Automatic Speech Recognition for Vietnamese using HTK system

NGUYEN Hong Quang, TRINH Van Loan, LE The Dat
School of Information and Communication Technology,
Hanoi University of Technology,
Hanoi, VIETNAM
(quangnh, loantv)@it-hut.edu.vn, thedat2001@yahoo.com

Abstract—**This paper presents our research on Vietnamese speech recognition using HTK[1] system. Our method is the adaptation of Vietnamese data into HTK system: text corpus was collected from Web and then bigram language model was built by the CMU SLM toolkit. The experiments were carried out with different acoustic models. Further more, the experiment using speaker adaptation technique was implemented by MLLR algorithm. For our results, WA (Word Accuracy) in the case of speaker independent recognition test is 71.37% and 75.96% with the test of speaker adaptation.**

Keywords-Automatic speech recognition, Vietnamese speech, acoustic model, HTK system, speaker adaptation

I. INTRODUCTION

Vietnamese is a language in the Viet-Muong group of Mon-Khmer branch of the Austro-Asiatic language family. In the opinion of the majority of linguists, Vietnamese is a tonal language with six tones. In each syllable, there is only a single tone. Tone is a very important component because the meaning of a word depends on tone. In recent years, researches in the area of speech recognition have been stepped up and recognition modules began to be deployed in real applications.

However, researches are aimed at the popular languages such as English, French, Mandarin Chinese, etc. The researches on the Vietnamese were just beginning. A quick approach is based on available systems developing for the languages such as English and French and then implementing appropriate adaptation for Vietnamese [1][5].

A fundamental difference characteristic between Vietnamese and the languages such as English and French is tonal property of the Vietnamese. For this reason, it is necessary to settle the question: how to represent the Vietnamese tones. There are two main methods for this problem. The first method integrated immediately tone information into Vietnamese phonemes [1][4] and the second method built an independent module to handle tones [5].

In addition, when speech recognition tests experimented with a language, the first necessary thing is to build data resources for this language. The most important resources of a language are speech corpus and text corpus. With a minority language like Vietnamese [2], a text corpus can be built quickly based on Web data sources [3][6] and speech corpus was built by transcription with data collected from radio stations [1][4] or by recording directly in a studio [2].

In this paper, we present at first the method of collecting language resources (text corpus and speech corpus) for Vietnamese. And then we focus on a method for building quickly the Vietnamese recognition system with the adaptation of Vietnamese data sources into a speech recognition system which has been developed for English and French. The method uses directly the representation of the tone information in Vietnamese phonemes. We have tested this method on a HTK system.

The next of this paper is organized as follows:

- In Section II, we describe the method of collecting resources for Vietnamese language: vocabulary, text corpus and speech corpus.
- In Section III, the method of Vietnamese recognition with HTK is represented, with special emphasis on the integration of tone information in the acoustic model. The recognition results are also given in this section.
- Finally, conclusion and future research are given in Section IV.

II. COLLECTING NATURAL LANGUAGE RESOURCES FOR VIETNAMESE

In this section, we present the methods to collect the most important resources for a speech recognition system: vocabulary, speech corpus and text corpus.

A. Vietnamese vocabulary

Firstly, the vocabulary is taken from an open source data. (http://www.informatik.uni-leipzig.de/duc/software/misc/wordlist.html). This vocabulary contains 22,418 words (single-syllable and multi-syllable words in Vietnamese). By taking the single syllable in these words, we have created the single-syllable vocabulary with 5943 syllables.

B. Vietnamese text corpus

Text corpus is widely used in speech recognition systems. One application of the text corpus is to create statistical language models. In this case, to obtain a good quality language model, a large text corpus is required, which covers a wide range of different types of documents.

However, the text corpus collected manually is impossible. Most studies have focused on the automatic construction of text corpus with data sources from the Internet [2].

[1]http://htk.eng.cam.ac.uk/

978-1-4244-8075-3/10/$26.00 ©2010 IEEE

1) Method of collecting text form Internet: to be able to collect the documents, we have relied on electronic sources of text collected from Internet. These texts will be converted to UTF-8 character code. Normally, newspaper pages collecting from electronic texts in HTML format have so many redundant information such as HTML tags, special symbols, foreign words, abbreviations, and data (integers, real numbers), date...

