T-61.182 Robustness in Language and Speech Processing

Astrid Hagen and Andrew C. Morris

Recent Advances in the Multi-Stream HMM/ANN Hybrid Approach to Noise Robust ASR, December 2002

explained by

Ramūnas Girdziušas

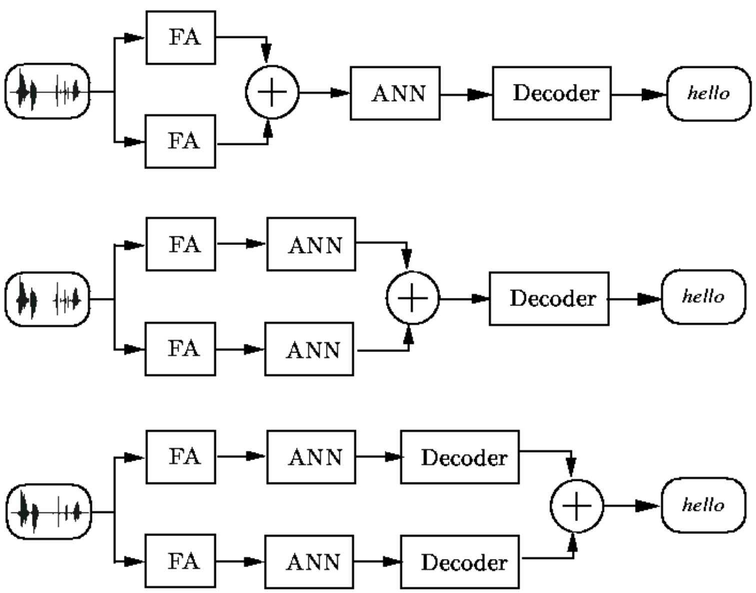
February 13th, 2003

1 – Introduction

- Standard HMM/GMM-based ASR systems perform well on clean speech.
- How to achieve noise robust ASR?
- Different features, context dependent speech units, various HMM extensions, phone and language models,...
- This paper:
  - an investigation of several hybrid ANN/HMM systems that use ANNs as multiple experts at various levels of a standard HMM-based ASR system;
  - mild focus on decision combination rules;
  - more on assessment of performance of several chosen ASR systems in the presence of non speech-like noise.

2 – Multi-expert systems in ASR

2.1 – Expert combination levels



Kuva 1: Multi-expert combination levels in HMM/ANN ASR systems. Top - feature level, middle - posterior probabilities level, bottom - hypothesis level.

2.2 – Combination rules

Example of combination rules at posterior probabilities expert combination level:

- q_k - speech state, $k = 1\ldots, K$.
- x - speech data for some given time frame.
- x_i - ith filter output.

Product rule

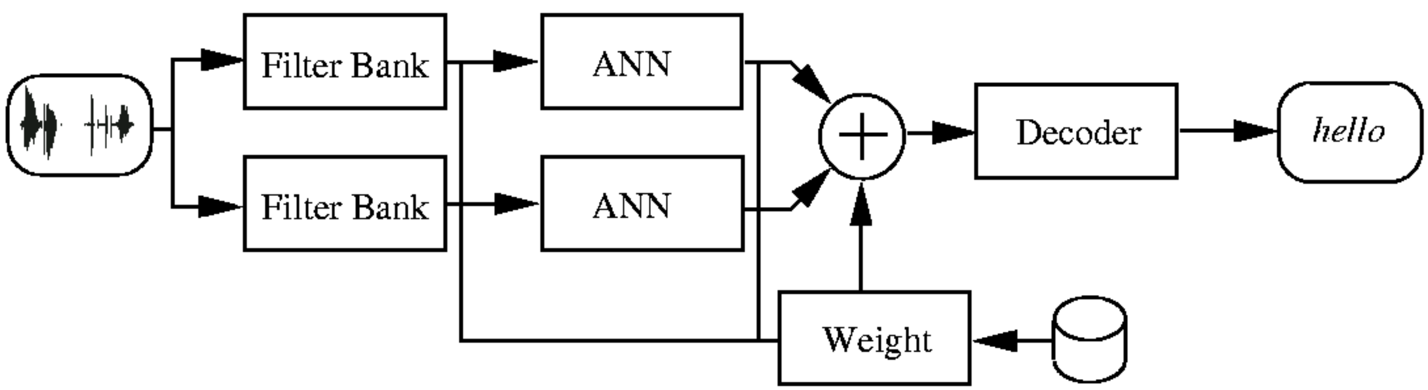
$$P(q_k|x) \propto \prod_{i=1}^{B} P(q_k|x_i), \tag{1}$$

