Landmark-Based Speech Recognition

Report of the Workshop Group, 8/16/2004

Mark Hasegawa-Johnson, *University of Illinois*
James Baker, *Carnegie-Mellon*
Steven Greenberg, *University of California*
Katrin Kirchhoff, *University of Washington*
Jen Muller, *Department of Defense*
Kemal Sonmez, *SRI*
Sarah Borys, *University of Illinois*
Ken Chen, *University of Illinois*
Amit Juneja, *University of Maryland*
Karen Livescu, *MIT*
Srividya Mohan, *Johns Hopkins University*
Emily Coogan, *University of Illinois*
Tianyu Wang, *Georgia Tech*

Special Thanks for Lattices, Software, and Research Results contributed by
Andreas Stolcke, John Makhoul, Spyros Matsoukas, Owen Kimball, Dimitra Vergyri, Luciana Ferrer, Jeff Bilmes, Partha Niyogi, Jim Glass, T.J. Hazen, Eric Fosler-Lussier, Carol Espy-Wilson, Shihab Shamma, Yanli Zheng, Tarun Pruthi, and Nima Mesgarani

Executive Summary: Landmark-Based Speech Recognition

Scientific Objective:

A recognizer capable of learning, from data, the information structures apparently used by human subjects in speech processing experiments.

Technological Objective:

Flexible acoustic and pronunciation models, in a high-dimensional observation space, with very low generalization error.

Systems Implemented and Tested:

1. Binary phonetic classifiers: place of articulation classification error dropped 10-50% relative to start of workshop
2. Dynamic Bayesian Network model of pronunciation variability: Computational complexity of an SVM-EBS-DBN hybrid model reduced from ~2000RT to ~100RT. Computational complexity of an SVM-DBN model is still ~1000RT, but dropping. No WER reduction yet on RT03 development set
3. Discriminative Pronunciation Model driven by analysis of word-lattice confusion networks
4. Maximum entropy score combination system for stream weight estimation in an augmented lattice

Current bottom line:

Systems 3 & 4 separately are each getting a non-significant WER reduction on the RT03 development set.

Outline of this talk

1. Motivation
   1. Why do we believe that landmark-based and gesture-based methods can reduce WER?
   2. Why test in a lattice rescoring paradigm?
2. System architecture
   1. System 1: a generative model (DBN+SVM) based on articulatory phonology
   2. System 2: a discriminative model (MaxEnt) targeted at word errors in a confusion network
3. Future plans
   1. … for the next twelve months
   2. … for the rest of the afternoon

Scientific motivation: Human speech perception is landmark-synchronous, and mediated by phonology

- “Landmark-Based Speech Perception” (Stevens):
  - Manner-Change Landmarks:
    - Human recognition of consonants requires 40ms excised after release or before closure (Furui)
    - Humans recognize vowels better if given vowel onset and offset (3 glottal pulses each) than if given the “steady-state” part of the vowel (all other glottal pulses) (Strange et al.)
    - Supported by our results for stops, nasals, fricatives (landmark place of articulation error: 10-20%, segment-internal place error: 20-50%)
  - Vowel-peak and Glide-dip landmarks:
    - Hillenbrand et al (1995): dynamic spectral measurements covering both vowel peak and offglide are necessary to classify the vowel
    - Supported by our results for vowels and glides (segment-internal place classification error: 9-15%, landmark error: 12-20%)
  - Errors in perception of nasality, frication, stridency, place, and voicing are independent (Miller and Nicely)
- “Articulatory Phonology” (Browman and Goldstein)
  - In VCV utterances: manner of C can change, never place
  - In VCCV: either C can assimilate features of the other, but new features are never created from scratch

Technological goal: Improved precision of the acoustic model and pronunciation model

