Zero Resource Speech Technologies: An Overview

Aren Jansen

JOHNS HOPKINS
UNIVERSITY

JHU CLSP Summer Workshop 2012
July 16, 2012

The Golden Age of Speech Recognition

- **iPhone's Siri:** is amazing (near perfect on me so far)
- **Google Voice Search:** 87,000 hours of transcribed speech (2000 manual + 85,000 automatic)

But…

- The success story fades when you consider a new (not seen in training):
  - Language, dialect, or accent
  - Domain
  - Channel/Environment

- **State-of-the-art:**
  - English: **15% error rate**
  - Mandarin: **30% error rate**
  - The same technology trained/tested on Cantonese: **70% error rate**

An Opportunity

- Transcribed speech is requires **time** and **money**
- Untranscribed speech is **unlimited** and **free:**
  - YouTube alone receives 60 hours of video on average every single minute (30 million hours per year)

The Zero Resource Setting

1. **No** transcribed training data
2. **No** language-specific models
3. **No** pronunciation dictionaries
4. **No** knowledge of what language it is

   (in extreme cases)

The Challenge

How can you automatically discover **linguistic structure** to aid downstream speech technologies?

Speech is Rich with Structure

Semantic: {book, reference, knowledge, wikipedia}

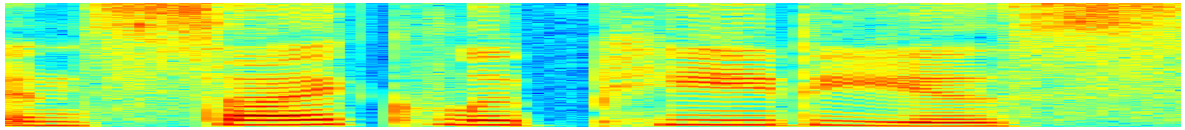
Grammatical: (he sold, NP, to her)

Lexical: encyclopedias

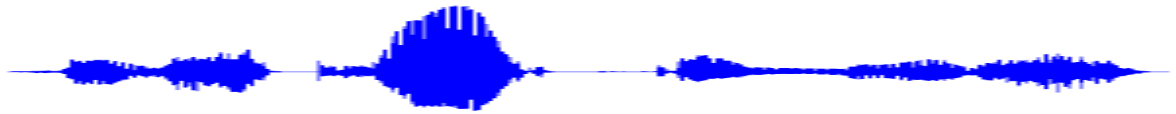
Phonetic: en s ai k l ow p iy d iy aa s

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

Acoustic:



Observed:

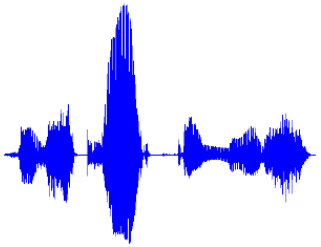


Applying Unsupervised Learning

- **In general, free to choose:**
  - 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:** 

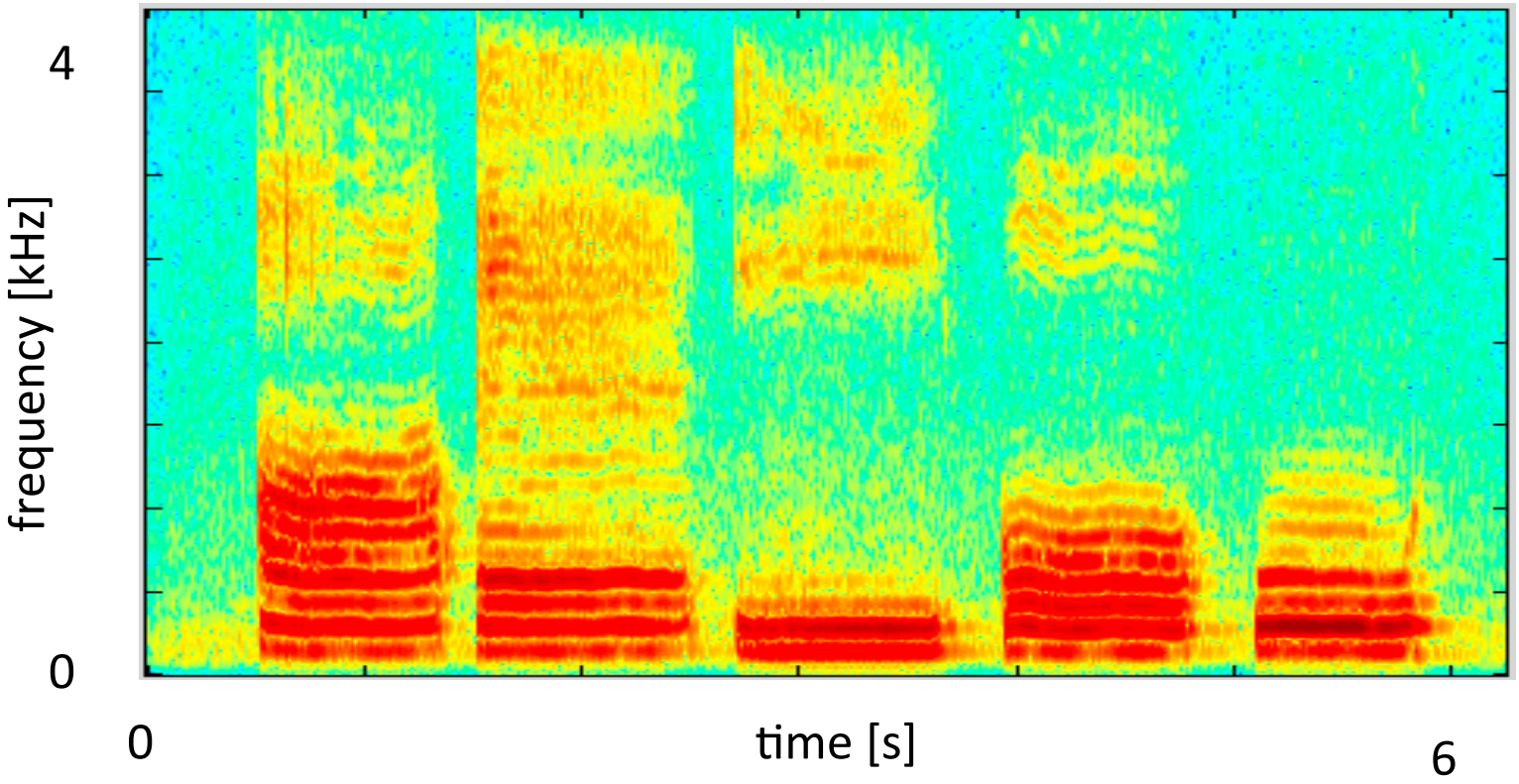
- **Scales of Analysis:**
  1. **Featural:** learn map from windowed signal to points in R^d
  2. **Phonetic:** learn to associate points in R^d with categorical subword unit inventory
  3. **Lexical:** learn to associate trajectories in R^d or strings of subword units with categorical words and phrases
  4. **Semantic:** learn to relate words and phrases according to topical content