Sum rule

$$P(q_k|x) \propto \sum_{i=1}^{B} w_i P(q_k|x_i), \tag{2}$$

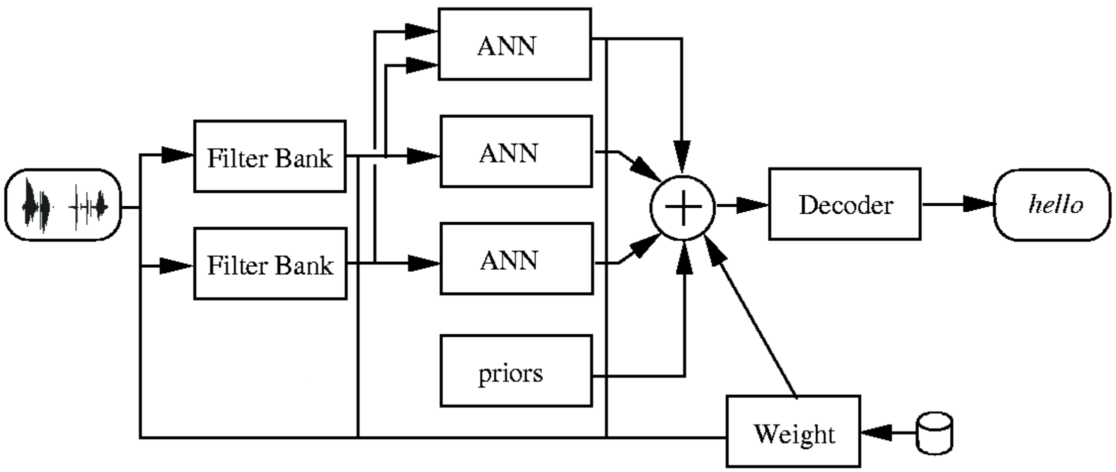
3 – Multi-band hybrid systems

3.1 – Standard Multi-band HMM/ANN Hybrid



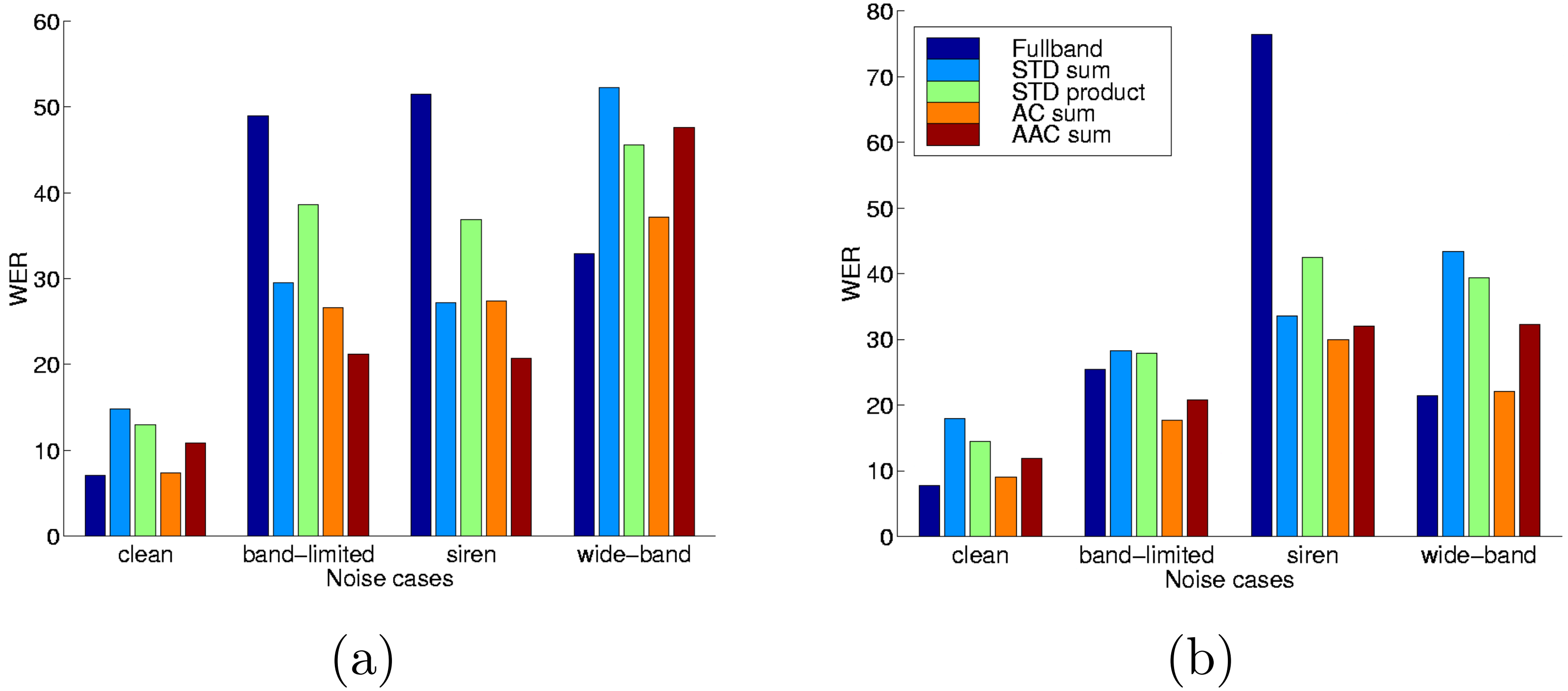
Kuva 2: Standard multi-band HMM/ANN hybrid. Posteriors level combination. Each expert sees narrow frequency range input.

