Influence of F0 on Vietnamese syllable perception

TRAN Do Dat[*,**], *Eric CASTELLI*[*], *Jean-François SERIGNAT*[**], *TRINH Van Loan*[*], *LE Xuan Hung*[*,**]

*International Research Center MICA - 1 Dai Co Viet, Hanoi, VIETNAM
(Do-Dat.Tran, Eric.Castelli,Van-Loan.Trinh, Xuan-Hung.Le)@mica.edu.vn
**CLIPS-IMAG Laboratory, UMR CNRS 5524, BP 53, 38041 Grenoble Cedex 9, FRANCE
(Do-Dat.Tran, Jean-Francois.Serignat, Le-Xuan.Hung)@imag.fr

Abstract

Understanding and managing tonal characteristics of Vietnamese language is one of the most difficult aspects in Vietnamese speech processing. However, at present, there is no common agreement about the influence of fundamental frequency (F0) on the perception of Vietnamese syllables. Thus, instead of analyzing the F0 of a limited number of Vietnamese syllables like other methods found in literature, this paper will present a new methodology based on, first the synthesis of arbitrary syllables, and secondly perception tests. This approach permits us to understand and define more precisely the role of F0 in the characterisation of Vietnamese tones.

1. Introduction

Nowadays, due to advances in vocal technologies, practical speech applications have been developed in various fields, such as building human machine interface modules for the disabled, for industrial control, or for multimedia applications, using speech synthesis and recognition. In Vietnam, speech processing has been studied in recent years and some results are now available [2][4][5]. One difficulty in Vietnamese speech processing consists in the characterization of tonal evolutions of F0 during the production of Vietnamese syllables. That is the reason why a disagreement on the understanding of the effect of F0 on Vietnamese syllables still exists between Vietnamese linguists. Two main opinions on this problem could be described as follow:

1) tone effects on the whole syllable [4][5];
2) tone effects essentially on the final part of the syllable [1][2].

It means that this disagreement focuses mainly on the influence of the initial consonant in Vietnamese syllables. However, both hypotheses are based on experimental results from the analysis of the F0 contour of limited Vietnamese syllables.

2. Characteristics of Vietnamese tones

As mentioned above, Vietnamese language is a mono-syllabic and tonal language with 6 tones (table 1). A syllable in full structure (a tonal syllable) has five parts: initial sound (consonant), medial sound (semi-vowel), nucleus sound (vowel or diphthong), final sound (consonant or semi-vowel) and tone (figure 1). Besides the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part. Each Vietnamese tone could contribute to construct the morpheme and meaning of word. The tone has the same function as a phoneme, it always assigns for syllable.

<table>
<tr><td colspan="4">TONAL SYLLABLE (6,492)</td></tr>
<tr><td colspan="4">BASE SYLLABLE (2,376)</td></tr>
<tr><td rowspan="3">Initial (22)</td><td colspan="3">Final (155)</td></tr>
<tr><td>Medial (1)</td><td>Nucleus (16)</td><td>Ending (8)</td></tr>
<tr><td colspan="3">TONE (6)</td></tr>
</table>

Figure 1: The phonological hierarchy of Vietnamese syllables with total numbers of each phonetic unit (in the hypothesis where tone effects on the whole syllable)

Table 1: The 6 Vietnamese tones (order number, Vietnamese name and used sign)

Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
ngang	huyền‘\’	ngã ‘~’	hỏi ‘?’	sắc ‘/’	nặng‘.’