Outline

- **Part 1:** Lexical Discovery
- **Part 2:** Phonetic Discovery
- **Part 3:** Semantic Discovery
- **Part 4:** Applications

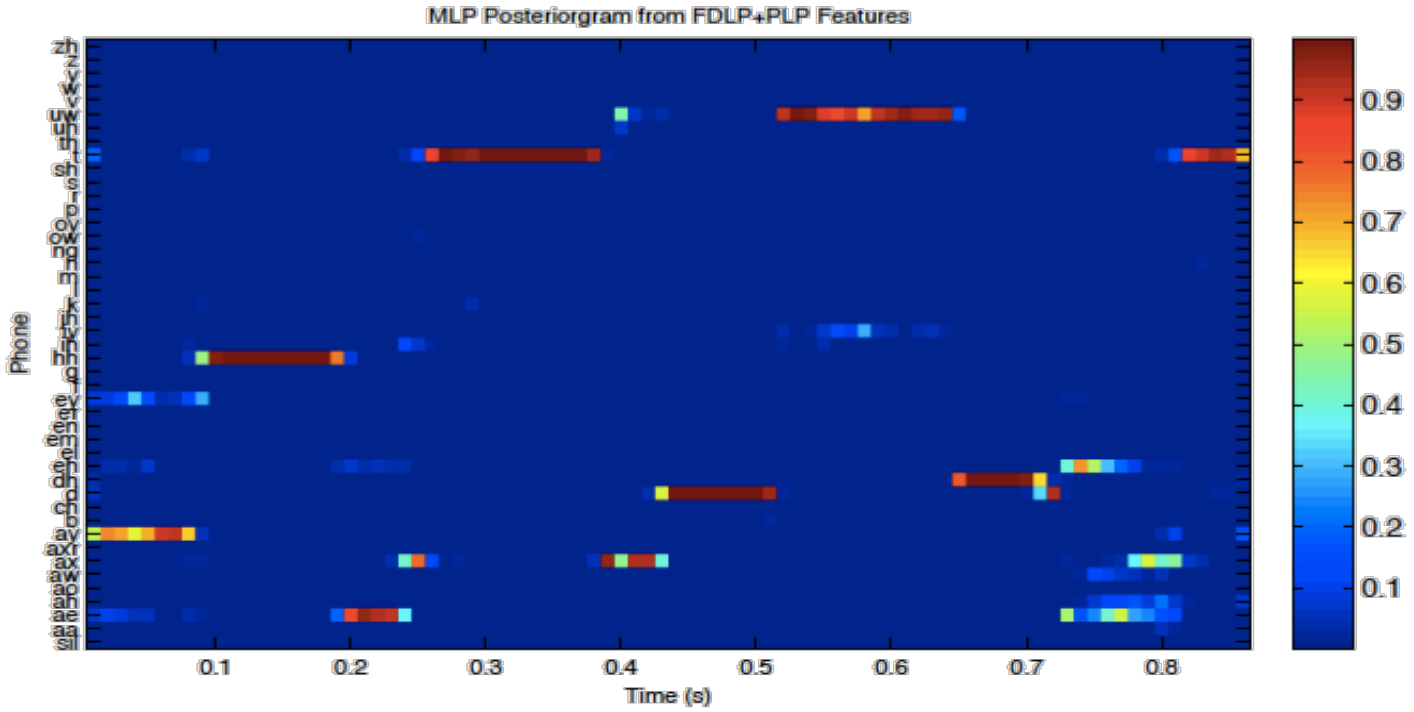
Preliminaries: Speech Features

- **Acoustic features:** PLP & MFCC = spectrograms + some post-processing



Preliminaries: Speech Features

- **Posteriorgrams:** posterior distribution over categorical units (e.g. phonemes) as a function of time



Part 1

LEXICAL DISCOVERY

Early Work from Speech Community

- **Multi-modal Computational Models:**
  - Discovery constrained by action in call routing system

    Allen Gorin, Stephen Levinson, et al. "An Experiment in Spoken Language Acquisition." *Trans. Speech and Audio Proc.* (1994)
  - Discovery constrained by vision

    Deb Roy. "Learning Words from Sights and Sounds: A Computational Model." PhD Thesis (1997)
- **Provided Noisy Phonetic Tokenization:**
  - Phone n-gram frequencies

    Allen Gorin et al., "Learning Spoken Language without Transcriptions", in *Proc. ASRU* (1999)

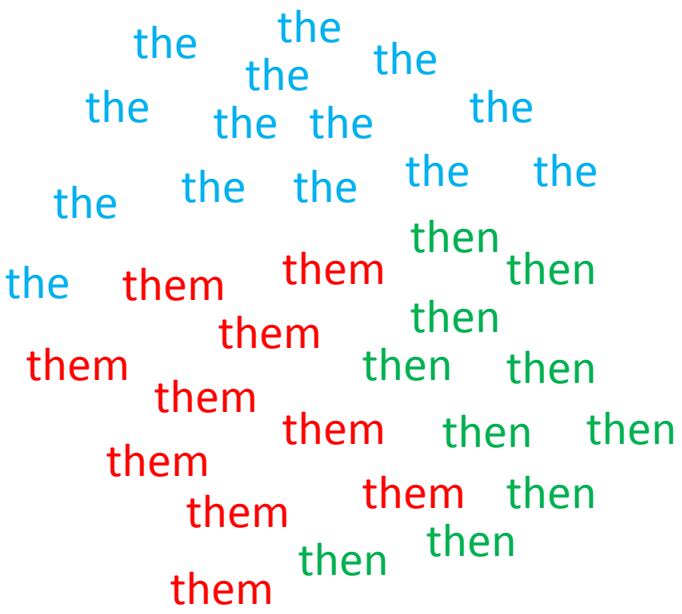
Lexical Discovery From Subword Unit Tokenizations

