Towards a Speaker Invariant Representation of Speech

Aren Jansen

with Ken Church, Hynek Hermansky, and Samuel Thomas

The Center For Language and Speech Processing

JOHNS HOPKINS
UNIVERSITY

JHU CLSP Summer Workshop 2012
July 17, 2012

Speech is Rich with Structure

Semantic: {book, reference, knowledge, wikipedia}

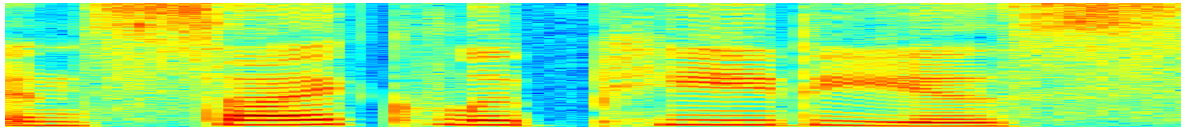
Grammatical: (he sold, NP, to her)

Lexical: encyclopedias

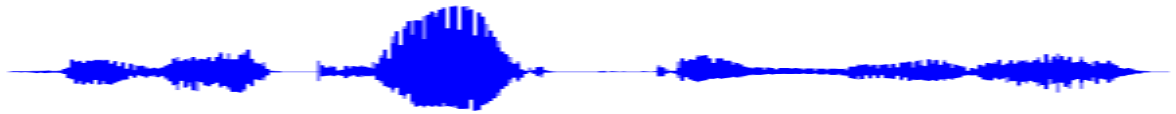
Phonetic: en s ai k l ow p iy d iy aa s

Acoustic-Phonetic: voiced unvoiced voiced unvoiced voiced unvoiced voiced unvoiced

Acoustic:



Observed:

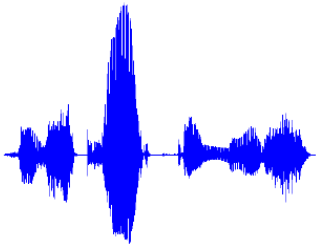


Applying Unsupervised Learning

- **In general, free to choose some combination:**
  - Representation (features)
  - Distance metric
  - Unsupervised learning (clustering) algorithm

- **But:** for speech these choices must be specialized for each level of linguistic structure

Where do you start?

- **Input:** 

- **Scales of Analysis:**
  1. **Featural:** learn maps from short time signal to points in R^d
  2. **Phonetic:** learn to associate regions in R^d with (categorical) subword unit inventory
  3. **Lexical:** learn to associate trajectories in R^d with (categorical) words and phrases
  4. **Semantic:** learn to relate words and phrases according to topical content

Two Constraints

- **Part 1:** Acoustic-phonetic constraints on feature space
- **Part 2:** Lexical constraints on phonetic inventory

Part 1

INTRINSIC SPECTRAL ANALYSIS

[Jansen & Niyogi, ICASSP 2006]

[Jansen, Thomas, & Hermansky, Interspeech 2012]

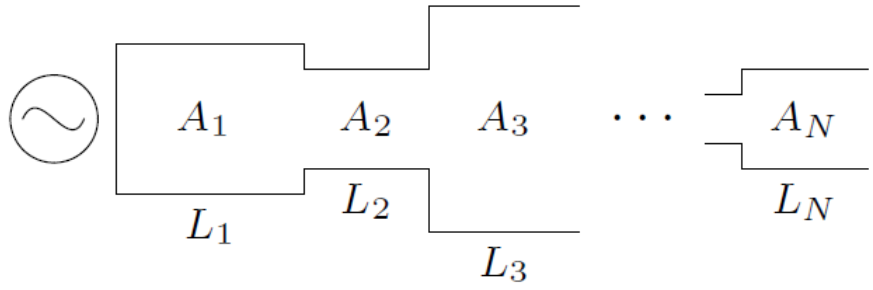
The Constraints of Speech Production

- We cannot produce arbitrary sounds:

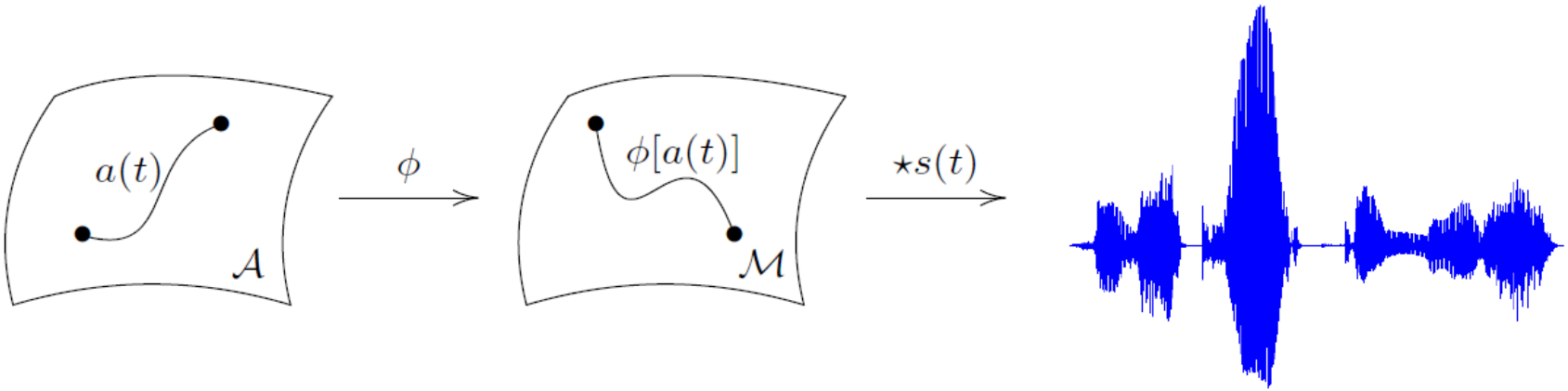
- The possible speech spectra are restricted to nonlinear subspaces

