Recent advances in Automatic Speech Recognition for Vietnamese

Viet-Bac Le, Laurent Besacier, Sopheap Seng, Brigitte Bigi, Thi-Ngoc-Diep Do

RECENT ADVANCES IN AUTOMATIC SPEECH RECOGNITION FOR VIETNAMESE

Viet-Bac Le, Laurent Besacier*, Sopheap Seng*'**, Brigitte Bigi*, Thi-Ngoc-Diep Do*,***

* LIG Laboratory, UMR 5217, BP 53, 38041 Grenoble Cedex 9, FRANCE
** International Research Center MICA, CNRS/UMI-2954, Hanoi, VIETNAM
email: viet-bac.le@imag.fr

ABSTRACT

This paper presents our recent activities for automatic speech recognition for Vietnamese. First, our text data collection and processing methods and tools are described. For language modeling, we investigate word, sub-word and also hybrid word/sub-word models. For acoustic modeling, when only limited speech data are available for Vietnamese, we propose some crosslingual acoustic modeling techniques. Furthermore, since the use of sub-word units can reduce the high out-of-vocabulary rate and improve the lack of text resources in statistical language modeling, we propose several methods to decompose, normalize and combine word and sub-word lattices generated from different ASR systems. Experimental results evaluated on the *VnSpeechCorpus* demonstrate the feasibility of our methods.

Index Terms *– ASR, Vietnamese, word, sub-word unit, acoustic modeling, language modeling.*

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. There are, however, more than 6,900 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 by languages for which large resources are available or which have suddenly become of interest because of the economic or political scene. On the other hand, 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 improvement 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, which is spoken by about 85 million people, but for which only very few usable electronic resources are available.

We can note that for Vietnamese, 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). In the beginning of our work, such resources were not available for languages like Vietnamese. One may also face other problems like the absence of linguistic or phonetic descriptions, few standards (character coding, IPA, …)…

This paper presents an overview of our recent activities concerning ASR for Vietnamese. 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 language (particularly for Vietnamese). We start in section 2 by presenting our methods and tools for text corpora acquisition and language modeling. In section 3, we present our work in acoustic modeling for under-resourced language. We present in section 4 an integration of multiple levels (word/sub-word) in an ASR system for Vietnamese. The experimental framework and results are presented in section 5. Section 6 concludes the work and gives some future perspectives.

2. TEXT CORPUS ACQUISITION AND LANGUAGE MODELING

2.1. Methodology for text corpora acquisition

2.1.1. Data collection

As for text resources, we have proposed and applied a new methodology for fast text corpora acquisition for under-resourced languages [Le 2003]. It is based on the use of Web resources to collect textual corpora. However, it is worth noticing 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 speed of communication is often lower (sometimes only several kilobits per second).

Consequently, we can not crawl all of the websites but we must focus on some which have more pages and higher speed than the others. Therefore, a non negligible time is needed to find out manually the websites to collect. For this purpose, the most interesting and easy to collect sites will generally be 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 targeted ASR task.

It is also important to note that most Web resources contain some redundant information (references, advertisements, announcements, menus, etc.) which is repeated on different pages.

This is due to the daily collection of news which may have a direct influence on the quality of the text corpus and also on the performance of the language modeling. By filtering the redundant information contained on the web pages collected from the same site, our corpus size was reduced by 50% but a significant reduction in its perplexity was observed [Le 2003].

2.1.2. Text corpora processing

Obtaining text corpora in usable form is not a trivial task. There are many ways to process text corpora but they are difficult to achieve [Habert 1998]. The toolkit described in this paper was initially developed to address the problem of processing text corpora for the language model training. In general, for a given task, the construction of a new language model requires a task-specific corpus. That means to create some new text processing and some new copies of data. Therefore, by proposing an XML format, we will obtain a complete, clean and unified version of different text corpora which allow us to easily create a specific language model.

Thus, we proposed in our work a general XML format for text corpus and developed some tools to convert, process and normalize text corpora. The developed toolkit, called *CLIPS-Text-Tk*, as well as the proposed XML format can be also used in other applications like: statistical linguistics, information retrieval, machine translation... At this moment, this toolkit can deal with French, Vietnamese and Khmer languages and it can be easily adapted to new languages. It also allows to rapidly process a very large text corpus with several millions of documents from different sources.

