Multivariate-State Models for Speech Recognition

Mark Hasegawa-Johnson

Lecture 3
Winter School on Speech and Audio Processing
Chennai, India
January 10, 2016
Speech is a one-dimensional signal which encodes multiple simultaneous partially independent information streams.
Outline

I. Background: Hidden Markov Models
II. Problem statement: Multivariate content of speech
III. A Multivariate-state HMM: Formant frequencies as state variables
IV. A bit of Phonology: Distinctive Features and Acoustic Phonetic Landmarks
V. Neural networks and support vector machines for landmark detection and distinctive feature classification
VI. Dynamic Bayesian Networks for pronunciation modeling
I. Background: Minimum Probability of Error (MPE) Classification

**Class Definition:**

- Functional Form with Trainable Parameters

**Training:**

- Modify Parameters of $p(\text{obs} \mid \text{class})$
- Create Lookup Table of $p(\text{class})$

**Classification:**

$$\text{class} = \text{argmax } p(\text{class} \mid \text{obs})$$
Hidden Markov Models
HMM Phone Models
HMM Word Models

[Phone String]

[Acoustic States]

Dictionary Expansion

Phone Expansion

\[
p(o \mid i), p(o \mid j), p(o \mid k)
\]
HMM Sentence Models

[Word String]
... did you bet your car ...

[Phone String]
... /y/ /uw/ /b/ /eh/ /t/ ...

[Acoustic States]

\[
p(o | i) \quad p(o | j) \quad p(o | k)
\]

obs
obs
obs
Recognition Scoring

Find Q to maximize the “Recognition Probability,”

\[ P(O,Q) = p(i) \ p(o_1|i) \ p(i|i) \ p(o_2|i) \ldots \]

Hidden Markov Model = [ i \ j \ k \ \ldots ]

State Sequence Q = [  i  i  i  j  j  k  k  k  \ldots ]

Observations O = [  o_1  o_2  o_3  o_4  o_5  o_6  o_7  o_8  o_9 ]
Implementation: the Viterbi Algorithm

**RECOGNITION:** \( \max_Q P(O, Q) = \max_{q_T} \delta_T(q_T) \)

where...

\[
\delta_t(q_t) = \max_{q_1} \ldots \max_{q_{t-1}} p(q_1, \tilde{o}_1, \ldots, q_t, \tilde{o}_t) \\
= \max_{q_{t-1}} \left[ \delta_{t-1}(q_{t-1}) \ p(q_t | q_{t-1}) \right] \ p(\tilde{o}_t | q_t) \\
1 \leq q_t \leq N, \ 1 \leq t \leq T
\]

**COMPLEXITY** = \( \mathcal{O} \left\{ TN^2 \right\} \)
Background: Stop Cons. Release

- **English has 3 places of articulation:**
  - Lips (b, p)
  - Tongue Blade (d, t)
  - Tongue Body (g, k)
II. Problem Statement: Content of Speech is Multivariate

1. Source Information: Prosody, Articulatory Features

<table>
<thead>
<tr>
<th>Breath Group</th>
<th>&quot;bet your car?&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Foot&quot;</td>
<td>&quot;bet your&quot;</td>
</tr>
<tr>
<td>Syllable</td>
<td>&quot;bet&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phone String</th>
<th>b</th>
<th>eh</th>
<th>t</th>
<th>y</th>
<th>ow</th>
<th>r</th>
<th>k</th>
<th>aa</th>
<th>r</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Articulator</th>
<th>lips</th>
<th>blade</th>
<th>blade</th>
<th>body</th>
<th>blade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue Tip</td>
<td>point?</td>
<td>flat</td>
<td>point</td>
<td></td>
<td>point</td>
</tr>
</tbody>
</table>
Content of Speech is Multivariate

2. Useful Non-Source Information: Composite Acoustic Cues

*Tongue Blade Consonant:* Peak Frequency at Release close to F4 or F5 of Vowel

Voice Onset Time > 30ms --> *Unvoiced Stop*

F2 onset vs. F2 in vowel --> *Tongue Blade Consonant*
Composite Cues: Traditional Solution