3.2 – All-combinations Multi-band Hybrid



Kuva 3: Expert is trained for every possible combination of sub-bands. Combination is at both the feature and posteriors level.

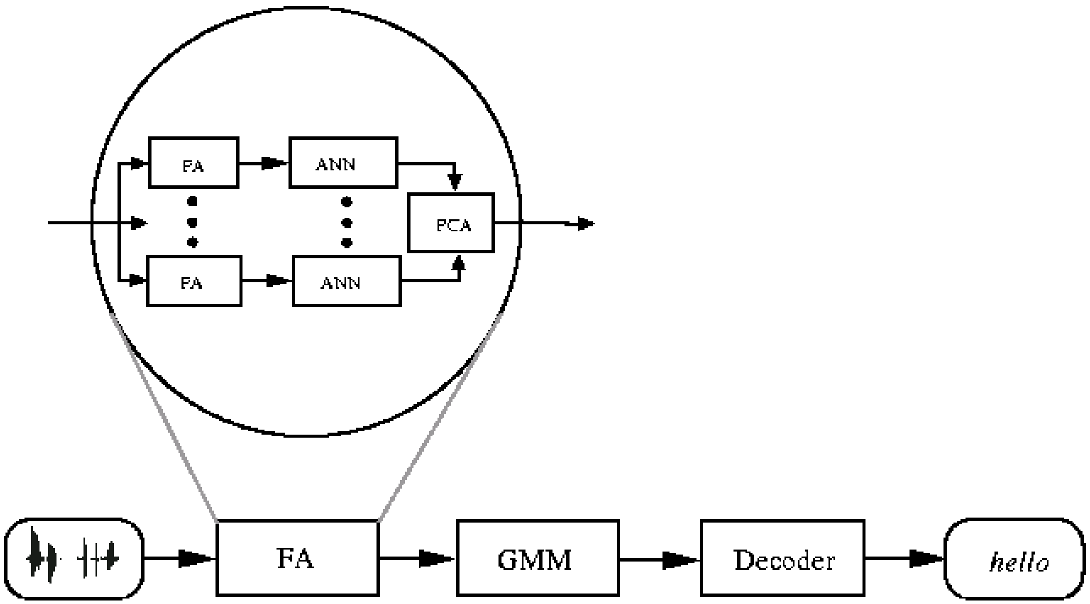
3.3 – Standard MB vs. All-combinations MB



