morris, mccowan, bourlard@idiap.ch
http://www.idiap.ch/

Microphone Arrays for Missing Data Mask Estimation: Two Ears is Enough

Iain McCowan, Andrew C. Morris

APPROACH / RESULTS / CONCLUSION

and

Low Cost Duration Modelling

Andrew C. Morris, Simon M. Payne

APPROACH / RESULTS / CONCLUSION

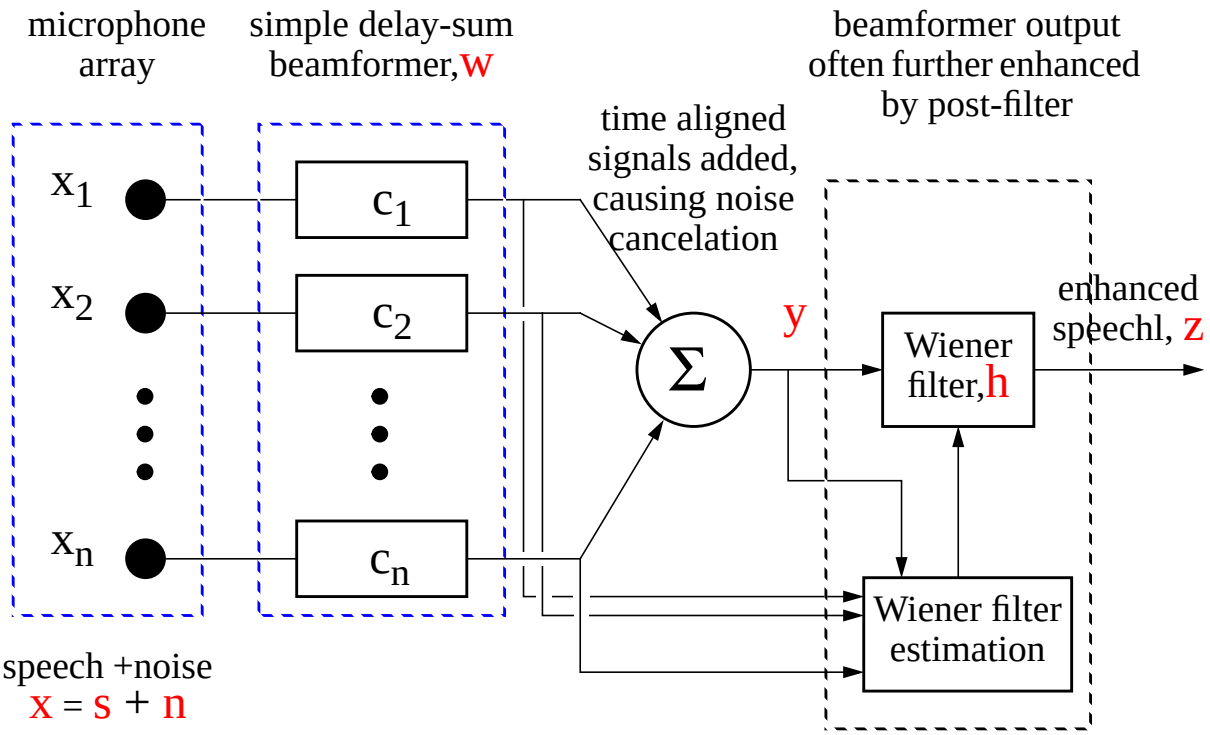
Microphone Arrays for MD Mask Estimation

Iain McCowan, Andrew C. Morris

MAs normally used for speech enhancement

Beamforming by MAs reduces level of undesired noise, permiting distant, hands-free signal acquisition.

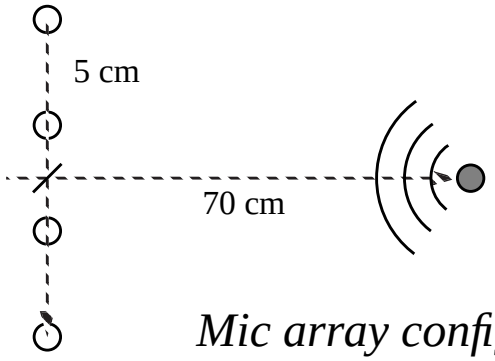
$y(f) = w^T(f)x(f)$ $\qquad$ $z(f) = h(f)y(f)$



Filter-sum beamformer with post-filter

Problem: High perormance ASR via MA based speech enhancement requires large numbers of microphones.

