SHAPE VECTOR CHARACTERIZATION OF VIETNAMESE TONES AND APPLICATION TO AUTOMATIC RECOGNITION

Nguyen Quoc-Cuong[1], *Pham Thi Ngoc Yên*[2], *Eric Castelli*[1]

[1]Laboratoire CLIPS-IMAG, équipe GEOD, Univ. Joseph Fourier B.P. 53, 38041 Grenoble cedex 9, France
Quoc-Cuong.Nguyen@imag.fr Eric.Castelli@imag.fr
[2]Faculty of Electricity- Hà Nôi University of Technology - 1 Dai Cô Viêt – Hà Nôi
ptnyen@vn.refer.org

ABSTRACT

In this paper, the tone recognition for Vietnamese standard language (Hanoi dialect) is described. The wavelet method is used to extract the pitch (F0) from a speech signal corpus. Thus, one feature vector for tone recognition of Vietnamese is proposed. Hidden Markov Models (HMMs) are then used to recognize the tones. Our results show that tone recognition seems independent of the vowel but presents better accuracy if one of both monotonous tones is used as pitch reference base. Finally, a first try of a complete isolated word recognition engine, adapted for Vietnamese, is presented.

1. INTRODUCTION

Very few studies have been carried out on the Vietnamese language, for the last 20 years. We can only quote few works as the most interesting on this field: studies of Doan on Vietnamese acoustic [1] and Han & Kim's work on Vietnamese tone variations [2]. Recently, appears the study of Vu on the characterization of Vietnamese tones [3].

Thus, at the level of the automatic recognition of this language, as far as we know no system have been built for Vietnamese and we think that recognition of Vietnamese is still an unresolved problem. However, a lot of studies were succeeded on other Asian languages, like Chinese (Mandarin) [4] [5] or Thai [6]. Our work about Vietnamese recognition has been derived from the techniques for Mandarin recognition and this paper presents our first results.

Like the Chinese language from which it derives partly, the Vietnamese is a mono and bisyllabic tonal language. It presents 16 vowels, 21 consonants and 2 semi-vowels. Generally, each isolated syllable is pronounced in one of the six tones, specified in Table 1, which can bring out six different words. However, several combinations do not exist, such as, for example, the association "ân + tone4" which does not have linguistic significance. Each syllable can be considered as combination of initial and final parts. Doan [1] proposed the syllable structure as follows:

$$\underbrace{[\text{ initial consonant }]}_{\text{initial}}\ \underbrace{[\text{ pretonal }]\,\text{vowel}\,[\text{ final consonant }]}_{\text{final}}$$

Initial consonant, pretonal and final consonant can be optional, i.e. could not exist in a word. Doan [1] and Han & Kim [2] studies have shown that tone information is superimposed on the final part of each syllable.

As in Chinese case, Vietnamese recognition can be divided into two parallel procedures, which are the recognition of tones and the recognition of syllables disregarding the tones. This paper focuses on our solution in tone recognition for Vietnamese. First, section 2 will present the corpus, the method using to detect the pitch and some comments of tones. Section 3 presents a series of experiment for tone recognition using HMMs technique, in order to evaluate the vowel nature dependency and its effect to recognition accuracy. After applied the feature vector used for Mandarin recognition, a new feature vector, specially adapted for Vietnamese is proposed. Finally, conclusions are given in the section 5.

Vietnamese tones	pattern
tone1	level tone
tone2	rising tone
tone3	broken tone
tone4	falling tone
tone5	curve tone
tone6	drop tone

Table 1. The six Vietnamese tones

2. CORPUS AND PITCH DETECTION

We have realized a speech material corpus with three aims:

- characterization of Vietnamese tones.
- study of tone recognition for Vietnamese.
- Vietnamese speech recognition in isolated word context applying into the vocal commands for simple process.

To carry out these objectives, the corpus was conceived to have items of each of the 16 vowels combined with the 6 tones, and presents number words and control command words for internet applications (for example “file”, “open”, “close”, etc…). We thus choose 135 words with 131 monosyllabic and 4 bisyllabic words.

Each syllable is pronounced 4 times in isolated word mode by 18 different speakers from North (Hanoi district), Central (Hue and Da Nang district) and South (Ho Chi Minh and Can Tho district): 6 females and 2 males from North, 2 females and 2 males from Central, 3 females and 3 males from South. We also point out that the pronunciation of Vietnamese is different in north in comparison from the south.

Thus, we have recorded a set of (4*135*18) syllables representing totally 9720 items for about 3 speaking hours. Speech signals are recorded at 16 kHz sampling rate, with an AD conversion precision of 16 bits. The analysis frame length and the frame shift was respectively 64 ms and 8 ms.

