Improved Soft Decisions in Missing Data ASR: Using Harmonicity in Conjunction with Local SNR Estimates

Speech and Hearing Research Group,

Dept. Computer Science,

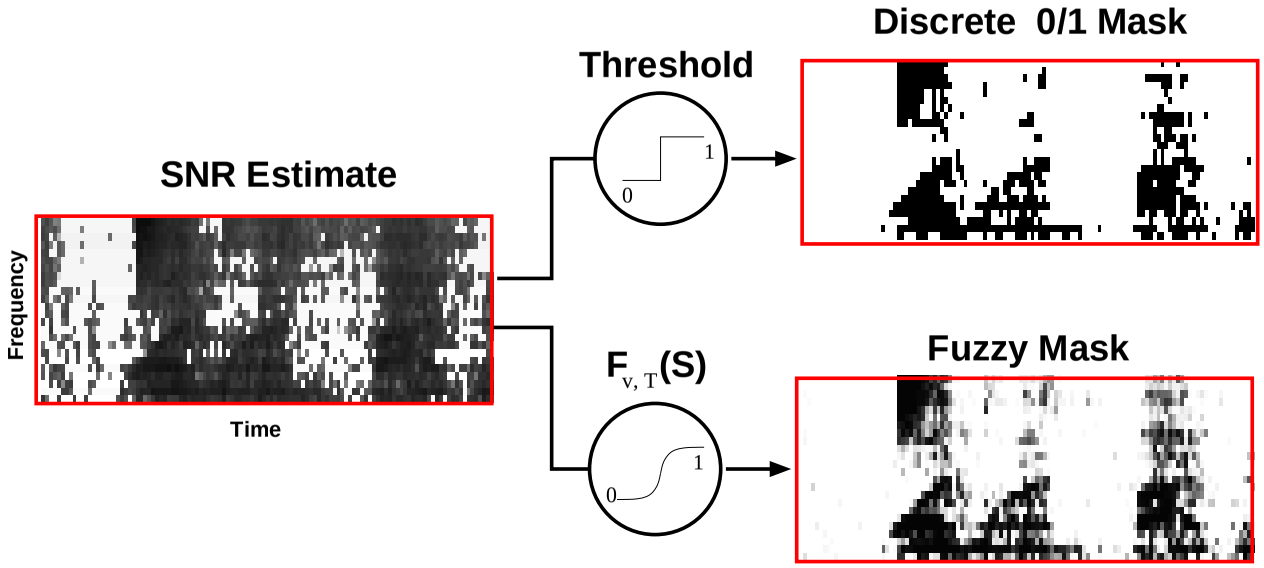
University of Sheffield, UK

January 24, 2001

Improved Soft Decisions in Missing Data ASR: Combining Masks

- Soft Decisions in Missing Data
- Harmonicity-based Fuzzy Masks
- Merging Local SNR and Harmonicity Masks
- Aurora 2000 Results
- Conclusions

Soft Decisions in Missing Data



Soft mask values are interpreted as "the probability that the data is reliable".

So rather than use the *present data likelihood* **OR** *the missing data 'induction constraint'*, every point uses weighted sum of **BOTH** terms.

Using Soft Decisions

Missing data probability calculation for discrete masks, showing the separate present and missing components:

$$\overline{f(x|S)} = \prod_{i \in p} f_i(x_i|S) \prod_{j \in m} \frac{1}{x_j} \int_0^{x_j} f_j(x_j|S) dx_j$$

With **soft decisions** the probability due to each feature vector component becomes a weighted sum of the present and missing probability terms:

