

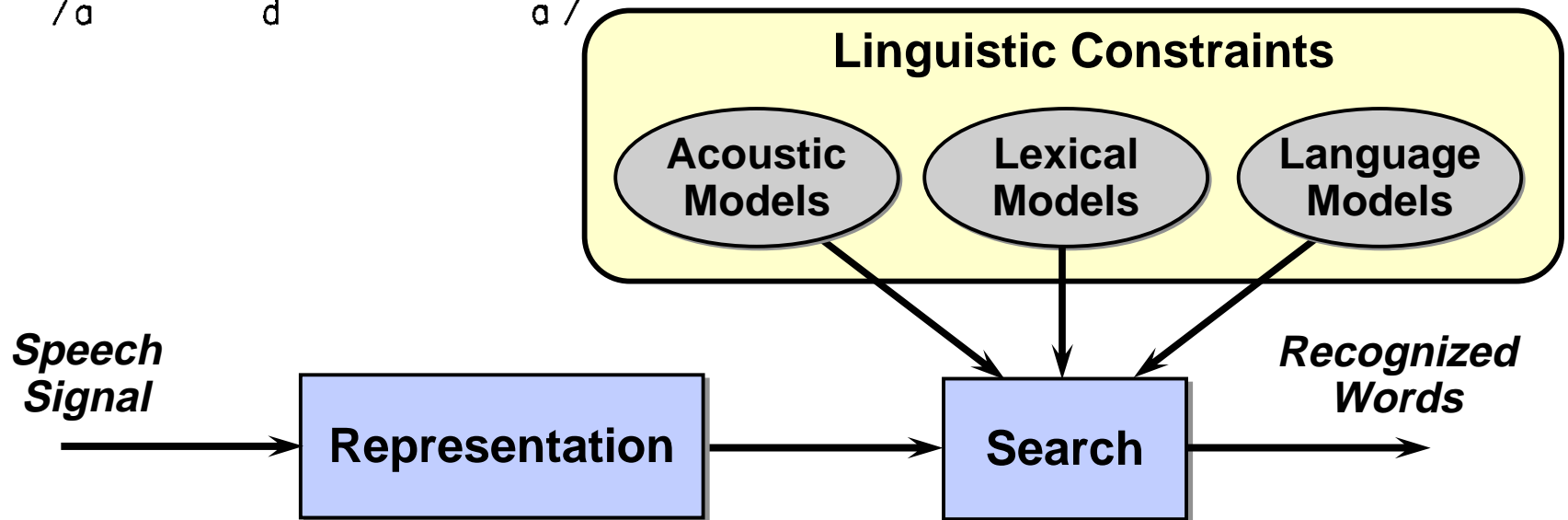
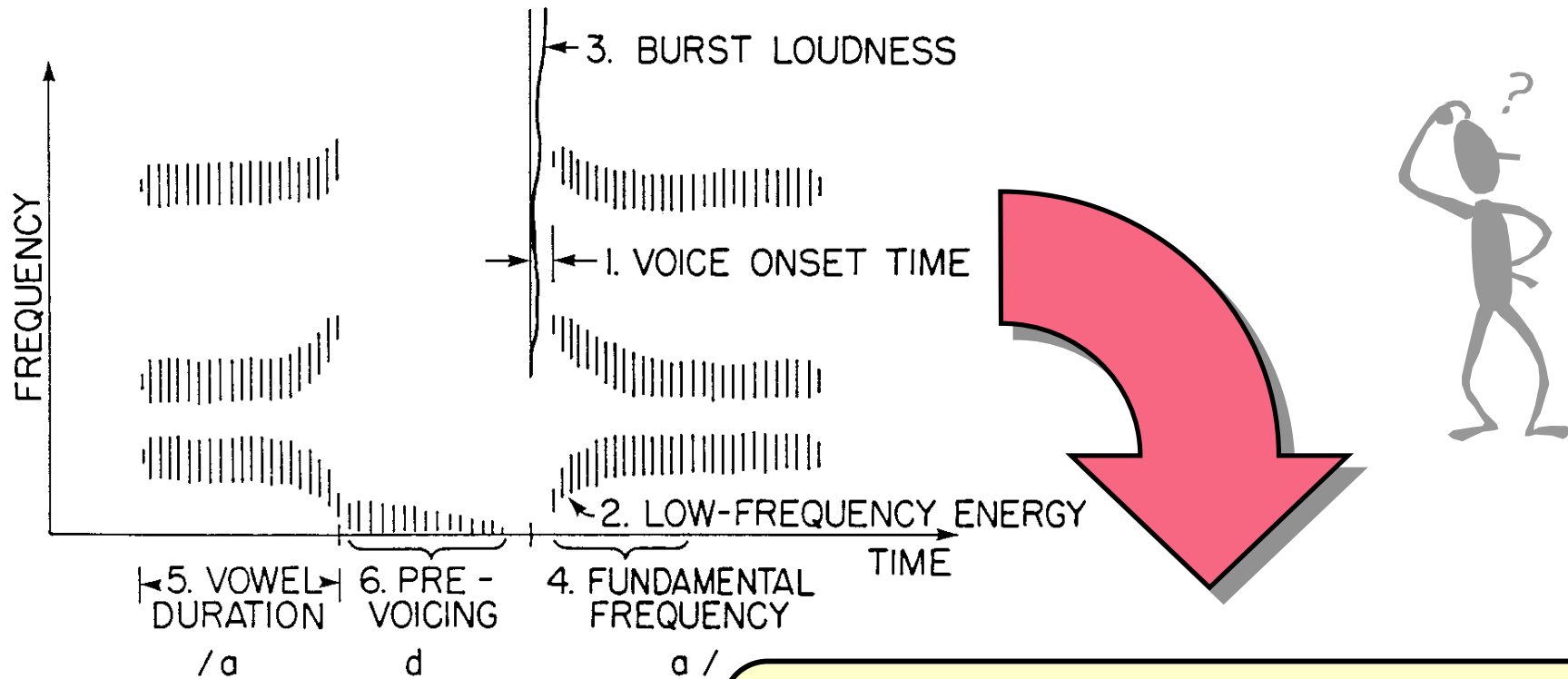
Putting Speech into Speech Recognition

Jim Glass (glass@mit.edu)

**Computer Science and Artificial Intelligence Laboratory
MIT**

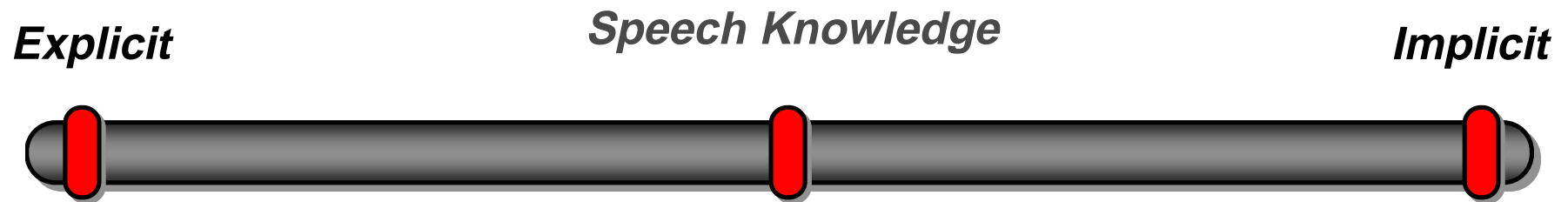
**NSF Symposium on Next Generation ASR
October 8, 2003**

Speech Science meets Speech Technology



Speech knowledge integration?