The *CLIPS-Text-Tk* toolkit consists of a set of tools which are applied sequentially to the text corpora. The advantage of this modular approach is that we can develop easily and rapidly. Moreover, we can also add some new tools, even modify and remove existent tools from the toolkit. For a new task, we can inherit from general processing tools, and adapt rapidly to create specific other tools. In the same way, the portability to a new language consists of heritage of all language independent tools and rapid adaptation of other language dependent tools.

To implement this approach, the original source language was French and the target language was Vietnamese. Some other target languages like English, Khmer and Chinese are also investigating in our work. All the tools in the toolkit were developed in the Linux environment and they were written in *gawk* language. The toolkit is freely downloadable[1], under GPL license.

2.2. Language modeling

An important problem in ASR is to accurately estimate statistical language models from insufficient amount of data, particularly for languages which have a very rich morphology where prefixes and suffixes augment word stems to form words. The problem is that 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. The above problem is then even more pronounced for dialects, due to the fact that additional prefixes, and sometimes suffixes, are informally introduced during the everyday use of language. Additionally, the amount of text data available for these dialects is usually much smaller than for standard languages, which will lead to poor estimates of the language model probabilities, and hence may hurt ASR performance. In the mean time, some languages like Chinese and Vietnamese, for instance, lack word separators. Then, word language models must be estimated from an error-prone word segmentation or they have to be estimated at a sub-word level (syllable for Vietnamese and character for Chinese) with potentially bad consequences on the word coverage of the n-gram models.

What is common between these two types of languages (*rich morphology* or *without word separators*)? One answer is *the use of sub-word units for language modeling*. Some previous works using sub-word units for language modeling have recently been published for Arabic and Turkish (morphological analysis). Data-driven or fully unsupervised [Kurimo 2006] word decomposition algorithms were used like in [Abdillahi 2006, Afify 2006] as well as working on the character level for unsegmented languages like in [Denoual 2006].

One aim of our work is to investigate how these two views of the data (word and sub-word) can be advantageously combined in an ASR system. We propose to work both at the language model level (by proposing hybrid vocabularies with both word and sub-word) as well as at the ASR output level (will be presented in section 4).

At the language model level, the general idea is that from the initial sub-word vocabulary (Vietnamese syllable vocabulary for example), we progressively add N most frequent words in the sub-word vocabulary. By increasing N, we have different hybrid sub-word/word vocabularies and different trigram LMs are obtained with these vocabularies. We will show the performance of these different LMs for Vietnamese in the experimentation section.

3. 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 [Schultz 2001]. In crosslingual acoustic modeling, previous approaches have been limited to context-independent (CI) models [Beyerlein 1999, Schultz 2001]. 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 [Imperl 2000] and used an agglomerative clustering process to define a multilingual set of triphones. T. Schultz [Schultz 2001] proposed PDTS method to overcome the problem of context mismatch in portability of CD acoustic models.

We have already proposed in [Le 2006] some methods for estimating similarities between acoustic-phonetic units (phonemes, polyphones, clustered polyphones). Using these similarity measures, we present rapidly 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.

[1] www-clips.imag.fr/geod/User/brigitte.bigi/

3.1. Crosslingual Context Independent Acoustic Modeling

For context independent acoustic modeling, the phonetic unit is the monophone and a distance 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 distance between Φ_S and Φ_T is calculated using the distance between two phonemes. We have:

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

where $d(s, t)$ is phoneme distance which can be calculated manually based on the IPA phoneme classification or automatically based on a *confusion matrix* [Le 2005].

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.

3.2. Crosslingual Context Dependent Acoustic Modeling

In this section, a context dependent acoustic model portability method is proposed based on the phonetic similarities described in [Le 2006].

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).

