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 monophonic models in source and target language is calculated. Let Φ_S and Φ_T be monophonic models in source and target language. The similarity between Φ_S and Φ_T is calculated by:

$$d(\Phi_S, \Phi_T) = d(s, t) \quad (1)$$

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 monophonic model in the target language, the nearest monophone model Φ_S^* in source language is obtained if it satisfies the following relation:

$$\forall \Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\,[d(\Phi_S, \Phi_T)] = \min[d(s, t)] \quad (2)$$

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 (PT_T) can be built. We suppose that such a decision tree (PS_S) is also available in the source language (figure 1).

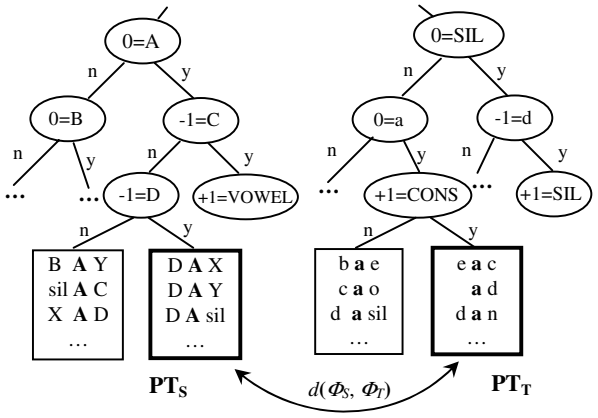

Figure 1 : *Clustered polyphone similarity across languages*

Let $\Phi_S = (P_{S1}, \ldots, P_{Sm})$ be a clustered polyphonic model of m polyphones in the source language and $\Phi_T = (P_{T1}, \ldots, P_{Tn})$ be a clustered polyphonic model of n polyphones in the target language, the similarity between Φ_S and Φ_T is calculated by:

$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m.n} \quad (3)$$

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

$$d(P_S, P_T) = \propto_0 .d(s_0, t_0) + \propto_1 .[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \propto_L .[d(s_{-L}, t_{-L}) + d(s_L, t_L)] \quad (4)$$

In (4), $\propto_k$ is contextual weight coefficient which represents the influence of contextual phonemes to the central phoneme and $d(s_k, t_k)$ is the phoneme similarity (k= -*L*,…*L*).

For each clustered polyphonic model in the target language, the nearest clustered polyphonic model Φ_S* in source language is obtained if it satisfies the following relation:

$$\forall \Phi_S, d(\Phi_S^*, \Phi_T) = \min [d(\Phi_S, \Phi_T)] \quad (5)$$

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.

Word	Pronunciation	Word	Pronunciation
a dua	a d u a	dần	d â n
am tường	a m t ư ơ n g	gả	g a
ban trưa	b a n t r ư a	giã	g i a
bút pháp	b u t p h a p	hồng quân	h ô n g q u â n
bơi thuyền	b ơ i t h u y ê n	quạng	q u a n g
chuyên gia	c h u y ê n g i a	rập khuôn	r â p k h u ô n

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[1]. Some dictionary entries for Khmer are presented in table 2.

Khmer Word	Pronunciation
កកិចកកុច	Ka Ka I Ca Ka Ka U Ca
កកិល	Ka Ka I Lo
គោងក្រពាត់	Ta OO NGo Ka Ro Po AA Ta
គោងទាម	Ta OO NGo To AA Mo
បិណ្ឌបាតចារិកវត្ត	Ba I NNo Do Ba AA Ta Ca AA Ro I Ka Vo Ta Ta
បិណ្ឌបាតទាន	Ba I NNo Do Ba AA Ta To AA No

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.

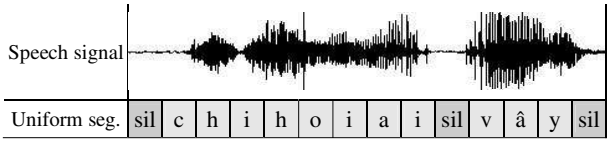

Figure 2. *Uniform segmentation of speech data*

Some previous works concluded that seed models perform better than random or flat starts [10]. In fact, by using the seed

[1] 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 utterance. Then, for each word, we uniformly segment speech data to every grapheme of the word. Figure 3 shows an example of speech data segmentation using word boundary detection.