Kuva 4: Standard Multi-band (STD) vs. All-combinations Multi-band (AC) on PLP features (a) and J-RASTA-PLP features (b). Numbers95 connected digits data.

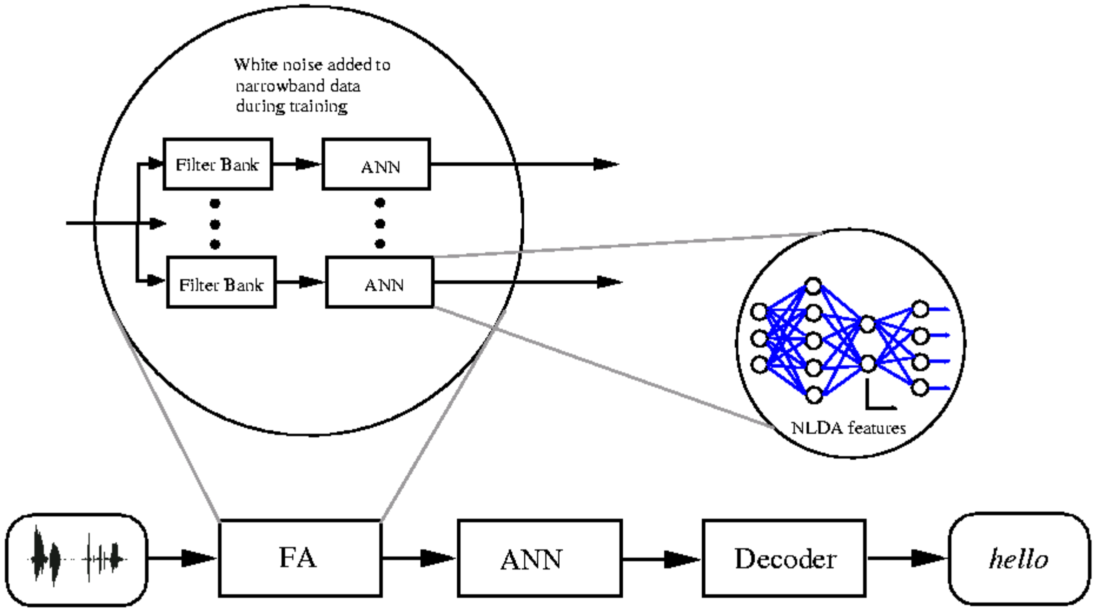
4 – Tandem HMM/ANN systems

4.1 – Multi-stream Tandem HMM/ANN hybrid



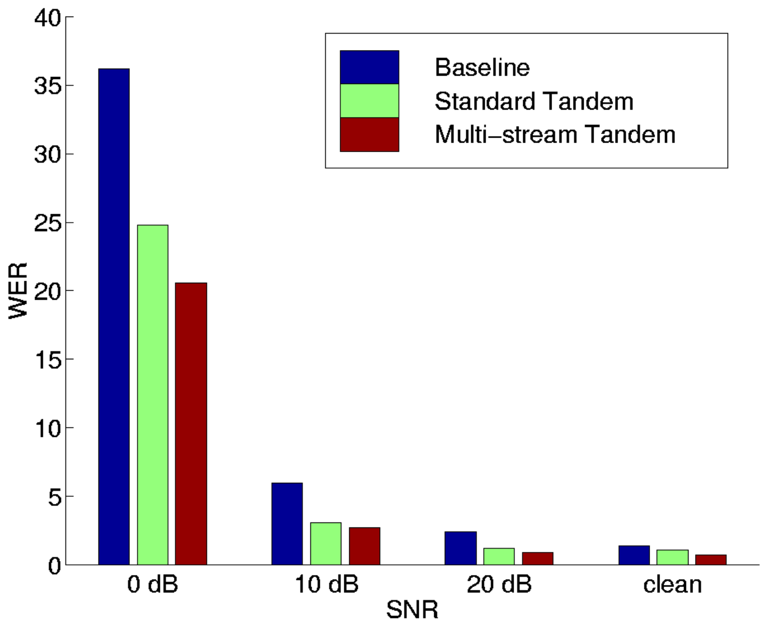
Kuva 5: Each ANN expert post-processes a separate stream of features.

