Tone Recognition of Vietnamese Continuous Speech using Hidden Markov Model

NGUYEN Hong Quang*†, NOCERA Pascal*, CASTELLI Eric† and TRINH Van Loan†
*Laboratoire Informatique dAvignon LIA, UAPV, Avignon, France
Email: (Quang.Nguyen, Pascal.Nocera)@univ-avignon.fr
†International Research Center MICA, HUT - UMI2954/CNRS - INP Grenoble, Hanoi, Vietnam
Email: (Hong-Quang.Nguyen, Eric.Castelli, Van-Loan.Trinh)@mica.edu.vn

***Abstract*—This paper presents our study on context-independent tone recognition of Vietnamese continuous speech. Each of the six Vietnamese tones is represented by a hidden Markov model (HMM for short) and we used VNSpeechCorpus to learn these models in terms of fundamental frequency, F_0, and short-time energy. We focus on evaluating the influence of different factors on the tone recognition. The experimental results show that the best method to learn F_0 and energy is to use a logarithmic transformation function and then normalization with mean and mean deviation. In addition, we show that using 8 forms of tones and the discrimination between male and female speakers increase the accuracy of the Vietnamese tone recognition system.**

***Keywords*—Vietnamese speech, tone recognition, pitch normalization of tone, energy normalization, pitch contour**

I. Introduction

Vietnamese language belongs to the Viet-Muong group, on the Mn-Khmer branch of the Austro-Asiatic language family. From linguist's points of view, Vietnamese is a syllabic tonal language with six lexical tones. In every syllable, there is one and only one tone. The tone is very important to decide the meaning of a word. If two similar monosyllabic-words have different tones, they have different meanings. Therefore, accurate tone recognition is essential for processing of tonal languages like Vietnamese, Mandarin, Thai, etc.

Vietnamese speech processing is only at its early stage of development. Some previous works addressed only isolated word [1], but not continuous speech. For isolated speech, syllables are pronounced singly and clearly, therefore the form of the F_0 contour of each tone doesn't change significantly and almost holds the canonical form (Fig. 1). However, in case of continuous speech, the F_0 contour of the tone is affected by many factors as sentence prosody, tone co-articulation, speaker's emotion, etc.

Although there are not many researches on continuous Vietnamese speech, there are a lot of works on other tonal languages like Mandarin [2], Cantonese [3] or Thai [4]. In tone recognition theory, two main approaches exist: in the first approach, each frame of the signal is represented by a vector and the HMM technique is used; in the second approach, more global, each whole tone is represented by a vector and ANN (Artificial Neural Network), GMM (Gaussian Mixture Models), SVM (Support Vector Machines) or decision tree techniques are used.

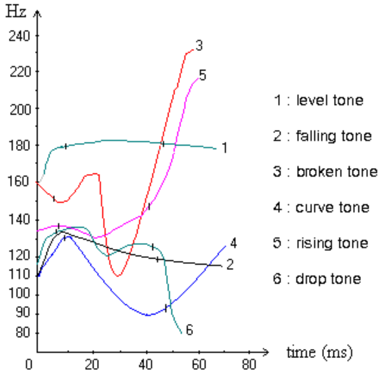

Fig. 1. The canonical form of the F_0 contour of the 6 Vietnamese tones

The phenomenon which seriously affects the F_0 contour of the tones is tone co-articulation. There are two major techniques to avoid this effect: using a contextual model of tone and using a context-independent model of tone while rejecting the segment influenced by the co-articulation effect.

In this paper, we first investigate normalization methods of F_0 and energy which help to reduce the influences of elements affecting the tones in continuous speech. In our experiments, we use HMM for tone modeling. Then, we apply a simple method to overcome the tone co-articulation effect. Lastly, we carry out a survey on the impact of using 8 tone forms instead of 6 and the discrimination between male and female speakers on tone recognition.

