A Novel Approach in Continuous Speech Recognition for Vietnamese, an isolating tonal language

NGUYEN Hong Quang$^{1,2}$, Pascal NOCERA$^1$, Eric CASTELLI$^2$, TRINH Van Loan$^2$

$^1$Laboratoire d’Informatique d’Avignon (LIA), University of Avignon, FRANCE
$^2$International Research Center MICA, Hanoi University of Technology, VIETNAM

Abstract
This paper proposes a new approach for the integration of the Vietnamese language characteristics into a Large Vocabulary Continuous Speech Recognition System (LVCSR) which was built for some European languages. Firstly, a new module of tone recognition using Hidden Markov model was constructed. Secondly, several methods were applied to transform a text corpus of monosyllabic words into text corpus of polysyllabic words and a statistical language model of polysyllabic words was built by using the new text corpus. Finally, all the knowledge has been included in the LVCSR system so that this system can be adapted for Vietnamese. Experiments are made on the VNSPEECHCORPUS. The results show that the accuracy of Vietnamese recognition system was increased, 46% of relative reduction of the word error rate is obtained by using Vietnamese language characteristics.

Index Terms: speech recognition, Vietnamese speech, tone recognition, statistic language model, polysyllabic words

1. Introduction
Vietnamese belongs to the Viet-Muong group on the Mon-Khmer branch of the Austro-Asiatic language group. According to the linguists’ opinion, Vietnamese is an isolating tonal language. Its main properties are the following:
- Vietnamese has special units, called word units (WU). These units constitute sometimes a syllable, a morpheme or a single word. However, WUs are always separated by spaces in written Vietnamese. For simplicity, we can consider WU as a “Vietnamese syllable” or “syllable” for short. Nevertheless there are also polysyllabic words composed of several WUs (mainly a sequence of two, three or four syllables).
- Vietnamese is also a tonal language with six tones. In every syllable, there is one and only one tone. The tone is very important to decode the meaning of a word. If two similar monosyllabic-words have different tones, their meaning is different.

Previous research works [10] show that speech recognition techniques validated for Western languages like English, French, etc. could not reach the same results if they are directly applied on Vietnamese because characteristics of Vietnamese language have to be taken into account in the recognition process to improve the results.

First, in the language modeling level, the word segmentation is different between Vietnamese and Western languages. English and French are polysyllabic languages and words are easily identifiable using spaces. However, spaces are not always a correct mark to segment a Vietnamese sentence into words.

Since words were used as base unit in natural language processing systems therefore the word segmentation is a very important for isolating language such as Mandarin Chinese, Thai language as well as Vietnamese. The use of polysyllabic words in a Vietnamese speech recognition system was already used in [7] where the maximum-matching word segmentation was used for word segmentation in the text corpus. In our paper, we present a new method to transform a text corpus of syllables into polysyllabic words.

At the acoustic level, the main difference is the importance of tones in speech. Integrating tone information into a speech recognition system for Vietnamese has attracted several researches in the recent years. On isolated speech, a module of base syllable recognition (base syllable is a syllable without tone) was constructed, then tone was found by an isolated module of tone recognition [1]. There are also some works for continuous speech recognition. In [3], the phonological transcription of words in the pronunciation dictionary contained only phonemes without tone, and the correct transcription (with tone) was chosen by the language model, like for the homophones in French. In [6], relative tone information (the fundamental frequency F0) was integrated in the acoustic vectors and the system used the models of tonal phonemes. For the method presented in [7], tones were used like special phonemes beside other phonemes without tone.

The above methods did not use all tone characteristics: there is only one tone in every syllable and tones are differentiated by the form of the F0 contour of the syllable. In this paper, we investigate a new method for integrating tone information and the statistical language model of polysyllabic words into a large vocabulary continuous speech recognition (LVCSR) system for Vietnamese.

Our paper is organized as follows:
- In Section 2, construction method of the tone recognition module using Hidden Markov Models (HMM) is described and in Section 3, word segmentation methods are investigated.
- In Section 4, our method using tone information in a LVCSR for Vietnamese is presented.
- The Section 5 describes the experimental framework and results of tone recognition and speech recognition for Vietnamese.
- Finally, conclusion and future research are given in Section 6.

