NSF Symposium on Next Generation ASR

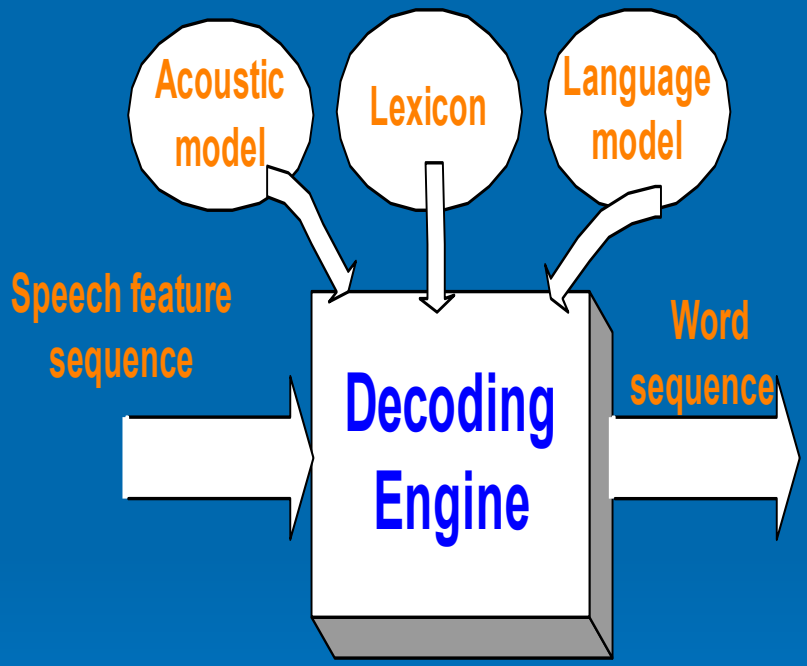
A Perspective of Present and Future Speech Modeling

Yunxin Zhao

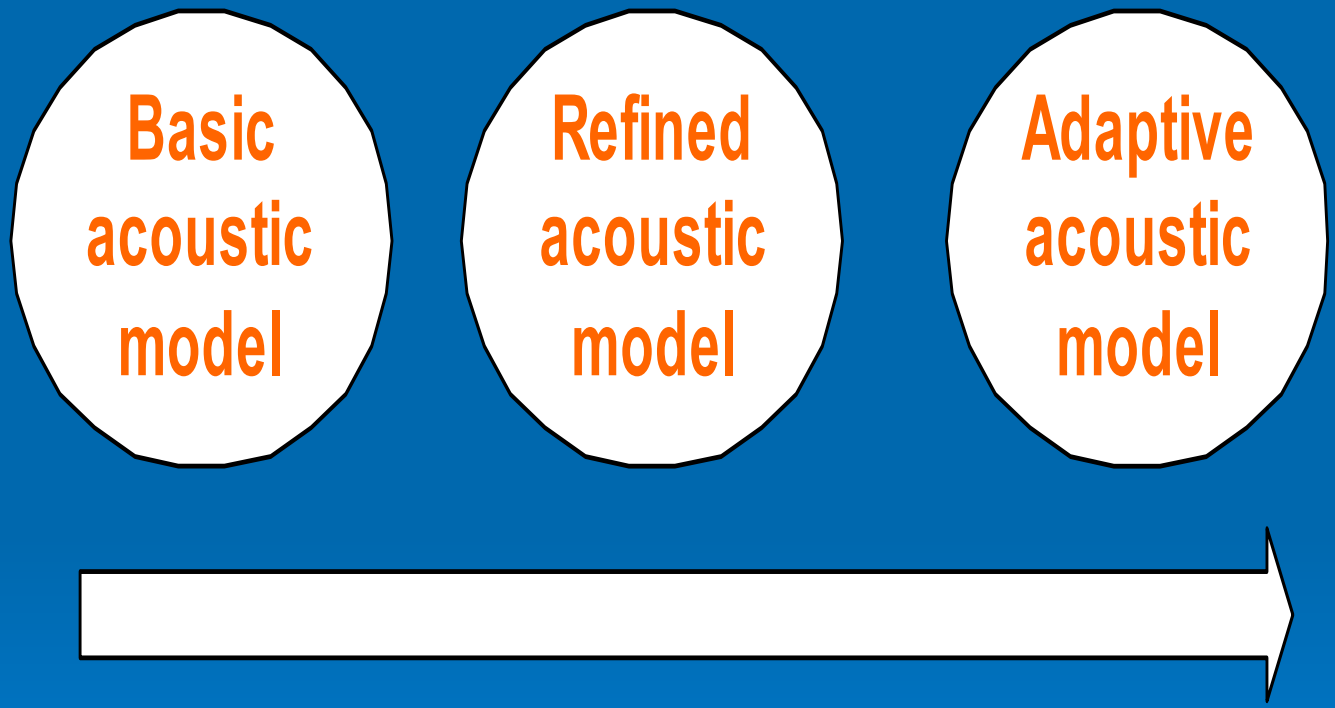
Department of Computer Science
University of Missouri-Columbia