A Simple Model

$\approx$



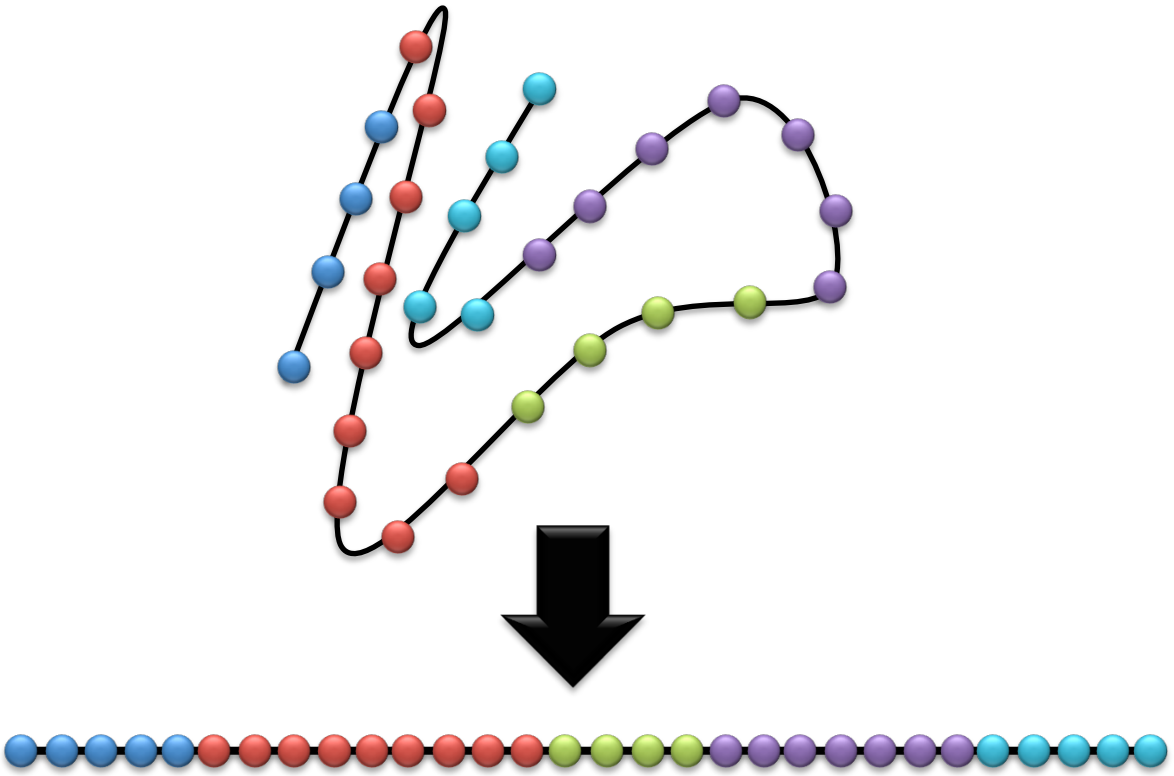
Low Dimension Manifold of Speech Sounds



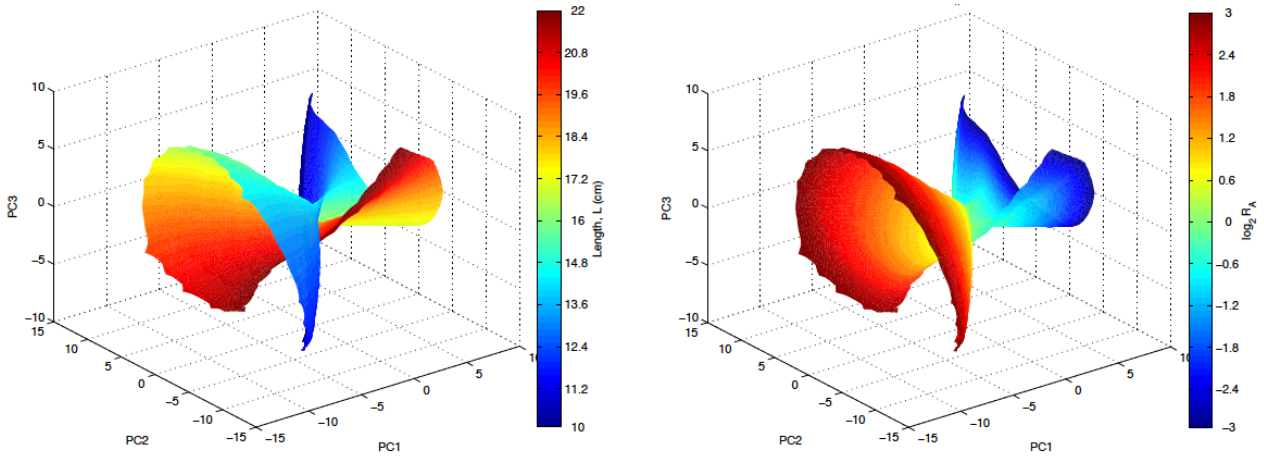
- $\mathcal{A}$ = set of vocal tract articulatory configurations
- $\mathcal{M}$ = set of vocal tract transfer functions
- Physics $\Rightarrow \phi : \mathcal{A} \rightarrow \mathcal{M}$ is a diffeomorphism
- Low $\dim(\mathcal{A}) \Rightarrow \mathcal{M}$ is a low-dimensional manifold

Manifold Learning

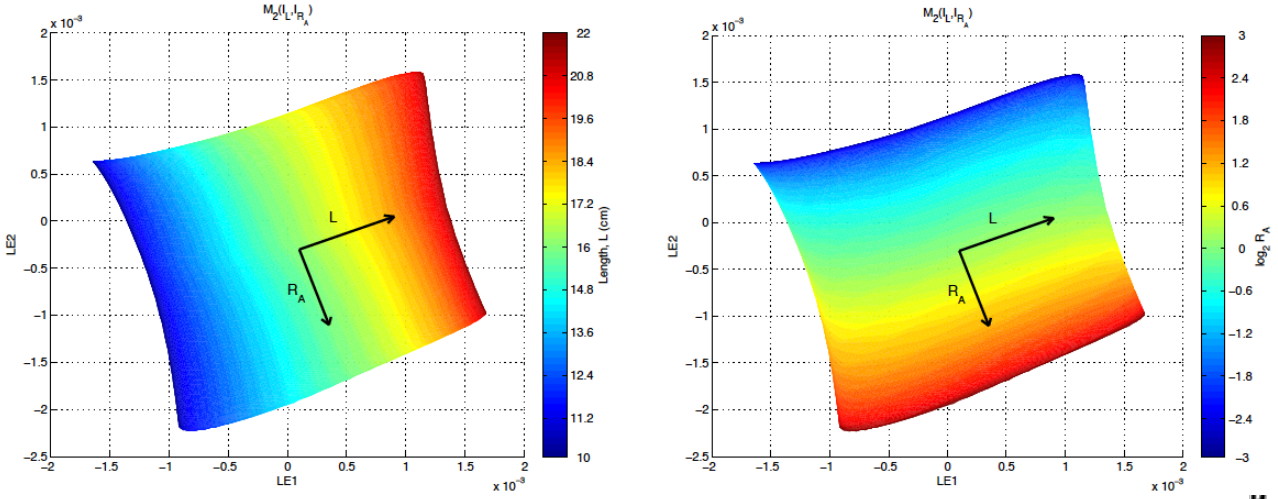
- Approximate nonlinear projection maps for an embedding that respects geodesic distance on the manifold (unsupervised)



The Vowel Manifold



Manifold Learning



Intrinsic Spectral Analysis

[Jansen & Niyogi, ICASSP 2006]

- Construct nearest neighbor graph over the unlabeled data sample and compute the (normalized) graph Laplacian:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

- Eigenvectors of the graph Laplacian define projection for in-sample data, eigenvalues define smoothness