- Parallel literature in the computational linguistics literature involving word segmentation from token sequences

- **Make sure to attend:**
  - **Mark Johnson's** talk this afternoon at 2:30pm
  - **Sharon Goldwater's** talk tomorrow at 9:50am

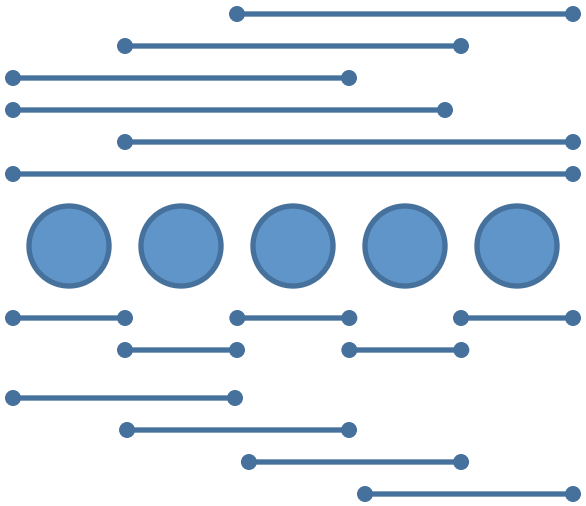
Words in Speech: No Longer Fixed Dimension

- Consider the space of spoken word utterances (not a vector space)
- Define a suitable distance metric characterizing acoustic phonetic similarity
- Perform clustering in this space



The Challenge

- The word segmentation is not provided!



- In *N* frames of speech, have O(*Nd)* possible intervals that can contain a word or phrase of max length *d*

The Key: Repetition

- Intervals corresponding to words and phrases are repeated:

  the king is dead long live **the king**

- Most of the intervals are not words and phrases and won't repeat (as much):

  thek ingis de adlon gli vet heking

- **Problem:** *Nd* possible intervals requires $O(N^2)$ interval comparisons

Prior Work in Unsupervised Word Discovery

- **Two Main Approaches:**

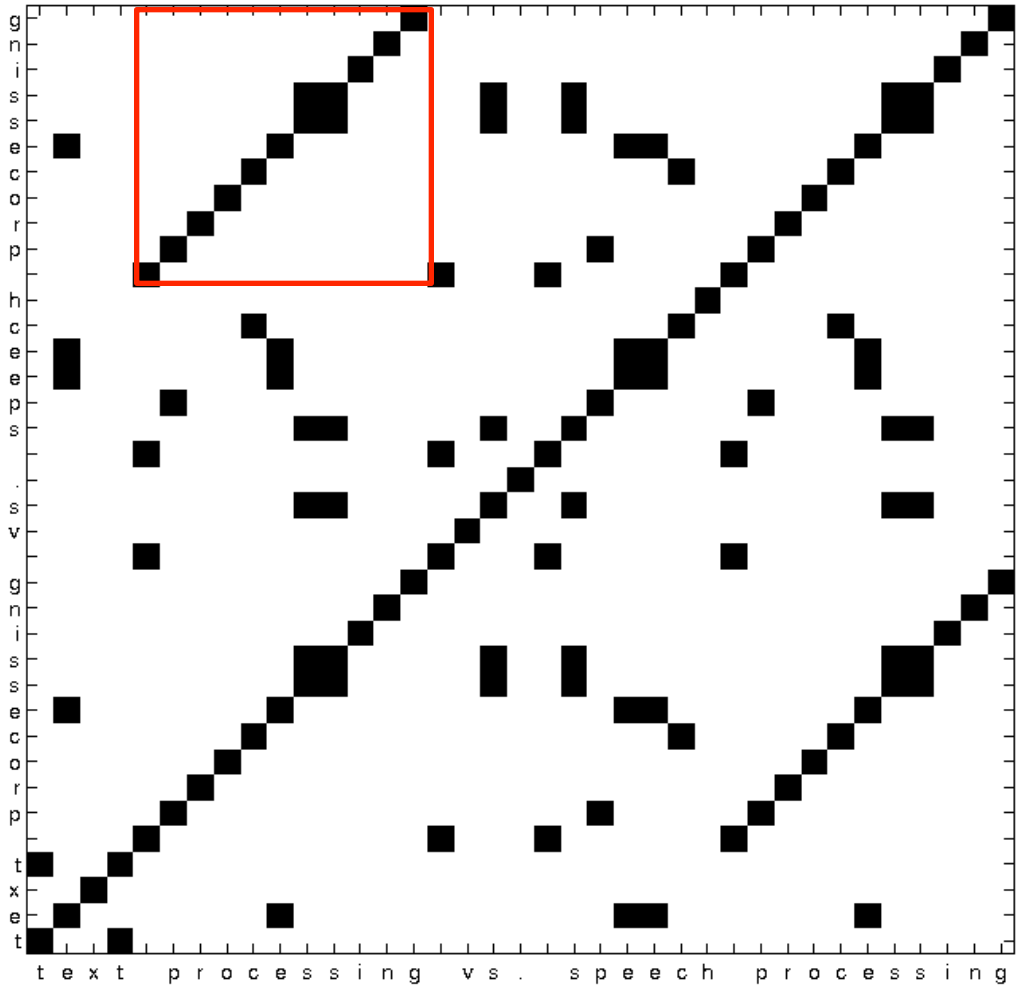
1. Search for repeated trajectories in acoustic feature space
   - **MFCC/PLP/FDLP:** [Park & Glass, TASLP 2008], [Muscariello, Gravier, & Bimbot, Interspeech 2009], [Jansen & Van Durme, ASRU 2011]
   - **GMM Posteriorgrams:** [Zhang & Glass, 2010]
   - **Mismatched Language Posteriorgrams:** [Jansen, Church, Hermansky, 2010]