- Automatic speech recognizers differ in the degree with which speech knowledge is incorporated into the system

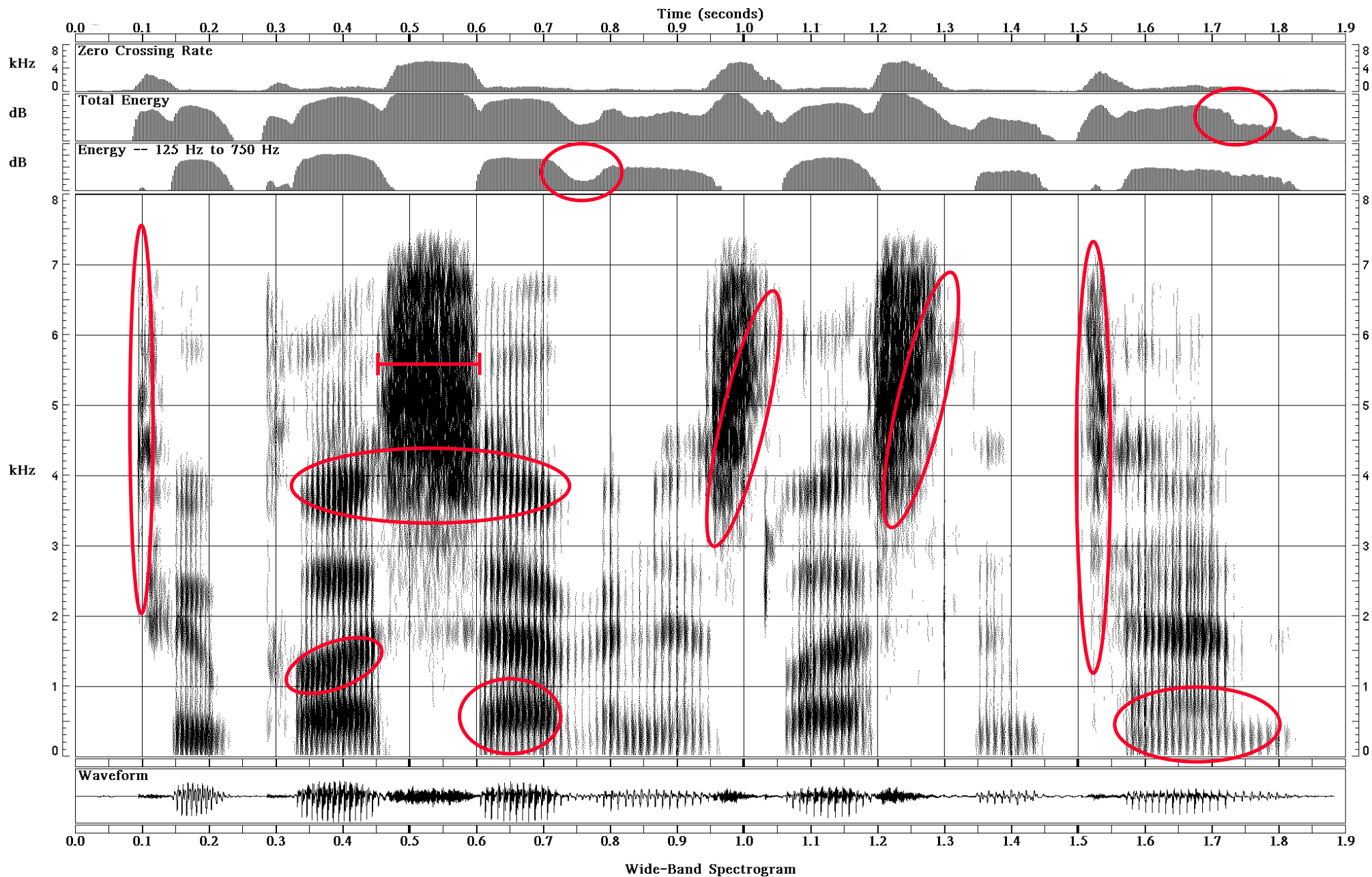


- Heuristic, rule-based models
- Heterogeneous and complex knowledge representation
- Intense knowledge engineering

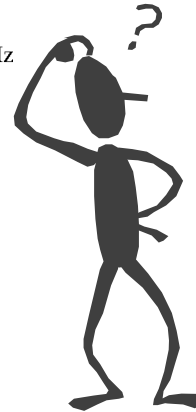
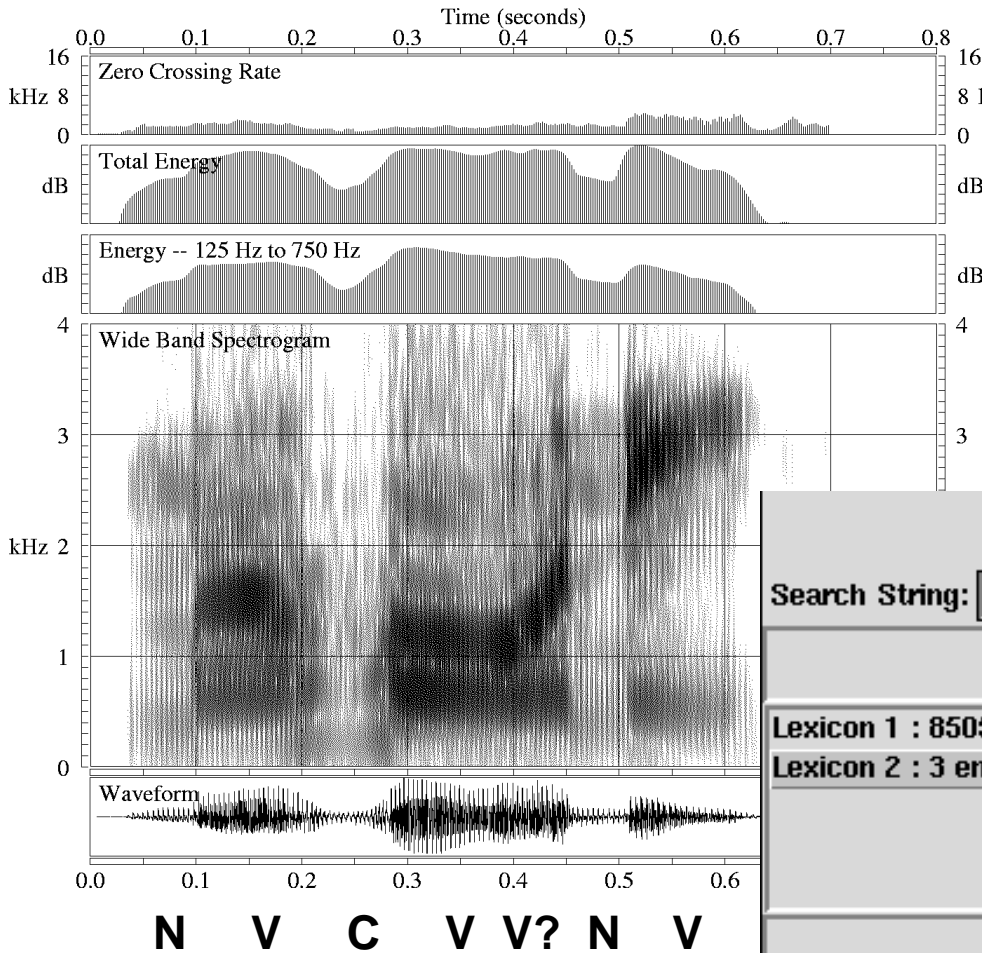
- Formal mathematical models
- Homogeneous and simple knowledge representations
- Automatic learning from data

- Is there a middle ground between these approaches?

