FIRST STEPS IN FAST ACOUSTIC MODELING FOR A NEW TARGET LANGUAGE: APPLICATION TO VIETNAMESE

Viet Bac Le, Laurent Besacier

CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE
{viet-bac.le, laurent.besacier}@imag.fr

ABSTRACT

This paper presents our first steps in fast acoustic modeling for a new target language. Both knowledge-based and data-driven methods were used to obtain phone mapping tables between a source language (french) and a target language (vietnamese). While acoustic models borrowed directly from the source language did not perform very well, we have shown that using a small amount of adaptation data in the target language (one or two hours) lead to very acceptable ASR performance. Our best continuous vietnamese recognition system, adapted with only two hours of vietnamese data, obtains a word accuracy of 63.9% on one hour of vietnamese speech dialog for instance.

1. INTRODUCTION

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this "linguistic divide" is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years, for instance in the SALTMIL[1] (Speech and Language Technology for Minority Languages) group. However, in SALTMIL, "minority language" mostly means "language spoken by a minority of people". We rather focus, in our work, on languages which have a "minority of resources usable in HLT". These languages are mostly from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

▪ Portability of acoustic models: this can be achieved, for example by using tools for performing a fast collection of

---

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL

speech signals [1] or by using Language Adaptive Acoustic Modeling [2],

▪ Language modeling for new languages: we have already proposed to use web-based techniques which have shown ability to collect large amount of text corpora [3]. For languages in which no usable text corpora exist, this is moreover the only viable approach to collect text data,

▪ Dictionaries: collaborative approaches like in [4] could be also proposed for ASR.

This paper addresses particularly our first steps in fast acoustic modeling for a new target language. At first, we had no speech data at all in the target language (vietnamese). Then, after having collected some vietnamese speech data, no phonetic alignment was available and so an automatic labeling process was firstly employed to phonetically align these acoustic data. Acoustic models for this labeling process were borrowed from French (source language) in which a huge amount of acoustic data was already available.

Concerning cross-lingual acoustic modeling, two approaches are proposed in section 2. The first approach is based on a pronunciation modeling with the source language phoneme set and the second is a model mapping approach. The main difference between both approaches lies in the phoneme set used in acoustic modeling. While the first approach uses a french phoneme set, the second uses a vietnamese phoneme set. Since the first acoustic models trained without any target data did not lead to acceptable performance, an adaptation process is also presented to improve the recognition rate by using a small amount of speech data in target language. Experimental framework and results are presented in section 3 and 4 respectively. Finally, section 5 concludes with work and gives some perspectives.

2. CROSS-LINGUAL ACOUSTIC MODELING

2.1. Phone mapping table generation

Some methods of phone mapping can be used to evaluate acoustic similarities across languages. The core of these methods is the phone mapping table that describes the similarity of sounds between two different languages. Both *knowledge-based* and *data-driven* methods [5] are used in our work to manually or automatically obtain these phone mapping tables. Table 1 shows an example of some phone units in Vietnamese mapped from a french phone set by both knowledge-based and data-driven phone mapping methods.

Vietnamese phoneme	French phoneme	
	Knowledge-based	Data-driven
t	t	t
G	g	g
X	k	R
NG	NG	n
w	w	o
u_o	u	o
...	...	...

Table 1: Sample of a Vietnamese/French phone mapping table

a) Manual Phone Mapping Table Generation (knowledge-based method)

The phone mapping table can be obtained by using acoustic-phonetic knowledge to categorize phonetic units with similar features of the individual languages. In a knowledge-based method, we find the IPA counterpart of target phonemes among phonemes in source language. This kind of method can be used if no data at all is available in the target language but a good knowledge of the target language is needed.

b) Automatic Phone Mapping Table Generation using Confusion Matrix (data-driven method)

By using small amounts of acoustic data in the target language, the phone mapping table can be automatically created with data-driven methods. In our work, a confusion matrix is calculated by applying a source language phoneme recognizer on a target language speech corpus already phonetized with the target language acoustic units.

Firstly, as in [6], a phoneme recognizer in the source language is applied on the development data set in target language to decode the phonetic representation of each utterance. A phonetic transcription in target language phone sets of these utterances must be already available. Then, to align phonetic hypotheses with their phonetic transcription references, we project them in a time scale (see figure 1).