4.2 – Narrow-band Tandem HMM/ANN hybrid

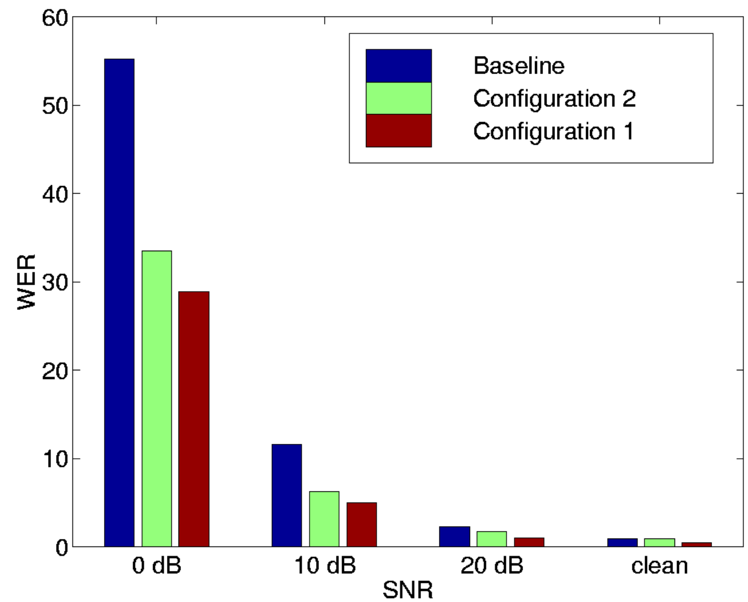


Kuva 6: Each narrow-band ANN expert is trained by adding white noise to its input. 127 HMM states, 1000 units in its hidden layer, 3 frames of context.

4.3 – Do 'tandem' systems improve the ASR?



(a)

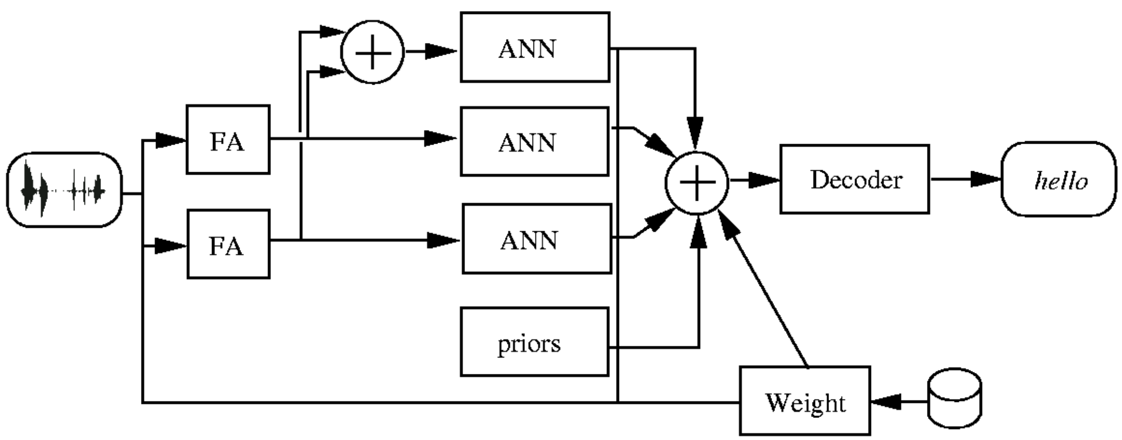


(b)

Kuva 7: Test results. Standard Tandem and Multi-stream tandem vs. baseline system (a) and Narrow-band Tandem vs. baseline system (b). Aurora 2.0 connected digits data.

