IEEE Catalog Number: CFP1523T-CDR
ISBN: 978-1-4673-6546-8

TABLE OF CONTENTS

Scroll to the title and select a **Blue** link to open a paper. After viewing the paper, use the bookmarks to the left to return to the beginning of the Table of Contents.

Cognitive/Wireless Networks

Chair: Dac-Binh Ha, *Duy Tan University, Vietnam*

Computing Technologies

Chair: Nguyen-Son Vo, *Duy Tan University, Vietnam*

Applications I

Chair: Ha H Kha, *Ho Chi Minh City University of Technology, Vietnam*

A novel approach for spoken term detection in Vietnamese

NGUYEN Hong Quang, TRINH Van Loan
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
quangnh, loantv@soict.hust.edu.vn

LE Xuan Thanh
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
thanhlx@soict.hust.edu.vn

***Abstract*— In this paper, we describe our results on spoken term detection (STD) for Vietnamese speech. A method for constructing the indexing module and the searching module for Vietnamese speech is presented. A new method for identifying detection threshold is also described. The experiments were carried out with different thresholds: normalized acoustic probability and posterior probability. For our results, optimal Actual Term-Weighted Value (ATWV) is 0.5982 (the probability of false alarm is 0.48% and the probability of miss detection is 20.85%), corresponding with the test using normalized posterior probability as detection threshold.**

Keywords*— Speech recognition, spoken term detection, Vietnamese speech, detection threshold

I. INTRODUCTION

Presently, several information retrieval systems have been developed, including some of world's most powerful systems like Google, Yahoo, etc. However, these systems only support searching information in text data format. Digitized voice sources are rapidly increasing from radio and television programs to online interviews, making information retrieval systems for voice more and more necessary.

Therefore, in recent years, the research of information searching directly from the source of multimedia data such as voice has seen very strong growth. NIST (National Institute of Standards and Technology) has designed a new idea called Spoken Term Detection [1]. Many research groups around the world have participated. An STD system consists of two main modules: speech indexing module and speech searching module. The first module is constructed based on a large vocabulary continuous speech recognition (LVCSR) system [3][4]. The second module is designed for maximizing speed of searching process.

In the STD system, we score each term. A term is accepted if the score of the term is greater than one threshold. There are several parameters used to identify the score of a term, the most popular of which is posteriori probability with and without normalization [12].

This paper presents our results of spoken term detection for Vietnamese speech. In this research, we used normalized posteriori probabilities for the score of the term.

The contents of the article include the following:

- Section 2 presents the method of constructing an STD system: indexing, using the search engine for Vietnamese speech, identifying the term scores and evaluating system performance.
- Section 3 presents the results of the experiments and the evaluations of the test results.
- Section 4 presents the conclusions and perspectives.

II. SPOKEN TERM DETECTION IN VIETNAMESE

A. *Overview of the STD system*

The goal of the STD system is to find all occurrences of a search term as quickly as possible in heterogeneous audio sources [1]. An STD system consists of two main modules: speech indexing module and speech searching module [Fig. 1].

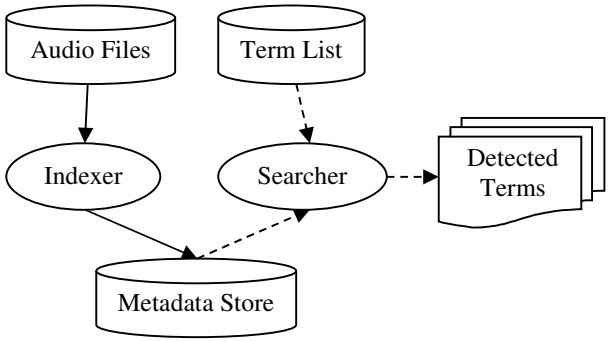

Fig. 1. The typical two-stage STD system architecture [1].