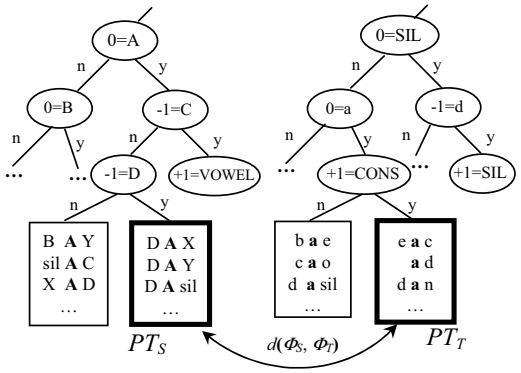

Figure 1: Clustered polyphone similarity across languages

Let $\Phi_S = (P_{S1}, ..., P_{Sm})$ be a clustered polyphonic model of m polyphones in the source language and $\Phi_T = (P_{T1}, ..., P_{Tn})$ 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:

$$d(P_S, P_T) = \alpha_0.d(s_0, t_0) + \alpha_1.[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \alpha_L.[d(s_{-L}, t_{-L}) + d(s_L, t_L)] \quad (4)$$

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.

4. COMBINATION OF WORD AND SUB-WORD UNITS IN ASR SYSTEM

As already said in section 2, the aim of this section is to investigate how word and sub-word can be advantageously combined in an ASR system. In fact, combining word graphs with sub-word graphs implies a correct way to decompose a word graph into its sub-word version, which is also proposed in this paper. Furthermore, since some previous works have shown the advantage of explicit WER minimization approach in a word lattice [Mangu 2000], we used confusion network (CN) in our work to decode the consensus hypothesis.

4.1. Word decomposition in the lattice

To deal with a language with a rich morphology or without explicit word separators, the use of classical word units in ASR and MT can be replaced by sub-word units like morphemes (case of Arabic) [Afify 2006, Besacier 2007] or syllables (case of Vietnamese). Such decomposition can reduce the high OOV rate and improve the lack of text resources in statistical language modeling. If a sub-word segmenter is already available, applying such decomposition is obvious on word strings (text corpora, *N*-best list). It is however more problematic when such decomposition must be applied to a word lattice at the output of an ASR system. The problem can be formulated as following: *how the word lattice should be modified when words are segmented into sub-word units?*

We propose in our work a new algorithm for splitting a word into a sequence of sub-words. Depending on the number of decomposed sub-words, some new nodes with sub-word labels are also inserted to the lattice. The main difference of our algorithm is that the duration and the acoustic score of each new sub-word can be looked up in a *sub-word information table*. If this kind of table is unavailable, the duration and the acoustic score may be approximately distributed as a function of the number of graphemes in each sub-word.

More precisely, the word lattice decomposition algorithm can be described with the following steps:

1. Based on a word/sub-word dictionary or a morphological analyzer, all decompoundable words in the word lattice are identified.

2. Each of these words is decomposed into a sequence of sub-words that depends on the number of sub-words in the word. Some new nodes and links are thus inserted in the word lattice.

3. By using a sub-word-based speech recognizer, a sub-word lattice is built for the same utterance. From this lattice, all sub-words with different timestamps, durations and acoustic scores are stored in a *sub-word information table*. For each new decomposed sub-word in the current word lattice, the new acoustic score and the duration is modified according to the appropriate values found in the *sub-word information table*. If such a sub-word recognizer is unavailable or the decomposed sub-words are not found in the *sub-word information table*, the duration and the acoustic score of the initial word are divided proportionally into sub-words as a function of the number of graphemes in the sub-words.

4. An approximation is made for the LM score: the LM score corresponding to the first sub-word of the decomposed word is equal to the LM score of the initial word, while we assume that after the first sub-word, there is only one path to the last sub-word of the word (so the following LM scores are made equal to 0).

In fact, a word lattice can be decomposed using the *lattice-tool* (v.1.5.2) of the SRILM toolkit [Stolcke 2002]. But with this tool, all the scores of the original word are retained on the first sub-word and the remaining sub-words get 0 scores and 0 duration (the total scores and the sentence posterior probability along the path are thus unchanged). Since the used lattice-to-CN algorithm [Mangu 2000] takes into account the duration of each word, this method might cause some wrong alignments during the converting process. This is the reason why we propose the above lattice decomposition technique.

