Ideas for Next-Generation ASR

1. **Model the Whole Speech Signal**
2. **Handle Mixtures**
3. **Respect Diversity**
4. **Other Remarks**

Dan Ellis <dpwe@ee.columbia.edu>

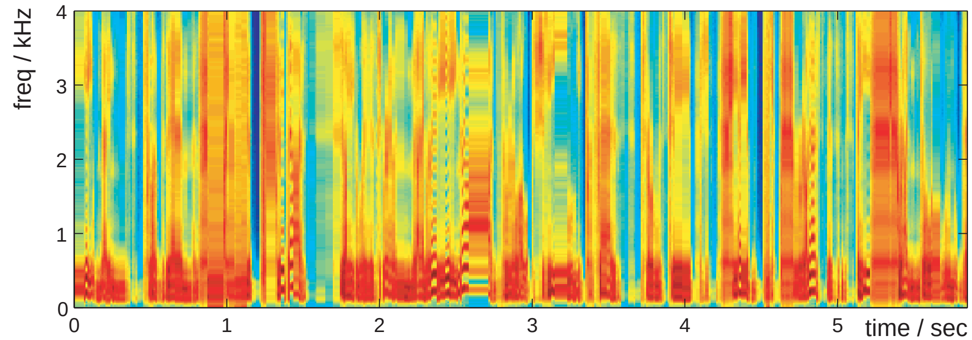
Laboratory for Recognition and Organization of Speech and Audio
(LabROSA)
Columbia University, New York
http://labrosa.ee.columbia.edu/

Outline

1. **Model the Whole Speech Signal**
   - Channel, accent, style
   - Timing/rate variation
   - Coarticulation
2. **Handle Mixtures**
3. **Respect Diversity**
4. **Other Remarks**

1 Model the Whole Speech Signal

- **HMM is a relatively weak model for speech**



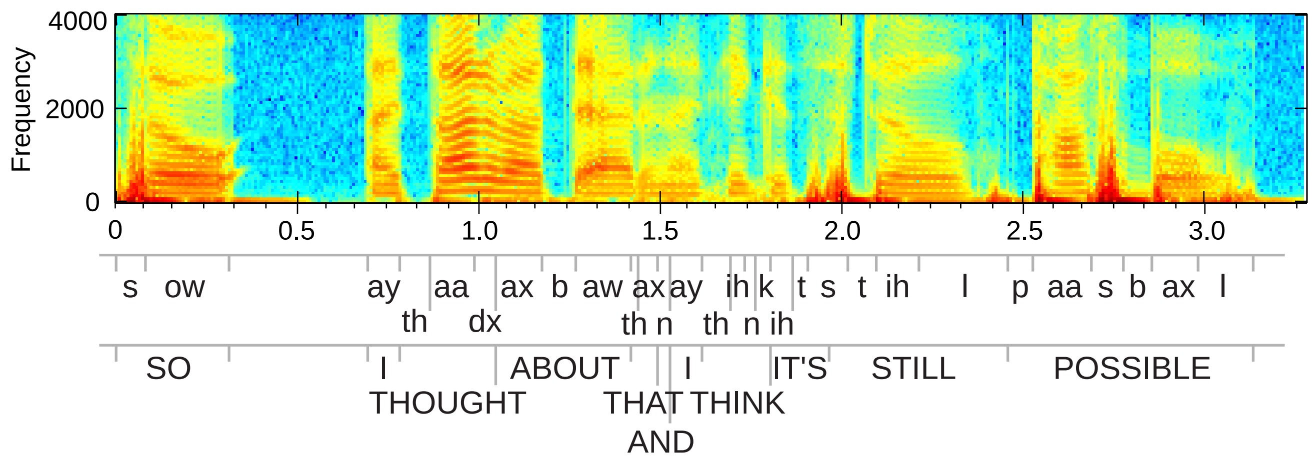
  - generates something speechlike, but
  - missing detail of real speech (...)
  - exponential segment durations
- **Only meant for** **inference** **of** $p(X|M)$
  - to choose between a few Ms
- **What would it take to model entire signal?**
  - capture perceptually sufficient information
  - e.g. speech coding quality