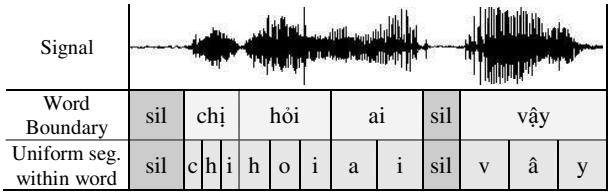

Figure 3. *Segmentation of speech data using word boundary detection*

Accuracy of proposed strategy and uniform segmentation strategy will be compared in the experimentation section.

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.

Phonetic questions	Phoneme	Grapheme
ALVEOLAR	t tʰ d n s z l	t d đ n x l s r
VELAR	k χ ŋ ɣ	c k g
FRICATIVE	f v s z ʂ ʐ χ ɣ h	v s x d r k g h
APPROXIMANT	w j	u o i y
FRONT	i e ɛ ĕ a ă	y i ê e a ă
BACK	ɯ u ɤ ŏ o ɔ ŏ	uw u ơ â oo o
CLOSE	i ɯ u	y i ɯ u

Table 3. *Grapheme-phoneme conversion*

4. Experiments

4.1. Experimental framework

All recognition experiments use the JANUS toolkit [11] developed at the ISL Laboratories. The model topology is a 3-state left-to-right HMM with 48 Gaussian mixtures per state. The pre-processing of the system consists of extracting a 43 dimensional feature vector every 16 ms. The features consist of 13 MFCCs, energy, the first and second derivatives, and zero-crossing rate. An LDA transformation is used to reduce the feature vector dimensionality to 32.

For language modeling (LM), since Vietnamese is a syllable-based writing system, a vocabulary of 6,492 Vietnamese syllables is collected. A vocabulary of 16,000 words is also obtained for Khmer. Then, for building a text corpus, documents were gathered from Internet and filtered. After data preparation, a text corpus of 868 MB for Vietnamese and 97 MB for Khmer are collected, respectively. Since Khmer language is a non-segmented language, a dictionary-based word segmentation tool is needed to segment a text sentence into words. The preliminary results obtained show 0.8% of segmented word error and 4.0% of segmented sentence error. Then, a syllable-based statistical trigram LM for Vietnamese and a word-based statistical trigram LM for Khmer are estimated from these text corpora using Katz backoff with Good-Turing discounting. It is important to note that in these LMs, the unknown words are removed since we are in the framework of closed-vocabulary models. The perplexity value evaluated on our speech test corpus is *109* for Vietnamese SLM and *84* for Khmer SLM.

For Vietnamese acoustic modeling, in order to build a polyphonic decision tree and to adapt the crosslingual acoustic models, 13 hours of speech data spoken by 36 speakers were used. The test set contains 400 utterances spoken by 3 speakers different from the training speakers. For Khmer acoustic modeling, we collected 3 hours of data spoken by 10 speakers. The training corpus contains 165 minutes and the test corpus contains 200 sentences spoken by all of 10 speakers.

4.2. Experimental results

For crosslingual experiments, we use multilingual context-independent models (MM7-CI) and context-dependent models (MM6-CD with 12,000 sub-quinphone models) developed by ISL Laboratories [1]. After the crosslingual transfer procedure, initial models were adapted with 2.25 hours (7 speakers) and 14 hours (36 speakers) of Vietnamese speech data. Figure 4 presents the syllable accuracies of crosslingual models with different amount of adaptation data. We note that VN-CI and VN-CD1000 are baseline systems (no use of crosslingual information for bootstrapping process) which correspond to CI and CD models with 1000 subtriphones. Similarly, MM7/VN-CI and MM6/VN-CD1000 are crosslingual CI and CD models. We find that when only 2-3 hours of data is available in target language, crosslingual CI models outperform crosslingual CD models but when we have more data (10-15 hours), crosslingual CD models are better. Anyway in both cases, the use of crosslingual approaches to bootstrap the systems outperforms the baseline. It is of course more clear when only a small amount of data is available (2.25h).