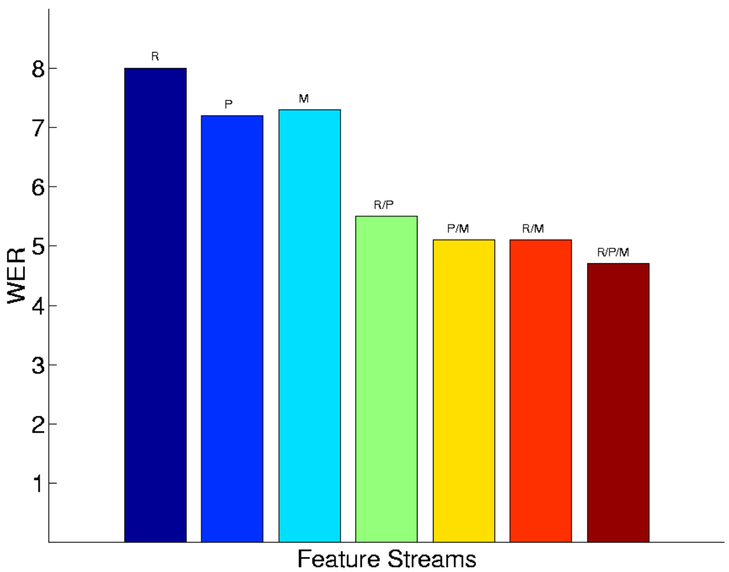
5 – All-combinations Multi-stream hybrid

5.1 – General idea

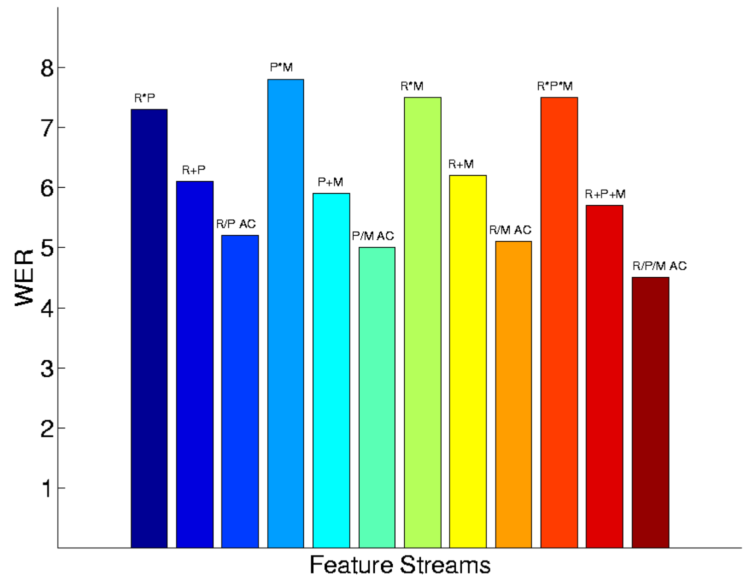


Kuva 8: Each expert acts as an independent speech state classifier with its own features and different model type.

5.2 – Test results



(a)