2. Decode with unsupervised acoustic model and look for repeated subword unit sequences (produced by unsupervised acoustic model)
   - [ten Bosch & Cranen, Interspeech 2007]
   - [Siu, Gish, Lowe & Chan, Interspeech 2011]

*** ACORNS project**

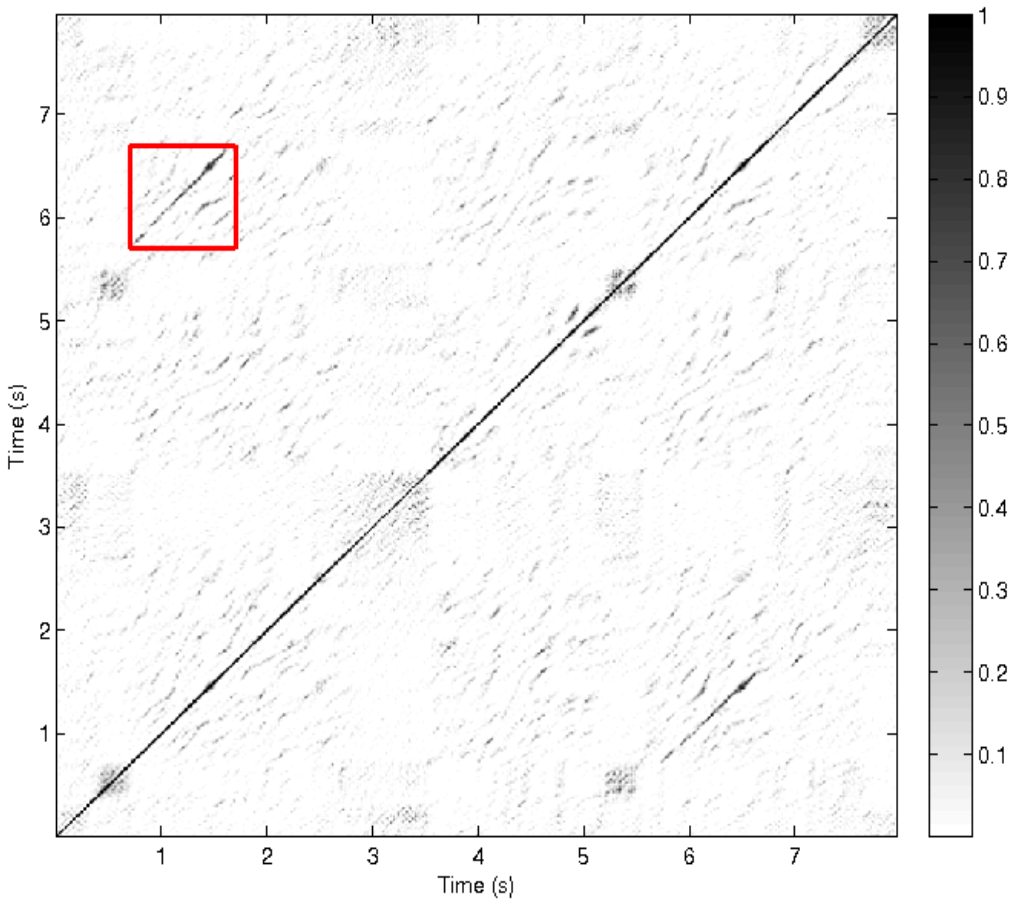
Text Dotplots

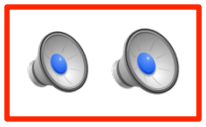
[Church & Helfman, 1993]



"*text processing vs. speech processing*"