The rest of this paper is organized as follows: in section 2, we describe the VNSpeechCorpus and a method to create tone corpus used in our experiments and evaluations; in section 3, our method to determine the fundamental frequency F_0 and the short-time energy is presented; in section 4, the tone recognition experiments are described; finally, conclusion and future research are given in section 5.

II. Speech Corpus

A. The VNSpeechCorpus

VNSpeechCorpus speech corpus used in our tests is a read speech corpus, recorded in a quiet studio [5]. There are two types of text: paragraph (80%) and conversation (20%). We use only the records of standard dialect speakers (North of Vietnam) with eighteen speakers: 10 men and 8 women corresponding to approximately 14.4 hours of speech. For our study, we divided 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).

B. Tone corpus

To realize tone recognition experiments, we completed the corpus with tone boundary description. Because each syllable has one tone, we first identify the boundary for VNSpeech-Corpus syllables, and then we can discover the tone boundary. Our approach is presented as follows:

- The syllable boundaries were aligned by using the LIA acoustic modeling toolkit [6].
- Then, the tone boundaries were manually corrected with the help of our tool developed using the Praat[1] environment.

We decided to use the voiced segment of the 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 and last points which present F_0 values.

III. F_0 IDENTIFICATION

In recent years, many F_0 identification algorithms have been presented. We decided to use Praat software to calculate the F_0. Praat is a well known tool which integrates some popular algorithms such as autocorrelation algorithm (AC), cross correlation algorithm (CC), etc. Users can select the input parameters as $minF_0$, $maxF_0$, $silence\ threshold$, etc.

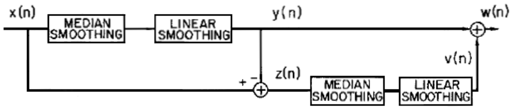

Fig. 2. Block diagrams of smoother and double-smoothing algorithms

Studying F_0 values calculated by AC and CC algorithms using Praat tool, we have realized that, AC algorithm doesnt calculate the F_0 around the broken point of tone 3 (broken tone) and at the end of drop tone (tone 6) while CC algorithm is able to calculate these points with a small silence threshold but makes an error by creating other points on some unvoiced segments. That is why we chose to use the CC algorithm with a small silence threshold and to suppress the false points using median smoothing and linear smoothing (Fig. 2 [8]). Based on our experience, we chose 3 points for the width of the median smoothing and the first linear smoothing (Hanning); and 5 points for the width of the second linear smoothing (Hanning).

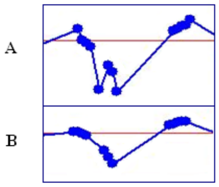

Fig. 3. F_0 contour of drop tone (tone 6) calculated by CC algorithm of Praat tool (A) before and (B) after filter the false points by double-smoothing algorithm.

To examine the above method, we manually calculated some F_0 contours and compared them with contours obtained by our proposed method. We chose 6 files of a speaker (a female speaker of Hanoi Television) which contains all tones, especially the tones 3 and 6. The F_0 contour automatically calculated will be accepted if the difference between this value and the manual value is less than 10 Hz. The results of these methods are presented in Table I.

TABLE I THE ACCURACY OF THE F_0 CALCULATED METHODS

	Tests					
Test's name	AC1	AC2	AC3	AC4	CC1	CC2
minF0 (Hz)	75	100	100	100	100	100
maxF0 (Hz)	600	400	400	400	400	400
silence threshold	0.03	0.03	0.02	0.02	0.02	0.02
using the filtrage	no	no	no	yes	no	yes
	F_0 **rate (in %)**					
F_0 correct rate	88.4	89.5	91.1	90.9	90.2	91.0
F_0 incorrect rate	1.1	0.8	0.8	1.0	2.5	1.8
F_0 deletion rate	9.9	9.1	6.8	6.9	4.3	4.3
F_0 insertion rate	0.7	0.6	1.3	1.3	3.1	3.0

The default parameters of Praat.were used in the AC1 test and the standard values of a female speaker was used in the AC2 test (100 Hz with minF0 and 400 Hz with maxF0). We noted that all results in the AC2 test were better than ones of AC1 test. When we compared the output of the AC algorithm (AC3 test) to the output of the CC algorithm (CC1 test), we noted that F_0 deletion rate is smaller for CC algorithm. In addition, the incorrect and insertion rates of F_0 detection are also reduced with the method based on CC algorithm (CC2 test).

