# An HMM Decomposition Approach applied to AURORA 2

Speech and Hearing Research Group,

Dept. Computer Science,

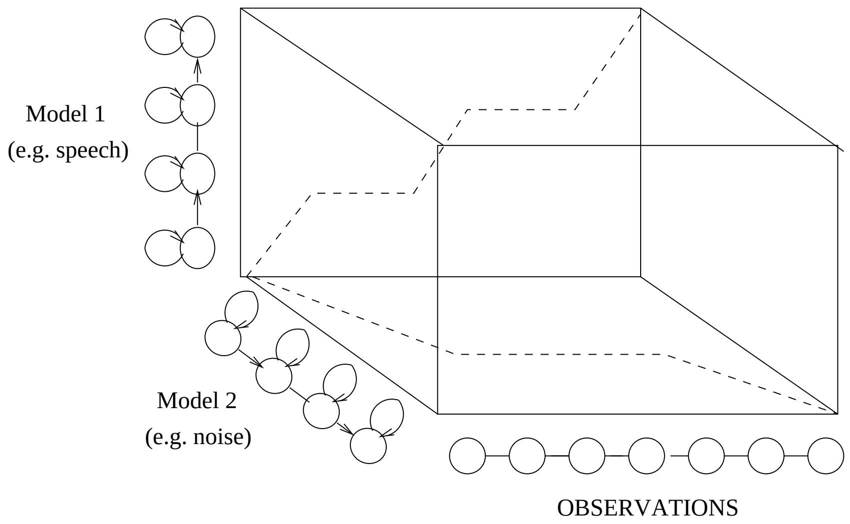
University of Sheffield, UK

January 8, 2002

## Overview

- HMM Decomposition
- Constructing a Noise Model
- Employing Delta Features
- Employing Harmonicity Information
- Results + Conclusions

# HMM Decomposition



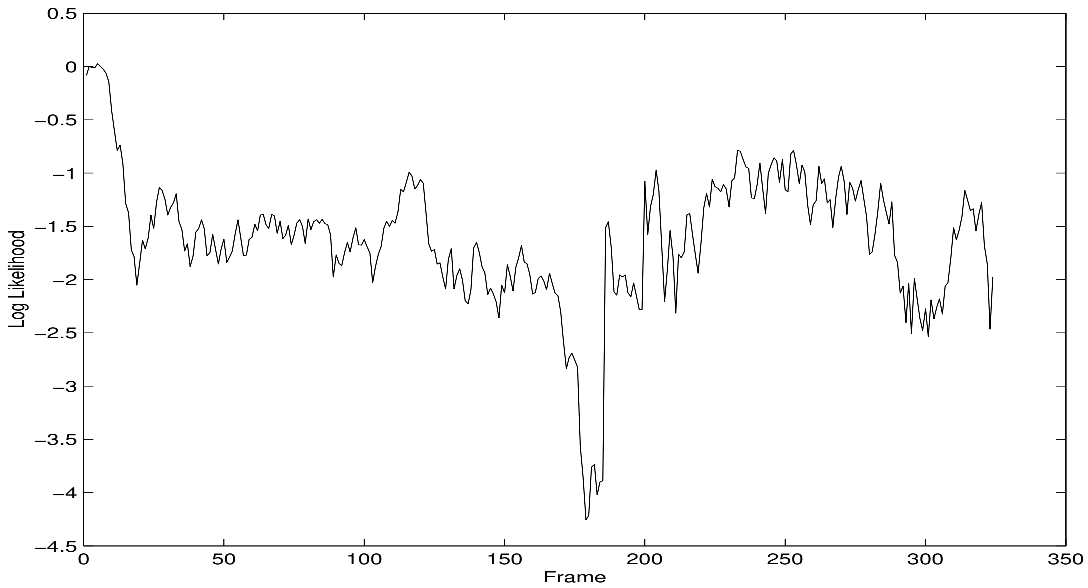

$$X_t = \log(\exp(X1_t) + \exp(X2_t)) \approx max(X1_t, X2_t)$$

$$P(X_t|i,j) =$$
$$C(X_t, \mu1_i, \sigma1_i)N(X_t, \mu2_j, \sigma2_j) + C(X_t, \mu2_i, \sigma2_i)N(X_t, \mu1_j, \sigma1_j)$$

# Estimating the Noise Model

**Stationary Noise Model:**

Assume Gaussian distribution and use first N frames to estimate noise $\mu$ and $\sigma$ for each frequency channel.

