Zero resource and modeling early language acquisition

Final Report
Overview

• Goals
  – simultaneous engineering (machines) and reverse engineering (infants) of the unsupervised discovery of linguistic structures from raw speech
  – link unsupervize clustering, spoken term discovery and Bayesian models of word segmentation
  – gather resources, discuss big picture, conduct preliminary experiments
Overview

• Report
  – Engineering & Cognitive evaluation of subword units (Task 1: Emmanuel)
  – Clustering of allophones (Task 5: Naomi)
  – Same-Different evaluation of features & acoustic models (Task 4: Aren)
  – Bayesian Word Segmentation on supervised and unsupervised units (Task 3-Sharon)
  – Samuel & Rick: zero resource and downstream application (Samuel & Rick)

• Wrap Up
1. Analysing the output of the SD task

- The word-level DTW similarity matrix
  - PROS: feature independent procedure; a single measure: average precision.
  - CONS: no idea of what the features get right or wrong (in terms of linguistic units)
- corpus: Timit
- 58M word pairs: normalized DTW-distance
- extract 102k near-minimal pairs
  - same length (>=4 phoneme)
  - edit distance <= 50%
modeling the DTW-distance as a linear combination of phoneme matches and mismatches

- eg: power – towel
  - mismatches: p-t, axr-el
  - matches: aw-aw

- Linear model: \( D \sim p-p + p-t + p-k + \ldots \)
  - 102000 word pairs, 1035 regressors (phoneme pairs)
  - fits \( \sim 50-60\% \) of the variance
• reliability:
  – select the regressors with at least 100 data points
  – split the word pairs in two random subsets
  – compute the correlation between the B for the two subsets

\[ R^2 = 0.9 \]
phoneme substitution distance matrix
extracting detailed linguistic information

- other possibilities:
  - interaction with stress
  - insertion/deletions
  - suprasegmental features
2. direct human/machine comparison: a new dataset

- Japanese dataset
  - 3-4 syllable CV(N) nonword minimal pairs
    - substitution minimal pairs C-C, V-V (one feature, two features, three features)
    - insertion-deletion minimal pair
    - length minimal pairs V vs V: C vs C:
    - pitch accent minimal pairs
  - 400 minimal pairs x 10 speakers
    - talker variation: 3 males, 3 females, 1 child
    - speaking style: 1 whispered, 1 emotional, 1 ‘sloppy’/fast
    - channel distortions (filtering, background noise, time reversal)

- Same in English/French

- Human data
  - SD (d’) scores in control Japanese Ss (4 hours)

- TASK:
  - cog scientists: model the confusion matrix (under all input variations)
  - engineering: match the performance of humans (under channel modifications)

cf: Ghitza & Sondhi, 1997; the CHIME challenge
http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html
3. Challenging tasks

• Psychoacoustical data
  – local and global real-time adaptation of identification as a function of
    • speaking rate, phonetic context, vocal tract, dialect
    -> Can machine models match this?

• Modeling language-specific listening
  – Q: what are the difficult contrasts when you are learning an L2?
task 5: allophone clustering

- CSJ core database (40 hours: lectures, spontaneous, broadcast)
- “allophonic” forced alignment
  - tied-triphone model (with linguistic questions)
  - each phone is modeled by a tri-state 17 gaussian mixture
    - 50, 100, 200, 500, 1000 contextual allophones
- distance matrix
  - distance matrix: DTW on the matrix of the three states (using KL distance between each state)
results (N=500)

- Baseline
  - hierarchical clustering (with 39 clusters)
  - purity: .65
- Spectral clustering
  - purity: .72

see also, Boruta, PHD diss (forthcoming)
wrap-up

• the engineering challenge:
  – build a complete system V1.0
  – tune it until it works better
  → do not hide the failures nor the tricks to make it work: expose it in public: it tells us something deep about the signal

• the scientific challenge:
  – critical stance
    • what is a point of a psychological model that only models an ‘easy’ problem (mini language, fake data)?
    • generative HMMs generate catastrophically bad speech; how could is possibly work?
  – cross field inspirations
    • how to combine multiple informational streams in the way humans do (Hynek)?
    • have we looked hard enough for invariant cues (Hynek)?
    • do infants perform raw signal spoken term discovery? if yes, when, how?
    • how to incorporate compensation for context effects in a non supervised model?