Mic arrays for MD mask estimation



Mic array configuration tested

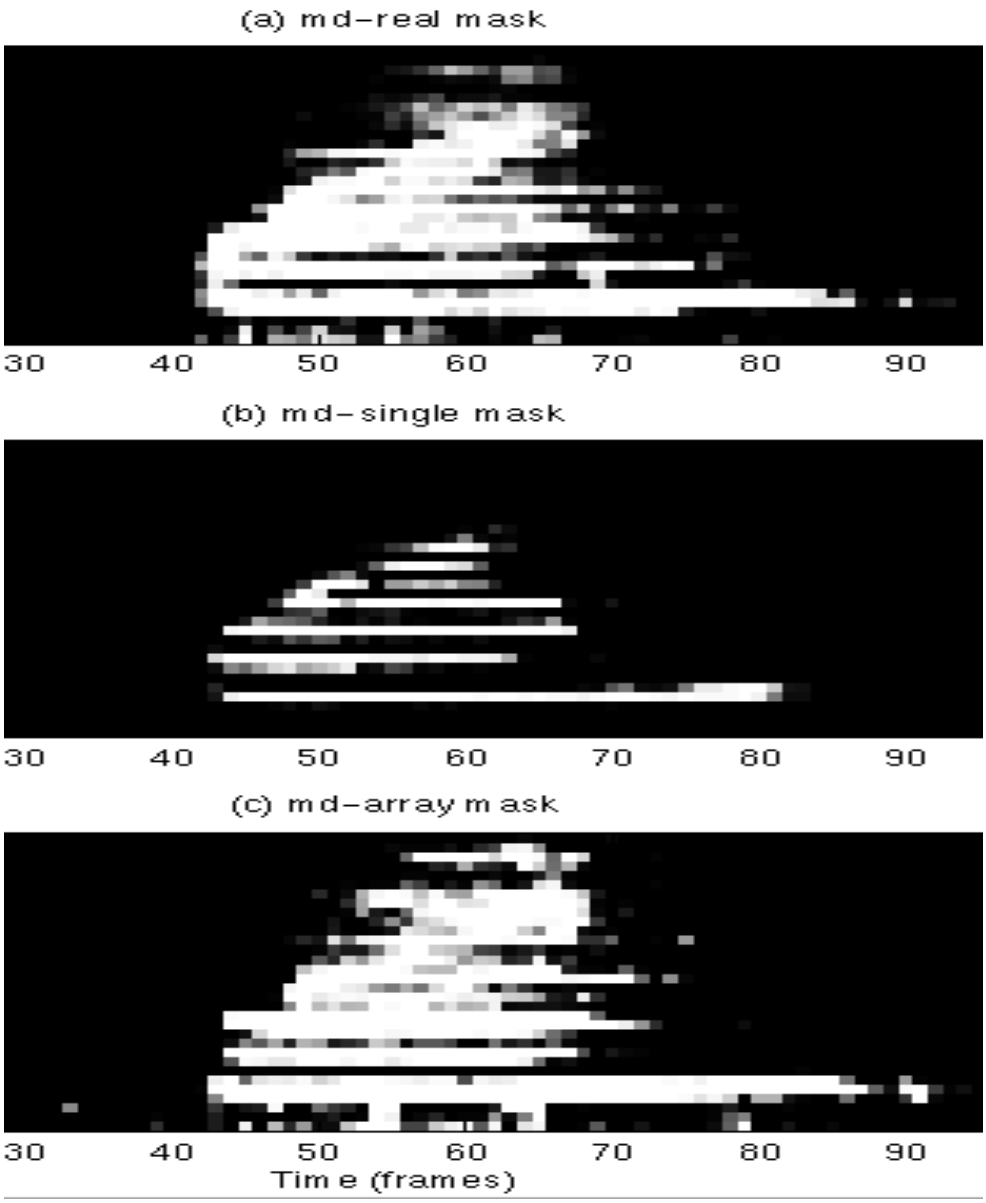
Noise estimation is by-product of speech enhancement. Estimate SNR and mask val. for frame *(k)* coeff *(i)* as follows.

