An overview of Speech Recognition research at ICSI

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

Outline

1. About ICSI
2. Hybrid connectionist-HMM speech recognition
3. Overview of current projects
4. Some details:
   Combinations, News retrieval, SpeechCorder
5. Conclusions

About ICSI

http://www.icsi.berkeley.edu/

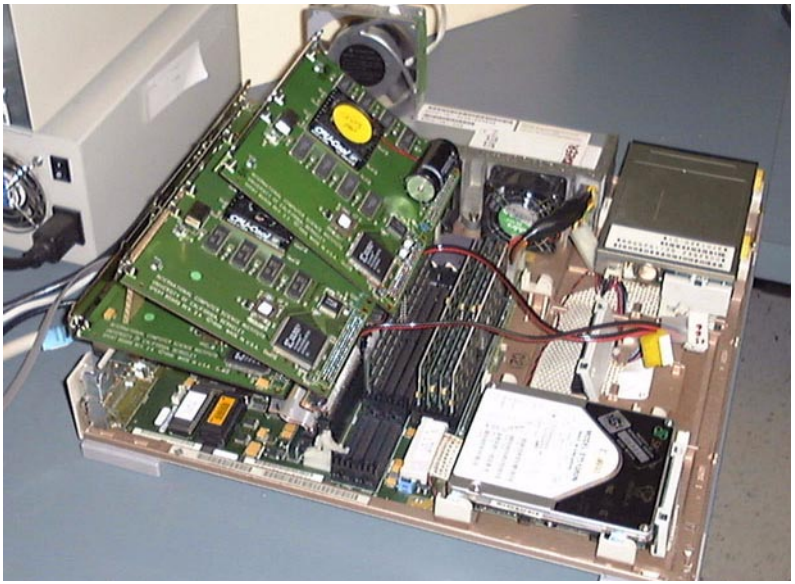
- **Founded 1988 as 'portal' between U.S. and European academic systems**
  - attached to UC Berkeley (but independent)
  - about 100 people at any time
- **Infrastructure funding from European government/industry**
  - .. in return for hosting visitors
  - Germany, Italy, Spain, Switzerland, Netherlands
- **Research groups consist of staff, visitors and UC Berkeley graduate students**
- **Original vision: 'massively parallel systems'**
  - has diversified since then

Groups at ICSI

- **“Realization” (Real Speech Collective?)**
  - Nelson Morgan, Steve Greenberg, Dan Ellis
  - Speech recognition for realistic conditions
  - Systems support for ASR applications
- **ACIRI**
  (AT&T Center for Internet Research at ICSI)
  - Network routing, traffic, security
- **Theory**
  - mathematical complexity, coding theory
- **Applications**
  - natural langauge understanding
  - connectionist models of cognition
- **Networks**
  - multimedia communications applications

ICSI Realization highlights

- **Connectionist framework for ASR (Morgan & Bourlard)**
- **RASTA front-end processing (Hermansky & Morgan)**
- **The Ring Array Processor (RAP)**
- **SPERT/T0**
  - SBUS boards with custom 8-way vector μ–proc



Outline

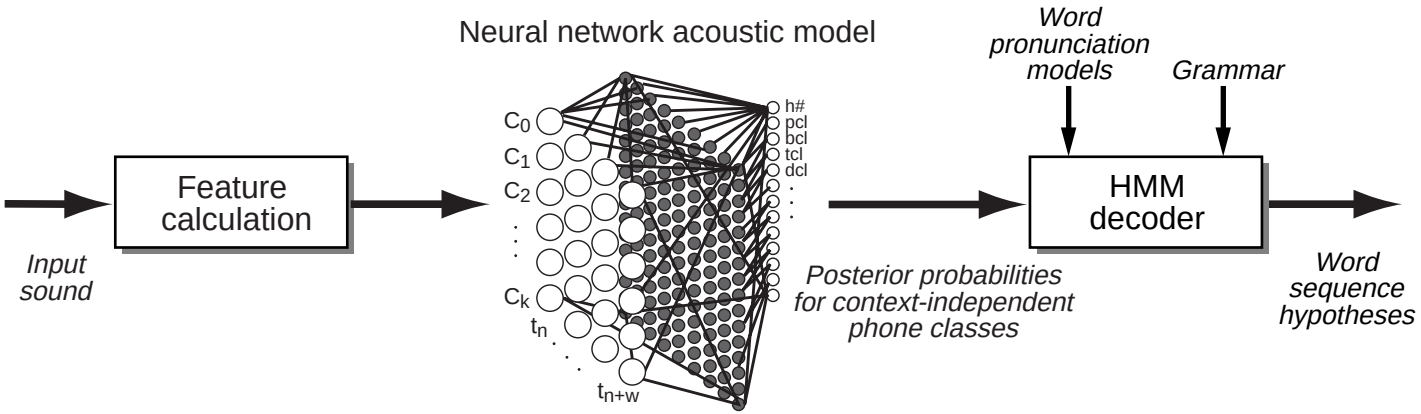
1. About ICSI
2. **Hybrid connectionist-HMM speech recognition**
   - visualizing speech recognition
   - building a recognizer
   - some issues in ASR
3. Overview of current projects
4. Some details
5. Conclusions

2 The Hybrid Connectionist-HMM system

- **Conventional ASR: Symbols *S*, observations *X***

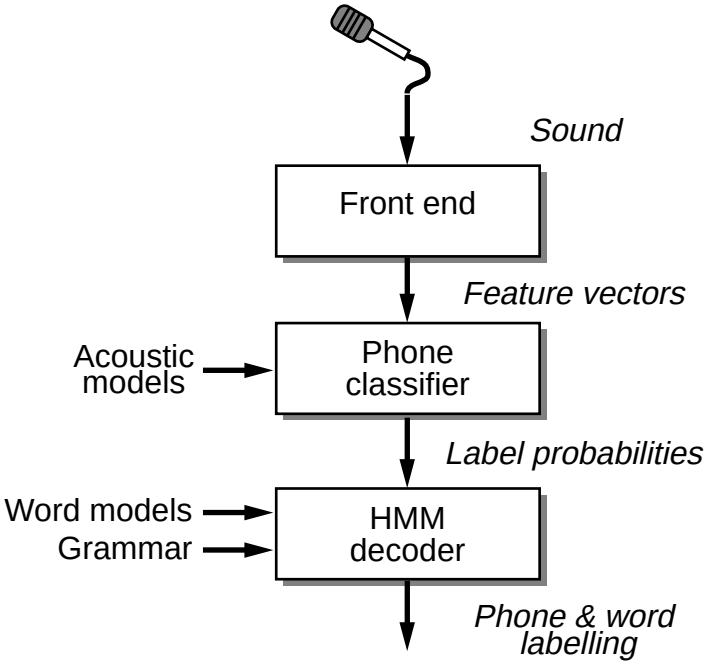
$$S^* = \underset{S}{\operatorname{argmax}}\, P(S|X)$$
$$= \underset{S}{\operatorname{argmax}}\, \frac{P(S,X)}{P(X)}$$
$$= \underset{S}{\operatorname{argmax}}\, \prod_i P(X_i|S_i) \cdot P(S_i|S_{i-1})$$