• concrete objectives:
  – publish a common paper(s) on V1.0
  – make some of the tools more user friendly (scripts for pipelines, documentation, internal wiki?)
  – setting up and validating psycho-engineering test battery/running some experiments for V2.0 (convolutional sparse coding; allophone clustering)
Unsupervised Clustering Methods for Learning Sound Categories
Phonological Alternations

[kʰ] at the beginning of a stressed syllable

[k] in an ‘sk’ cluster
Spectral Clustering

• Graph-based clustering method

• No parametric assumptions on the distributions associated with sound categories

• Assumes distances between near neighbors are reliable, but longer distances are not

(Ng, Jordan, & Weiss, 2002; Luxberg, 2007)
How Connected are Allophones?

k → $k^h$ at the beginning of a stressed syllable

æ → before a nasal consonant
How Connected are Allophones?

\[ k \rightarrow k^h \text{ at the beginning of a stressed syllable} \]

\[ æ \rightarrow \text{ before a nasal consonant} \]
Questions

• Zero-resource speech technologies
  Can this method recover phonemes from a large-scale similarity matrix?

• Models of early language acquisition
  Are allophones structured such that this method will be effective?
TIMIT Database (English)

• Take one example of each phone-context pair from TIMIT (21275 phones)

• Extract a feature vector time series of cepstral coefficients, deltas, and double deltas for the phone

• Compute pairwise similarities using dynamic time warping (DTW)
## Similarity Matrix

```
Similarity Matrix

<table>
<thead>
<tr>
<th></th>
<th>aa_#h_bcl</th>
<th>aa_#h_l</th>
<th>aa_#h_n</th>
<th>aa_#h_r</th>
<th>aa_ao_v</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa_#h_bcl</td>
<td>.4048</td>
<td>.3856</td>
<td>.4821</td>
<td>.3406</td>
<td>...</td>
</tr>
<tr>
<td>aa_#h_l</td>
<td>.4048</td>
<td>.3707</td>
<td>.4935</td>
<td>.4617</td>
<td>...</td>
</tr>
<tr>
<td>aa_#h_n</td>
<td>.3856</td>
<td>.3707</td>
<td>.3144</td>
<td>.2959</td>
<td>...</td>
</tr>
<tr>
<td>aa_#h_r</td>
<td>.4821</td>
<td>.4935</td>
<td>.3144</td>
<td>.4223</td>
<td>...</td>
</tr>
<tr>
<td>aa_ao_v</td>
<td>.3406</td>
<td>.4617</td>
<td>.2959</td>
<td>.4223</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

21275 x 21275
```
CSJ Database (Japanese)

• CSJ core database (40 hours: lectures, spontaneous, broadcast)

• “allophonic” forced alignment
  – tied-triphone model (with linguistic questions)
  – each phone is modeled by a tri-state 17 gaussian mixture
  – 50, 100, 200, 500, 1000 contextual allophones

• distance matrix
  – distance matrix: DTW on the matrix of the three states (using KL distance between each state)
Results (N=500)

- **Baseline**
  - hierarchical clustering (with 39 clusters)
  - purity: .65

- **Spectral clustering**
  - purity: .72

see also, Boruta, PHD diss (forthcoming)
Evaluating Speaker Independence of Features and Unsupervised Acoustic Models

Aren Jansen, Pascal Clark, Samuel Thomas, Keith Levin, David Harwath, Ian MacGraw, Thomas Schatz, Vijay Peddinti, Ken Church, Hynek Hermansky
Same/Different Task

1. Compute representation for the corpus

2. Extract representation for ~10k word examples

3. Compute all pairwise DTW distances (~60 Million) and rank

4. Compute average precision for distinguishing same vs. different type
Proxy for Phone Error Rate When There Are No Phones

In supervised case it tracks phone recognition accuracy perfectly!
Corpora

• **Switchboard:**
  – English
  – Multi-speaker
  – Conversational telephone speech
  – 500 hr

• **TIMIT:**
  – English
  – Multi-speaker
  – Prompted
  – 4 hr
Feature Types

• **Acoustic features**, e.g. PLP, MFCC
  – Frame level distance: cosine
  – **Participants**: Vijay, Pascal, Thomas

• **Posteriorgrams**: discrete distribution over unsupervised subword unit set
  – Frame level distance: symmetrized KL divergence
  – **Participants**: Aren, Samuel, Dave
Universal GMM: Bottom Up Baseline

1. Use expectation maximization to train a Universal Gaussian mixture model over a large collection of speech

2. Assume each Gaussian defines a subword unit

3. Compute posteriorgram over Gaussian components
Adding Weak Top Down Constraints

• Train detailed 1024-component Universal GMM over speaker dependent subword units