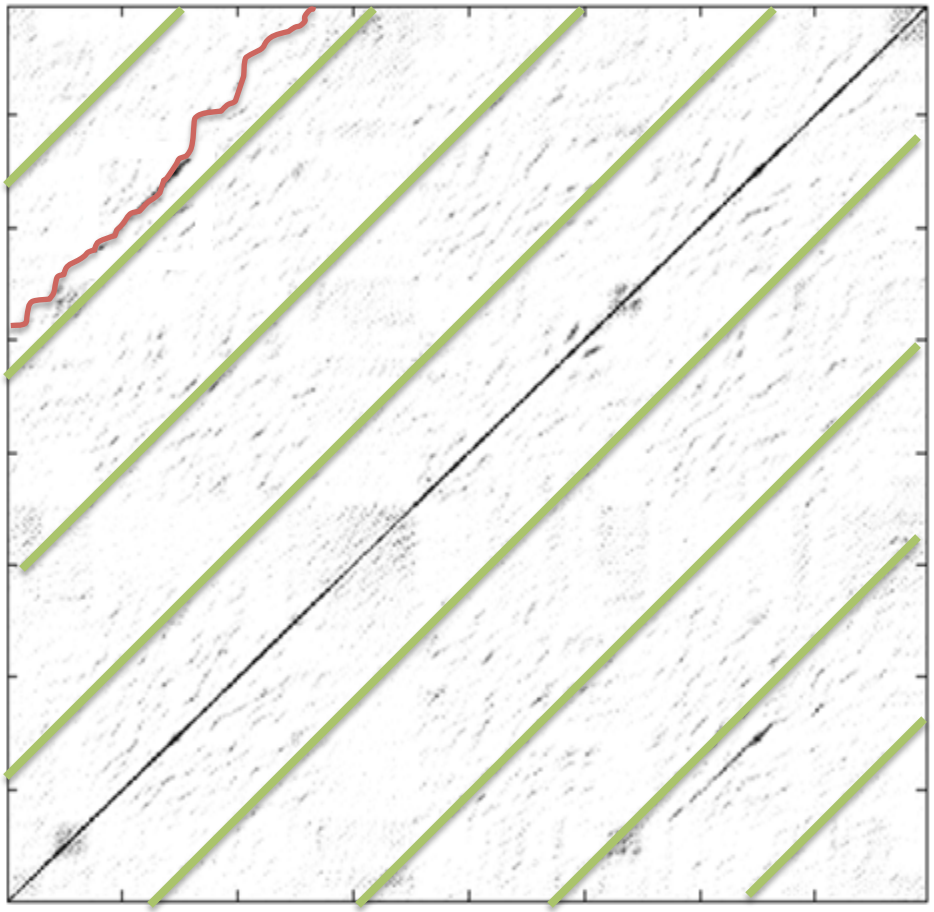
Acoustic Dotplots



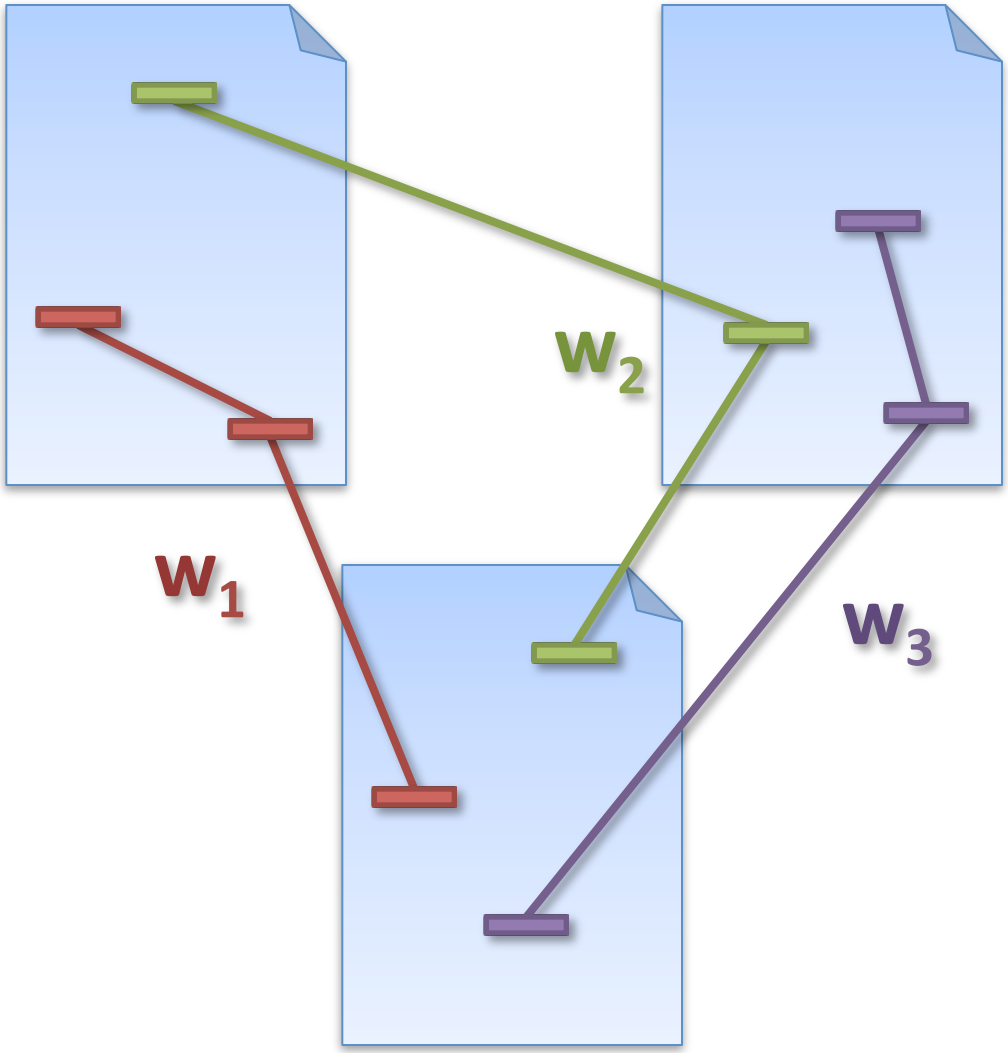


Segmental Dynamic Time Warping

[Park & Glass, 2008]



Pseudo-word Clusters



Limitation #1: Speed

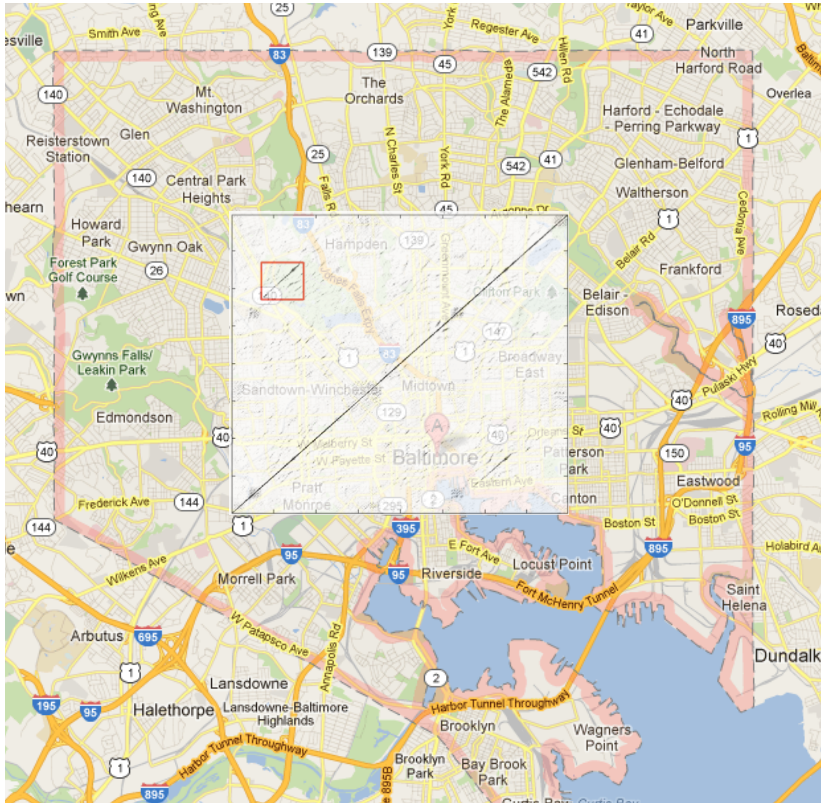
Two Computational Bottlenecks:

1. Compute the self similarity matrix (acoustic dotplot)

2. Search the similarity matrix for matching paths

Current Scalability State-of-the-art

[Jansen & Van Durme, ASRU 2011]



60 hours = **3.5 mi x 3.5 mi** dotplot at monitor resolution (100 dpi)