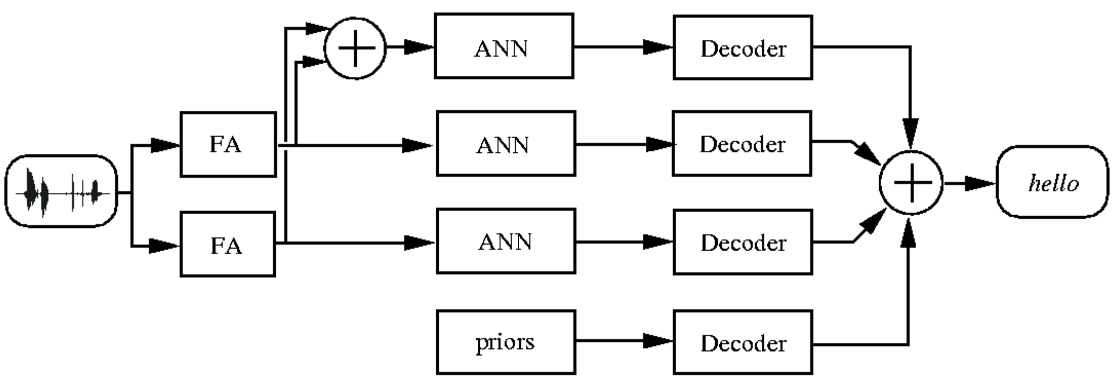


(b)

Kuva 9: Test results for the All-combinations Multi-stream hybrid, employing PLP (P), RASTA-PLP (R) and MSG (M) features. Results are presented for single streams and feature concatenation (a) and for posteriors combination (b). Portuguese SPEECHDAT.

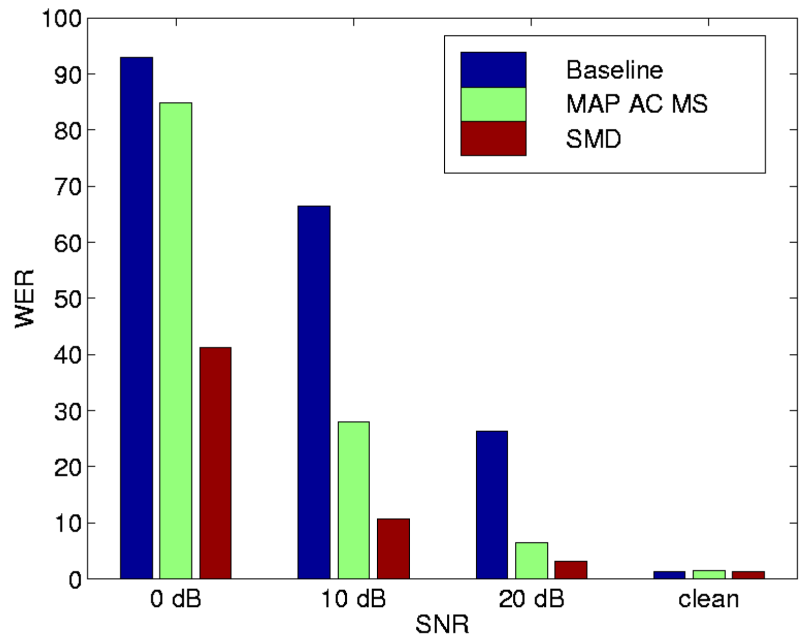
6 – Hypothesis level combination hybrids

