AUTOMATIC SPEECH RECOGNITION FOR UNDER-RESOURCED LANGUAGES

Laurent BESACIER, Viet-Bac LE

Laboratoire d'Informatique de Grenoble, France

Laurent.Besacier@imag.fr

This paper presents our methodology for ASR in the context of under-resourced languages. Our data collection methodology is explained. Then, different techniques for bootstrapping acoustic models are presented: cross-lingual and grapheme-based acoustic modelling. Firstly, we present the potential of cross-lingual independent and dependent acoustic modelling for Vietnamese language. Experimental results on Vietnamese ASR show that when we have only a few hours of speech data in the target language, cross-lingual context-independent (CI) modelling works better. However, when we have more speech data, cross-lingual CI modelling is outperformed by cross-lingual context-dependent (CD) modeling. We also conclude that, in both cases, cross-lingual systems are better than monolingual baseline systems. We also investigate some techniques of grapheme-based acoustic modeling. To improve the performance of the graphemic acoustic models initialization, we use a word boundary detector to segment an utterance into words. This technique eliminates some inter-word segmentation errors. Moreover, results obtained both from Vietnamese and Khmer ASR demonstrated the feasibility of the grapheme-based approach. Finally, we also present preliminary experiments in statistical language modelling for reducing the complexity of the models using sub-word units. The potential of such an approach is shown for dialectal Arabic where very few text data are available for training a statistical language model.

Key words: Speech recognition, under-resourced languages, acoustic model bootstrapping, sub-word units for language modelling.

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, there are more than 6000 languages in the world and only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned with languages for which large resources are available or that have suddenly become of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities have been less worked on in the past years. One way of improving this "language divide" is to do more research on portability of HLT for multilingual applications.

Among HLT, we are particularly interested in Automatic Speech Recognition (ASR). Therefore, we are interested in new techniques and tools for rapid development of ASR systems for under-resourced languages or π-languages when only limited resources are available. These languages are typically spoken in developing countries, but can nevertheless have many speakers. In this work, we investigate especially Vietnamese and Khmer, which are respectively spoken by 67 million and 13 million people, but for which speech processing services do not exist at all. In fact, measuring the availability of these services for a given language allows one to define its "computerization level". An example of such a metric is presented in [4]: a list of services is evaluated, for a given language, by an expert and a mean score is calculated (marks for each service are weighted by the "criticity" or importance of the service). Table 1 gives an example of this metric applied to Khmer, a language mainly spoken in Cambodia (6.2/20). The same metric evaluated for Vietnamese (table 2) gives 10/20. In [4], an under-resourced language (or π -language) is thus defined as a language that has a score below 10/20.

From these tables, we can note that for both languages, speech processing services do not exist at all. The reason is mainly that for developing such systems, a large amount of resources is needed (text,

transcribed speech corpora, phonetic dictionaries). Such resources are not available for languages like Vietnamese and Khmer. One may also face other problems like the absence of linguistic or phonetic descriptions, few standards (character coding, IPA, etc.), languages mainly spoken (and not really written).

Table 1. Computerization level for the Khmer language

	Services / resources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
Text processing				
	Basic input	10	16	160
	Visualization / printing	10	14	140
	Searching and replacing	8	12	96
	Text selection	6	12	72
	Lexicographical sort	5	0	0
	Spelling correction	2	0	0
Speech processing				
	Text-To-Speech	5	0	0
	Automatic Speech Recognition	5	0	0
Translation				
	Tools for automatic translation	8	4	32
OCR				
	Optical Character Recognition	9	0	0
Resources				
	Bilingual dictionary	10	4	40
	Usability dictionary	10	0	0
Total		88		540
Average				**6,2 / 20**

Table 2. Computerization level for the Vietnamese language

	Services / resources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
Text processing				
	Basic input	10	16	160
	Visualization / printing	10	16	160
	Searching and replacing	8	17	136
	Text selection	6	17	102
	Lexicographical sort	5	6	30
	Spelling correction	2	6	12
Speech processing				
	Text-To-Speech	5	0	0
	Automatic Speech Recognition	5	0	0
Translation				
	Tools for automatic translation	8	6	48
OCR				
	Optical Character Recognition	9	12	108
Resources				
	Bilingual dictionary	10	13	130
	Usability dictionary	10	0	0
Total		88		886
Average				**10 / 20**

This paper presents an overview of our activities concerning ASR for under-resourced languages. We have thus developed a methodology and tools to collect, process and model linguistic and acoustic resources in order to quickly develop ASR systems for new under-resourced languages.

