Insights into Pronunciation Modeling and ASR
Using Mixed Unit Pronunciation Models

Arthur Kantor
Department of Computer Science
University of Illinois - Urbana Champaign

September 12, 2008
Outline

1. Insights into Pronunciation Models
   - Pronunciation Perplexity
   - Context

2. Mixed Unit Models
   - Motivation
   - Methods
   - Duration Conditional Acoustic Models

3. Timeshrinking
   - Motivation
   - Method
### Pronunciation Perplexity

#### Recognizing from Synthesized Data [McAllaster et al. 1998]

<table>
<thead>
<tr>
<th>Accoustic Observations</th>
<th>Dictionary</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthesized from word trans.</td>
<td>base</td>
<td>10.8</td>
</tr>
<tr>
<td>synthesized from triphone trans.</td>
<td>base</td>
<td>43.9</td>
</tr>
<tr>
<td></td>
<td>base+test</td>
<td>33.5</td>
</tr>
<tr>
<td>real acoustics</td>
<td>base</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>base+test</td>
<td>60.8</td>
</tr>
</tbody>
</table>

- Seemingly the pronunciation difference accounts for most of the errors.
Recognizing from Synthesized Data [McAllaster et al. 1998]

<table>
<thead>
<tr>
<th>Accoustic Observations</th>
<th>Dictionary</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthesized from word trans.</td>
<td>base</td>
<td>10.8</td>
</tr>
<tr>
<td>synthesized from triphone trans.</td>
<td>base</td>
<td>43.9</td>
</tr>
<tr>
<td></td>
<td>base+test</td>
<td>33.5</td>
</tr>
<tr>
<td>real acoustics</td>
<td>base</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>base+test</td>
<td>60.8</td>
</tr>
</tbody>
</table>

- Seemingly the pronunciation difference accounts for most of the errors.
- If there is acoustic model/data mismatch, growing the lexicon in the optimal way hurts.
Insights into Pronunciation Models

Mixed Unit Models

Timeshrinking

Pronunciation Perplexity

**DBN for a typical recognizer**

- Squares/Circles denote Discrete/Continuous R.V.’s
- Dashed node edges denote deterministic R.V.’s
- Filled in nodes are observed R.V.’s

Word {one, seven, ...}

Word Transition {true, false}

Pronunciation Index {1 ... 6}

Pronunciation {/s eh v eh n/, /s eh v n/}

Sub-phone State Transition {true, false}

Sub-phone index {1 ... 30}

Sub-phone State {s1, s2, s3, eh1, ...}

Observation {Gaussian Mixture}
**The pronunciation model**

- **Word** \{one, seven, ...\}
- **Word Transition** \{true, false\}
- **Pronunciation Index** \{1 ... 6\}
- **Pronunciation** \{/s eh v eh n/, /s eh v n/\}
- **Sub-phone State Transition** \{true, false\}
- **Sub-phone index** \{1 ... 30\}
- **Sub-phone State** \{s1, s2, s3, eh1, ...\}
- **Observation** \{Gaussian Mixture\}
WER is used as proxy of acoustic model/data mismatch
The trend is true across a variety of languages, datasets and recognizers ($R^2 = .80$).
Recognition from phonetic transcriptions benefits greatly from growing the lexicon.

On human-made phonetic transcriptions of Switchboard:
- Allowing asynchrony between articulatory features [Livescu 2004]
  - Error rate 64.8% → 29.7% on isolated word recognition task.
- Surfical Pronunciation Model [Ristad 1998]
  - WER 18.61% → 12.63% on recognition task.
Let $N$ be a network of HMMs representing a set of pronunciations for a word $W$.

Under high acoustic model/data mismatch conditions, there exists a single path (single pronunciation) through $N$ that will yield a lower WER when used as the pronunciation for $W$ instead of $N$ in an ASR system.
The pronunciation variability for common words is large.
- There are 80 different pronunciation for the word ‘and’ (entropy=4.6 bits).

Pronunciation Perplexity must be kept low for complex recognition tasks.
- e.g. Conversational speech

Multiple pronunciation models are necessary, but must be almost completely disambiguated by context.
Outline

1. Insights into Pronunciation Models
   - Pronunciation Perplexity
   - Context

2. Mixed Unit Models
   - Motivation
   - Methods
   - Duration Conditional Acoustic Models

3. Timeshrinking
   - Motivation
   - Method
Motivation

Problems with Phone-Based Pronunciation Models

‘know’ vs. ‘no’
- Phonetic transcriptions are identical: /N OW/
- The expected duration of ‘know’ is 80 ms shorter than ‘no’ [Ganapathiraju 2001].
- Recognizer must rely solely on the language model to chose the right word.