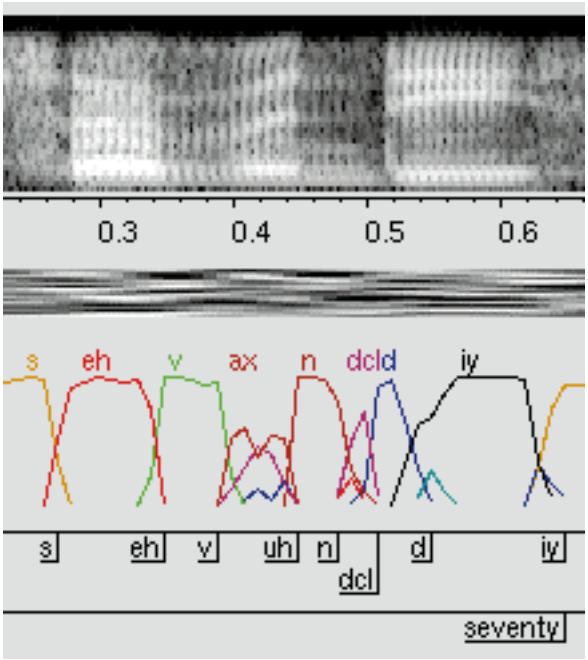
- **$P(X_i|S_i)$ is acoustic *likelihood* model e.g. GMM**
- **Connectionist replaces with *posterior*, $P(S_i|X_i)$:**



Visualizing speech recognition

- **Speech as a sequence of discrete symbols q_i**



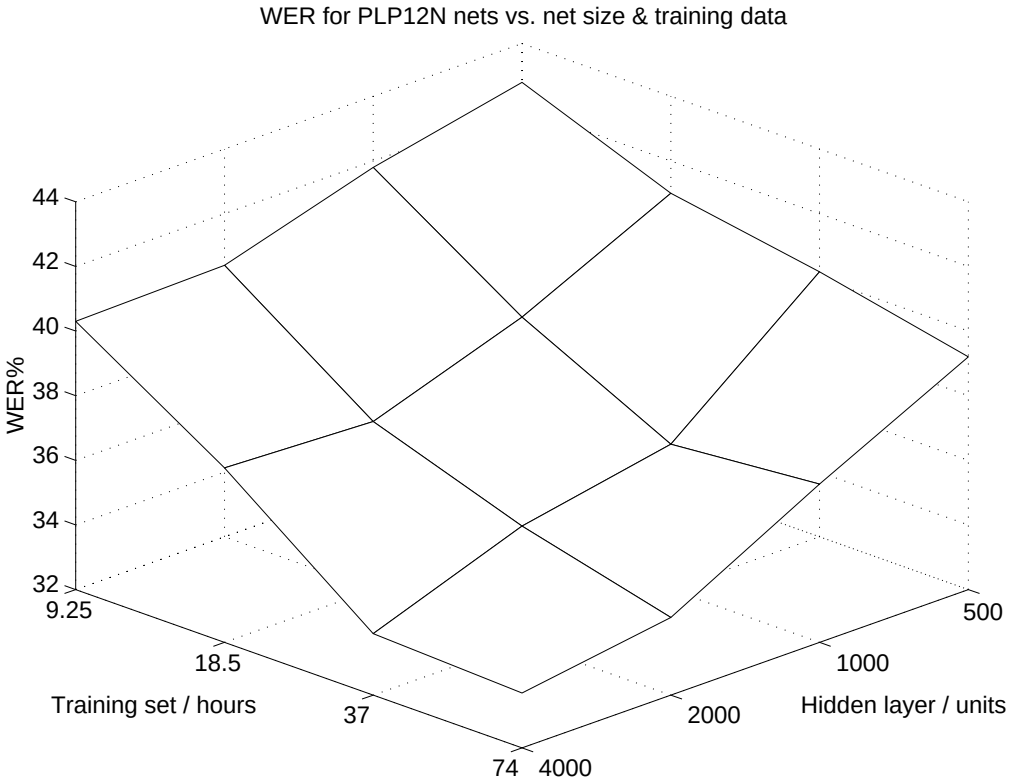


Building a recognizer

- **Define pronunciation models**
  - application vocabulary
  - standard dictionaries + phonetic rules?
- **Build language model**
  - $P(W_i|W_{i-1},...)$
  - count n-grams in example texts?
- **Train acoustic model**
  - choose features to suit conditions
  - train neural net on *large* labeled corpus
  - relabel & retrain?

How much training data?

- **The bulk of recent improvements derives from larger training sets**



- **Largest model: 2.5M parameters, 32M patterns**

  = 24 days to train on TetraSPERT, 10^{15} ops

Some issues in ASR

- **'Spectrogram reading' paradigm**
  - short-time features, concatenative models
- **Goal: classifier accuracy / Word Error Rate (WER)**
  - objective measures, but quite opaque
  - normalization vs. generalization
- **HMM requires search over all word sequences**
  - dominates processing time in large-vocabulary
- **Best solutions (e.g. features) depends on task**
  - RASTA plus delta-features good for small vocab
  - plain normalized PLP best for Broadcast News
  - modulation spectrum features best for combo...
- **Key challenge = Robustness**
  - to: task, acoustic conditions, speaking rate, style, accent ...