Firstly, our methodology for rapid text and speech corpora acquisition for under-resourced languages is proposed in section 2. For fast developing of text processing tools for a new under-resourced language, an open source generic toolkit named CLIPS-Text-Tk was developed in our work.

For acoustic modeling, we address particularly in section 3 the usage of acoustic-phonetic unit similarities for multilingual acoustic models portability to new languages. Especially, an estimation method of the similarity between two phonemes is first proposed. Based on these phoneme similarities, some

estimation methods for polyphone similarity and clustered polyphonic model similarity are investigated. For a new language, a source/target acoustic-phonetic unit mapping table can be constructed with these similarity measures. Then, clustered models in the target language are duplicated from the nearest clustered models in the source language and adapted with limited data to the target language. Results obtained for Vietnamese demonstrate the efficiency of these methods to bootstrap acoustic models.

In section 4, we describe our attempts to reduce statistical language model complexity, which is particularly needed when few text data are available for a given language. A sub-word modeling approach is proposed and preliminary results are presented in this section. Finally, section 5 concludes this work.

2. COLLECTING RESOURCES

Nowadays, almost all of the techniques and methods in ASR systems, in particular the Large Vocabulary Continuous Speech Recognition (LVCSR) systems, use statistical approaches. However, given the statistical nature of these methods, a large amount of resources (vocabularies, text corpora, transcribed speech corpora, phonetic dictionaries) is crucial and required to train models (acoustic models, language models) and to test the performances of the systems. Consequently, a large speech corpus which contains hours of signals recorded by hundreds speakers (for acoustic modeling) and a text corpus with millions of words (for language modeling) is currently necessary for building an ASR system for a new language.

However, these crucial resources are not directly available for under-resourced languages. Thus, a methodology for rapidly building them is necessary.

2.1. Collecting text

Concerning text resources, a new methodology for fast text corpora acquisition for π-languages was already proposed and applied in [11]. Documents are gathered from Internet by web robots. Then, these web pages are filtered and analyzed for building a text corpus. An open source toolkit is currently proposed by our laboratory[1]. It is important to note that text corpora for language modeling cannot be collected easily for under-resourced languages for the following reasons:

- There are less pages and websites than for well-resourced languages.
- The debit of communication is often lower (sometimes only several kilobits per second).

Table 3 illustrates this problem by comparing different figures between a well-resourced language (French) and under-resourced languages: number of Internet hosts, number of Internet users and Internet connectivity bandwidth.

Table 3. Problems when using the web to collect text data for under-resourced languages

Language	Number of internet hosts (cf. INTERNIC[2] 2001)	Number of internet users (cf. CIA[3], 2003)	Internet bandwidth [UIT-ICS 2001]
French	325 103	21 900 000	191 898 Mbps
Vietnamese	9 037	3 500 000	34 Mbps
Khmer	818	30 000	6 Mbps
Lao	937	15 000	2 Mbps

Consequently, we cannot crawl all of the websites but we must focus on some which have more pages and higher debit than the others. So, a non-negligible time is needed to find out the websites to collect. For

[1] http://www-clips.imag.fr/geod/User/viet-bac.le/outils/
[2] International Network International Center http://www.internic.net
[3] CIA The World Factbook: http://www.cia.gov/cia/publications/factbook/index.html

this purpose, the most interesting and easy to find web sites will be generally news websites. Most of them have archives that can be collected in order to have a larger amount of data. The drawback is that the data collected will be "newspaper like" and not necessarily representative of the spoken language. Table 4 illustrates the amount of data that can be collected from such a web site (data collected from the "Voice of America" web site[4] for different languages).

Table 4. Example of the amount of text data that can be collected on a news web site

Language	#sentences	#words	#size
Indonesian	116k	2.4M	17M
Korean	405k	7M	67M
Pashto	7k	0.2M	2M
Kurdish	24k	0.6M	8M
Hindi	73k	2M	28M
Persian (farsi)	212k	5.8M	54M

This table shows that it is possible to collect a reasonable amount of data for some languages; however, it shows also that big differences exist between under-resourced languages concerning the amount of data that can be collected with such a method.

2.2. Collecting speech

Concerning speech resources, we have first built local collaborations (MICA/Hanoi, ITC/Phnom-Penh) and performed local recordings. For the acquisition and management of speech signals during recording, we use the EMACOP system, developed in our laboratory [18]. EMACOP is a Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows environment. This system also provides verification facilities, through an integrated interface, to check signal quality and consistency between resulting signals and what was presented to the speaker during recording. EMACOP runs in a multithreading client-server architecture and meets SAM specifications in input and in output. An example of the EMACOP interface for khmer is presented in Figure 1.

