Dialectal Chinese Speech Recognition

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

Progress Report, July 28, 2004
Dialects ( ) vs. Accented Putonghua

- Linguistically, the “dialects” are really different languages.
- This project treats Putonghua (PTH - Standard Mandarin) spoken by Shanghainese whose native language is Wu: Wu-Dialectal Chinese.
Project Goals

• Overall goal: find methods that show promise for improving recognition of accented Putonghua speech using minimal adaptation data.
• More specifically: look at various combinations of pronunciation and acoustic model adaptation.
• Demonstrate that “accentedness” is a matter of degree, and should be modeled as such.
Data Redivision

• Original data division has proved inadequate since attempts to show differential performance among test-set speakers failed.

• We redivided the corpus so that the test set contained ten strongly accented and ten weakly accented speakers.

• New division has 6.3 hours training and 1.7 hours test data for spontaneous speech.
Baseline Experiments

- Two acoustic models:
  - Mandarin Broadcast News (MBN)
  - Wu-Accented Training Data
- Language model built on HKUST 100 hour CTS data, plus Hub5, plus Wu-Accented Training Data Transcriptions
- AM’s with smaller # of GMM’s per state generalize better and yield better separation of two accent groups.

Dialectal Chinese Speech Recognition

Workshop 2004
The Center for Language and Speech Processing
Baseline Experiments

Dialectal Chinese Speech Recognition
Oracle Experiment I

Add test-speaker-specific pronunciations to the dictionary:

- sang hai `Shanghai’
- sang he 1.39
- suo `speak’
- shuo 1.67
- ze zong `this kind’
- zei zong 1.10
- e men 1.10 `we’
- uo men

Run recognition using the modified dictionary
Preliminary Oracle Results

• So far we have been unable to show any improvement using the Oracle dictionaries.
“Accentedness” Classification

• General idea: accentedness is not a categorical state, but a matter of degree.

• Can we do a better job of modeling accented speech if we distinguish between levels of accentuation?
Younger Speakers More Standard: Percentage of Fronting (e.g. sh -> s)
“Accentedness” Classification

• Two approaches:
  – Classify speakers by age, then use those classifications to select appropriate models.
  – Do direct classification into accentedness

• The former is more “interesting”, but the latter seems to work better.
Age Detection

• Shafran, Riley & Mohri (2003) demonstrated age detection using GMM classifiers including MFCC’s and fundamental frequency. Overall classification accuracy was 70.2% (baseline 33%)
• The AT&T work included 3 age ranges: youth (< 25), adult (25-50), senior (>50)
• Our speakers are all between 25 and 50. We divided them into two groups (<40, >=40)
Age Detection

- Train single-state HMM’s with up to 80 mixtures per state on:
  - Standard 39 MFCC + energy feature file
  - The above, plus three additional features for (normalized) f0: $f_0$, $\Delta f_0$, $\Delta \Delta f_0$
  - Normalization: $f_0\text{norm} = \log(f_0) - \log(f_0\text{min})$ (Ljolje, 2002)

- Use above in decoding phase to classify speaker’s utterances into “older” or “younger”

- Majority assignment is assignment for speaker
Age Detection (Base = 11/20)

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
<th>Spontaneous</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MFCC</td>
<td>MFCC+f0</td>
<td>MFCC</td>
</tr>
<tr>
<td>Spontaneous</td>
<td></td>
<td>13</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Read</td>
<td></td>
<td>13</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

Dialectal Chinese Speech Recognition

Workshop 2004
An NSF Sponsored Event
The Center for Language and Speech Processing
Accent Detection

• Huang, Chen and Chang (2003) used MFCC-based GMM’s to classify 4 varieties of accented Putonghua.

• Correct identification ranged from 77.5% for Beijing speakers to 98.5% for Taiwan speakers.
# Accent Detection (Base = 10/20)

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th><em>Spontaneous</em></th>
<th></th>
<th><em>Read</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MFCC</td>
<td>MFCC+f0</td>
<td>MFCC</td>
<td>MFCC+f0</td>
</tr>
<tr>
<td></td>
<td><em>Spontaneous</em></td>
<td>12</td>
<td>15</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td><em>Read</em></td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Dialectal Chinese Speech Recognition

Workshop 2004
An NSF Sponsored Event
The Center for Language and Speech Processing
# Correlation between Errors

| 008 | YOUNGER | 2 |
| 009 | YOUNGER | 2 |
| 011 | YOUNGER | 2 |
| 012 | YOUNGER | 2 |
| 016 | YOUNGER | 2 |
| 032 | YOUNGER | 3 |
| 035 | YOUNGER | 3 |
| 043 | OLDER   | 3 |
| 046 | OLDER   | 3 |
| 047 | OLDER   | 3 |
| 053 | OLDER   | 3 |
| 054 | OLDER   | 2 |
| 059 | OLDER   | 3 |
| 061 | YOUNGER | 2 |
| 064 | YOUNGER | 2 |
| 066 | YOUNGER | 2 |
| 067 | YOUNGER | 2 |
| 076 | OLDER   | 3 |
| 098 | OLDER   | 3 |
| 099 | OLDER   | 3 |