Step 1: Indexing speech file. The indexer used an automatic speech recognition system to generate a word lattice, because searching for terms in the lattice is more accurate than searching the 1-best transcript [6][7]. This module indexes each speech file and stores word hypotheses to the metadata store. The indexer never gets access to the terms.

Step 2: Searching. The searcher looks up each term (a term is a sequence of one or more words) in the metadata store (the searcher never gets access to the audio), and extracts the term hypotheses: file name, start-time, and end-time of term if

scores of these hypotheses are greater than a detection threshold. The most popular value used for term score is posterior probability [7].

Basic detection performance will be characterized in the usual way via standard detection error tradeoff (DET) curves of missed probability (P_{Miss}) versus false alarm probability (P_{FA}):

- P_{Miss}: probability of items in the database that cannot be found by the system.
- P_{FA}: probability of terms found by the system, but these terms do not exist in the metadata store.

Missed and false alarm probabilities are functions of the detection threshold.

To evaluate the performance of the system, the primary evaluation measure will be the "Actual Term-Weighted Value" (ATWV) [1]. The ATWV is the detection value attained by the system as a result of the system output and the binary "YES/NO" decisions output for each putative occurrence.

B. STD system for Vietnamese speech

An STD system consists of two components: indexer and searcher modules.

Indexer is constructed from a Vietnamese large vocabulary speech recognition (LVCSR) system [2]. For the given acoustic observation $\mathbf{X} = X_1X_2\ldots X_n$, the goal of speech recognition is to find the corresponding word sequence $W^* = w_1w_2\ldots w_m$ that has the maximal posterior probability $P(\mathbf{W} \mid \mathbf{X})$, as expressed by Eq. (1) [11].

$$W^* = \text{argmax}_w P(W|X) = \text{argmax}_w P(X|W).P(W)/P(X) \quad (1)$$

Since the maximization of Eq. (1) is carried out with the observation X fixed, the above maximization is equivalent to maximization of the Eq. (2).

$$W^* = \text{argmax}_w P(W|X) = \text{argmax}_w P(W).P(X|W) \quad (2)$$

Following Eq. (2), the score of word (in logarithm scale) to use in the decode process is expressed by Eq. (3).

$$\text{Score}(W) = \log P(W) + \log P(X|W) \quad (3)$$

The first component P(W) is extracted from the language model; otherwise, the second one is extracted from the acoustic model. To find the influence of the language model in the decode process, we define the weight γ of the language model and then the score of word is expressed by Eq. (4).

$$\text{Score}(W) = \gamma.\log P(W) + \log P(X|W) \quad (4)$$

Training the tri-phone acoustic model effectively often requires a large speech corpus. This is currently not feasible for Vietnamese. Therefore, we built mono-phone acoustic models for Vietnamese.

Overview of our Vietnamese speech indexing module is described in Figure 2. The system consists of two stages:

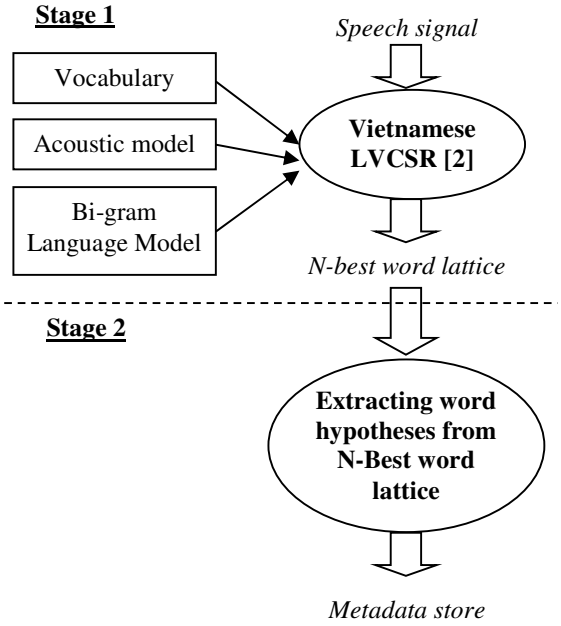

Fig. 2. Overview of Vietnamese speech indexing module

- Stage 1. Bi-gram language model is created by using reference text of train corpus. Then, speech signal (in WAV file format) is decoded on our Vietnamese speech recognition system. The outputs of this stage are a lattice N-best hypothesis.
- Stage 2. Extracting word hypotheses from N-Best word lattice and storing these word hypotheses into the metadata store.