Automatic Peak Tracking

Statistical Classification

/b,p/ OR /d,t/ OR /g,k/
Types of Measurement Error

- Small Errors: Spectral Perturbation
- Large Errors: Pick the Wrong Peak

![Graph showing frequency and amplitude in dB]
Large Errors are 20% of Total

Std Dev of Small Errors = 45-72 Hz
Std Dev of Large Errors = 218-1330 Hz
P(Large Error) = 0.17-0.22

Measurement Error (Hertz)  
re:  Manual Transcriptions
Measurement Error Predicts Classification Error

<table>
<thead>
<tr>
<th>Correct Answer</th>
<th>Errors (Manual Measures)</th>
<th>Errors (Automatic, Actual)</th>
<th>Errors (Automatic, Predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/b/, /p/</td>
<td>20%</td>
<td>34%</td>
<td>40%</td>
</tr>
<tr>
<td>/d/, /t/</td>
<td>5%</td>
<td>18%</td>
<td>17%</td>
</tr>
<tr>
<td>/g/, /k/</td>
<td>7%</td>
<td>19%</td>
<td>22%</td>
</tr>
</tbody>
</table>
III. Solution: Composite Cues as State Variables

Multivariate State Sequence, \([Q,F]\)

Observations \(O = [o_1 \ o_2 \ o_3 \ o_4 \ o_5 \ o_6 \ o_7 \ o_8 \ o_9]\)
Complexity of Solution Without Additional Constraints

\[ \delta_t(\vec{q}_t) = \max_{\vec{q}_{t-1}} \left[ \delta_{t-1}(\vec{q}_{t-1}) \ p(\vec{q}_t | \vec{q}_{t-1}) \right] \ p(\vec{o}_t | \vec{q}_t) \]

\(\vec{q}_t\) takes \(N^P\) different values, so

**COMPLEXITY** = \(\mathcal{O}\left\{ TN^{2P} \right\}\)
Useful Constraint #1: State Independence

\[ p(\vec{q}|\vec{r'}) = \prod_{p=1}^{P} p(q_p|r_p), \quad p(\vec{\sigma}_t|\vec{q}) = \prod_{p=1}^{P} p(\vec{\sigma}_t|q_p) \]

then for each \( p \in [1, P], \)

\[ \delta_t(q_p) = \max_{r_p} [ \delta_{t-1}(r_p) p(q_p|r_p) ] p(\vec{\sigma}_t|q_p) \]

**Complexity** = \( \mathcal{O}\left\{ PTN^2 \right\} \)
Useful Constraint #2: Hierarchical Dependence

\[ \text{COMPLEXITY} = \mathcal{O} \left( TP_F (N_q N_F)^2 \right) \]
## Description of the Test System

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>50ms Consonant Releases, 12 Tokens/Class</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>DFT Normalized Amplitude, DFT Convexity</td>
</tr>
<tr>
<td><strong>Transition PDF</strong></td>
<td>Gaussian</td>
</tr>
<tr>
<td><strong>Observation PDF</strong></td>
<td>Gaussian</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>3 Classes: Lips, Blade, Body, Closed-Set Classification</td>
</tr>
</tbody>
</table>
## Test System Results

<table>
<thead>
<tr>
<th></th>
<th>100% Accurate (Closed-Class, 36 Tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>$4T_P N_Q N_F^2 = 864$ K multiplies</td>
</tr>
<tr>
<td></td>
<td>(18ms at 50 Mflops)</td>
</tr>
<tr>
<td>Fringe Benefit</td>
<td><em>a Posteriori</em> $p(F1 \mid O, Q)$ gives</td>
</tr>
<tr>
<td></td>
<td>“confidence limits” on the</td>
</tr>
<tr>
<td></td>
<td>automatic measurement of $F1$</td>
</tr>
</tbody>
</table>
a Posteriori Measurement Distributions: 10ms After /d/ in “dark”

DFT Amplitude

DFT Convexity

P(F | O, Q)

Frequency (0-4000 Hertz)
IV. a bit of Phonology: landmark-based speech recognition

- What is a phonetic landmark?
- MLP-based and SVM-based landmark detection
- Event-based and DBN-based pronunciation modeling for landmark-based speech recognition
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%
2. Dynamic Bayesian Network model of pronunciation variability: Computational complexity ~2000RT to ~100RT
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 each achieve a small non-significant WER reduction on Switchboard
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 nearly 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
What are Landmarks?