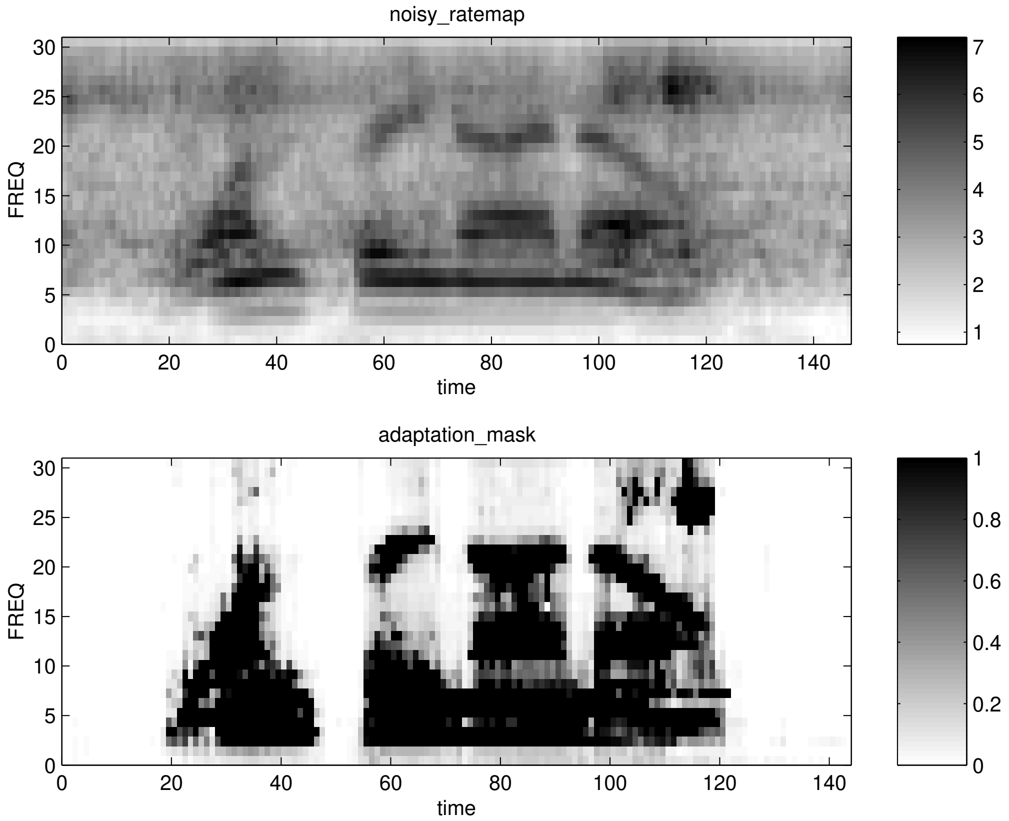


# Adaptive Noise Models

“1390” +

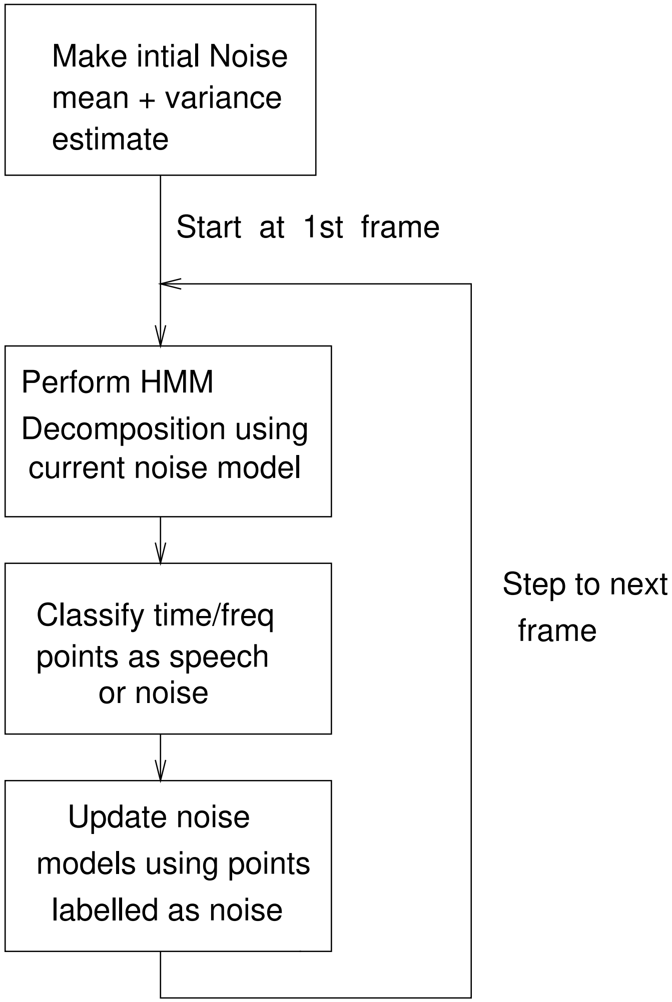
5dB Subway

P(Speech)




Adapt noise models using data in the low P(Speech) regions.

