Out-of-Vocabulary Spoken Term Detection

PhD Thesis Talk

Dong Wang

Supervisors: Simon King, Joe Frankel, Jim Scobbie
CSTR, University of Edinburgh
December 8th, 2009
Contents

• Research subject and hypothesis
• Research contributions
  ❑ Stochastic pronunciation modeling
  ❑ Term-dependent discriminative decision making
  ❑ Direct posterior confidence estimation
• Summary and future work
Contents

• Research subject and hypotheses

• Research contributions
  - Stochastic pronunciation modeling
  - Term-dependent discriminative decision making
  - Direct posterior confidence estimation

• Summary and future work
Information Headache and Audio Retrieval

How to manage the flood of information?
Techniques for Audio Retrieval

- Large Vocabulary Continuous Speech Recognition (LVCSR)
  - Applying Automatic Speech Recognition (ASR) technology, transcribe speech to text for documentation and search.
  - Manual corrections may apply.

- Spoken Document Retrieval (SDR)
  - Transcribe spoken documents to word/subword –based intermediate representation applying ASR, and then retrieve relevant documents exploiting conventional information retrieval techniques.
  - Focus on entire documents

- Keyword Spotting (KS)
  - Spot occurrences of some keywords from audio streams applying ASR technology.
  - Keywords are usually pre-defined in a spotting task.
Spoken Term Detection: Task Definition

Spoken Term Detection (STD):
NIST: Search vast, heterogeneous audio archives for occurrences of spoken terms.

Features:
- Focus on spoken terms or keywords (unlike LVCSR)
- Concern individual terms (unlike SDR)
- Open-vocabulary (unlike KS)

... Kuwait ... car ...
Spoken Term Detection: Framework

\[ d = (K, s = (t_{start}, t_{end}), v_a, v_l, \ldots) \]

\[ c_{lattice}(d) = P(K_{start}^{end} | O) \]

\[ \sum_{\alpha, \beta} P(O | K_\alpha, K_{start}^{end}, K_\beta) P(K_\alpha, K_{start}^{end}, K_\beta) \]

\[ \frac{P(O)}{P(O)} \]

<table>
<thead>
<tr>
<th>TERM</th>
<th>ST</th>
<th>ET</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuwait</td>
<td>20.2</td>
<td>20.8</td>
<td>0.99</td>
</tr>
<tr>
<td>Kuwait</td>
<td>22.4</td>
<td>23.1</td>
<td>0.87</td>
</tr>
<tr>
<td>car</td>
<td>20.1</td>
<td>20.3</td>
<td>0.35</td>
</tr>
<tr>
<td>car</td>
<td>28.5</td>
<td>28.8</td>
<td>0.32</td>
</tr>
<tr>
<td>car</td>
<td>30.8</td>
<td>31.2</td>
<td>0.05</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ assert(d) = \begin{cases} 
1 & c_{lattice}(d) \geq \theta \\
0 & c_{lattice}(d) < \theta 
\end{cases} \]
Spoken Term Detection: Evaluation Metrics

\[ ATWV = \sum_{K \in \Delta} \frac{P_{\text{miss}}(K) + \beta P_{\text{FA}}(K)}{|\Delta|} \]
Challenges from Out-of-Vocabulary Terms

- **Out-of-Vocabulary (OOV) words and terms**
  - OOV words are those words absent from the system dictionary and training data.
  - In STD, OOV terms are defined as terms consisting of one or more OOV words.
  - E.g., Kuwait, Mujeeb-u-Rahman, Taliban.

- **OOV sources**
  - Limited dictionaries.
  - Human language dynamics (names, addresses, brands, techniques, diseases, etc).