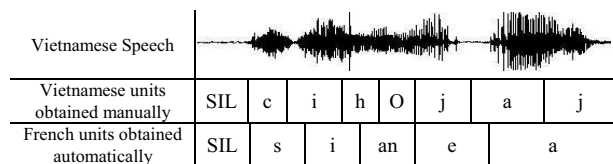| Vietnamese Speech								
Vietnamese units obtained manually	SIL	c	i	h	O	j	a	j
French units obtained automatically	SIL	s	i	an	e	a		

Figure 1: Time-based source/target phoneme scoring

The phonetic hypotheses and the references are thus compared frame by frame to count the co-occurrences between a phoneme in source language and a phoneme in target language. These counts make up a matrix where each entry gives the number of times a reference phoneme in the target language has been confused with a phoneme in source language. To obtain the final *confusion matrix*, each entry is normalized by dividing it through the total of occurrences of all corresponding phonemes in source language. Finally, the phone mappings are derived from this confusion matrix by selecting each phoneme in target language with the correspondence phoneme in source language which has the highest confusion value.

Furthermore, we can also calculate the Phoneme Correct Rate (PCR) of a phonetic recognizer based on this confusion matrix. Phonetic references and hypotheses in a same phone set are aligned to count the co-occurrences $C_{i,j}$ between phoneme i

and phoneme j (at the frame level) in a set of n phonemes. Then the PCR is calculated as shown on the following equations.

$$\text{Normalization of } i^{th} \text{ phoneme: } N_i = \sum_{j=1}^{n} C_{i,j}$$

$$\text{Phone Correct Rate: } PCR = \sum_{i=1}^{n} \frac{C_{i,i}}{N_i}$$

In the experiments of section 4, we will report PCR to evaluate the quality of our acoustic models, but also Phone Accuracy (PA) rate which is also used in the literature and which corresponds to the count of the correct hypothesis phonemes found on a signal compared to an aligned reference (same scoring as the word accuracy using sclite tool for instance).

2.2. Pronunciation modeling with the source language phoneme set

The purpose of the pronunciation modeling approach is to describe the pronunciation of a word or a dictionary entry in the target language in terms of the symbols associated with the acoustic units of the source language(s). That means that target language words are described in term of source language units.

Cross-lingual pronunciation modeling techniques were previously proposed in [6, 7, 8] to phonetically transcribe a word in target language in terms of the symbols used in source languages acoustic models. To automatically associate each of the target language words with a sequence of phonemes in the source language(s), a phonetic recognition system in the source language(s) was applied on the words to be phonetized. A pronunciation model of each word in the training data was obtained via a N-best hypotheses list. One drawback of these approaches is that one have to apply a phoneme recognizer on each target language word to be phonetized.

In our work, an important difference to note is that we already have a pronunciation dictionnary for the target language where each word is phonetized in target language phone units, because a vietnamese phonetizer was already designed in a previous work [9]. Then, the pronunciation modeling process is based on the transformation of each entry of the target language pronunciation dictionary into dictionary entries described with the source language acoustic units. The process used to transform a pronunciation dictionary from a target language phoneme set to a source language phoneme set is achieved using the following steps:

1. Use data-driven phone mapping techniques to build the confusion matrix. Each phoneme in target language can be mapped into N-best phonemes in source language depending on their confusion values.
2. Transform the target language pronunciation dictionary by replacing its phone units by their corresponding units in source language. Thus, each target language dictionary entry may be transformed to one or more entries in source language (table 2).

Vietnamese word	Vietnamese pronunciation	French pronunciation
gánh(1)	g EX J	g a J
gánh(2)	–	g in J
gánh(3)	–	g in n
hiên(1)	h ie n	R e n
hiên(2)	–	R i e n

Table 2: Pronunciation modeling using phone mapping

After obtaining a pronunciation dictionary in terms of the symbols associated with the acoustic units of the source language, we can directly use acoustic models in source language to decode the speech of an utterance in the target language. Moreover, for this approach it is interesting to use several source languages instead of one to better cover the phone inventory of the target language. Thus, this approach can be extended and used with multilingual acoustic models [2].

2.3. Acoustic model mapping

In that case, the difference with section 2.2 approach is that the final phoneme set used is the target language one. A phone mapping table is first created by using both knowledge-based and data-driven phone mapping methods.
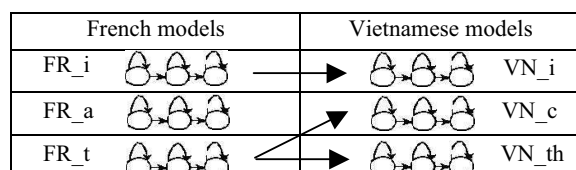

Figure 2: Vietnamese model mapped from French model

Cross-lingual acoustic models are then built by borrowing source language acoustic models (see figure 2). The obtained cross-lingual models can be used to decode the acoustic representation of each utterance by using a pronunciation dictionary in target language. This approach can be extended to use sub-phone units mapping [2, 5] where HMM states of a phoneme in target language can be borrowed from HMM states of some different phonemes in source language.