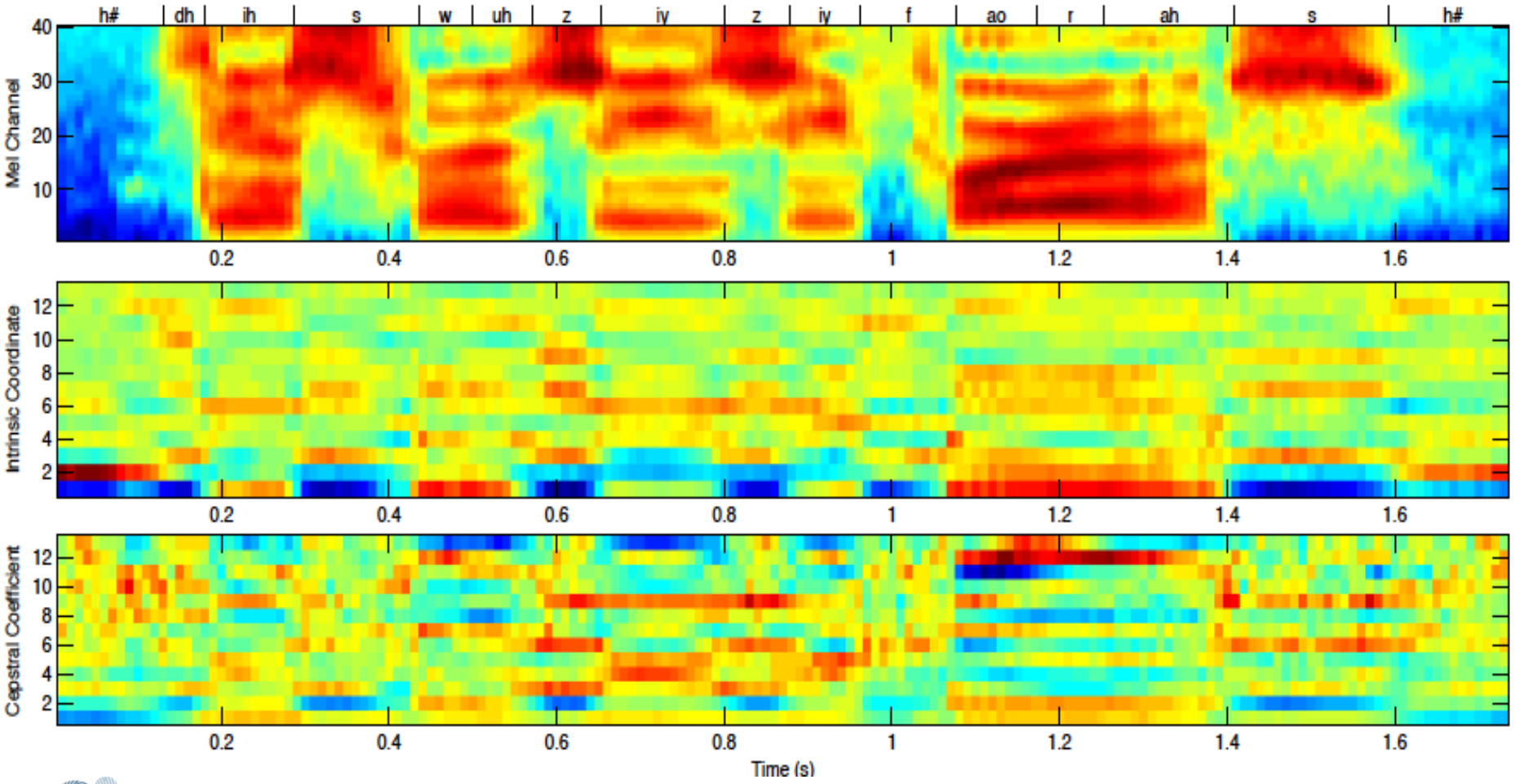
- Extend out-of-sample with kernel methods (essentially interpolation)

$$f^* = \arg\min_{f \in \mathcal{H}_K} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L}\mathbf{f},$$

$$f_j^*(v) = \sum_{i=1}^{n} \alpha_i^{(j)} K(x_i, v)$$

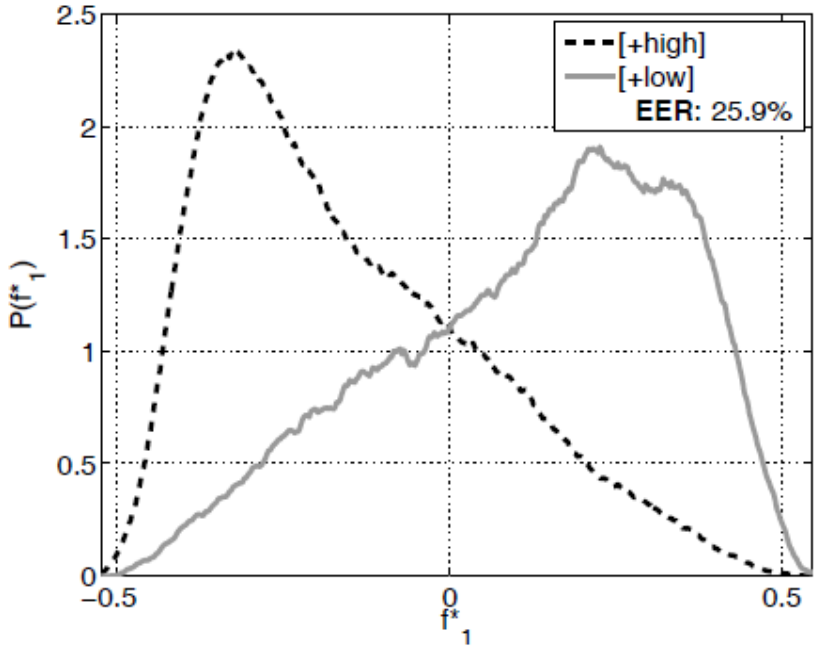
Intrinsic Spectral Analysis

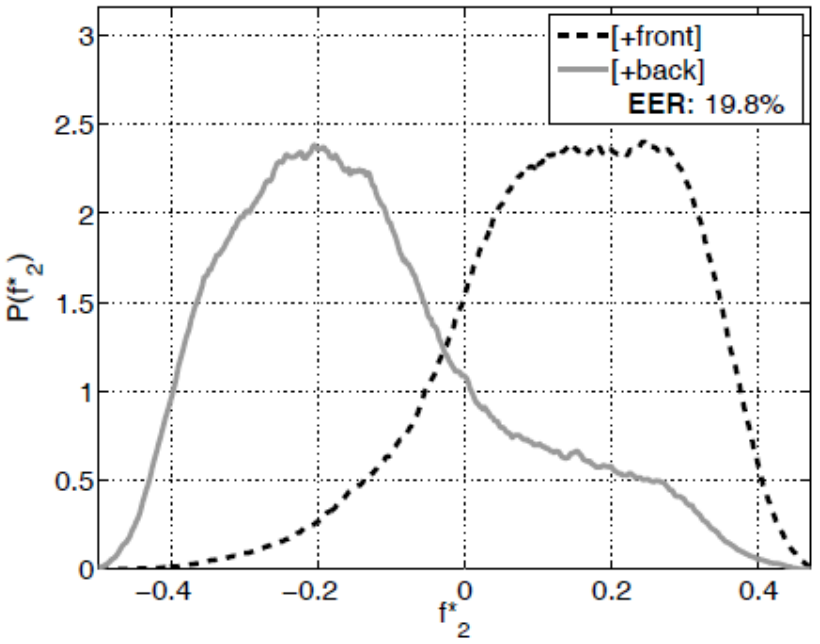
[Jansen & Niyogi, ICASSP 2006]