- **OOV for STD**
  - OOV is an intrinsic problem of STD which is an open-vocabulary task.
  - The OOV issue, if not addressed, will become more and more serious as time goes on, and will finally devalues a STD system.
Out-of-Vocabulary STD

- Current approaches to OOV term detection
  - Convert search terms to subword representations, usually phonemes
    - Rules, Decision trees, HMMs, MLPs, etc.
  - Search the subword representations in subword lattices generated by subword ASR.
## Baseline System

<table>
<thead>
<tr>
<th></th>
<th>ASR</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Meetings/IHM</td>
<td>INV/OOV terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>256/484</td>
</tr>
<tr>
<td>Features</td>
<td>MFCC+CMN/CVN</td>
<td>Term search</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exact match</td>
</tr>
<tr>
<td>HMMs</td>
<td>3-state CD phones</td>
<td>Pron. prediction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CART(Festival)</td>
</tr>
<tr>
<td>AM training data</td>
<td>80.2 hours</td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lattice posterior</td>
</tr>
<tr>
<td>LM training data</td>
<td>521 MW</td>
<td>Decision strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Term-independent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WER/PER (ASR)</th>
<th>ATWV (INV STD)</th>
<th>ATWV (OOV STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WORD 3-gram</strong></td>
<td>39.5</td>
<td>0.5661</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>PHONE 4-gram</strong></td>
<td>44.53</td>
<td>0.2438</td>
<td>-0.1647</td>
</tr>
<tr>
<td><strong>PHONE 5-gram</strong></td>
<td>41.65</td>
<td>0.3670</td>
<td>-0.1232</td>
</tr>
<tr>
<td><strong>PHONE 6-gram</strong></td>
<td>40.49</td>
<td>0.4173</td>
<td><strong>-0.1010</strong></td>
</tr>
</tbody>
</table>
Thesis Hypothesis

- A major shortcoming of current OOV treatment is that **special** properties of OOV terms are not considered **specially**.

- **OOV special properties**
  - High uncertainty in pronunciation
    - Acoustic uncertainty interweaved with linguistic uncertainty
    - E.g., Kwait, Skype, Linux, EBLUL, Iraq, Buccleuch Place…
  - High diversity in term properties
    - Occurrence rate, phonemic structure, linguistic background, morphological form, etc.
  - Weakness in acoustic and language modelling

- **We hypotheses that treating these OOV special properties will improve OOV term detection.**
Thesis Contributions

- Stochastic Pronunciation Modelling
  - Term Search
  - Confidence Estimation
  - Decision Making

- Direct Posterior Confidence Estimation

- Discriminative Decision Making

ASR

ATWV DET

NIST Tool
Contents

• Research subject and hypotheses

• **Research contributions**
  - Stochastic pronunciation modeling
  - Term-dependent discriminative decision making
  - Direct posterior confidence estimation

• Summary and future work
Stochastic Pronunciation Modelling

ASR

STD subsystem

NIST Tool

Term Search

Confidence Estimation

Decision Making

Direct Posterior Confidence Estimation

Discriminative Decision Making

Page • 15
Stochastic Pronunciation Modelling

- A high degree of uncertainty exists in OOV pronunciations.
- This uncertainty can be represented by a probability distribution \( P(Q) \), which we call a stochastic pronunciation model, or SPM.

\[
c_{\text{pron}} = P(Q \mid K)
\]

\[
c = (1 - \gamma)c_{\text{lattice}} + \gamma c_{\text{pron}}
\]
**Soft Match**

- Soft Match: Mismatches are allowed in lattice search to compensate ASR errors and pronunciation variation.

\[
c_{\text{match}} = \prod_{q \in Q, q \Leftrightarrow q'} P(q' \mid q)
\]

\[
c = (1 - \nu)c_{\text{lattice}} + \nu c_{\text{match}}
\]

Example:

- \(c_{\text{lattice}} = 0.32\)
- \(c_{\text{match}} = 1.0\)
- \(c_{\text{lattice}} = 0.54\)
- \(c_{\text{match}} = 0.2\)
Joint-multigram Model-Based Pronunciation Prediction

