Pairwise Learning using Multi-lingual Bottleneck Features for Low-resource Query-by-example Spoken Term Detection

Y Eugen Yuan\textsuperscript{1,2}, Cheung-Chi Leung\textsuperscript{2}, Lei Xie\textsuperscript{1}, Hongjie Chen\textsuperscript{1}, Bin Ma\textsuperscript{2}, Haizhou Li\textsuperscript{2,3}

\textsuperscript{1}School of Computer Science, Northwestern Polytechnical University, Xi’an, China
\textsuperscript{2}Institute for Infocomm Research, A*STAR, Singapore
\textsuperscript{3}Department of ECE, National University of Singapore, Singapore

March 9, 2017
Outline

1 Introduction
   - Background
   - Motivation and contribution

2 Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3 Experiments
   - Data and evaluation
   - Results and analysis

4 Conclusions
Outline

1. Introduction
   - Background
   - Motivation and contribution

2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3. Experiments
   - Data and evaluation
   - Results and analysis

4. Conclusions
Problem description of low-resource query-by-example spoken term detection (QbE-STD)

- A search problem for the occurrence of a spoken query in audio archives.
- Limited training data in low-resource scenarios.
- Difficult to give utterances with labels if no prior linguistic knowledge in the language.
Previous work

- Extract unsupervised acoustic features directly in low-resource target languages [1, 2, 3, 4].
- Extract posterior or bottleneck features (BNFs) from neural networks (NNs) trained using high-resource non-target languages [5, 6, 7, 8, 9].
Outline

1 Introduction
   - Background
   - Motivation and contribution

2 Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3 Experiments
   - Data and evaluation
   - Results and analysis

4 Conclusions
Motivation

- **Pairwise learning**
  - Training NNs with paired examples.
  - Successful for various tasks, including face verification [10], sentence similarity [12], phone discrimination [11], and our previous study [13] on a word discrimination task.

- **Multi-lingual BNFs**
  - A kind of compact representations.
  - More language-independent and more flexible for rapid language adaptation; especially in low-resource languages.
Motivation

- **Pairwise learning**
  - Training NNs with paired examples.
  - Successful for various tasks, including face verification [10], sentence similarity [12], phone discrimination [11], and our previous study [13] on a word discrimination task.

- **Multi-lingual BNFs**
  - A kind of compact representations.
  - More language-independent and more flexible for rapid language adaptation; especially in low-resource languages.
Contribution

- The first attempt to use pairwise learning based on multi-lingual BNFs.
- The first attempt to use pairwise learning for QbE-STD.
Outline

1. Introduction
   - Background
   - Motivation and contribution

2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3. Experiments
   - Data and evaluation
   - Results and analysis

4. Conclusions
Multi-lingual BNF extraction

- Train a multi-lingual bottle-type NN from non-target languages.
Outline

1 Introduction
   - Background
   - Motivation and contribution

2 Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3 Experiments
   - Data and evaluation
   - Results and analysis

4 Conclusions
Pairwise learning with an autoencoder

- Align two sequences of multi-lingual BNFs with DTW.
- Train a pre-trained AE with Mean Squared Error (MSE) using aligned frame pairs.
- Extract newly learned feature representation from an internal hidden layer of trained NN.

![Diagram](image-url)
Outline

1. Introduction
   - Background
   - Motivation and contribution

2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3. Experiments
   - Data and evaluation
   - Results and analysis

4. Conclusions
NN-based template matching method for QbE-STD
Outline

1 Introduction
   - Background
   - Motivation and contribution

2 Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3 Experiments
   - Data and evaluation
   - Results and analysis

4 Conclusions
Data

Target language (for QbE-STD)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Training set (No. of word pairs)</th>
<th>Keyword set (No. of examples)</th>
<th>Test set (No. of utterances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT [3, 4]</td>
<td>10,000</td>
<td>346</td>
<td>944</td>
</tr>
<tr>
<td>Switchboard</td>
<td>100,000</td>
<td>346</td>
<td>100</td>
</tr>
</tbody>
</table>

Non-target languages (for multi-lingual BNFs extractor)
- HKUST Mandarin Chinese (LDC2005S15; 170hr)
- Fisher Spanish (LDC2001S01; 152hr)
Data

- **Target language (for QbE-STD)**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Training set (No. of word pairs)</th>
<th>Keyword set (No. of examples)</th>
<th>Test set (No. of utterances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT [3, 4]</td>
<td>10,000</td>
<td>346</td>
<td>944</td>
</tr>
<tr>
<td>Switchboard</td>
<td>100,000</td>
<td>346</td>
<td>100</td>
</tr>
</tbody>
</table>