Outline

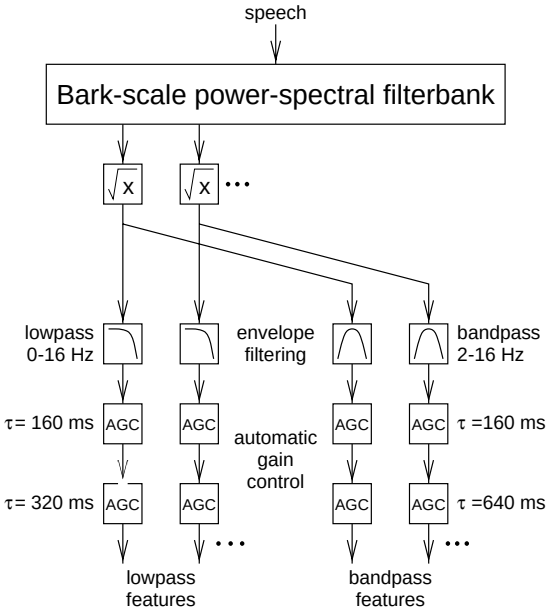
1. About ICSI
2. Hybrid connectionist-HMM speech recognition
3. **Overview of current projects**
   - Front-end features
   - Pronunciation and language modeling
   - Modeling alternatives
   - Other topics
4. Some details
5. Conclusions

The modulation-filtered spectrogram

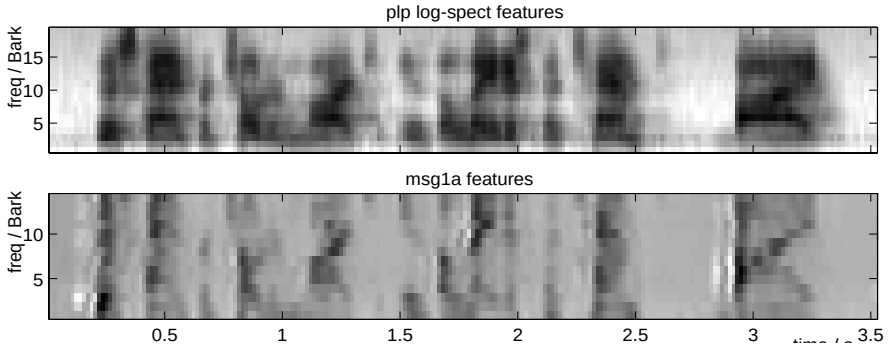
(Brian Kingsbury)

- **Goal: invariance to variable acoustics**

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



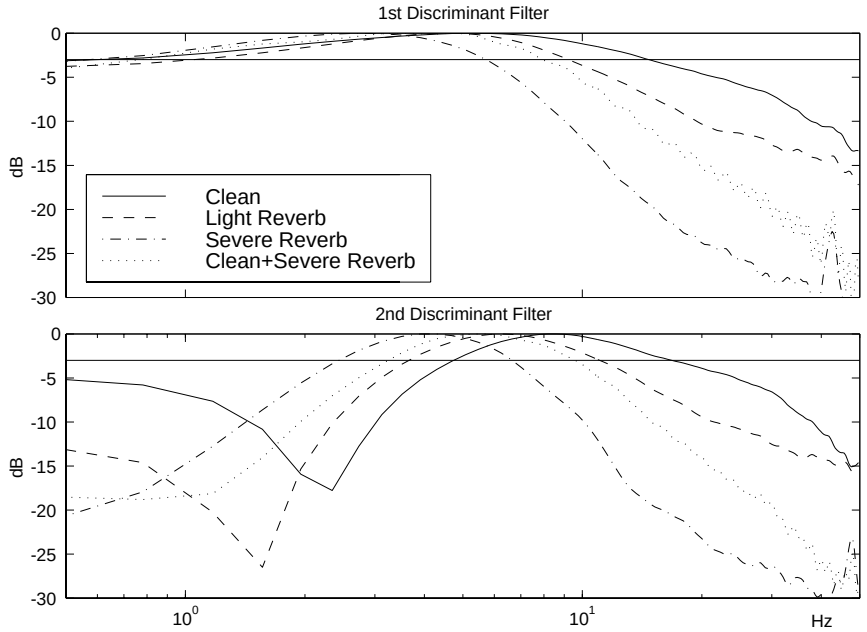
- **Comparison:**



Data-driven feature design

(Mike Shire)

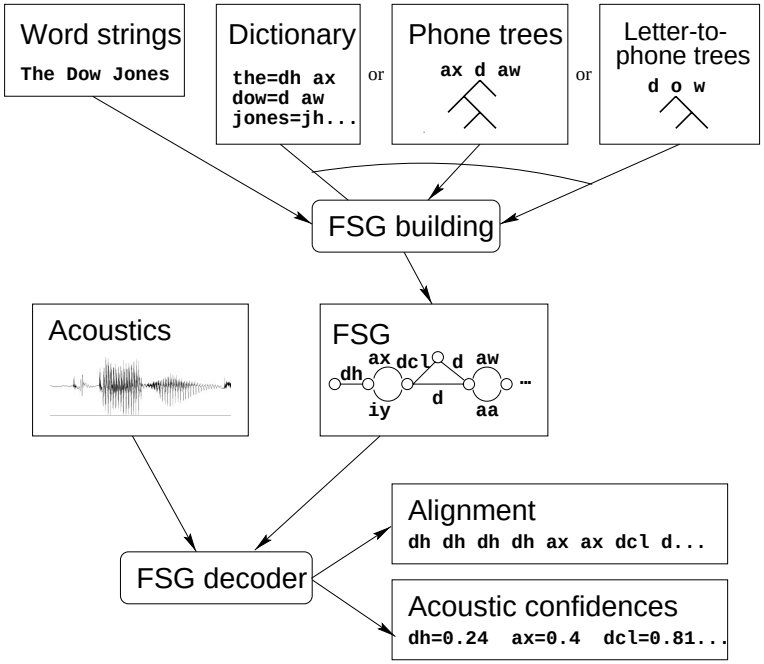
- **'Optimal' features for different conditions**
  - subband envelope domain
  - linear-discriminant analysis (LDA) for filter coeffs
  - separation of labeled classes is optimized
- **Modulation-frequency domain responses for clean, reverb, mixture:**



Automatic pronunciation extraction

(Eric Fosler)

