1. Introduction

◆ Bilingual speakers are known for their ability to code-switch their languages during communication.
◆ This study proposes a data-driven approach to phone set construction by integrating acoustic features and cross-lingual context-sensitive articulatory features into distance measure between phone units.

2. Data Sparseness Problem in Code-Switching Speech Recognition

◆ Hard to collect sufficient utterances of L2 spoken by speakers to train code-switching ASR for context-dependent acoustic model training.

3. Proposed Idea

4. System Diagram

5. Used Features

➢ Traditional acoustic features
  - MFCC, ΔMFCC, Δ ΔMFCC
  - energy, Δenergy, Δ Δenergy
➢ Articulatory features

<table>
<thead>
<tr>
<th>Articulatory Attributes</th>
<th>vowel</th>
<th>fricative</th>
<th>nasal</th>
<th>voiced</th>
<th>back</th>
<th>coronal</th>
<th>labial</th>
<th>approximant</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>Glottal</td>
<td>low</td>
<td>vocalic</td>
<td>dental</td>
<td>high</td>
<td>mid</td>
<td>continuant</td>
<td></td>
</tr>
<tr>
<td>velar</td>
<td>Anterior</td>
<td>retroflex</td>
<td>round</td>
<td>tense</td>
<td>silence</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


6. Hierarchical Triphone Model Clustering

➢ Distance estimation
  - Among triphones having the same central phone
  - Different triphone acoustic models of the current triphones
  - Articulatory models of their preceding and succeeding states

Distance between triphones, X and Y

\[ D(X, Y) = w_{AC} D_{AC}(X, Y) + w_{AA} D_{AA}(X, Y) \]

where

\[ D_{AC}(X, Y) = \frac{1}{2} \sum_{i=0}^{5} w_{s} KL(x_{i}, y_{i}) \]

\[ D_{AA}(X, Y) = \frac{1}{2} (D_{AAAC}(X, Y) + D_{AASw}(X, Y)) \]

\[ D_{AC} \] and \[ D_{AA} \] are calculated by using KL-Divergence.

➢ After merging closest triphone-pair, update the mean and covariance vectors of the new merged triphone

7. Code-Switching Language Model Building

➢ Generate word bi-gram counts by using a machine translation-based method for unseen code-switched word bigrams

8. Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simply Combined Phone Set</td>
<td>80.79%</td>
</tr>
<tr>
<td>IPA-based Phone Set</td>
<td>81.31%</td>
</tr>
<tr>
<td>Traditional Data Driven-Based Method</td>
<td>81.53%</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>84.27%</td>
</tr>
</tbody>
</table>

9. Conclusions

◆ The proposed method outperforms other three traditional methods by 3.48%, 2.96% and 2.74%, respectively.
◆ The evaluation results confirm that the context-sensitive articulatory attributes can help eliminate the data sparseness problem and increase reliability in estimating distances among triphones.