Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR

Viet Bac Le, Laurent Besacier
CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE
email: viet-bac.le@imag.fr, laurent.besacier@imag.fr

Abstract
This paper presents a comparison of some different acoustic modeling strategies for under-resourced languages. When only limited speech data are available for under-resourced languages, we propose some crosslingual acoustic modeling techniques. We apply and compare these techniques in Vietnamese ASR. Since there is no pronunciation dictionary for some under-resourced languages, we investigate grapheme-based acoustic modeling. Some initialization techniques for context independent modeling and some question generation techniques for context dependent modeling are applied and compared for Khmer ASR.

Index Terms: ASR, acoustic modeling, Vietnamese, Khmer.

1. Introduction
Nowadays, computers are heavily used to communicate via text and speech. Text processing tools, electronic dictionaries, and even more advanced systems like text-to-speech or dictation are readily available for several languages. However, the implementation of Human Language Technologies (HLT) requires significant resources, which have only been accumulated for a very small number of the 6,900 languages in the world. Among HLT, we are particularly interested in Automatic Speech Recognition (ASR). We are interested in new techniques and tools for rapid portability of speech recognition systems when only limited resources are available. Resource sparse languages are typically spoken in developing countries, but can nevertheless have many speakers. In this paper, we investigate Vietnamese and Khmer languages, which are spoken by about 70 million people in Vietnam and 13 million people in Cambodia, but for which only very few usable electronic resources are available.

In this paper, we present different strategies of acoustic modeling for under-resourced languages. We start in section 2 by proposing different techniques in crosslingual acoustic modeling. When there is no pronunciation dictionary available in target language, we investigate, in section 3, some techniques of grapheme-based acoustic modeling. The experimental framework and some comparative results for Vietnamese and Khmer ASR are presented in section 4. Section 5 concludes the work and gives some future perspectives.

2. Crosslingual Acoustic Modeling
The research in crosslingual acoustic modeling is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language [1]. In crosslingual acoustic modeling, previous approaches have been limited to context-independent (CI) models [1, 2, 3]. Monophonic acoustic models in target language were initialized using seed models from source language. Then, these initial models could be rebuilt or adapted using training data from the target language. Since the recognition performance is increased significantly in wider contexts, the crosslingual context-dependent (CD) acoustic modeling can be investigated. A triphone similarity estimation method based on phoneme distances was first proposed in [4] and used an agglomerative clustering process to define a multilingual set of triphones. T. Schultz [1] proposed PDTS method to overcome the problem of context mismatch in portability of CD acoustic models.

We have already proposed in [5] some methods for estimating similarities between acoustic-phonetic units (phonemes, polyphones, clustered polyphones). Using these similarity measures, we propose in this section two crosslingual acoustic schemes in which the similarities between two models (monophonic or polyphonic) can be determined by phoneme similarity or clustered polyphone similarity.

2.1. Crosslingual CI Acoustic Modeling
For CI acoustic modeling, the phonetic unit is the monophone and a similarity between monophonetic models in source and target language is calculated. Let $\Phi_S$ and $\Phi_T$ be monophonetic models in source and target language. The similarity between $\Phi_S$ and $\Phi_T$ is calculated by:

$$d(\Phi_S, \Phi_T) = d(s, t)$$

where $d(s, t)$ is phoneme similarity which can be calculated manually based on the IPA phoneme classification or automatically based on a confusion matrix [5].

For each monophonetic model in the target language, the nearest monophonetic model $\Phi_T^*$ in source language is obtained if it satisfies the following relation:

$$\forall \Phi_S, d(\Phi_S, \Phi_T^*) = \min[d(\Phi_S, \Phi_T)] = \min[d(s, t)]$$

By applying equation (2), a phoneme mapping table between source and language can be obtained. Based on this mapping table, the acoustic models in the target language can be borrowed from the source language and adapted by a small amount of target language speech data.

2.2. Crosslingual CD Acoustic Modeling
In this section, a CD acoustic model portability method is proposed based on the phonetic similarities described in [5].

Firstly, by using a small amount of speech data in the target language, a decision tree for polyphone clustering (PTF) can be built. We suppose that such a decision tree ($PS_T$) is also available in the source language (figure 1).
where the similarity between a clustered polyphonic model of \( n \) polyphones in the target language, the previous works [6, 7, 8] for well resourced languages. Firstly, In our work, we investigate some techniques in grapheme-based models of \( n \) polyphones in the source language and influence of contextual phonemes to the central phoneme and sub-word units. The selected unit is usually phoneme.