Let s, n, x, z *= clean signal, noise, noisy sig, enhanced sig.*

Let $x(f)$ *=> energy domain,* $x(k)$ *=> log energy domain.*

- $\hat{s}(f) = z(f) = h(f)w^{T}(f)x(f)$
- $\hat{n}(f) = x(f) - \hat{s}(f)$
- $\hat{snr}_i(k) = \hat{s}_i(k) - \hat{n}_i(k)$
- Hard MD mask $r_i(k) = 1$ if $\hat{snr}_i(k) > \beta$, else $= 0$
- Soft MD mask $r_i(k) = 1 / \left(1 + e^{-\alpha(\hat{snr}_i(k) - \beta)}\right)$

MA MD masks better than simple MD masks

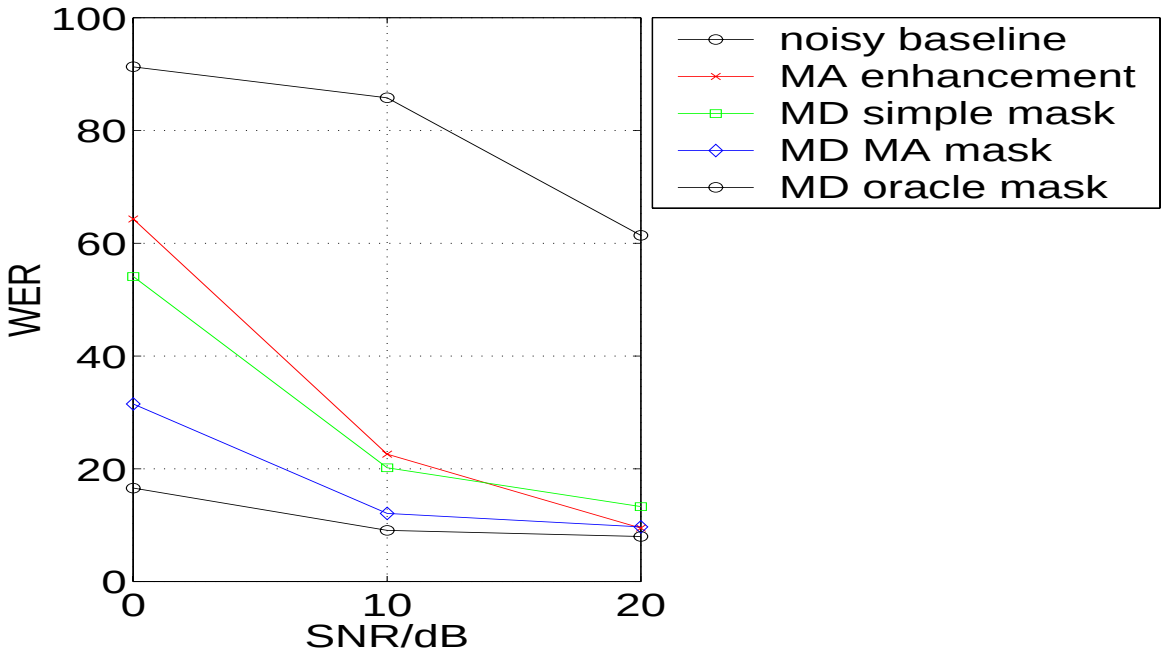


Top Fig shows oracle mask, utterance “one”, (a = 1, β = 0)

Mid Fig shows “noise = first 10 frames” mask

Bot Fig shows 4-mic array mask

Reco with MA MD masks better than with simple MD masks or with MA enhanced speech



Experiment: Training data = Aurora clean training set. Test data = Aurora clean test set 1a. Noise = artificially added office noise plus convolution with room response. 4 mic array.

Figure shows that:

- MA for speech enhancement has comparable performance to MD using simple masks (SMD, with bounds constraint).
- MD with MA masks gives a further 40% rel WER reduction at snr 10 and 0 dB.

Discussion and Conclusion

Experiments were repeated with 2-mic array. Performance fell significantly with enhanced speech, but not with MD.

Tests were proof of concept only. Advantage of MAs in MD ASR should increase further with highly non stationary noise.

