RESPITE progress report

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

Outline

1. **Hybrid AURORA system**
2. **Using hybrid results with HTK**
3. **Multifeature design**
4. **Multistream pronunciation modeling**

1

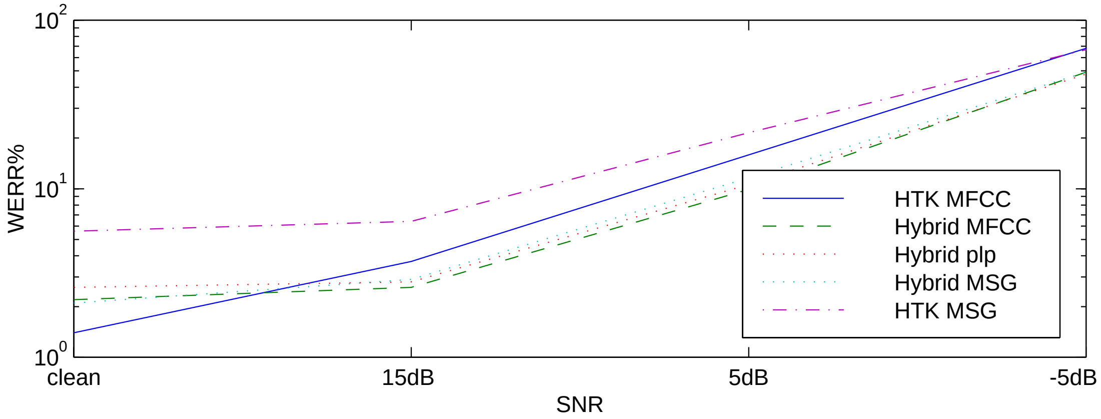
Hybrid AURORA system

- **AURORA noisy digits task**
  - TIDIGITS + 4 kinds of noise x 7 SNR levels
  - standard HTK back-end provided
  - objective: standard features for mobile phones
- **ICSI’s small-vocab techniques**
  - modulation-filtered spectrogram (MSG) features
  - posterior probability combination (multistream)
- **Can we combine them?**
  - hybrid NN-HMM baseline system for AURORA
  - use a TIDIGITS lexicon & phone models
  - bootstrap labels from NUMBERS95 network
  - use 480 hidden-unit net as N95

Baseline AURORA results

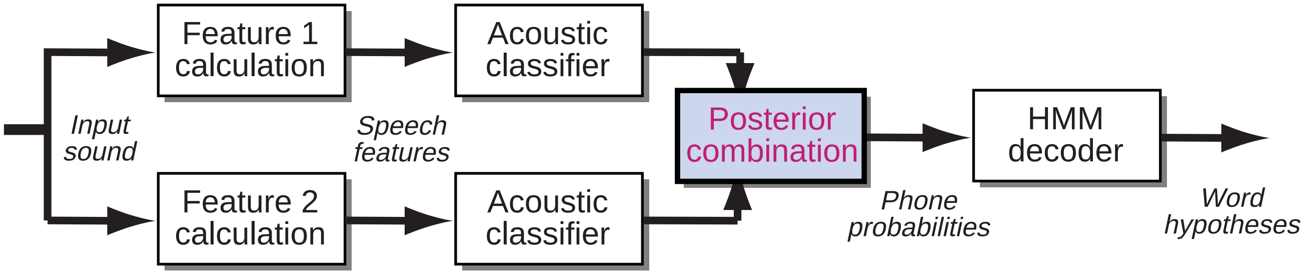
- **AURORA test has 28 numbers...**
- **...report just a few**
  - mean WER % for ∞, 15, 5, -5 dB SNR
  - + overall mean ratio to HTK MFCC baseline

System	Feature	Clean	SNR15	SNR5	SNR-5	Avg. ratio
HTK	**MFCC+d**	1.4%	3.7%	15.9%	68.0%	100.0%
Hybrid	**MFCC+d**	2.2%	2.6%	9.9%	49.1%	82.1%
Hybrid	**plp12N+d**	2.6%	2.8%	10.6%	47.9%	89.6%
Hybrid	**msg3N**	2.1%	2.9%	11.6%	49.2%	87.1%
HTK	**msg3NKG**	5.6%	6.4%	21.5%	66.8%	184.5%



