Current work at ICSI

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

Outline

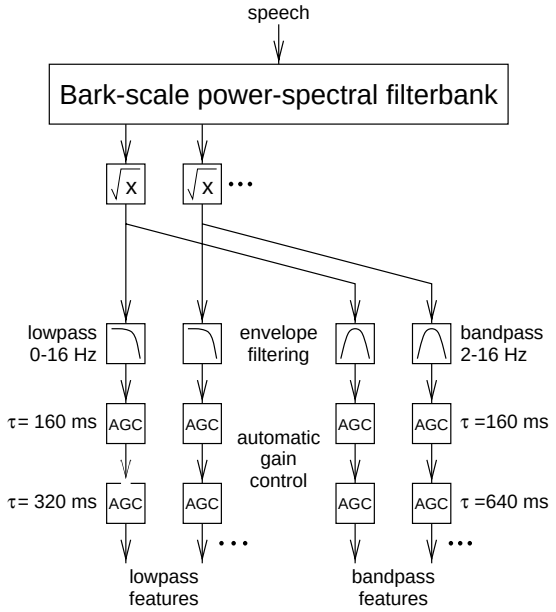
1. **Broadcast News MLP recognizer**
2. **Topic modeling**
3. **Acoustic segment classification**
4. **Thisl demonstrator front-end**

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%

Broadcast News recognizer

- **1998 evaluation - RNN + MLP**
- **8000 HU nets trained for MLP-only system:**

combo WER%	RNN98	MSG-8kHz	PLP-16kHz
RNN98	27.2	24.9	24.5
MSG-8kHz		29.7	24.4
PLP-16Khz			25.5

  - RNN+MSG+PLP: 23.7%
  - plp 8000HU forward-pass ~0.7x real time (spert)

- **Gender-dependent versions:**

net set	WER_F%	WER_M%	WER%
plp-GD	20.3	27.2	24.6
msg-GD			
plp+msg-GD			

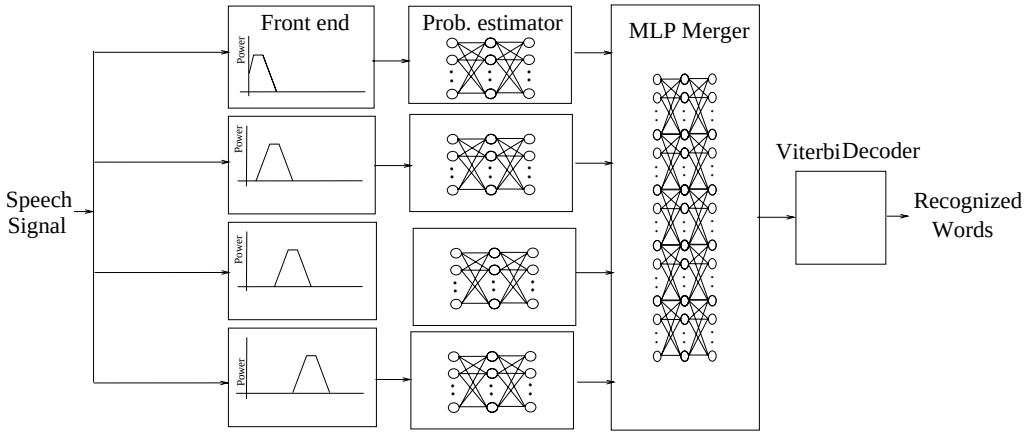
Broadcast News: ongoing

- **Dynamic pronunciations** (Eric Fosler)
  - data-derived rules for context-dependent pronunciations:
    phones, syllables, words, rate ...
  - rescored N-best output from 1st pass
  - ~ 3% RER improvement
- **Multiband** (Adam Janin / Nikki Mirghafori)
  - 20% RER for small-vocabulary (Numbers)
  - no significant improvement yet for BN
  - features: MSG, cepstra, KLT, plp
  - all-way possible combinations & weights

Multiband for Broadcast News

(Adam Janin / Nikki Mirghafori)