Channel, Accent, Style

- **Factors affecting spectral distributions**
  - just absorbed into model variance?
  - or: adapted generically e.g. MLLR
- **Channel**
  - typically fixed per session
- **Accent**
  - typically fixed per speaker
- **Style**
  - particular to application?
- **Modeling these factors explicitly**
  - improves generalization
  - reduces variance of models, hence..
  - allows better discrimination of voice from other random stuff

Timing/Rate variation

- **Timing is weakly modeled with current HMMs**
  - duration models have little influence on WER
- **Some baseline variation**
  - small improvements with rate-dependent models
- **Greatest variation is within phrase?**



  - models for within-phrase rate-contours
  - use as constraints on recognition
    (contrast likelihoods of alternate hypotheses
    e.g. in rescoring)

Coarticulation

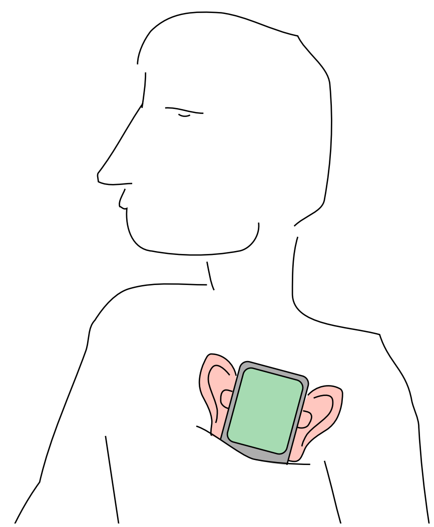
- **HMMs are piecewise-constant feature models**
  - ... at least with Viterbi decoding
  - delta features can map trajectories to more constant values, but just a ‘patch’
- **More states, more context-dependence reduces mismatch**
  - but model is too general: adjacent states are in fact strongly related
  - consequence: insatiable hunger for 1000s of hours of training data
- **Generative models of coarticulation not particularly hard**
  - e.g. HDM, SSM (Deng, Bridle, ...)
  - inference is hard...
  - ... but many new techniques from Machine Learning community (MCMC, variational, ...)

Outline

1. Model the Whole Speech Signal
2. **Handle Mixtures**
   - Frontier applications
   - Auditory Scene Analysis
   - Multisource models
3. Respect Diversity
4. Other Remarks

2 Handle Mixtures

- **Historically, speech recognition was made tractable by limiting domain to pure speech**
  - limited problem still hard enough
  - as a consequence: systems discard information that distinguishes speech/nonspeech
- **Many (most?) "frontier applications" involve nonspeech and mixtures**
  - meeting recordings
  - multimedia indexing
  - "Lifelog" audio diary



Approaches to handling sound mixtures

- **Separate signals, then recognize**
  - Computational Auditory Scene Analysis (CASA), Independent Component Analysis
  - nice, if you can make it work
- **Recognize combined signal**
  - ‘multicondition training’
  - combinatorics seem daunting
- **Recognize with parallel models**
  - optimal inference from full joint state-space

$$p(O, x, y) \to p(x, y|O)$$

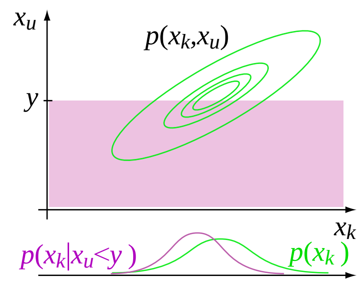
  - or: skip obscured fragments, infer from higher-level context
  - or do both: missing-data recognition

Missing Data Recognition

(Barker, Cooke & Ellis ’03)

