Heterogeneous Features in Multi-Stream ASR using the FC Approach

Astrid Hagen

IDIAP
Martigny - Valais - Suisse

Dalle Molle Institute for Perceptual Artificial Intelligence

Martigny • Valais Switzerland

http://www.idiap.ch

Overview

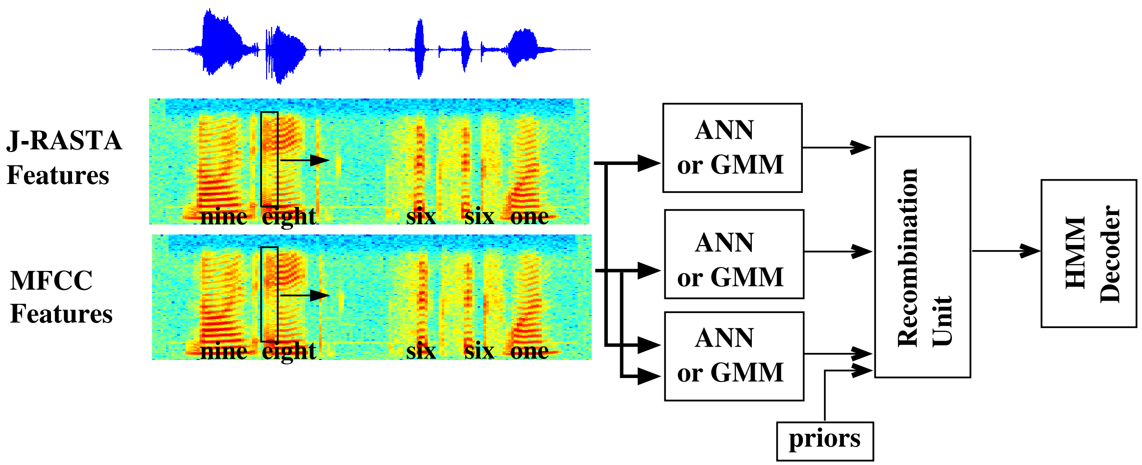
- From Multi-*Band* FC to Multi-*Stream* FC
- Heterogeneous Features in Multi-Stream FC
  - J-RASTA, PLP, MFCC, AMM Features

    Experimental Results
  - Multiple Time Scales
    - Variable Window Size Features
    - Difference Features

    Experimental Results
- Summary and Outlook

Multi-Stream Speech Recognition

FC in Multi-Stream



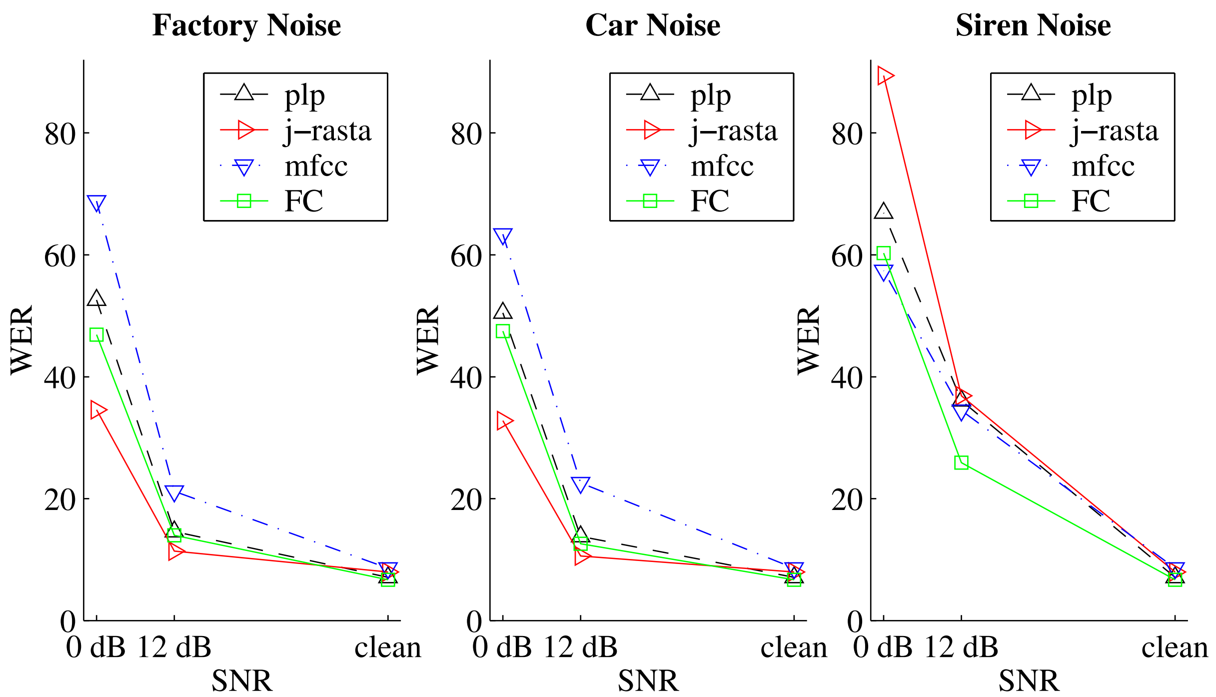
$P(q_k|x) \simeq \sum_{i=1}^{2^d} w_i P(q_k|x_{c_i})$ or $P(q_k|x) \simeq \prod_{i=1}^{2^d} P(q_k|x_{c_i})$

with d streams $\Rightarrow 2^d$ possible combinations

1. **Concatenation of feature vectors** into all possible combinations.
2. After independent processing of each stream, do **posterior combination**.