Metadata store is organized into two sets of files:

- Content files contain the word hypotheses, which are start time, end time, and acoustic score of each word hypothesis.
- Index files contain the total number of hypotheses and indexes of each word hypothesis (used for expediting the searching process).

These word hypotheses are stored in our binary format. All items in the index files and content files are sorted alphabetically.

For each term, the search module will search occurrences of this term in the metadata store and compare the score of word with the threshold. Occurrences with scores larger than the threshold will be outputted with necessary information: file name, start-time, and end-time of term. In this research, we use different methods to calculate the score of term:

- Using directly acoustic score (this score is log likelihood which is created by the Vietnamese speech recognition system).
- We normalized acoustic score by time (the length of occurrences of search term).

- Using posterior probability. This probability is calculated by Eq. (4).

III. EXPERIMENTS

A. *Vietnamese Natural Resources*

The most common natural language resources are text corpus and speech corpus.

For text corpus: At first, we collect HTML files from two electronic newspapers and two electronic literatures. We see that redundant information removed from electronic newspapers is more than that removed from electronic literatures. This is due to electronic newspaper pages in HTML format containing much more redundant information, such as HTML tags, menus, images, advertising, etc. We implemented a filter to remove redundant information and to normalize text [2]. Text corpus collected is named BKVTEC (Vietnamese BachKhoa Text Corpus), with a capacity of 535 MB, including 4 million sentences with 90 million syllables.

For speech corpus, we use BKSPEC corpus [2]. The total number of sentences recorded is 3,044, corresponding to 5.93 hours of speech. This corpus is divided into two parts: the training part (TRAIN) including 15 speakers and the development part (TEST), consisting of four speakers. Total speaking time of the test is 88.8 minutes and the average time of each speaker is 22 minutes.

B. *Vietnamese speech recognition experiments and analysis results*

The basic resources to build a speech recognition system are: vocabulary, acoustic model and language model.

1) *Vocabulary and Pronunciation*

Vietnamese is a monosyllabic language. The basic component in Vietnamese writing is the syllable.

Each Vietnamese syllable has four main components:

- **I**nitial, one of 22 consonants.
- **M**edial, one semi-vowel.
- **N**ucleus, one of 16 main vowels.
- **E**nding, one of 8 phonemes.

Our vocabulary contains only single syllables with 8,557 syllables.

2) *Acoustic model and language model*

An important specific characteristic of Vietnamese is the tones. One can integrate the tonal information into the acoustic model. Nevertheless, the number of acoustic units will be more important and we have to increase the size of training data. We have thus used the acoustic model with basic acoustic units such as non-tonal di-phones [2]. Therefore, each Vietnamese syllable is a combination of two di-phones: /I/ - /M/ - /N/ and /N/ - /E/, and there are 589 di-phones. These acoustic models do not contain tonal information. The tonal information of each syllable will be determined during decoding processing by language model.

Each phoneme is represented by a five-state HMM (Hidden Markov Model) model, which has three emitting states; the 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. We used TRAIN corpus to train our acoustic model.

For the Vietnamese bi-gram language model, we used CMU SLM toolkits to construct tri-gram language model. BKVTEC text corpus is used to train the language model. In this language model, there are 1.72 millions bi-grams.

The experiments are tested on a PC (personal computer) with processors Intel Dual-Core E6600 3.06 GHz, RAM 1GB. WER (Word Error Rate) has been applied to estimate the results.