- **Scheme that worked best for small vocab:**
  - 4-way frequency split
  - plp cepstra+deltas within each band
  - MLP classifier for each band + MLP combiner



- **Weighted average of all possible combos**
  - $p(q \mid a,b,c,d) = \sum_S p(q \mid S,a,b,c,d) \cdot p(S)$

    S ranges over 16 possible combinations
  - $p(S)$ from? constant, local feature (entropy)
  - oracle best $p(S) \rightarrow$ WER=19% (25%RER)

Topic modeling

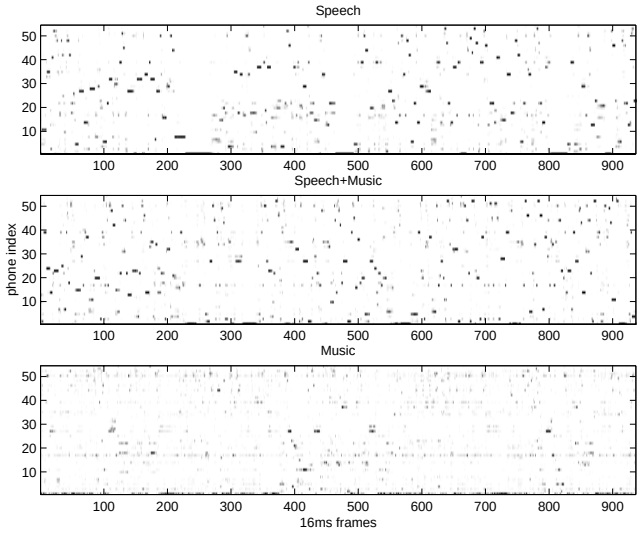
(Dan Gildea & Thomas Hofmann)

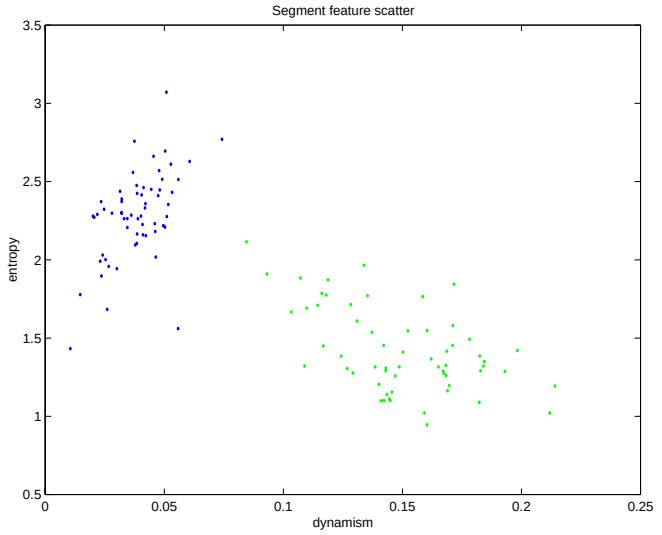
- **Bayesian model:**
  - p(*word* | *doc*) = $\sum_t$ p(*word* | *topic*) p(*topic* | *doc*)
  - EM modeling of p(*word* | *topic*) & p(*topic* | *doc*) over training set
  - p(*topic* | *doc*) estimated from context in recognition
- **Use to modify language model weights**
  - p(*word*) $\propto$ p_{tri}(*word*) p_{top}(*word*) / p_{uni}(*word*)
  - WSJ: trigram perplexity of 109 reduced 17%
  - use for BN recognition?
- **Use for topic segmentation?**

Acoustic Segment Classification

(Gethin Williams (SU) & Dan Ellis)

- **Features from posteriors show utterance type:**
  - average per-frame entropy
  - 'dynamism' - mean squared 1st-order difference
  - average energy of 'silence' label
  - covariance matrix distance to clean speech





- **100% on Scheirer/Slaney speech-music testset**
- **Use for acoustic segmentation?**

Thisl demo development

- **Stand-alone Tcl/Tk implementation**
  - doesn't require httpd
  - speech-input ready