Speech Recognition at ICSI: Broadcast News and beyond

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

Outline

1. **The DARPA 'Broadcast News' task**
2. **Aspects of ICSI's BN system**
3. **Future directions for speech recognition**

1 DARPA 'Broadcast News'

- **DARPA standard speech tasks**
  - Resource Management (1980s)
  - Wall Street Journal (early 1990s)
  - Broadcast News (1996 on)
  - Switchboard (1996 on)
  - Call Home (1997 on)
- **Distinguishing features**
  - vocabulary size, grammar perplexity
  - speaking style: read, spontaneous, familiar
  - acoustic conditions, variability
  - accent, dialect, language
- **Annual evaluation 'bakeoffs'**
  - unseen common evaluation set
  - key result is overall Word Error Rate

Broadcast News details

- **Training material recorded off-air**
  - ABC, CNN, CSPAN, NPR
  - 50 hours for 1996, 1997 +50h, 1998 +100h
  - word transcriptions + speaker time boundaries
  - excluding commercials $\rightarrow$ 74 h training set
- **7-way acoustic condition classification**
  - F0: prepared studio speech (~40%)
  - F1: spontaneous studio speech (20%)
  - F2: telephone-bandwidth (20%)
  - F3: background music (5%)
  - F4: degraded acoustics (5%)
  - F5: foreign accents (5%)
  - Fx: combinations/other (5%)

Broadcast News history

- **Best WER results:**
  - 1996: HTK: 27%
  - 1997: HTK: 16% (but: easier; 22% on 1996 eval)
  - 1998: November
- **Some clear conclusions**
  - one classifier for all conditions (or male/female)
  - feature adaptation (VTLN, MLLR, SAT)
  - importance of segmentation
  - hard to improve grammar
  - 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.:

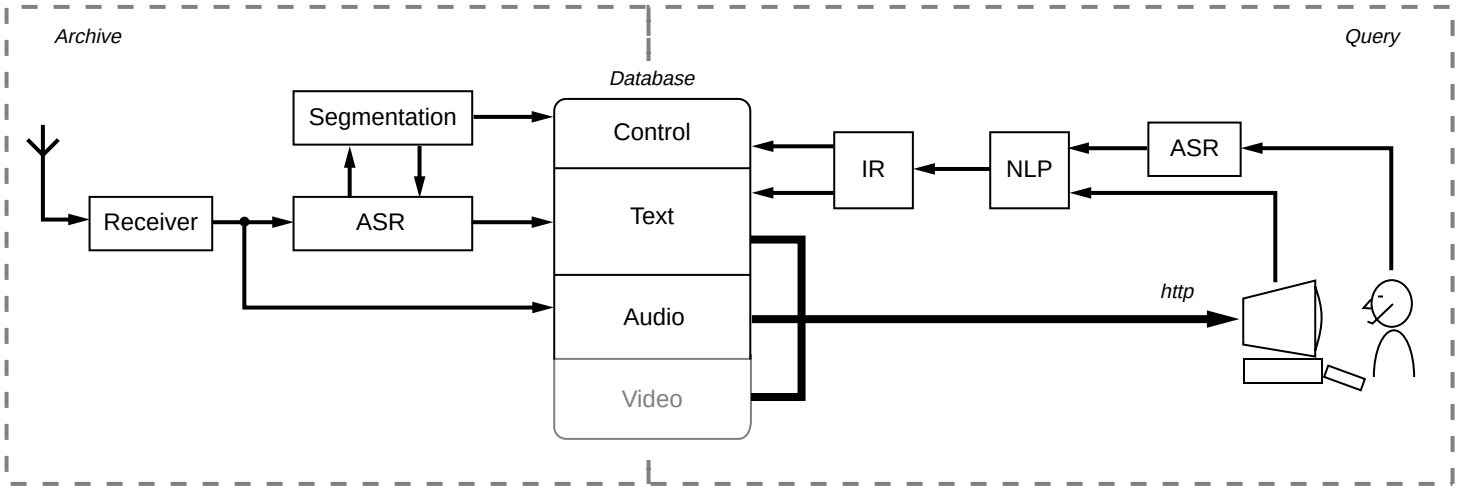
F0: THE VERY EARLY RETURNS OF THE NICARAGUAN PRESIDENTIAL ELECTION SEEMED TO FADE BEFORE THE LOCAL MAYOR ON A LOT OF LAW

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**
- **> 500 hr archive data**
- **IR is key factor**
  - stop lists
  - weighting schemes
  - query expansion