THOUGHT [th ao t]

\[ \begin{array}{c}
\text{TH} \\
\text{OU} \\
\text{GH} \\
\text{T}
\end{array} \]

\[ \begin{array}{c}
\text{[th]} \\
\text{[ao]} \\
\text{[ ]} \\
\text{[t]}
\end{array} \]

\[ U = (u_1 u_2, \ldots, u_L) = \left( \tilde{g}_1 \tilde{g}_2, \ldots, \tilde{g}_L \right) \]

\[ P(U) = P(u_1 u_2, \ldots, u_L) = \prod_i P(u_i | u_{i-1} u_{i-2} \ldots) \]

\[ \frac{P(Q | G)}{P(U)} = \sum_{G(U) = G, Q(U) = Q} P(U) \]

Model | WER% | PER%
--- | --- | ---
CART (stop=1) | 35.2 | 8.7
Joint-multigram | 33.2 | 8.2
+Insertion comp. | 32.7 | 8.1
+Reverse dec. | 31.3 | 7.8
+Pron. Unification | **30.3** | **7.5**

1-best prediction

n-best prediction
Experiments: Joint-multigram Model-based STD

<table>
<thead>
<tr>
<th>Model</th>
<th>Norm.</th>
<th>ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART(Baseline)</td>
<td>No</td>
<td>-0.1010</td>
</tr>
<tr>
<td>CART(Baseline)</td>
<td>Yes</td>
<td>0.2126</td>
</tr>
<tr>
<td>Joint-multigram</td>
<td>No</td>
<td>0.0273</td>
</tr>
<tr>
<td>Joint-multigram</td>
<td>Yes</td>
<td>0.2761</td>
</tr>
<tr>
<td>CART+JM</td>
<td>Yes</td>
<td>0.3030</td>
</tr>
</tbody>
</table>

Confidence Normalisation[1]:

\[
\zeta_k(c(d)) = \frac{c(d) \times \alpha + \gamma}{N_{true}^K} - \beta \frac{1 - c(d) \times \alpha - \gamma}{T - N_{true}^K}
\]

\[
c = 1 - (1 - c_1)(1 - c_2)^{\alpha}
\]

**Experiments: Stochastic Pronunciation Modelling**

<table>
<thead>
<tr>
<th>Model</th>
<th>ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>0.2126</td>
</tr>
<tr>
<td>Joint-multigram</td>
<td>0.2761</td>
</tr>
<tr>
<td>SPM</td>
<td>0.3153</td>
</tr>
<tr>
<td>Soft Match</td>
<td>0.3275</td>
</tr>
<tr>
<td>SPM+Soft Match</td>
<td><strong>0.3427</strong></td>
</tr>
</tbody>
</table>

![Graph showing performance metrics](image-url)
Summary

- Joint-multigram models provide better performance than decision trees in terms of both pronunciation prediction and OOV term detection.

- The stochastic pronunciation modelling approach substantially improves OOV term detection.

- Compared to soft match, a widely used approach to pronunciation uncertainty treatment, stochastic pronunciation modelling is more efficient to control false alarms and thus achieves better performance in the case of low FA rate. Combining these two approaches provides further gain.
Contents

• Research subject and hypotheses

• **Research contributions**
  - Stochastic pronunciation modeling
  - Term-dependent discriminative decision making
  - Direct posterior confidence estimation

• Summary and future work
Discriminative Decision Making

Stochastic Pronunciation Modelling

Term Search

STD subsystem

Confidence Estimation

Decision Making

Direct Posterior Confidence Estimation
Discriminative Decision Making

- Decision theory shows that decisions based on classification posterior probabilities are optimal in terms of decision errors.

\[
assert(d) = \begin{cases} 
C_{hit} & P(C_{hit} \mid d) \geq \theta \\
C_{FA} & P(C_{hit} \mid d) < \theta 
\end{cases}
\]

- The lattice-based confidence does not lead to optimal decisions.
Discriminative Decision Making

- A discriminative mapping is desired to convert the lattice-based confidence to a classification posterior, so that a discriminative decision can be achieved.