An Abundance of Acoustic-Phonetic Cues



Spectrogram Reading Experiments



“lionhearted”
 “marauding”
 “midmorning”
 “immortal”
 “memorials”



Crystal Query

Search String:

Lexicon History

Lexicon 1 : 8505 entries

Lexicon 2 : 3 entries (PHONEMIC search for "NVCVV?NV") searched from Lexicon

Search Results

Spelling

Pronunciation

mahoney

məhoni

~~mckinney~~

~~məkini~~

~~newcomer~~

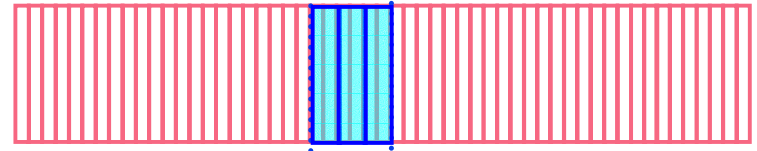
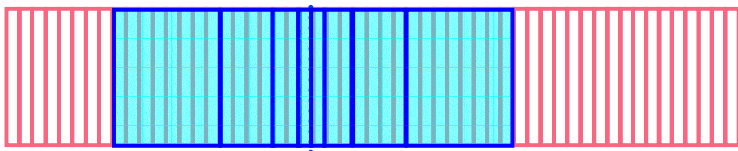
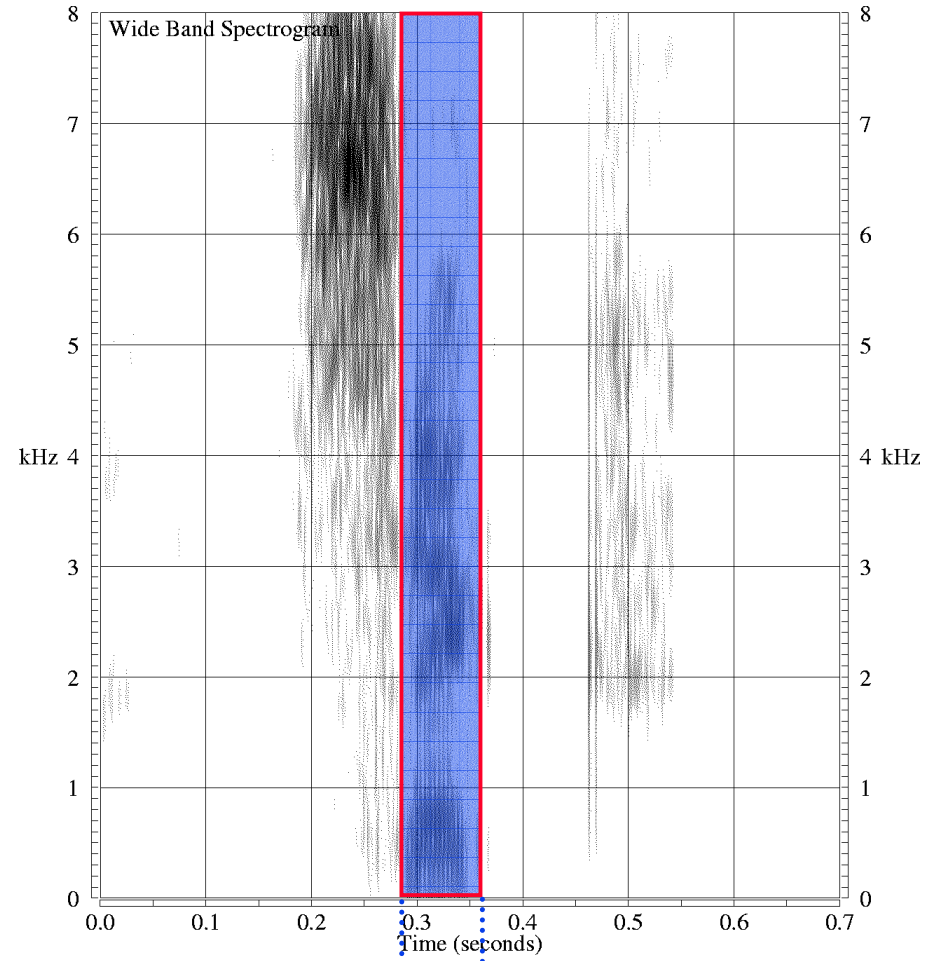
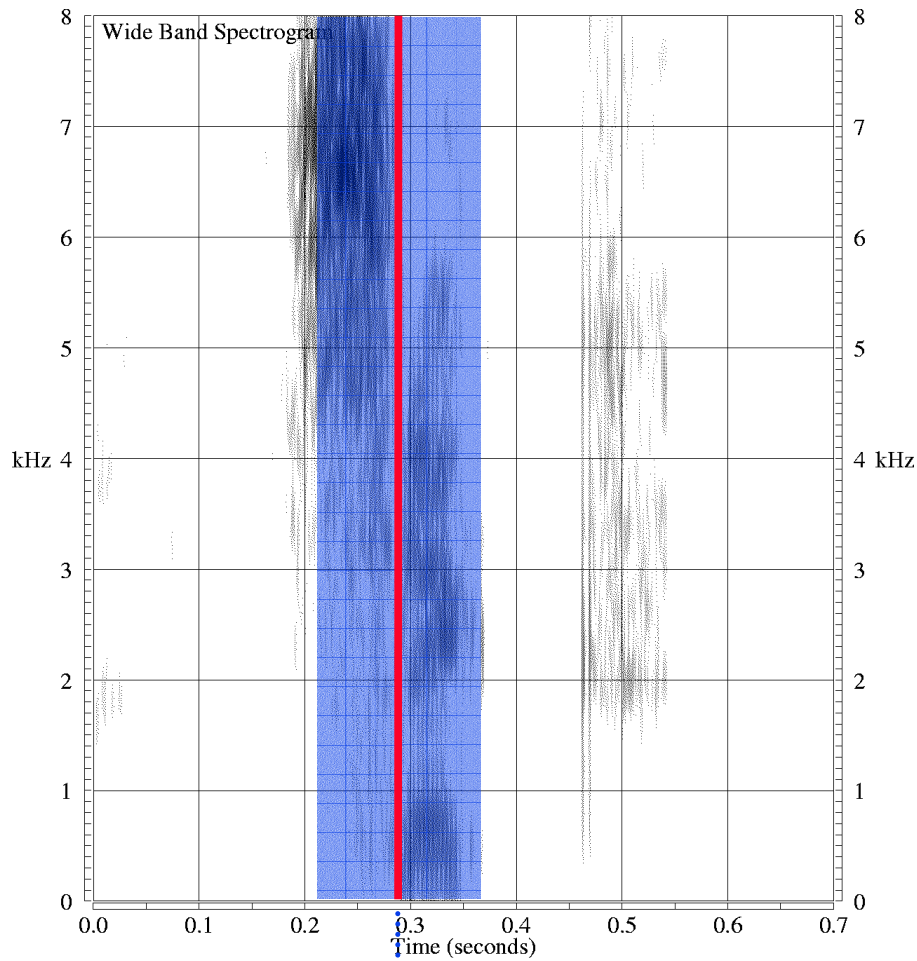
~~nukamə~~