‘right’ vs. ‘I’
- ‘right’ is frequently pronounced as ‘I’: /AY/
- One of the most common confusion pairs in JHU06 experiments.
- The language model has to fight the pronunciation model to correctly recognize the word.
Variability is greatest for common, short words.

but

There is a lot of training data for common, short words.

- Monosyllabic words cover 80% of conversational speech.
- 100 most common words cover 65% of speech (1000 cover 90%).

We can make more detailed models for common words.
Possible Solution: Mixed Unit Models

- Model words with phone, syllable and word units
  - The choice of a word’s model depends on the amount of training data for that word.
- Common syllables get their own pronunciation models.
- A word is pronounced as a single sequence of models.
  - No pronunciation perplexity

**examples**

- no = /NO\_w/
- know = /KNOW\_w/
- nobody = /N\_OW\_s B\_p AA\_p D\_IY\_s/
A word model is used only if
- There are enough training examples.
- The likelihood of the word tokens is higher using the word model than using the syllable model.

Use decision trees to cluster contexts.
DT questions are based on word, syllable, phone and duration features.
Otherwise model the word as a sequence of syllables.
Select syllable models analogously.
The identity of the word/syllable/phone to the left/right of target model.

The phonetic features of the phone (fricative, voiced, etc.)

Duration: Is the hypothesized duration of the example longer or shorter than the median duration of the hypothesized unit?
Duration Conditional Acoustic Models

Duration

- The decoder must hypothesize the duration of the units. The duration is discarded before the transcription is returned.

- Improved duration modeling yields slight but consistent improvements.
  - Easy to improve on a geometric distribution of duration.
  - One of the best approaches is to put minimum and maximum limits on unit duration.
  - The acoustic models are not changed.

- We can also condition the acoustic models on the hypothesized duration.

- Requires explicit tracking of duration.
Promise of the Approach

- A simpler approach simply allows has two sets of phone models: can switch between models at word boundaries
  - It works! [Zhang 2000] is screwing up my plot. →
  - Pronunciation perplexity is increased.
  - In this case, short acoustic models are allowed to model a long example word.
- We can enforce short acoustic models for short examples.
  - If duration distributions of short and long models don’t overlap, there is no perplexity.
Duration Conditional Graphical Model

- Word \{one, seven, ...\}
- Word Transition \{true, false\}
- Pronunciation Index \{1 \ldots 6\}
- Pronunciation \{ (/seh/ /vehn/), ...\}
- Syllable Transition \{true, false\}
- Syllable index \{1 \ldots 6\}
- Syllable \{/seh\_fast/, /seh\_slow/, /vehn/...\}
- Sub-syllable State Transition \{true, false\}
- Sub-syllable index \{1 \ldots 30\}
- Sub-syllable State \{seh1, seh2, ...\}
- Cumulative Duration \{1 \ldots N frames\}
- Observation \{Gaussian Mixture\}
- Duration \{1 \ldots N frames\}
Outline

1. Insights into Pronunciation Models
   - Pronunciation Perplexity
   - Context

2. Mixed Unit Models
   - Motivation
   - Methods
   - Duration Conditional Acoustic Models

3. Timeshrinking
   - Motivation
   - Method
Beam pruning of hypotheses space is always used in large vocabulary decoding.

The pruning beam contents are similar throughout recognition.
  - The differences are often in the durations of the phones, and not the sequence of the phones.

Forcing similar hypotheses out of the beam should improve recognition.
Method

Timeshrink the Observations

1. Classify frames into phonemes
2. Find substrings (segments) of frames where:
   - The most likely phone is the same and
   - The classifier has high confidence
3. Summarize these segments with a single representative frame.
4. Low confidence frames remain as their own summaries.
5. Perform recognition over summary frames, exponentiated by the duration of the segment.

- Hopefully this makes the contents of the beam more diverse.
- The downside is that we may have merged two different phones together into one summary frame.
Segmentation
## Preliminary Results

<table>
<thead>
<tr>
<th>experiment</th>
<th>summary frames</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = 1$ (baseline)</td>
<td>100.0%</td>
<td>58.1</td>
</tr>
<tr>
<td>$\tau = .95$</td>
<td>70.3%</td>
<td>55.6</td>
</tr>
<tr>
<td>$\tau = .9$</td>
<td>62.8%</td>
<td>54.9</td>
</tr>
<tr>
<td>$\tau = .8$</td>
<td>55.0%</td>
<td>55.0</td>
</tr>
<tr>
<td>$\tau = .8$ narrow</td>
<td>65.6%</td>
<td>55.1</td>
</tr>
<tr>
<td>$\tau = .7$</td>
<td>49.2%</td>
<td>55.8</td>
</tr>
<tr>
<td>$\tau = .5$</td>
<td>37.5%</td>
<td>62.5</td>
</tr>
</tbody>
</table>
Summary