Our Runtime: 6 hours on 100 cores

Exhaustive S-DTW: 3 months on 100 cores

Produces: Order 100k-1M units, not all distinct, not all words

Limitation #2: Speaker Independence

- **Task:** Multi-speaker same word/different word discrimination

Features	Average Precision	
	Same Spkr	Overall
PLP	52.3	19.0
English AM	59.8	40.2

← Raw Acoustics

← Supervised (200 hours)

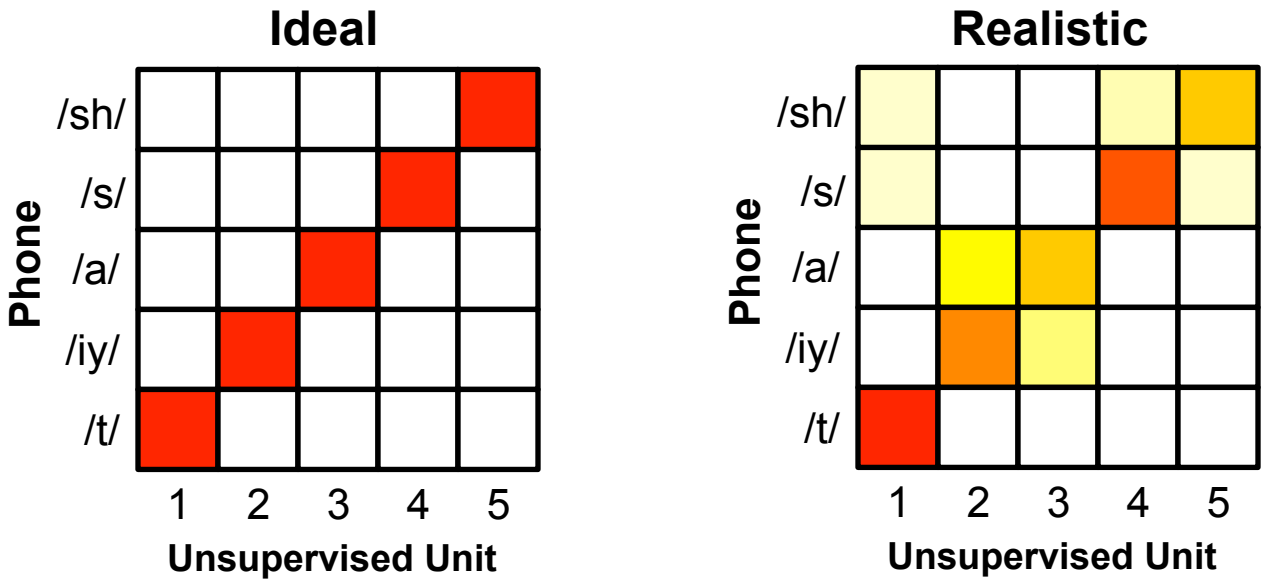
- **Efforts**
  - Self-similarity matrix-based distance metrics [Muscariello, Gravier & Bimbot, ICASSP 2011]
  - Post-processing of unsup. acoustic models [Anguera, ICASSP 2012]
  - Manifold learning [Jansen, Thomas, & Hermansky, Interspeech 2012]

Part 2

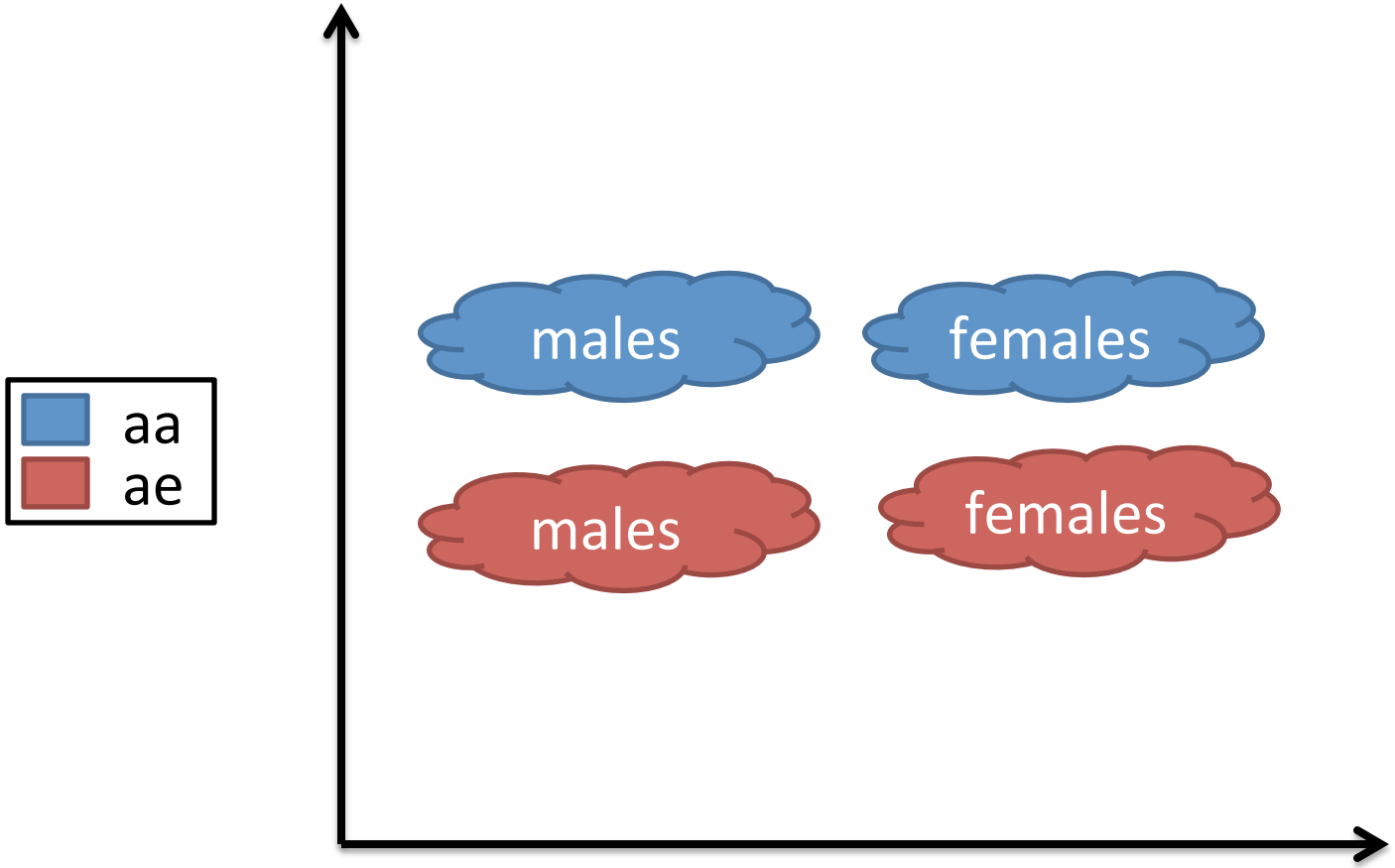
PHONETIC DISCOVERY

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



Bottom-up Unsupervised Training (1/3)