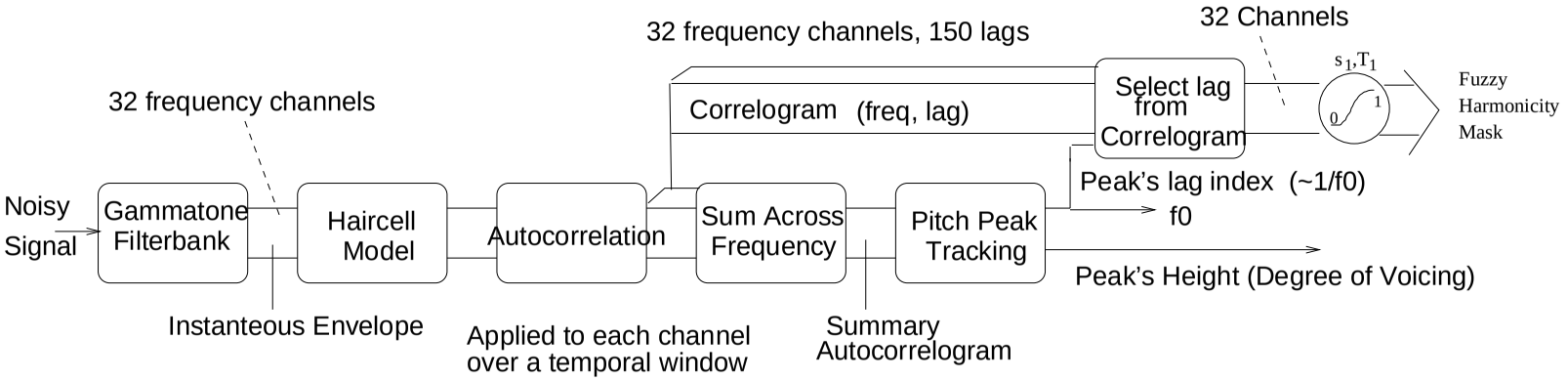
$$\overline{f(x|S)} = \prod_{i=1}^{N} \left(w_i f_i(x_i|k) + (1 - w_i) \frac{1}{x_i} \int_0^{x_i} f_i(x_i|k) dx_i \right)$$

Using Soft Decisions

Generalising to models employing Gaussian mixtures:

$$\sum_{k=1}^{M} P(k|S)\Big(\prod_{i=1}^{N} (w_i f_i(x_i|k,S) + (1-w_i)\frac{1}{x_i}\int_0^{x_i} f_j(x_i|k,S)dx_i)\Big)$$

Harmonicity Masks



- The Harmonicity Mask is designed to mark **voiced speech** regions.
- It works well when noise is inharmonic or the SNR is favourable.
- Refinements necessary when noise is harmonic and dominant:
  –> pitch tracking, multisource decoding?

Mask Combination

We now have two fuzzy masks:

- **Fuzzy SNR-based mask -** Works well in stationary noise.
- **Fuzzy Harmonicity-based mask** - Highlights voiced speech regions.

We also have a ‘degree of voicing’ parameter, V.

How do we combine the masks?

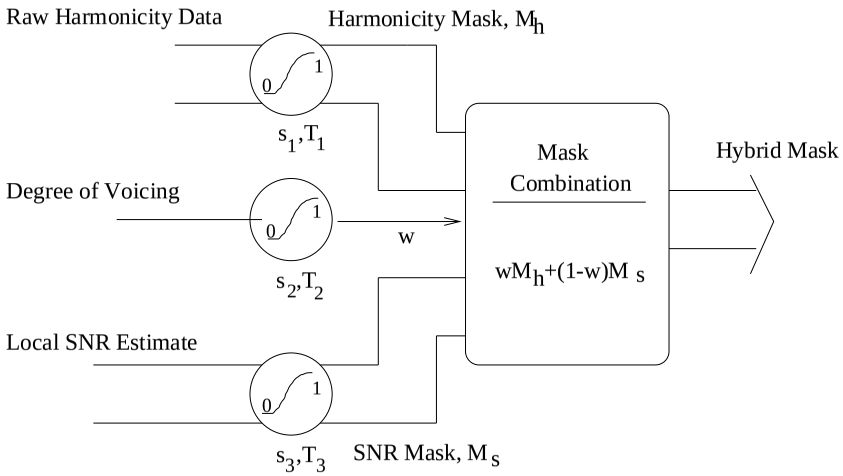
Mask Combination

Discrete combination: (One parameter)