4.2. Word and sub-word lattices combination

4.2.1. Lattice combination

In this section, the use of multiple levels of lexical units (word, morpheme, syllable …) during the ASR decoding process is proposed. By using different word and sub-word units in the lexicon, different LMs are built and different word and sub-word lattices are thus outputted by different speech recognizers. The question is *what the benefit is, if we merge these different lattices in a common lattice.*

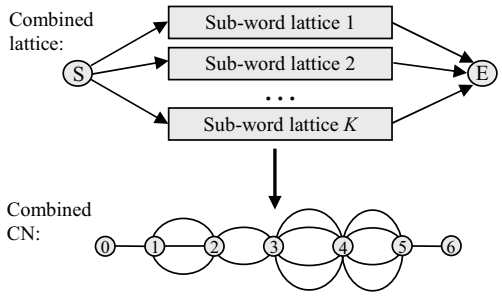

Figure 2: Combined sub-word lattice and the corresponding CN

Figure 2 presents our combination scheme which can be described with the followings steps:

1. By applying the lattice decomposition algorithm presented above, all words and sub-words in different lattices are decomposed into a unique sub-word set.

2. Create a new starting node S and a new ending node E for the common lattice. Then, we link the node S with starting nodes of all lattices and link ending nodes of all lattices with E. After this step, all lattices are merged into a common lattice.

3. The obtained lattice is then converted into CN and the consensus hypothesis can be decoded.

Another lattices combination scheme was also presented in [Li 2002] where they used an initial step (similar to step 2 of our scheme) to merge lattices together. Then, merged lattice was edited by merging similar links, building new links among nodes and renormalizing acoustic scores from different lattices. The sentence MAP hypothesis was finally decoded from this merged lattice. The difference of our combination scheme is that we do not modify the nodes and the links of the merged lattice because it is converted into CN in order to decode the consensus hypothesis.

4.2.2. Normalization of posterior probabilities

Since word and sub-word lattices are generated by different systems, a normalization step is needed. Sentence posteriors can be normalized by the sum of the sentence posteriors in the lattice:

$$P(W^k|A) = \frac{P(W^k)P(A|W^k)}{\sum_{k=1}^{N} P(W^k)P(A|W^k)} \quad (6)$$

where k ranges over the set of hypotheses outputted by the speech recognizer [Mangu 2000]. In a lattice, the total of the sentence posteriors can be computed by the Forward-Backward algorithm.

This normalization step can be used in the lattices combination scheme presented above. Before combining into a common lattice in step 2, word and sub-word lattices are decomposed and then normalized by equation (6). In next section, performances of the combination scheme with and without normalization are compared.

5. EXPERIMENTS AND RESULTS

5.1. Experimental framework

5.1.1. ASR system

All recognition experiments use the IBIS decoder of the JANUS toolkit [Soltau 2001] developed at the ISL Laboratories. The model topology is a 3- state left-to-right HMM with 32 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.

The ASR performance is measured with Syllable Error Rate (SLER) since Vietnamese word segmentation is not a trivial task and segmentation errors may prevent a fair comparison of different ASR hypotheses.

5.1.2. Vietnamese Text and Speech Resources

Since syllable plays an important role in Vietnamese language (it is both morphological and phonological base units), a vocabulary of about 6,500 syllables (called *V*0 since there is no word in this vocabulary) was extracted from a 35k word vocabulary (called *V*35k). Then the syllable-based and the word-based pronunciation dictionaries were built by applying our *VNPhoneAnalyzer* [Le 2004].

Documents were gathered from Internet and filtered for building a *Broadcast news* text corpus. After the data preparation steps, the text corpus has a size of 317 MB, i.e. 55 million words. A syllable-based and a word-based trigram LMs were trained from this text corpus using the SRILM toolkit [Stolcke 2002] with a Good-Turing discounting and Katz backoff for smoothing. It is important to note that with this toolkit, the unknown words are removed in our case, since we are in the framework of closed-vocabulary models.

Speech data was extracted from the *VNSpeechCorpus* [Le 2004], which was built at LIG and MICA laboratories. For acoustic modeling, 13 hours of speech data spoken by 36 speakers were used. In order to evaluate the performance of the crosslingual acoustic modeling methods, we used 400 "dialogue-like" sentences spoken by 3 speakers (called *Eval01* set). The second test set contains 277 long sentences spoken by 2 speakers (called *Eval02* set). The *Eval02* test set will be used to evaluate the performance of word/sub-word decomposition and combination schemes. We note that all of speakers in *Eval01* and *Eval02* test set are different from the training speakers.