Outline

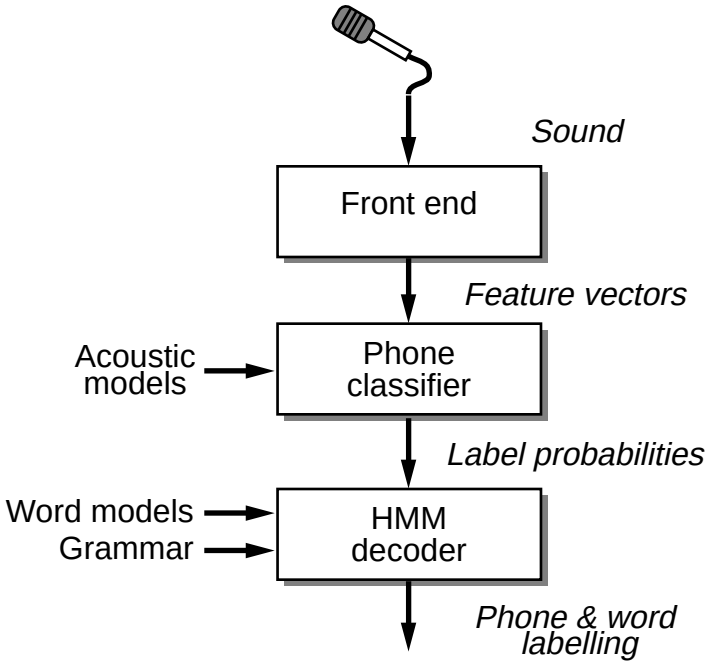
1. **The DARPA ‘Broadcast News’ task**

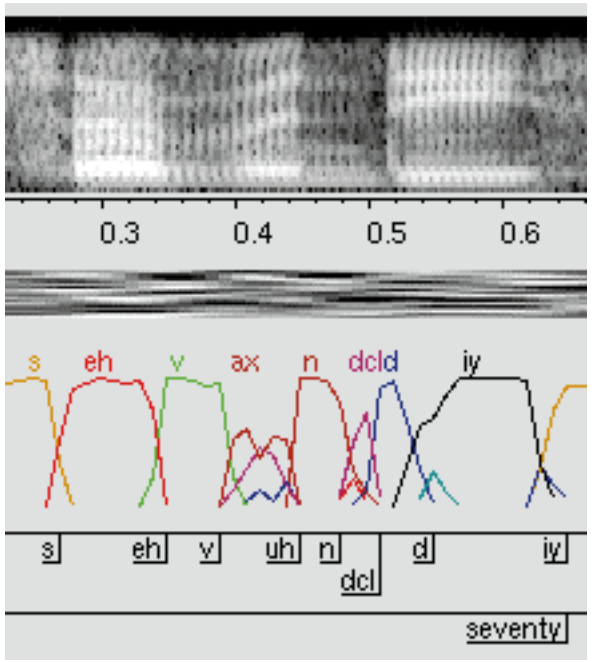
2. **Aspects of ICSI’s BN system**
   - the standard speech recognition architecture
   - front-end, classifier & HMM decoder issues
   - adaptation & segmentation
   - lessons: ‘size matters’

3. **Future directions for speech recognition**

Standard speech recognition

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



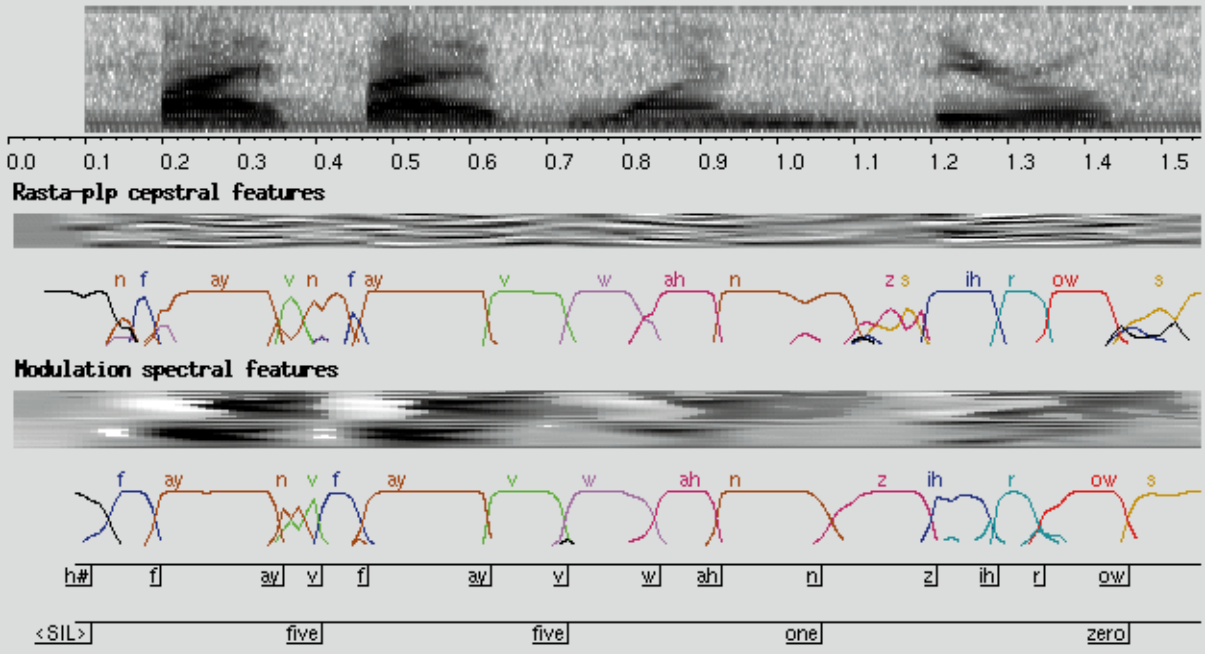


Front-end issues

- **'Spectrogram reading' paradigm**
  - short-time spectral features
  - (perceptual) frequency-warping helps
  - normalization e.g. RASTA
- **Goal = classifier accuracy**
  - objective measure, but quite opaque
  - the right space for generalization
  - tension between detail & blurring
- **Best solution depends on task**
  - RASTA plus delta-features good for small vocab
  - plain normalized PLP best for BN
  - modulation spectrum features best for combo...
- **Normalizing...**
  - ... in training
  - ... unseen speech