The method using Dyadic Wavelet Transform (D_yWT) is used to detect pitch period on Vietnamese speech signals [7][8]. Event based pitch detectors estimate the pitch period by locating the instant at which the glottis closes (event) and then measuring the time interval between two such glottal closures. The event is detected based on the assumption that during glottal closure, the vocal tract is strongly excited, that causes an abrupt change in a speech signal. We present in Figure 1 an example from ours results, consisting of six tones of a North female subject (PNY).

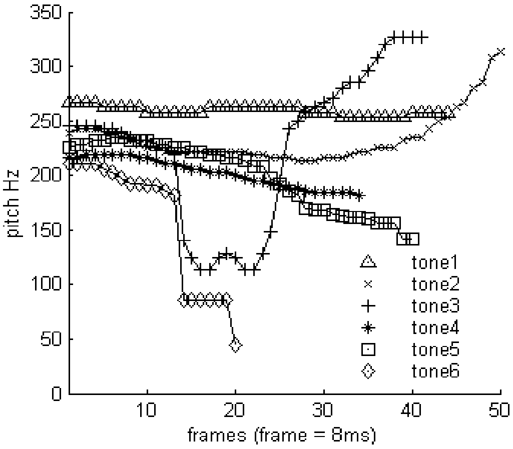

Figure 1. Example of pitch evolutions for six VN tones (female subject PNY)

In Vietnamese, two kind of syllable can be distinguished: open and closed syllable. Closed syllables are finished by one of the three consonants /p, t, k/ and could only be combined with ton5 and ton6. Open syllables, without end-consonant /p, t, k/, could be combined with all six tones. Thus, we distingue 8 representations of the six tones: ton1, ton2, ton3, ton4, ton5a, ton6a (ton5 and ton6 with open syllable), ton5b and ton6b (ton5 and ton6 with close syllable). In order to characterize the tones, we study each contour with general and simple shapes as shown on Figure 2, on which, start, middle and end measure points are quoted.

Results coming from tour study are given below:

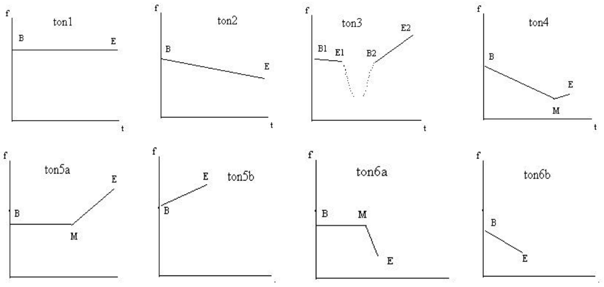

Figure 2. Simple shapes of each tones

- Duration of the tone6 is shorter than others. For example, the average duration of tone6 of the speaker PNY is 0.12 s, while the average duration of five tones 1,2,3,4,5 is between 0.26 s and 0.36 s.

- Onsets, (i.e. the initial point of final part), of three first tones are higher than the others: for the speaker PNY average onsets of tone1, tone2 and tone3 are 30 Hz higher than average onsets of the three others.

- Endpoints, (i.e. the final point of final part), of tone1, tone2 and tone3 are higher than the endpoints of tone4, tone5 and tone6. For the subject PNY, the average endpoint values of tone1, tone2, tone3 are greater than 245 Hz, while average endpoint values for tone4, tone5 and tone6 are lower than 175Hz.

- Endpoints of tone2 and tone3 are maximal of their pitch contour.

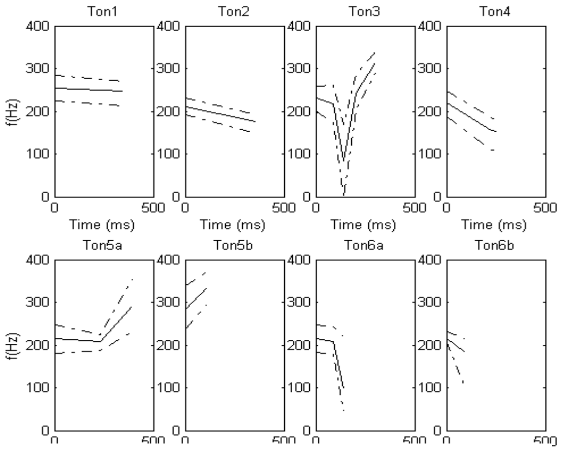

Figure 3. Tone shapes for the subject PNY

An example of shape characterization of the 8 representations of Vietnamese tones is given, for the subject PNY, on Figure 3, where average contours are printed in plain lines and standard deviations in dotted lines.

These results were very useful to help us improving the characterization of the vector used in our tone recognition method, as it will be shown it the paragraph 3.