5.2. Experimental Results

5.2.1. Crosslingual acoustic modeling experiments

For crosslingual 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 (Schultz 2001). 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 3 presents the syllable error rates of crosslingual models with different amount of adaptation data (evaluated on the *Eval01* test set). VN-CI and VN-CD1000 are baseline systems (no use of crosslingual information for bootstrapping process) which correspond to context independent and context dependent models with 1000 subtriphones. Similarly, MM7/VN-CI and MM6/VN-CD1000 are crosslingual context independent and context dependent models. We found out that when only 2-3 hours of data were available in the target language, crosslingual context independent models outperformed crosslingual context dependent models but when there were more data (10-15 hours), crosslingual context dependent models were better. In both cases, however, the use of crosslingual approaches to bootstrap the systems outperformed the baseline. It is of course more clear when only a small amount of data is available (2.25 hours).

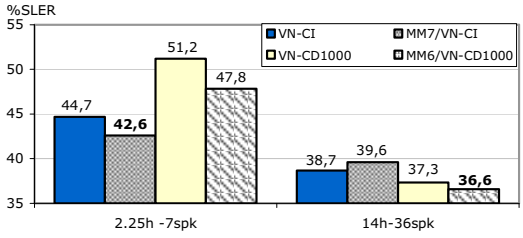

Figure 3: Comparison of acoustic modeling techniques with different amount of adaptation data for Vietnamese ASR

5.2.2. Word decomposition experiments

In order to test the performance of the word lattice decomposition method, we use the following test protocol: firstly, from the initial syllable vocabulary (*V*0), we progressively add *N* most frequent words in the *V*0. By increasing *N* from 0 to 35k, we have 10 different hybrid syllable/word vocabularies (called *V*0, *V*0.5k, *V*1k, …*V*35k) and 10 different trigram LMs are trained with these vocabularies. Secondly, words in lattices outputted from 10 speech recognizers (called *original lattices*) are decomposed into syllables by proposed algorithm (called *decomposed lattices*). Finally, these lattices are converted into CNs by the *lattice-tool* of the SRILM toolkit. We note that in these experiments, the crosslingual context dependent models (MM6/VN-CD1000) presented below are used. Figure 4 shows a comparison of the consensus hypothesis decoded from the *original CN* and the *decomposed CN*. Even if results evaluated on the *Eval02* test set show that the syllable-based LM is never outperformed by hybrid word/syllable based LMs, the *decomposed CN* works systematically better than *original CN*. It results in an absolute SLER reduction of 0.5% over the *original CN* when the *V*25k vocabulary is used.

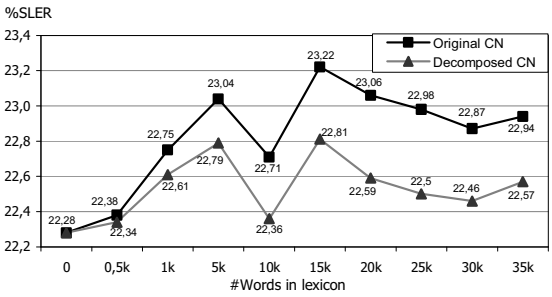

Figure 4: Comparison of the original lattices and the decomposed lattices as a function of the number of words added to the initial syllable vocabulary (V0)

5.2.3. Lattices combination experiments

In this experiment, the word and sub-word lattices combination scheme presented in section 4.2 is used. Syllable-based lattice and word-based lattice are first decoded from *V*0 and *V*35k system, respectively. Every word in the word-based lattice is then decomposed into syllables. Before converting to CN, both lattices are combined with (called *CN_Norm*) and without (called

CN_NoNorm) the normalization of the posterior probabilities.

Figure 5 presents an overview of the results evaluated on the *Eval02* test set: sentence MAP baseline hypotheses for *V*0 and *V*35k systems, consensus hypotheses decoded from CNs for both systems, consensus hypotheses decoded from *CN_NoNorm* and *CN_Norm*. The results show the benefit of the lattices combination (when done with normalization) compared to the sentence MAP baseline. This lattices combination approach leads to a significant improvement compared to the sentence MAP baseline approach, which shows the interest of using multiple units (word, sub-word) for LM in ASR.