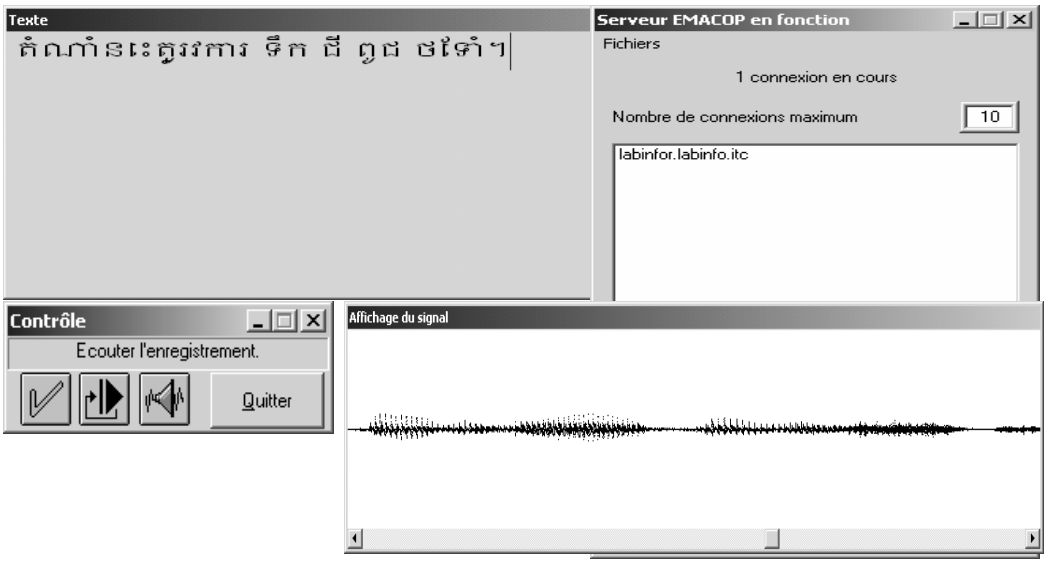

Figure 1. Example of the EMACOP interface for khmer.

[4] www.voanews.com

3. ACOUSTIC MODEL BOOTSTRAPPING

3.1. Acoustic-Phonetic Unit Similarities

The research in cross-lingual 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 [14]. Based on this assumption, we propose in this section some methods for estimating the similarities of some phonetic units (phoneme, polyphone, clustered polyphone) which will be further used in cross-lingual context dependent acoustic modeling.

Phoneme Similarity

In our work, both data-driven and knowledge-based methods are applied and proposed to automatically or manually obtain the phoneme similarities across languages.

a) Data-driven methods. The acoustic similarity between two phonemes can be obtained automatically by calculating the distance between two acoustic models (HMM distance [10], Kullback-Leibler distance, Bhattacharyya distance, Euclidean distance [16] or by calculating a confusion matrix [3, 5]). A confusion matrix is calculated by applying a source language phoneme recognizer to a small amount of target language acoustic data, which was already phonetically annotated with the target language acoustic units. Note that in the basic phoneme recognizer we use, all phonemes have the same probability to appear. Then, each entry of the confusion matrix is normalized by dividing it through the number of occurrences of all corresponding phonemes in the source language [12].

Normally, the confusion matrix represents the likelihood of the confusion between two phonemes. Thus, we can use these phoneme confusions to evaluate phoneme similarities. Let M, N be numbers of phonemes in source and target language. Let $A(M,N)$ be the confusion matrix. The similarity $d(s_i, t_j)$ between phoneme t_j in the target language and phoneme s_i in the source language is calculated as:

$$d(s_i, t_j) = A_{i,j} \tag{1}$$

where $A_{i,j} \in [0,1]$, i=1..M, j=1..N.

b) Proposed knowledge-based method. Traditionally, knowledge-based methods have been applied to find the phoneme of the source language that best matches a phoneme in the target language [5, 14]. However, no knowledge-based method is known that allows computing the similarity between two phonemes. Thus, in this section, we propose a new knowledge-based method to calculate the phoneme similarity. As we know, similarities between sounds are documented in international phonetic inventories like the International Phonetic Alphabet (IPA)[5] which classifies sounds based on phonetic knowledge.