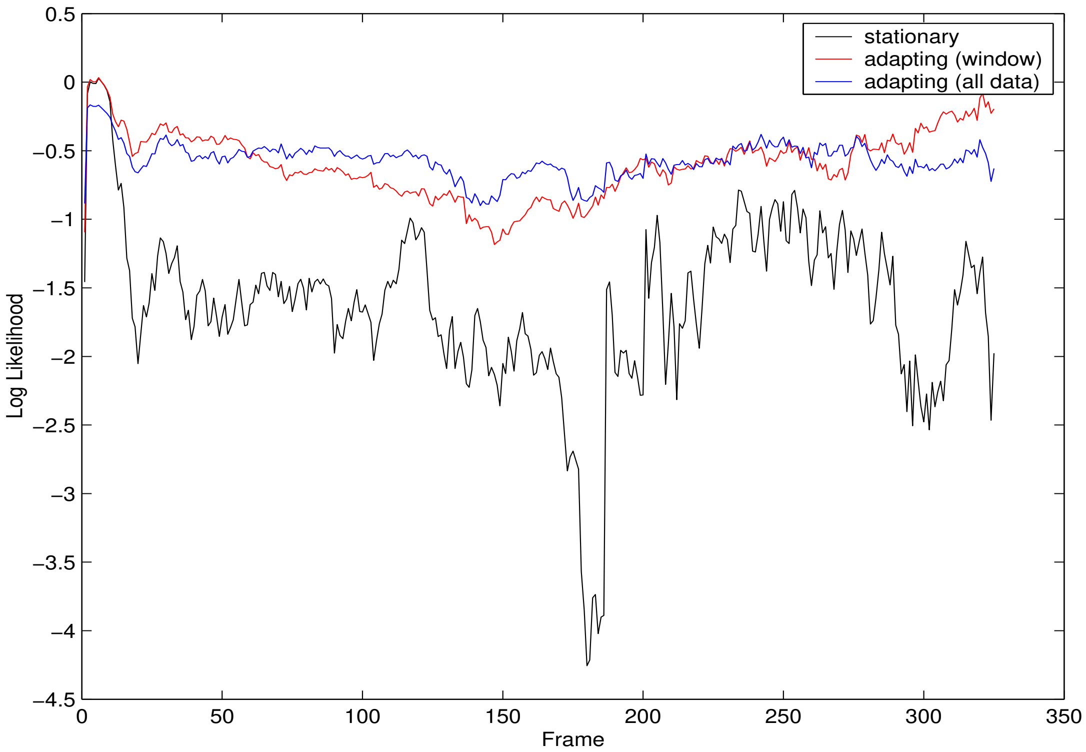
# Adaptive Noise Models




Noise mean and variance adaptation strategies:

- employ all points classified as noise,
- employ previous N points classified as noise.