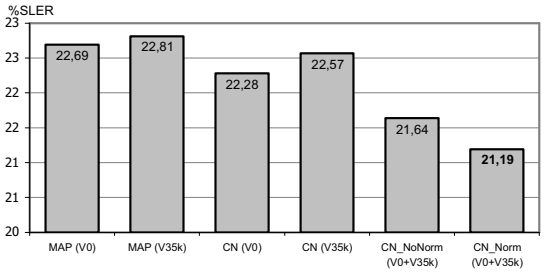

Figure 5: Comparison of the syllable-based lattices, word-based lattices and the combined lattices.

6. CONCLUSIONS

This paper presented our methodology for ASR in the context of under-resourced languages, particularly for Vietnamese. Our data collection methodology was explained. To obtain text corpora, we propose to use a general XML format and developed some tools to convert, process and normalize text corpora. For language modeling, we investigated word, sub-word and hybrid word/sub-word models.

For acoustic modeling, a crosslingual acoustic modeling is presented. We presented the potential of crosslingual independent and dependent acoustic modeling for the Vietnamese language. Experimental results on the Vietnamese ASR showed that when there were only a few hours of speech data in the target language, crosslingual context independent modeling worked better. However, when more speech data had been used, crosslingual context independent modeling was outperformed by crosslingual context dependent modeling. We also noticed that in both cases, crosslingual systems were better than monolingual baseline systems.

Moreover, a word/sub-word lattices decomposition and combination approach was proposed in order to exploit the use of multiple units in ASR. This approach was tested in an ASR system for Vietnamese. We conclude that our lattices combination method outperformed the sentence MAP baseline. Moreover, the lattices decomposition and combination tools are made available by the authors for any person who is interested in. In the future, we plan to apply these methods in Khmer language in which more lexical units (word, syllable, characters cluster and character) can be investigated.

7. REFERENCES

[Abdillahi 2006] N. Abdillahi et al., "Automatic transcription of Somali language", *Interspeech'06*, pp. 289-292, Pittsburgh, PA, 2006.

[Afify 2006] M. Afify et al., "On the use of morphological analysis for dialectal Arabic Speech Recognition", *Interspeech'06*, pp. 277-280, Pittsburgh, PA, 2006.

[Besacier 2007] L. Besacier et al., "The LIG Arabic/English Speech Translation System at IWSLT 07", *IWSLT'07*, Trento, Italy, 2007.

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

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

[Kurimo 2006] M. Kurimo et al., "Unsupervised segmentation of words into morphemes - Morpho Challenge 2005: Application to Automatic Speech Recognition", *Interspeech'06*, pp. 1021-1024, Pittsburgh, PA, 2006.

[Le 2003] V-B. Le et al., "Using the Web for fast language model construction in minority languages", *Eurospeech'03*, pp. 3117-3120, Geneva, Switzerland, 2003.

[Le 2004] V-B. Le et al., "Spoken and written language resources for Vietnamese", *LREC'04*, pp. 509-602, Lisbon, Portugal, 2004.

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

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

[Li 2002] X. Li, R. Singh, R. M. Stern, "Lattice Combination for Improved Speech Recognition", *ICSLP'02*, Denver, CO, 2002.

[Mangu 2000] L. Mangu et al., "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks", *CSL*, vol. 14, no. 4, pp. 373-400, 2000.

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

[Soltau 2001] H. Soltau et al., "A One Pass-Decoder Based On Polymorphic Linguistic Context", *ASRU'01*, pp. 214-217, Trento, Italy, 2001.

[Stolcke 2002] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *ICSLP'02*, vol. 2, pp. 901-904, Denver, CO, 2002.

[Denoual 2006] E. Denoual, Y. Lepage, "The character as an appropriate unit of processing for non-segmenting languages", *NLP Annual Meeting*, pp.731-734, Tokyo, Japan, 2006.

[Habert 1998] B. Habert, C. Fabre, F. Issac, "De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques", *InterEditions*, Masson, Informatiques, Paris, France, 1998.