Broadcast News: Features & acoustic modelling

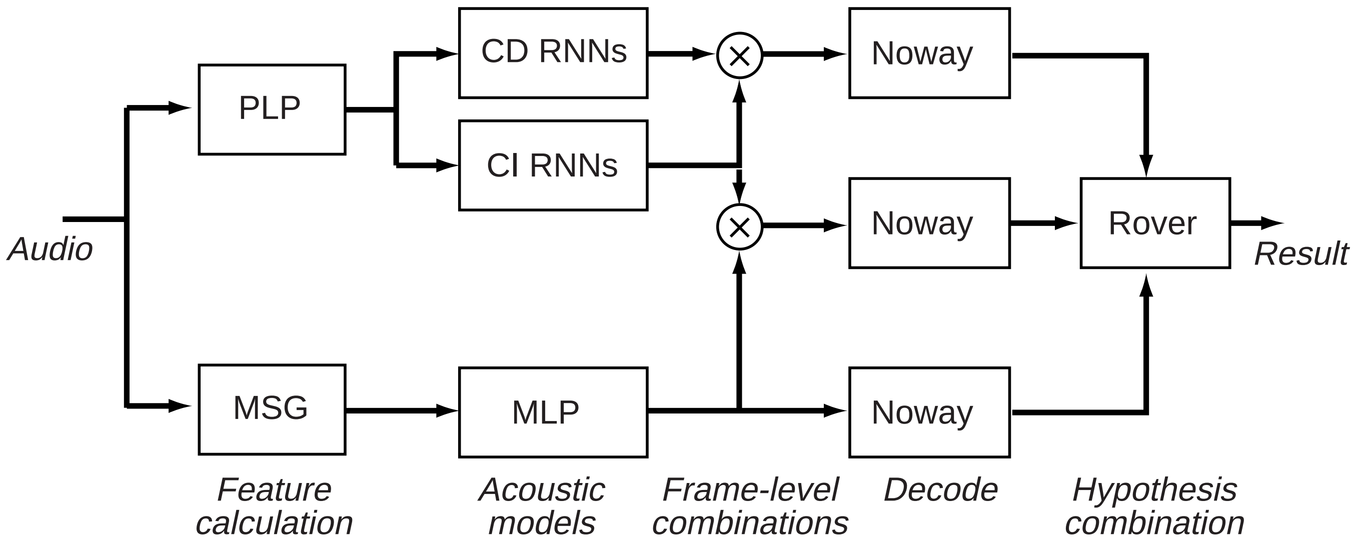
Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. **The modulation-filtered spectrogram**
2. **Features and combinations**
3. **Net size and training size**
4. **Results by condition**
5. **Whole-utterance features**
6. **Gender-dependence**

SPRACH BN System Overview

- **Abbot + 2nd acoustic model + ...**

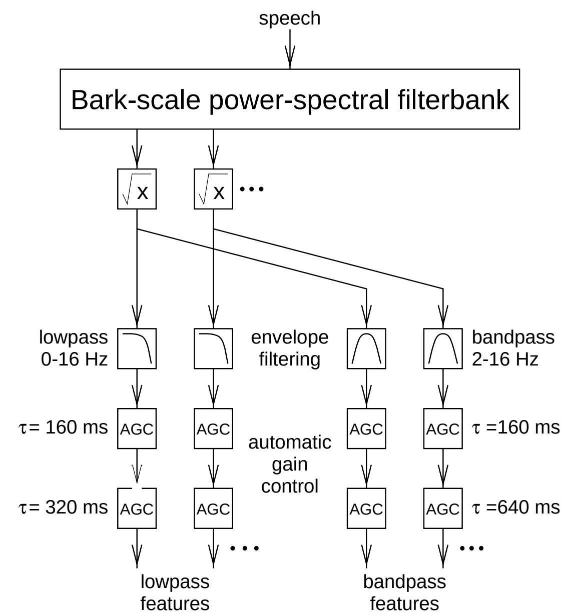


The modulation-filtered spectrogram

(Brian Kingsbury)

- **Goal: invariance to variable acoustics**

  - filter out irrelevant modulations
  - channel adaptation (on-line auto. gain control)
  - multiple representations



- **Results (small vocabulary):**

Feature	Clean test WER	Reverb test WER
plp	5.9%	22.2%
msg	6.1%	13.8%

Feature choice

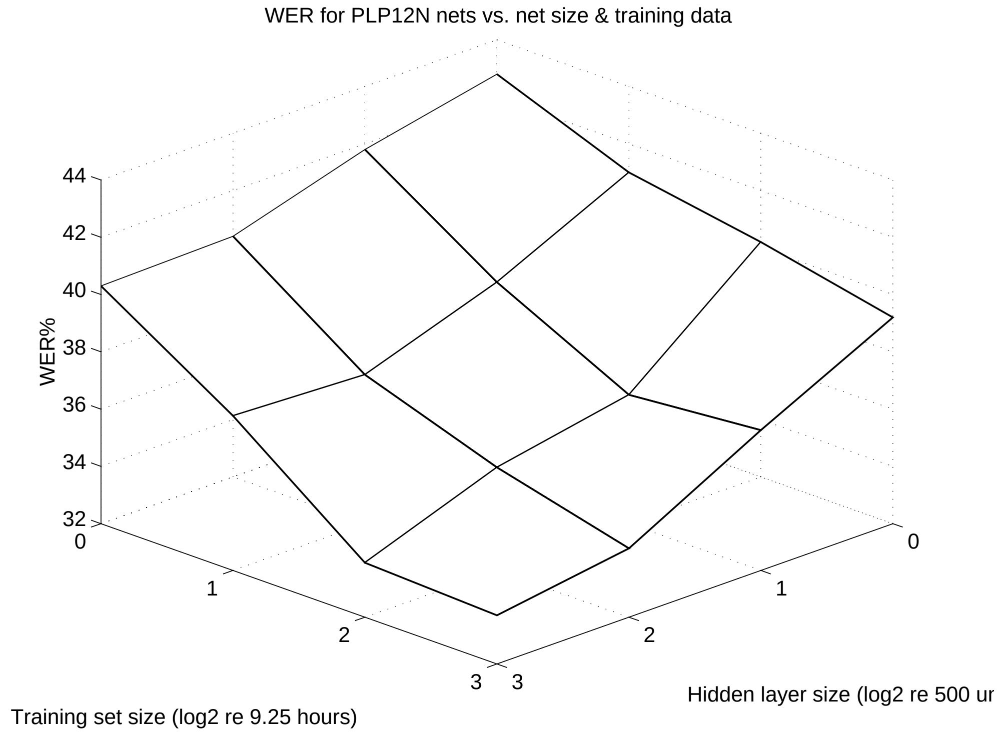
- **Base ABBOT system: normalized PLP**
- **Additional ftrs band-limited to 4kHz**
  - help with telephone speech
  - just to be ‘different’
- **Searched over features, deltas, context window**
  - plp12N-8k best alone
  - rasta performed poorly (16ms windows)
  - msg1N-8k best for combination with RNN