Classifier issues

- **Find $p(q_i|X)$**
  - directly by (discriminant) neural-net estimation
  - by likelihood i.e. model $p(X|q_i)$ with Gaussians
  - more data permits finer detail in q_i
- **Combining classifiers helps:**



HMM decoder issues

- **Define all allowable output q_i sequences**
  - phone models
  - word pronunciations (lexicon)
  - word sequences (grammar)
- **Search for best matching sequence**
  - dominates processing time in large-vocab systems
  - variation of pronunciation with speaking rate
  - data-derived pronunciations
  - handling poor acoustics

Adaptation, segmentation & confidence

- **Big gains from adaptation & normalization**
  - e.g. VTLN, MLLR
  - typ. 10-20% relative WER improvement
- **Requires marking of homogeneous segments**
  - hand-labelled
  - ‘rate of change’ metric for automatic boundaries
  - clustering models for segments
- **Confidence metrics**
  - typically elusive
  - help indicate errors
  - help to segment material
  - conserve decoding effort
- **$p(q_i|X)$ should correlate with confidence**

Status of the ICSI BN project

- **WER:**
  - started out (April) ~ 50%
  - best single net ~ 33%
  - best combination ~26%
- **'Size matters'**
  - biggest gain from large classifiers & lots of data
  - e.g. 200k parameters, 4M patterns = 40%
    800k parameters, 16M patterns = 33%
  - training time = 11days (special hardware)
  - (other approaches reach similar conclusion)
- **Innovations**
  - combinations
  - multiband?
  - segmental features?
  - time windows?

Outline

1. The DARPA ‘Broadcast News’ task
2. Aspects of ICSI’s BN system
3. **Future directions for speech recognition**
   - removing the ‘grammar crutch’
   - the signal model & what is thrown away
   - a research agenda

The crutch of grammar

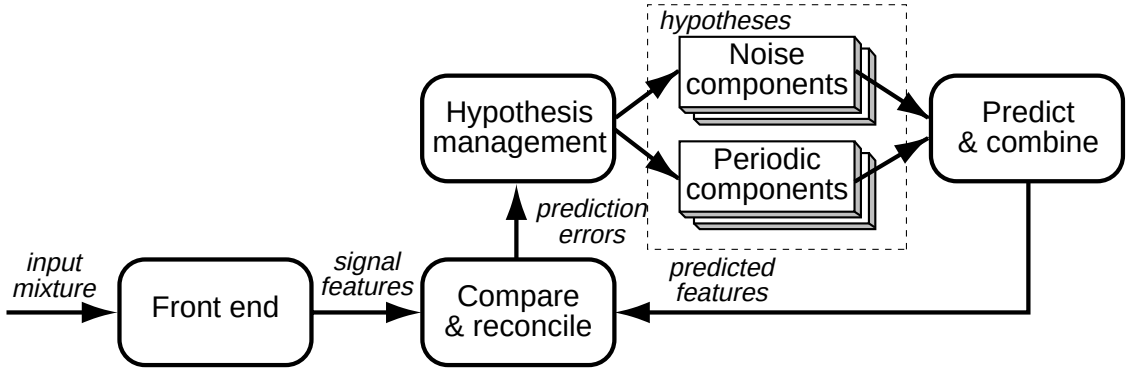
- **The downside of objective evaluation**
  - research priority has been pragmatic goal of reducing WER
  - human speech recognition results from many constraints
  - grammatic/semantic constraints implicit in word sequence statistics (grammar)
  - automatic analysis of large corpora is possible & helpful
- **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

The prediction-driven approach

- **Originally for non-speech auditory scene analysis**
- **Analysis-by-synthesis model**
  - representation is generative parameters
  - analysis is search & tracking of models



Prediction-driven analysis of speech/nonspeech mixtures

- **Speech just another class of models...**
- **Account for all (speech) perceptual features**
  - phoneme identity
  - speaker identity
  - speaking rate & style
- **Informed by speech coding & synthesis**
- **Problem: efficiency of analysis**
  - currently: direct evaluation of label likelihoods, search over discrete lexical space
  - proposed: implies search of continuous speech-quality space

Conclusions

- **Broadcast News: interesting task**
- **ICSI's BN system: useful framework**
  - significant 'infrastructure investment'
  - large, well-known, interesting, real problem
  - carries implicit research priorities
- **'Sore thumbs' in current speech recognition & some research directions**
  - separating the effects of different constraints (acoustic model & language model)
  - signal models that can incorporate nonspeech
  - track all perceptual attributes, don't just discard them