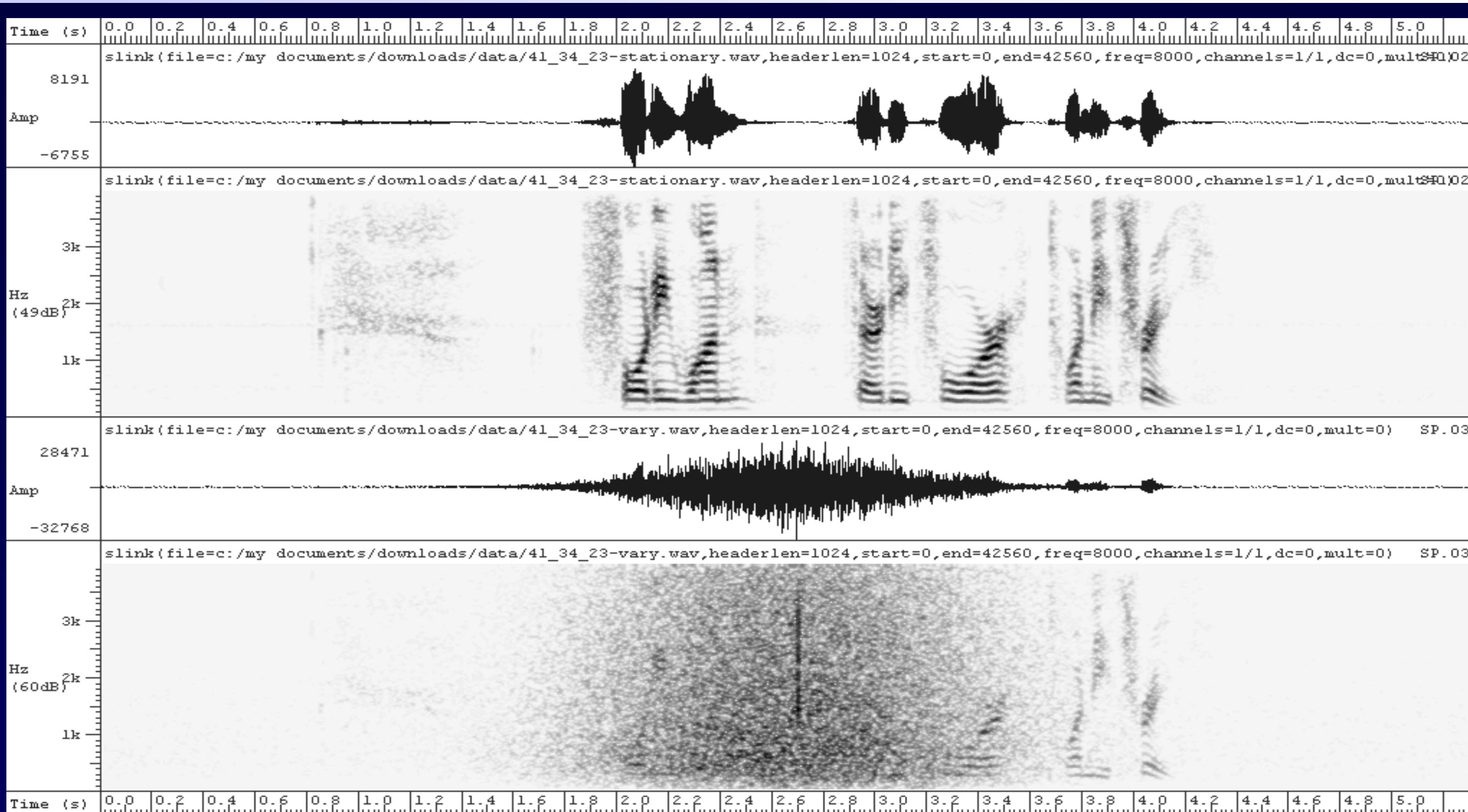
THE UNIVERSITY OF BIRMINGHAM

Stage Two Summary

- Limiting the number of SNR specific PMC models to 7 does not affect SV performance on unknown SNR

- Better performance is achieved by automatic selection of models

Varying SNR Task



Modelling SNR Dynamics

- Operating models in parallel assumes that SNR changes occur at model boundaries

- Create one model from multiple models, with the SNR dynamics embedded within the transition probabilities

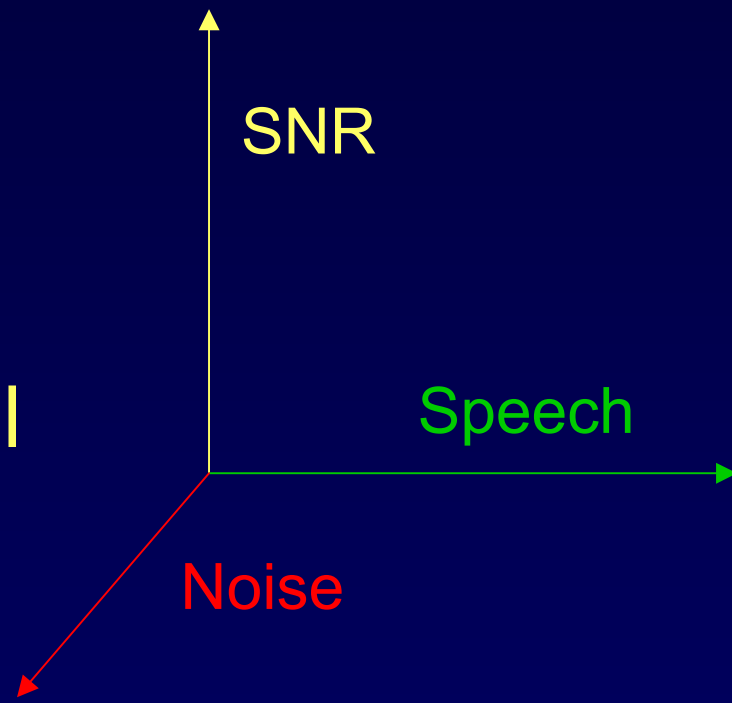
Implementation of a Composite HMM

- Rows and columns correspond to different SNR, 1st row = entry probability

$$\begin{array}{r} \textit{Entry} \\ +\,18dB \\ +\,12dB \\ +\,6dB \\ 0dB \\ -\,6dB \\ -\,12dB \\ -\,18dB \end{array} \begin{bmatrix} 0.3 & \boxed{0.2} & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.4 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.4 \end{bmatrix}$$