Speech modeling in ASR

- Acoustic model assigns likelihood scores to speech features, e.g. MFCC or PLP, for each acoustic unit.
- Scoring needs to be robust to speakers' voice characteristics, accent, style, channel, and environment.
- In general, acoustic model is the most significant knowledge source among the three that determines the accuracy of ASR system.



Progress of Acoustic modeling



Basic techniques of acoustic modeling

- Hidden Markov modeling of context-dependent phone units.
- Phonetic decision tree based clustering of allophones within aligned HMM states of each phone unit.
- Gaussian mixture density sharing by tied HMM states.
- Maximum likelihood estimation as the standard optimization criterion, model complexity controlled by BIC or MDL.
- Acoustic models become standardized in many research sites due to the popularity of HTK and shared data corpora of LDC.

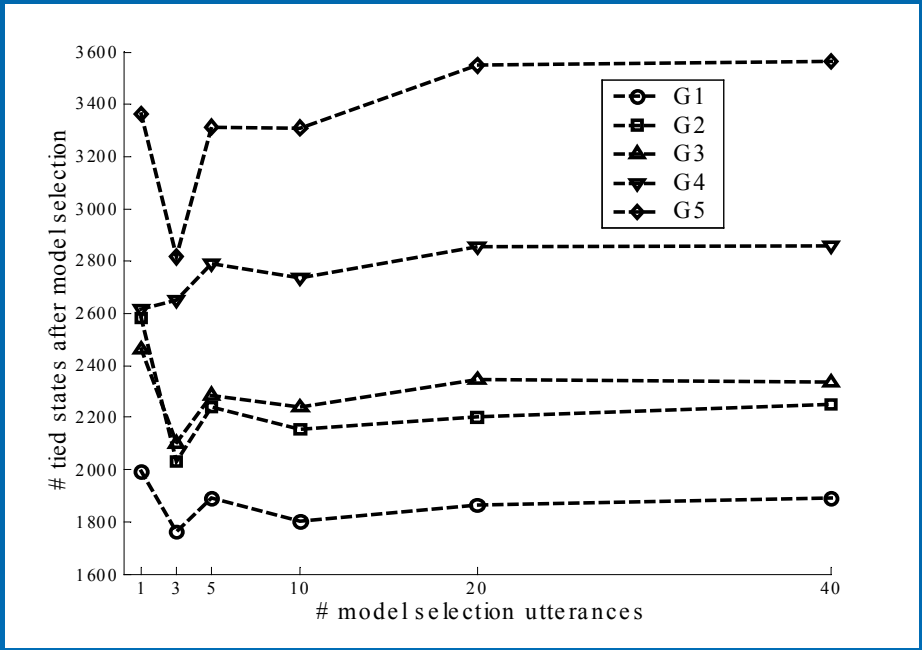
Refined techniques of acoustic modeling

- Largely motivated by the challenging task of Switchboard.
- Conditioning speech features to reduce variations:
  - vocal tract length normalization
  - cepstral mean and variance normalization
  - feature transformation, e.g. HLDA.
- Extended context dependency: triphones to quinphones.
- Discriminative training criterion: MCE, MMIE
- Detailed modeling of nonspeech events.