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

IV. VIETNAMESE TONE RECOGNITION

Based on the fundamental frequency F_0 contour and short-time energy, we built a tone recognition system using HMMs.

A. Tone model

In our tests, we used an HMM for each tone (6 HMMs for 6 tones). In each HMM, there are 3 emitting states and each emitting state consists of 16 Gaussian mixtures. HMMs are context-independent models. Each model was 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). We used the training corpus and the test corpus described in section II-A With a sample X in

[1] http ://www.fon.hum.uva.nl/praat/

the test corpus (one speech signal segment of a tone in the test corpus), we calculated the probability on the HMMs of tones. The resulting tone was identified following (1).

$$\widehat{T} = arg \underbrace{max}_{i=1\to6} P(T_i|X) = arg \underbrace{max}_{i=1\to6} P(X|T_i).P(T_i) \quad (1)$$

P(Ti) was calculated using training corpus following (2).

$$P(T_i) = \frac{\text{number of Tone } \boldsymbol{i} \text{ in the training corpus}}{\text{Total number of tone in the training corpus}} \quad (2)$$

According to our experience, tone recognition results are not homogeneous between male and female speakers. In the next experiments we only used male speakers: 8 men in the training part and 2 men in the test part. We will present in the section IV-E the influence of gender on the tone recognition.

B. Influence of F_0 normalization

In continuous speech, F_0 contour of tone is affected by many factors. Therefore, F_0 normalization is necessary, even between sentences of one speaker. The following F_0 normalization methods are proposed:

- F_0 normalization by minF0 and maxF0 of each sentence (NORM_F0_MIN_MAX).

$$f_0(t) = \frac{F_0(t) - minF_0}{maxF_0 - minF_0} \quad (3)$$

- Using logarithm of F_0 value and normalizing this logarithmic value of F_0 by **min** and **max** of each sentence (NORM_LOG_F0_MIN_MAX).

$$f_0(t) = \frac{logF_0(t) - min\ logF_0}{max\ logF_0 - min\ logF_0} \quad (4)$$

- F_0 normalization by F_0 mean of each sentence (NORM_F0_MEAN).

$$f_0(t) = F_0(t)\ /\ \overline{F_0} \quad (5)$$

- Using logarithm of F0 value and normalizing this logarithmic value of F_0 by mean and max of each sentence (NORM_LOG_F0_MEAN).

$$f_0(t) = logF_0(t)\ /\ \overline{logF_0} \quad (6)$$

- F_0 normalization by mean and standard deviation of the F_0 of each sentence (NORM_F0_MEAN_DEV).

$$f_0(t) = \frac{F_0(t) - \overline{F_0}}{\delta_{F_0}} \quad (7)$$

- Using logarithmic value of F_0 and normalizing this new value by mean and standard deviation of each sentence (NORM_LOG_F0_MEAN_DEV).

$$f_0(t) = \frac{logF_0(t) - \overline{logF_0}}{\delta_{logF_0}} \quad (8)$$

We then performed a series of experiments to evaluate the influences of the above F0 normalization methods. Each speech signal frame was represented by a vector with 3 features $[f_0(t), \Delta f_0(t), \Delta\Delta f_0(t)]$. The experiment results are given in Table II.