At the first step, we implemented a filter to remove redundant information and to normalize text. Requirements for this step are to create the text that includes only the words in the Vietnamese dictionary. For this reason, we built a dictionary including common Vietnamese syllables (approximately 6000 syllables). Then in every HTML format article, we extracted the main contents, and then we normalized the text as follows:

- Remove HTML tags
- Exclude words which are not in the dictionary
- Replace abbreviations by full words. For example, "DHBK") => "Đại Học Bách Khoa" (University of Technology)
- Conversion of special characters. Example: "%" => "phần trăm" (percent)
- Conversion of data from number forms into text. Example: "2008" => "hai ngàn không trăm linh tám" (two thousand eight).
- Conversion of date to text. For example, "12/1/1999" => "ngày mười hai tháng một năm một ngàn chín trăm chín chín" (Twelfth, January, one thousand nine hundred and ninety-nine); "3/1" => "mồng ba tháng một" (the third January)
- ...

Finally, each sentence is written down in a text file with the format : <s> "Content" </ s>. For example: <s> Chúng tôi là sinh viên</s> (<s> We are students </s>).

TABLE I. THE SIZE OF TEXT CORPUS COLLECTED FROM VIETNAMESE ELECTRONIC DOCUMENTS

Type of text page	Electronic Source	Original size ((in HTML format)	Size after filtering	Filtration rate (times)
Electronic Newspapers	Vnexpress.net	2,79 GB	168 MB	16.6
	Vietnamnet.net	1.26 GB	113 MB	11.2
Electronic Literatures	Vanhoc.xitrum.net	415 MB	97 MB	4.3
	Vnthuquan.net	1.2 GB	162 MB	7.4
Totals		4.814 GB	540 MB	8.9

2) Building text corpus: the method presented above is used to build a text corpus. First, we collect HTML files from two electronic newspapers and two electronic literatures [Tab. I]. With the results of table I we see that redundant information removed from electronic newspaper is more than from electronic literatures. This is due to an electronic newspaper pages in HTML format will contain much more redundant information such as HTML tags, menus, images, advertising, etc.

Text corpus collected is named as BKVTEC (Vietnamese BachKhoa Text Corpus), with a capacity of 535 MB, including 4 million sentences with 90 million syllables.

3) Construction of statistical language model: we have used the above BKVTEC to build statistical language models. BKVTEC is divided into two parts: 90% for the training part and 10% for the test part [Tab. II].

TABLE II. TRAINING AND TEST USED FOR CONSTRUCTION OF STATISTICAL LANGUAGE MODELS

Type of text page	Training Part	Test Part
Newspapers	259 MB	22 MB
Literatures	237,9 MB	21 MB
Newspapers + Literatures	496,9 MB	43 MB

CMUSLM Toolkits (http://www.speech.cs.cmu.edu/SLM_info.html) were used to build statistical language models. The unigram, bigram, trigram LM were experimented and the results are described in Tab. III.

TABLE III. PERPLEXITY OF VIETNAMESE LANGUAGE MODELS ON THE TEST CORPUS

Language Model (LM)		Perplexity	
		Literature test corpus	***Newspaper test corpus***
Bigram LM	*Literature* LM	**198.94**	326.54
	Newspaper LM	497.41	**108.57**
	Newspaper + Literature LM	227.94	122.22
Trigram LM	*Literature* LM	**131.42**	237.24
	Newspaper LM	279.26	65.45
	Newspaper + Literature LM	159.89	**62.43**

Normally, daily information source are found on electronic newspapers. Results in Tab. III shows that the addition of data collected from the pages of electronic literatures improves the perplexity of trigram language model on the electronic newspaper corpus.

C. Vietnamese speech corpus

The building Vietnamese speech corpus is performed in two steps (Fig. 1). At the step 1, we collected text documents, and then at the step 2, we recorded Vietnamese sentences in a studio with specialized equipment.

We use the same method presented in Section II to build the text for recording. Data source is collected from the electronic newspaper "www.vnexpress.net". The topics collected are daily life, science, business, and motor-cars. Then at the final step, the text files were checked and edited manually. Total number of sentences collected is 8047 with about 208000 syllables.

Recording process was realized at the Computer Engineering Lab of Hanoi University of Technology, with KayLab (http://www.kayelemetrics.com) recording equipment CSL Model 4500. We also used specialized equipment (Real-

time EGG Analysis) that allows recording EGG (Electroglottography) signal.