\[ c_{disc}(d) = P(C_{hit} | d) = f(c_{lattice}(d), c_1, c_2, \ldots) \]

- A MLP and a SVM are used to implement the discriminative mapping.
Term-Dependent Confidence

- Term-dependent factors can be taken by the discriminative mapping, leading to a term-dependent discriminative confidence which is supposed to compensate the high diversity among OOV terms.

\[
c_{\text{disc}}(d) = f(c_{\text{lattice}}(d), R_0(K_d), R_1(K_d), \ldots)
\]

\[
R_0(K) = \frac{\sum_i c_{\text{lattice}}(d^K_i)}{T}
\]

\[
R_1(K) = \frac{\sum_i (1 - c_{\text{lattice}}(d^K_i))}{T}
\]
## Experimental Results

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Term Dependence</th>
<th>Model</th>
<th>ATWV(INV)</th>
<th>ATWV(OOV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lattice</td>
<td>NO</td>
<td>-</td>
<td>0.4743</td>
<td>0.2761</td>
</tr>
<tr>
<td>Disc.</td>
<td>NO</td>
<td>MLP</td>
<td>0.5453</td>
<td>0.2927</td>
</tr>
<tr>
<td>Disc.</td>
<td>NO</td>
<td>SVM</td>
<td>0.5432</td>
<td>0.2894</td>
</tr>
<tr>
<td>Disc.</td>
<td>YES</td>
<td>MLP</td>
<td><strong>0.5466</strong></td>
<td><strong>0.2931</strong></td>
</tr>
<tr>
<td>Disc.</td>
<td>YES</td>
<td>SVM</td>
<td>0.5434</td>
<td>0.2921</td>
</tr>
</tbody>
</table>

1. Discriminative confidence, based on either MLP or SVM, performs better than the lattice-based confidence.
2. Integrating term-dependent factors provides further performance improvement.
Term-Dependent Discriminative Decision for Uncertainty Treatment

\[
c_{\text{disc}}^{\text{spm}} (d) = f(c_{\text{lattice}} (d), c_{\text{pron}}, R_0(K_d), R_1(K_d), \ldots)
\]

\[
c_{\text{disc}}^{\text{softmatch}} (d) = f(c_{\text{lattice}} (d), c_{\text{match}}, R_0(K_d), R_1(K_d), \ldots)
\]
Summary

- Term-dependent discriminative confidence improves STD performance significantly, for both INV and OOV terms.
- Discriminative decision making works well with SPM and Soft Match – based pronunciation uncertainty treatment. The discriminative confidence-based combined system achieves better performance than individual systems and the combined system based on lattice-based confidence.
Contents

- Research subject and hypotheses
- **Research contributions**
  - Stochastic pronunciation modeling
  - Term-dependent discriminative decision making
  - Direct posterior confidence estimation
- Summary and future work
Direct Posterior Confidence

Stochastic Pronunciation Modelling

Term Search

Confidence Estimation

Decision Making

Direct Posterior Confidence Estimation

Discriminative Decision Making
Weakness of Lattice-based confidence

- Till now, all the confidence measures are based on lattice-based posterior probabilities and therefore heavily relies on the power of the acoustic models and language models which are usually weak for OOV terms.

\[
C_{lattice} = P \left( K_{t_{\text{start}}}^{t_{\text{end}}} \mid O \right) = \sum_{\alpha, \beta} P \left( O \mid K_{\alpha}, K_{t_{\text{start}}}^{t_{\text{end}}}, K_{\beta} \right) P \left( K_{\alpha}, K_{t_{\text{start}}}^{t_{\text{end}}}, K_{\beta} \right) / P \left( O \right)
\]
MLP-based Frame-wise Acoustic Confidence

Generative model

\[ P(q_t \mid O) = P(q_t \mid o_{t-4}, \ldots, o_t, \ldots o_{t+4}) \]
\[ = P(q_t \mid \chi_t) \]