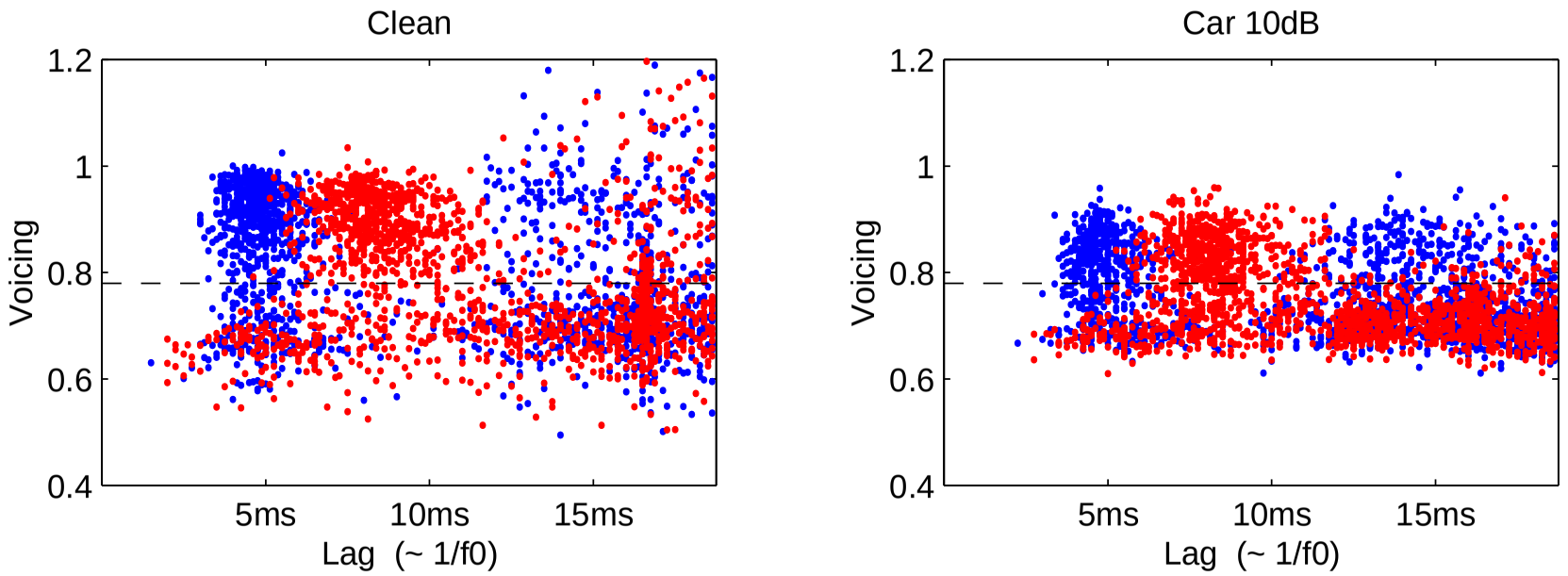
If $V > V_0$ frame is Voiced, else frame is Unvoiced. Then,

- Voiced frames –> Use harmonicity-based mask
- Unvoiced frames –> Fall back on SNR masks

Fuzzy combination: (Two parameters)