Proposed research areas

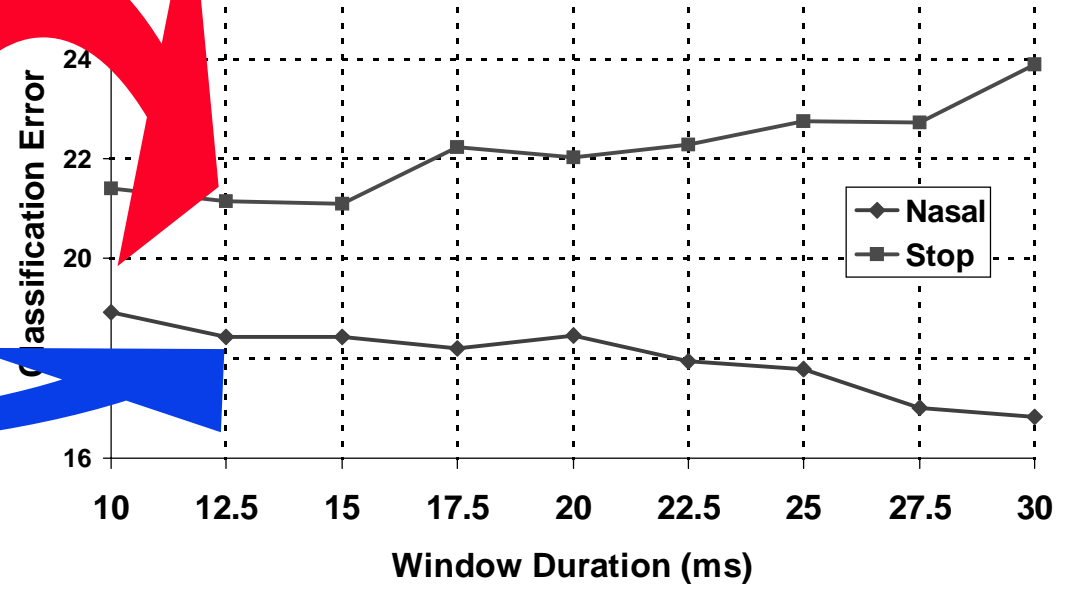
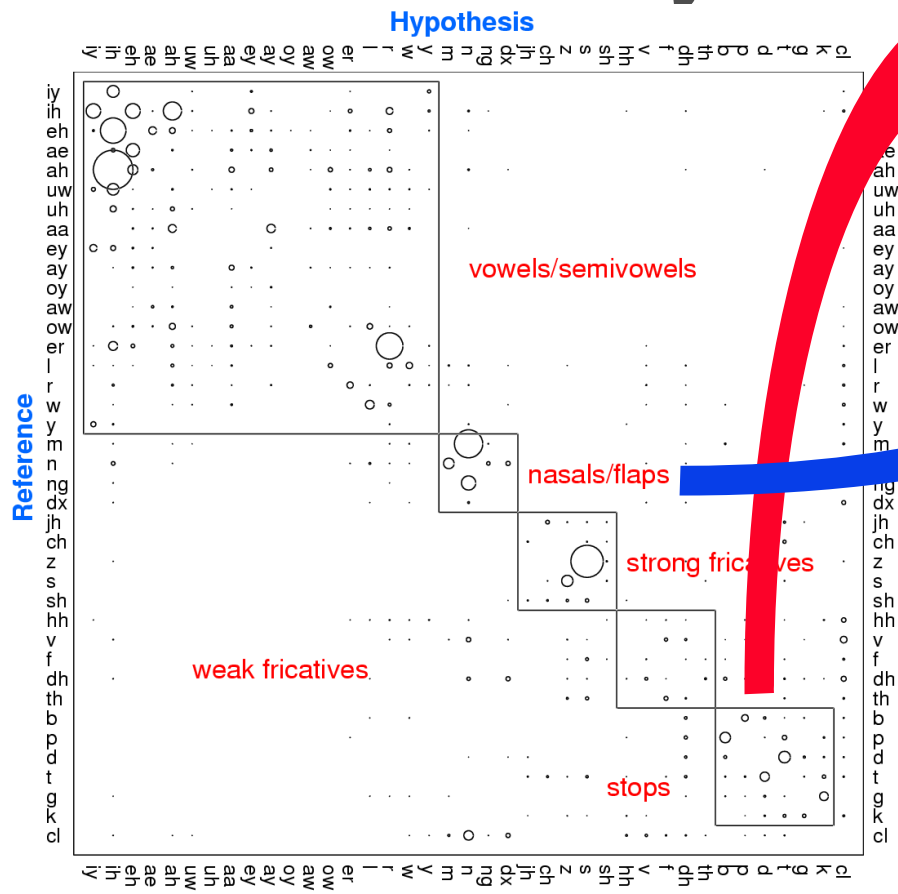
- **Potential areas for speech knowledge integration:**
 - Beyond frames?
 - Beyond homogeneity?
 - Beyond phonetics?
 - Beyond phonemes?
 - Beyond words?
 - Beyond WER?
 - Beyond audio?
 - Beyond speech?
 - Beyond tasks?

Beyond frames?

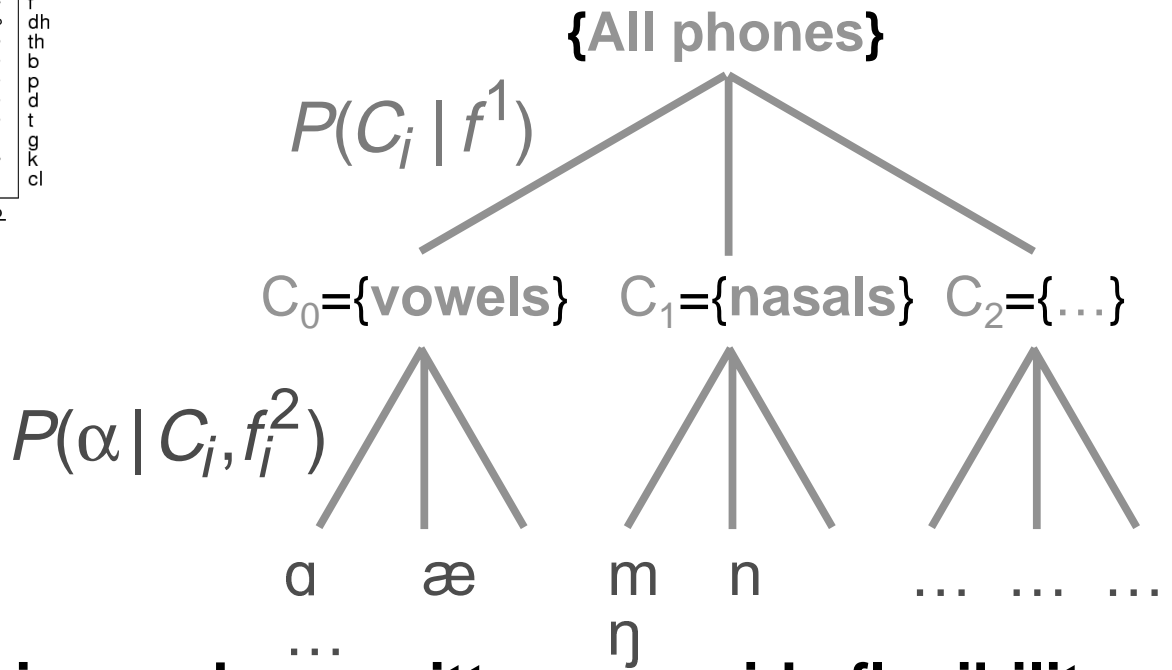


- **Landmark and segment models complement frames**

Beyond homogeneity?