Based on the IPA phoneme classification, we propose a *bottom-up algorithm* to determine a distance-based similarity between two phonemes. This algorithm consists of two steps: *top-down classification using a hierarchical graph* and *bottom-up phoneme distance estimation.*

Step 1: Top-down classification. Figure 2 shows a hierarchical graph where each node is classified into different layers. To each node we manually assigned a group of phonemes following the IPA phoneme classification scheme. Each group of phonemes has a user-defined similarity value assigned, which represents the similarity between the elements within this group. All nodes corresponding to the same layer obtain the same similarity value. Let k be the number of layers and G_i be the user-defined similarity value for layer i (i = 0...k-1). In our work, we investigated several

[5] http://www2.arts.gla.ac.uk/IPA/ipa.html

settings of k and G_i and set $G = \{0.9; 0.45; 0.25; 0.1; 0.0\}$ with $k = 5$ based on a cross-evaluation in cross-lingual acoustic modeling experiments.

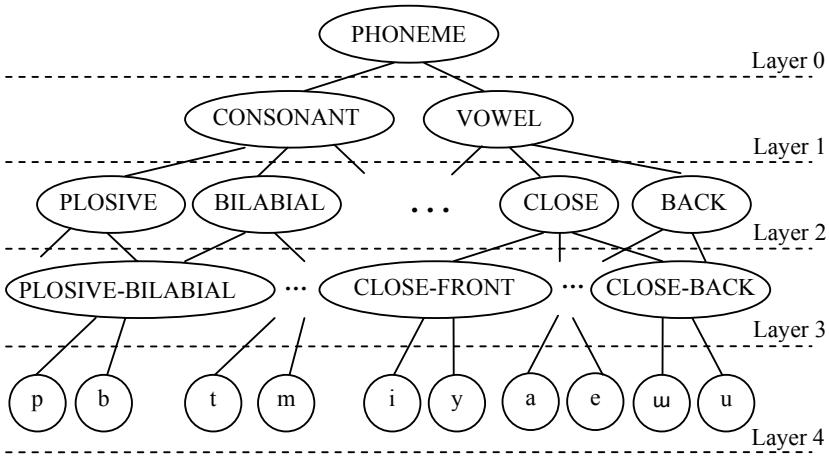

Figure 2. Hierarchical graph for phoneme similarity

To grow this graph, we start with the group PHONEME, which contains all the phonemes, at layer 0 and divide it into a CONSONANT group and VOWEL group at layer 1. This top-down classification is applied with increasingly specified grouping criteria until each group contains only one phoneme.

Step 2: Bottom-up estimation. To estimate the distance between two phoneme s and t, we locate them in the leaves of the graph and then trace back from their respective leaves until the nearest common parent node is reached. The similarity between s and t is thus given by the similarity value of layer i, which contains this parent node, we have:

$$d(s, t) = G_i \tag{2}$$

For example, the parent node of vowel [i] and [u] is CLOSE, we have:

$$d([\text{i}], [\text{u}]) = G_2 \ \ (= 0.25 \text{ in our experiment }).$$

Polyphone Similarity

Let L be the left and the right context length of a polyphone. We assume that the context length of polyphones in source and target language are the same. If not, a context normalization procedure is needed. Let S be the phoneme set in source language, T be the phoneme set in target language.

Let $P_S = (s_{-L}, s_{-L+1},\ldots, s_{-1}, \boldsymbol{s_0}, s_1,\ldots, s_L)$ and $P_T = (t_{-L}, t_{-L+1},\ldots, t_{-1}, \boldsymbol{t_0}, t_1,\ldots, t_L)$ be polyphones in source and target language, where $s_{-L},\ldots, s_{-1}, s_0, s_1,\ldots, s_L \in S$ and $t_{-L},\ldots, t_{-1}, t_0, t_1,\ldots, t_L \in T$ denote the central phoneme, left phonemes or right phonemes of P_S and P_T.

The distance-based between P_S and P_T is calculated as a weighted sum of distances between corresponding source/target phonemes along their context:

$$d(P_S, P_T) = \propto_0.d(s_0, t_0) + \propto_1.[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \propto_L.[d(s_{-L}, t_{-L})+d(s_L, t_L)] \tag{3}$$

where $\propto_0, \propto_1,\ldots \propto_L$ are contextual weight coefficients which represent the influence of contextual phoneme to the central phoneme; $d(s_k, t_k)$ is the phoneme distance ($k = -L,\ldots L$). In the same way, the triphone similarities are calculated in [7].