Combination systems

- **Posterior combination has worked well**



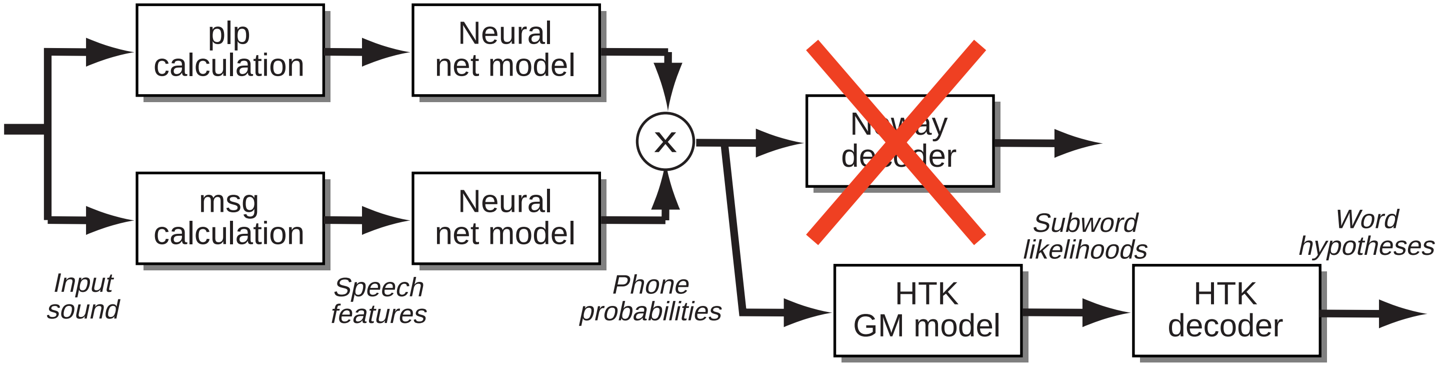
$$P(q_i|X_1,X_2) \propto P(q_i|X_1)\cdot P(q_i|X_2) \,/\, P(q_i) \quad \textbf{... if } X_1 \perp X_2 | q$$

- **But it depends on features**

Features	Clean	SNR15	SNR5	SNR-5	Avg. ratio
plp12Nd	2.6%	2.8%	10.6%	47.9%	89.6%
msg3N	2.1%	2.9%	11.6%	49.2%	87.1%
plp12Nd-msg3N	1.7%	2.4%	9.5%	47.3%	74.1%
plp12N-msg3aN **• dplp12N-msg3bN**	1.7%	2.1%	8.8%	46.9%	70.1%
plp12Nd • msg3N	1.5%	1.9%	8.2%	43.0%	63.0%

2 Using hybrid results with HTK

- **AURORA specification: use HTK recognizer**
- **How to put combinations into HTK**
  - feature combination (with LDA?)
  - posteriors as features (only 24 phone classes)