2.4. Model adaptation

While both the knowledge-based and data-driven phone mappings can be used without modification of the original source language acoustic models, HMM adaptation using MLLR or MAP [5] techniques can also be used to improve the systems using a small amount of target language adaptation data if available.

In our work, after the cross-lingual acoustic modeling step, limited data from the target language was phonetically aligned using the initial acoustic models. Then, these models were adapted with this limited data using MAP adaptation.

However, after adapting the cross-lingual models, a first evaluation showed high error rates for some phonemes. A reason can be that for these error prone phonemes, there was an important difference between the source data (used to build the initial model) and the target data (used for adaptation).

Thus, we decided to propose a *model collective adaptation* method to improve these phoneme error rates. The following algorithm is used:

1. By evaluating the individual phone error rates on a development set, we eliminate cross-lingual acoustic models which have a high error rate (using a threshold).

2. We reinitialize these eliminated models by running a k-mean clustering process on the training data of the target language. That means that only target language data is used to train these few models.

3. We perform some iterations of training on the data set of target language.

The performances of cross-lingual acoustic modeling and adaptation processes are shown in the section 4.

3. EXPERIMENTAL FRAMEWORK

3.1. Vietnamese pronunciation dictionary creation

Vietnamese language is a monosyllabic and tonal language with 6 tones (see figure 3). Except the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part.

Tonal syllable (6,492)				
Base syllable (2,376)				Tone (6)
INITIAL(22)	FINAL (155)			
	Medial(1)	Nucleus(16)	Ending(8)	

Figure 3: The phonological hierarchy of Vietnamese syllables with the total number of each phonetic unit

A vocabulary of 6,492 isolated-words was firstly extracted from 40,000 full-words vocabulary. Then a pronunciation dictionary for Vietnamese was built by applying our *VNPhoneAnalyzer* [9] on this isolated-word vocabulary. The pronunciation dictionnary was finally verified with the help of the Linguistic Institute in Vietnam.

3.2. Text corpus and language modeling

A new methodology for fast text corpora acquisition for minority languages was already proposed and used in [3]. Documents were gathered from Internet by some web robots. Then, these web pages were filtered and analyzed for building a text corpus. The Vietnamese data collection is composed of more than 2.5 GB of web pages. After data preparation, the text corpus is made of 868 MB, i.e. 10,020,267 sentences.

By using the vocabulary of 6,492 isolated-words, a Vietnamese language model was constructed and estimated from this text corpus. The perplexity value evaluated on our future test corpus is 108.5.

3.3. Speech corpora

To calculate the confusion matrices (for data-driven approach), to train the acoustic models in baseline and adapted systems and to test the performance of ASR systems, a vietnamese speech corpus was needed. *VNSpeechCorpus* [9], which have been built in CLIPS-IMAG and MICA[2] Laboratories, was used. The speech is recorded in both quiet and office environment. The speakers are from 3 major dialect regions of Vietnam: the South, the North, and the Middle. Each speaker has been asked for recording about 55-60 minutes, which includes 25 minutes of phonemes, tones, digits, application words, 6-7 minutes of short dialogue and 25 minutes of about 40 common and private text-based paragraphs.

At this time, 15 speakers have been recorded in quiet environment. We use the text-based paragraphs subset as development and adaptation corpus (about 3 hours). The dialogue subset is used as test corpus (about 1 hour).

---

[2] www.mica.edu.vn

The BREF80[3] and BRAF100 speech corpora [1] were used to train the French ASR system. The BREF80 corpus, designed at LIMSI, contains about 10 hours of speech data of 80 speakers. The BRAF100 corpus, which was recorded in CLIPS-IMAG laboratory by 100 speakers, contains about 25 hours of speech data. Both Vietnamese and French speech data used a sampling frequency of 16 KHz and a sampling rate of 16 bits.

4. EXPERIMENTATIONS

4.1. ASR System

All recognition experimentations use the JANUS Speech Recognition Toolkit (JRTK) [10] developed by the ISL Laboratories. The model topology is 3 states left-to-right, 64 Gaussian mixtures. The pre-processing of the system consists of extracting a feature vector every 10 ms. The feature vector of 43 dimensions contains zero-crossing, 13 MFCC, energy and their first and second derivatives. A LDA transformation is used to reduce the feature vector size to 24. For the moment, we deal with context-independent acoustic models only for Vietnamese. In the experiments described in this paper, the phones are modeled independently of the tones. The decision between two different words corresponding to a same phone sequence but to different tones is made by the language model.