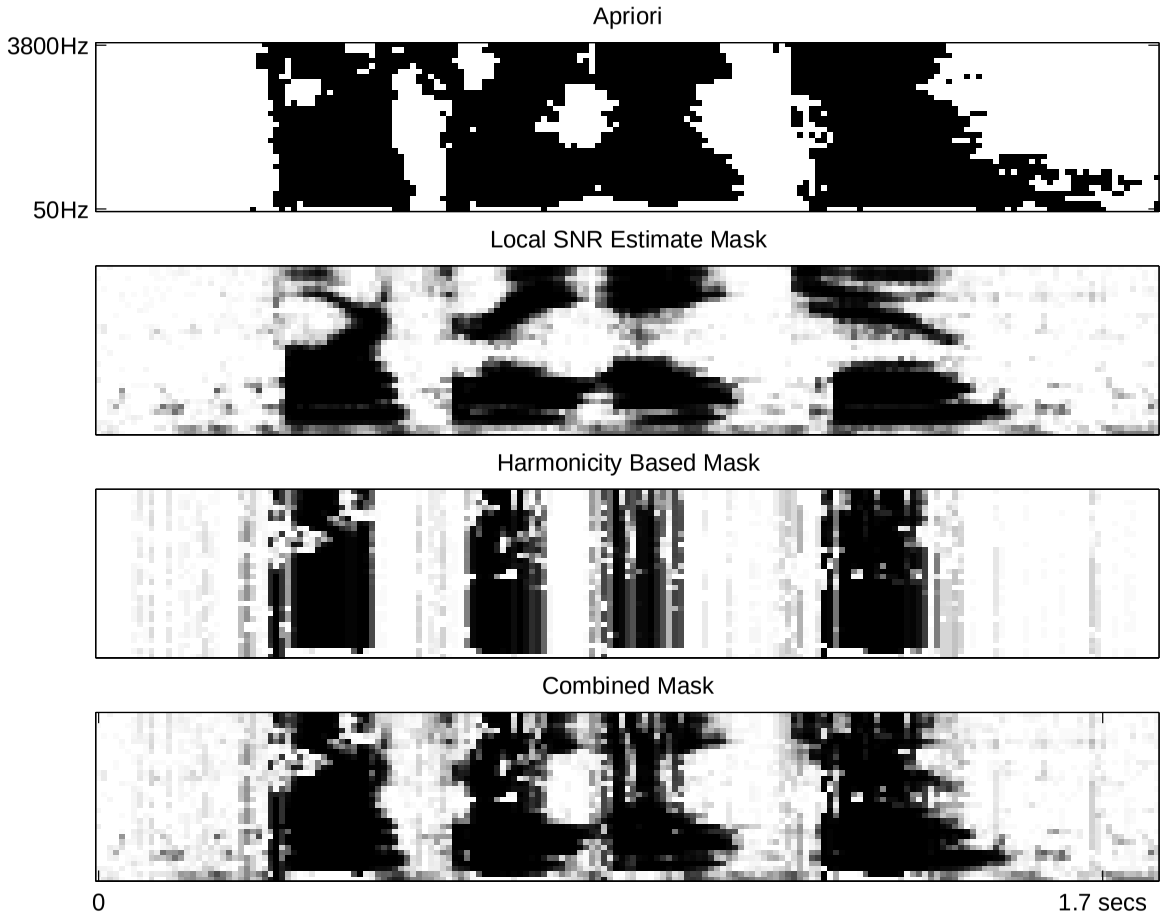
Tuning the Voicing Sigmoid



Voicing vs. Lag for female and male speakers.