We must first define the weight γ of bi-gram language model. These weights are defined based on the experiments of Vietnamese recognition with the set TEST. In this experiment, our Vietnamese LVCSR is used to find the best hypothesis for sentences. We have used the different values of γ. The results are given in Fig. 3. The optimal value of γ is 14 with WER (Word Error Rate) = 26.96 %.

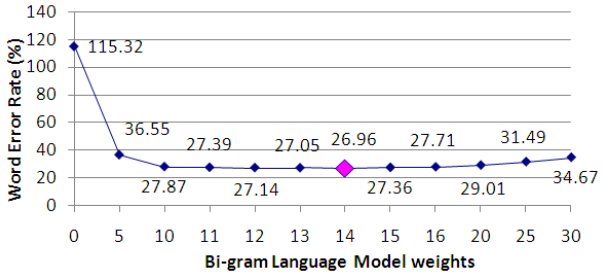

Fig. 3. WER (Word Error Rate) following by bi-gram language model weights the test experiments of Vietnamese recognition with the set DEV using our Vietnamese LVCSR for different values of γ and for bi-gram language model.

C. *STD Tests for Vietnamese speech*

1) *Building the search list*

TABLE I. DISTRIBUTION OF SEARCH TERMS USED IN TESTING

Term	Search term	Number of occurrences
1-syllables	40	5429
2-syllables	33	704
3-syllables	20	162
4-5-syllables	7	127
Total	**100**	**6422**

The search item list is built from the description of the files in the TEST set. In the list, there are 100 terms. First, we count the number of occurrences of each term in the BKVTEC, and then we sorted the list of terms by number of

occurrences and by number of syllables. We chose terms with numbers of occurrences greater than 5. We then manually check these terms. The statistical results of the search term list are described in Table 1.

2) STD Experiments

First we used the Vietnamese speech recognition system to create the N-best word lattice hypotheses for which N is 2 and 3. Then, experiments with different search threshold were conducted:

- **Test 1.** Using log likelihood of term, which is created by the Vietnamese speech recognition system.
- **Test 2.** In this test, we normalized acoustic score by time (the length of occurrences of search term) for threshold.
- **Test 3.** Using posterior probability for threshold (Eq. 3).

TABLE II. ATWV OF THE STD TEST WITH THE DIFFERENT THRESHOLD (USING N-BEST WORD LATTICES CREATED BY VIETNAMESE SPEECH RECOGNITION SYSTEM, N=2)

Type of threshold	ATWV
Acoustic score (log likelihood)	0.5388
Acoustic score normalized by length of search term	0.5805
Posterior probability	0.5389
Posterior probability normalized by length of search term	0.5982

These tests were experiments with values of 14 for language model weight γ (Fig. 3) and N is 2 or 3. In each experiment, we calculated P_{Miss}, $P_{FalseAlarm}$ and ATWV for all detection threshold. The optimum threshold is the value corresponding with the maximum ATWV.

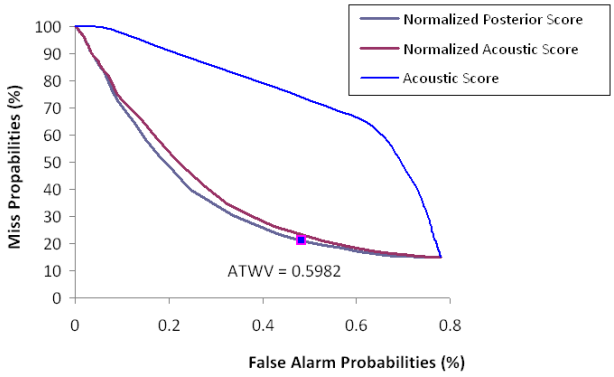

Figure 4. DET curves of the STD tests using different score of term: acoustic score, normalized acoustic score, normalized posterior score. Optimum ATWV is 0.5982 (probability of false alarm is 0.48% and probability of miss detection is 20.85%), corresponding with the test using normalized posterior probability as score of terms.