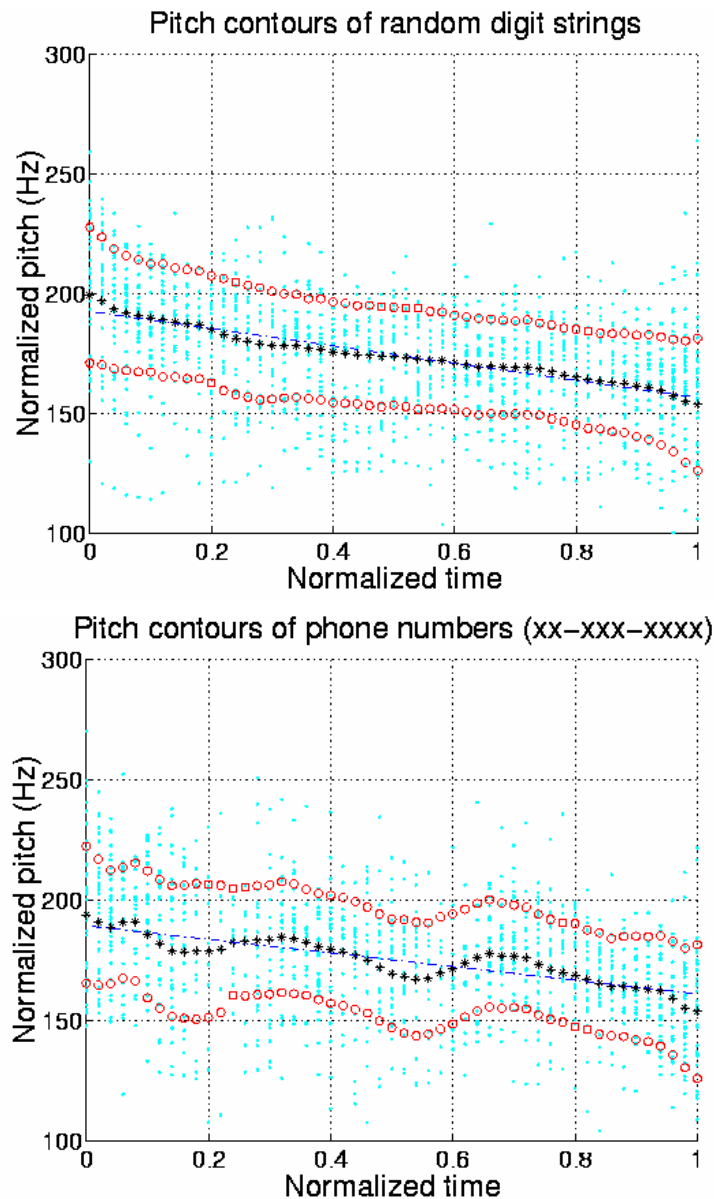
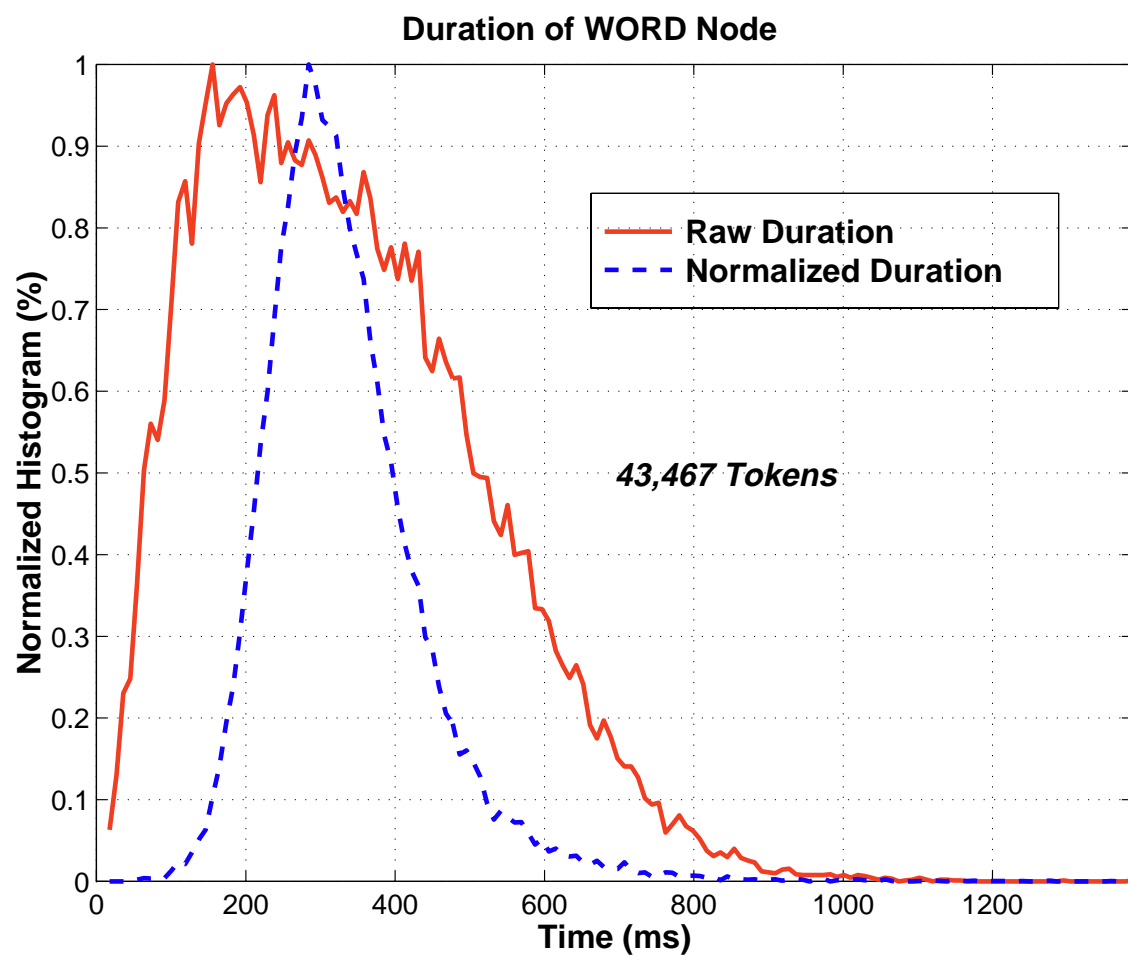