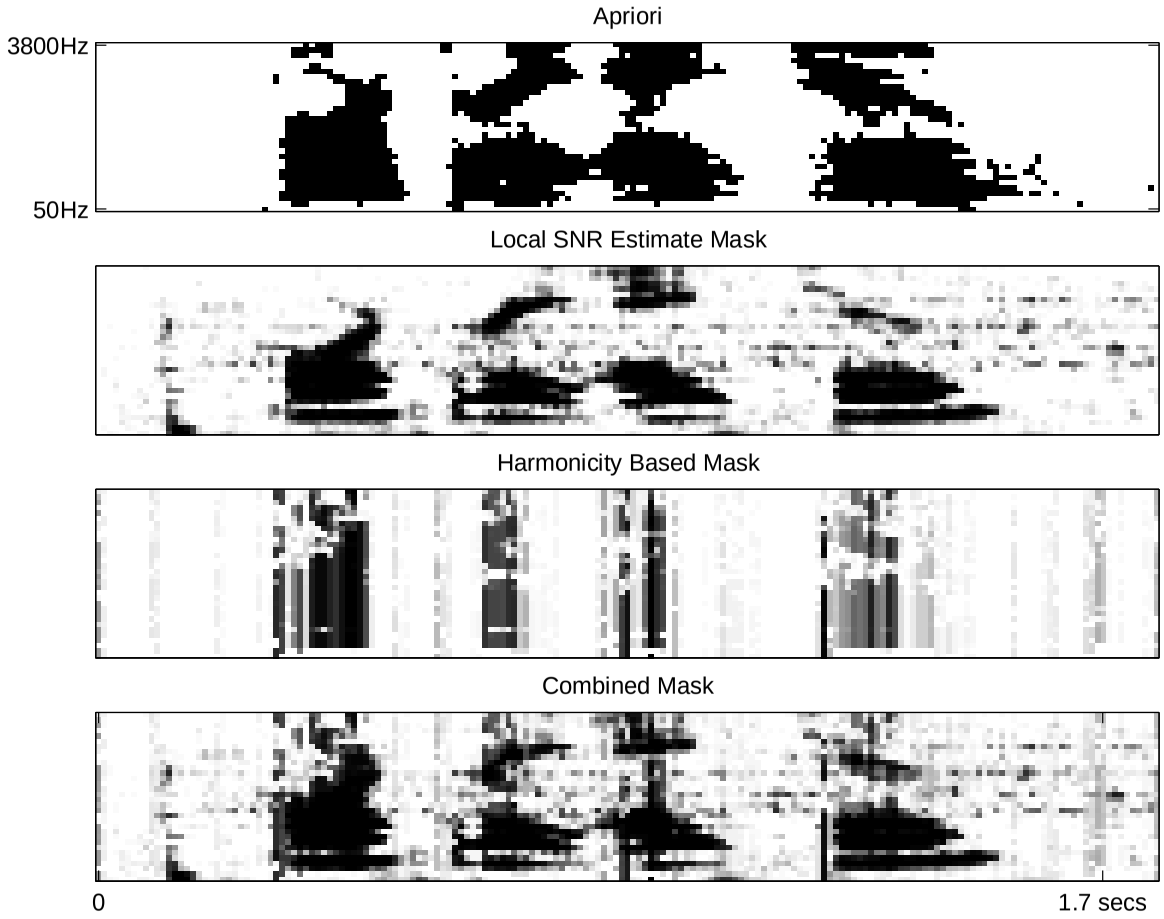
Comparison with Apriori Masks

Male "4382" + Car @ 20dB SNR



Comparison with Apriori Masks

Male "4382" + Car @ 10dB SNR



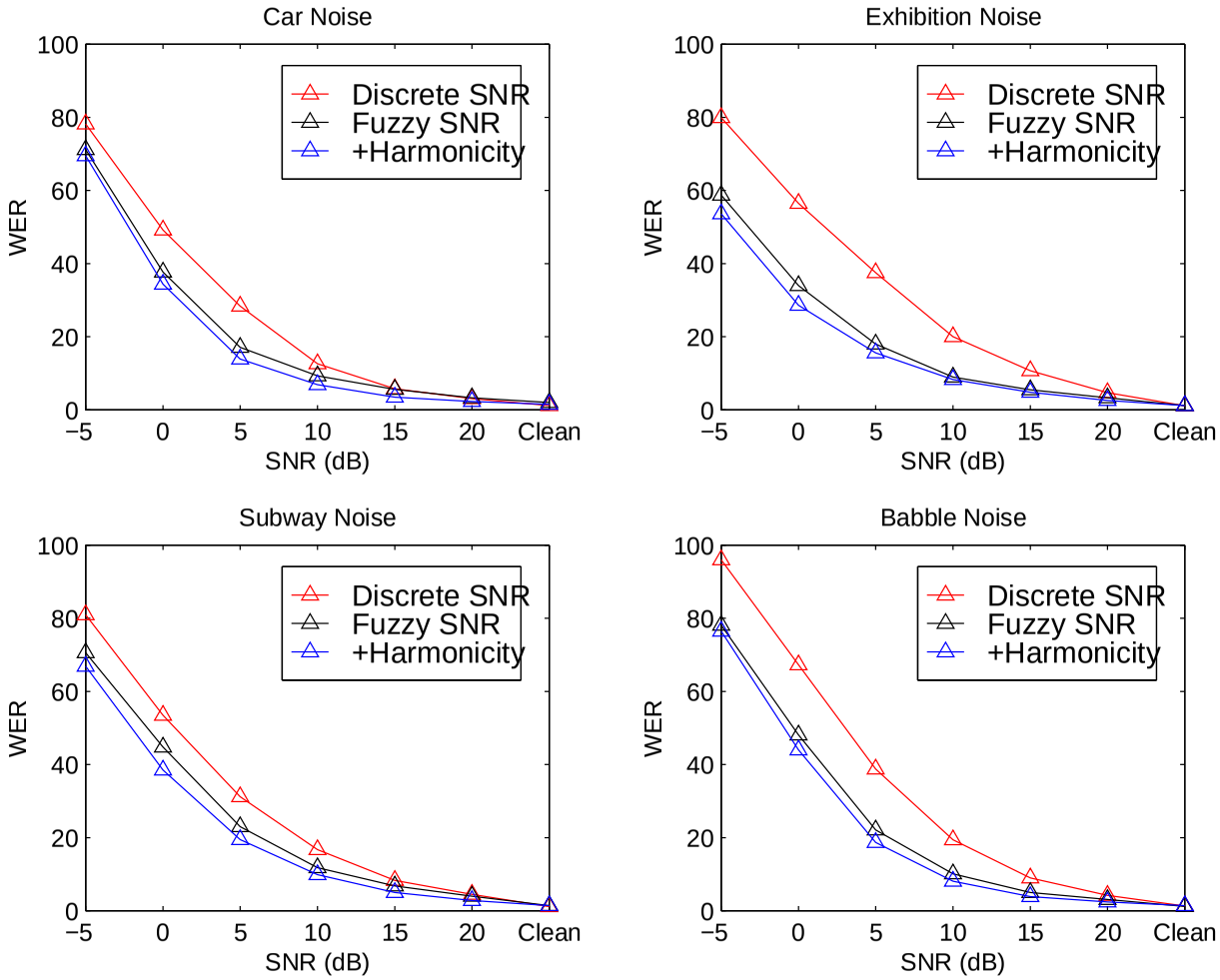
Aurora 2000 Experiments

- Trained on **clean data**.
- Testing using Set A

  (i.e. subway, exhibition, babble and car noises).
- Features: 32 channel gammatone filter bank, + deltas.
- Two slightly different sets of models
  - + Aurora Models: 16 states per digit,
  - + DC Models: 11.5 states per digit on average.
- 7 mixtures per state