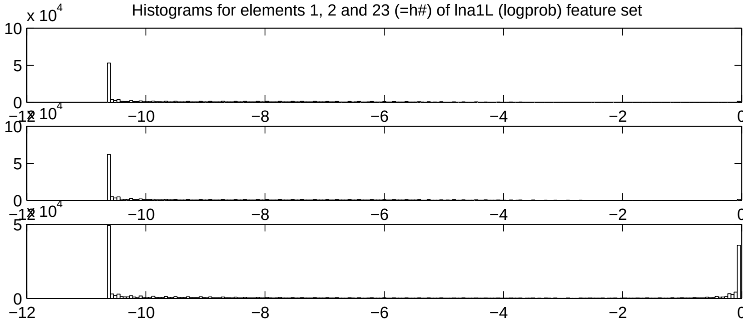


- **HTK handles it!**

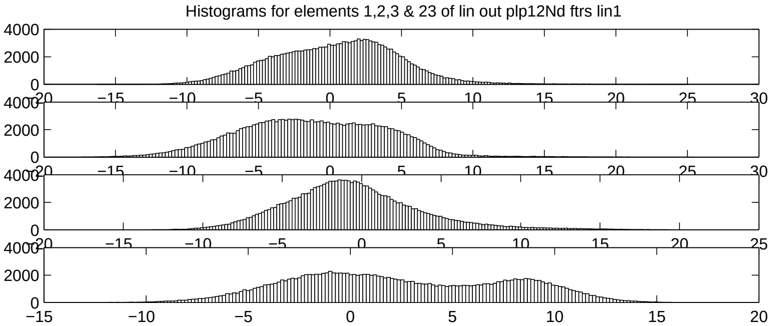
System	Feature	Clean	SNR15	SNR5	SNR-5	Avg. ratio
Hybrid	plp • msg	1.5%	1.9%	8.2%	43.0%	63.0%
HTK	**posteriors**	1.1%	1.9%	8.2%	46.1%	59.1%

Tailoring posteriors for HTK

- **Posteriors are very un-Gaussian**
  - log-transform doesn’t help much
- **A linear output layer helps a lot**
  - remove softmax: $y_i = \exp(x_i)/\Sigma_j(\exp(x_j))$



Histograms for elements 1, 2 and 23 (=h#) of lna1L (logprob) feature set



Histograms for elements 1,2,3 & 23 of lin out plp12Nd ftrs lin1

- **Do combinations by summing linear outputs**

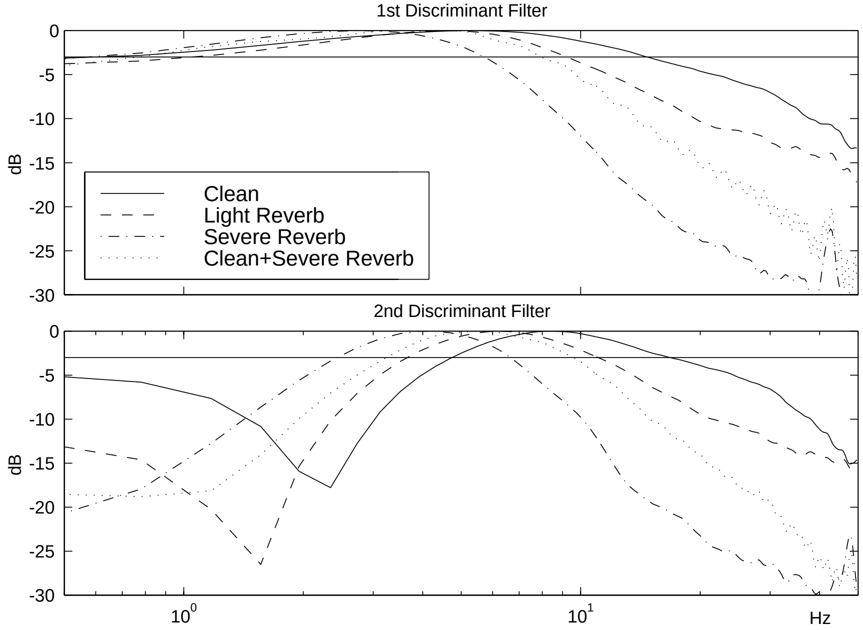
System	Feature	Clean	SNR15	SNR5	SNR-5	Avg. ratio
HTK	posteriors	1.1%	1.9%	8.2%	46.1%	59.1%
HTK	**log(p)**	0.9%	1.8%	8.9%	48.8%	58.6%
HTK	**Σ(lin. o/p)**	0.9%	1.6%	7.7%	44.1%	51.6%

3

Multifeature design

(Mike Shire)

- **‘Optimal’ features for different conditions**
  - subband envelope domain
  - linear-discriminant analysis (LDA) for filter coeffs
- **Modulation-frequency domain responses for clean, reverb, mixture:**



4 Multistream pronunciation models

(Barry Chen)

- **Combine streams in the decoder**
  - ‘HMM combination’
  - separate state assignment for each stream
  - constrain (disallow?) asynchrony
- **Are particular asynchronies important?**
  - between certain bands?
  - between certain sounds?
  - in particular directions?
- **Re-estimate transition probabilities in 1-state asynchrony 4-band models**
  - no improvement yet