$$P(\alpha | f_1 \dots f_n) = \sum_i P(\alpha | C_i, f_i^2) P(C_i | f_i^1)$$



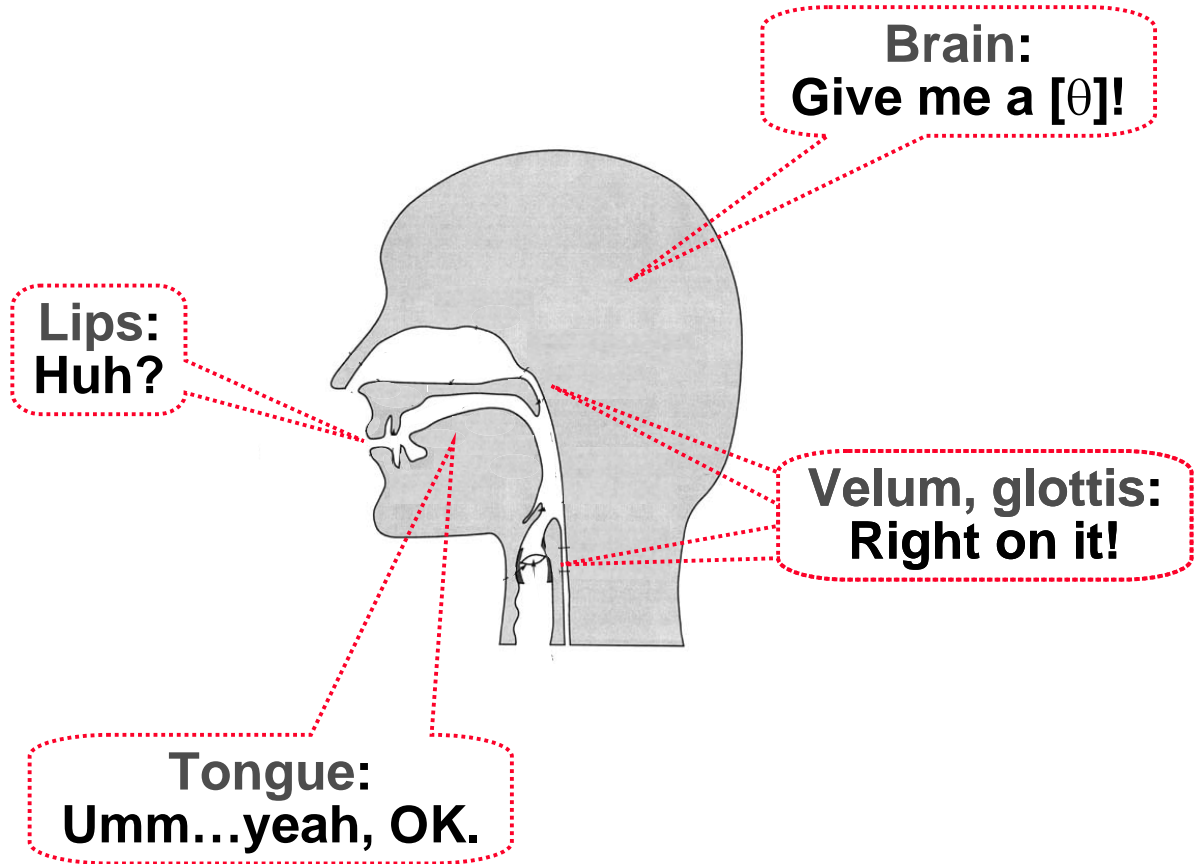
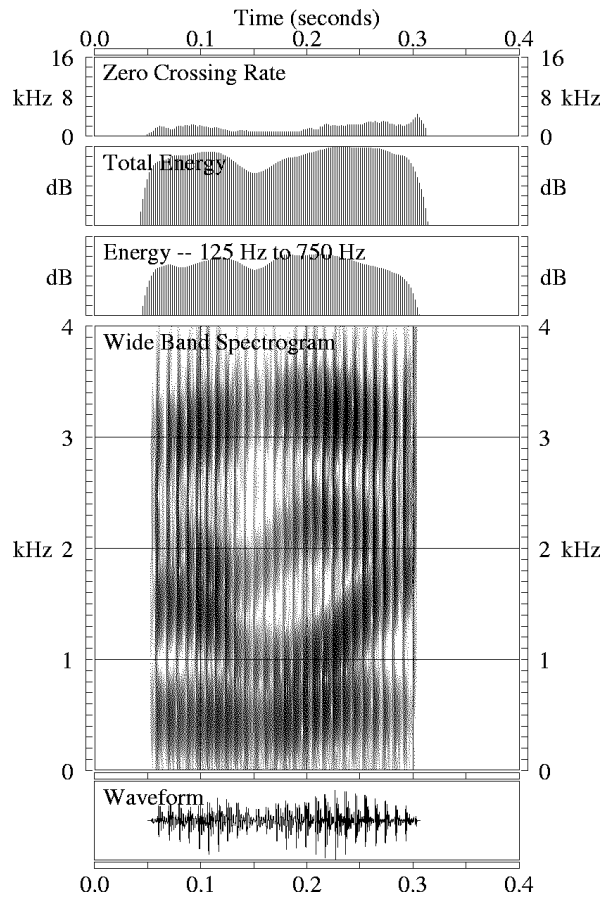
- **Heterogeneous hierarchies and committees provide flexibility**

Beyond phonetics?



- **Why is prosody so important for TTS, but not for ASR?**

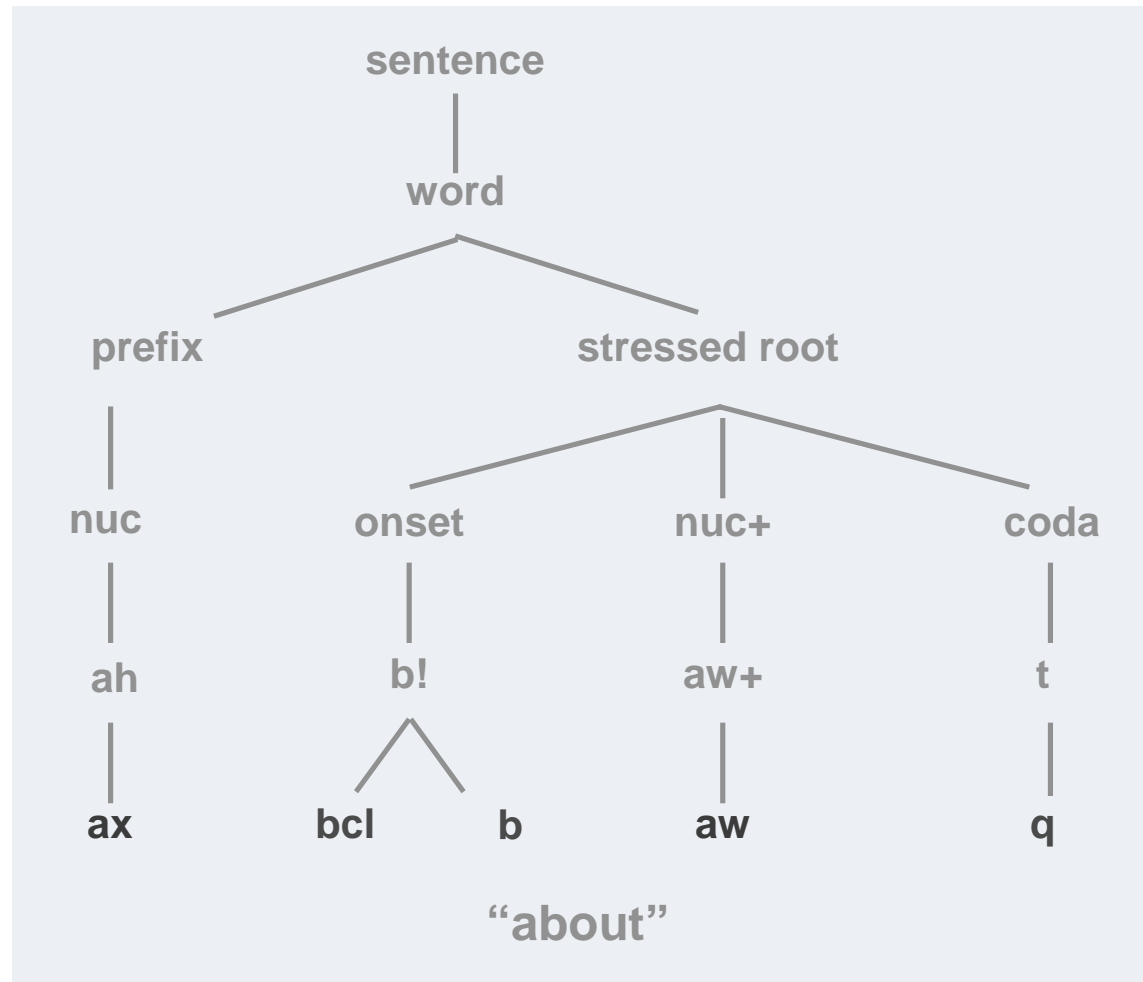
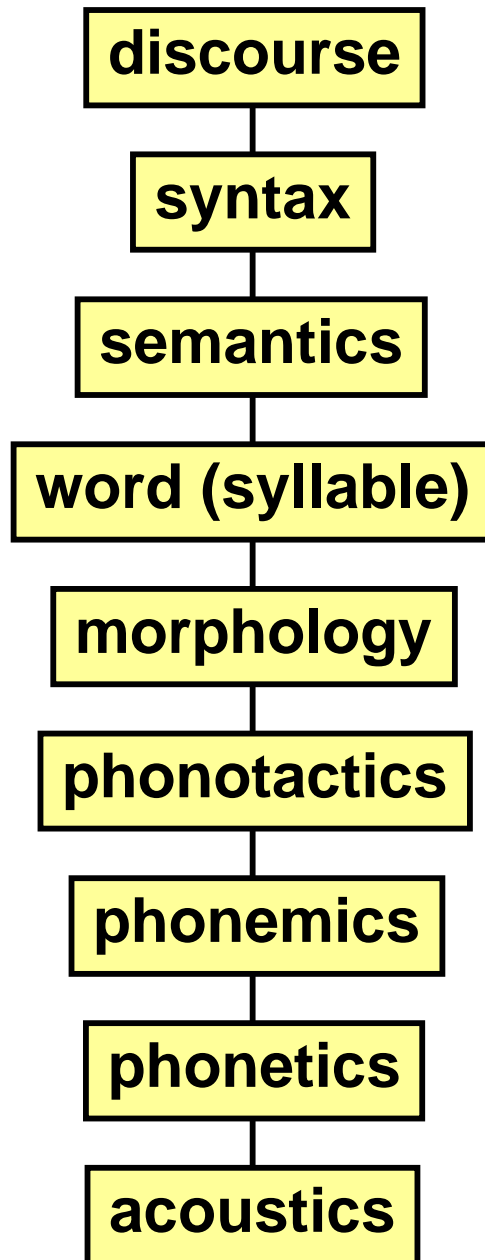
Beyond phonemes?