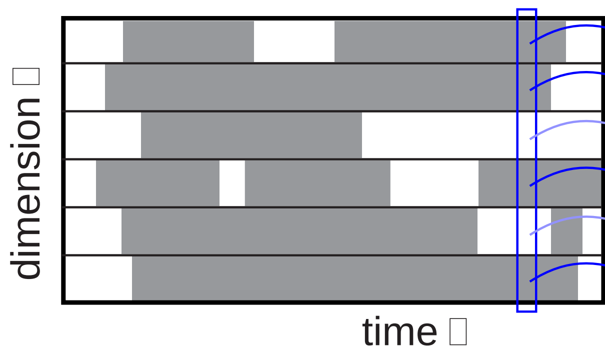
- **Can evaluate speech models** $p(\mathbf{x}|m)$ **over a subset of dimensions** x_k

$$p(\mathbf{x}_k|m) = \int p(\mathbf{x}_k, \mathbf{x}_u|m)\,d\mathbf{x}_u$$



- **Hence, missing data recognition:**

Present data mask



$$P(\mathbf{x}\,|\,q) = P(x_1\,|\,q)\cdot P(x_2\,|\,q)\cdot \cancel{P(x_3\,|\,q)}\cdot P(x_4\,|\,q)\cdot \cancel{P(x_5\,|\,q)}\cdot P(x_6\,|\,q)$$

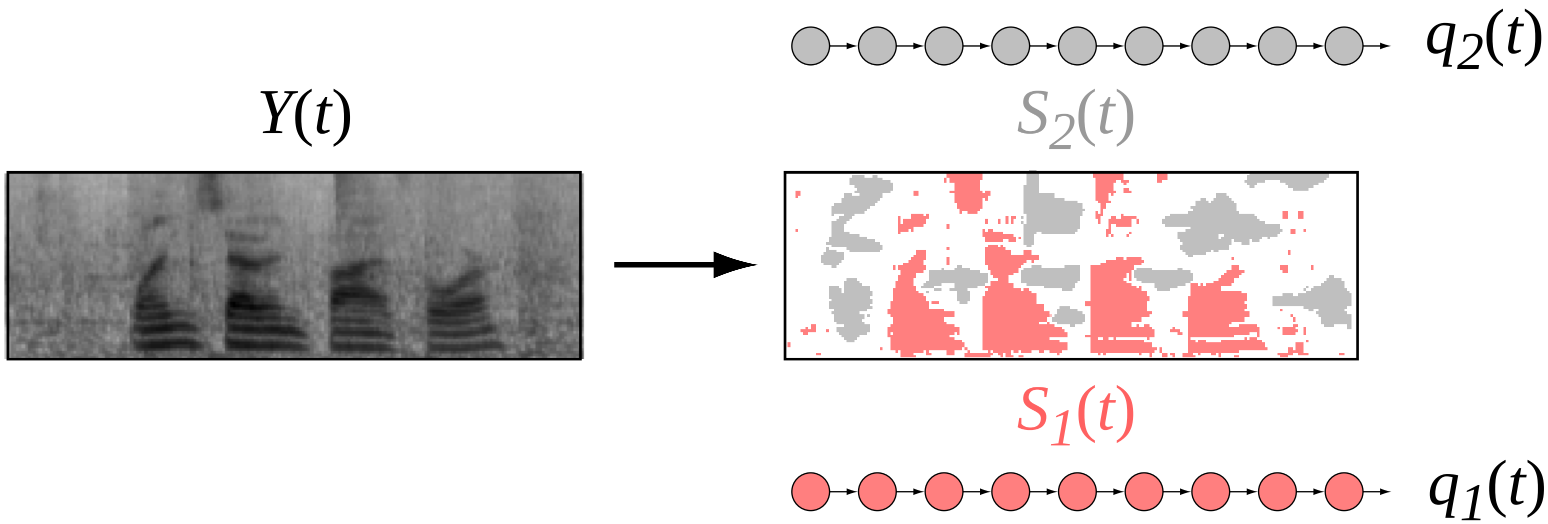
.. but need segregation mask

- **Fit model and segregation given obs’n:**

$$P(M, S|Y) = P(M)\int P(X|M)\cdot\frac{P(X|Y,S)}{P(X)}dX\cdot P(S|Y)$$

Multi-source decoding

- **Search for more than one source**



- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
  - locally coherent regions
- **Lots of issues in models, representations, matching, inference...**

Outline

1. Model the Whole Speech Signal
2. Handle Mixtures
3. **Respect Diversity**
    - Class-specific classifiers
    - Using different information differently
4. Other Remarks