- Time-frequency regions of high mutual information between phone and signal (maxima of $I(q;X(t,f))$)
- Acoustic events with similar importance in all languages, and across all speaking styles
- Acoustic events that can be detected even in extremely noisy environments

Where do these things happen?

- Syllable Onset $\approx$ Consonant Release
- Syllable Nucleus $\approx$ Vowel Center
- Syllable Coda $\approx$ Consonant Closure

$I(q;X(t,f))$ experiment: Hasegawa-Johnson, 2000
Landmark-Based Speech Recognition

Phoneticontent:

MAP transcription:
... backed up ...

Search Space:
... buck up ...
... big dope ...
... backed up ...
... bagged up ...
... big doowop ...

Syllable Structure:
ONSET  NUCLEUS  CODA
ONSET  NUCLEUS  CODA

Time (s)
Frequency (kHz)
Lexical Notation: What are “Distinctive Features?”

MANNER FEATURES:  
+sonorant +continuant = Vowel, Glide  
+sonorant –continuant = Nasal, /l/  
–sonorant +continuant = Fricative  
–sonorant –continuant = Stop
Each Distinctive Feature relies on different acoustic observations.

Acoustic diversity can improve word recognition accuracy in noise:

10% WRA improvement at 0dB SNR.

(Kirchhoff, 1999)
Noise Robustness of Distinctive Features: Pink Noise

Articulatory feature classification more robust than phone classification at low SNRs

(Chang, Shastri and Greenberg, 2001)
Some features better in white noise, some better in pink noise; diversity improves overall performance

(Chang, Shastri and Greenberg, 2001)
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: “articulatory phonology:” represent parameter tying among pronunciation variants using a dynamic Bayesian network
Support Vector Machines

Kernel: Transform to Infinite-Dimensional Hilbert Space

(SVM Discriminant Dimension = argmin(error(margin)+1/width(margin)))

(Niyogi & Burges, 2002: Posterior PDF = Sigmoid Model in Discriminant Dimension)

An Equivalent Model: Likelihoods = Gaussian in Discriminant Dimension
Stop Detection using Support Vector Machines

False Acceptance vs. False Rejection Errors, TIMIT, per 10ms frame
SVM Landmark Detector: Half the Error of an HMM

(1) Delta-Energy ("Deriv"):
Equal Error Rate = 0.2%

(2) HMM (*): False Rejection Error = 0.3%

(3) Linear SVM:
EER = 0.15%

(4) Kernel SVM:
Equal Error Rate = 0.13%

Niyogi & Burges, 1999, 2002
Acoustic Feature Vector: A Spectrogram Snapshot (plus formants and auditory features)
Two types of SVMs: landmark detectors \( p(\text{landmark}(t)) \), landmark classifiers \( p(\text{place-features}(t) | \text{landmark}(t)) \)

2000-dimensional acoustic feature vector

SVM

Discriminant \( y_i(t) \)

Sigmoid or Histogram

Posterior probability of distinctive feature \( p(d_i(t)=1 | y_i(t)) \)
**SVM Training: Switchboard vs. NTIMIT, Linear vs. RBF**

- **NTIMIT:**
  - Read speech = reasonably careful articulations
  - Telephone-band, with electronic line noise
  - Transcription: phonemic + a few allophones

- **Switchboard:**
  - Conversational speech = very sloppy articulations
  - Telephone-band, electronic and acoustic noise
  - Transcription: reduced to TIMIT-equivalent for this experiment, but richer transcription available
Distinctive Feature Lexicon

- Based on ICSI train-ws97 Switchboard transcriptions
- Compiled to a lexicon using Fosler-Lussier’s babylex lexical compiler
- Converted to landmarks using Hasegawa-Johnson’s perl transcription tools