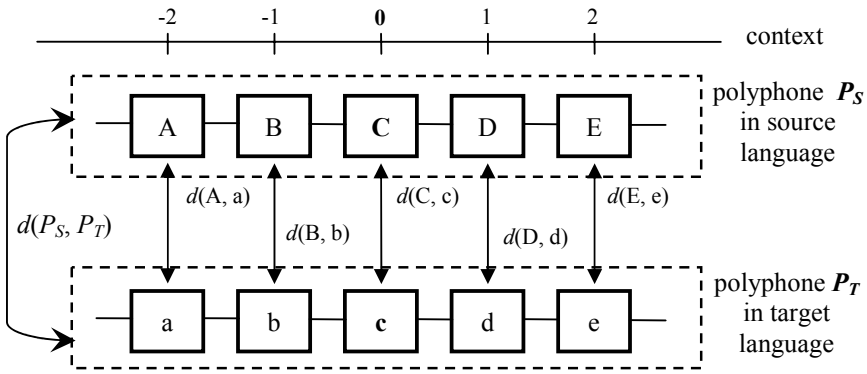

Figure 3. Example of similarity between two polyphones.

Figure 3 shows an example of the similarity between polyphone P_S = (A B **C** D E) and P_T = (a b **c** d e) in the source and target language. For each polyphone of the target language, the nearest polyphone P_S* in source language is obtained that satisfies the following relation:

$$\forall\ P_S \in S,\ d(P_S^*, P_T) = \min[d(P_S, P_T)\ ] \tag{4}$$

Clustered Polyphonic Model Similarity

Since the number of polyphones in a language is very large (e.g., over 100,000 triphones for English), a limited training corpus usually does not cover enough occurrences of every polyphone. As a consequence many polyphones in the test set have never been seen in training. Thus, we need to find models that are accurate and trainable in acoustic modeling. A decision tree-based clustering (illustrated in figure 4), or an agglomerative clustering [7] procedure is needed to cluster and model similar polyphones in a clustered polyphonic model.

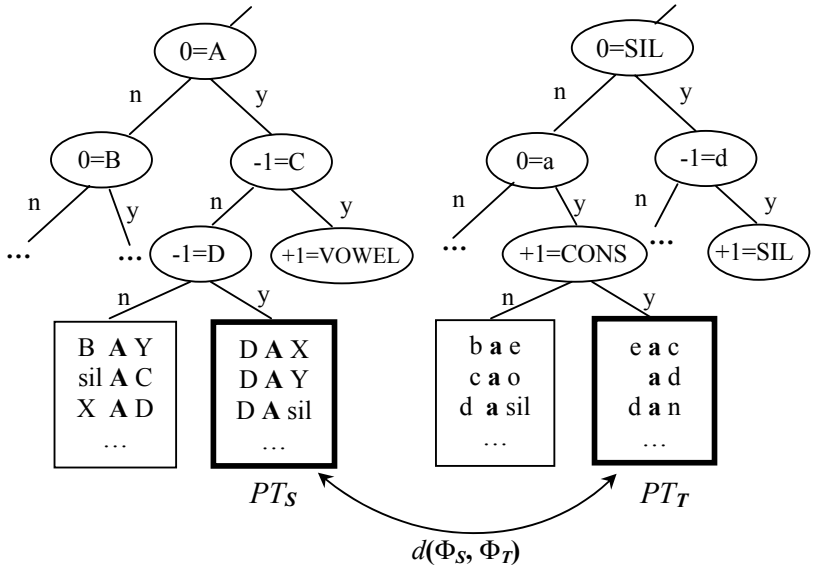

Figure 4. Clustered polyphone similarity across languages.

Therefore, for cross-lingual context dependent modeling, a clustered polyphonic model similarity evaluation method must be proposed to find two nearest clustered polyphonic models across languages (figure 4).

Let $\Phi_S = (P_{S1}, P_{S2} \ldots P_{Sm})$ be a clustered polyphonic model of m polyphones in the source language and $\Phi_T = (P_{T1}, P_{T2}, \ldots, P_{Tn})$ be a clustered polyphonic model of n polyphones in the target language, the similarity between Φ_S and Φ_T is the average of all distances between any two polyphones in Φ_S and Φ_T. We have:

$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m \cdot n} \tag{5}$$

For each clustered polyphone set in the target language, the nearest clustered polyphone set P_S* in source language is obtained if it satisfies the following relation:

$$\forall \Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)] \tag{6}$$

3.2. Crosslingual Acoustic Modeling

Context Independent Acoustic Model Portability

For context independent acoustic modeling, the phonetic unit is the monophone and a distance between monophone models is calculated. Φ_S and Φ_T are calculated using the distance between two phonemes. Equation (5) leads to:

$$d(\Phi_S, \Phi_T) = d(P_S, P_T) = d(s, t) \tag{7}$$

where $d(s, t)$ is calculated by equation (1) or (2).
Equation (6) leads to:

$$\forall \Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)\ ] = \min[d(s, t)] \tag{8}$$

By applying equation (8), 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 (see [12] for more details).

Context Dependent Acoustic Model Portability

In this section, a context dependent acoustic model portability method is proposed based on the phonetic similarities described in the previous section.