How "warmth" gets its 'p'

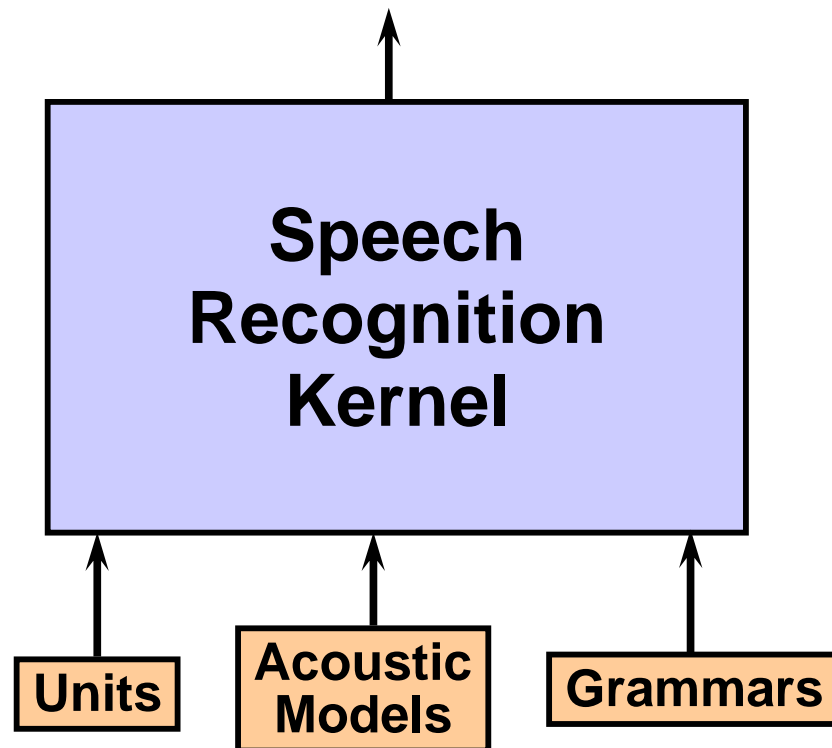
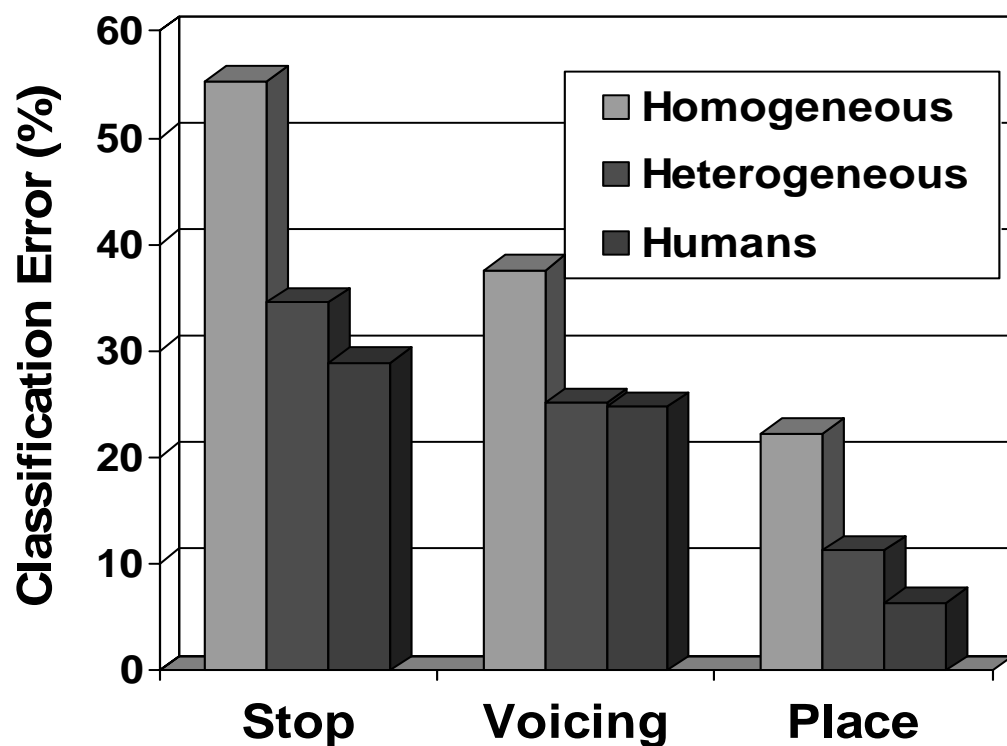
- **Wanted:** lexical representations that can accurately describe surface realizations

Beyond words?



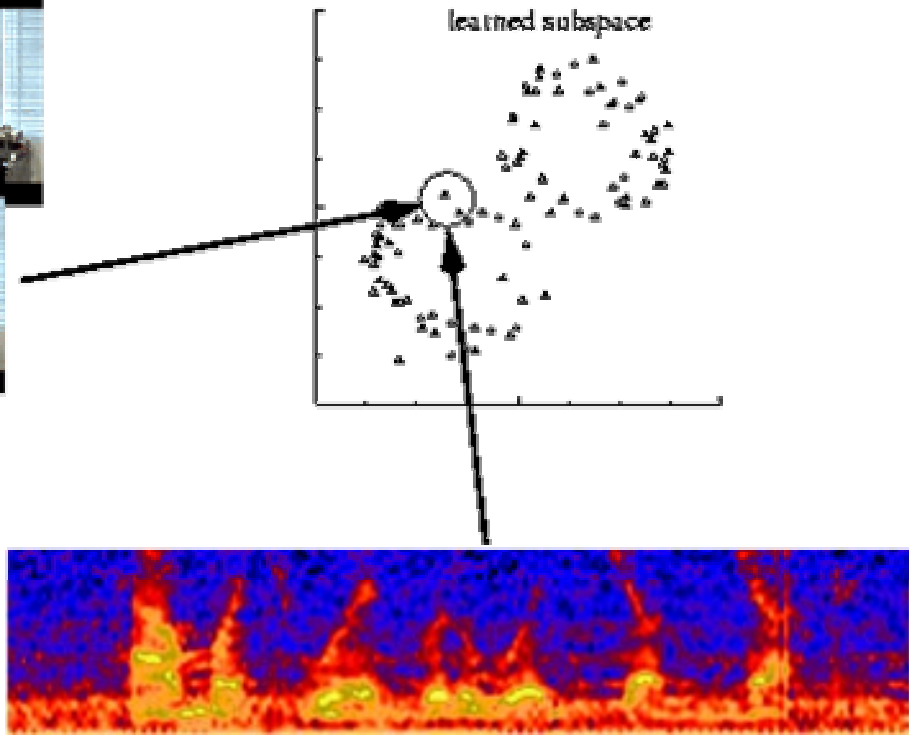
- Opportunities exist for more explicit modelling of linguistic hierarchies (e.g., language & phonological models)

Beyond WER?



