morris@idiap.ch
http://www.idiap.ch/

Missing-data masks in all-combinations multi-band decoding

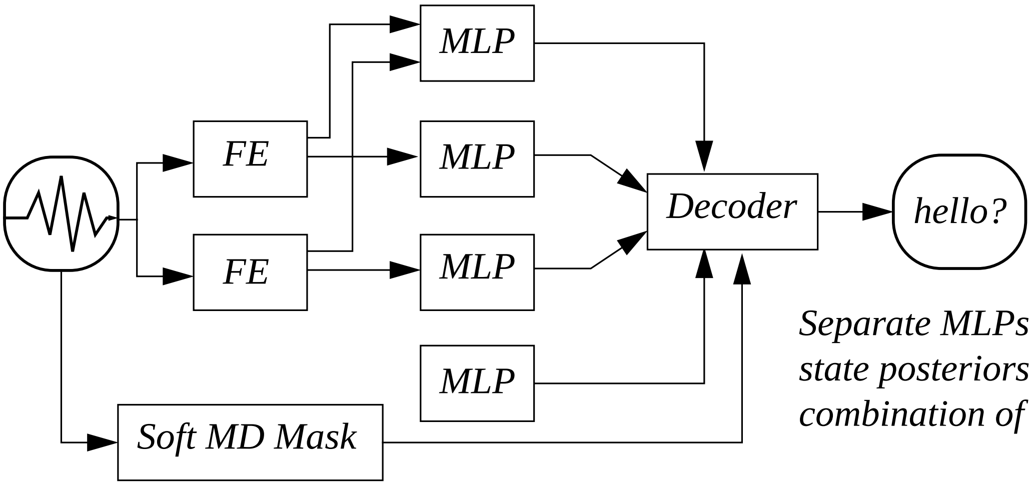
It is shown that for MAP decoding with all-comb experts

- multi-band expert weighting can make use of **same** soft missing-data mask as used with “missing-data” ASR
- experts must be combined **during**, not before, decoding

MAP decoder architectures

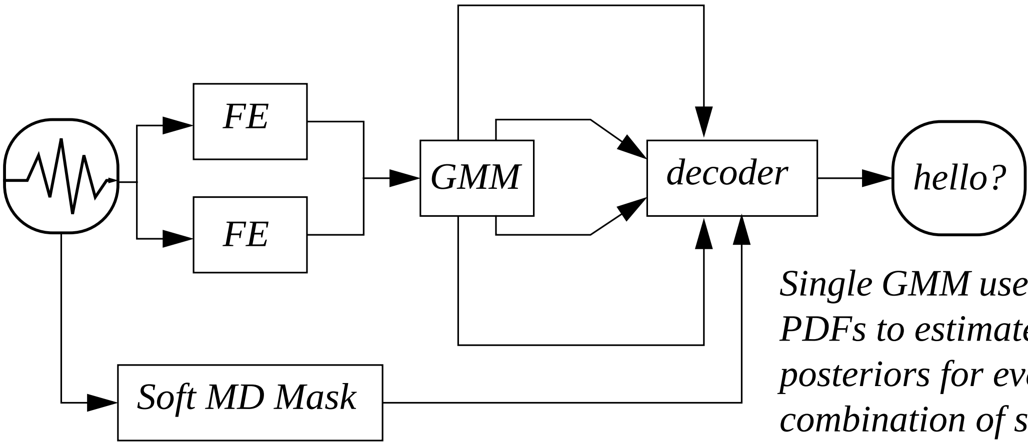
MAP decoding => experts combined during Viterbi

All-combination multi-band SMD HMM/MLP



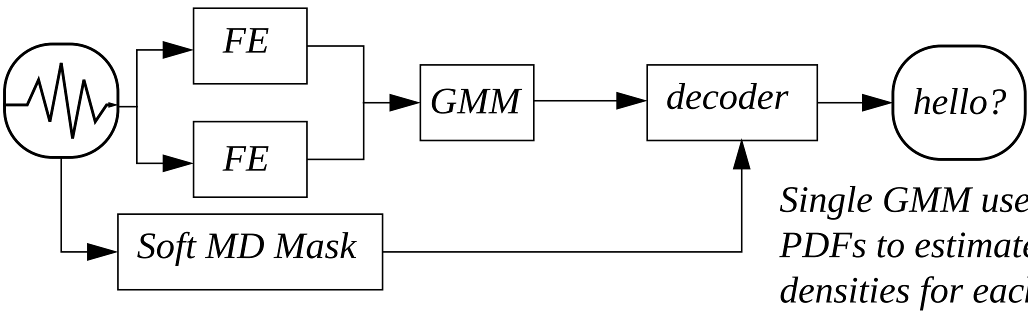
Separate MLPs estimate state posteriors for every combination of sub-bands

All-combination multi-band SMD HMM/GMM



Single GMM uses marginal PDFs to estimate state posteriors for every combination of sub-bands

Usual missing-data ASR = SMD HMM/GMM



Single GMM uses marginal PDFs to estimate state densities for each data coefficient

All-combinations posteriors based decoder can make use of same mask as usual missing-data decoder