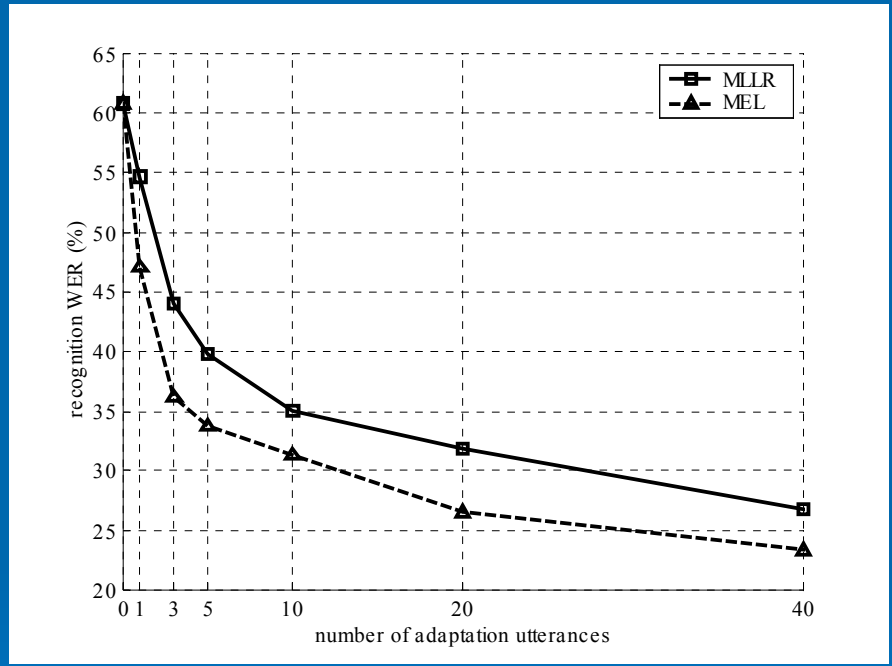
Adaptive acoustic modeling

- Speaker-independent acoustic models are still inefficient to cover individualities of speakers.
- Further improvement for individual speakers are made through speaker adaptation:
  - model transformation: e.g. MLLR
  - model estimation: e.g. MAP
  - model complexity selection for nonnative speakers
  - combinations of the above.
- Acoustic models are made noise-robust by
  - integrating specific models of channel and noise with
    - pre-measured parameters
    - blind estimation of parameters
  - using general speaker adaptation techniques

Complexity selection was performed on phonetic decision trees by a maximum expected likelihood criterion



Complexity selection results on 5 speaker groups, G1-G3 are nonnative



Word error of G1 resulting from complexity selection & MLLR

(figures from X. He & Y. Zhao, IEEE Trans. on SAP, July, 2003)

Future directions?

Have we reached the limit of the current modeling framework?

- NIST regular evaluation events, e.g. Switchboard task, show performance improvements of research systems over the years.
- With the availability of large training speech corpora and vast computing power, modeling techniques that were deemed impractical, e.g., long range context-dependent modeling, have been explored and led to higher ASR accuracy.
- State-of-the-art ASR systems have shown recognition word error rate to be a linearly decreasing function of logarithmic training data size, while compared with children's exposure to speech, as young as two years old, our speech models are still very much under-trained.

