Web-derived Pronunciations for Spoken Term Detection
Doğan Can  
Erica Cooper  
Arnab Ghoshal  
Martin Jansche  
Sanjeev Khudanpur  
Bhuvana Ramabhadran  
Michael Riley  
Murat Saraçlar  
Abhinav Sethy  
Morgan Ulinski  
Christopher White

Boğaziçi University  
MIT  
Johns Hopkins University  
Google Inc.  
IBM T. J. Watson Research  
Boğaziçi University  
IBM T. J. Watson Research  
Cornell University  
Johns Hopkins University
Spoken Term Detection (STD): open-vocabulary search over spoken document collections

Classic Large-Vocabulary Continuous Speech Recognition (LVCSR) assumes a closed vocabulary
Speech signal
Sampled waveform
Waveform windows
Cepstral features
Hidden Markov model states
Contextual phones
Phones
Words

} Pronunciation model
Overview

Speech signal
Sampled waveform
Waveform windows
Cepstral features
Hidden Markov model states
Contextual phones
Phones
Spoken Term Detection (STD):
open-vocabulary search over spoken document collections
Build phone index instead of word index
Search by (approximate) phonetic match
Need word pronunciations during search
Need word pronunciations during search
For an open-ended vocabulary
For proper names from a variety of origins
Continually evolving

Ahmadinejad, Blagojevich, Sotomayor, ...
Models over pairs of strings:

Letter-to-phone (L2P, pronunciation) models
Phone-to-phone (P2P) model
Letter-to-letter (L2L, transliteration) models
Latent alignment models, like in SMT

\[ \Pr[\lambda, \pi] = \sum_a [\lambda, \pi | a] \]

Alignments \( a \) assumed to be monotonic

Train on parallel data \((\lambda_1, \pi_1), \ldots, (\lambda_n, \pi_n)\):

Impute latent alignments with a 1-gram model, EM trained from flat start

Train \( n \)-gram language model on imputed alignments \((n = 2, 3, 4, 5)\)
Call these “pair $n$-gram models”

All models are joint models $Pr[\lambda, \pi]$

For 1-gram models, can derive conditional models $Pr[\lambda | \pi]$ or $Pr[\pi | \lambda]$ from joint ones in closed form

Expressed as finite-state transducers (FSTs) using the OpenFst library (openfst.org)

Operations on models are well-known FST manipulations
The Web is a rich source of pronunciations:

IPA transcription
The Ctenophora (pronounced /tɪˈnɛəfərə/, singular ctenophore, pronounced /ˈtɛnəfərə/ or /ˈtiːnəfərə/), commonly known as comb jellies, are a phylum of animals that live in marine waters worldwide. en.wikipedia.org

Ad-hoc transcription
Two species of ctenophores (pronounced TEN-uh-fores), can be found just off shore in the Chesapeake Bay: Mnemiopsis and Beroe. nationalzoo.si.edu

The Moonjelly is a small sea creature about the size of a child’s hand. It looks like a blob of clear, colorless jelly. Its scientific name is "Ctenophore" (pronounced tee-ne-for.) markshasha.com
The Web is a rich source of pronunciations
Finding them involves:
  Extracting a superset of candidates
  Validating the extracted candidates
  Normalizing the pronunciations
Find candidate pronunciations by pattern matching over billions of Web pages:

...(pronounced …)
...pronounced “…”
..., pronounced ..., 
...
...
...

Extraction
IPA predates computers, the Web, and modern notions of phonetics/phonology
IPA is difficult to use even by experts
IPA symbols are scattered across several Unicode code blocks
Cannot tell just by looking at a character whether it is part of an IPA transcription
IPA characters are often misappropriated

You can write upside down like this...
For each pronunciation candidate, find the most likely matching orthographic string.

The Ctenophora (pronounced /tɪˈnæfərə/, singular ctenophore, pronounced /ˈtɛnəfɔər/) uses a very simple pronunciation model to score orthographic strings.
Extraction had to be simple and fast to allow it to run at Web scale.

Extraction validation examines a few million (orthography, pronunciation) candidates and removes candidates with invalid or undesirable pronunciations.

removes candidates with wrong or undesirable orthographies
Rain Water, the product, comes from Dripping Springs, where it is collected and bottled by Richard Heinichen, a 57-year-old former blacksmith. ... Mr. Heinichen (pronounced like the beer) said he sold about 170,000 16-ounce bottles last year... nytimes.com

So, that said, I thought I'd talk a little about the towns of Dharamsala (pronounced Dar-am-Shala) and Pushkar (pronounced like the thing you would do when your automobile breaks down). strangebenevolent.blogspot.com
Annotate a few hundred candidates

Extract a few dozen features, in particular alignment-based features that count e.g. vowel mismatches or consonant matches

Train and apply Support Vector Machine (SVM) classifiers
Normalization is necessary to homogenize the extracted raw pronunciations.

For IPA pronunciations, transcription conventions and/or skills vary.

For ad-hoc pronunciations, need to generate phones.
For extracted IPA pronunciations, consider the subset of words found in Pronlex (PL)

Check what happens when we train L2P models on one source (PL, IPA) and evaluate it on another

Compute phone error rate (PhER) by 5-fold parallel cross-validation

Do this for the top 7 websites in our data
Focus on the IPA-PL evaluation

Train phone-to-phone (P2P) normalization models on parallel (IPA, Pronlex) data

Vary the $n$-gram order of the P2P models

Use P2P models to normalize IPA data, train L2P models on normalized IPA

Compare with L2P model trained directly on Pronlex
Phonetic transcription conventions vary by data source.