Inverting the Production Mechanism

- The intrinsic coordinates are articulatory (distinctive) feature correlates





A Speaker Independent Front-end

[Jansen, Thomas, & Hermansky, Interspeech 2012]

- **ISA:** computed from log mel spectrogram, **fully unsupervised**
- **MFCC:** log mel spectrum + cepstral analysis (equiv. to PCA)
- **Evaluation:** TIMIT corpus, same/different task

Features	Average Precision
MFCC	33.8
ISA	48.5
English AM	75.4

← **35% of gap**

PLUS: No Harm Done in Highly Supervised Setting

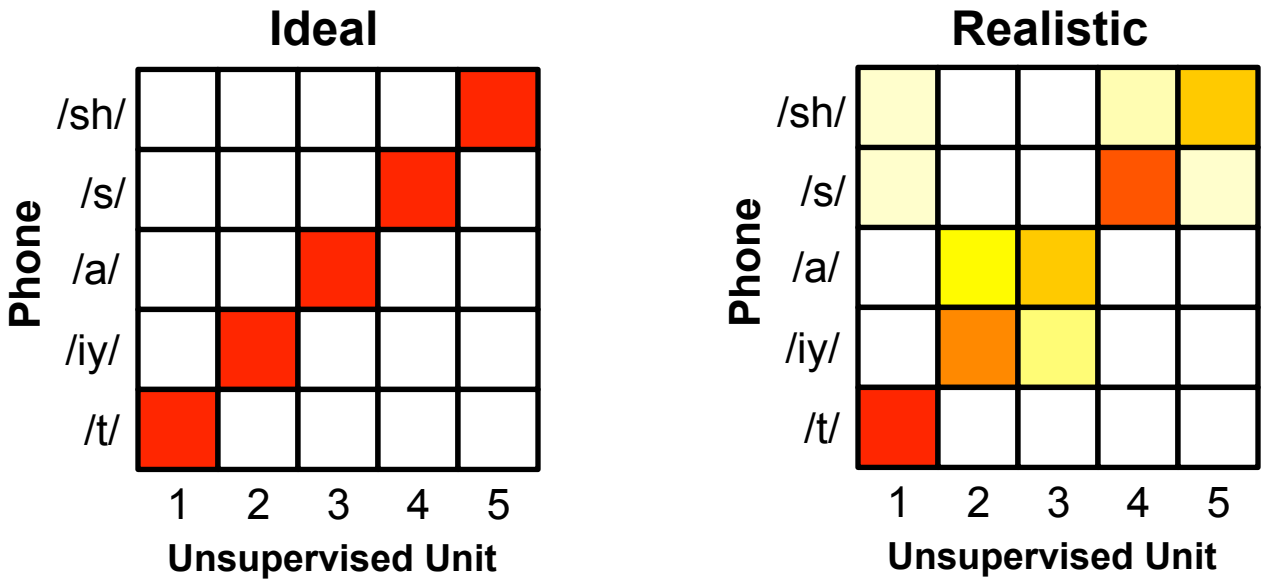
Part 2

PHONETIC DISCOVERY

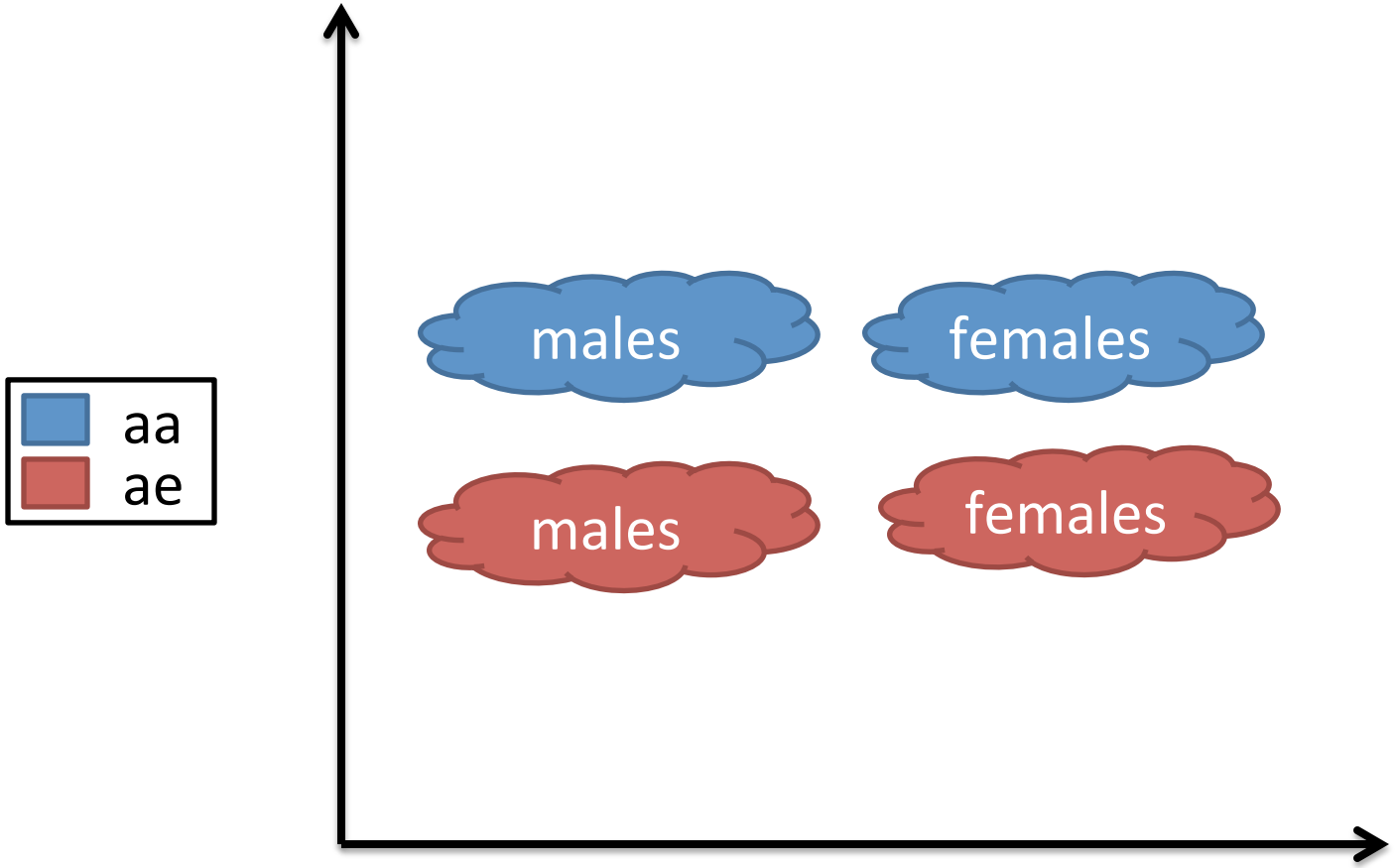
[Jansen & Church, Interspeech 2011]