Traditionally, ASR systems represent lexical units in terms of Traditionally, ASR systems represent lexical units in terms of graphemes, some previous works concluded that seed models perform better than random or flat starts [10]. In fact, by using the seed

\[ d(\Phi_S, \Phi_T) = \sum_{i=1}^{n} d(P_S, P_T) \]

where \( d(P_S, P_T) \) is the contextual similarity between polyphones:

\[ d(P_S, P_T) = \sum_{i} \left( d_s(P_S, P_T) \right) \]

In (4), \( \omega_s \) is contextual weight coefficient which represents the influence of contextual phonemes to the central phoneme and \( d_s(t_k, t_{k+1}) \) is the phoneme similarity (\( k = -L, \ldots, L \)).

For each clustered polyphonic model in the target language, the nearest clustered polyphonic model \( \Phi_T^v \) in source language is obtained if it satisfies the following relation:

\[ \Phi_S \Rightarrow \Phi_T^v \]

This nearest clustered polyphonic model is then copied into the correspondent model in the target language.

Finally, while acoustic models borrowed directly from the source language do not perform very well, an adaptation procedure (MLLR, MAP, ...) can be applied with a small amount of speech data in the target language. We will compare these crosslingual techniques in the experimentation section.

3. **Grapheme-Based Acoustic Modeling**

Traditionally, ASR systems represent lexical units in terms of sub-word units. The selected unit is usually phoneme (monophone or polyphone). Thus, the performance of the ASR systems depends on the quality of the pronunciation dictionary.

However, for under-resourced languages, the design of a pronunciation dictionary may be a problem because of the following reasons:

- No expert knowledge of the target language may be available (non native developers, badly described languages, ...).
- Handcrafting process is time and cost consuming.

In our work, we investigate some techniques in grapheme-based acoustic modeling. This modeling approach was already used in previous works [6, 7, 8] for well resourced languages. Firstly, since several under-resourced languages are script languages, a character Romanization process is generally needed to build a pronunciation dictionary. Then, a word boundary detector is used to initialize the grapheme-based CI acoustic models. For CD modeling, we also propose and compare different techniques of question generation for decision trees.

**3.1. Pronunciation dictionary creation**

For Latin alphabet languages like English, French, Vietnamese, grapheme-based pronunciation dictionary is built by simply splitting a word into its graphemes. Table 1 presents a sample of a grapheme-based pronunciation dictionary for Vietnamese.

<table>
<thead>
<tr>
<th>Word</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>a dua</td>
<td>dân</td>
</tr>
<tr>
<td>am tương</td>
<td>gã</td>
</tr>
<tr>
<td>ban trụ</td>
<td>bút pháp</td>
</tr>
<tr>
<td>bối thủy</td>
<td>bối thủy</td>
</tr>
<tr>
<td>chuyên gia</td>
<td>chuyể nga</td>
</tr>
</tbody>
</table>

Table 1. Grapheme-based pronunciation dictionary for Vietnamese

However, for other script writing systems like Chinese, Korean, Arabic, Thai, Khmer, the creation procedure needs a character Romanization extra step. In our work, since Khmer language makes use of an alphabetic writing system (Khmer alphabet), the Romanization step contains a character conversion by using the Unicode Character Name Table. Some dictionary entries for Khmer are presented in table 2.

<table>
<thead>
<tr>
<th>Word</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ka</td>
<td>Ka I Ca Ka U Ca</td>
</tr>
<tr>
<td>Ka</td>
<td>Ka I Lo</td>
</tr>
<tr>
<td>Ts</td>
<td>Ts OO Ngo Ka Ro Po AA Ta</td>
</tr>
<tr>
<td>Ts</td>
<td>Ts OO Ngo To AA Mo</td>
</tr>
<tr>
<td>Bu I N No Do Ba AA Ta Ca Ca Ro I Ka Vo Ta Ta</td>
<td></td>
</tr>
<tr>
<td>Bu I N No Do Ba AA Ta To AA No</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Grapheme-based pronunciation dictionary for Khmer

**3.2. Initialize the acoustic models**

Since there are no labeled training data for graphemes, some alternative initialization strategies must be used to initialize the acoustic models: random start, flat start or uniform segmentation, etc. With flat start, we can make all models equal initially. With uniform segmentation, acoustic models are started by uniformly segmenting the speech data and associating each successive segment with successive states (like in HTK toolkit [9]). Figure 2 presents a uniform segmentation of speech data to initialize the grapheme-based models.

![Uniform segmentation of speech data](http://www.unicode.org/charts/PDF/U1780.pdf)
models, we can provide sub-word unit transcriptions of speech data by an automatic time alignment procedure. In phoneme-based acoustic modeling, seed models in target language can be borrowed from other languages (called crosslingual acoustic models). In grapheme-based acoustic modeling, the use of seed models borrowed from a multilingual grapheme-based system can speed up the bootstrapping procedure in comparison with flat start but the performances of two methods are similar [7]. This is due to the poor sharing of graphemes across languages.