Notation

Q state sequence for one utterance

X spectrotemporal signal for one utterance

$\hat{\Omega}$ estimated SMD mask, $\omega_{f,t} = \hat{P}(x_{f,t} clean)$

M MD indicator mask, $(\mu_{g,t} = 1) \Leftrightarrow band_{g,t}$ clean

P_c $P(Xclean)$

Usual missing-data (GMM) MAP objective

$$\hat{Q} = \arg max_Q E[P(Q|X,\Theta)]$$

$$E[P(Q|X,\Theta)] \propto P(Q)\int p(X|Q)p(X|X_{obs})|dX$$

$$p(X|X_{obs})| = P_c\delta_{(X-X_{obs})} + (1-P_c)U(0, X_{obs})$$

Posteriors based (MLP) MAP objective previously tested

$$\hat{Q} = \arg max_{Q,M} P(Q|X,M) \qquad \text{assumes } \Omega = 0.5$$

Posteriors based (MLP) MAP objective using MD mask

$$\hat{Q} = \arg max_{Q,M} P(Q,M|X) \qquad \text{makes use of } \hat{\Omega}$$

$$= \arg max_{Q,M} P(M|X)P(Q|X,M)$$

$$P(M|X) \cong \prod_{g,t \in M} \omega_{g,t} \prod_{g,t \notin M} (1-\omega_{g,t})$$

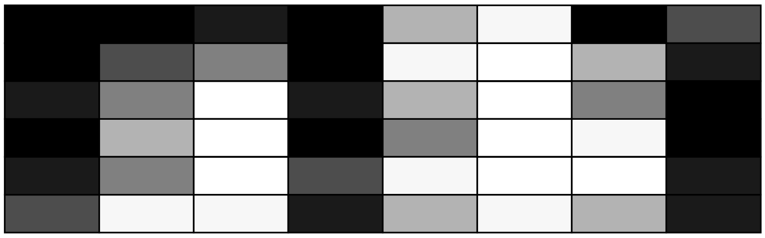
During Viterbi, each frame selects single expert

Soft missing data mask for posteriors based decoder

Per coefficient soft mask, P(x coeff(f,t) missing)

$$\hat{\Omega}_{coeffs},\ \omega_{f,t} = \hat{P}(x_{f,t} clean)$$

$f = 1\ldots6$



$t = 1\ldots8$

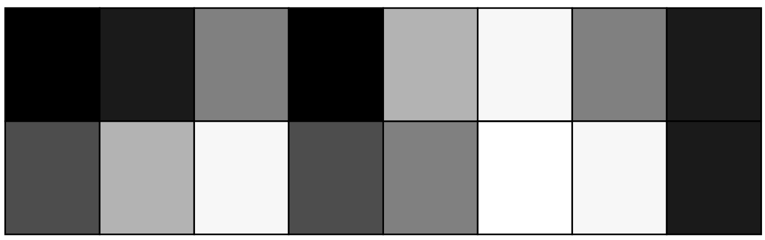
$$P_c = P(Xclean) \cong \prod_{f,t} \omega_{f,t}$$

Per band mask, P(x band(g,t) missing)

If P(band clean) = P(all components in band are clean),

$$\hat{\Omega}_{band},\ \omega_{g,t} = \prod_{f \in g} \omega_{f,t}$$

$g = 1\ldots2$



$t = 1\ldots8$

$$P(M|X) \cong \prod_{g,t \in M} \omega_{g,t} \prod_{g,t \notin M} (1 - \omega_{g,t})$$

Test results promising, even when $\Omega = 0.5$

$$\hat{Q} = \arg max_{Q,M} P(Q|X,M)$$

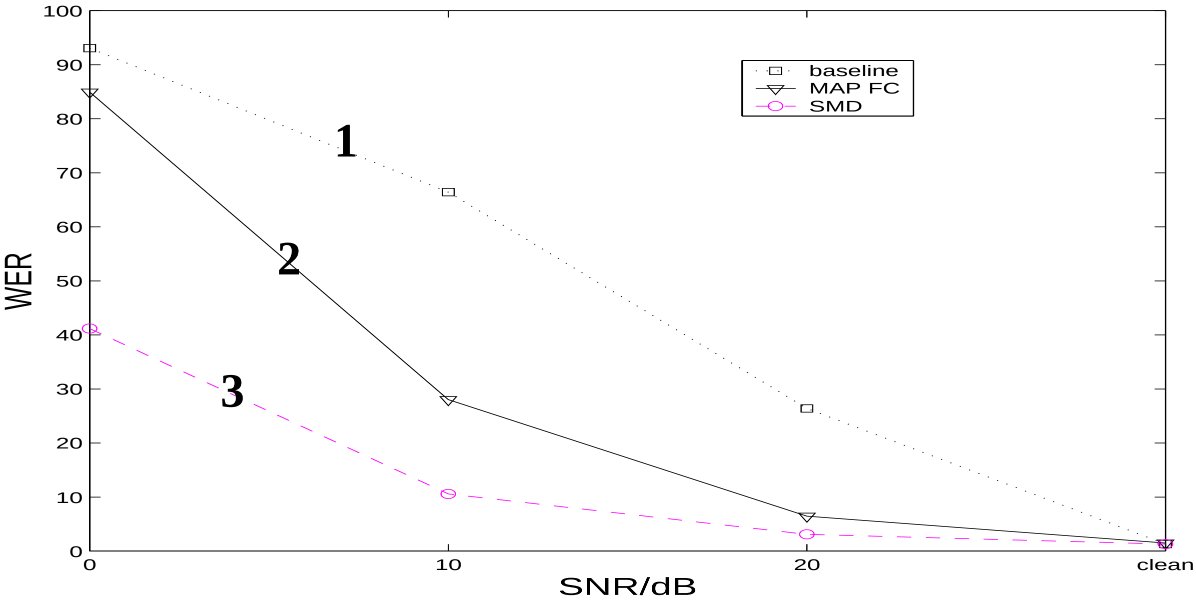


Fig shows WER (Aurora, av. over 4 noise conditions) for

1. baseline HMM/GMM
2. HMM/GMM AC multi-stream

   $\hat{Q} = \arg max_{Q,M} P(Q|X,M)$ (assumes $\Omega = 0.5$)

   Stationary band mask.

   Streams are MFCC with 1st & 2nd differences.

3. Usual missing-data ASR = SMD HMM/GMM SMD