# Comparison of Adaptive and Stationary Noise Models




| Model | Stationary | Windowed | All Data |
|---|---|---|---|
| Mean Log Lik | -1.62 | -0.60 | -0.57 |

# Adaptation Example

"1390" +

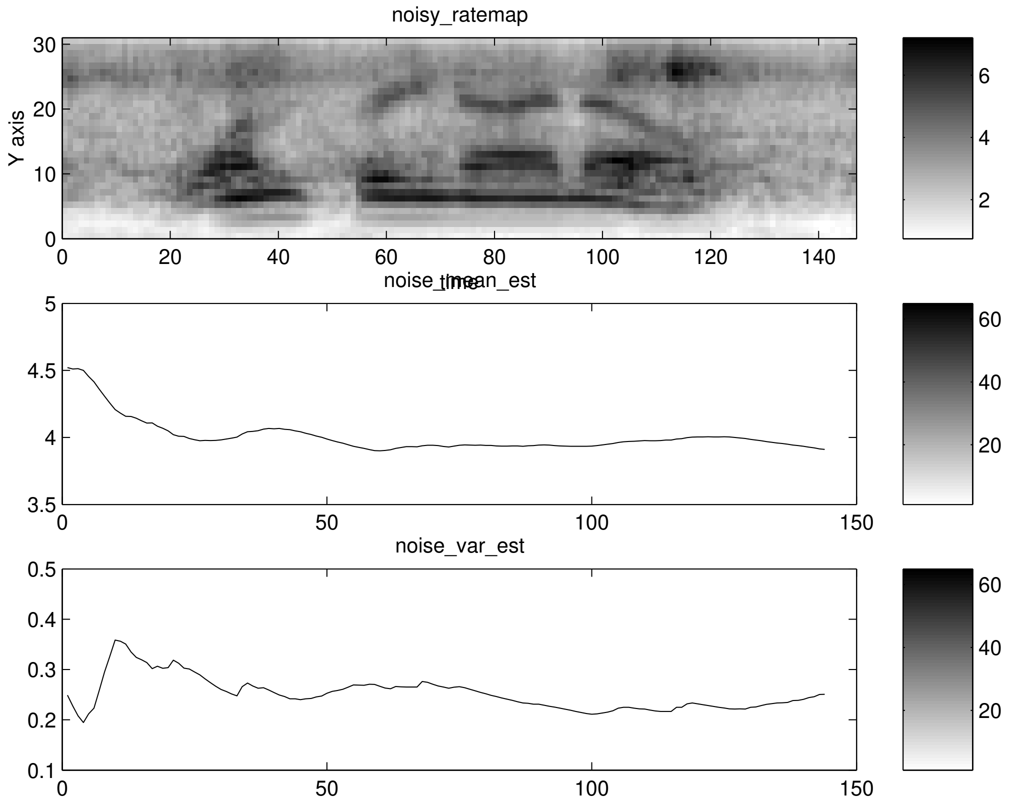
5dB Subway

Noise Mean

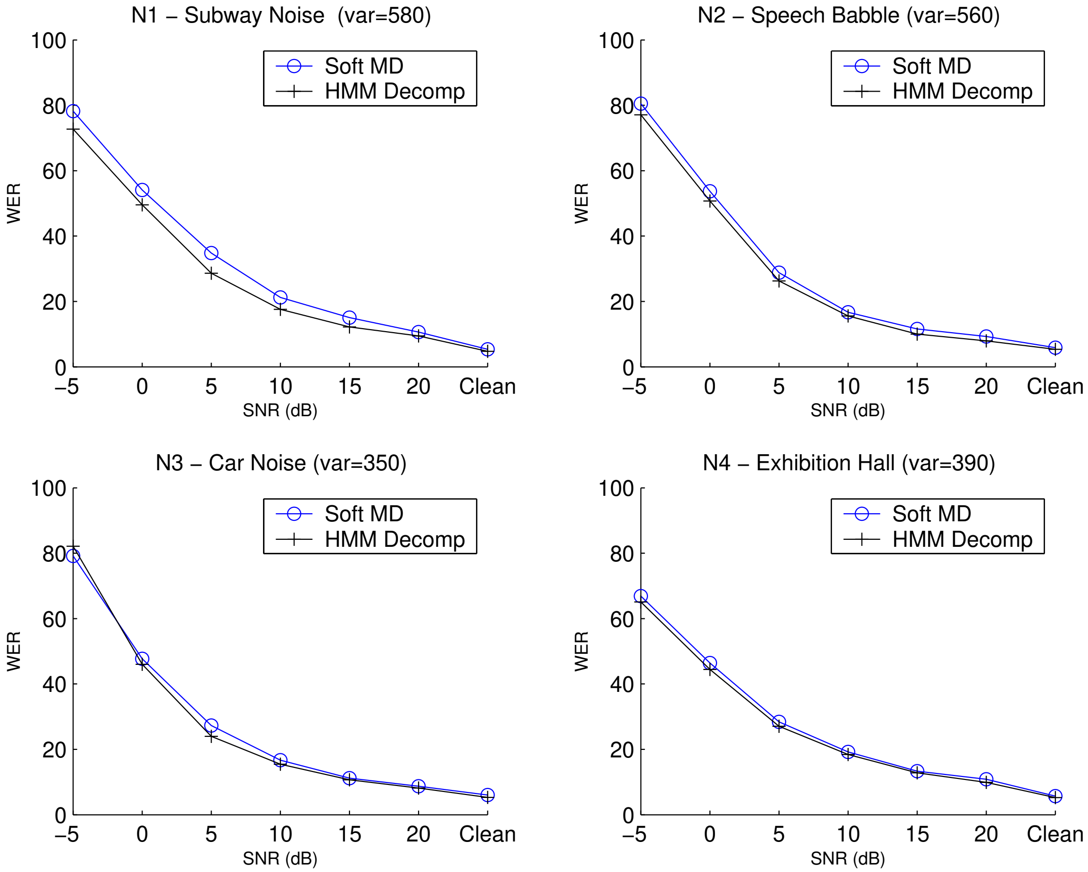
(channel 25)