Implementation of a Composite HMM

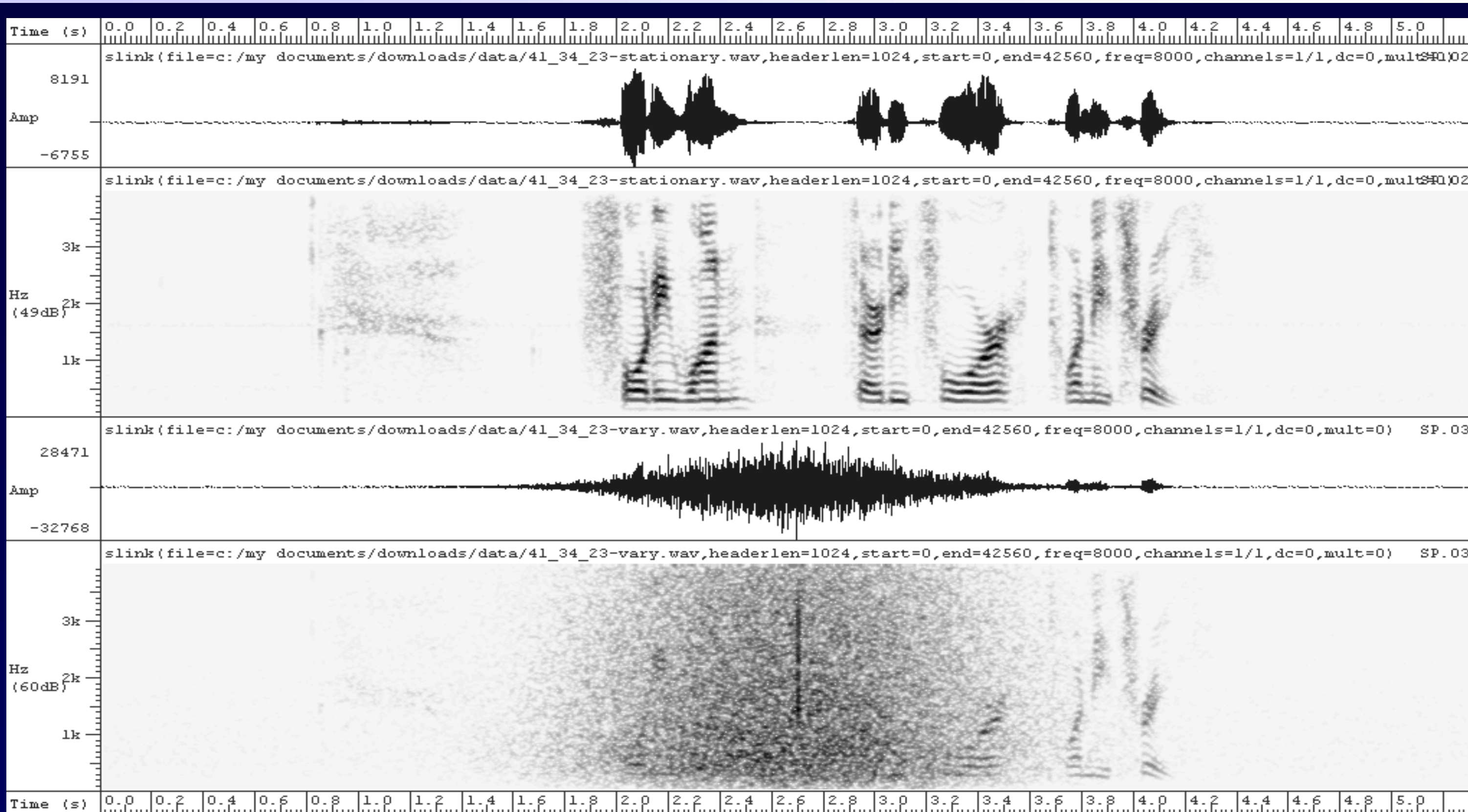
- 3 dimensional model
- 1 state noise model
- 3 state speech model
- 7 state SNR model



Expectations

- Extracting true SNR dynamics and embedding it into the transition probabilities will further improve performance [to be evaluated]

Varying SNR Task



Evaluation Using Non-stationary SNR Utterances

- Clean speech models tested on non-stationary SNR phrases
  - Speech noise : 38.55% EER
  - Operations room noise : 34.78% EER

- Performance of compensated models
  - Speech noise : 19.92% EER
  - Operations room noise : 18.84% EER

Stage Three Summary

- Composite 3-D HMM created
- SNR dynamics embedded into transition probabilities
- Improvement in performance observed

Conclusion

- PMC improves SV performance under both stationary and varying speech SNR
- SNR dynamics can be embedded into the HMM structure, providing additional information

Work In Progress

- Currently : known noise, unknown SNR
- Ideally : unknown noise, unknown SNR
- Tracking SNR transitions
- Comparison with other robust methods
- Establishing another baseline using matched recognition