- **Basic acoustic-phonetic modelling can be much improved**
 - Compare to benchmarked human performance (e.g., phones, syllables)
 - A next generation of TIMIT could help support this research
- **Better sub-word models should help new word recognition**

Beyond audio?



(from Fisher and Darrell)



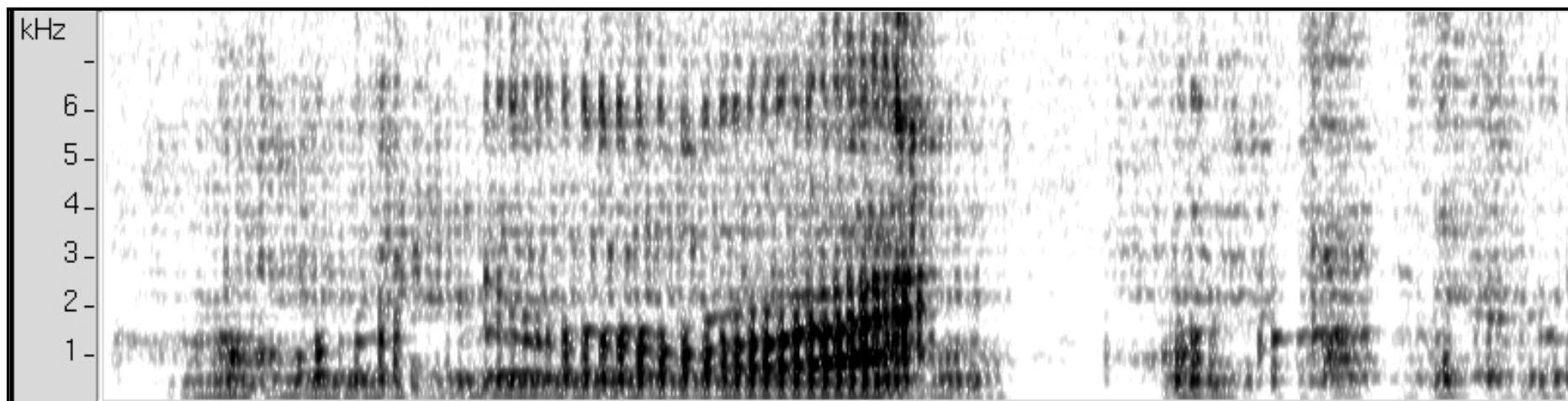
Image Variance



Audiovisual Mutual Information

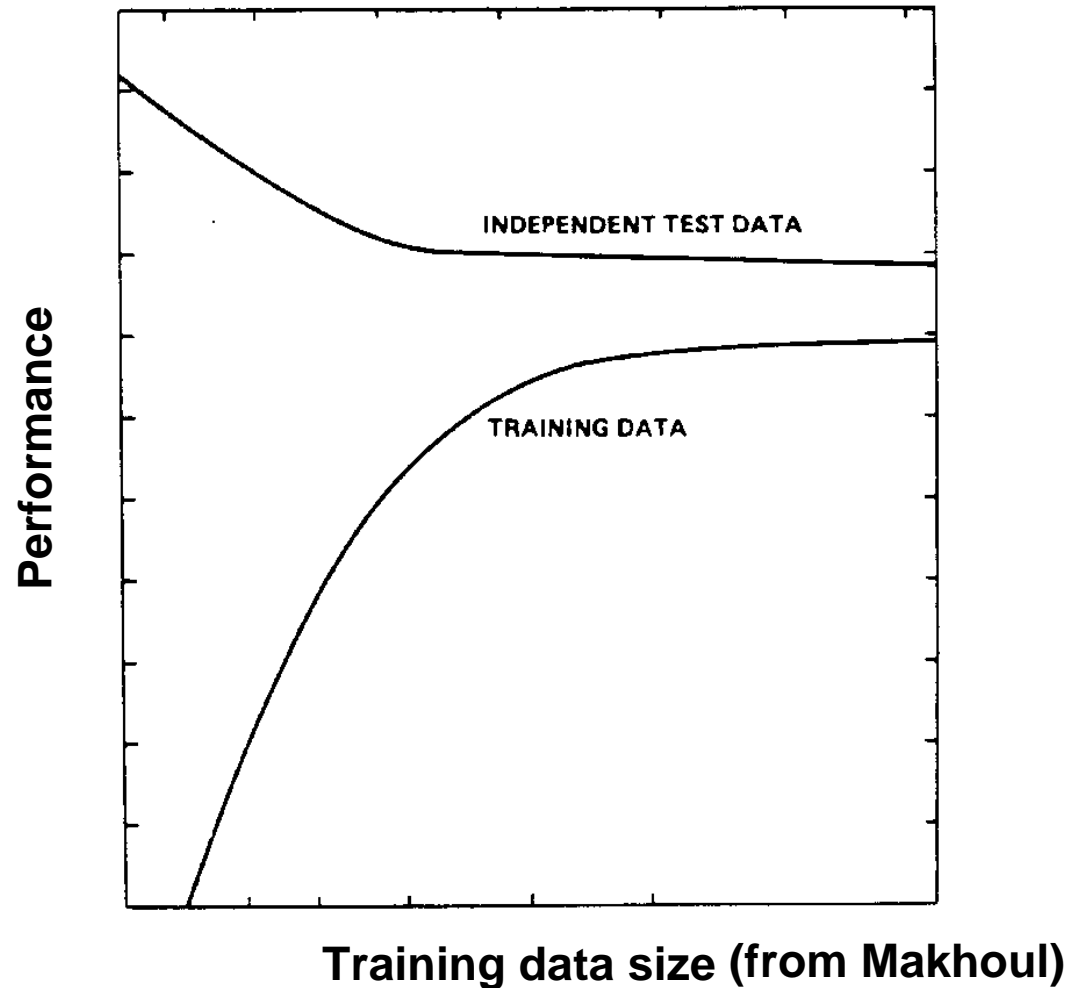
- **Audio-visual information is complementary, consistent**
- **Will non-audio information help ASR go beyond humans?**

Beyond speech?



- **Robustness will come from processing general audio**
- **Speech knowledge can help us determine non-speech**

Beyond tasks?



- **Better use of speech knowledge might help us generalize faster**
- **Should we have a “Turing torture test” to evaluate generality?**
- **Do we need to see an example of everything beforehand?**

Summary

- **Although it is difficult to determine how to incorporate speech knowledge into modern speech recognizers, there exist opportunities at many levels**
- **Speech knowledge should also play an important role when we process audio-visual and/or non-speech data**