Large vocabulary continuous speech recognition for Vietnamese, an under-resourced language

Nguyen Hong Quang
Pascal Nocera
Eric Castelli
Trinh Van Loan

SLTU, May 5th 2008
Contents

✓ Vietnamese characteristics
✓ Vietnamese corpus
✓ Our first Vietnamese system
✓ Using Linguistics Characteristics
  ✓ Syllabics vs. Polysyllabic words
✓ Using Acoustics Characteristics
  ✓ Tone recognition
✓ New Results
✓ Conclusion.
Vietnamese language

- An isolating tonal language
- Structure of Vietnamese syllable:

<table>
<thead>
<tr>
<th>Tonal syllable (6698)</th>
<th>Initial (22 consonants)</th>
<th>Final</th>
<th>Tone (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medial (1 semi-vowels)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nucleus (obligatory)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(13 vowels / 3 diphthong)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ending (6 consonants, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>semi-vowels)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## The differences between Vietnamese and Western languages

<table>
<thead>
<tr>
<th>Vietnamese</th>
<th>Western languages (English, French)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Under-resourced language</td>
<td>✓ “Full” resourced language</td>
</tr>
<tr>
<td>✓ Isolating language</td>
<td>✓ Multi-syllabic language</td>
</tr>
<tr>
<td>✓ Tonal language</td>
<td>✓ Non-tonal language</td>
</tr>
</tbody>
</table>

Developing new modules to integrate Vietnamese characteristics into a speech recognition system built for western languages.
Speech corpus

✓ VNSPEECHCORPUS (developed in the MICA center)

- Read speech.
- Text:
  - Paragraph (80%), conversation (20%).
  - Common part: read by all speakers (70%) and private part:
    selected randomly for each speaker (30%).
- Recorded in a studio.
- 25 male speakers and 25 female speakers of 3 major
dialect regions: the North, the Middle and the South
Speech corpus in our tests

✓ 18 speakers from the North of Vietnam
  ▪ 10 male speakers.
  ▪ 8 female speakers.
✓ 14,4 hours : average 0,8 hours / speaker
✓ Split to two corpus :
  ▪ Training corpus (11,2 hours) : 8 men + 6 women.
  ▪ Test corpus (3,2 hours) : 2 men + 2 women.
Text corpus

✓ Collected from Internet.
✓ Superfluous data are removed.
✓ 46 millions words, 2.8 millions sentences.
✓ 255 Mbytes.
Vietnamese acoustic model

✓ Characteristics:
  - Context-independent models
  - 45 HMM models (for Vietnamese phonemes)
  - 39 coefficients (13 coefficients MFCC + Δ + ΔΔ)
  - 3 emitting states for each model
  - 64 Gaussian mixture for each state

✓ Method:
  - Initial model: from French model
  - Alignment and training iterations
  - Phone Error Rate PER = 43.8%
Lexicon and language model

✓ Lexicon
  ▪ 6700 syllables.

✓ Statistical language model:
  ▪ Using CMU SLM Toolkits
  ▪ 2.5 millions bigram, 5.6 millions trigram

✓ Using Speeral system for all the experiments.
Our first LVCSR for Vietnamese

✓ Method:
  ▪ Using phonemes without tone:
    ▪ Ex: ba, bà, bá, bả, bã, bạ ➔ b a
  ▪ Using language model of tonal syllables.
  ▪ Speaker adaptation : MLLR

✓ Results (Word Error Rate : WER):
  ▪ Speaker independent : WER = 34.7%
  ▪ Speaker adaptation : WER = 25.2%
Using polysyllabic words

✓ Polysyllabic lexicon : 6700 syllables + 33000 polysyllabic words
✓ Text corpus of polysyllabic words : word segmentation by Maximum Matching algorithm on the text corpus of syllables.
✓ Method :
  ▪ Using phonemes without tone
  ▪ Using language model of tonal words (mono and polysyllabic words).
✓ Results :
  ▪ Speaker independent : WER = 30.5%
  ▪ Speaker adaptation : WER = 22.6%
Tone recognition module

✓ Tone corpus:
  ▪ Viterbi algorithm with acoustic model
  ▪ Manually corrected with our tool on Praat environment
  ▪ Tone segment -> voiced segment in the syllable (F0 calculated by AC algorithm of Praat)

✓ Each tone is represented by an HMM model
  ▪ 39 coefficients (13 coefficients MFCC + Δ + ΔΔ)
  ▪ 3 emitting states for each model
  ▪ 64 Gaussian mixture for each state

✓ Tone Accuracy Rate: 75.8%
Integrating tone information in the Speeral system

- The used scores
  - Using phoneme without tone $\Rightarrow$ Score$_{ac}$
  - Language model $\Rightarrow$ Score$_{lm}$
  - Tone recognition module $\Rightarrow$ Score$_{tone}$

- Principal idea: tone information will be added in the linguistic scoring phase
  - Syllable $\Rightarrow$ tone $\Rightarrow$ tone score
  - \[ S_{syl} = \alpha \cdot \text{Score}_{ac} + (\beta \cdot \text{Score}_{lm} + \gamma \cdot \text{Score}_{tone}) \]

- Results:
  - Speaker independent: WER = 33.5%
  - Speaker adaptation: WER = 24.7%
New Results (1/2)

✓ New Tone Recognition Module
  ▪ Using $F_0$ and short-time energy
  ▪ HMM (and neural network)
  ▪ Segment reduction and context dependent tone model

✓ New Polysyllabic Words Lexicon
  ▪ Linguistic lexicon : maximum matching
  ▪ Automatic lexicon
    • Criteria : perplexity and mutual information
    • Dynamic programming
New Results (2/2)

<table>
<thead>
<tr>
<th>Experiments (male speakers)</th>
<th>Acoustic model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speaker Independent</td>
</tr>
<tr>
<td>LM_SL without tones</td>
<td>34.6</td>
</tr>
<tr>
<td>LM_MM without tones</td>
<td>29.3</td>
</tr>
<tr>
<td>LM_DP_MI without tones</td>
<td>25.6</td>
</tr>
<tr>
<td>LM_SL with tones</td>
<td>24.7</td>
</tr>
<tr>
<td>LM_MM with tones</td>
<td>21.2</td>
</tr>
<tr>
<td>LM_DP_MI with tones</td>
<td>18.7</td>
</tr>
</tbody>
</table>

- LM_SL : syllabic language model
- LM_MM : polysyllabic language model (word segmentation by MM)
- LM_DP_MI : polysyllabic language model (word segmentation by dynamic programming using mutual information)
Conclusions

✓ Results with polysyllabic words are better than syllabic. Results with automatic lexicon are better than with linguistic lexicon.

✓ Using tone information based on MFCC coefficients improve slightly the performance of the system. Best results are obtained with tone information based on $F_0$ and short-time energy.