Because our dictionary contains only single-syllable words, for multi-syllable terms, the searcher looked for acceptance sequences of single-syllable words, and these words must satisfy adjacency timing constraints.

From the results (ATWV) in the Table 2 and from the DET curves in the Fig. 4, we see that:

- The result (ATWV) of the test using posterior probability for score of term is better than the one tested using score acoustic.
- The results of the tests using method of normalization score (acoustic score and posterior probability) are better than the one using normal threshold (which is not normalized by time).
- Then the, best result is obtained in the test using normalized posterior probability for search threshold.

After that, we implemented the STD tests using different N-best word lattices, which are created by Vietnamese speech recognition system (N=2 and N=3). The test results are described in Table. 3.

TABLE III. ATWV OF THE STD TEST WITH THE DIFFERENT THRESHOLD (USING DIFFERENT N-BEST WORD LATTICES CREATED BY VIETNAMESE SPEECH RECOGNITION SYSTEM, N=2 AND N=3)

Threshold	N=2	N=3
Acoustic score normalized by length of search term	0.5805	0.4076
Posterior probability normalized by length of search term	0.5982	0.4382

From the results in the Table 3, we can see that the performance (ATWV) of the system descent when the value of N increased. We can explain these results as the following: when N is increased, the number of word hypotheses also increase (missed probability decreased), but the wrong hypotheses (false alarm hypotheses) also increased so much so that ATWV decreased.

IV. CONCLUSIONS

In this paper, we have presented our results of the construction of spoken term detection for Vietnamese speech. For our results, optimum ATWV is 0.5982 (probability of false alarm is 0.48% and probability of missed detection is 20.85%), corresponding with the test using normalized posterior probability as threshold.

In the future research, we will research methods of integrating Vietnamese speech characteristics like tonal information to the STD system.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Education and Training of Vietnam, Edx corporation (Vietnamese company) for supporting this research.

References

[1] Jonathan Fiscus, Jérôme Ajot, George Doddington, Spoken Term Detection Evaluation Overview, 2006 Spoken Term Detection Workshop, December 14-15, 2006 http://www.nist.gov/speech/tests/std

[2] Nguyen Hong Quang, Trinh Van Loan, Le The Dat, Automatic Speech Recognition for Vietnamese using HTK system, RIVF 2010, Hanoi, Vietnam, November, 2010

[3] Quan VU, et al., "A Spoken Information Based Approach for the Retrieval of Soccer Video Events", Speech Technologies, Book 2, Intech Open Access Publisher, 2011

[4] Quan VU, et al., "Advances in Acoustic Modeling for Vietnamese LVCSR", IALP/IEEEE 2009, Singapore

[5] Igor Szöke, Michal Fapšo, Martin Karafiát, Lukáš Burget, František Grézl, Petr Schwarz, Ondřej Glembek, Pavel Matějka, Stanislav Kontár, Honza Černocký : NIST STD 2006 workshop, Gaithersburg, United Stated of America, December 14-15, 2006

[6] Olivier Siohan, Bhuvana Ramabhadran, "The IBM 2006, Spoken Term Detection System", STD 2006 Evaluation Workshop, Gaithersburg, United Stated of America, December 14-15, 2006

[7] David R. H. Miller, "Rapid and Accurate Spoken Term Detection, BBN Technologies", STD 2006 Evaluation Workshop, Gaithersburg, United Stated of America, December 14-15, 2006

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

[9] 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.

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

[11] X. Huang, A. Acero, H. Hon, "Spoken Language Processing, A Guide to Theory, Algorithm, and System Development", Carnegie Mellon University, 2001

[12] Dimitra Vergyri, Izhak Shafran, Andreas Stolcke, Ramana R. Gadde, Murat Akbacak, Brian Roark, Wen Wang, The SRI/OGI 2006 Spoken Term Detection System, EUROSPEECH 2007, Antwerp, Belgium, 2007