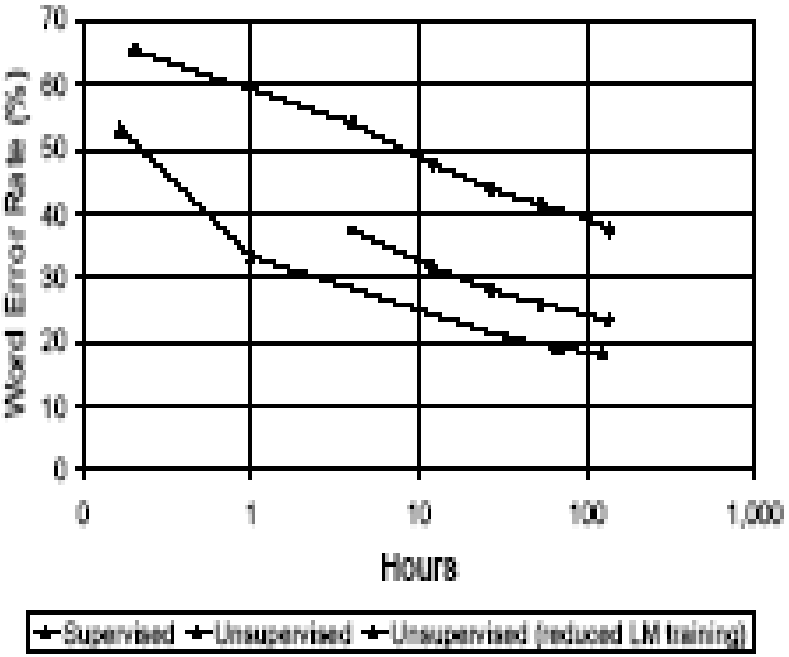


Figure 3. The results from Table 2 plotted using a logarithmic scale for the quantity of training material.

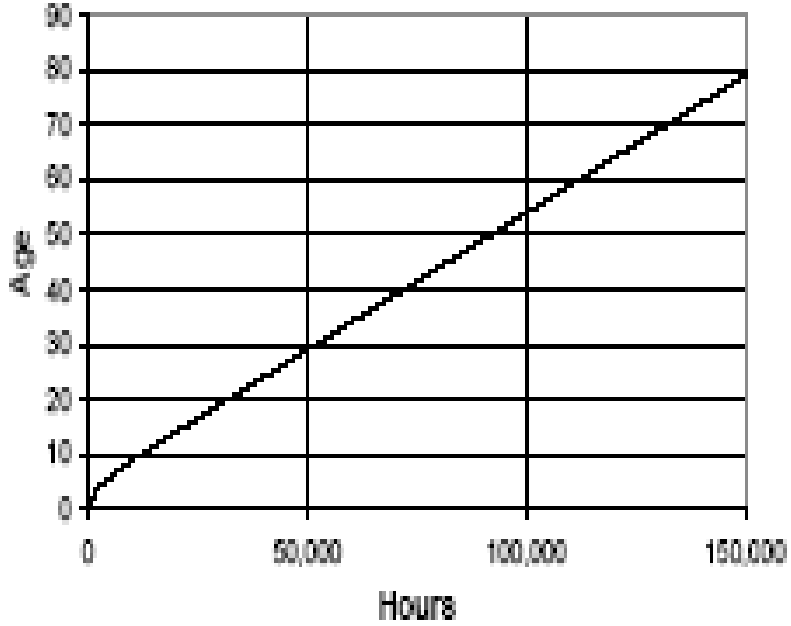


Figure 4: Estimated amount of speech a human being hears as a function of age.

(figures are from paper by R. Moore in Eurospeech 03)

- Envisioning unlimited training speech data becoming available:
  - We can build acoustic model on larger units than phones, such as syllables, morphems, or even multiwords (short words), which are known to be capable of better capturing coarticulation effects.
  - We can design acoustic model to be more amenable to self learning, including structure, parameter, and complexity.
- Envisioning computing power becoming comparable to human brain:
  - We should integrate more knowledge sources (KS) in ASR, such as prosody, environment.
  - We should make model adaptation to operating states a generic function of ASR system.
- Effective adaptive learning algorithms are essential.

Bio-mimetic speech modeling

- At present, it seems that we have better utilized properties of human auditory system in the representation of speech features than the design of speech recognition system.
- Cross-disciplinary research will likely stimulate revolutionary new methods of speech modeling.
- Through a full understanding of human cognitive functions we can build bio-mimetic speech recognition systems into bio-mimetic computers.