Segment, Cluster, & Model

1. Segment speech into stationary segments
2. Cluster segments
3. Train model for each cluster

Examples:

- Self-Organizing Units (SOUs)
  - [Garcia & Gish, 2006]
- Statistical Word Discovery (SWD)
  - [ten Bosch & Cranen, Interspeech 2007]
  - Recent segmentation work by Odette Scharenborg

Bottom-up Unsupervised Training (2/3)

Unsupervised Training of Standard ASR Components

- Universal Gaussian Mixture Models
  - Zhang & Glass, ASRU 2009
- Successive State Splitting (SSS): HMM-GMM training
  - Varadarajan, Khudanpur & Dupoux, ACL 2008

Bottom-up Unsupervised Training (3/3)

The Next Generation:

- Switching Linear Dynamical Systems
  - Bala Varadarajan, Sanjeev Khudanpur (talk tomorrow at 10:50a)
- Non-parametric Bayesian Modeling
  - Lee & Glass, ACL 2012 (Bayesian Segment/Cluster/Model)
  - Shinji Watanabe
- Deep Learning pre-training
  - Mike Seltzer
  - Yann LeCun, Fei Sha

Main Hurdle: Speaker Independence

- Bottom-up approaches predisposed to learn **speaker-specific** allophones
- **Task:** Multi-speaker word matching

Features	Average Precision
PLP	14.6
SOU Tokenization	12.4
GMM Posteriors	14.7
English AM (200hr)	45.6

Incorporating Top-Down Constraints

- **Jansen & Church, Interspeech 2011** (me, tomorrow 11:40a):
  1. Discover words from acoustics using S-DTW style algorithm
  2. Train whole word model for each (context dependent phones)
  3. Cluster context dependent phones across models
  4. Compute posteriorgrams and iterate

- **Non-parametric Bayesian version: Naomi Feldman's** talk tomorrow morning at **9am**

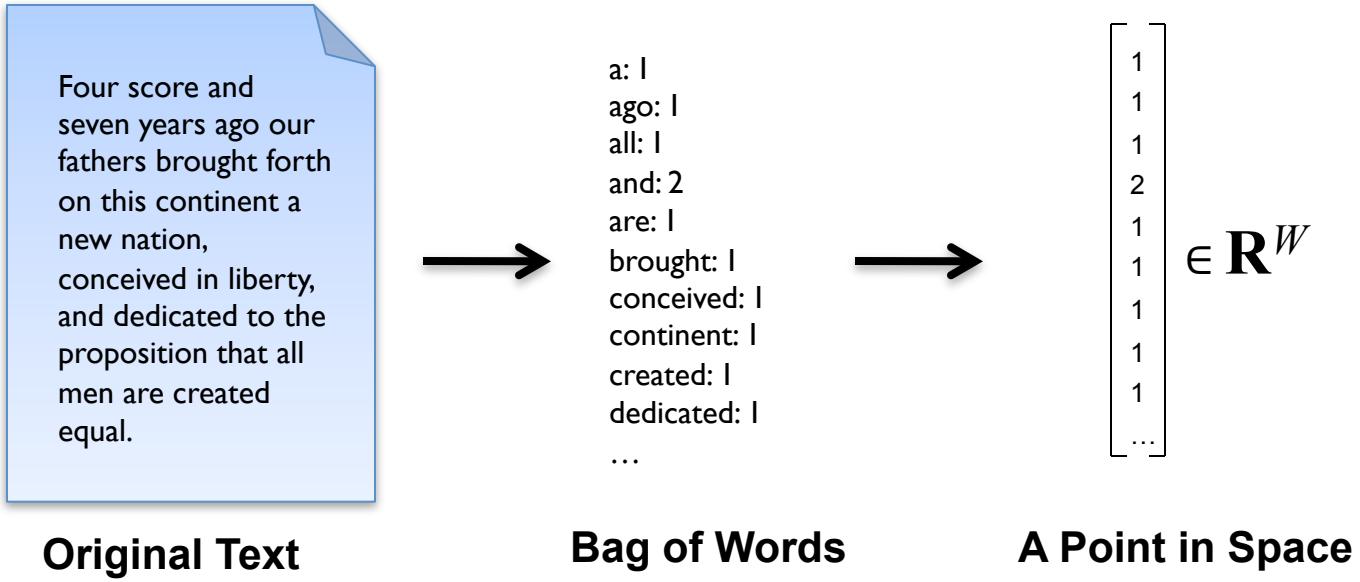
- **Hazen, Siu, Gish, Lowe & Chan, ASRU 2011**:
  1. Train acoustic model for SOUs and decode
  2. Discover pseudo-words in SOU transcript/lattice
  3. Augment LVSCR lexicon, train language model, iterate

Part 3

SEMANTIC DISCOVERY

Words are Defined By the Company They Keep

- **The Trick:** Characterize document content w/o understanding
- **The Computation:** Convert text into a "Bag of Words"



Bags of SOU n-grams/Pseudo-terms

[Gish, Siu, Chan, Belfield – Interspeech 2009]

[Dredze, Jansen, Coppersmith & Church -- EMNLP 2010]

Regular
Text Document

Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.