**Landmarks in blue, Place and voicing features in green.**

AGO(0.441765)  
+syllabic  +reduced  +back  
↓continuant↓sonorant  +velar  +voiced  
↑continuant↑sonorant  +velar  +voiced  
+syllabic  –low  –high  +back  +round  +tense

AX

G closure

G release

AG(0.294118)  
+syllabic  +reduced  –back  
↓continuant↓sonorant  +velar  +voiced  
↑continuant↑sonorant  +velar  +voiced  
+syllabic  –low  –high  +back  +round  +tense

IX

G closure

G release

OW
## SVM Training: Accuracy, per frame, in percent

<table>
<thead>
<tr>
<th>Train Test Kernel</th>
<th>NTIMIT Linear</th>
<th>NTIMIT RBF</th>
<th>NTIMIT&amp;SWB Linear</th>
<th>NTIMIT&amp;SWB RBF</th>
<th>NTIMIT Switchboard Linear</th>
<th>NTIMIT Switchboard RBF</th>
<th>Switchboard Linear</th>
<th>Switchboard RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech onset</td>
<td>95.1</td>
<td>96.2</td>
<td>86.9</td>
<td><strong>89.9</strong></td>
<td>71.4</td>
<td>62.2</td>
<td>81.6</td>
<td>81.6</td>
</tr>
<tr>
<td>speech offset</td>
<td>79.6</td>
<td>88.5</td>
<td>76.3</td>
<td><strong>86.4</strong></td>
<td>65.3</td>
<td>78.6</td>
<td>68.4</td>
<td>83.7</td>
</tr>
<tr>
<td>consonant onset</td>
<td>94.5</td>
<td>95.5</td>
<td>91.4</td>
<td>93.5</td>
<td>70.3</td>
<td>72.7</td>
<td>95.8</td>
<td><strong>97.7</strong></td>
</tr>
<tr>
<td>consonant offset</td>
<td>91.7</td>
<td>93.7</td>
<td>94.3</td>
<td><strong>96.8</strong></td>
<td>80.3</td>
<td>86.2</td>
<td>92.8</td>
<td><strong>96.8</strong></td>
</tr>
<tr>
<td>continuant onset</td>
<td>89.4</td>
<td>94.1</td>
<td>87.3</td>
<td><strong>95.0</strong></td>
<td>69.1</td>
<td>81.9</td>
<td>86.2</td>
<td>92.0</td>
</tr>
<tr>
<td>continuant offset</td>
<td>90.8</td>
<td>94.9</td>
<td>90.4</td>
<td><strong>94.6</strong></td>
<td>69.3</td>
<td>68.8</td>
<td>89.6</td>
<td>94.3</td>
</tr>
<tr>
<td>sonorant onset</td>
<td>95.6</td>
<td>97.2</td>
<td><strong>97.8</strong></td>
<td>96.7</td>
<td>85.2</td>
<td>86.5</td>
<td>96.3</td>
<td>96.3</td>
</tr>
<tr>
<td>sonorant offset</td>
<td>95.3</td>
<td>96.4</td>
<td>94.0</td>
<td><strong>97.4</strong></td>
<td>75.6</td>
<td>75.2</td>
<td>95.2</td>
<td>96.4</td>
</tr>
<tr>
<td>syllabic onset</td>
<td>90.7</td>
<td>95.2</td>
<td>91.4</td>
<td><strong>95.5</strong></td>
<td>69.5</td>
<td>78.9</td>
<td>87.9</td>
<td>92.6</td>
</tr>
<tr>
<td>syllabic offset</td>
<td>90.1</td>
<td>88.9</td>
<td>87.1</td>
<td><strong>92.9</strong></td>
<td>54.4</td>
<td>60.8</td>
<td>88.2</td>
<td>89.7</td>
</tr>
</tbody>
</table>
**Acoustic Feature Selection: MFCCs, Formants, Rate-Scale**