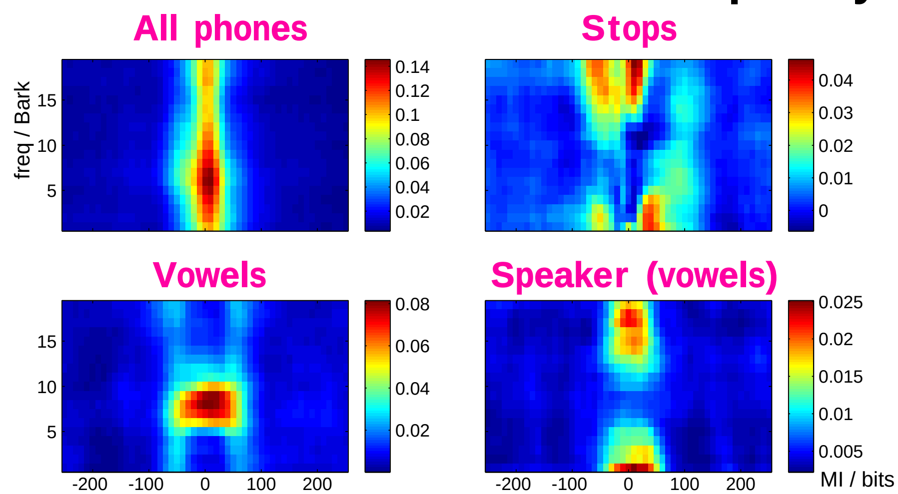
3 Respect Diversity

- **Speech signal is very diverse**
  - different kinds of phonemes (vowels, stops...)
  - different kinds of information (lexical, affective...)
  - different timescales (phones, words, phrases...)
- **Information needs are diverse**
  - phoneme classification
  - syllable detection
  - phrase detection
- **Technical approaches are diverse**
  - the more different they are, the bigger the gain from combination
  - ‘Rover effect’

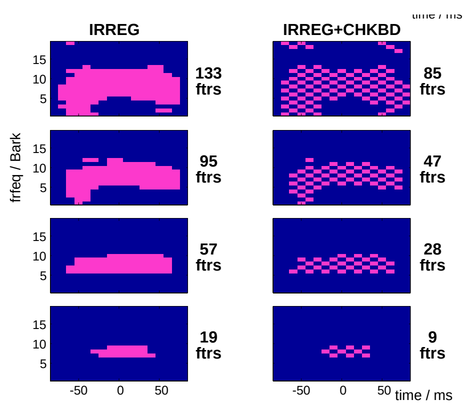
Finding the Information in Speech

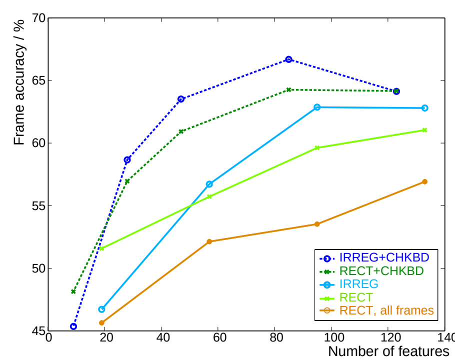
(Scanlon & Ellis, Eurospeech '03)