• Given only word example pairs (e.g. from context info, term discovery), but no types:
  1. Use DTW to align acoustic features of each pair
  2. Accumulate co-occurrence statistics for GMM components
  3. Cluster components into speaker-independent subword units
Top-Down Constraints: Switchboard (Aren and Sam)

<table>
<thead>
<tr>
<th>Features</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>17.7 %</td>
</tr>
<tr>
<td>100-GMM + PLP</td>
<td>18.1 %</td>
</tr>
<tr>
<td>+ top down constraints: 1024-GMM to 100 units</td>
<td>27.9 %</td>
</tr>
<tr>
<td>English AM</td>
<td>51.6 %</td>
</tr>
</tbody>
</table>

Weak word level constraints buy you 30% of the supervised gap
Modeling Manifold Features: TIMIT
(Aren and Dave)

• Evaluate manifold features in unsupervised acoustic model:

<table>
<thead>
<tr>
<th>Features</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>33.4 %</td>
</tr>
<tr>
<td>Intrinsic Spectral Analysis (ISA)</td>
<td>49.6 %</td>
</tr>
<tr>
<td>50-GMM + MFCCs</td>
<td>23.6 %</td>
</tr>
<tr>
<td>50-GMM + ISA</td>
<td>32.0 %</td>
</tr>
</tbody>
</table>

ISA improvement over MFCC translates to unsupervised acoustic model
Tackling Speaker Variance in zero resource setting

• Speaker variance is critical issue in unsupervised clustering algorithms.

• Phonetic and lexical discovery approaches are predisposed to learn speaker specific allophones.
Hermansky and Broad, Proc. ICASSP
How to separate the front cavity response from the back cavity response?

Effective Formant Hypothesis

Hermansky and Broad, Proc. ICASSP
Distance Histograms

Blue — Distances between different words
Red — Distances between same words
Other factors explored

• Temporal Smoothing – Uniform Gain of 2% for various degrees of smoothing

• Gain Normalization – Removes the channel effects 2% increase
Waveforms to Tokenizations

Dave Harwath

Ian McGraw
From Waveforms to Tokenizations

... t (cat) w ah z (was) ...  

... t w ah z ...
From Waveforms to Tokenizations

Speech

... 029 (cat) 014 007 043 (was) ...

Reinserted boundaries for evaluation

... 029 014 007 043 ...

Token String for training
From Waveforms to Tokenizations

- MFCC_39 + CMN/CVN + PCA Rotation
- 50, 25 component GMMs (Diag. cov. matrix)
- 0.5 self-transition probability assumed for each component during decoding, remaining transition probability mass spread uniformly across other components

… 029 (cat) 014 007 043 (was) …

Reinserted boundaries for evaluation

… 029 014 007 043 …

Token String for training
Word segmentation with noisy phonetic input
Word segmentation

• Given a speech corpus, find meaningful units.
  – Word-spotting (e.g., DTW):

    ![Waveform for word-spotting example]

  – Word segmentation:

    ![Waveform for word-segmentation example]

      ─ Look at the doggie
      ─ Where’s the doggie
Word segmentation

• Given a speech corpus, find meaningful units.
  – Word-spotting (e.g., DTW):
    
    ![Waveform example 1](image1)
    ![Waveform example 2](image2)
  – Word segmentation:
    
    Look at the doggie
    Where’s the doggie

\[\text{Look at the doggie} \quad \text{Where’s the doggie}\]
Goal

• Current segmentation systems assume phonemic input.
• Eventually, want systems that can work with noisy tokenizations.
• Here: establish baseline performance of current systems on tokenizations with different degrees of noise.
  – How bad is the problem?
  – Use 1-best output; worst-case scenario.
Models

• Word-level dependencies:
  – Unigram (DP) model
  – Bigram (HDP) model
  – Collocation (Adaptor Grammar) model

• Sublexical dependencies:
  – All results use either phone bigrams or morphs.
Data (Switchboard)  (Aren, David, Ian, Sanjeev, Rick)

- Phonemic transcriptions
- Phonetic transcriptions (30-35% PER)
- Phone recognizer output (50% PER)
- Four unsupervised phonetic tokenizations:
  - MIT2: 25 categories
  - MIT1: 50 categories
  - Jansen2: 100 categories with smoothing
  - Jansen1: 100 categories
Switchboard vs. IDS

For about 35k words each:

<table>
<thead>
<tr>
<th></th>
<th>Words/Utt</th>
<th>Phones/Wd</th>
<th>Lexicon</th>
<th>Disfluent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swbd</td>
<td>12</td>
<td>3.1</td>
<td>3227</td>
<td>Yes</td>
</tr>
<tr>
<td>IDS</td>
<td>3.3</td>
<td>2.9</td>
<td>1321</td>
<td>No</td>
</tr>
</tbody>
</table>
Switchboard vs. IDS

**Token F-score**

- Uni
- Bi
- Colloc.