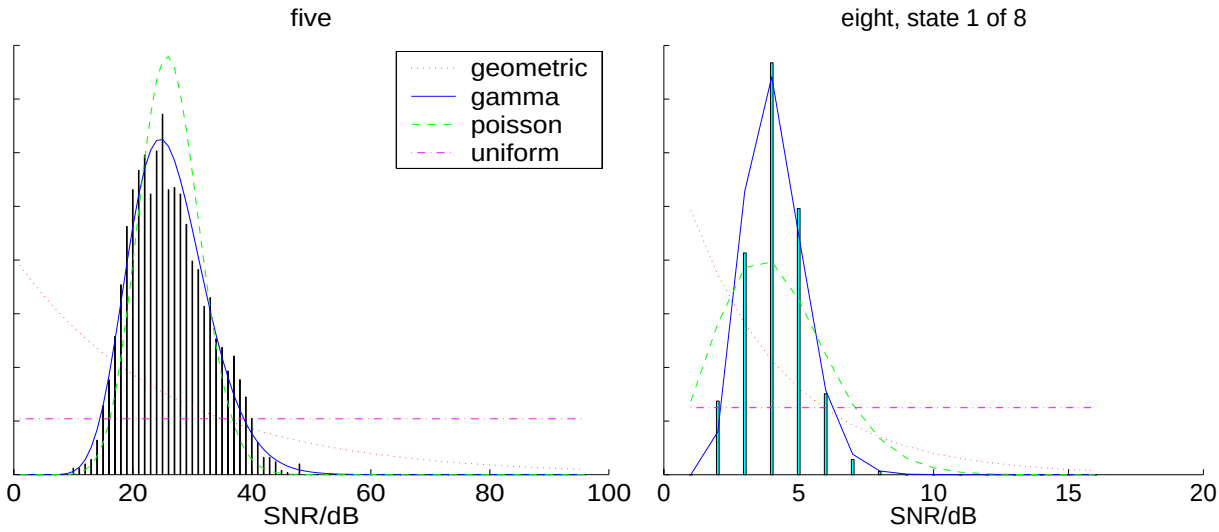
Further tests should include:

- more advanced beamformers (than delay-sum)
- adaptive beamforming (speaker position not fixed)
- different noise types (including highly non stationary)
- different array configurations

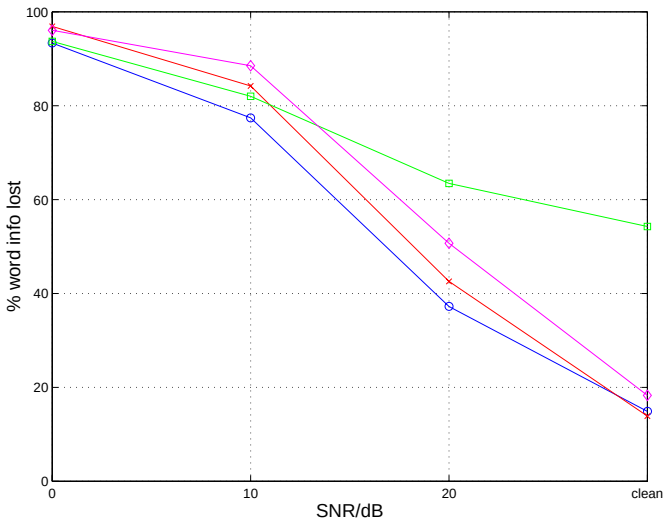
Low Cost Duration Modelling

Andrew C. Morris, Simon M. Payne

Duration Histogram Smoothing



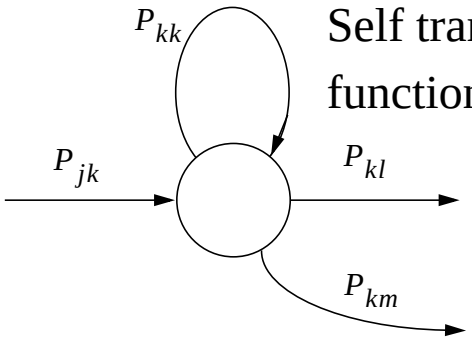
Duration histograms for typical word and state, + fitted pdfs