- Must not allow pronunciation perplexity on difficult tasks, such as large vocab speech recognition.
- Mixed unit models seem promising.
- Timeshrinking seems promising.
- Can watch my progress (all code and data) on http://mickey.ifp.uiuc.edu/speechWiki/index.php/Fisher_Corpus
Method

Difficulties so far

- The Size of the fisher Corpus.
  - 20 cluster days to train 64-gaussians mixture models
  - OOV words need transcriptions (done)
- Syllable models are risky.
  - syllables risk: context-free syllable models perform comparably with triphones.
  - tri-syllables risk: Quinphones give only minor improvements over triphones on a well tuned system.
extra slides

Outline

4 extra slides

- Preliminaries
The Language Model

- The language model is assumed to be an $N$th order Markov chain, or $N$-gram.

$$p(W_1 ... K) = \prod_{k=1}^{N-1} p(W_k | W_1, ... W_{k-1}) \prod_{k=N}^{K} p(W_k | W_{k-1}, ... W_{k-N})$$

- The $N$-gram is backed off to $N-1$-gram to handle data sparsity
- Common multi-words are used (‘going to’ → ‘gonna’)

← Back
Representation of Speech

- The speech waveform is segmented into overlapping frames of fixed duration.
- Each frame is represented by a feature vector designed to capture only the information relevant to speech and discard all other acoustic information.
- Traditional features are PLPs and MFCCs, their deltas and delta-deltas.
- Sometimes the features are the phoneme labels generated by per-frame MLP classifiers trained on PLPs or MLPs of the neighboring frames.
The Observation Model

- In the simplest form, the acoustics of each sub-word state are modeled in isolation.

- Let $S_{lt}$ be the sub word state at time $t$.

  $$p(O_{1...T}|S_{1...L}) = \prod_{t=1}^{T} p(O_t|S_{lt})$$

- Can also condition the acoustics of the sub-word state on the neighbors of that sub-word state.
- This describes the triphone and tri-syllable models.
Why Syllables?

Perhaps triphone and word models are enough?

- Phoneme pronunciation variability depends on its position within the syllable [Greenberg 1998].
- Systems with all three models perform better than systems with just word and triphone models [Ganapathiraju 1997]
Breaking up a unit model into sub-unit models: Assumptions

- The context choices for each word are the same in both representations.
- Each word has the same number of sub word states in whole-word and syllable representations.
- Each state has a single Gaussian observation.
Breaking up a unit model into sub-unit models: Algorithm

- Choose K most frequent words, and hold out some of them for testing.
- Train word models for most frequent K words, remember their gammas at the last iteration.
- Train syllable models on all data besides most frequent K words and also remember their gammas.
- For each word k in each possible context:
  - For each state q in k:
    - Create a new obs model by adding the mean and variance of respective word and syllable q’s weighted by the gammas
    - Calculate the likelihood of the held out words using the new obs model and the word model
    - If the new model is better, break the word into syllables
SYt may only be *true* if the cumulative duration lies between some syllable dependent thresholds \( l_{SY} \leq CD < h_{SY} \).

- If \( CD < l_{SY} \) then \( SYt \) is set to *false*, preventing a transition out of the syllable.
- If \( CD \geq h_{SY} \) then \( SYt \) is set to a special 0-probability state, knocking the hypothesis out of the search beam.

If a syllable has two models \( SY = sy_1 \) and \( SY = sy_2 \), and their thresholds don’t overlap \( h_{sy_1} \leq l_{sy_2} \), the pronunciation perplexity is identical to that of the baseline.
Phone Classifiers
Given an observation frame and its neighbors a phone classifier predicts the most likely phone or a phonetic feature.
- Can model interdependence between nearby observations.
- Can segment the observation string into substrings belonging to the same phone.
- Classifiers have been implemented as MLPs.

Tandem Models
A tandem model uses the classifier outputs in tandem with the original observations.
- Significantly outperforms monophones and triphones.
A **segmental model** can generate a string of observations from a single hidden state.

- Can also model interdependence between nearby observations.
- A generalization of HMM.
- Timeshrinking is a very simple Segmental Model.
Main Contributions

- Literature review provides some guidelines for when pronunciation models with increased pronunciation perplexity will be helpful.
- A set of experiments modeling words with mixed unit models.
- A set of experiments with timeshrinking segmental models.
Very Rough Timeline

- 4 months for mixed unit models and baseline
- 1 month for timeshrinking
- 1 month for duration conditional models
- 3 months to write up the results in a thesis