In our work, we investigate another initialization strategy of acoustic models. Firstly, we use a word boundary detector to decode the lower and upper boundary of every word in the acoustic models. Firstly, we use a word boundary detector to investigate and compared in our experiments: techniques for each language. Thus, two techniques are comparative experiments help us to find the appropriate techniques in context of under-resourced languages are. These paper, we try to see how appropriate these question generation these techniques depend on each language. Therefore, in this

3.3. Grapheme-Based CD Acoustic Modeling

Obviously, the grapheme is not an appropriate unit in acoustic modeling and its pronunciation strongly depends on its writing context. Thus, some CD acoustic modeling techniques must be investigated to improve the accuracy of grapheme-based acoustic models. In fact, graphemic questions for decision tree state tying can be built manually or automatically by borrowing some techniques from phoneme-based acoustic modeling. Some previous works compared some question generation techniques in the grapheme-based systems [6, 7] but the performance of these techniques depend on each language. Therefore, in this paper, we try to see how appropriate these question generation techniques in context of under-resourced languages are. These comparative experiments help us to find the appropriate techniques for each language. Thus, two techniques are investigated and compared in our experiments:

- **Singleton**: each graphemic question contains a grapheme.
- **Grapheme-phoneme conversion**: a grapheme is assigned to a phonetic question if the grapheme is part of the phoneme.

Table 3 presents a graphemic conversion of some phonetic questions in Vietnamese language.

<table>
<thead>
<tr>
<th>Phonetic questions</th>
<th>Phoneme</th>
<th>Grapheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALVEOLAR</td>
<td>tʰ d n z i</td>
<td>tʰ d n i z r</td>
</tr>
<tr>
<td>VELAR</td>
<td>k ŋ y</td>
<td>k ŋ</td>
</tr>
<tr>
<td>FRICATIVE</td>
<td>f v s z y h</td>
<td>v s d r k g h</td>
</tr>
<tr>
<td>APPROXIMANT</td>
<td>w j</td>
<td>w o i y</td>
</tr>
<tr>
<td>FRONT</td>
<td>i e t ə a</td>
<td>y i e e a</td>
</tr>
<tr>
<td>BACK</td>
<td>w v ə ə s</td>
<td>w u o s o o</td>
</tr>
<tr>
<td>CLOSE</td>
<td>i u u</td>
<td>y ɨ u</td>
</tr>
</tbody>
</table>

Table 3. Grapheme-phoneme conversion
In addition, we also compare performances of phoneme-based (VN-CD1000) and grapheme-based (VN-CD1000-GP) approaches for Vietnamese in figure 4. Although grapheme-based approach is slightly outperformed by phoneme-based approach, the grapheme-based approach shows a good potential when no pronunciation lexicon is available, we investigate a grapheme-based ASR system. Firstly, we compare two initialization strategies of grapheme-based CI acoustic models: uniform segmentation (baseline) strategy and word boundary initialization strategies of grapheme-based ASR system. Secondly, we investigated some techniques of grapheme-based acoustic modeling. To improve the performance of the graphemic acoustic models initialization, we used a word boundary detector to segment an utterance into words. This technique reduced some inter-word segmentation mistakes. Moreover, results obtained both from Vietnamese and Khmer ASR demonstrated the feasibility of the grapheme-based approach. In the future, we will investigate word-based ASR systems for Vietnamese to obtain the most likely recognition unit in Vietnamese language. We will collect more text and speech data and build a phonetic dictionary for Khmer language in order to compare a phone-based approach with the grapheme-based approach presented here.

5. Conclusions

This paper presented different techniques of acoustic modeling for under-resourced languages: crosslingual and grapheme-based acoustic modeling. Firstly, we presented the potential of crosslingual independent and dependent acoustic modeling for Vietnamese language. Experimental results on Vietnamese ASR showed that when we have only a few hours of speech data in target language, crosslingual CI modeling works better. However, when we have more speech data, crosslingual CI modeling is outperformed by crosslingual CD modeling. We can also conclude that in both cases, crosslingual systems are better than monolingual baseline systems. Secondly, we investigated some techniques of grapheme-based acoustic modeling. To improve the performance of the graphemic acoustic models initialization, we used a word boundary detector to segment an utterance into words. This technique reduced some inter-word segmentation mistakes. Moreover, results obtained both from Vietnamese and Khmer ASR demonstrated the feasibility of the grapheme-based approach. In the future, we will investigate word-based ASR systems for Vietnamese to obtain the most likely recognition unit in Vietnamese language. We will collect more text and speech data and build a phonetic dictionary for Khmer language in order to compare a phone-based approach with the grapheme-based approach presented here.

6. References