Phonetic Structure

- Want units that map similar speech sounds produced by **different speakers** to the same categorical subword unit (or at least to same distribution over units)

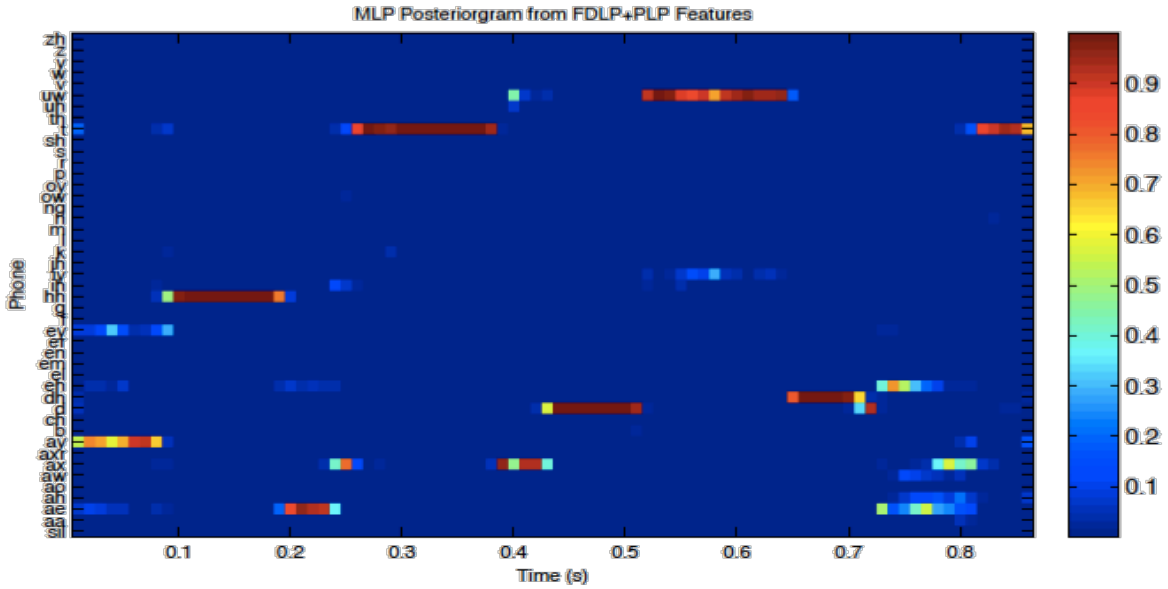


Why This is Very Hard



A Note about Tokenization

- **Ideal:** Produce a phonetic transcript
  - speaker independent tokenization
- **Easier goal:** Produce a phonetic posteriorgram
  - Speaker independent **distribution** over unit set
  - Supported by DTW-based word discovery methods, but not Bayesian methods (yet!)

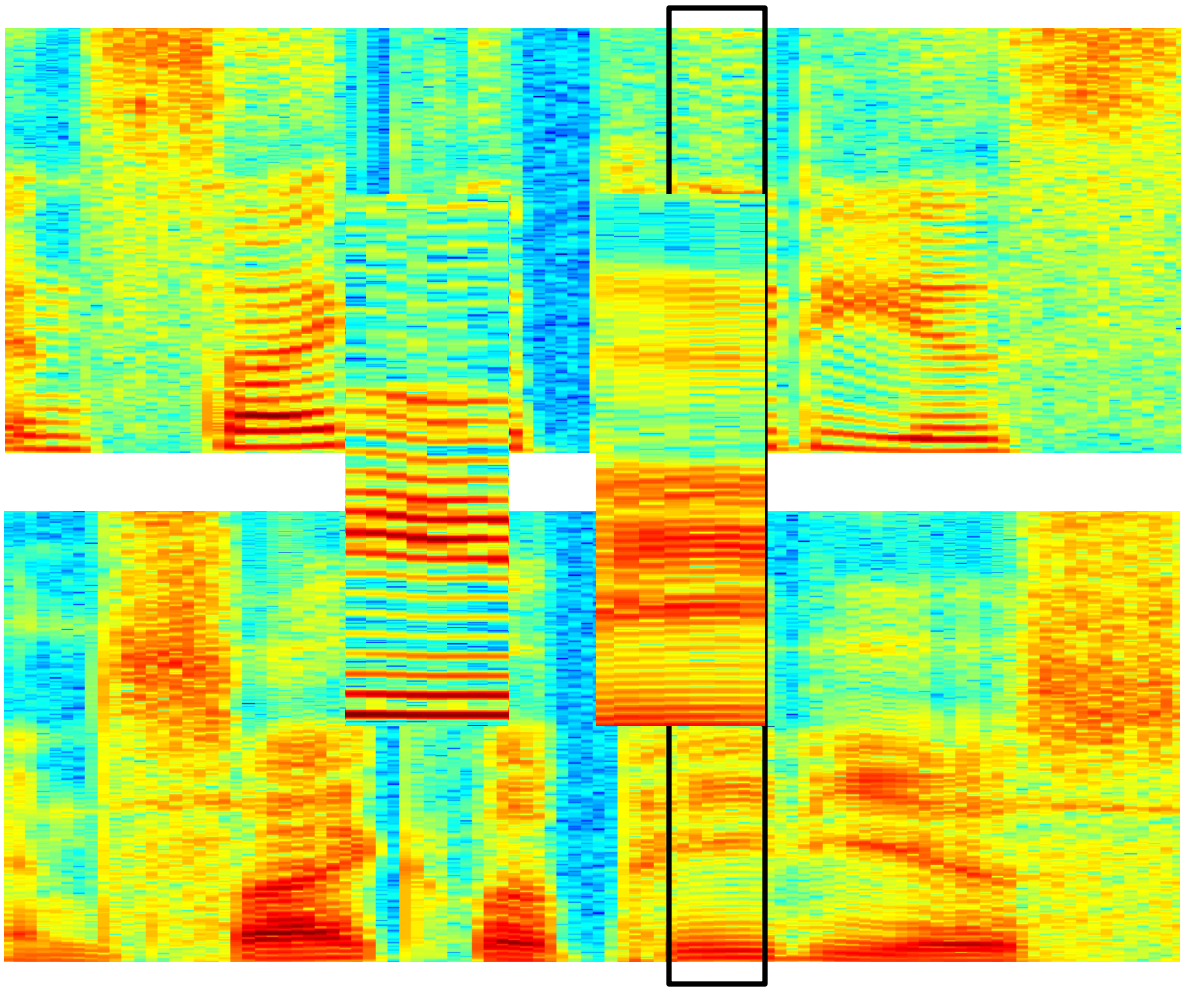


Related Work

- Bottom-up EM-style unsupervised training procedures
  [Garcia & Gish, 2006], [Varadarajan et al., 2008], [Siu et al., 2010], [Zhang & Glass, 2010], [Lee and Glass, 2012]

- **Not Addressed:** Speaker independence constraints to learn phone-like units for *consistent output across speaker*
- **Our Idea:** Exploit saliency of whole word patterns to provide top-down speaker independence constraints

Same or Different?