TABLE II THE TONE RECOGNITION RESULTS WITH DIFFERENT F_0 NORMALIZATION METHODS

Method	TAR (tone accuracy rate %)
Without normalisation	63.20
NORM_F0_MIN_MAX	60.74
NORM_LOG_F0_MIN_MAX	61.57
NORM_F0_MEAN	67.92
NORM_LOG_F0_MEAN	67.85
NORM_F0_MEAN_DEV	70.01
NORM_LOG_F0_MEAN_DEV	***70.44***

As the experimental data show, F0 normalization by ***mean*** and ***standard deviation*** of the F_0 of each sentence gave better result than the others and to transform F_0 value into a logarithmic value improves slightly this result. So, we chose the NORM_LOG_F0_MEAN_DEV F_0 normalization method for the next experiments.

C. Influence of short-time energy

In order to characterize short-time energy influence, we added this parameter in the feature vector. This short-time energy parameter is calculated by using the (9).

$$e(t) = \frac{logE(t) - \overline{logE(t)}}{\delta_{logE(t)}} \quad (9)$$

where $E(t)$ is the short-time energy in the 0.01 second window, $\overline{logE(t)}$ and $\delta_{logE(t)}$ are the mean and standard deviation of the $logE(t)$ in a sentence.

Then each speech signal frame was represented by a 6 feature vector: $[f_0(t), \Delta f_0(t), \Delta\Delta f_0(t), e(t), \Delta e(t), \Delta\Delta e(t)]$, where $f_0(t)$ is the $F_0(t)$ normalized by NORM_LOG_F0 MEAN_DEV method. The tone recognition result (TAR was 75.80%).

D. Effect of speaker's gender

Three experiments were executed to evaluate the effect on the tone recognition accuracy of the speaker's gender:

- we used only male speakers: 8 speakers in the training part and 2 speakers in the test part;
- we used only female speakers: 6 speakers in the training part and 2 speakers in the test part.
- all speakers: 14 speakers in the training part and 4 speakers in the test part.

The corresponding results are presented in Table III. The experimental data show that the recognition accuracy of tone 3 is not homogeneous between male speakers (TAR = 19.9%) and female speakers (TAR = 47.4%) and that the recognition rate of tone 3 is globally poor (TAR = 32.8%). The same phenomenon also appears for tone 4.

Looking at the F_0 contours of these two tones, we can explain those results. In tone 3, two-thirds of the contour is characterized by an abrupt dip caused by a heavy laryngealization [7]. This segment is very clear for female speakers but not for male ones. In tone 4, the low onset falls further gradually until the point at two-thirds of contour from the onset. From this point, the extremely low F0 starts to rise toward the end

TABLE III TONE RECOGNITION RATE (TAR) FOLLOW SPEAKER'S GENDER CHARACTERISTICS

Tone	Men	Women	Men and Women
Tone 1 TAR (%)	95.4	94.7	95.1
Tone 2 TAR (%)	77.1	74.4	76.5
Tone 3 TAR (%)	**19.9**	**47.4**	**32.8**
Tone 4 TAR (%)	**62.0**	**27.6**	**41.4**
Tone 5 TAR (%)	81.5	76.6	77.9
Tone 6 TAR (%)	45.9	37.7	42.2
Average TAR (%)	**75.80**	**70.41**	**72.83**

[7]. This segment is very clear for male speakers but not for female ones. In addition, the slope of the first segment of male speaker is bigger than for the female speaker. So the tone recognition is affected by speakers gender, especially on tone 3 and tone 4.

E. Influence of tone co-articulation

Tone co-articulation effects are a common phenomenon appearing in continuous speech. More specifically, the F_0 contour of a tone is generally affected by its left tone and its right neighbours: the F_0 contour could be then very different from the canonical form of F_0 contour obtained in isolated mode. There are two major tone co-articulation effects: carry-over effect (influenced by the left tone) and anticipatory effect (influenced by the right tone). Several researches on Mandarin and Vietnamese have showed that the carry-over effect is stronger than the anticipatory effect. In some cases, even 50% of the tone from its beginning is influenced by the carry-over effect [7].