- **Mutual Information in time-frequency:**



- **Use to select classifier input features**





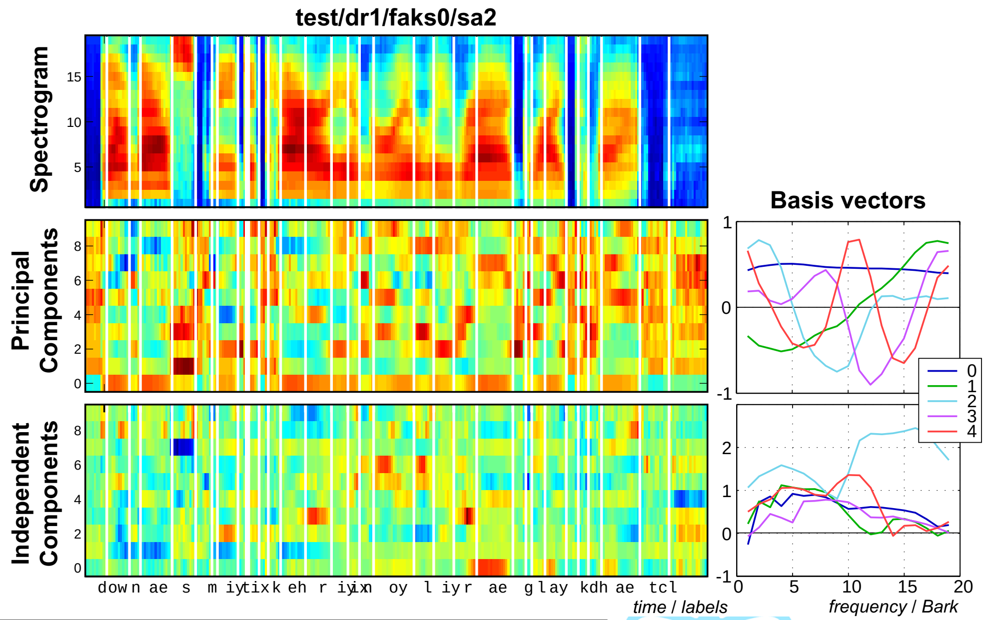
Using different information differently

- **Integrating paradigms have lots of power**
  - e.g. HMM does it all: time warp ... LM

  ... but can we gain by breaking up the tasks?
  - separate vowel center detection
    & consonant “adornment” classification

→ **“Event-based” recognition**
  - separate specialized
    detectors

- acoustic-phonetic features
  .. or data-derived near-equivalents
  from e.g. Independent Component Analysis



Outline

1. Model the Whole Speech Signal

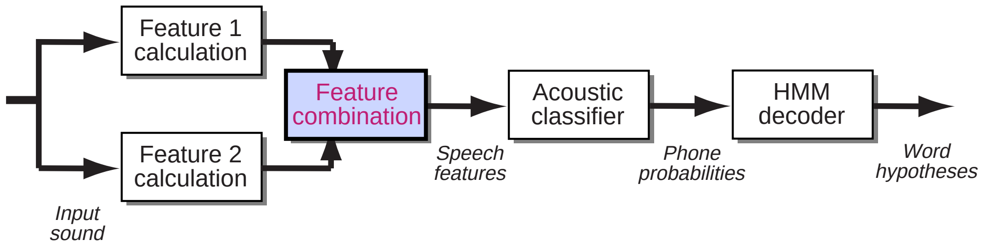
2. Handle Mixtures

3. Respect Diversity

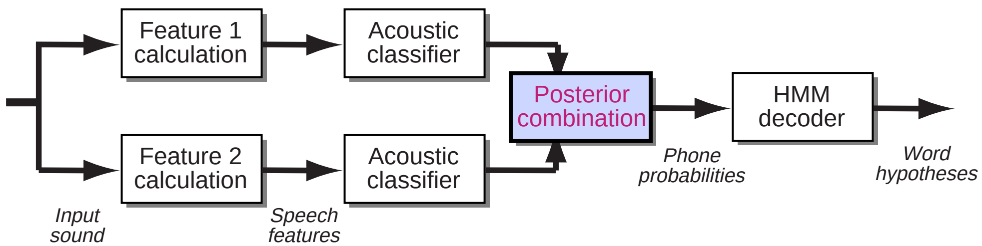
4. **Other Remarks**
    - Combinations & infrastructure
    - How much data?
    - Blackboards

4 Other Remarks: Different ways to combine systems

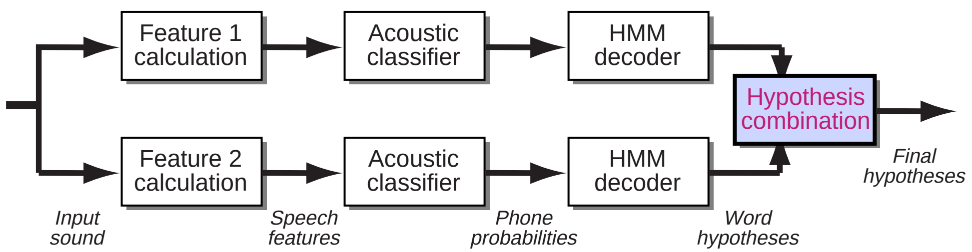
- **After each stage of the recognizer**



10% RER improvement (2 streams)