2. Vietnamese Tone Recognition
In our approach, a tone recognition system is built using Hidden Markov Model. First, given the fundamental frequency and the
short-time energy at frame $t$ ($F_0(t)$ and $E(t)$), we normalize
these values by using the following equations:
$$ f_0(t) = \frac{\log F_0(t) - \log \bar{F}_0(t)}{\sigma_{\log F_0(t)}} $$
$$ e(t) = \frac{\log E(t) - \log \bar{E}(t)}{\sigma_{\log E(t)}} $$

$\bar{F}_0(t)$, $\sigma_{\log F_0(t)}$ and $\bar{E}(t)$, $\sigma_{\log E(t)}$ are the mean and the
standard deviation of $\log F_0(t)$ and $\log E(t)$ on each sentence.

Then, each speech signal frame is represented by a 6 feature vector:
$[f_0(t), \Delta f_0(t), \Delta \Delta f_0(t), e(t), \Delta e(t), \Delta \Delta e(t)]$

A context-independent model is used; each tone is modelled by
an HMM. With a sample $X$ in the test corpus (one speech signal
segment of a tone in the test corpus), the probabilities on the
HMM of tones are calculated as follow (3).
$$ P(T_i \mid X) = P(X \mid T_i) \cdot P(T_i) / P(X) $$

$P(T_i)$ is calculated on the training corpus with (4).
$$ P(T_i) = \frac{\text{Number of tone } i \text{ in the training corpus}}{\text{Total number of tones in the training corpus}} $$

Using logarithmic transformation of these probabilities, we
have (5).
$$ \log P(T_i \mid X) = \log P(X \mid T_i) + \log P(T_i) - \log P(X) $$

Because the likelihood (LL) of HMM of tone $i$ is substituted for
$P(X \mid T_i)$, the tone result is identified by (6).
$$ \hat{T} = \arg \max_{i=1, \ldots, 6} \left[ \log LL(X \mid \text{HMM}_i) + \theta \cdot \log P(T_i) \right] $$

$\theta$ is the value corresponding to the best tone recognition result
on the training corpus.

3. Word Segmentation

To use polysyllabic words in a LVCSR for Vietnamese, we need
a polysyllabic lexicon and a text corpus of polysyllabic words.
Two methods are presented here to build such data: the first use
a given semantic lexicon and the second an automatic lexicon.

3.1. Use of a given semantic vocabulary

For this method, we suppose to have a lexicon of polysyllabic
words [2]. Then, we look for polysyllabic words from begin-
ning to the end of each sentence, syllable by syllable. In case of
ambiguity (there are several words beginning from the same
syllable), the longest grouping (MM: Maximum Matching) is
chosen.

3.2. Use of an automatic vocabulary

A new method to build automatic vocabulary of polysyllabic
words was implemented. This method uses a text corpus of syl-
lables and extracts a polysyllabic lexicon. The objective is to
extract and to isolate the most frequent sequences of syllables,
whether they have a semantic meaning or not. In that case, se-
quences can be a part of word, a word, either word group. In the
first step, the word segmentation of sentence is performed by
a dynamic programming algorithm with the syllabic language
model [4] (the duration of a polysyllabic word cannot exceed
4 syllables). In the second step, every list of syllables group is
analyzed and the most significant are retained (in term of mu-
tual information). The mutual information of syllables group is
calculated on the training text corpus and the mutual informa-
tion of the groups having a number of appearances lower than a
threshold K is set to 0. The K is selected to have the same word
number in the new vocabulary as in the semantic vocabulary.
This method is called DP_MI.

4. Structure of Vietnamese Continuous
Speech Recognition system

Diagram of LVCSR system for Vietnamese is outlined in Fig. 1.
Our system, adapted from the Speeral system [5] contains two
layers. In the first layer, the acoustic decoding uses the back-
ward Viterbi algorithm to create a phoneme lattice. In the sec-
ond layer, words hypotheses are generated by using an adapted
A* algorithm. The first frame, last frame and acoustic score of
each phoneme in the word are noted. For each word, the score of
language model, acoustic model and tone model are then cal-
culated. To identify the score of tone model, first, the boundary
of each syllable in the word is identified by using the pronun-
ciation dictionary, and then the tone of each syllable is detected
according to the syllable writing. Next, the voiced segment is
found based on the first and the last frame of the syllable; finally
the information is used by the tone recognition module to bring
tone score.

5. Experiments and results

5.1. Speech Corpus

5.1.1. The VNSPEECHCORPUS

VNSPEECHCORPUS speech corpus used in our tests is a read
speech corpus, recorded in a quiet studio [2]. There are two
types of text used during recording: paragraph (80%) and con-
version (20%). The speech data are digitalized in Wave format with 16 KHz sampling rate and A/D conversion precision of 16 bits. We use only records of standard dialect speakers (North of Vietnam) of eighteen speakers: 10 men and 8 women corresponding to approximately 14.4 hours of speech. For our study, we divide the corpus into two parts: 8 men and 6 women for the training corpus (11.2 hours of speech), 2 men and 2 women for the test corpus (3.2 hours of speech).