TABLE IV THE TONE RECOGNITION ACCURACY WITHOUT K% OF TONE FROM ITS BEGINNING

K	**TAR (tone accuracy rate %)**
0	75.80
10	76.44
15	76.22
20	76.82
25	***77.02***
30	76.88
35	76.64
40	77.00
45	76.37
50	75.62

A simple approach to reduce the influence of tone co-articulation was presented for Mandarin language in [2]. The method does not use segments influenced by tone co-articulation, but segments containing the most critical information for tonal perception. We propose here a similar approach: we don't use K% of tone from its beginning (we only use the remaining part for tone recognition). The experiments with different values of K are given in table IV.

According to the experimental data, we can see that tone recognition accuracy in the above experiments is better than using all voiced segments. In addition, K = 25% give the best accuracy (TAR = 77.02%). We chose this value for our next experiments.

F. Experiments on 8 tone forms

In Vietnamese, almost of syllables can appear with all tones except syllables ending with stop consonants (/p/, /t/, /k/) which only appear with tone 5 or tone 6. These syllables are shorter than the other syllables and the F_0 contours rise or drop more sharply. Therefore, most linguists consider that there are 8 tone forms in Vietnamese : tone 1, tone 2, tone 3, tone 4, tone 5a and tone 6a for syllables ending with voiced phoneme, tone 5b and tone 6b for syllables ending with stop consonants [1]. We have realized experiments to evaluate the influence of 8 tone forms on tone recognition accuracy.

These experiments only used the voice of male speakers. We used 8 HMMs for 8 tone forms. We considered that we knew characteristics of the phoneme ending so the result tone was found as follows: if the syllable ending with voiced phoneme, we only used 6 HMMs of tone 1, 2, 3, 4, 5a, 6a; otherwise, if the syllable ending with stop consonants, we only used 2 HMMs of tone 5b and 6b. The result of the experiment (TAR) was 81.02%. If we used the knowledge about syllable ending in the 6 tone recognition system of the previous section, we obtained only 79.62%.

V. CONCLUSION

In this paper, we have described some experiments on Vietnamese tone recognition in continuous speech. The experimental results show that applying logarithmic transformation function followed by normalization with mean and mean deviation is the best method for F0 and energy. We have also shown that the recognition accuracy of tone 3 and tone 4 is not homogeneous between male speakers and female speakers. In addition, as our experiment showed, with added information of the syllable phoneme ending, using 8 form tones is better than one using 6 form tones. For future research, we will integrate the tone recognition module in an automatic speech recognition system to improve its performance.

REFERENCES

[1] Q.C. Nguyen, N.Y. Pham, and E. Castelli, *Shape vector characterization of Vietnamese tones and application to automatic recognition*, ASRU 2001, Trento, Italy, 2001.

[2] J. Zhang and K. Hirose, *Tone nucleus modeling for chinese lexical tone recognition*, Speech Communication, vol. 42, pp. 447466, 2004.

[3] Y. Qian, F.K. Soong, and T. Lee, *Tone-enhanced generalized character posterior probability (gcpp) for cantonese LVCSR*, ICASSP 2006, Toulouse, France, 2006.

[4] L. Tan, M. Karnjanadecha, and T. Khaorapapong, *A study of tone classification for continuous Thai speech recognition*, INTERSPEECH 2004, Jeju Island, Korea, 2004.

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

[6] P. Nocera, G. Linares, D. Massoni, and L. Lefort, *Phoneme lattice based a search algorithm for speech recognition*, TSD 2002, Brno, Czech Republic, 2002.

[7] 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.

[8] L.R. Rabiner, M.R. Sambur, and C.E. Schmidt, *Applications of a nonlinear smoothing algorithm to speech processing*, IEEE Transactions on acoustics, speech, and signal processing, vol. ASSP-23, pp. 552557, 1975.