- **Canonical pronunciations are too limited**
- **Phonetic rules overproduce (→ homonyms)**
- **Filter candidates against acoustics:**



- **Decision trees for canonical→real mapping**

Topic modeling (Latent Semantic Analysis)

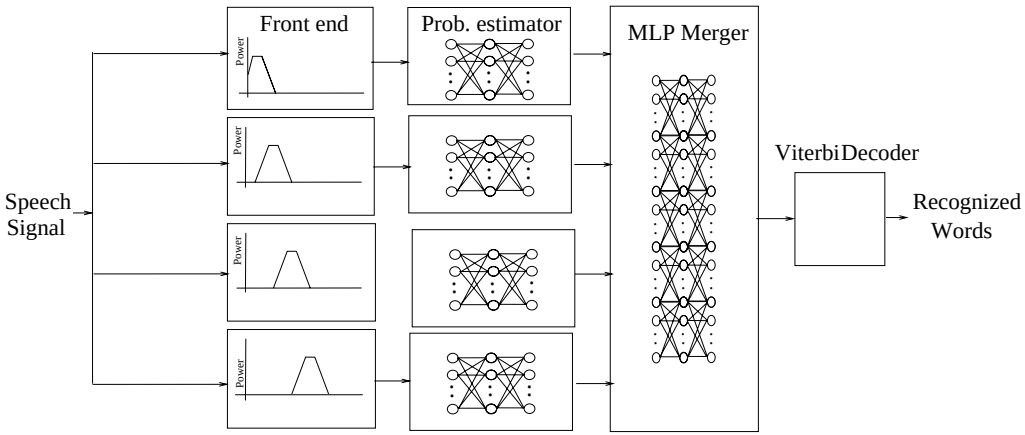
(Dan Gildea & Thomas Hofmann)

- **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*)
  - Trigram language model perplexity of 109 reduced 17%
- **Use for topic segmentation?**

Multiband systems

(Adam Janin / Nikki Mirghafori)

- **Separate recognizers look at different bands**
  - Fletcher/Allen model of human speech recog.
  - noise/corruption in one channel is limited
  - how to combine results?



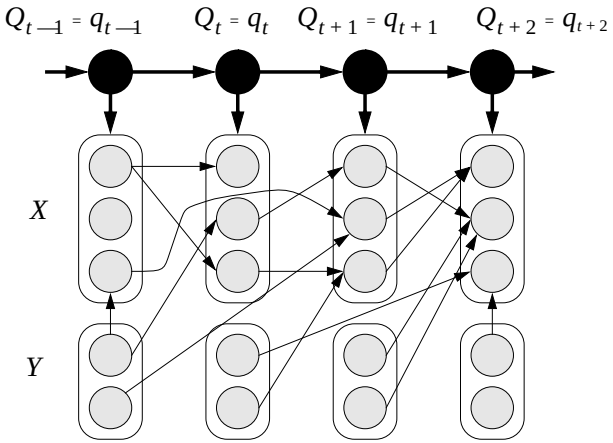
- **Weighted average of all possible combos**
  - $p(S \mid a,b,c,d) = \sum_B p(S \mid B,a,b,c,d) \cdot p(B)$

    B ranges over 16 possible band combinations
  - $p(B)$ from? constant, local feature (entropy)

Buried Markov Models

(Jeff Bilmes)

- **Increasing the scope of the classifier input**
- **Add state-dependent sparse links:**

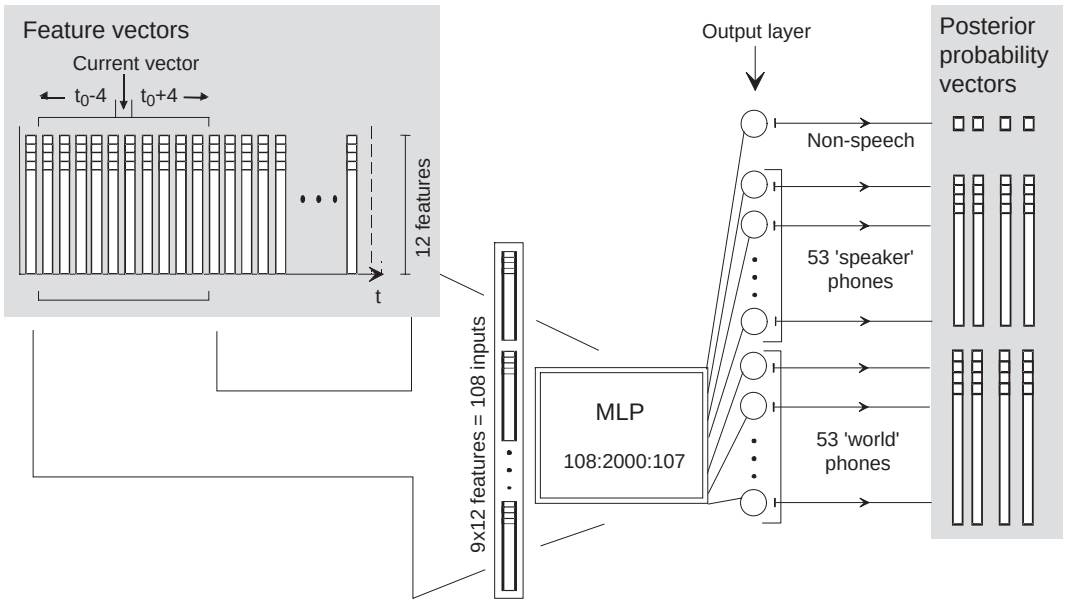


- **How to add links?**
  - maximum conditional mutual information
  - greedy algorithm
- **How to model?**
  - linear dependence as first attempt

Connectionist speaker recognition

(Dominique Genoud)

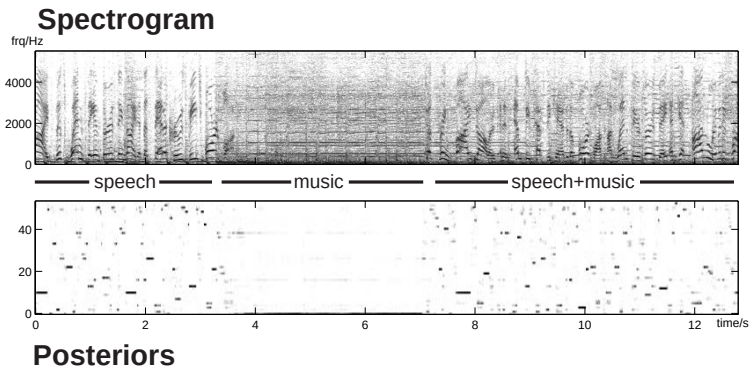
- **Use neural networks to model speakers rather than phones?**
- **Specialize a phone classifier for a particular speaker?**
- **Do both at once for “Twin-output MLP”:**