5.1.2. Tone corpus

To implement tone recognition experiments, we completed tone boundary description in the corpus. Because each syllable has a tone, we first identify the boundary for VNSPEECHCORPUS syllables, and then we find the tone boundary. Our approach is presented as follows:

- The syllable boundaries were aligned by using the LIA acoustic modeling toolkit [5].
- Then, the syllable boundaries were manually corrected with the help of our development tool using the Praat \(^1\) environment.

We decided to use the voiced segment of syllable as tone segment. This voiced segment was tagged based on the fundamental frequency \(F_0\) as follows: the beginning and ending points of the voiced segment are respectively the first point and the last point which present \(F_0\) values.

5.2. Vietnamese tone recognition experiment

5.2.1. \(F_0\) and energy identification

Fundamental frequency \(F_0\) is calculated using the CC (cross-correlation) algorithm in Praat tool. A small silence threshold is used in this algorithm, and then \(F_0\) is filtered by median smoothing and linear smoothing filter [9].

Another parameter generally used in a tone recognition system is the short-time energy. In our experiment, short-time energy is extracted every 0.01 second with the Praat tool.

5.2.2. Tone recognition experiment

In our tests, we use an HMM for each tone (6 HMMs for 6 tones). In each HMM, there are 3 emitting states of 16 Gaussian mixtures. HMMs are context-independent models. Each model is trained by using all speakers in the training corpus and speakers in the test corpus were used to calculate the tone recognition result (TAR: tone accuracy rate).

Tone recognition results are not homogeneous between male and female speakers. In this experiment we only used male speakers: 8 men for the training part and 2 men for the test. Using the training corpus and the test corpus described in section 5.1.1 we obtained a TAR of 75.80%.

5.3. Vietnamese LVCSR system framework

This section presents principal elements used by our large vocabulary continuous speech recognition system for Vietnamese.

5.3.1. Vietnamese Pronunciation Dictionary Creation

The structure of Vietnamese syllable is illustrated in the Table 1. Firstly, a vocabulary of 6698 syllables is extracted from 40000 full-words vocabulary [2]. Then a pronunciation dictionary for Vietnamese is built by applying VNPhone Analyzer [2] on this isolated-word vocabulary. At last, tone information in the syllable pronunciation is deleted. Thus, the tonal syllables which have the same base syllable would have the same pronunciation. The pronunciation form of a polysyllabic word is obtained by the concatenation of the pronunciation forms of the syllables in the word.

5.3.2. Text Corpus and Language Model

For an under-resourced language like Vietnamese, a method of constructing the text corpus and language model is presented in [4]. According to this method, the text corpus is built on the data extracted from Vietnamese electronic newspapers web sites. Next, superfluous data (menus, references, advertisements, announcements present in different pages) are removed from this database. Finally, all sentences containing only in-vocabulary words (monosyllabic words lexicon) are kept. In the filtered corpus, there are approximately 2.7 million sentences with 45 million monosyllabic words.

In order to train our statistical language model of syllables, we use the CMU SLM toolkit \(^2\) with Good-Turing discounting and Katz back-off for smoothing method and the perplexity in the test speech corpus is 72.74.

5.3.3. Acoustic model

We dealt with context-independent acoustic models for Vietnamese. So a hidden Markov model for each Vietnamese phoneme is used. We have chosen 5 state left-right topology with 3 emitting states, the entry and exit states of a HMM are non-emitting. Each emitting state consists of 64 Gaussian mixtures. The feature vectors are extracted each 10 ms and contain 39 dimensions of 13 MFCC, plus their first and second derivatives. This acoustic model was trained with all the data on the training corpus by using expectation-maximization (EM) algorithm.

For speaker adaptation, the phonetic alignment of the test corpus was firstly taken using the LVCSR system results with the speaker independent acoustic model on the test corpus. Next, in order to create a new model for each speaker, the common model was adapted with all the data of each speaker by using MLLR adaptation technique. Then, the LVCSR system was reused with the new model. This process was repeated several times to get the model which will be used in the speaker adaptation experiments [10].