- Acoustic Model
  - Place of articulation is encoded by the whole pattern of change in spectral, formant, and rate-scale features (70ms following consonant release)
  - Dynamic spectrum is a large observation vector (200-10000 dim)
  - Generalization from a high-dimensional observation: use SVMs
  - Result: well-selected new observation dimensions reduce classification error up to the point where number of observation dimensions is almost equal to number of training frames
- Pronunciation Model
  - Switchboard contains dozens of pronunciations per word
  - Multiple-pronunciation dictionaries reduce WER after ~1.5/word
  - Model 1, “articulatory phonology:” represent parameter tying among pronunciation variants using a dynamic Bayesian network
  - Model 2, “discriminative pronunciation model:” find a small number of landmarks whose overlap or sequence distinguishes the word from competing words

Why lattice rescoring is a useful test…

- The goal of precise acoustic and precise pronunciation modeling
  - … is to improve our ability to correctly recognize words
  - Standard evaluation metric for this capability is WER
- Complementary information
  - Objective of the landmark-based system: explicit models of spectral dynamics, in a 2000-dimensional observation space (spectrogram+short-time-energies+formants+auditory model) that is (we believe) different from the observation space modeled by the HMM
  - Augmenting lattice edge scores with complementary information can sometimes reduce WER
- Simplified problem
  - System 1, articulatory phonology: computational complexity too high for first-pass recognition
  - System 2, discriminative pronunciation model: constrained use of landmarks to fix errors in the first-pass system without introducing new errors

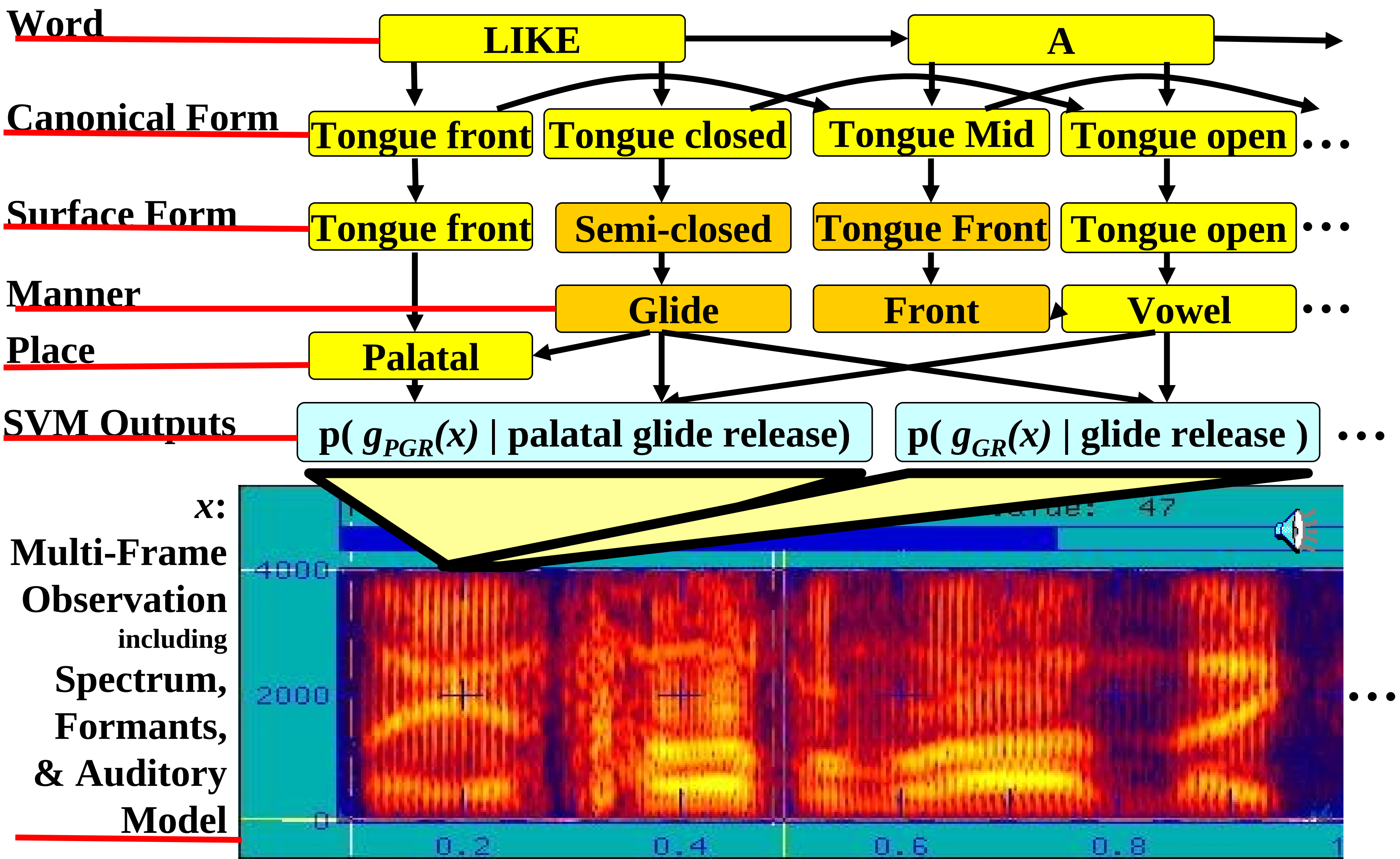
… and why lattice rescoring is not a perfect test