“encyclopedias”

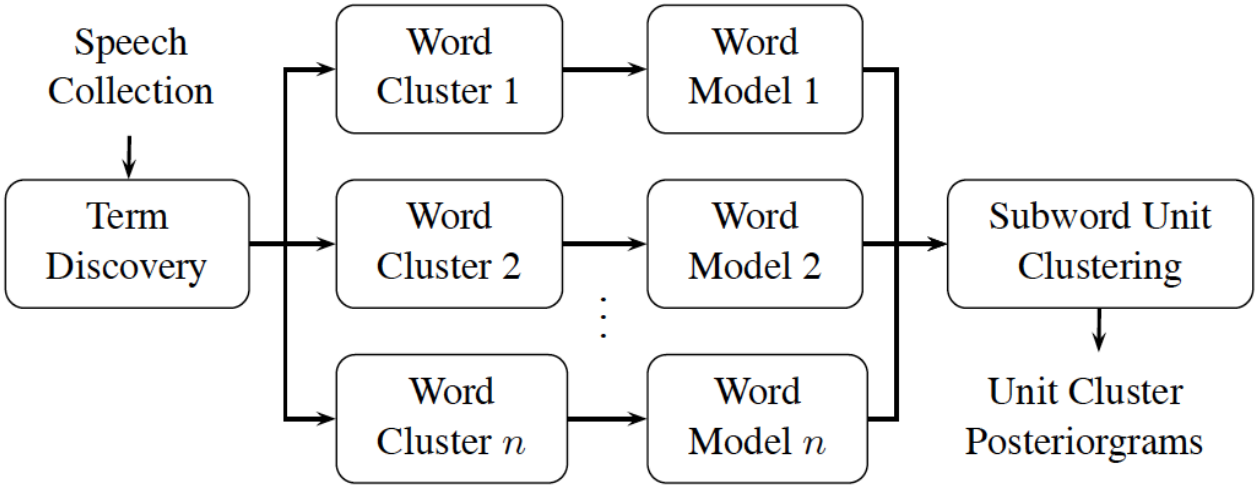
A Top-Down Strategy

[Jansen & Church, 2011]

1. Discover words repeated throughout the collection by multiple speakers (i.e., do the easier part first, now that we can afford it computationally)

2. Learn a subword unit acoustic model that decodes consistent subword unit sequence for all instances of each (unknown) word type

Unsupervised Training Procedure



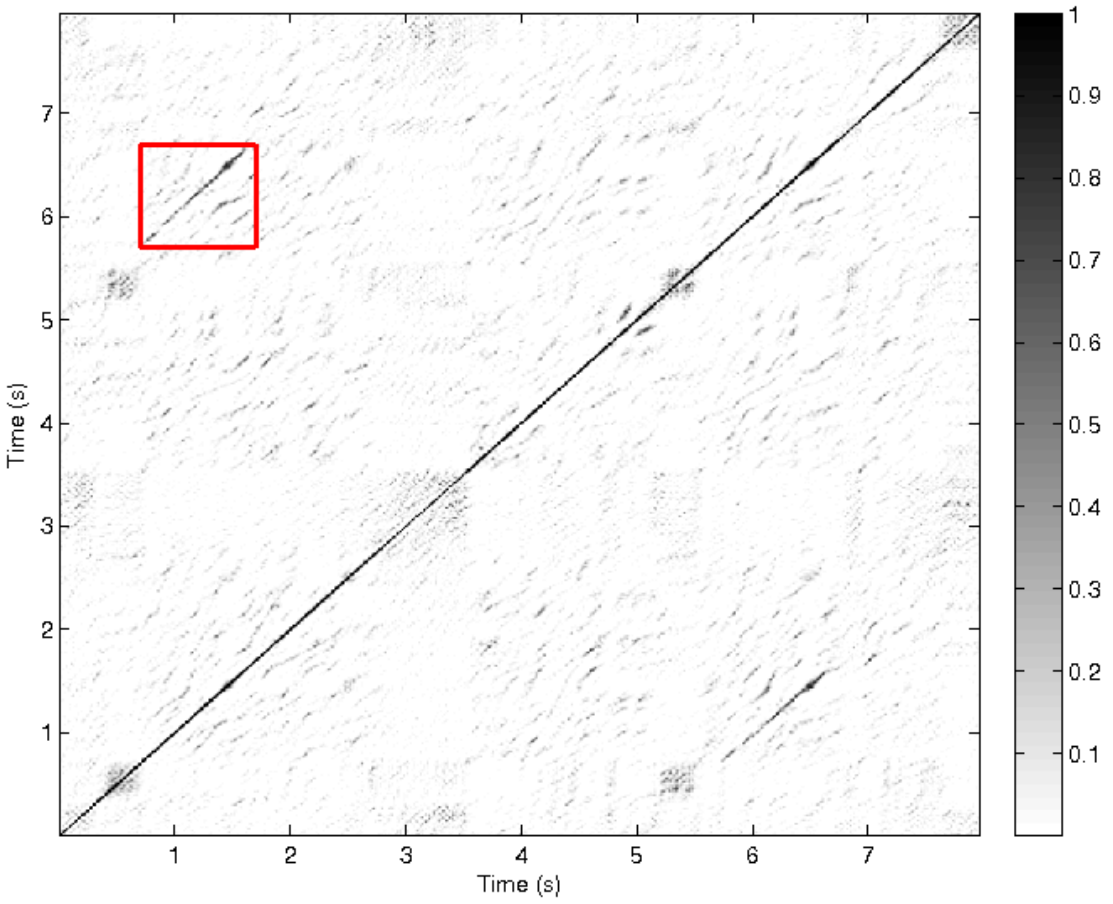
Step 1: Run spoken term discovery across corpus

Step 2: Use each discovered word/term cluster to train a whole-word HMM with GMM emission densities