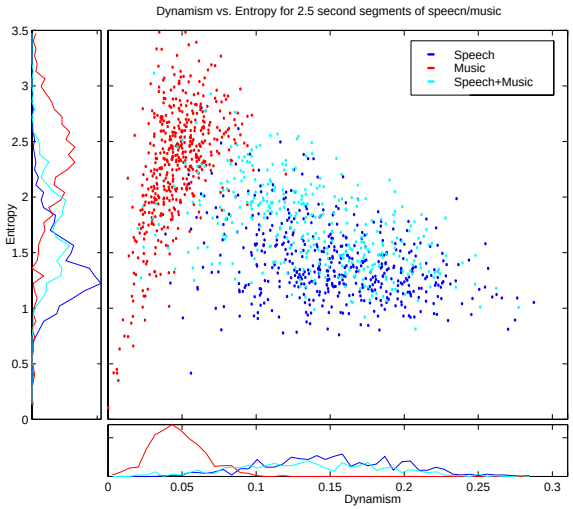


Acoustic Segment Classification

(Gethin Williams)

- **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



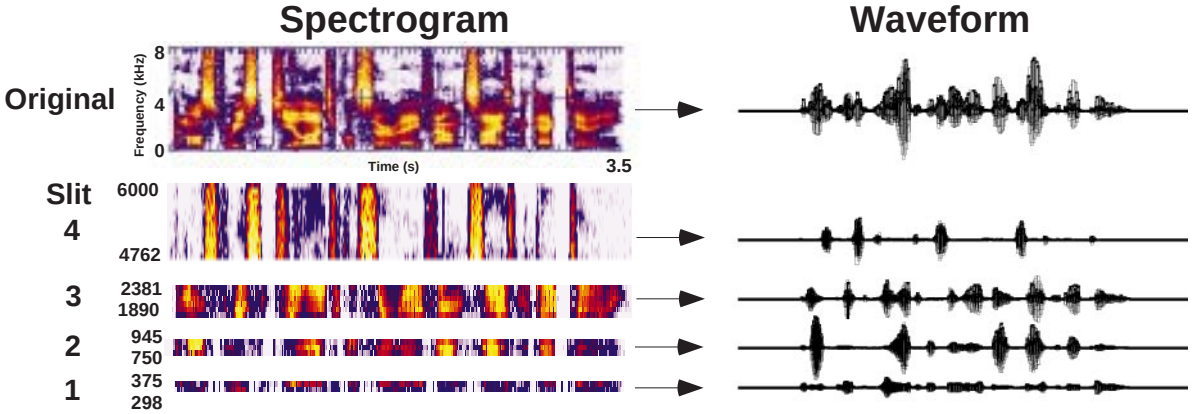


- **1.3% error on 2.5 second speech-music testset**
- **Use for finding segment boundaries?**

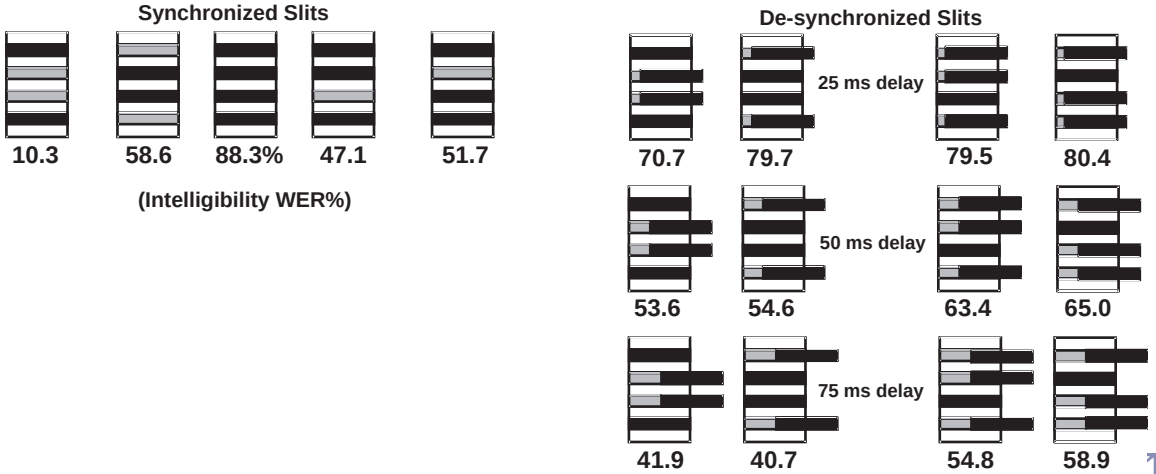
Perceptual experiments

(Steven Greenberg et al.)

- **Use spectrally-sparse speech:**



- **Effect of temporal alignment between bands:**



Outline

1. About ICSI
2. Hybrid connectionist-HMM speech recognition
3. Overview of current projects
4. **Some details:**
   - Information stream combination
   - Broadcast News spoken-document retrieval
   - SpeechCorder PDA
5. Conclusions

