Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban
Sarah Samson, Laurent Besacier, Benjamin Lecouteux, Mohamed Dyab

To cite this version:
Sarah Samson, Laurent Besacier, Benjamin Lecouteux, Mohamed Dyab. Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban. Interspeech 2015, Sep 2015, Dresden, Germany. 2015. hal-02015501

HAL Id: hal-02015501
https://hal.archives-ouvertes.fr/hal-02015501
Submitted on 12 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban

Sarah Samson Juan\textsuperscript{1,2}, Laurent Besacier\textsuperscript{2}, Benjamin Lecouteux\textsuperscript{2}, Mohamed Dyab\textsuperscript{2}

\textsuperscript{1}Faculty of Computer Science and Information Technology, UNIMAS, Malaysia
\textsuperscript{2}Grenoble Informatics Laboratory (LIG), Univ. Grenoble Alpes, Grenoble, France

contact: laurent.besacier@imag.fr

---

**Introduction**

- **Context**
  - Exploit data from closely-related language for developing ASR for very under-resourced language
  - Case study on Iban, a language spoken in Sarawak, Borneo Island.
  - The language is close to Malay, which is largely spoken in Malaysia.

- **Objectives**
  - Fast-bootstrapping approach for building Iban pronunciation dictionary
  - Improve performance of (low-resourced) Iban acoustic models

- **Problems**
  - Building an Iban pronunciation dictionary from scratch
  - Very limited training data for acoustic model training

---

**Methodology**

- Semi-supervised approach for Iban pronunciation dictionary
- Semi-supervised lexicon design using Malay data
- Cross-lingual acoustic modelling with limited training data
- Cross-lingual SGMMs - porting Universal Background Model (UBM) across languages
- Cross-lingual DNNs - language-specific top layer for DNN

---

**Using a closely related language (Malay) for Iban pronunciation dictionary**

**Hybrid G2P:**

-pure Iban?

---

**Pronunciation dictionary evaluation on Iban ASR**

- **Acoustic model**
  - Context dependent
  - HMM triphone state
  - 39 MFCCs
  - Train on 7hr news data

- **ASR tool**
  - Kaldi

---

**Baseline ASR using Hybrid G2P pronunciation dictionary**

- **No speaker adaptation**

---

**Cross-lingual acoustic modelling for low-resourced Iban ASR**

- **Training data**
  - Malay - 120h MASS corpus
  - English - 118h TED corpus
  - Iban - 1h condition ; 7h condition

- **Using SGMM**
  - No speaker adaptation
  - UBM Gaussians : 600
  - No. of substates : 805 (1h) and 10K (7h)
  - Approach: initialize Iban SGMM using UBM trained on source language data (monolingual or multilingual)

- **Using DNN**
  - 6 hidden layers, each with 1024 units
  - LDA, MLLT, SAT-MLLR (speaker adaptation)

**Evaluation of cross-lingual/multilingual SGMM on Iban ASR**

<table>
<thead>
<tr>
<th>Cross-lingual SGMM</th>
<th>Amount of training data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1h</td>
</tr>
<tr>
<td>Using monolingual UBM:</td>
<td></td>
</tr>
<tr>
<td>a. Malay</td>
<td>28.3</td>
</tr>
<tr>
<td>b. English</td>
<td>30.8</td>
</tr>
</tbody>
</table>

**Using multilingual UBM:**

<table>
<thead>
<tr>
<th></th>
<th>1h</th>
<th>7h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Iban + Malay</td>
<td>27.2</td>
<td>19.6</td>
</tr>
<tr>
<td>b. Iban + English</td>
<td>29.8</td>
<td>19.2</td>
</tr>
<tr>
<td>c. English + Malay</td>
<td>29.4</td>
<td>19.1</td>
</tr>
<tr>
<td>d. Iban + Malay + English</td>
<td>28.3</td>
<td>19.2</td>
</tr>
</tbody>
</table>

# of substates: 805 (1h) and 10K

**Evaluation of language specific DNN on Iban ASR**

<table>
<thead>
<tr>
<th>DNN with lang. specific top layer</th>
<th>Amount of train data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1h</td>
</tr>
<tr>
<td>a. Hidden layers from English</td>
<td>19.1</td>
</tr>
<tr>
<td>b. Hidden layers from Malay</td>
<td>18.9</td>
</tr>
</tbody>
</table>

**Towards a zero-shot ASR**

- **Approach and setup**
  - Train Iban ASR on automatic transcripts - obtained from decoding Iban training data with Malay acoustic models
  - Malay acoustic models - 120h training data; SGMMs
  - Iban ASR (from automatic transcripts) - 7h training data; train GMM, SGMM and DNN models

**Results**

<table>
<thead>
<tr>
<th>ASR system (7h)</th>
<th>GMM</th>
<th>SGMM</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised (no spkr adapt.)</td>
<td>36.0</td>
<td>18.9</td>
<td>18.4</td>
</tr>
<tr>
<td>Supervised (with spkr adapt.)</td>
<td>19.7</td>
<td>16.6</td>
<td>15.8</td>
</tr>
<tr>
<td>Unsupervised (no spkr adapt.)</td>
<td>21.4</td>
<td>18.6</td>
<td>18.9</td>
</tr>
</tbody>
</table>

**Conclusions**

- Built first ASR system for Iban - corpus and Kaldi scripts available on github: [https://github.com/sarahjuan/iban](https://github.com/sarahjuan/iban)
- Using Malay (closely-related) data in the lexicon design for Iban is better than using English (not a close language)
- Cross-lingual effect on acoustic model is more evident on SGMM experiment for 1h training data (very limited condition)
- Language specific top layer for DNN (English and Malay source languages do not make a difference for Iban DNN)
- Improving Zero-shot ASR: conf measures to select the best transcripts