5.4. Vietnamese LVCSR experiment

During the linguistic decoding, the score of tonal syllable is inserted to define the score of each word as in (7).

\[
\text{Score}_{\text{word}} = \alpha \cdot \text{Score}_{\text{ac}} + \beta \cdot \text{Score}_{\text{ML}} + \gamma \cdot \text{Score}_{\text{tone}} \quad (7)
\]

\(\text{Score}_{\text{ac}}, \text{Score}_{\text{ML}}, \text{Score}_{\text{tone}}\) are the score of acoustic model, language model and tone model. These scores are logarithmic values of corresponding probabilities.

\(^1\)http://www.fon.hum.uva.nl/praat/

\(^2\)http://www.speech.cs.cmu.edu/SLM_info.html

<table>
<thead>
<tr>
<th>Table 1: Structure of Vietnamese syllable.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(22 consonants)</strong></td>
</tr>
<tr>
<td><strong>Medial</strong></td>
</tr>
<tr>
<td>(1 semivowels)</td>
</tr>
<tr>
<td><strong>Initial</strong></td>
</tr>
</tbody>
</table>
5.4.1. Baseline system

The context-independent acoustic model and the syllabic language model are used in this first experiment. The tone model is not used ($\gamma = 0$), so the tone of syllable is decided only by the syllabic language model (which is named LM_{SL}). $\alpha, \beta$ are selected empirically to obtain the best recognition results.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>SER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male speakers</td>
<td>34.6</td>
</tr>
<tr>
<td>Female speakers</td>
<td>34.9</td>
</tr>
<tr>
<td>All speakers</td>
<td>34.7</td>
</tr>
</tbody>
</table>

Three results are presented in Table 2: on the male speakers, on the female speakers and on all speakers.

5.4.2. Using tone information

This experiment uses the tone model in recognition system. In this case, tone information and linguistic information contribute together to the tone of syllable.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Acoustic model</th>
<th>Speaker independent</th>
<th>Speaker adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM_{SL} without tones</td>
<td>34.6</td>
<td>21.3</td>
<td></td>
</tr>
<tr>
<td>LM_{MM} without tones</td>
<td>29.3</td>
<td>18.6</td>
<td></td>
</tr>
<tr>
<td>LM_{DP,MI} without tones</td>
<td>25.6</td>
<td>16.0</td>
<td></td>
</tr>
<tr>
<td>LM_{SL} with tones</td>
<td>24.7</td>
<td>13.5</td>
<td></td>
</tr>
<tr>
<td>LM_{MM} with tones</td>
<td>21.2</td>
<td>12.4</td>
<td></td>
</tr>
<tr>
<td>LM_{DP,MI} with tones</td>
<td>18.7</td>
<td>10.7</td>
<td></td>
</tr>
</tbody>
</table>

Because we use the tone model described in section 5.2.2, we only use male speakers. $\alpha$ and $\beta$ of the baseline system are used and the experiments with different $\gamma$ are tested. The best results are described in the Table 3.

5.4.3. Using language model of polysyllabic words

The MM algorithm and DP_{MI} algorithm (section 3.1 and 3.2) are used to transform the text corpus of syllables (section 5.3.2) into polysyllabic words text corpus. The semantic lexicon has 40342 polysyllabic words and the automatic lexicon has 39823 polysyllabic words. CMU SLM toolkit is applied on the new corpus to create new language models of polysyllabic words. These language models are called LM_{MM} and LM_{DP,MI}.

The results in the table 3 showed that the use of polysyllabic words and tone information increased the performance of system, 46% relative reduction in syllable error rate. In the test corpus, there are 20% conversation and 80% paragraph. By using polysyllabic language model, the result on the paragraph is changed more significantly than on the conversation (in the table 4, SER is 33.0% in the case of LM_{SL} and only 32.6% with LM_{DP,MI}). There are two reasons for that. The first is that polysyllabic words were extracted from electronic newspapers more linguistically near paragraphs corpus than conversation corpus. The second is that there are many short phrases in the conversations, which are even composed of only two or three words. In this case, polysyllabic language model has no much influence. Otherwise, the results on the tests with tone information increased in both cases: paragraph and conversation.

6. Conclusions

In this paper, we have presented our study on the use of tone information and polysyllabic words in a large vocabulary continuous speech recognition system for Vietnamese. The result with our word segmentation method (DP_{MI}) is better than the one with MM algorithm. We realize that by using polysyllabic language model, the result on the paragraph changed significantly than on the conversation. Otherwise, the results on the experiments with tone information are enhanced in all cases. Finally the use of polysyllabic words and tone information increase the performance of system of 46% relative reduction in syllable error rate. For the future research, context-dependent tone models will be done and will be integrated in the LVCSR system for Vietnamese.

7. References