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 4).

Secondly, by applying the equation (5), we can evaluate the distance between any two source/target clustered polyphonic models. That allows us, by applying the equation (6), to determine for each model in the target language, the most similar model in the source language. This model is then copied into the acoustic model in the target language.

Finally, while acoustic models borrowed directly from the source language did not perform very well, an adaptation procedure (MLLR, MAP, etc.) can be successfully applied with a small amount of speech data in the target language (see also [14]).

3.3. 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, etc.),

– Handcrafting process is time and cost consuming.

In our work, we investigated some techniques in grapheme-based acoustic modeling. This modeling approach was already used in previous works [8], [9], [1] for well-resourced languages. It does not use any pronunciation dictionary since the words of the vocabulary are described in terms of grapheme units. Then, a word boundary detector is used to initialize the grapheme-based CI acoustic models. For CD modeling, we also propose and compare different question generation techniques for decision trees.

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 identical initially. With uniform segmentation, acoustic models are started by uniformly segmenting the speech data and associating each successive segment with successive states (like in the Hidden Markov Modeling Toolkit (HTK) [20]). Figure 5 presents a uniform segmentation of speech data to initialize the grapheme-based models.

Figure 5. Uniform segmentation of speech data.

Some previous works concluded that seed models perform better than random or flat starts [19]. In fact, by using the seed 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 cross-lingual 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 [9]. 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 6 shows an example of speech data segmentation using word[6] boundary detection.

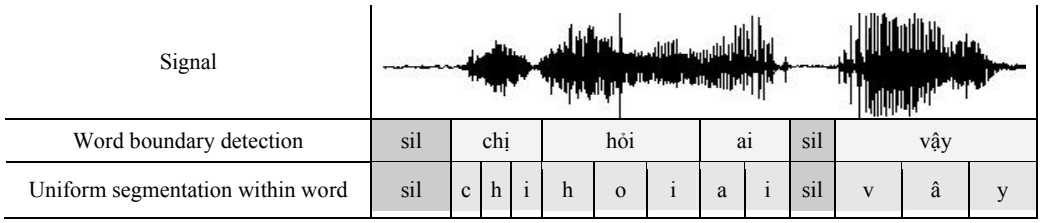

Figure 6. Segmentation of speech data using word boundary detection.

The accuracy of proposed strategy and uniform segmentation strategy will be compared in the experiments section.

[6] with is actually a *syllable* boundary detection for this example in vietnamese language

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 [8], [9] 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 5 presents a graphemic conversion of some phonetic questions in Vietnamese language.

Table 5. Grapheme-phoneme conversion

Phonetic questions	Phoneme	Grapheme
ALVEOLAR	t t^h 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 ɛ̆ a ă	y i ê e a ă
BACK	ɯ u ɤ ɤ̆ o ɔ ɔ̆	uw u ơ â oo o
CLOSE	i ɯ u	y i ư u

3.4. Experimental results

Experimental framework

All recognition experiments use the JANUS toolkit [6] 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 is collected, respectively [13]. 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 (use of SRILM toolkit [17]). 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 the same 10 speakers.

Results

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

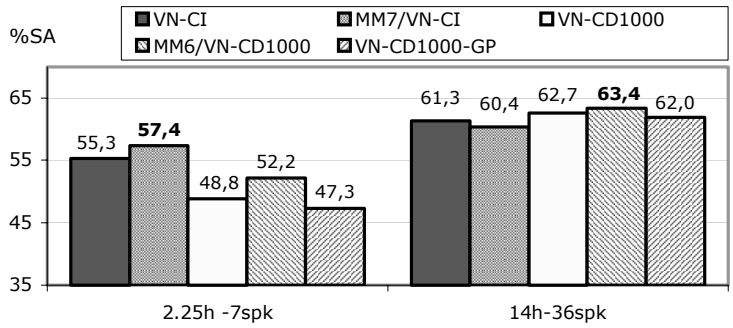

Figure 7. Comparison of acoustic modeling techniques with different amount of adaptation data for Vietnamese ASR (% syllable accuracy).

In addition, we also compare performances of phoneme-based (VN-CD1000) and grapheme-based (VN-CD1000-GP) approaches for Vietnamese in figure7. Although the grapheme-based approach is slightly outperformed by the 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 the 7 iterations of bootstrapping and presented in figure 8. The word boundary detection strategy significantly outperforms the baseline strategy during the 5 first iterations and continues to be slightly better in the last iterations. We can conclude from this result that the word boundary detection can be efficiently applied to initialize grapheme-based models.