- **Non-target languages (for multi-lingual BNFs extractor)**
  - HKUST Mandarin Chinese (LDC2005S15; 170hr)
  - Fisher Spanish (LDC2001S01; 152hr)
Metrics of evaluation

**MAP**: the mean average precision of each query in the test set.

**P@N**: the average precision of the top N utterances where N is the number of the correct hit utterances in test set.

**P@5/P@10**: the average precision of the first five or ten ranked utterances.
Outline

1. Introduction
   - Background
   - Motivation and contribution

2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)

3. Experiments
   - Data and evaluation
   - Results and analysis

4. Conclusions
# QbE-STD on TIMIT and Switchboard

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Representation</th>
<th>No pairwise training (MAP/P@N/P@10)</th>
<th>Pairwise training (MAP/P@N/P@10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>MFCCs</td>
<td>0.285/0.289/0.247</td>
<td>0.297/0.293/0.257</td>
</tr>
<tr>
<td></td>
<td>BNFs (Mandarin)</td>
<td>0.494/0.459/0.413</td>
<td>0.571/0.538/0.467</td>
</tr>
<tr>
<td></td>
<td>BNFs (Spanish)</td>
<td>0.540/0.512/0.446</td>
<td><strong>0.594/0.561/0.484</strong></td>
</tr>
<tr>
<td></td>
<td>BNFs (Multi-lingual)</td>
<td><strong>0.552/0.524/0.461</strong></td>
<td><strong>0.594/0.561/0.490</strong></td>
</tr>
<tr>
<td>Switchboard</td>
<td>MFCCs</td>
<td>0.232/0.200/0.232</td>
<td>0.258/0.236/0.260</td>
</tr>
<tr>
<td></td>
<td>BNFs (Mandarin)</td>
<td>0.370/0.338/0.446</td>
<td>0.417/0.382/0.451</td>
</tr>
<tr>
<td></td>
<td>BNFs (Spanish)</td>
<td>0.388/0.358/0.475</td>
<td>0.430/0.398/<strong>0.484</strong></td>
</tr>
<tr>
<td></td>
<td>BNFs (Multi-lingual)</td>
<td><strong>0.400/0.365/0.485</strong></td>
<td><strong>0.435/0.404/0.473</strong></td>
</tr>
</tbody>
</table>
Analysis

- **Multi-lingual BNFs**
  - Are much better than MFCCs.
  - Usually outperform the cross-lingual BNFs.

- **Pairwise learning**
  - Provides a more efficient feature representation for QbE-STD.
  - Usually hold the best performance with multi-lingual BNFs in the QbE-STD tasks.
Analysis

- Multi-lingual BNFs
  - Are much better than MFCCs.
  - Usually outperform the cross-lingual BNFs.

- Pairwise learning
  - Provides a more efficient feature representation for QbE-STD.
  - Usually hold the best performance with multi-lingual BNFs in the QbE-STD tasks.
Dependence on the amount of word-pair supervision

- With more word pairs, pairwise learned NN feature representation gives a better performance.
- With 10,000 word pairs, pairwise learned features give comparable performance to those using all word pairs.
Effect of input features and frame alignment

Regardless of either MFCCs or multi-lingual BNFs are used for frame-level DTW alignment, multi-lingual BNFs consistently provide much better QbE-STD results than MFCCs as input features.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Input features of AE</th>
<th>Features for alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCCs</td>
<td>MFCCs</td>
</tr>
<tr>
<td>TIMIT</td>
<td>0.285/0.289/0.247</td>
<td>0.320/0.314/0.274</td>
</tr>
<tr>
<td></td>
<td>0.587/0.556/0.486</td>
<td>0.594/0.561/0.490</td>
</tr>
<tr>
<td>Switchboard</td>
<td>MFCCs</td>
<td>0.258/0.236/0.260</td>
</tr>
<tr>
<td></td>
<td>BNFs (Multi-lingual)</td>
<td>0.432/0.395/0.483</td>
</tr>
<tr>
<td></td>
<td>BNFs (Multi-lingual)</td>
<td>0.435/0.404/0.473</td>
</tr>
</tbody>
</table>
We have proposed to perform pairwise learning using multilingual BNFs of word pairs for QbE-STD. Pairwise learning makes the resulted features more capable in phonetic discrimination for a new target language. Brings further performance improvement on low-resource QbE-STD tasks.

In future work, we will investigate methods of word-level pairwise learning for this task, which avoids frame-level alignment of word pairs.
References I


References II