Noise Var

(channel 25)



# Initial Results for AURORA 2




| | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|
| MD Soft | 49.5 | 70.2 | 81.5 | 87.2 | 87.2 |
| HMM Decomp | 52.3 | 73.5 | 83.2 | 88.5 | 91.1 |

## Employing Delta Features

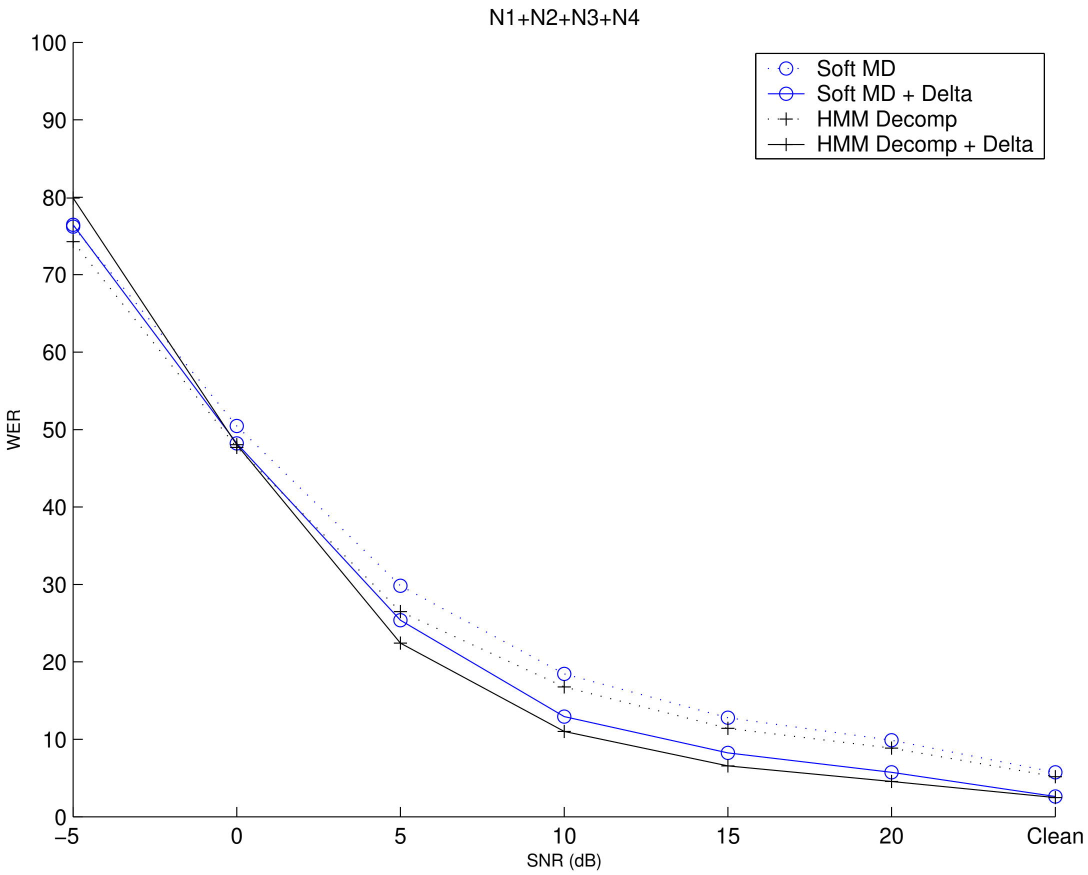
- Without delta features:

$$P(X) =$$
$$C(X, \mu_S, \sigma_S)N(X, \mu_N, \sigma_N) + C(X, \mu_N, \sigma_N)N(X, \mu_S, \sigma_S)$$