3. TONE RECOGNITION

For our study of tone recognition of standard Vietnamese (Hanoi dialect), we have extracted from our corpus speech signals of 5 female subjects from Hanoi district. The database can be separated into 3 sets: training set, first and second tests set. Training set contains 8 vowels; each syllable is repeated 3 times, that constitute 1395 syllables. First test set contains only 8 vowels chosen among the 16 of the training set, but each syllable is repeated 1 times, in a total of 465 syllables. Second test set contains the 8 remaining vowels, each syllable is repeated 4 times, to present a set of 800 syllables.

In the literature, methods based on HMMs or neural network techniques are most often used to carry out the recognition of the tone. We choose to use the method using HMMs (with HTK toolkit [9]) where the idea consists in capturing the temporal variation of the pitch contour [4].

The implementation of HMMs includes two parts: the training and the classification phases. In training phase, each HMM is trained for each tone, i.e. there are 6 HMMs for 6 tones. At the classification phase, the test sequence of vector feature is given to all of these 6 HMMs. The tone is then identified by the HMM which gives the highest probability score.

The tone information is superimposed on the final part of each syllable. However, for the task of automatic tone recognition, the final part detection is more difficult when initial part is voiced: measuring the pitch, we have the same difficulties in detecting the limits between initial and final part as to differentiate a consonant voiced from the vowel which follows it. Thus, our solution consists using the pitch contour on the voiced part of the syllable, i.e. the initial and final parts together.

In the tone recognition for Mandarin, Yang et al. [4] defined the feature vector as follows:

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1})] \qquad (1)$$

where f_t is the pitch frequency at frame *t*.

First component represents local slope of the pitch frequency while the second is the level in the pitch frequency. In the same study, in order to achieve a tone recognition system that is independent of speakers, the pitch is normalized by the pitch reference base (bias) F_0 that is the average pitch of a specified tone (the constant one) depending on the speaker. Hence, each value f_t in the feature vector is replaced by the ratio (f_t / F_0 *)*. We also see that there is only second component of feature vector is affected by F_0.

In Vietnamese, the tone1 is monotone from the high register, and the tone4 is also monotone but from the low register. The reference base F_0 can be chosen between the mean value of tone1 and tone4 for ours experiments. In a first step, we proceed to 3 experiments in order to validate this Mandarin feature vector for Vietnamese tone recognition and to determine the best mean value for F_0:

Experiment 1: using the unbiased feature vector with form defined for Mandarin by equation (1).

Experiment 2: using the same feature vector biased by mean tone1;

Experiment 3: using feature vector biased by tone4.

Use of the pitch reference base is necessary. For vectors defined by (1), the recognition accuracy with biased vectors in the experiments 2 and 3 is higher than in the experiment 1 using unbiased vectors. However, use of tone1 and tone4 like the reference base gives the same accuracy.

As one expected it, the maximal recognition rate is only 83.1% whereas the results of Yang et al. [4] for four Mandarin lexical tones have been equal to 96.5%. In first rapid conclusion, we deduce from it that the Mandarin's feature vector is not an optimal representative for Vietnamese and must be adjusted. An explanation could be founded if we remember that there are only 5 tones in Mandarin (6 in Vietnamese with 8 representations) and that the Vietnamese tone3 present a complex contour with important variation between high and low registers. Pitch contours for Mandarin's tones seem to be simplest [4].

Following comments of tones that presented in section 2, the duration of voiced part and the difference between endpoint and onset will be joined into the feature vector, in order to complete its description. However, the initial part is influenced by the voiced consonant if it exists in the syllable, the difference is thus calculated between middle and final points of voiced part. Therefore, we propose an adapted feature vector defined as:

$$\begin{aligned} y_t = [&\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(e_t) - \log(e), \\ &a, \log(f_e) - \log(f_m)] \end{aligned} \qquad (2)$$

where e_t is the energy at frame *t*, *e* is the maximal energy of the voiced part, *a* is the duration of the voiced part, f_e is the pitch frequency at the final frame and f_m is the pitch frequency at the middle frame. The last three components of vectors in expression (2) are the normalized short-time energy, the duration and the global slope. The vector is automatically extracted. Three new experiments using this adapted feature vector has been carried out:

Experiment 4: using unbiased feature vector with form defined by (2);

Experiment 5: using adapted feature vector biased by tone1.

Experiment 6: using adapted feature vector biased by tone4.

Results of Tables 2, 3 and 4 show that for the new adapted vector defined by (2), the recognition accuracy with biased vectors in the experiments 5 and 6 is higher than in the experiment 4 using unbiased vectors. However, like in the Mandarin's vector case, use of tone1 and tone4 like the reference base gives nearly the same accuracy.

Recognition accuracy using our adapted feature vector defined by expression (2) is better than those obtained in the case of a use of feature vector defined by expression (1) Tone recognition increased by ~10%. The last three components in expression (2), i.e. the normalized short-time energy, the duration and the global slope seem very helpful for improving performance.