- Word boundary times in lattice may include landmarks from neighboring words, or leave out landmarks from target word
- Correct transcription is not always in the lattice
- Word errors in the lattice are caused by a combination of many factors affecting both language model and acoustic model
  - Language model score of incorrect transcription is often much better than that of correct transcription
  - Large difference in language model scores may swamp small improvements in the acoustic score

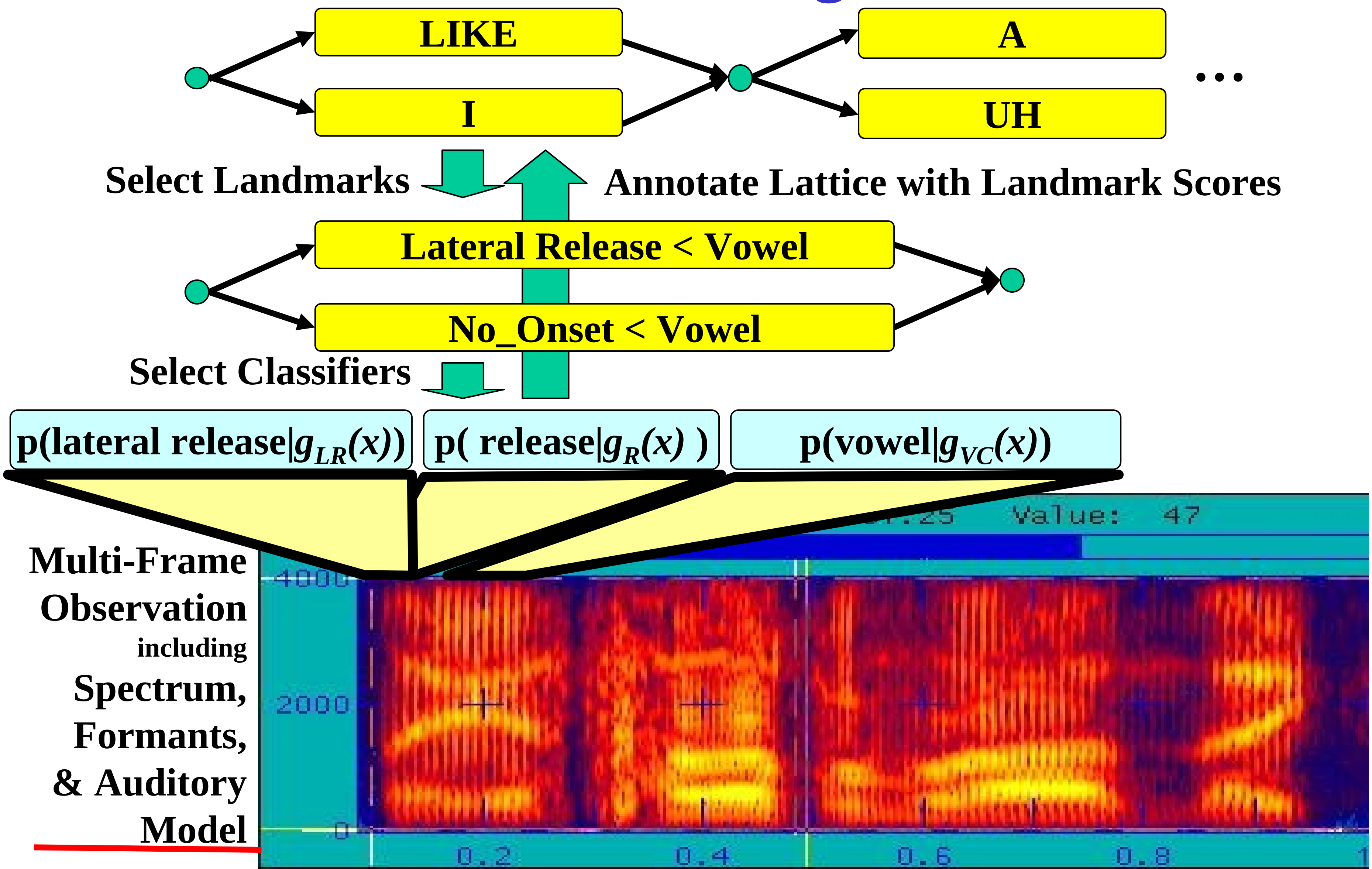
System architectures developed during WS04