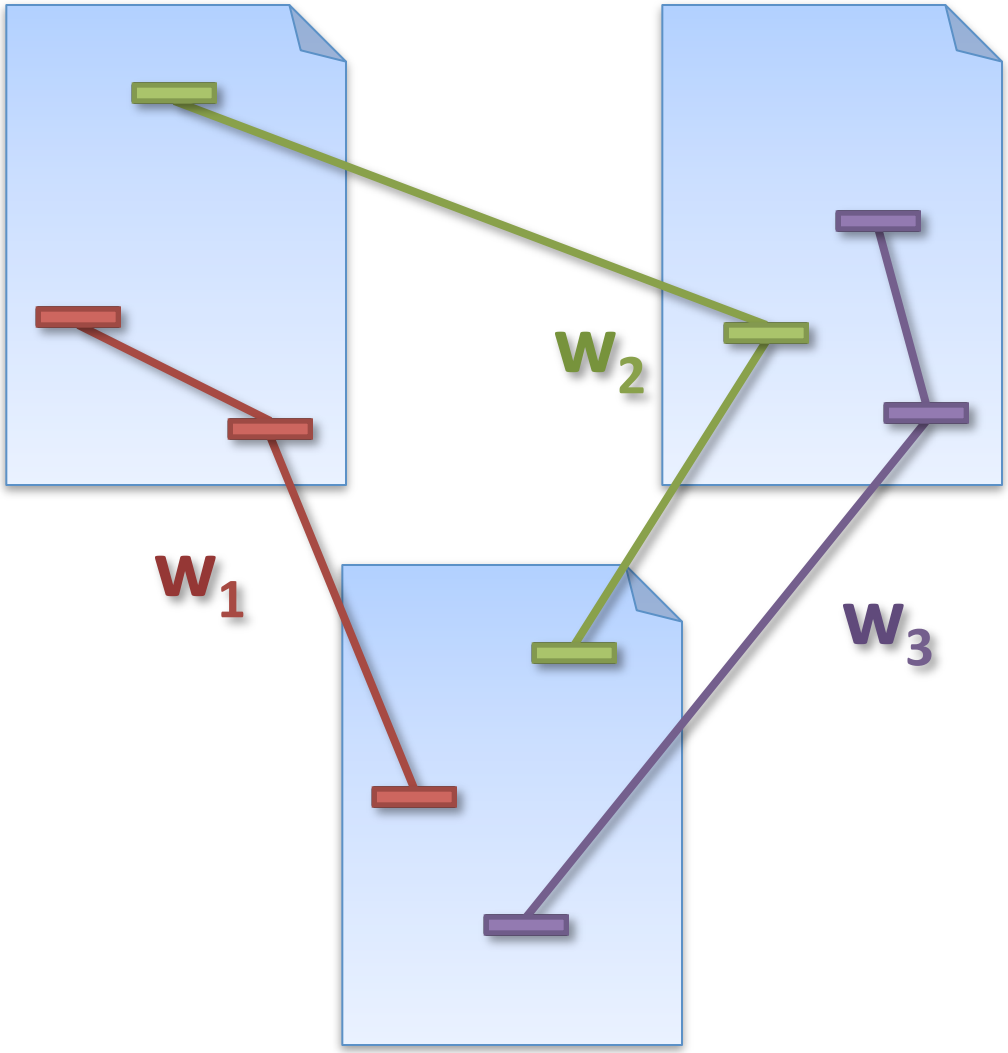
Step 3: Cluster HMM states across word HMMs to produce speaker independent subword unit models (spectral clustering)

Step 4: Compute posteriorgrams using state cluster GMMs and optionally use to repeat steps 1-3

Step 1a: Spoken Term Discovery



Step 1b: Cataloging Pseudo-terms

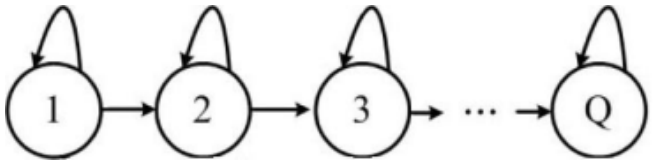


Step 2: Training Whole-Word HMM-GMMs

- **Assume:** all instances in each word cluster share the same unknown sequence of subword units

1. Predict the appropriate number of units for each word cluster

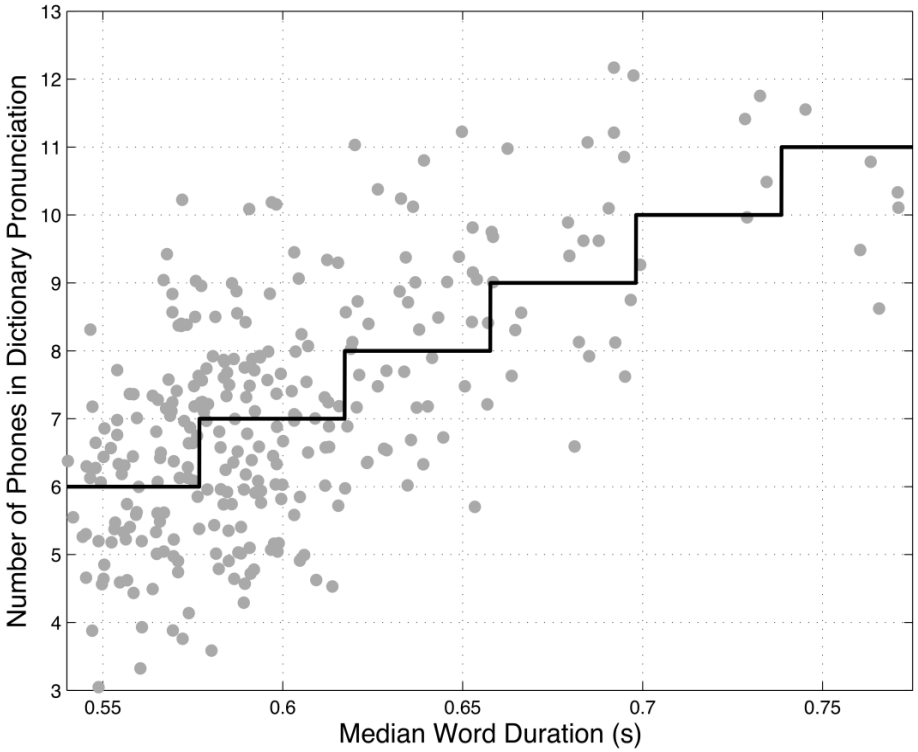
2. Model each word with a simple left-to-right HMM with one state per unit and one 8-component GMM per state



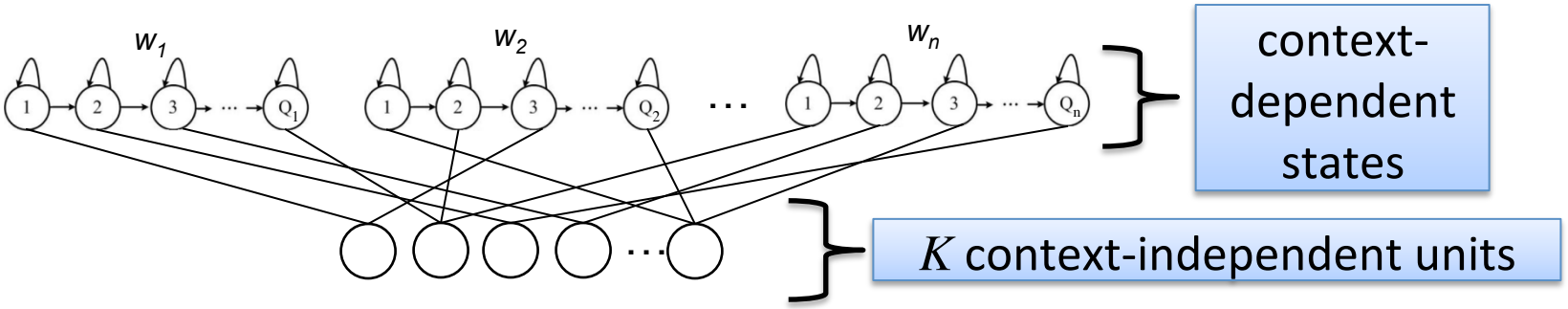
3. Perform standard Baum-Welch training

Predicting the Number of States

Simple: linear regression on median word duration



Step 3: Clustering Subword States Across Word Types



1. Discard the state priors and transition probabilities

2. Given a large sample of speech $X=x_1 \ldots x_T$ construct a state similarity (modified correlation) matrix *M*

3. Using the similarity matrix *M*, perform spectral clustering [Shi & Malik, 2000] into *K* context independent classes

Results

- **Proposed:** posteriorgrams over unsupervised subword units
- **PLP:** standard speech recognition features
- **Evaluation:** Telephone speech, same/different task