For the 6 experiments examined together, recognition accuracy of first and second test sets are similar. First test was proceed with 8 of the 16 Vietnamese vowels, moreover the second test was proceed with the 8 remained vowels. Thus we can conclude that the tone recognition performance is independent of the vowels.

tone	Tone Accuracy	
	1st test	2e test
tone1	82.5	89.3
tone2	89.0	94.4
tone3	82.5	75.7
tone4	89.0	85.0
tone5	81.6	87.5
tone6	90.9	87.1
average	85.9	86.5

Table 2. Results of the experiment 4 in %

tone	Tone Accuracy	
	1st test	2e test
tone1	97.2	95.6
tone2	95.4	96.6
tone3	82.5	81.0
tone4	94.5	90.0
tone5	83.3	92.5
tone6	88.3	85.5
average	90.2	90.2

Table 3. Results of the experiment 5 in %

tone	Tone Accuracy	
	1st test	2e test
tone1	97.2	96.8
tone2	96.3	92.7
tone3	82.5	84.2
tone4	96.3	92.0
tone5	81.6	91.2
tone6	96.1	91.9
average	91.6	91.4

Table 4. Results of the experiment 6 in %

4. COMPLETE ISOLATED WORD RECOGNITION SYSTEM

We coupled our Vietnamese tone recognition system, as described above, to a classical HMMs recognition system of syllables disregarding the tones, in order to build a complete isolated word recognition system.

For syllables recognition, MFCC_E_D_A acoustic vector with 39 components which contain three types of parameters:

- 12 MFCC coefficients + 1 energy coefficient
- 12 Delta MFCC coefficients + 1 delta energy coefficient
- 12 acceleration MFCC coefficients + 1 acceleration energy coefficient

The HMMs topology presents 6 states with 4 emitting states and we used 1 stream and 1 mixture Gaussian.

For this first test, we used a small corpus made up of 131 syllables (with the tone) whose each syllable is repeated 4 times by 5 speakers. In this set of 131 syllables, 13 couples of these are identical based-syllable but with different tones. The training set contains all syllables, which were repeated 3 times. Testing set contains all syllables only with 1 repetition.

For the based-syllable (i.e. syllable without the tone), we reach a recognition accuracy of 91.3% For each systems in parallel, the 2 most probable candidates are kept. A small dictionary allows us eliminate the non-possible tone/syllable couples (i.e. a closed syllable with the tone3 for example). The maximum recognition accuracy for our complete system is equal to 93,8 %.

5. CONCLUSION

In this paper, we have presented a first try of the tone recognition for Vietnamese. One new feature vector is described, derived from the one proposed for Mandarin but adapted to give a better characterization of complex Vietnamese tones. The tone recognition of Vietnamese seems independent of the vowels. A high recognition accuracy (~90%) was reached by using our adapted feature vector with the pitch reference base chosen among both monotonous tones (tone1 or tone4). By coupling this tone recognition system with a state of the art engine, we built a first multi-speaker automatic isolated words recognition system, adapted for Vietnamese. Testing it on 5 speakers this system reached a recognition accuracy of approximately 94 %.

For the continuation of our study, we will test our Vietnamese recognition system with the 18 speakers of our database. We will build an application of Internet navigation type in order to test in real situation the operation of our recognition engine.

This study was carried out within the framework of the co-operation program "International Research Center MICA" between laboratory CLIPS-IMAG and the Hanoi University of Technologies.

6. REFERENCES

[1] T.T. Doan, "Ngu am tiêng viêt (Vietnamese phonetic)", Nha Xuat Ban Editions, 1977.

[2] M.S. Han, K.O Kim , "Phonetic variation of Vietnamese tones in disyllabic utterances tones", Journal of Phonetics, vol. 2, 1974, pp 223-232

[3] B. H. Vu, "Ve dac trung co ban cua thanh dieu tiêng viêt o trang thai tinh (characterisation of Vietnamese tones in static mode)" , Journal of Linguistic Institute of Vietnam, Vol. 6, 1999, pp 34-53.

[4] W.J Yang et all, "Hidden Markov Model for Mandarin Lexical Tone Recognition", IEEE Trans. ASSP, vol36, no 7, July 1988, pp 988-992

[5] L.S Lee et all, "Golden Mandarin (I) - A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary", IEEE Trans. ASSP, vol. 1, no 2, April 1993

[6] A. Tungthangthum, "Tone Recognition for Thai", Circuits and Systems, IEEE APCCAS 1998, Asia-Pacific Conference, p. 157-160.

[7] S. Kadambe, G.F Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals", IEEE Trans. Information Theory, vol. 38, no 2, March 1992.

[8] S. Kadambe, G.F Boudreaux-Bartels, "A Comparison a Wavelet Functions for Pitch Detection of Speech Signals", in Proc. IEEE ICASSP 1991

[9] Steve Young et all, "The HTK Book", the Cambridge University Engineering Department, July 2000.