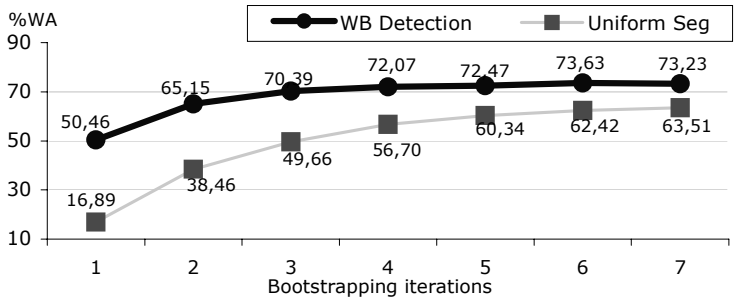

Figure 8. Comparison of initialization strategies for grapheme-based acoustic modeling for Khmer ASR (%word accuracy).

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 9. The singleton questions are slightly better in 500 and 1500 subtriphones models but they are outperformed by grapheme-phoneme questions in 1000 subtriphones models. From this result, it is hard to conclude which method is the more efficient.

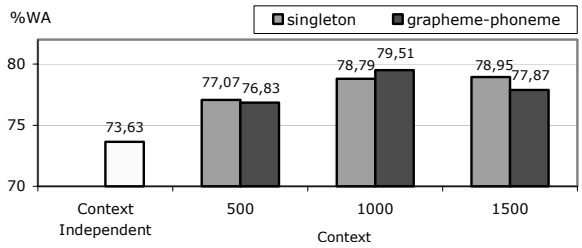

Figure 9. Comparison of question generation techniques in grapheme-based CD modeling for Khmer ASR.

4. REDUCING MODELS COMPLEXITY

4.1. Statistical language modeling for non-written or little-written languages and dialects

As already said, a frequent problem for ASR in under-resourced languages is the lack of text resources. Among the languages worldwide, a lot of them do not have a literary tradition and also many dialects are used for everyday conversations but not for written communication. Chinese and Arabic are examples of languages with a large number of unwritten regional dialects that differ significantly from the standard language [15]. In that case, the collection of resources for spoken language technology is more difficult since the only way to obtain "text" material is to record and transcribe data in that language.

Then, the problem is to accurately estimate statistical language models from such a small amount of data. This is particularly true for dialects of arabic which has a very rich morphology. For them, prefixes and suffixes augment word stems to form words. The problem is that for automatic speech recognition (ASR), a word is often defined as a string of characters separated by space. Hence, this word definition is not aware of morphological relationships between different words. In practice this leads to a high out-of-vocabulary (OOV) rate. For example, 64K word lexicons which typically lead to around a 0.5% OOV rate for English result in about 5% OOV for Arabic and hence lead to higher word error rate during recognition. The above problem is then even more pronounced for dialectal arabic due to the fact that additional prefixes, and sometimes suffixes, are informally introduced during the everyday use of language and that the amount of text data available for dialectal Arabic is usually much smaller than that for standard arabic, which will lead to poor estimates of the language model probabilities, and hence may hurt ASR performance.

The next subsection describes our first attempt to try to reduce the complexity of the models (vocabulary size, N-gram models) by segmenting LM training data into sub-word units. Here we address the use of an automatic morphological segmentation for dialectal Arabic speech recognition.

4.2. Morphological LM for ASR

We have developed a finite state transducer based automatic dialectal arabic segmenter. In this model, arabic characters, spaces and the automatically inserted prefix and suffix markers appear on the arcs of the finite state machine. The language model is conditioned to insert prefix and suffix markers based upon the frequency of their appearance in relation to the adjacent n-gram character contexts that appear in the training data (we used 5-grams). Our dialectal arabic segmenter was trained on tokenized data of about 100K words. The accuracy of this model, computed as the percentage of words in a final segmentation that is in agreement with the words provided in a manually segmented reference, is 95.4%. Then this automatic tokenizer was

used instead of a fixed set of affixes, to create the ASR lexicon. Keeping the N most frequent words unsegmented allows to keep a good LM coverage while controlling, in the mean time, the vocabulary size.

Figure 10 shows the ASR performance for a dialect of Arabic where the size of the ASR vocabulary (and thus the model complexity) is controlled using the method described above. The experimental setup is precisely described in [2] and 'seg' means no tokenization of the LM data (corresponding to a vocabulary beyond 64k).