**Lexicon F-score**

- Uni
- Bi

IDS
Swbd

IDS
Swbd
Evaluating automatic phones

- **Phonemic transcriptions:**

<table>
<thead>
<tr>
<th>Gold:</th>
<th>Found:</th>
</tr>
</thead>
<tbody>
<tr>
<td>dh</td>
<td>dh</td>
</tr>
<tr>
<td>ax</td>
<td>ax</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>ih</td>
<td>ih</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>ae</td>
<td>ae</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
</tr>
</tbody>
</table>

- **Automatic tokenizations:**

<table>
<thead>
<tr>
<th>Gold:</th>
<th>Found:</th>
</tr>
</thead>
<tbody>
<tr>
<td>dh</td>
<td>12</td>
</tr>
<tr>
<td>ax</td>
<td>27</td>
</tr>
<tr>
<td>b</td>
<td>11</td>
</tr>
<tr>
<td>ih</td>
<td>3</td>
</tr>
<tr>
<td>g</td>
<td>24</td>
</tr>
<tr>
<td>k</td>
<td>17</td>
</tr>
<tr>
<td>ae</td>
<td>6</td>
</tr>
<tr>
<td>t</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>13</td>
</tr>
</tbody>
</table>
Phonemic vs. Phonetic

Token F-score

- Phonemic
- Phonetic
- Recog.

 Uni
 Bi
 Colloc
Why so bad (esp. bigram)?

- Number of distinct words (correct boundaries):

![Bar chart showing the number of distinct words for different categories: Phonemic, Phonetic, Recog., MIT2, MIT1, Jansen2, and Jansen1. The y-axis represents the number of distinct words ranging from 0 to 40000, and the x-axis lists the categories. The chart indicates a significant difference in the number of distinct words across the categories.]
Accomplishments

• Data sets for testing with differing noise.
• Baselines showing:
  – Lots of room for improving word seg to account for/correct for noise.
  – Lots of room for improving phonetic tokenizations using improved clustering or top-down info.
Issues/Future work

• Evaluation:
  – Is boundary measure appropriate? How to score lexicon?

• Modeling:
  – How much better if using lattices instead of 1-best?
  – Other ways to incorporate top-down information?
Reducing Demand for Transcribed Speech: *Bootstrapping* from other languages
& Self training *selectively*

Samuel Thomas

Joint work with Mike Seltzer, Ken Church and Hynek Hermansky

The Johns Hopkins University
Microsoft Research
IBM Research
Reducing Demand for Transcribed Speech

- Language Models
- Acoustic Models
- Lexicon
- Feature Extraction

Transcribed Speech
“1++” Hour Transcribed Speech

- 1++ >> 0
- Plausible scenario for IARPA Babel
  - Limited transcribed speech in target language
  - Plus more in non-target languages
  - Plus more untranscribed
- Data
  - CallHome English (15 hours transcribed)
    - 1 hour transcribed
    - 14 hours “untranscribed”
    - 31 hours of CallHome German/Spanish
  - Plus other resources: dictionary, language model, etc…
Baseline: There is no data like more data

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Word Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 hour transcribed English</td>
<td>71.2</td>
</tr>
<tr>
<td>1 hour transcribed English + 31 transcribed German/Spanish + 14 untranscribed</td>
<td>TBD</td>
</tr>
<tr>
<td>15 hours transcribed English</td>
<td>53.5</td>
</tr>
<tr>
<td>System</td>
<td>Word Error Rate (%)</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>1 hour transcribed English</td>
<td>71.2</td>
</tr>
<tr>
<td>1 hour transcribed English + 31 hours German/Spanish (3 layer MLP)</td>
<td>62.8</td>
</tr>
<tr>
<td>1 hour transcribed English + 31 hours German/Spanish (Deep Neural Net)</td>
<td>59.0</td>
</tr>
<tr>
<td>Self Training <em>Selectively</em> to improve posteriorigrams</td>
<td>57.6</td>
</tr>
<tr>
<td>Self Training <em>Selectively</em> to improve acoustic models</td>
<td>55.3</td>
</tr>
<tr>
<td>15 hours transcribed English</td>
<td>53.5</td>
</tr>
</tbody>
</table>
Using Zero Resource Tools to Enhance High Resource Spoken Term Detection