Feature	Elements	WER% alone	WER% RNN combo
RNN baseline			33.2
plp12N	13	36.7	31.1
ras12+dN	26	44.4	32.5
msg1N	28	39.4	29.9

(2000HU, 37h trainset, align2 labels, 7hyp decode)

Net and training set: Size matters

- **Huge amount of training data available**
  - 142h = 32M training patterns @ 16ms
- **Search over net size / training set size**



Evolution of the acoustic model

- **Increasing size of classifier, training set**
- **Iterative re-alignment of target labels**
  - in combination with RNN base
- **Steady improvement:**

HUs	Trnset	Labels	Trn time	WER%	ComboWER %
2000	37h	align1	4days	39.4	29.9
2000	37h	align2	4 days	38.6	30.4
4000	74h	align2	7 days	35.3	29.3
8000	142h	align4	21 days	31.6	26.8

(msg1N, 7hyp decode)

Results by acoustic condition

- **Evaluation results broken down into 6 spoke categories**
- **4kHz audio should help F2 (telephone)**
- **msg features might help poor acoustics**

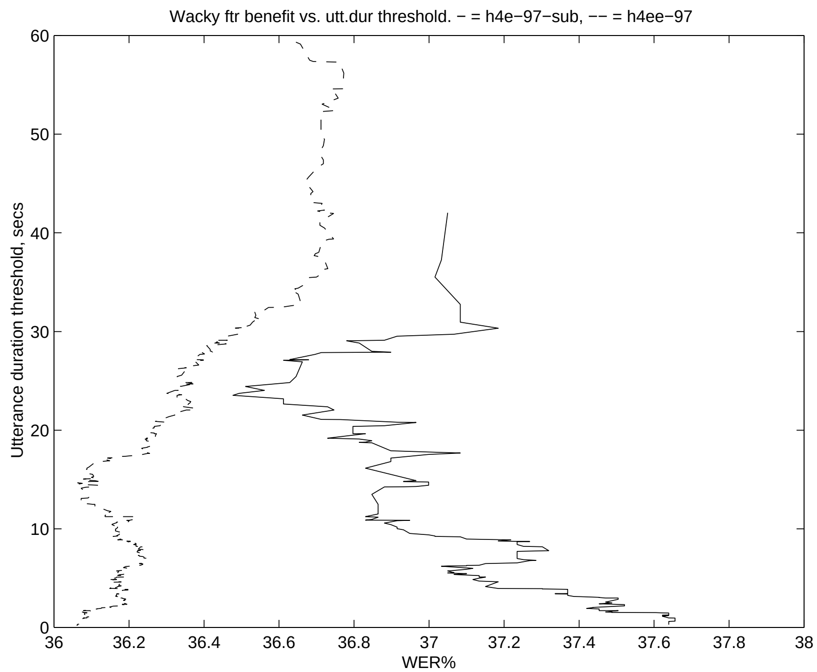
System	ALL	F0	F1	F2	F3	F4	F5	Fx
RNN	29.9	15.2	29.3	**51.6**	33.0	32.8	**19.3**	56.7
MSG (8000)	29.7	17.7	31.9	**39.4**	32.5	33.3	**29.8**	49.4
RNN+MSG	25.4	14.3	24.4	**38.0**	31.0	28.7	**18.5**	49.0
Sprach'98-1	21.7	11.6	24.7	32.4	33.8	15.5	27.9	29.6
Sprach'98-2	20.0	13.6	23.8	28.4	18.9	23.0	15.7	48.3

(full decode)

Whole-utterance features

- **BN groups have focussed on adaptation & normalization**
  - VTLN, MLLR, SAT
- **Maybe do similar thing with extra net inputs**

→ **Whole-utterance pre-normalization feature-dimension variances as constant inputs to net**



Gender dependence (GD)

- **Train separate nets on Female/Male data**
  - males represented 2:1 in BN
  - oracle labels?

Net	F% (2224)	M% (3711)	WER%(5938)
2000HU/25h UF	28.3	53.1	43.8
2000HU/25h UM	42.1	33.3	36.6
4000HU/50h U	29.6	33.6	32.1
Oracle best	25.7	30.5	28.7
Combination scheme	27.7	32.2	30.5

(plp12N-8k, 7hyp decode)

- **Best practical scheme**
  - use classifier entropy to choose M or F net
  - use decoder likelihood to choose GD or GI
- **Now training on full set**