- With delta features:

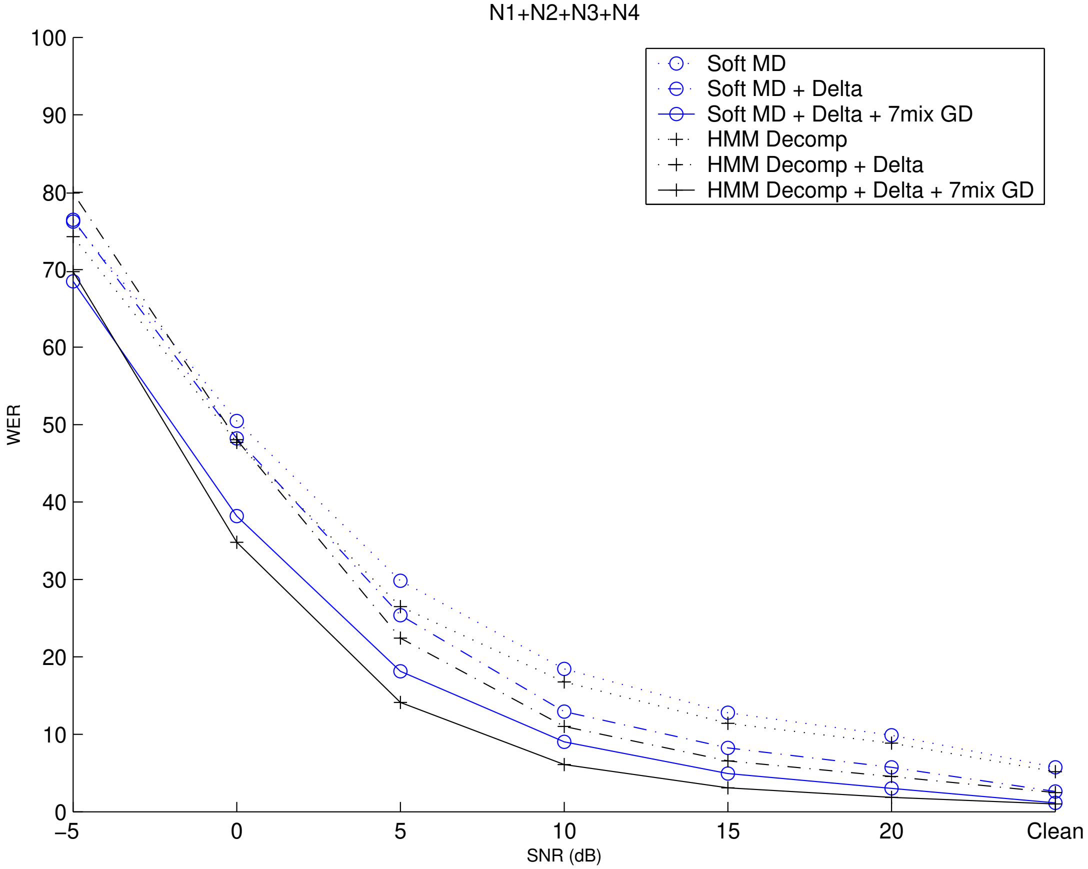
$$C(X, \mu_S, \sigma_S)(N(X, \mu_N, \sigma_N) + N(X', \mu_{N'}, \sigma_{N'}))$$
$$+ C(X, \mu_N, \sigma_N)(N(X, \mu_S, \sigma_S) + N(X', \mu_{S'}, \sigma_{S'}))$$

# Results with Delta Features (3mix)




| | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|
| MD Soft | 51.8 | 74.6 | 87.1 | 91.8 | 94.3 |
| HMM Decomp | 52.0 | 77.6 | 89.0 | 93.4 | 95.4 |