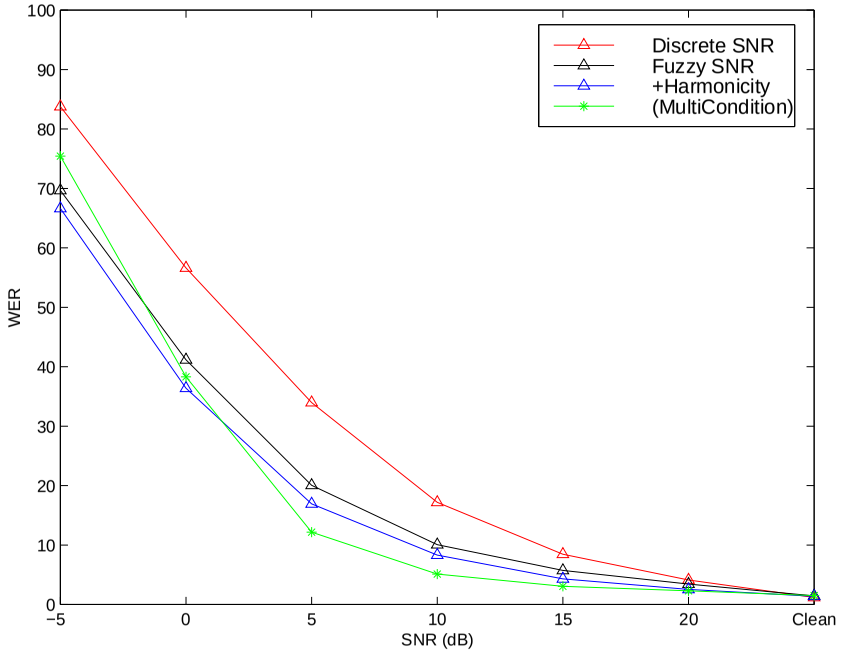
  (note, relatively large num. of mixes needed for spectral features).

Aurora Results: Test Set A



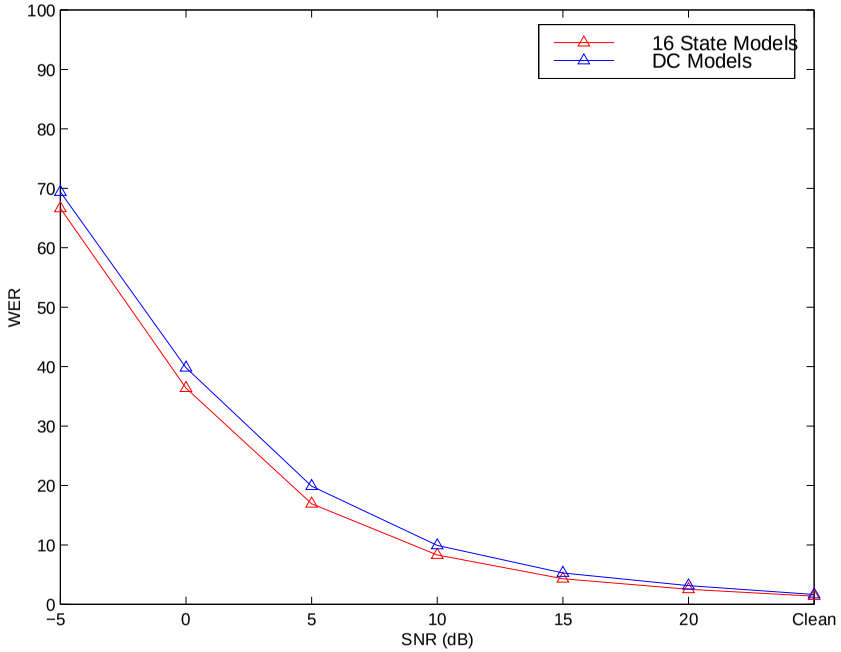
(32 channel filter bank + deltas)

Aurora Results: WER averaged over noise condition



MASK / SNR	-5dB	0dB	5dB	10dB	15dB	20dB	Clean
Discrete SNR	83.8	56.6	34.0	17.2	8.5	4.1	1.2
Fuzzy SNR	69.7	41.2	20.1	10.1	5.7	3.4	1.5
+ Harmonicity	**66.6**	**36.4**	**16.9**	**8.3**	**4.3**	**2.5**	**1.4**