20% RER improvement (2 streams)

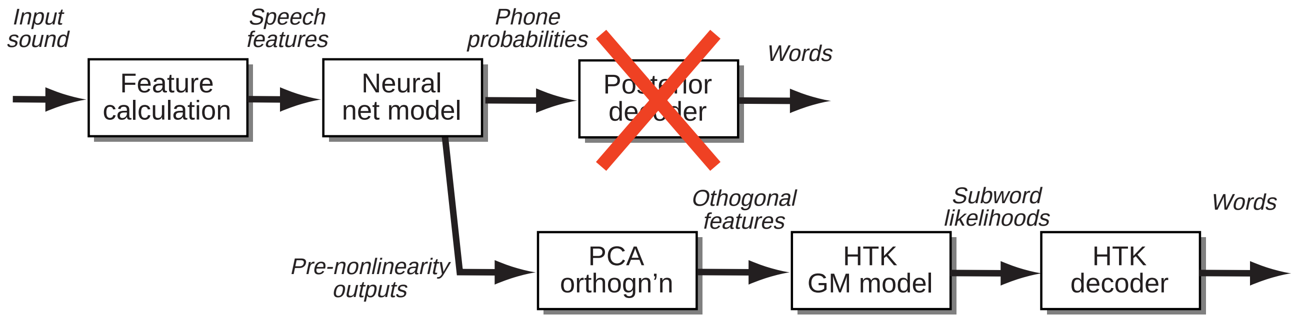


25% RER improvement (5 stream ROVER)

Combining modeling techniques: ‘Tandem’ acoustic modeling

(Hermansky, Ellis, Sharma, ICASSP’00)

- **To combine Neural Net models with HMMs:**

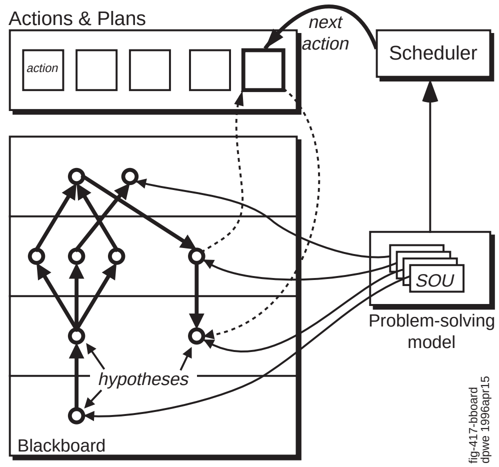


- **Result: better performance than either alone!**
  - Tandem alone: 20% RER improvement
  - Posterior combination + Tandem: 50% RER improvement
- **Excellent infrastructure for feature experiments**
  - nets are tolerant of feature eccentricities
  - e.g. MSG features→HTK has double the WER of Tandem version, MSG→net→HTK

How Much Data?

- **Near-unanimous calls for more data**
  - sure-fire way to improve accuracy .. a little
  - labeled data is expensive, hence limited
- **How much data do we need?**
  - see examples of 'all' speech variants?
  - infant example: 6 hr/day = 2000 hr/yr
    = 10,000 hr by age 5 (Moore graph)
  - brute force recognition-by-matching:
    every possible syllable? word? phrase? x voices
    100 voices x 2k syllables x 4/sec x 100 contexts
    = 1,600 hr (6k syl/hr)
    (but: distribution of examples)
- **What about generalization???**
  - goal should be abstraction of patterns
    from examples corpus
  - i.e. marginals, not full volume of examples

Blackboards?



- **Events,**
  Tiers,
  Hypothesis Generation &
  Verification
  = Hearsay Blackboard (1973)

- **What went wrong last time?**
  - bad knowledge, blame allocation
  - inefficient decoding
  - how to incorporate training?
- **So, this time around?**
  - new mechanisms for blame?
  - inefficiencies don't matter so much?
  - induction of rules from data?

Summary & Conclusions

- **Accept that sound is often/usually a mixture**
  - combine models and/or carve up features
- **Use more detailed models of speech**
  - so we can still recognize after carving up
- **Tandem models as enabling infrastructure**
  - able to glean value from wacky features
- **Find novel approaches for recognition**
  - vowel nuclei + adornments?