- Binary acoustic phonetic classifiers for
  - Detecting a manner-change landmark
  - Classifying place of articulation at each landmark and at each segment-internal frame
- DBN-SVM model of pronunciation variability
- Discriminative pronunciation model for rescoring of confusion networks
- Maximum Entropy method for estimating stream weights for lattice rescoring

DBN-SVM model of pronunciation variability



Discriminative pronunciation modeling



Maximum entropy estimation of stream weights for lattice rescoring

$$f_1(obs, hyp) = \log p_{AM}(obs \mid hyp)$$

$$f_2(obs, hyp) = \log p_{LM}(hyp)$$

$$f_3(obs, hyp) = [\#words(hyp)]$$

$$f_4(obs, hyp) = \log p_{LANDMARK-PRONUNCIATION-MODEL}(obs \mid hyp)$$

$$\log P(hyp \mid obs) = \left(\sum_i \lambda_i f_i(obs, hyp)) \right) - \log Z(obs)$$

Current results:

Training corpus: 20% WER (17% reduction)

Development test corpus: 12 word reduction in WER (<0.1%)

Future Plans: for the next twelve months

- Full DBN+SVM hybrid system, with all classifier context dependencies encoded as edges in the DBN, will be made practical and then tested. Proposed task: lattice rescoring on Hub-5 data
- Systems intermediate between HMM and DBN+SVM will be developed and tested
- Progressively improved acoustic classifiers will be tested in both MaxEnt and DBN+SVM systems
- Maximum entropy lattice rescoring will be tested with prosodic, syntactic, and other word-level side information
- Mathematical analysis will study DBN+SVM integration in both training and test

Future Plans: for the rest of the afternoon

- Technical presentations
  - Amit Juneja: Distinctive feature detection and landmark-based rescoring
  - Karen Livescu: Feature/Landmark-based pronunciation modeling using dynamic Bayesian networks
  - Katrin Kirchhoff: Discriminative rescoring using landmarks
  - Kemal Sonmez: Maximum entropy techniques for min-WER score combination with sausages
  - Steve Greenberg: Beyond landmarks
- Coffee break
- Student proposals for post-workshop research
  - Srividya Mohan: Automatic identification and classification of words using phonetic and prosodic features
  - Emily Coogan: Pronunciation variability
  - Tianyu Wang: Glottalization and vowel nasalization detection