6.1 – Idea



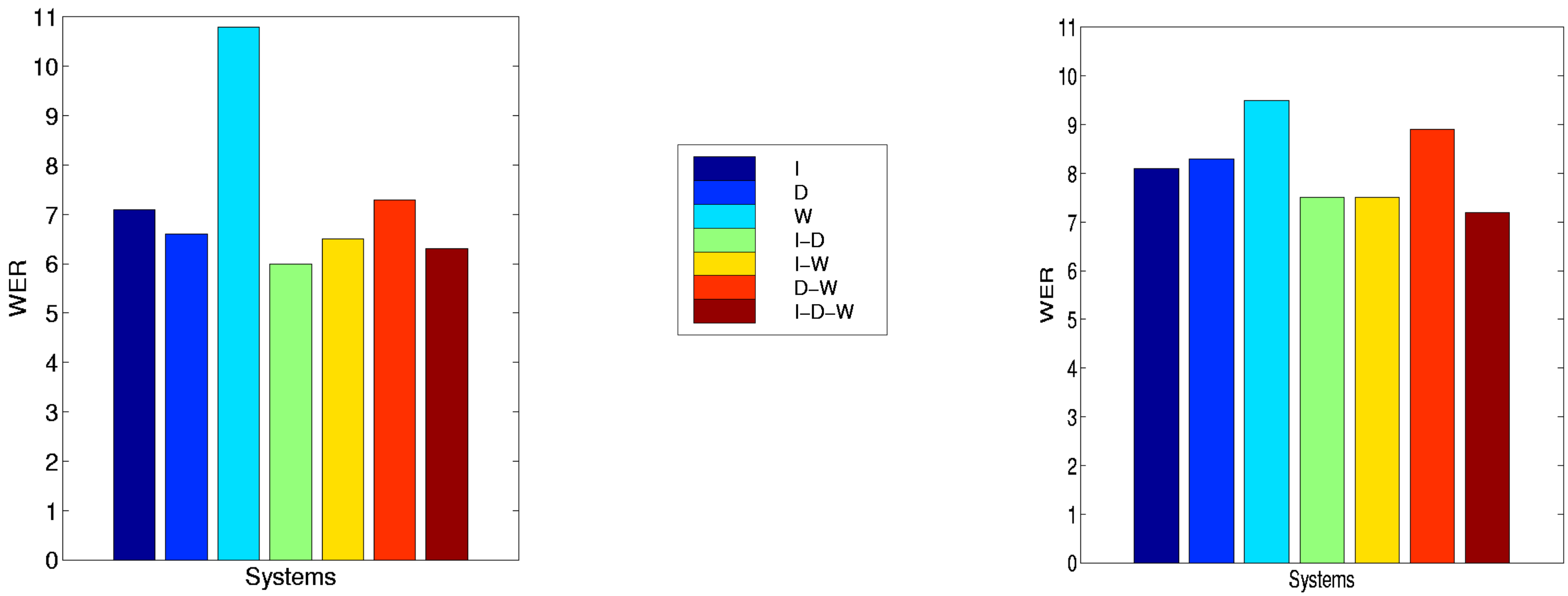
Kuva 10: All-combinations Multi-stream hybrid with hypothesis level combination. Each expert is trained on every possible combination of feature streams.

6.2 – ACMS hypothesis combination performance



Kuva 11: HMM-based ACMS MAP hypothesis level combination vs. baseline single-stream system and the 'soft missing data' (SMD) technique. Aurora 2.0 TIDIGITS connected digits corpus.

6 – Hypothesis level combination hybrids



(a)

(b)

Kuva 12: 3-ANN-based ACMS MAP hypothesis level hybrid, employing context-independent monophone models (I), context-dependent triphone models (D) and word models (W) vs. hypothesis level MS systems. PLP features (a) and RASTA-PLP features (b). SPEECH-DAT corpus.

7 – Comparative system performance

	SNBand	DNBnd	WBand	Clean	RFtrs	NoMat
MB	+	+	-	-	-	+
ACMB	+	+	0	0	+	+
MST	+	+	+	+	+	0
NBT	+	+	+	+	+	+
ACMS	+	+	+	+	+	+

Kuva 13: SNBand=static narrow-band noise; DNBand=dynamic narrow-band noise; WBand=wide-band noise; Clean=no noise; RFtrs=advantage persists with noise robust features; NoMat=advantage persists with non-matched noise types.

8 – Conclusions

- **Standard Multi-band**: simple, yet problems with the product rule, degrades ASR performance on clean speech!
- **All-combinations Multi-band**: pairwise sub-band dependence increases ASR performance in case of wide-band noise and clean speech.
- **Multi-stream Tandem**: processes different representations of full-band signal, useful in data fusion.
- **Narrow-band Tandem**: somehow disassembles noise, though immunity is only from the non speech-like noises.
- **All-combinations Multi-stream**: improves ASR performance on matched and non-matched noises, even without expert weighting. Unlike All-combinations Multi-band hybrids, improves ASR performance on clean speech.

9 – New directions

- New features.
- Multi-condition training.
- New classifier architectures.
- New combination rules and weighting schemes.
- Asynchronous decoding.
- One-stage multi-expert training.
- HMM/GMM based recognition with missing-data.