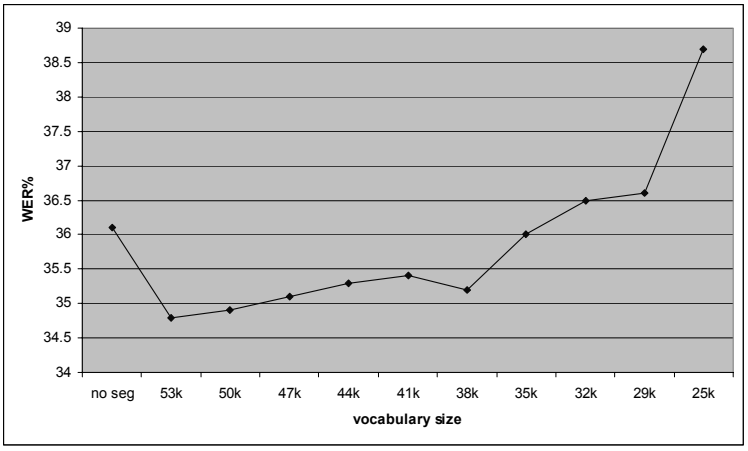

Figure 10. ASR performance (%WER) for different LM complexities (vocabulary size).

These results show that it is possible to reduce the vocabulary (and then the complexity) of an ASR system by applying a sub-word segmentation while keeping the same performance and even slightly decreasing the word error rate. This shows the potential of such a sub-word modeling approach when only sparse text resources are available for training language models. Such an approach could also be applied to other tasks like machine translation for instance.

5. CONCLUSION

This paper presented our methodology for ASR in the context of under-resourced languages. Our data collection methodology was explained. Then different techniques for bootstrapping acoustic models were presented: 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. We also 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. Finally we also presented preliminary experiments in statistical language modeling for reducing the complexity of the models using sub-word units. The potential of such an approach was shown for dialectal Arabic where very few text data was available to train a statistical language model.

REFERENCES

1. ABDOU S. *et al*., *The 2004 BBN Levantine Arabic and Mandarin CTS Transcription Systems*, RT-04 Workshop, Palisades, NY, USA, 2004.
2. AFIFY M., SARIKAYA R., KUO J. BESACIER L. and GAO Y., *On the use of morphological analysis for dialectal Arabic speech recognition*, Proc ICSLP 2006, Pittsburgh, USA, September 2006.

3. BAYEH R. *et al*., *Towards multilingual speech recognition using data driven source/target acoustical units association*, ICASSP'04, **Vol. I**, pp. 521-524, Montreal, Canada, May 2004.

4. BERMENT V. *Méthodes pour informatiser des langues et des groupes de langues peu dotées*, PhD Thesis, J. Fourier University – Grenoble I, May 2004.

5. BEYERLEIN P. *et al*., *Towards language independent acoustic modeling*, ASRU'99, Keystone, CO, USA, December 1999.

6. FINKE M. *et al*., *The Karlsruhe-Verbmobil Speech Recognition Engine*, ICASSP'97, **Vol. 1**, pp. 83-86, Munich, Germany, 1997.

7. IMPERL B. *et al*., *Agglomerative vs. Tree-based clustering for the definition of multilingual set of triphones*, ICASSP'00, **Vol. 3**, pp. 1273-1276, Istanbul, Turkey, June 2000.

8. KANTHAK S., NEY H., *Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition*, ICASSP'02, pp. 845-848, Orlando, 2002.

9. KILLER M., STUKER S., SCHULTZ T. *Grapheme based Speech Recognition*, Eurospeech'03, Geneva, Switzerland, September 2003.

10. KOHLER J., *Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds*, ICSLP'96, pp. 2195-2198, Philadelphia, PA, USA, October 1996.

11. LE V.B, BIGI B., BESACIER L., CASTELLI E., *Using the Web for fast language model construction in minority languages*, Eurospeech 2003, pp. 3117-3120, Geneva, September 2003.

12. 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.

13. LE V. B. *et al*, *Spoken and written language resources for Vietnamese*, LREC'04, pp. 509-602, Lisbon, Portugal, May 2004.

14. 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.

15. SCHULTZ T., KIRCHHOFF K., *Multilingual Speech Processing*, Academic Press, Elsevier, 2006.

16. SOOFUL J.J., BOTHA E.C., *An acoustic distance measure for automatic cross-language phoneme mapping*, PRASA'01, pp. 99-102, South Africa, November 2001.

17. STOLCKE A., *SRILM - An Extensible Language Modeling Toolkit*, ICSLP'02, **Vol. 2**, pp. 901-904, Denver, CO, USA, September 2002.

18. VAUFREYDAZ D., *et al.*, *EMACOP: Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole*, JEP'98 (Speech Processing Workshop), Martigny, Switzerland, pp. 175-178, June 1998.

19. 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.

20. YOUNG S., *et al*., *The HTK Book*, Cambridge University Engineering Department, 2002.