4.2. Experimental results

In order to test both the phoneme recognizer and the whole ASR system, we systematically report phoneme correct rate (PCR) and word accuracy (WA) obtained on our vietnamese test data (about 1 hour). The phoneme accuracy (PA) is also given in table 3 and 4 but it is obviously very correlated with the PCR.

Table 3 shows the performance obtained without any adaptation data. In model mapping approach, data-driven method produces slightly better results in comparison to knowledge-based (KB) method. However, the difference is very small and in any case, the performance is not really acceptable.

Approaches	%PA	%PCR	%W.A
DictMap-DataDriven	-	-	16.54
ModelMap-KB	**27.18**	24.51	16.13
ModelMap-DataDriven	26.04	**25.89**	**18.52**

Table 3: Phone recognition and ASR performances without adaptation data

Table 4 shows the performance obtained with adaptation data aligned with acoustic models obtained with the *model mapping* approach with either knowledge-based or data-driven method. The performance of our *model collective adaptation* process is also reported. Adaptation data comes from a part of the *VNSpeechCorpus* which contains about 3 hours of speech data. We divide this corpus into 3 adaptation subsets (each contains about 1 hours of data). It is important to note that in these 3 adaptation subsets, the speakers are the same as the ones in the test data. So, in addition to language adaptation, we have to admit that we also do speaker adaptation at the same time. This fact can explain the big difference in performance observed between "no adaptation data" and "one hour adaptation data". The adaptation process is then performed in 3 cycles over 3 subsets. It is interesting to note that with only one hour of signal

---
[3] http://www.elda.fr/catalogue/en/speech/S0006.html

we already reach acceptable ASR performance: 62.6% with the *model collective* adaptation which gives the best result. Quite surprisingly, the improvement of the results is then not very important when using 2 and 3 hours of adaptation data.

Models	Adapt 1h			Adapt 2h			Adapt 3h		
	PA	PCR	WA	PA	PCR	WA	PA	PCR	WA
K-Based	58.0	57.1	60.4	59.2	58.9	63.6	58.9	58.8	62.2
D-Driven	58.7	57.4	61.6	59.4	58.8	63.8	**59.8**	**59.4**	63.2
Collective	**59.3**	**58.9**	**62,6**	**59.9**	**58.9**	**63.9**	59.6	58.9	**63.4**

Table 4: Phone recognition and ASR performances with adaptation data

5. CONCLUSIONS AND PERSPECTIVES

This paper presented our first steps in fast acoustic modeling for a new target language (vietnamese). Both knowledge-based and data-driven methods were used to obtain phone mapping tables between a source and a target language. While acoustic models borrowed directly from the source language did not perform very well, we have shown that using a small amount of adaptation data in the target language (one or two hours) lead to very acceptable ASR performance. Our best vietnamese recognition system, adapted with only two hours of vietnamese data, obtains a word accuracy of 63.9% on a speech dialog test set for instance. These cross-lingual acoustic modeling and adaptation techniques are also currently applied on Mexican and Khmer languages.

6. REFERENCES

[1] D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, M. Akbar, "A New Methodology For Speech Corpora Definition From Internet Documents", *LREC 2000*, Athens, Juin 2000.

[2] 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.

[3] V.B. Le, B. Bigi, L. Besacier, E. Castelli, "Using the Web for fast language model construction in minority languages", *Eurospeech 2003*, pp. 3117-3120, Geneva, September 2003.

[4] V. Berment, "Several technical issues for building new lexical bases", *PAPILLON Workshop*, Tokyo, 2002.

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

[6] S. Stüker, "Automatic Generation of Pronunciation Dictionaries For New, Unseen Languages by Voting Among Phoneme Recognizers in Nine Different Languages", *Master thesis*, Carnegie Mellon University, April, 2002.

[7] T.Martin, T.Svendsen, S. Sridharan, "Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition", *Eurospeech 2003*, pp. 3125-3128, Geneva, September 2003.

[8] R. Bayeh, S.Lin, G.Chollet, C.Mokbel, "Towards multilingual speech recognition using data driven source/target acoustical units association", *ICASSP 2004*, vol. I, pp. 521-524, Montreal, Canada, May 2004.

[9] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, J-F. Serignat, "Spoken and written language resources for Vietnamese", *LREC 2004*, Lisbon, May 2004.

[10] M.Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine", *ICASSP'97*, Munich, 1997.