Website-specific IPA normalization makes extracted pronunciations look more like those found in Pronlex.

L2P models trained on normalized Web-IPA pronunciations are as good as models trained on comparable amounts of Pronlex.
For extracted ad-hoc pronunciations, we need to derive phones from the two available orthographies.

From last Wednesday’s *New York Times*: Phthalates (pronounced THAL-ates) are among the most common endocrine disruptors, and among the most difficult to avoid.

Ambiguities remain in the simplified orthography (which *th* sound?)
Experiment with 4 ways of generating phones for ad-hoc pronunciations

L2P model trained on orthography
L2P model trained on ad-hoc prons
Factored generative model with conditional independence
Full model over aligned triples
Ad-hoc transcriptions are easier to produce than IPA transcriptions. We found 80% more ad-hoc transcriptions than IPA on the Web. L2P models trained on ad-hoc data are better than L2P models trained on comparable amounts of data in standard orthography.
Indexation of weighted finite automata

Used in Spoken Utterance Retrieval and Spoken Term Detection

Related to suffix and factor automata

Implemented with OpenFst

Also see *Spoken Information Retrieval for Turkish Broadcast News* by Parlak and Saraçlar in tonight’s poster session
Goal of Spoken Term Detection is to find the time interval containing the query, for each occurrence of the query.

Retrieval is based on the posterior probability of substrings (factors) in a given time interval.

Need to index the (preprocessed) output lattices of an automatic speech recognition (ASR) system.
Preprocessing of ASR output lattices:

Cluster non-overlapping occurrences of each word (or sub-word)

Assign other occurrences to the cluster with which they maximally overlap

Time interval of each cluster is the union of all its members

Adaptively quantize the time intervals
Index construction:

Union of preprocessed FSTs

Optimized for efficiency

Factor-automaton introduces a new start state and a new final state, plus transitions to and from every other state

Normalized to form a proper posterior probability distribution
Searching for a user query is as simple as:

Representing the query as an FSA, which may represent multiple pronunciations

Composing the query FSA with the index FST

Projecting onto the output labels (time intervals) and ranking by best path

Produces results ordered by decreasing posterior probability
Analyze the impact of web-derived pronunciations on the retrieval of out-of-vocabulary (OOV) queries in an STD task.

Held out 1290 names of persons and places and rare or foreign words with 5+ occurrences in the Broadcast News corpus.

Removed those words from the vocabulary of the speech recognizer.

Removed all utterances containing the held-out data from the BN training data.
Trained a recognizer using the IBM Speech Recognition Toolkit on 300 hours of BN

Word error rate on standard BN test set was 19.4%

100 hours containing OOV terms held out for further experiments, transcribed by the recognizer and indexed by the STD system

Experiment with different pronunciations during retrieval, report ATWV metric from NIST 2006 STD Evaluation
Results with reference pronunciations in terms of ATWV (higher is better)
Experiments with Web-derived pronunciations added to a baseline L2P system

- L2P
- L2P + Webpron

Experiments:
- Ad-hoc raw
- Ad-hoc manual
- IPA
Examples of Webprons with positive impact

<table>
<thead>
<tr>
<th></th>
<th>L2P</th>
<th>Webpron</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBRIGHT</td>
<td>ae l b r a y t</td>
<td>ao l b r a y t</td>
</tr>
<tr>
<td>GREENSPAN</td>
<td>gr iy n s p a a n</td>
<td>gr iy n s p a e n</td>
</tr>
<tr>
<td>SHIMON</td>
<td>sh ih m ax n</td>
<td>sh ih m ow n</td>
</tr>
</tbody>
</table>
Fraction of correctly detected occurrences

<table>
<thead>
<tr>
<th></th>
<th>L2P</th>
<th>Webpron</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBRIGHT</td>
<td>0.238</td>
<td>0.475</td>
</tr>
<tr>
<td>GREENSPAN</td>
<td>0.713</td>
<td>0.950</td>
</tr>
<tr>
<td>SHIMON</td>
<td>0.950</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Experiments
Examples of Webprons with negative impact

<table>
<thead>
<tr>
<th>Name</th>
<th>L2P</th>
<th>Webpron</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREUND</td>
<td>fr oy nd</td>
<td>fr eh nd</td>
</tr>
<tr>
<td>SANTO</td>
<td>s ae n t ow</td>
<td>s ax/ey/eh n t</td>
</tr>
<tr>
<td>THIERRY</td>
<td>th iy ax r iy</td>
<td>t eh r iy</td>
</tr>
</tbody>
</table>
Number of false alarms

FREUND: 375
SANTO: 750
THIERRY: 1,125

Experiments
People sometimes use nearest-neighbor pronunciations, where the pronunciation of a familiar word is used for a similar unfamiliar word.

For cases like Thierry / Terry, which occurs as a suffix in *military*, or *voluntary*, false alarms increase dramatically.

Overall, Web-derived pronunciations have a net positive impact.
Large quantities of human-supplied pronunciations are available on the Web

Our methods yield more than 7M occurrences of raw English pronunciations

After validation and normalization, extracted pronunciations have a positive impact on a Spoken Term Detection task

Our approach can be used to bootstrap pronunciation dictionaries for other tasks and languages