Dialectal Chinese Speech Recognition

Workshop 2004
An NSF Sponsored Event
The Center for Language and Speech Processing
Dialectal Chinese Speech Recognition

# Utterances Needed for Classification

Dialectal Chinese Speech Recognition

Workshop 2004
An NSF Sponsored Event
The Center for Language and Speech Processing
Rule-based Pronunciation Modeling (1)

- Motivation: using less data to obtain dialectal recognizer from PTH recognizer

- Data:
  - devtest set - 20 speakers' dialectal data taken from the 80-speaker train set
  - test set - 20 speakers' dialectal data (10 more standard plus 10 more accented)

- Mapping: \((pth, wdc [, Prob])\)
  - \(pth\): a Putonghua IF (PTH-IF)
  - \(wdc\): a Wu dialectal Chinese IF (WDC-IF), could be either a PTH-IF, or a Wu dialect specific IF (WDS-IF) unseen in PTH.
  - \(\{WDC-IF\} = \{PTH-IF\} + \{WDS-IF\}\)
  - \(Prob = Pr \{WDC-IF \mid PTH-IF, WDS-IF\}\), can be learned from WDC devtest
Rule-based Pronunciation Modeling (2)

• Observations on WDC data:
  – Mapping pairs almost the same among all three sets (train, devtest, test)
  – Mapping pairs almost identical to experts' knowledge;
  – Mapping probabilities also almost equal;
  – Syllable-dependent mappings consistent for three sets.

• Remarks:
  – Experts' knowledge can be useful;
  – Can use less data to learn rules, and adapt the acoustic model
  – Feasible to generate pronunciation models for dialectal recognizer from a standard PTH recognizer with minimal data
Rule-based Pronunciation Modeling (3)

• Observations on more standard vs. more accented speech:
  – **Common points:**
    • As a whole, the mapping pairs and probabilities (as high as 0.80) are the same, and quite similar to those summarized by experts, for 35 out of 58.
  – **Differences:**
    • More standard speakers can utter some (but not most!) IFs significantly better;
    • Over-standardization more often for more accented speakers.
  – **Remarks:**
    • Pairs \((zh, z), (ch, c), (sh, s), (iii, ii)\) as well their corresponding reverse pairs seem to be important to identify the PTH level;
    • We don't see other significant differences. Still unclear what features people use in identifying “standardness” in a speaker.
Rule-based Pronunciation Modeling (4)

- Preliminary experimental results (w/o AM adaptation)

<table>
<thead>
<tr>
<th></th>
<th>Word (%C, %A)</th>
<th>Char (%C, %A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.49 3.04</td>
<td>14.78 8.70</td>
</tr>
<tr>
<td>+ bigram</td>
<td>23.91 20.91</td>
<td>30.81 27.83</td>
</tr>
<tr>
<td>+ PTH-IF mapping</td>
<td>7.58 4.22</td>
<td>15.06 8.71</td>
</tr>
<tr>
<td>+ PTH-IF mapping + bigram</td>
<td>24.31 21.69</td>
<td>31.52 28.38</td>
</tr>
<tr>
<td>+ PTH-IF mapping + ProbLex + bigram</td>
<td>24.23 21.67</td>
<td>31.45 28.34</td>
</tr>
</tbody>
</table>

%C: %Correct, %A: %Accuracy
Work in Progress: Phonetic Substitutions

• Ratio of certain phones – s/sh, c/ch, z/zh, n/ng – is indicative of accentedness.

• How confident can one be of the true ratio within a small number of instances. For 20 instances:
  s/sh: 76% confident within 10% of true ratio
  z/zh: 88% .......................... 10% ..............
  c/ch: 75%.........................10%................
  n/ng: 81%..........................10%..............

• Number of utterances required to get 20 instances:
  s/sh 9; z/zh 14; n/ng 3.5
Further Dictionary Oracles

• “Whole dialect” oracle: use pronunciations found in all of training set for Wu-accented speech.
• “Accentenedness” oracle: have two sets of pronunciations, one for more heavily accented and one for less heavily accented speakers.
MAP Acoustic Adaptation

• Use Maximum a posteriori (MAP) adaptation to compare results of adapting to:
  – All Wu-accented speech
  – Hand-classified groups
  – Automatically-derived classifications
Minimum Perplexity Word Segmentation

- Particular word segmentation for Chinese has an effect on LM perplexity on a held-out test-set. E.g.:

  Character bigram model: \( \text{perp} = 114.78 \)
  Standard Tsinghua dictionary: \( \text{perp} = 90.11 \)
  Tsinghua dictionary + 191 common words: \( \text{perp} = 90.71 \)

- Is there a “minimum perplexity” segmentation?