Richard Rose, Atta Norouzian, Benjamin
Thanks to Sharon Goldwater

July 27, 2012

McGill University Dept. of ECE
Outline

- Zero-resource graph re-ranking of terms from STD system
  - Feature-based acoustic dot-plots
- Unsupervised Bayesian word segmentation on lecture speech task
  - Using reference vs. decoded transcriptions
Graph Re-ranking for a Lecture Speech Task

• Graph re-ranking with posterior-gram based dot-plots was shown to improve term detection performance w.r.t. baseline STD system
• Investigate the effect of replacing posterior-based dot-plots with feature-based (zero-resource) dot-plots:
  • Generate features (perceptual linear prediction) – All from a single speaker
  • Regenerate dot-plots for all segment pairs
  • Re-run graph re-ranking
• Problem:
  • Presence of silence intervals compromises performance
  • Silence has not yet been removed from our data
Low Resource Context Constraints: Graph Based Re-Ranking

- **High Resource**: Treat scores from spoken term detection as node potentials [Chen et al, 2011]

  \[ x_1, x_2, x_3, x_4, x_5, p_{1,2} \]

- **Zero Resource**: Use feature based dot-plots to discover self-similar intervals [Jansen et al, 2010] [Parks and Glass, 2005]

- **Graph-Based Re-Ranking**: Allows interval similarity to constrain relationship among scores [Hsu et al 2007]
  - Increase / decrease confidence of hypothesized terms based on similarity to other hypotheses
  - Discover new hypotheses?
Graph Based Re-Ranking

Query: $Q$

Segment Retrieval / Term Verification

High Resource
Spoken Term Detection:
Detected Intervals

Zero Resource
Interval Similarities:
Zero Resource
Acoustic Dot Plots

Constraints
Graph Re-ranking

$$S(x_1 | Q)$$ $x_1$

$$S(x_2 | Q)$$ $x_2$

$$S(x_3 | Q)$$ $x_3$

$$S(x_4 | Q)$$ $x_4$

$$S(x_5 | Q)$$ $x_5$

Arc weights:
$$p_{i,j} = \frac{s_{i,j}}{\sum_k s_{i,k}}$$

Normalized Scores:
$$u_i = \frac{S(x_i | Q)}{\sum_{x_i} S(x_i | Q)}$$
Graph Re-ranking for a Lecture Speech Task

- ROC for STD scores before and after re-ranking:

- Re-ranking with zero resource dot-plots provides an improvement in detection performance similar to posterior based dot-plot
- Better preprocessing of features (silence removal) should improve performance
Unsupervised Bayesian Word Segmentation on a Lecture Speech Task

- Subset of chemistry lectures – spontaneous speech and imperfect transcriptions:
  “… Ampicillin is a broad spectrum eh ah antibiotic and will …”
  “… Using porn they managed to grow the mold through the …”
- Batch sampler using Sharon’s unigram model with and without vowel constraints
  • Segments forced to include vowels

- Task: Discover boundaries from multiple token sequences
  • Baseform phonemic expansion of reference word transcriptions
  • Baseform phonemic expansion of decoded word transcriptions (word accuracy: 57%)
  • Unconstrained decoded phoneme sequence (phone accuracy: 41%)
Unsupervised Bayesian Word Segmentation on a Lecture Speech Task Task

- Run on 18,000 words, 1200 utterances, ~2 hours of speech

<table>
<thead>
<tr>
<th>Transcriptions</th>
<th>Vowel Constraint</th>
<th>Token F-Score</th>
<th>Boundary Precision</th>
<th>Boundary Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>No</td>
<td>54</td>
<td>75</td>
<td>84</td>
</tr>
<tr>
<td>Reference</td>
<td>Yes</td>
<td>60</td>
<td>87</td>
<td>78</td>
</tr>
<tr>
<td>Const. Decoded*</td>
<td>No</td>
<td>59</td>
<td>80</td>
<td>85</td>
</tr>
<tr>
<td>Unconst. Decoded</td>
<td>Yes</td>
<td>29</td>
<td>42</td>
<td>58</td>
</tr>
</tbody>
</table>

*Gold standard was decoded word string

- Results are consistent with those obtained for Switchboard transcriptions
Achievements

• Common resources
  – databases:
    • English: TIMIT, SWITCHBOARD
    • English+French (IDS): Providence Corpus
    • Japanese: CSJ
  – programs
    • SD DTW Evaluation
    • Bayesian word segmentation
    • spoken term DTW

• Tutorials

• Preliminary experiments
  – baseline performance of WS on more variable inputs
  – baseline allophone clustering
  – baseline spoken word discovery on child directed speech
  – preliminary experiments on zero resource applications