→ $v \in \mathbf{R}^W$

Actual Lexical Labels

Pseudo-term
Spoken Document

P_4 P_6 P_9 P_{19} P_2 P_{12}

→ $v \in \mathbf{R}^P$

Label Placeholders

Part 4

APPLICATIONS

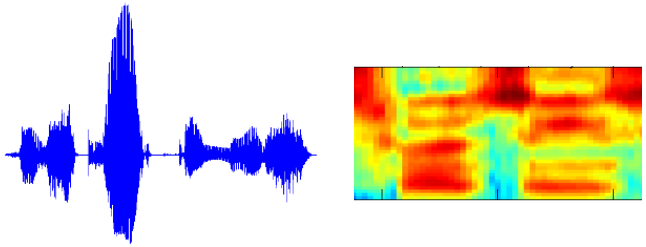
Query-by-Example (QbyE) Search

- Instead of a text query (word or phone string), you are given a short snippet of audio

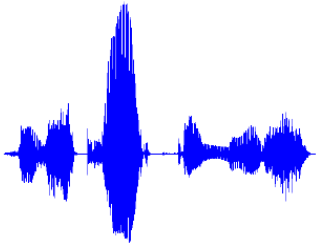
Normally

New York City
president
/p r eh z ih d eh n t/

Query-by-example



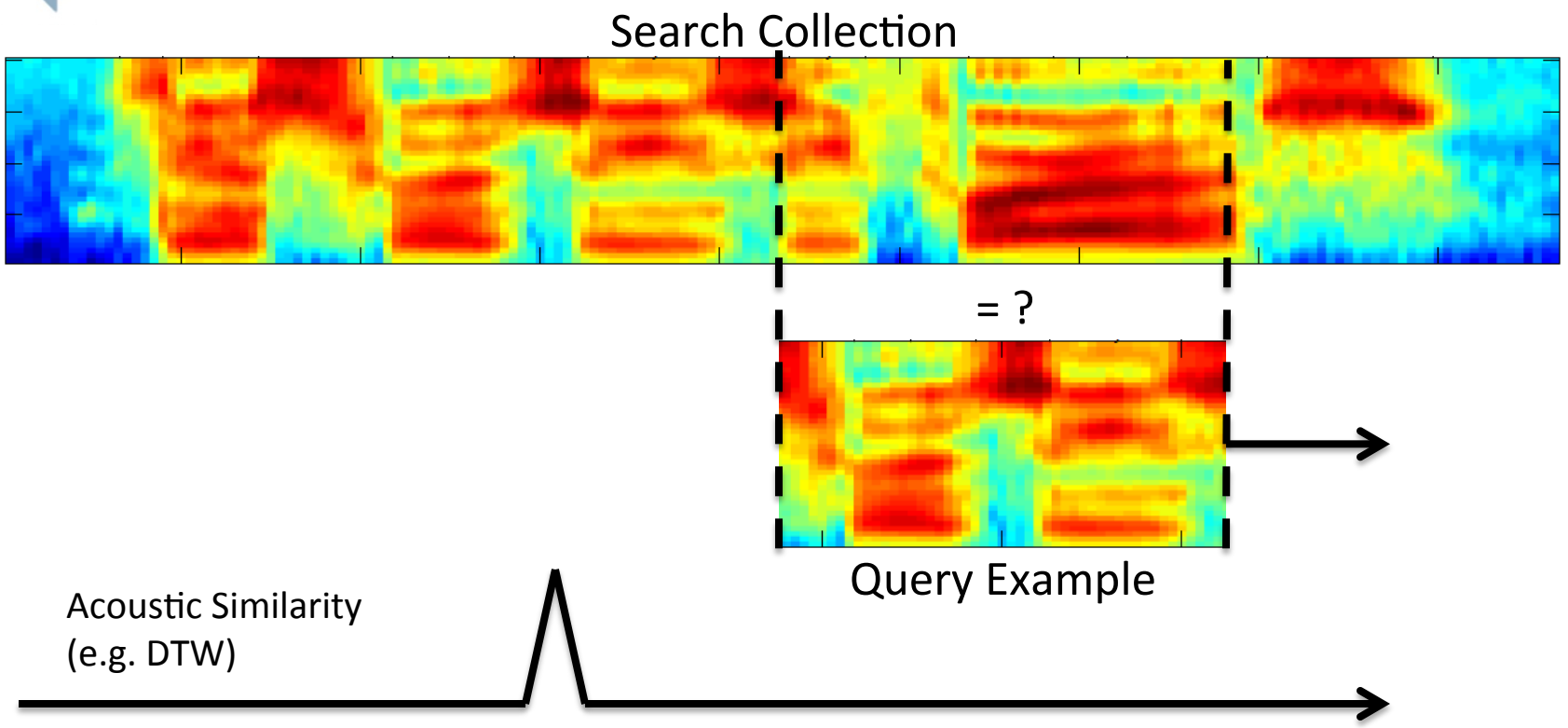
- **Past approaches:** Phonetically decode query and search index
  e.g [Shen et al, Interspeech 2009],[Parada et al, ASRU 2009]



/p r eh z ih d eh n t/

- **MediaEval 2011/2012:** Florian Metze, Xavier Anguera ,et al.

Zero Resource QbyE: DTW Search



- **Optimized DTW:** **200X real-time** [Zhang & Glass, Interspeech 2011]
- **Ported to a GPU:** **1,000X real-time** [Zhang & Glass, ICASSP 2012]
- **Randomized Approx:** **Up to 500,000X real-time** [Jansen & Van Durme, Interspeech 2012]

Topic Identification/Clustering

- Apply classification/clustering algorithms to bags of units vectors:
  1. **Bags of pseudo-words**

     Dredze, Jansen, Coppersmith & Church, EMNLP 2010
  2. **Bags of SOU n-grams**

     Gish, Siu, Chan, Belfield, Interspeech 2009

     Hazen, Siu, Gish, Lowe, & Chan, ASRU 2011

With a Little Annotation…

- Discovery methods learn units, but **does not provide labels**
- **Ideal case:** labeling a single example of each unit completes the process, enabling other applications
- **But:** as soon as you any labels, you are in semi-supervised regime (and other methods may be more suitable)

Area Most in Need of Attention
(my view as an engineer)

- **Speaker invariance** is most important issue at all levels of analysis for downstream technologies

- Unsupervised learning at each level of analysis cannot be optimally performed in isolation, e.g.:
  - Lexical structure constrains phonetic structure
  - Semantic structure constrains lexical structure