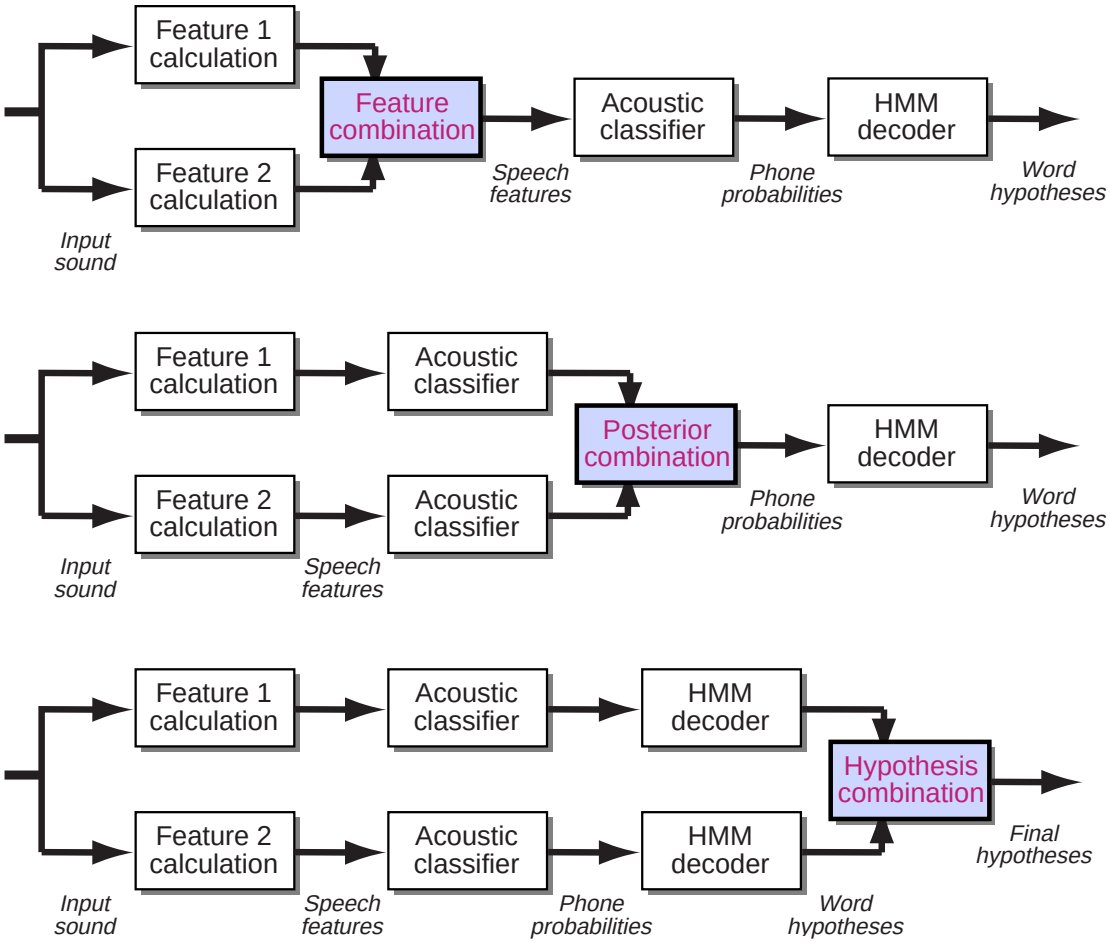
4.1 Information stream combinations

- **Task: AURORA noisy digits**
  - continuous digits
  - test: 4 noise types x 7 SNR levels
  - train: mixed clean/noisy data
- **Feature design evaluation**
  - intermediate representation for mobile phones
  - evaluation specifies GMM-HMM (HTK) system
- **Baseline results:**

Feature	System	WER ratio
mfcc	HTK	100.0%
plp	Hybrid	89.6%
msg	Hybrid	87.1%
msg	HTK	205.0%
msg KG	HTK	184.5%

- **Can we combine features advantageously?**

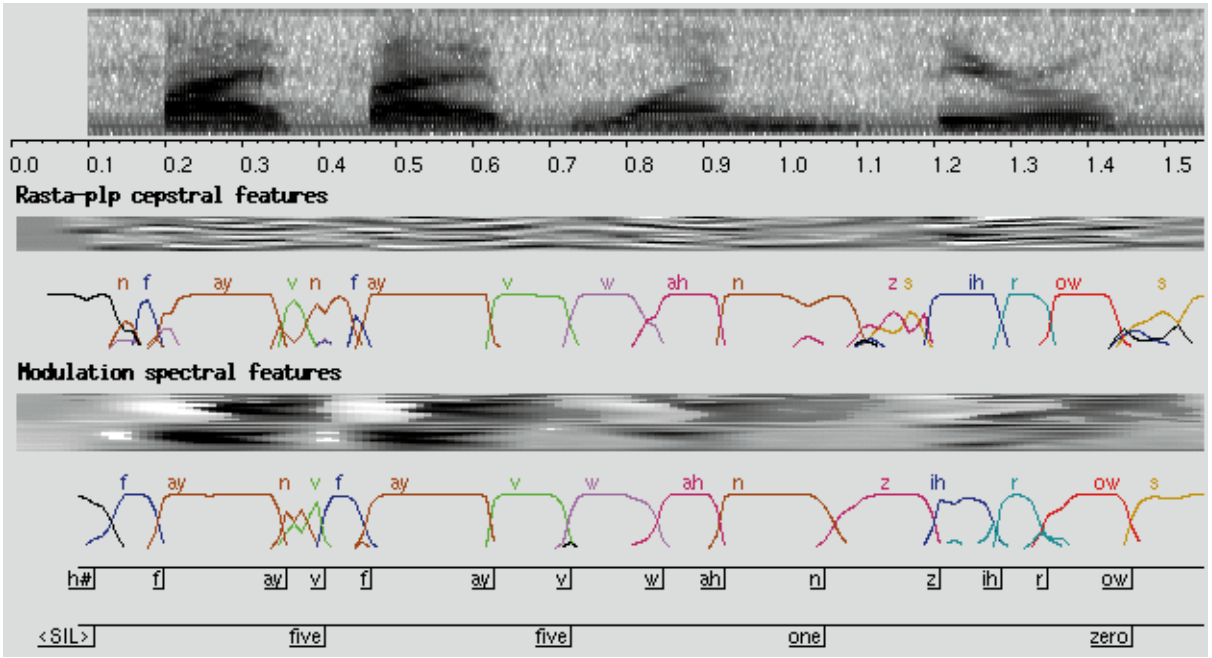
Combination schemes



- **Simple probability combination works best:**

  $P(q_i|X_1,X_2) = P(q_i|X_1) \cdot P(q_i|X_2) \,/\, P(q_i)$ **... if** $X_1 \perp X_2 | q$

Posterior multiplication



Features	System	WER ratio
plp + msg	Feature combo	74.1%
plp + msg	Prob. combo	63.0%
plp + msg	HTK on probs.	51.6%