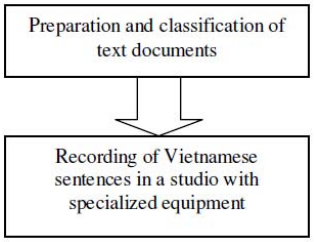

Figure 1. Diagram for the construction of Vietnamese speech corpus

TABLE IV. VIETNAMESE CHARACTER REPRESENTATION USED FOR HTK

Order	UTF-8	VIQR	BKTC	Examples
1	À	A`	A2	Bà (Grand mother)
2	Ã	A~	A3	Xã (Commune)
3	Ả	A?	A4	Ảnh (Image)
4	Á	A'	A5	Cá (Fish)
5	Ạ	A.	A6	Tạ (Quintal)
6	Ă	A(	A7	Ăn (Eat)
7	Â	A^	A8	Sân (Yard)
8	Ơ	O+	A9	Cơm (Rice)
9	Đ	Đ	DDi	Đi (Go)

Speech was recorded with sampling frequency of 16000Hz, mono mode, 16 bit per sample. Speech data recorded contain 19 male voices. The average age of speakers is 23. They have the same standard Vietnamese dialect (in the North of Vietnam). Corresponding to each topic, each speaker reads about 20-40 sentences which are randomly selected. Total number of sentences recorded is 3045 with the capacity 1.32 GB (both speech signal and EGG signal), corresponding to 5.93 hours of speech. This Vietnamese speech corpus is named BKSPEC.

III. VIETNAMESE SPEECH RECOGNITION USING HTK SYSTEM

The first thing to do is to make the HTK understand Vietnamese characters. We then tested a number of different acoustic models for Vietnamese.

A. Adaptation of Vietnamese characters to HTK system

A new problem to resolve is that HTK was originally designed to recognize English, so the characters are stored as 8 bit ASCII standard code. Meanwhile, Vietnamese are now stored in UTF-8 format. So we used the following method to adapt the Vietnamese characters to HTK.

First, Vietnamese characters were converted from UTF-8 to VIQR code (Vietnamese Quoted-Readable - this is a convention for writing Vietnamese using ASCII 7 bit) using Unikey tool (Fig 2).

Nevertheless, in HTK system characters '(' and '+' coincide with keywords: symbol '(' is used in the definition of grammar, '+' is used when creating triphone. So at the next step, these special characters were converted into the digits [Tab. IV]. This character coding method is named BKTC (BachKhoa Text Code).

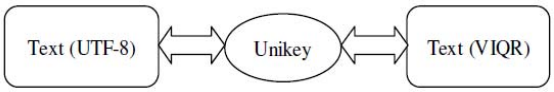

Figure 2. Conversion between UTF-8 code and VIQR code by using Unikey software

B. Vietnamese phonemes

1) Structure of Vietnamese syllables: a typical Vietnamese syllable [Tab. V] is composed of four main parts: Initial (one of 22 consonants), Medial (one of semi-vowels), Nucleus (one of 16 main vowels) and Ending (one of 8 phonemes). /M-NE/ is commonly known as Rhyme or Final part. For the four parts above, only Nucleus is required. The other parts may be present or absent in a syllable.

TABLE V. STRUCTURE OF VIETNAMESE SYLLABLE

Tonal syllable (5943)				
Initial (22)	**F**inal			**T**one (6)
	Medial (1)	**N**ucleus(16)	**E**nding (8)	

In addition, each Vietnamese syllable has only one of 6 tones: level tone, failing tone, broken tone, curve tone, rising tone and drop tone. Tone is a super-segment characteristic and could be on the full syllable [9] or just on the final part of syllable [10]. In the next section, we present some methods of integration of tone information into acoustic model for aVietnamese.

2) Representation of Vietnamese phonemes: to represent the phonemes used for acoustic model (AC-model) without using tone information in syllables [Tab VI], we have the following methods:

- Method AC_MN: each syllable consists of monophones: /I/ - /M/ - /N/ - /E/.
- Method AC_IF: each syllable consists of main vowel and the rhyme: /I/ - /MNE/.
- Method AC_DP: each syllable is a combination of two diphones: /I/ - /M/ - /N/ and /N/ - /E/

TABLE VI. REPRESENTATION OF VIETNAMESE PHONEME (EXAMPLE FOR SYLLABLE "/TOÁN/ - MATH", DIGIT '5' CORRESPONDING TO RISING TONE)

Order	Method	Example
1	AC_MN	**/t/ - /o/ - /a/ - /n/**
2	AC_MNT	**/t/ - /o/ - /a5/ - /n/**
3	AC_IF	**/t/ - /oan/**
4	AC_IFT	**/t/ - /oan5/**
5	AC_DP	**/toa/ - /an/**
6	AC_DPT	**/toa/ - /an5/**
7	AC_DPTA	**/toa5/ - /an5/**

To add tone information to syllable, we used the following methods to represent Vietnamese phonemes [Tab VI]:

- Method AC_MNT: tone is on the main vowel, so for each vowel, we have at most 6 new vowels with 6 tones.

- Method AC_IFT: tone is only on the Rhime
- Method AC_DPT: tone is only on the last diphone of syllable.
- Method AC_DPTA: tone is on both diphones of syllable.