Heterogeneous Features in Multi-Stream FC

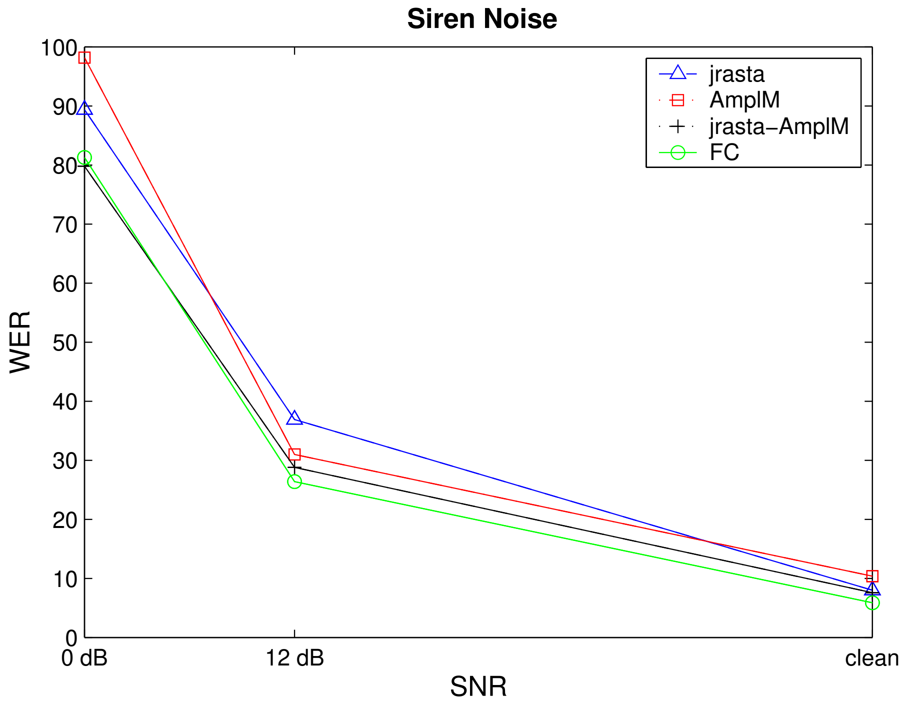
J-Rasta, MFCC and PLP Features



Word error rates for j-rasta, **plp** and mfcc features as single input streams and the multi-stream FC system using all their combinations as input streams.

Heterogeneous Features in Multi-Stream FC

Amplitude Modulation Map and J-Rasta Features



Word error rates for all combinations of amplitude modulation and j-rasta features in the multi-stream FC system. Amplitude Modulation Map features stem from Georg Meyer, Keele.

Multiple Time Scales in Multi-Stream FC

Variable Window Size Features

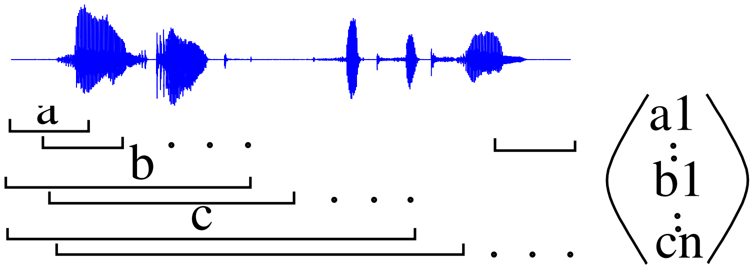


Illustration of multiple time scale features extracted from a regular-sized (a), triple- (b) and quintuple-sized (c) data window.

