Dialectal Chinese Speech Recognition

Richard Sproat, University of Illinois at Urbana-Champaign
Thomas Fang Zheng, Tsinghua University
(Bill Byrne, Johns Hopkins University)
Liang Gu, IBM
Dan Jurafsky, Stanford University
Jing Li, Tsinghua University
Yi Su, Johns Hopkins University
Yanli Zheng, University of Illinois at Urbana-Champaign
Haolang Zhou, Johns Hopkins University
Philip Bramsen, MIT
David Kirsch, Lehigh University

Opening Day Presentation, July 6, 2004

Dialectal Chinese Speech Recognition

Workshop 2004
An NSF Sponsored Event
The Center for Language and Speech Processing
Dialects ( ) vs. Accented Putonghua

- Linguistically, the “dialects” are really different languages.
- Common (mis)conception: Chinese write the same but speak differently. (Well, actually this is true, but it’s because people usually write in Standard Chinese.)
- This project treats Putonghua (PTH - Standard Mandarin) spoken by Shanghainese whose native language is Wu: Wu-Dialectal Chinese.
Wu vs. PTH vs. Wu-Accented PTH

Wu vs. PTH

“There are over 1200 students.”

PTH vs. Wu-Accented PTH

“Hua Temple --- Longhua Temple, how did it come about, right? I, that is, I saw a story that is often told about this.”
Project Goals

• Develop a general framework for dialectal Chinese ASR which models:
  – Phonetic variability
  – Lexical variability
  – Pronunciation variability

• Find methods to modify baseline PTH recognizer to obtain a recognizer for the dialect of interest:
  – dialect-related knowledge (syllable mapping, cross-dialect synonyms, …)
  – adaptation data (in small quantities, or even lacking)
Background on Data Collection

• Wu-Dialectal Chinese Speech Database
  – 11 hours/100 speakers, with phonetic transcriptions
  – Coded for gender, age, education, Putonghua (PTH) level, fluency
  – Read speech (5.5 hours):
    • Type I: each sentence contains PTH words only (5-6k)
    • Type II: each sentence contains one or two most commonly used Wu dialectal words while others are PTH words
  – Spontaneous Speech (5.5 hours)
    • Conversations with PTH speaker on self-selected topic from: sports, policy/economy, entertainment, lifestyles, technology
  – 20 Beijing speakers (character and pinyin transcriptions only)
• 50k-word Electronic Dictionary with each word having:
  – PTH pronunciation in PTH initial-final (IF) string
  – Wu dialect pronunciation in Wu IF string
Data Set Division

Data were split according to age (younger, older), education (higher, lower), and PTH level

<table>
<thead>
<tr>
<th>Set</th>
<th># of files</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>DevTrain</td>
<td>3,689</td>
<td>6.42h / 23,101s</td>
</tr>
<tr>
<td>DevTest</td>
<td>465</td>
<td>0.79h / 2,835s</td>
</tr>
<tr>
<td>Test</td>
<td>451</td>
<td>0.82h / 2,950s</td>
</tr>
</tbody>
</table>
Baseline System

• Standard Chinese AM for spontaneous speech (JHU)
  – 39 dimensional MFCC_E_D_A_Z
  – diagonal covariance matrix
  – 4 states per unit
  – 103,041 units (triIF), 10,641 real units (triIF)
  – 3,063 different states (after state tying)
  – 16 mixtures per state, 28 mixtures per state for silence unit
• Single lexical entry for each Chinese syllable
• Connected syllable network: no LM
## Baseline System

<table>
<thead>
<tr>
<th>Test set</th>
<th># of files</th>
<th>Syl Corr %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hub-4NE</td>
<td>1,263</td>
<td>69.27</td>
</tr>
<tr>
<td>Wu-dialectal</td>
<td>451</td>
<td>28.54</td>
</tr>
</tbody>
</table>
Pronunciation Variation
(Rebecca Starr and Dan Jurafsky)

• Focus on sh/zh/ch => s/z/c and s/z/c => sh/zh/ch

• Sibilants in Wu-PTH Corpus:
  – 19,662 tokens of s/z/c/sh/zh/ch
  – Each token coded for predictive factors:
    • Age
    • Gender
    • Education
    • Phone (sh, zh, ch)
    • Phonetic context

• Logistic Regression
Results

• Massive variation between speakers:
  – 15%-100% use of standard pronunciation
  – Age/education best predictors of standard sh.zh.ch
  – Younger speakers more standard
Younger Speakers More Standard

Dialectal Chinese Speech Recognition
Results

• Massive variation between speakers:
  – 15%-100% use of standard pronunciation
  – Age/education best predictors of standard sh/zh/ch
  – Younger speakers more standard
  – Conclusions:
    • Need speaker-specific pronunciation adaptation.
    • Or cluster by accent “severity”.
Three Kinds of Adaptation

• Acoustic model (AM) adaptation
• Lexicon adaptation (pronunciation modeling)
• Language model (LM) adaptation
Acoustic Model Adaptation

• Purpose
  – Highly accurate and rapidly applicable recognition of accented/dialectal PTH speech
  – Innovative acoustic modeling algorithms that can effectively and efficiently use limited accented/dialectal training data

• Strategies
  – Cluster speakers with accents/dialects
  – Adapt acoustic models during recognition
  – Automatically bootstrap existing accented/dialectal acoustic training data; retrain acoustic models using bootstrapped data
Proposals for AM Adaptation

• Unsupervised clustering of accented speakers
  – Cluster speakers into “accent types” using acoustic training data
  – Map test speakers to one of these clusters
  – Use information from the cluster to adapt to a given test speaker

• Generalized Acoustic Model Adaptation
  – Multi-stream HMM using "super information set”:
    Acoustic characteristics    Sub-dialectical accents
    Lexicon pronunciation set   Start/end pronunciation style
  – Adaptation of Multi-stream HMMs using MLLR algorithms

• Iterative Data Bootstrapping and AM Optimization
  – Enhance dialectal acoustic training data by seeking “dialect-similar” utterances in generic PTH acoustic training corpora
  – Iteratively improve dialectal AMs using expanded training data
Lexicon Adaptation: Standard Approach

• Create rules/CARTs to add pronunciation variants.
  – Hand-written rules or
  – Rules induced from phonetically transcribed data
• Use rules to expand lexicon
• Force-align lexicon with training set to learn pronunciation probabilities.
• Prune to small number of pronunciations/word.

Lexicon Adaptation: Problems

• Limited success on dialect adaptation:
  – Mayfield Tomokiyo 2001 on Japanese-accented English: *no WER reduction*
  – Huang et al. 2000 on Southern Mandarin: *1% WER reduction over MLLR*

• Probable main problems:
  – Most gain already captured by triphones and MLLR
  – Speakers vary widely in their amount of accent so *dialect*-specific lexicons are insufficient
Lexicon Adaptation Goals

• Speaker-specific lexicon adaptation:
  Given small amounts accented PTH
  – Learn which pronunciation changes are characteristic of a given speaker/speaker cluster
  – Automatically detect appropriate “strength of accent” speaker cluster for a given speaker to determine how to dynamically set pronunciation probabilities in lexicon.
Language Model Adaptation

• Little gain expected from LM: no Wu-specific syntax, except some final particles.

• However we will do some MAP adaptation using standard PTH LM and transcribed Wu-accented training data.

(cf. Roark and Bacchiani, 2003)
Summary

• Research will focus mainly on two areas:
  – Acoustic modeling
  – Lexicon Adaptation/Pronunciation Modeling

• Two main themes will be:
  – Adaptation
  – Clustering into speaker “types”