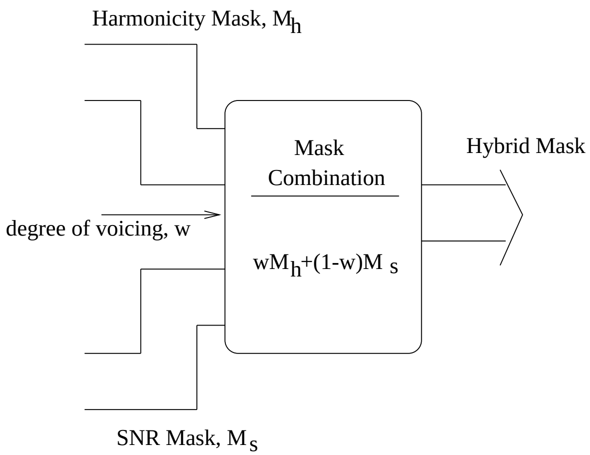
# Results with Delta Features (7mix GD)




| | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|
| MD Soft | 61.8 | 81.9 | 91.0 | 95.1 | 97.0 |
| HMM Decomp | 65.2 | 85.9 | 93.9 | 96.9 | 98.2 |

# Employing Harmonicity Information

- Soft MD

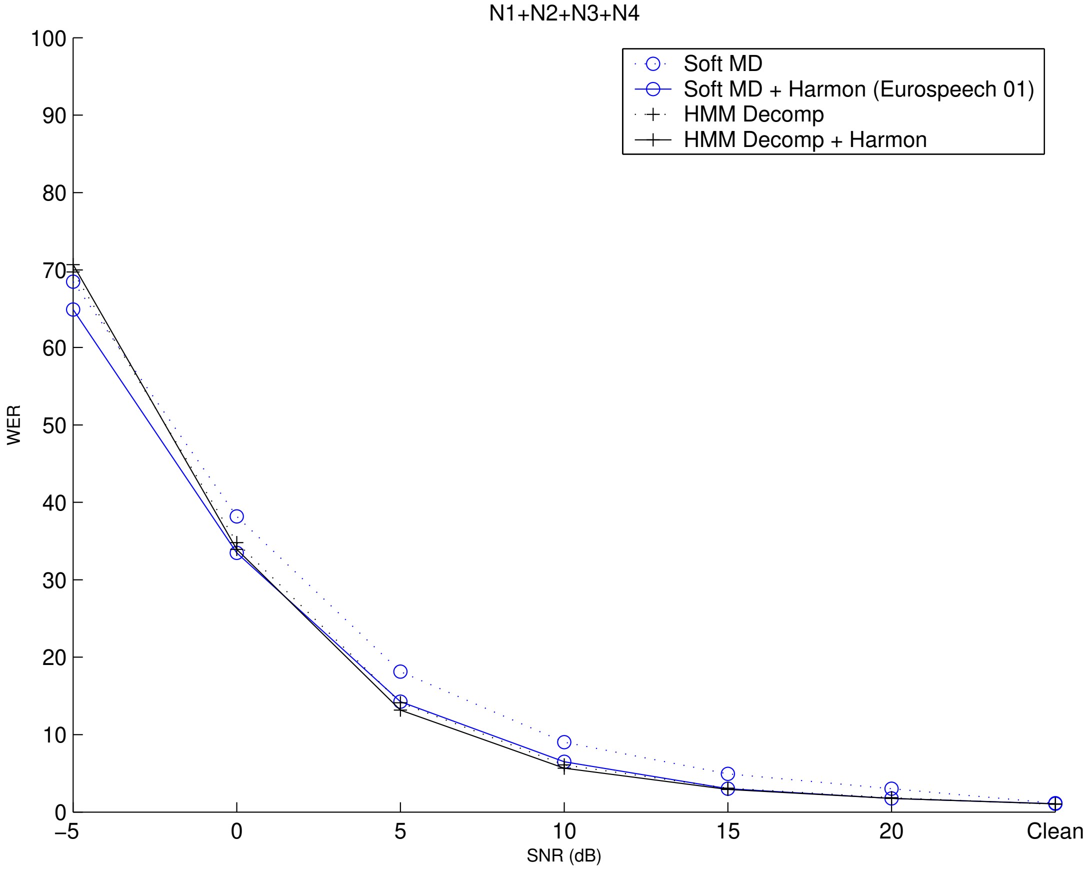



- HMM Decomposition

$$P_h(Speech) = wM_h + (1-w)0.5$$

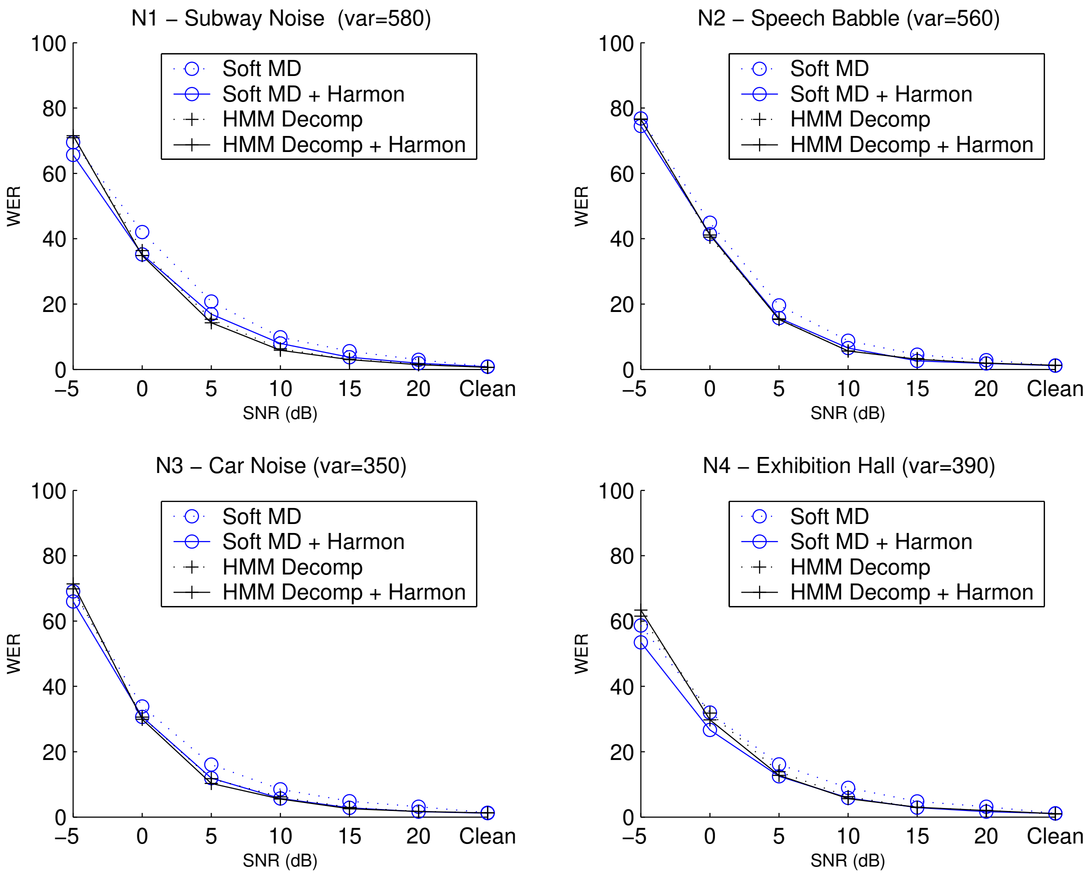
$$P(X) = C(X, \mu_S, \sigma_S)N(X, \mu_N, \sigma_N)(1 - P_h) + C(X, \mu_N, \sigma_N)N(X, \mu_S, \sigma_S)P_h$$

# Results for Full System




| | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|
| MD Soft | 66.5 | 85.8 | 93.5 | 97.0 | 98.2 |
| HMM Decomp | 66.1 | 86.8 | 94.3 | 97.1 | 98.2 |

# Results for Full System




| | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|
| MD Soft | 66.5 | 85.8 | 93.5 | 97.0 | 98.2 |
| HMM Decomp | 66.1 | 86.8 | 94.3 | 97.1 | 98.2 |

## Model Decomposition vs Soft Missing Data

- **Missing Data with Soft Masks**

  P($x$)= $w$ P(Speech=$x$) + $\frac{1-w}{x}$ P(Speech<$x$)

  Where $w$ = P(Observation is dominated by Speech)

- **Speech/Noise Model Decomposition**

  P($x$)= P(Noise<$x$) P(Speech=$x$) + P(Noise=$x$) P(Speech<$x$)

where $x$ is the observed data.