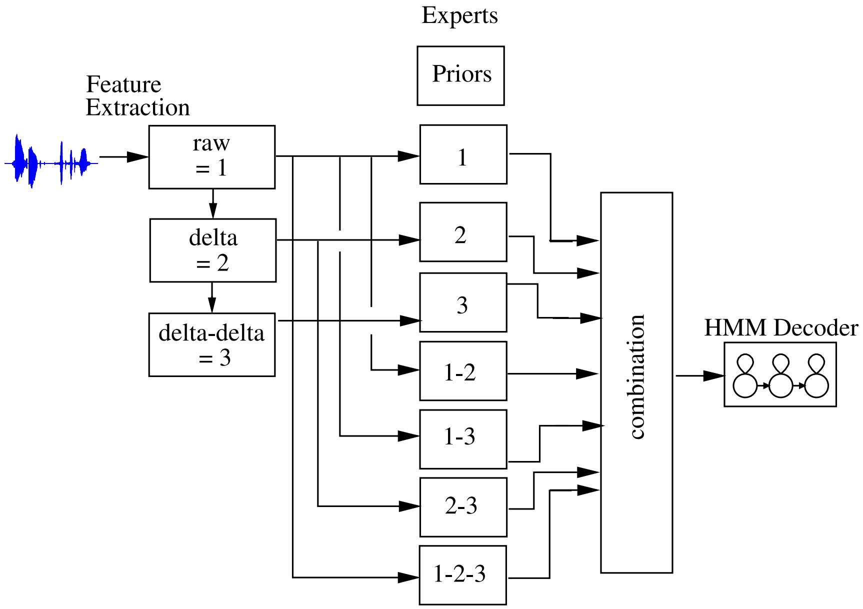
Experiments

	Car		Factory		Clean
	12 dB	0 dB	12 dB	0 dB	100 dB
(1)	13.8	50.5	14.6	52.6	7.1
(1)*(1-3)*(1-5)*(1-3-5)	14.2	45.5	14.7	49.3	8.1

Word error rates on clean speech and speech with car and factory noise for posterior combination with features extracted on **window sizes of 25 ms (1), 75 ms (3) and 125 ms (5)** (frame shift 12.5 ms).

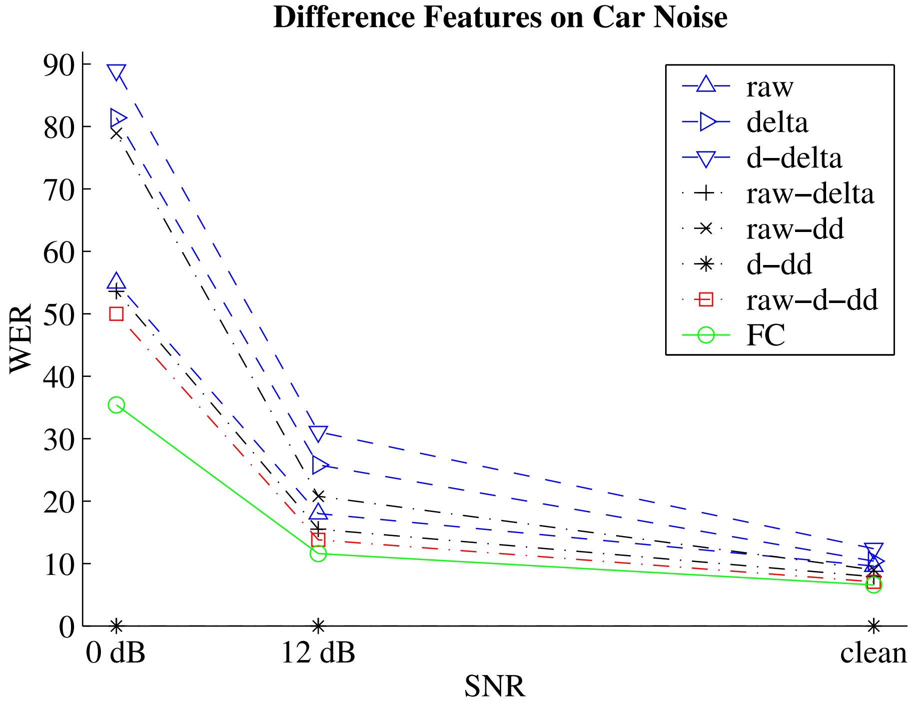
Difference Features for Multi-Stream FC

Instantaneous, Delta and Delta-Delta features taken as separate input streams. ⇒ Feature and posterior combination give 7 input streams to FC system :



Difference Features for Multi-Stream FC

Car Noise



Instantaneous PLP and Difference Features as Multiple Time Scales in the Full Combination (FC) Multi-Stream Approach.

("raw-d-dd" corresponds to the usually used full-band system incorporating raw, delta and delta-delta features as one input stream.)

Summary and Conclusion

- FC multi-*band* gives higher robustness than
  - the original multi-band approaches on clean and all noise conditions
  - the full-band system on band-limited noise (competitive in clean)
- FC multi-*stream* needed for
  - performance improvement in clean (as compared to full-band)
  - higher robustness to real environmental noise (–"–)
- FC multi-stream can incorporate
  - different acoustic feature streams (plp, j-rasta, mfcc; AMM)
    [Hagen,Morris,Bourlard: ASR2000]
  - different time scales (variable window size features, *difference features*)
    [Hagen,Bourlard: ICSLP2000]

Outlook

- Further work on multi-stream FC using more divers features, such as
  - Difference features or Amplitude Modulation Map features.
  - Phonetic or visual feature streams.
- Use of mixture of experts for multi-stream FC.
- Still need better (on-line) adaptive weights for both multi-band and multi-stream; ongoing work on ML weights adaptation.