1. **Accuracy per Frame, Stop Releases only, NTIMIT**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>MFCCs+Shape</th>
<th>MFCCs+Formants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>RBF</td>
</tr>
<tr>
<td>+/- lips</td>
<td>78.3</td>
<td>90.7</td>
</tr>
<tr>
<td>+/- blade</td>
<td>73.4</td>
<td>87.1</td>
</tr>
<tr>
<td>+/- body</td>
<td>73.0</td>
<td>85.2</td>
</tr>
</tbody>
</table>

2. **Word Error Rate: Lattice Rescoring, RT03-devel, One Talker**

Baseline (PER, GMM-HMM ASR): 15.0% (113/755)

Rescoring, place based on:
- MFCCs + Formant-based params: 14.6% (110/755)
- Rate-Scale + Formant-based params: 14.3% (108/755)
Event-Based Smoothing (EBS) of SVM outputs

- Maximize $\Pi_i p(\text{features}(t_i) \mid X(t_i)) p(t_{i+1} - t_i \mid \text{features}(t_i))$
- Forced alignment mode: computes $p(\text{word} \mid \text{acoustics})$; rescores the word lattice
- Manner class recognition mode: smooths SVM output; preprocessor for the DBN
Small-Vocabulary Word Recognition Using Landmarks: Results on TIDIGITS

TIDIGITS recognition, using models trained on TIMIT. Word recognition accuracy given only MANNER CLASS FEATURES:

- Manner-Class HMMs: 53% WRA
- SVM EBS: 76% WRA

(Juneja and Espy-Wilson, 2003)
Pronunciation Modeling based on Articulatory Phonology (Livescu, Ph.D. thesis 2005)

Many pronunciation phenomena can be parsimoniously described as resulting from asynchrony and reduction of sub-phonetic features

- **One set of features based on articulatory phonology** [Browman & Goldstein 1990]:

  - **warmth** → [w aʊ r m p θ] - Phone insertion?
  - **I don’t know** → [əh də x uː n oʊ n] - Phone deletion??
  - **several** → [s əh r v ək l] - Exchange of two phones???
  - **instruments** → [ɪh_n s ʃ ɛm ɪh_n n s]  
    **everybody** → [eθ_ r ɔːw ə_ay]

---

**t** | **ch** | **k** | **s** | **ih_n** | **s** | **ch** | **em** | **ih_n** | **n** | **s**
---|---|---|---|---|---|---|---|---|---|---
**TExAS**

**eθ** | **r** | **uə_əy** | **l** | **ay** | **ih_ɪd**
---|---|---|---|---|---
**EVERYBODY**
Pronunciation Model: Dynamic Bayesian Network, with Partially Asynchronous Articulators
Pronunciation Model: DBN, with Partially Asynchronous Articulators

- \( \text{word}_t \): word ID at frame \#t
- \( \text{wdTr}_t \): word transition?
- \( \text{ind}_{t}^{i} \): which gesture, from the canonical word model, should articulator \#i be trying to implement?
- \( \text{async}_{t}^{i:j} \): how asynchronous are articulators \#i and \#j?
- \( \text{U}_{t}^{i} \): canonical setting of articulator \#i
- \( \text{S}_{t}^{i} \): surface setting of articulator \#i
A feature-based pronunciation model (Livescu, 2005)

- The model is implemented as a dynamic Bayesian network (DBN):
  - A representation, via a directed graph, of a distribution over a set of variables that evolve through time
- Example DBN with three features:

$$
\text{Pr}(\text{async}^i = a) = \text{Pr}(|\text{ind}^i - \text{ind}^2| = a)
$$

$$
\text{checkSync}^{|i} = 1 \text{ if } |\text{ind}^i = \text{ind}^2| = \text{async}^{|i}
$$

given by baseform pronunciations
**DBN-SVM landmark-based speech recognizer**

<table>
<thead>
<tr>
<th>Word</th>
<th>LIKE</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical Form</td>
<td>Tongue front</td>
<td>Tongue closed</td>
</tr>
<tr>
<td>Surface Form</td>
<td>Tongue front</td>
<td>Semi-closed</td>
</tr>
<tr>
<td>Manner</td>
<td>Glide</td>
<td>Front</td>
</tr>
<tr>
<td>Place</td>
<td>Palatal</td>
<td></td>
</tr>
<tr>
<td>SVM Outputs</td>
<td>$p(g_{PGR}(x) \mid \text{palatal glide release})$</td>
<td>$p(g_{GR}(x) \mid \text{glide release})$</td>
</tr>
</tbody>
</table>