% info loss performance of different parametric pdfs, noise = subway, snr20

i.e. Gamma pdfs work best

Transition Probability Calculation



Self transition probability now function of state duration

Other transition probabilities from transition matrix scaled to sum to 1

usual Markovian assumption

$$P(q_t | q_1, q_2, \ldots q_{t-1}) \approx P(q_t | q_{t-1})$$

becomes

$$P(q_t | q_1, q_2, \ldots q_{t-1}) \approx P(q_t | q_{t-1}, d_{t-1})$$

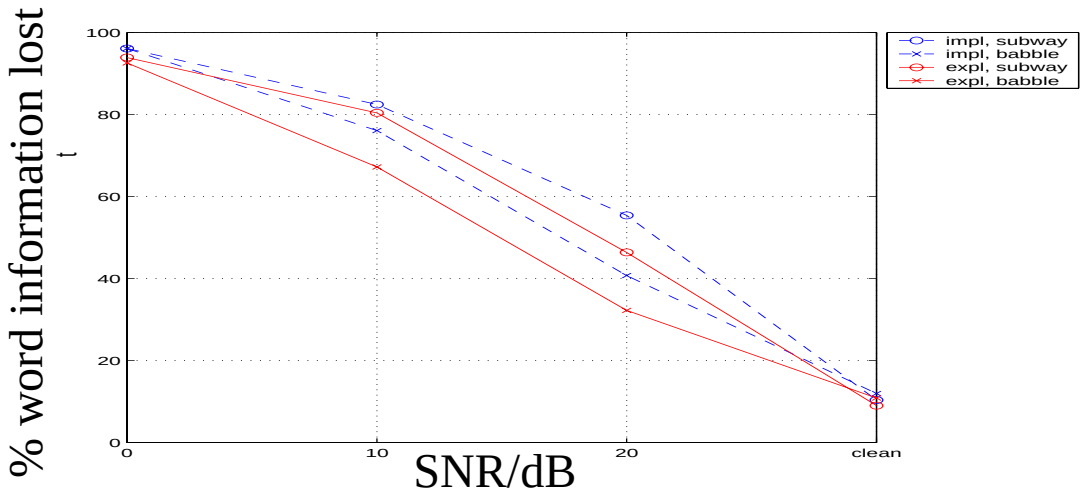
Fixed self transition probabilities become

$$P(notrans(d)) = (fd > d | fd \geq d)$$

$$= \frac{P(fd > d+1)}{P(fd \geq d)}$$

$$P(fd > d) = P(fd > d \wedge fd \geq d) = P(fd > d | fd \geq d) P(fd \geq d)$$

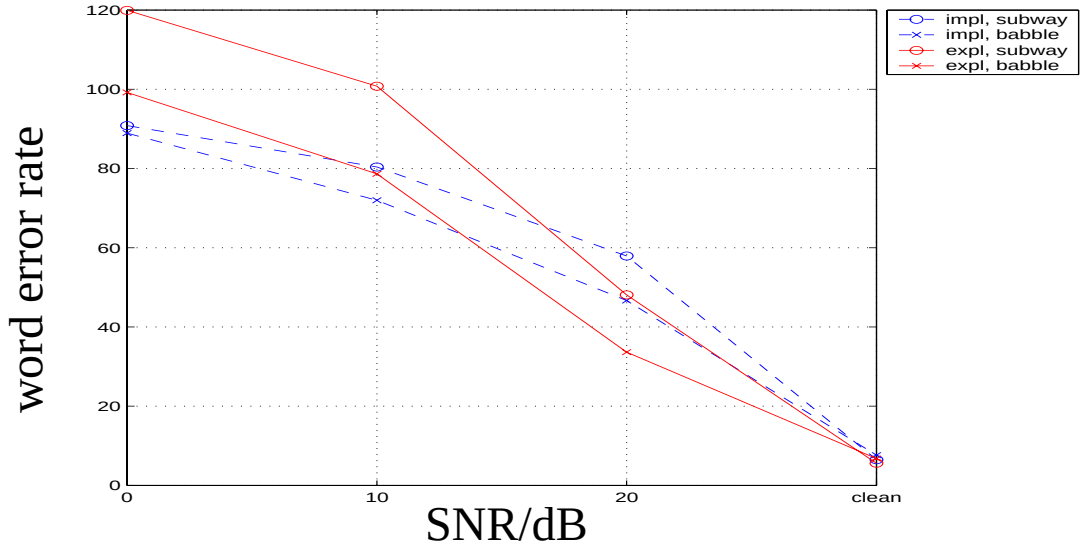
Explicit vs Implicit Models Performance



$$WIP = 1 - H/(H+S+D)(H+S+I)$$

Top Fig. compares reco performance using implicit (blue) and explicit (red) duration models, on subway (‘o’) and babble (‘x’) noise, at SNR clean to 0 dB.

Bot Fig shows same, but using WER instead of WIP



$$WER = (S+D+I)/(H+S+D)$$