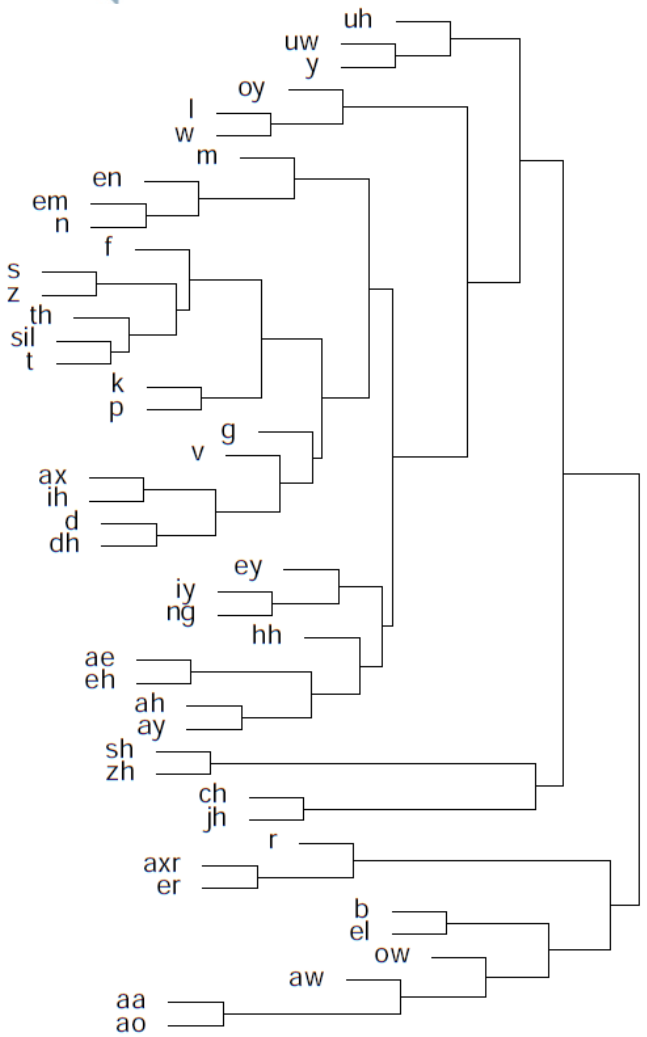
Features	Average Precision
PLP	16.9
Proposed	31.2
English AM	51.6

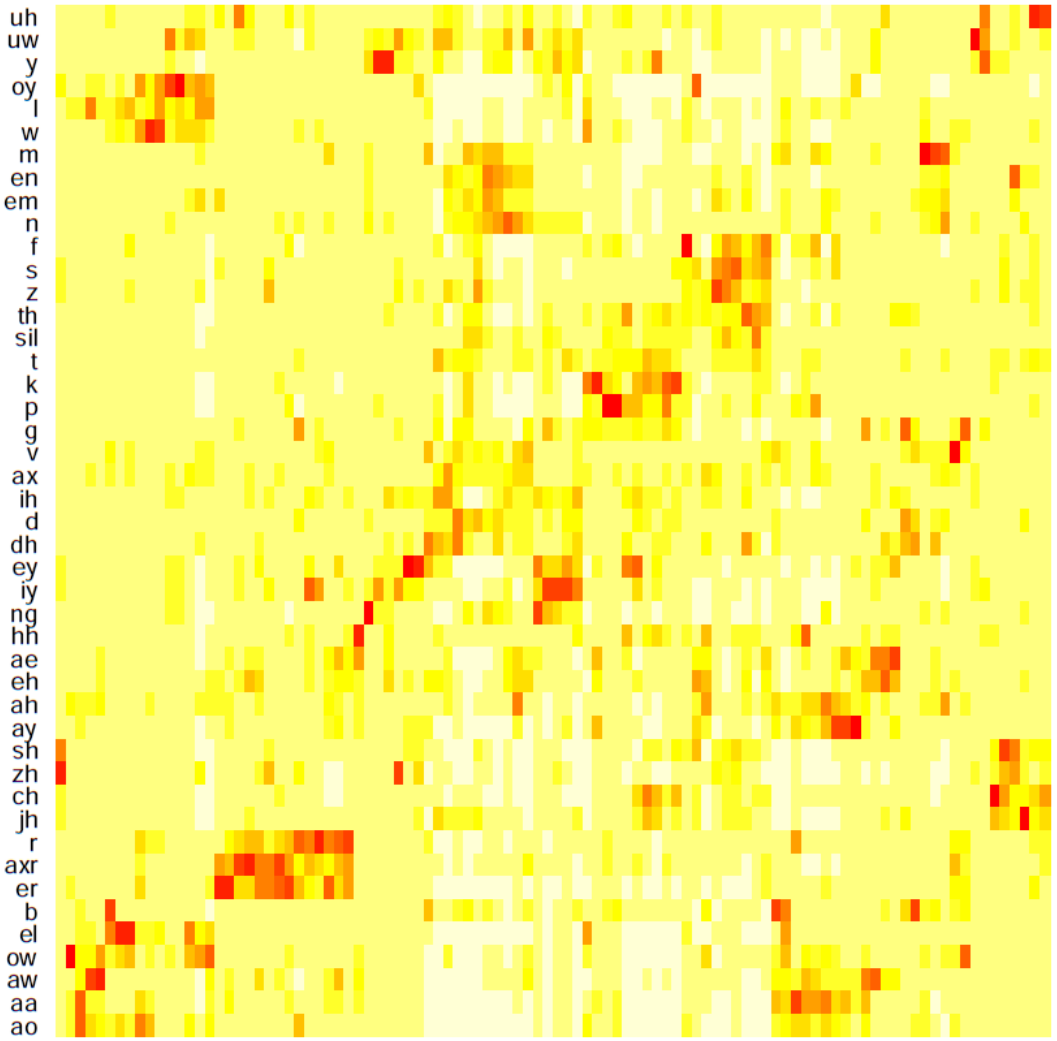
Unsupervised

Supervised (200 hr)

Our method bridges up to 41% of gap between raw features and fully supervised AM!

Did we learn phones?





Similar phones map to similar unsupervised subword units

human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Conclusion

- Unsupervised learning at each level of analysis need not be performed in isolation
- Words are the most salient entry point and top-down constraints have clear value
- How can we do better?
  - deep learning
  - nonparametric Bayesian modeling
  - more cross level constraints