Aurora WER Results: Aurora vs. DC Word Models



Models / SNR	-5dB	0dB	5dB	10dB	15dB	20dB	Clean
16 State Models	66.6	36.4	16.9	8.3	4.3	2.5	1.4
DC Word Models	69.4	39.8	19.9	9.9	5.3	3.2	1.7

Conclusions

- **In combination, Harmonicity and Local SNR masks perform better than either mask individually**, i.e:
  - + better approximation to the apriori ('cheating') mask,
  - + better recognition results.
- **The mask generation parameters are robust**,
  i.e. one set of parameters will perform well over a large range of noise types, and noise levels.
- **Sensible values can be estimated from clean speech.**

Further Work

- **Temporal Smoothing.**

  Smoothing the masks appears to improve results for some noise types - but seriously damages results for others.

- **Using F0 Information.**

  Using F0 to distinguish between voiced speech and harmonic noise. F0 tracking. ‘Multi-pitch’ decoding.

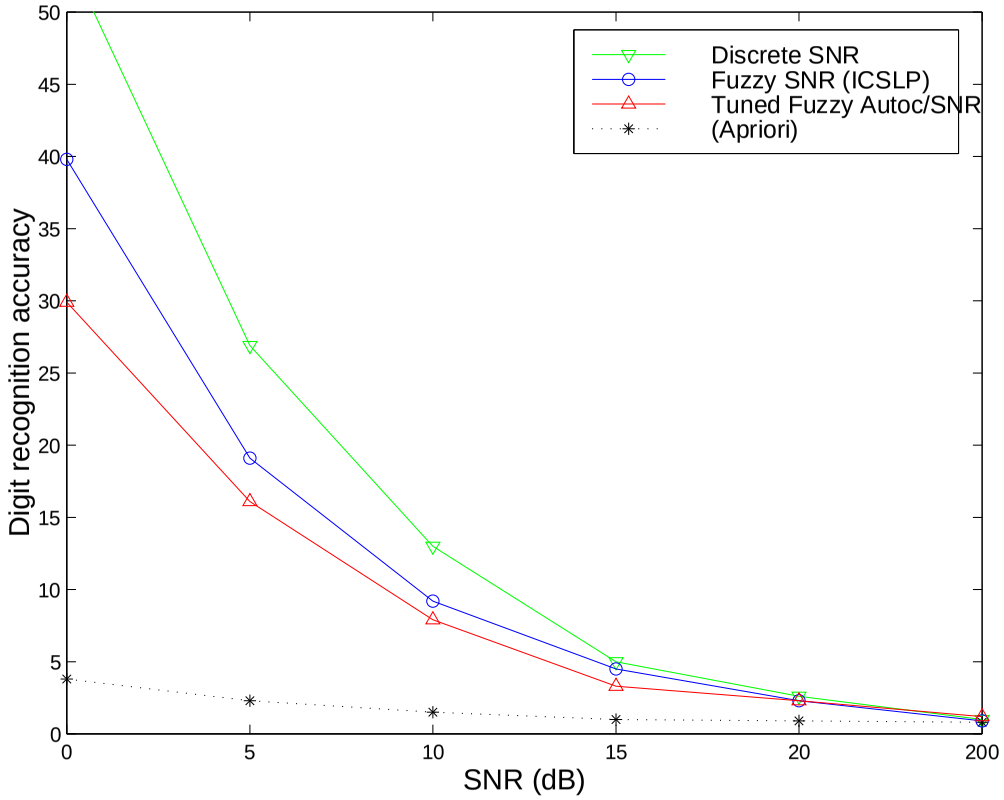
- **Adaptive Sigmoid Parameters.**

  Techniques for fine tuning the mask generation parameters according to the noise estimate.

- **More General Mask Combination Techniques.**

Learning Noise Specific Parameters

20 KHz TIDigits + Factory Noise



Parameters tuned to minimise distance to Apriori masks at 0 & 5 db.