4.2 Spoken document retrieval

- **Based on DARPA/NIST Broadcast News**
- **Training material recorded off-air**
  - ABC, CNN, CSPAN, NPR
  - 200 hour training set (TREC: 550 hour archive)
  - training:
    word transcriptions + speaker time boundaries
- **Best WER results:**
  - 1996: HTK: 27%
  - 1997: HTK: 16% (but: easier; 22% on 1996 eval)
  - 1998: LIMSI: 14% (SPRACH: 21%)
- **Some clear conclusions**
  - one classifier for all conditions (or male/female)
  - feature adaptation (VTLN, MLLR, SAT)
  - importance of segmentation
  - more data is useful

Applications for BN systems

- **Live transcription**
  - subtitles
  - transcripts
  - but: more than words?
- **Video editing**
  - precision word-time alignments
  - commercial systems by IBM, Virage, etc.
- **Information Retrieval (IR)**
  - TREC/MUC ‘spoken documents’
  - tolerant of word error rate, e.g.:

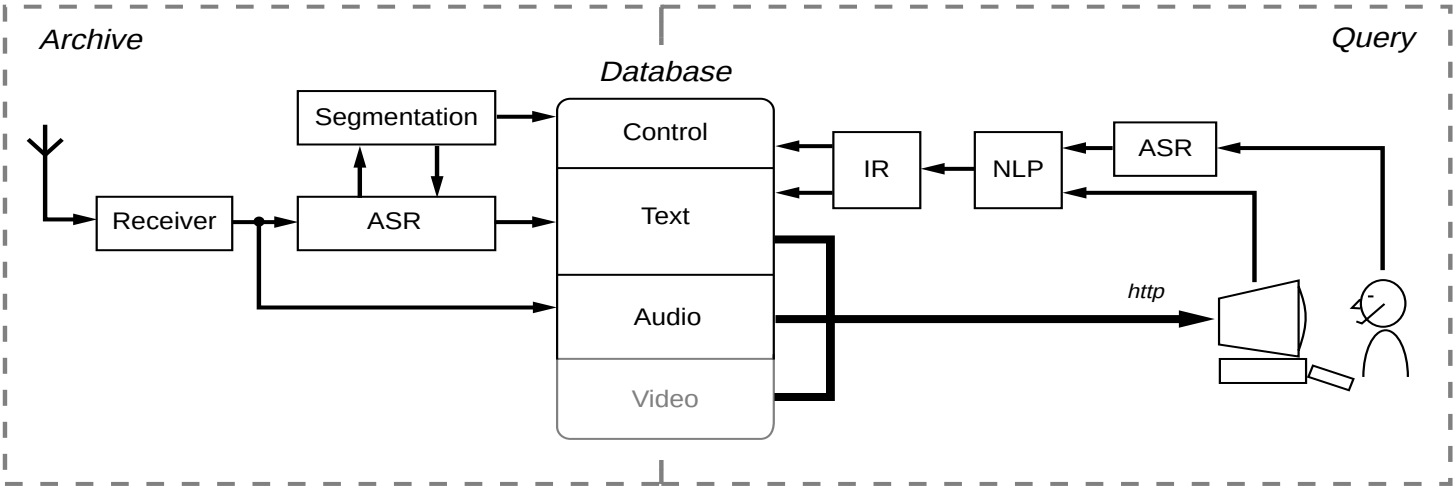
F0: THE VERY EARLY RETURNS OF THE NICARAGUAN PRESIDENTIAL ELECTION SEEMED TO FADE BEFORE THE LOCAL MAYOR ON A LOT OF LAW

F4: AT THIS STAGE OF THE ACCOUNTING FOR SEVENTY SCOTCH ONE LEADER DANIEL ORTEGA IS IN SECOND PLACE THERE WERE TWENTY THREE PRESIDENTIAL CANDIDATES OF THE ELECTION

F5: THE LABOR MIGHT DO WELL TO REMEMBER THE LOST A MAJOR EPISODE OF TRANSATLANTIC CONNECT TO A CORPORATION IN BOTH CONSERVATIVE PARTY OFFICIALS FROM BRITAIN GOING TO WASHINGTON THEY WENT TO WOOD BUYS GEORGE BUSH ON HOW TO WIN A SECOND TO NONE IN LONDON THIS IS STEPHEN BEARD FOR MARKETPLACE

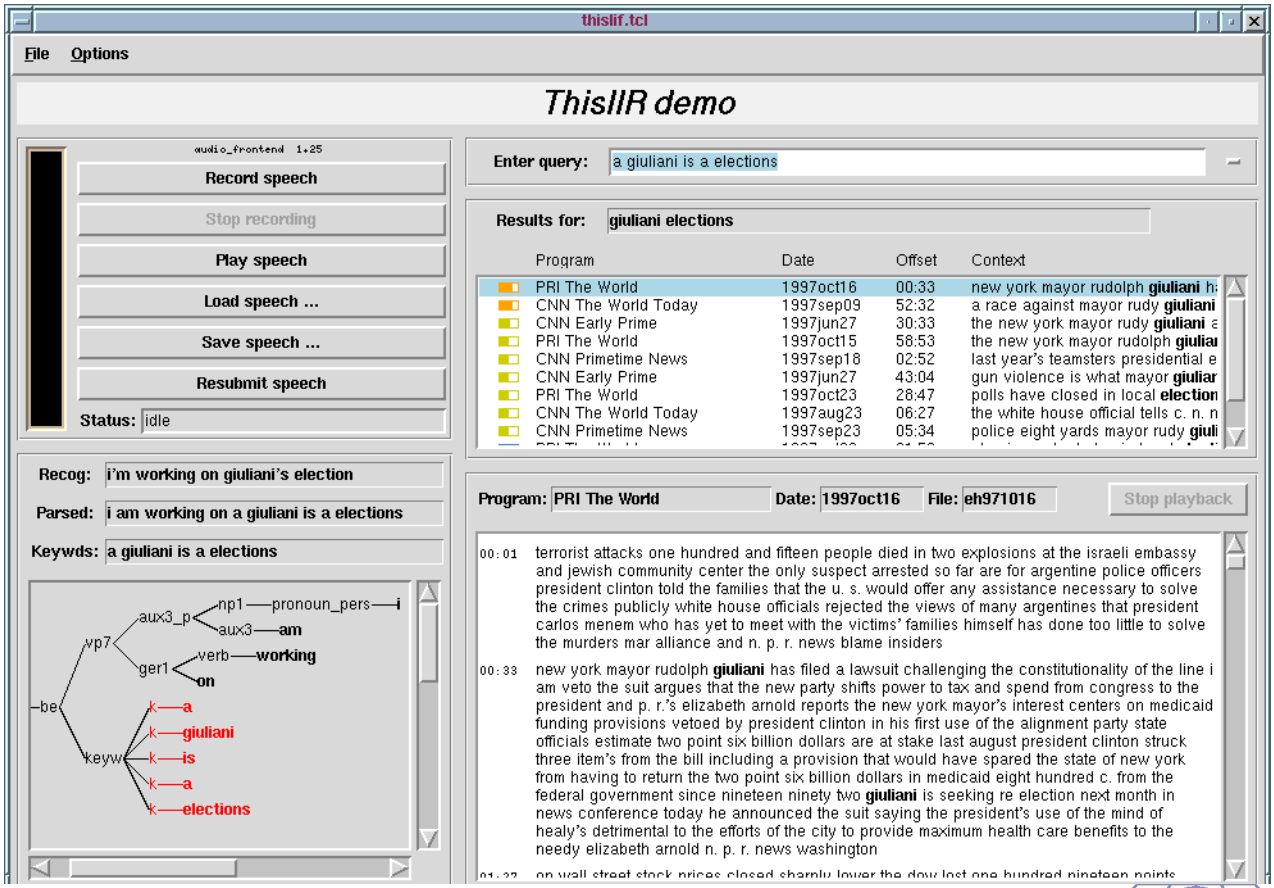
Thematic Indexing of Spoken Language (Thisl)