$x$: Multi-Frame Observation including Spectrum, Formants, & Auditory Model
Design decisions

- What kind of SVM outputs should be used in the DBN?
  - Method 1 (EBS/DBN): Generate landmark segmentation with EBS using manner SVMs, then apply place SVMs at appropriate points in the segmentation
  - Method 2 (SVM/DBN): Apply all SVMs in all frames, allow DBN to consider all possible segmentations
    - In a single pass
    - In two passes: (1) manner-based segmentation; (2) place+manner scoring

- How to train DBN (transcriptions vs. SVM outputs)?

- Distinctive feature hierarchy: place features irrelevant unless [+consonantal]. Should DBN model this?

- Should reduction probability $p(S_t^i|U_t^i)$ depend on context (e.g., values of other features, or word frequency, or syllable position)?
Lattice Rescoring

- Pronunciation model (computed from train-ws97):
  - CAT(0.5) ↑sonorant↑continuant +syllabic
  - ↓sonorant↓continuant
  - CAT(0.5) ↑sonorant↑continuant +syllabic

- Rescore the edge:

\[
p(w_i|X) = \sum_{\overline{\lambda}} \left( \frac{p(\overline{\lambda}|w_i)}{\sum_{j=1}^{N} p(\overline{\lambda}|w_j)} \right) p(\overline{\lambda}|X, SNR)
\]
DBN/SVM rescoring experiments

For each lattice edge:
- SVM probabilities computed over edge duration and used as soft evidence in DBN
- DBN computes a score $S \propto P(\text{word} \mid \text{evidence})$
- Final edge score is a weighted interpolation of baseline scores and EBS/DBN or SVM/DBN score

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Experimental setup</th>
<th>3-speaker WER (# errors)</th>
<th>RT03 dev WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>- $\infty$</td>
<td>Baseline</td>
<td>27.7 (550)</td>
<td>26.8</td>
</tr>
<tr>
<td>1</td>
<td>EBS/DBN, “hierarchically-normalized” SVM output probabilities, DBN trained on subset of ICSI transcriptions</td>
<td>27.6 (549)</td>
<td>26.8</td>
</tr>
<tr>
<td>2</td>
<td>+ improved silence modeling</td>
<td>27.6 (549)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EBS/DBN, unnormalized SVM probs + fricative lip feature</td>
<td>27.3 (543)</td>
<td>26.8</td>
</tr>
<tr>
<td>4</td>
<td>+ DBN trained using SVM outputs</td>
<td>27.3 (543)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>+ full feature hierarchy in DBN</td>
<td>27.4 (545)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>+ reduction probabilities depend on word frequency</td>
<td>27.4 (544)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>+ retrained SVMs + nasal classifier + DBN bug fixes</td>
<td>27.4 (544)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SVM/DBN, 1 pass</td>
<td>Miserable failure!</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>SVM/DBN, 2 pass</td>
<td>27.3 (542)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>SVM/DBN, 2 pass, using only high-accuracy SVMs</td>
<td>27.2 (541)</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- **Phonetic Landmarks & Distinctive Features**
  - Better accuracy than phones (because there are fewer target labels over which to perform the classification; target 90%+ accurate for most features)
  - Better noise robustness than phones
  - Use SVMs with small training corpora (up to ~10k examples), use neural net with large training corpora

- **Multivariate-State Models (DBNs)**
  - Formant frequencies as hidden state variables: improved stop consonant classification accuracy (relative to HMM)
  - Articulatory Phonology models:
    - Slightly improved WER on 3 Switchboard speakers
    - Audiovisual speech recognition: complementary to CHMM