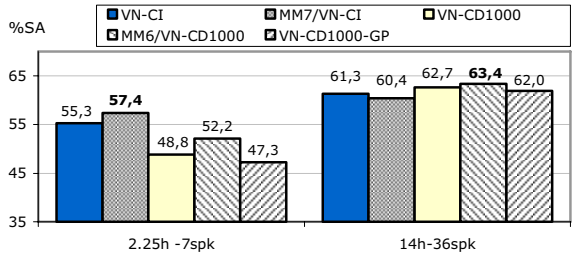

Figure 4. *Comparison of acoustic modeling techniques with different amount of adaptation data for Vietnamese ASR*

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. For Khmer ASR, since there is no phonetic dictionary 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 detection strategy. Performances (word accuracy) of two strategies are tested after each of 7 iterations of bootstrapping and presented in figure 5. The word boundary detection strategy significantly outperforms the baseline strategy in 5 first iterations and continues to be slightly better in the last iterations. We concluded that the word boundary detection can be efficiently applied to initialize grapheme-based models.

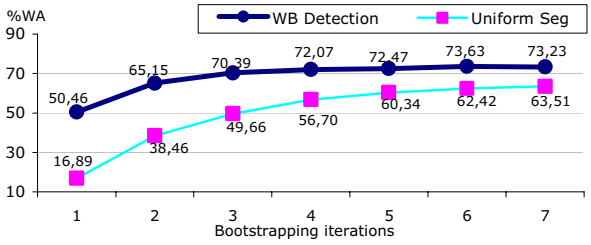

Figure 5. *Comparison of initialization strategies for grapheme-based acoustic modeling for Khmer ASR*

Then, from CI acoustic models, we continue to build CD triphone models by a decision tree based clustering procedure with different techniques of question generation. Performance of acoustic models for Khmer is shown in figure 6. The singleton questions are slightly better in 500 and 1500 subtriphones models but they are outperformed by grapheme-phoneme questions in 1000 subtriphones models.

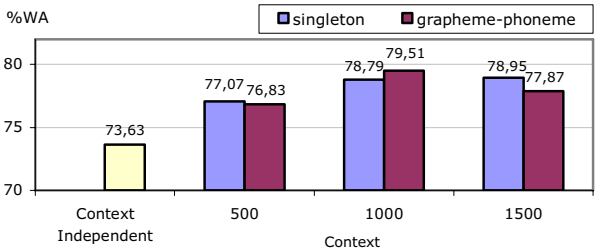

Figure 6. *Comparison of question generation techniques in grapheme-based CD modeling for Khmer ASR*

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

[1] Schultz, T., Waibel, A., "Language independent and language adaptive acoustic modeling for speech recognition", Speech Communication, vol. 35, no. 1-2, pp. 31-51, August 2001.

[2] Beyerlein, P., et al., "Towards language independent acoustic modeling", ASRU'99, Keystone, CO, 1999.

[3] Le, V-B., Besacier, L., "First steps in fast acoustic modeling for a new target language: application to Vietnamese", ICASSP'05, vol. 1, pp. 821-824, Philadelphia, PA, USA, March 2005.

[4] Imperl, B., et al., "Agglomerative vs. Tree-based clustering for the definition of multilingual set of triphones", ICASSP'00, vol. 3, Istanbul, Turkey, 2000.

[5] Le, V-B., Besacier, L., Schultz, T., "Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability", ICASSP'06, Toulouse, France, May 2006.

[6] Kanthak, S. and Ney, H., "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition", ICASSP'02, pp. 845-848, Orlando, 2002.

[7] Killer, M., Stüker, S., Schultz, T., "Grapheme Based Speech Recognition", Eurospeech'03, pp. 3141-3144, Geneva, Switzerland, September 2003

[8] Abdou, S. et al., "The 2004 BBN Levantine Arabic and Mandarin CTS Transcription Systems", RT-04 Workshop, Palisades, NY, USA, 2004.

[9] Young, S., et al., "The HTK Book", Cambridge University Engineering Department, 2002.

[10] Wheatley, B., et al., "An evaluation of cross-language adaptation for rapid HMM development in a new language", ICASSP'94, pp. 237-240, Australia, 1994.

[11] M. Finke et al., "The Karlsruhe-Verbmobil Speech Recognition Engine", ICASSP'97, vol. 1, pp. 83-86, Munich, Germany, 1997.