MLP-based frame-wise confidence:
- Discriminative
- Local score

Discriminative model
MLP-based Acoustic Posterior Confidence

\[
\prod_{t=t_0}^{t_1} p(g \mid \chi_t) \prod_{t=t_1}^{t_2} p(uw \mid \chi_t) \prod_{t=t_2}^{t_3} p(g \mid \chi_t) \prod_{t=t_3}^{t_4} p(el \mid \chi_t)
\]

\[
c_{mlp} = P(K_{t_{\text{end}}}^{t_{\text{end}}} \mid O) = \prod_{t_{\text{start}}}^{t_{\text{end}}} P(q_t \mid O) = \prod_{t=t_{\text{start}}}^{t_{\text{end}}} P(q_t \mid \chi_t)
\]
Lattice-based Language Model Posterior Confidence

Context-dependent discriminative model

\[ c_{im} = P(K^l | L) \]

\[ = \sum_{\alpha_K^l, \beta_K^l} \frac{P(\alpha_K^l, K^l, \beta_K^l)}{P(L)} \]

Lattice-based LM posterior
Confidence Integration

- Acoustic-Language posterior confidence

\[ c_{mlp+lm} = 1 - (1 - c_{mlp})^\alpha (1 - c_{lm}) \]

- MLP-Lattice confidence integration

\[ c_{mlp+lattice} = 1 - (1 - c_{mlp})^\alpha (1 - c_{lattice}) \]
## Experiments: Direct Posterior Confidence

<table>
<thead>
<tr>
<th>Confidence</th>
<th>ATWV (INV)</th>
<th>ATWV (OOV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lattice</td>
<td>0.4743</td>
<td>0.2761</td>
</tr>
<tr>
<td>MLP</td>
<td>0.4902</td>
<td>0.2971</td>
</tr>
<tr>
<td>MLP+LM</td>
<td>0.4963</td>
<td>0.2941</td>
</tr>
<tr>
<td>MLP+Lattice</td>
<td><strong>0.5344</strong></td>
<td><strong>0.2973</strong></td>
</tr>
</tbody>
</table>

1. Acoustic posterior confidence outperforms lattice-based confidence.
2. LM integration provides remarkable improvement for INV terms, but does not help OOV terms.
3. Integrating lattice-based confidence provides more improvement for INV terms.
Experiments: Direct Posterior Confidence For SPM and Soft Match

<table>
<thead>
<tr>
<th>Uncertainty Treat</th>
<th>ATWV (Lattice)</th>
<th>ATWV (MLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONE</td>
<td>0.2761</td>
<td>0.2971</td>
</tr>
<tr>
<td>SPM</td>
<td>0.3153</td>
<td>0.3288</td>
</tr>
<tr>
<td>Soft Match</td>
<td>0.3275</td>
<td><strong>0.3387</strong></td>
</tr>
</tbody>
</table>

\[
c^{spm} = (1 - \gamma)c_{mlp} + \gamma c_{pron}
\]

\[
c^{softmatch} = (1 - \nu)c_{mlp} + \nu c_{match}
\]
Experiments: Direct Posterior Confidence and Discriminative Decision with Uncertainty Treatment

\[ c_{disc}^{spm} = f (c_{mlp+lattice} , c_{mlp} , c_{lattice} , c_{pron} , R_0 , R_1) \]
\[ c_{disc}^{softmatch} = f (c_{mlp+lattice} , c_{mlp} , c_{lattice} , c_{match} , R_0 , R_1) \]
Contents

• Research subject and hypotheses

• Research contributions
  - Stochastic pronunciation modeling
  - Term-dependent discriminative decision making
  - Direct posterior confidence estimation

• Summary and future work
Summary

- This thesis proposed three approaches to enhance OOV term detection. Experimental results attested our hypothesis that OOV special properties should be treated specially when detecting OOV terms.
Future Work

- Fast search
- Variable-sized lattice units
- Learning OOV properties
Thanks to my supervisors.
Thanks to all CSTR members.
Thanks to my Chinese friends and my family.
Thanks to the EdSST fellowship.