- **EC collaboration, BBC providing data**
- **1000+ hr archive data**
- **IR is key factor**
  - stop lists
  - weighting schemes
  - query expansion



ThislGui

- **Tcl/Tk front-end to Thisl IR engine**
- **Spoken query input:** SPRACHdemo/AbbotDemo
- **NLP integration:** prolog lattice parser

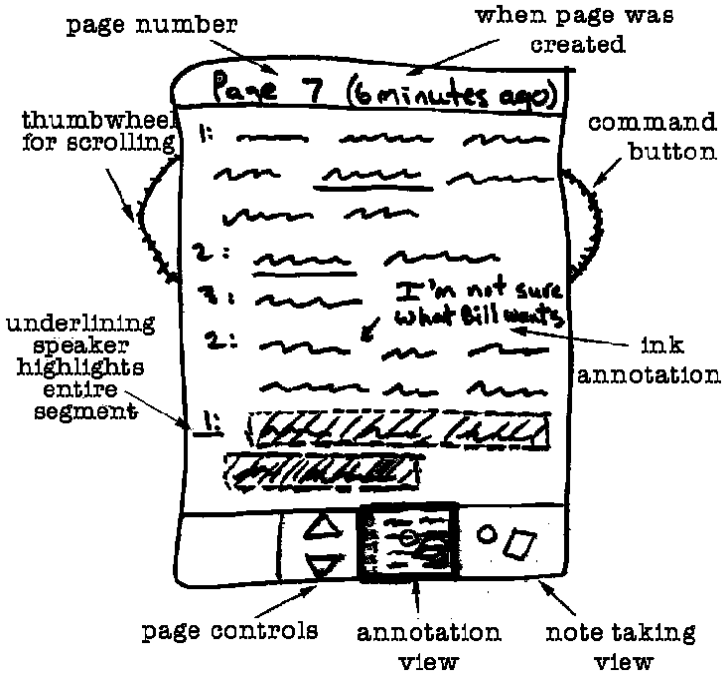


4.3 SpeechCorder

- **Convergence of interesting problems:**
  - ubiquitous PDAs
  - multimedia processing
  - very fast, low-power CPU design
  - resource-bound speech recognition
- **ICSI / UC Berkeley / MIT collaboration**
  - ICSI: speech & audio processing
  - UCB: user interface design
  - MIT: new low-power CPUs
- **Current proposal**
  - PDA
  - meeting / memo recorder
  - .. also for sociological study?
    (new Human Centered Computing consortium)

The SpeechCorder GUI

- **Live annotation of recognized speech**



- **Application issues:**
  - correcting errorful transcriptions
  - finding places in the recording
  - annotations (speaker, notes, emphasis)

SpeechCorder: Audio

- **Speech recognition**
  - not close-mic'd
  - speaker ID
  - low-power / low-memory / vectorized
  - non-local processing for 'best' transcript?
- **Other audio issues**
  - identifying speech versus nonspeech
  - finding / indexing nonspeech events...

Outline

1. About ICSI
2. Hybrid connectionist-HMM speech recognition
3. Overview of current projects
4. Some details
5. **Conclusions**
   - more criticisms of ASR
   - future work

5 Conclusions

- **The downside of objective evaluation**
  - research priority has been pragmatic goal: reduce WER
  - human speech recog. uses many constraints
  - grammatic/semantic constraints implicit in word sequence statistics (grammar)
  - automatic analysis of large corpora is possible & helpful
- **The problems with a grammar**
  - unexpected (unseen) phrases are discounted
  - highly brittle alternatives
  - masks underlying performance
- **A more scientific approach**
  - first work on the underlying phoneme classifier
  - follow nonsense syllable performance (Fletcher)

The signal model in speech recognition

- **Systems & approach have been optimized for speech-alone situation**
  - minimize classifier parameters, maximize use of 'feature space'
  - e.g. cepstra [example]
- **Possibly non-lexical data thrown away**
  - pitch
  - timing/rhythm
  - speaker identification
- **Dire consequences**
  - .. dealing with nonspeech sounds
  - .. distinguishing success & failure
- **Popular focus of research**
  - e.g. segmental models, pitch features
  - fail to obtain robust improvements

Future work

- **Continue improving robustness**
  - better features
  - better pronunciations
  - better modeling
- **Still looking for a good architecture**
  - multiband
  - multistream
  - more adaptation
  - more contextual dependence
- **Speech recognition: useful for applications**
  - archive indexing, summarizing
  - personal devices, new interfaces
  - tie-in to general audio analysis...