C. Vietnamese speech recognition test

1) Acoustic model: we used the vocabulary described in section II. This vocabulary contains only single syllables (2521 syllables) which exists in the BKVTEC text corpus. Then we developed a program to build a pronunciation dictionary for Vietnamese.

Each phoneme is represented by a five-state HMM (Hidden Markov Model) model, which has 3 emitting states, start state and end state are non-emitting states. Each state is represented by a Gaussian Mixture Densities with 16 Gaussian components. Speech signal is parameterized by sequences of MFCC vectors with 39 parameters (12 MFCC coefficients, energy coefficient, with their first and second derivatives). The frame period is 10msec.

Speaker adaptation is a technique for increasing system performance when we have identified speaker. Two adaptive algorithms used widely today are MAP [8] and MLLR [7]. Both these algorithms are supported in HTK.

2) Language model: at present, HTK support bigram language model with monophones. We used the CMU SLM Toolkits on BKVTEC text corpus to create this language model in ARPA format. We then used HBuild tool of HTK to create the language model corresponding to HTK format. This model has 35812 bigrams.

3) Test results: BKSPEC Speech Corpus is divided into two parts: the training part including 15 speakers and the test part consists of four speakers. Total speaking time of the test is 88.8 minutes and the average time of each speaker is 22 minutes.

TABLE VII. WORD ACCURACY OF VIETNAMESE SPEECH RECOGNITION

Order	Method	Independent speaker	Adaptation speaker
1	AC_MN	59.61	65.60
2	AC_MNT	57.95	64.77
3	AC_IF	70.72	75.36
4	AC_IFT	70.39	75.73
5	AC_DP	**71.37**	**75.96**
6	AC_DPT	70.44	75.69
7	AC_DPTA	65.89	71.47

In the first test, all data in training corpus are used to create speaker independent acoustic models. Then this model is used for the speakers in the test corpus. The test is done with all of acoustic models presented in section III.B.

With speaker adaptation test, the general acoustic model is adapted to all data of a speaker in the test corpus by using MLLR technique and then the new model is used to experiment on the data of this speaker.

The test results are described in Tab. VII. We can see that the best result corresponds to the case of independent speaker recognition with AC_DPT method (Section III.B.2). The results also indicate that the adding tone information to syllable may not improve the performance of the system.

We can explain these results as the following: the acoustic model is trained by using MFCC parameters, but these parameters don't contain so much information about tons. In the literature, there are two methods which can improve the performance of system: adding tonal parameters (fundamental frequency F0) on parameter vector or using supra-segmental ton information [5]. These two methods will be applied in our future researches.

IV. CONCLUSION

In this paper, we have presented our study of building a Vietnamese speech recognition system based on HTK system. The results show that applying tone information directly to monophones representation does not help to improve the system performance Furthermore, the experiment also showed that the best result obtained when we used diphones in Vietnamese acoustic model, corresponding to a 71.37% word accuracy in the test of independent speaker and 75.96% with the adaptation speaker by MLLR algorithm.

For the future research, at first we will integrate trigram language model into our system. And then, construction of multi syllabic language models will be done and a tone recognition module will be integrated in the system using language models.

REFERENCES

[1] T. T. Vu, D. T. Nguyen, M. C. Luong, J-P. Hosom, Vietnamese large vocabulary continuous speech recognition, Interspeech 2005, Lisbon, Portugal, September, 2005.

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

[3] V. B. Le, D. D. Tran, E. Castelli, L. Besacier, J-F. Serignat, First steps in building a large vocabulary continuous speech recognition system for Vietnamese, RIVF 2005, Can Tho, Vietnam, February, 2005.

[4] Q. Vu, K. Demuynck, D. V. Compernolle, Vietnamese Automatic Speech Recognition: the FlaVoR Approach, ISCSLP 2006, Kent Ridge, Singapore, 2006.

[5] H. Q. Nguyen, P. Nocera, E. Castelli, V. L. Trinh, A Novel Approach in Continuous Speech Recognition for Vietnamese, an isolating tonal language, Interspeech 2008, Brisband, Australia, September, 2008.

[6] V.B. Le, B. Bigi, L. Besacier, and E. Castelli, Using the web for fast language model construction in minority languages, EUROSPEECH 2003, Geneva, Switzerland, September 2003.

[7] C. J. Leggetter, P. C.Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 1995, 171–185.

[8] Gauvain, C. Lee, Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions SAP 2, 1994, 291–298.

[9] T. T. Doan, Vietnamese Phonetic (in vietnamese). Editions of Hanoi University of Education, Hanoi, Vietnam, 1999.

[10] D.D. Tran, E. Castelli, J-F. Serignat, V.L. Trinh, and X.H. Le, Influence of f0 on vietnamese syllable perception, INTERSPEECH 2005, Lisbon, Portugal, September, 2005.