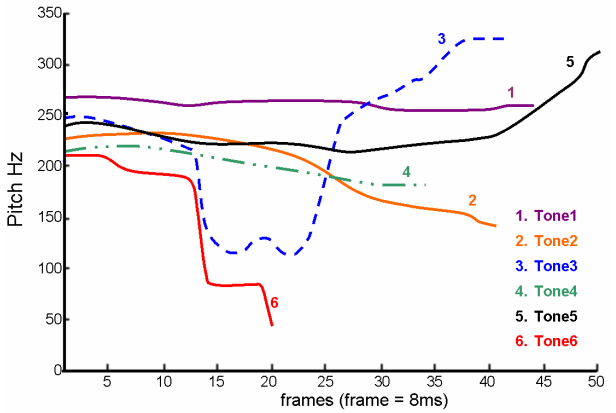


Figure 2: Example of the contours of six tones (female subject PNY), as described in [2]

The F0 contours of the 6 Vietnamese tones (examples are shown in Figure 2), are described as follows:

- Tone 1- Level tone. (“ngang”) is a high tone. At the beginning of syllable, it is the highest tone. The steady state of the level contour is observed consistently [1][2][3].
- Tone 2 - Falling tone. (“huyền”), the onset of the falling tone is lower than tone 1, tone 5 and tone 3. The low f0 at the onset gradually falls toward the end [1][2][3].
- Tone 3 - Broken tone. (“ngã”), the onset is as high as that of the level of tone 5, it is higher than the falling tone [1][2][3]. The second third of the contour of this tone is characterized by an abrupt dip caused by a heavy laryngealization. In most cases, the bottom of the dip occurs between the mid-point and the point two-thirds

from onset. A creaky voice is heard during this dip [1][3][5].

- Tone 4 - Curve tone ("hỏi"), the onset is the lowest among the six tones. The low onset falls further gradually until the point two-thirds from the onset. From this point, the extremely low f0 starts to rise toward the end [1][2][3].
- Tone 5 - Rising tone (tone "sắc"), the onset is also high. Starting from high onset, the F0 gradually rises for the first two thirds of the duration. After this point, the rise becomes more rapid. [1][2][3].
- Tone 6 - Drop tone ("nặng"), the onset is usually higher than that of the falling of curve tone but considerably lower than the tone 1, tone 5 and tone 3. This tone is characterized by a heavy laryngealization at the end and also by its considerably shorter duration than the other tones. The duration of this tone is approximately two thirds of the other tones [2][3]. The main body of this tone is almost leveled or slightly falling.

These descriptions are only for the Hanoi dialect, the standard dialect of Vietnamese language. They would be changed with the other dialects in the South and the Centre of Vietnam. In these regions, there are only 5 tones [1][2] instead of 6 like the Hanoi dialect, because tone 3 and tone 4 are pronounced identically.

3. Vietnamese Speech Synthesis system

Several Vietnamese speech synthesis systems have been developed in recent years [4][5]. The system of [4] is a parametric and rule based speech synthesis system, which is based on the source-filter-model of speech production. It does not permit to have a high quality synthetic speech. Besides, a disadvantage of [5] is that the F0 and duration of syllable could not be manipulated. In our work, we chose the concatenation method with TD-PSOLA [9] algorithm for our speech synthesis system. It has been widely used in recent years for many languages with good results and applied successfully for speech synthesis systems of other tonal languages like Chinese [6][7] and Thai [5].

With the TD-PSOLA technique, the position of pitch marks needs to be precise. Thus, in order to have high quality synthetic speech, a pitch marking tool which uses a simple propagating algorithm was developed (figure 3).

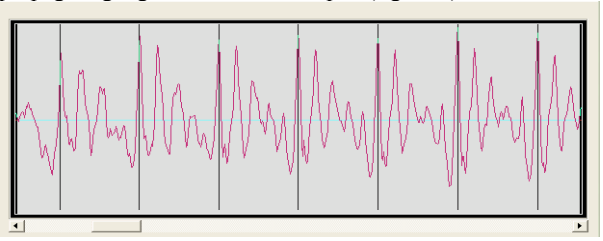


Figure 3: Pitch marks of vowel /a/ in a syllable /ba_5/

By applying recent results [2] about the shapes of fundamental frequency of the six Vietnamese tones, we have succeeded in synthesizing syllables with six different tones (figure 4). Based on TD-PSOLA algorithm, synthetic syllables can be obtained by concatenation of different acoustic units such as diphone, half-syllable, initial/final part, and non-tonal monosyllable. For each of acoustic unit, the tonal syllable is synthesized with a F0 contour and duration desired.

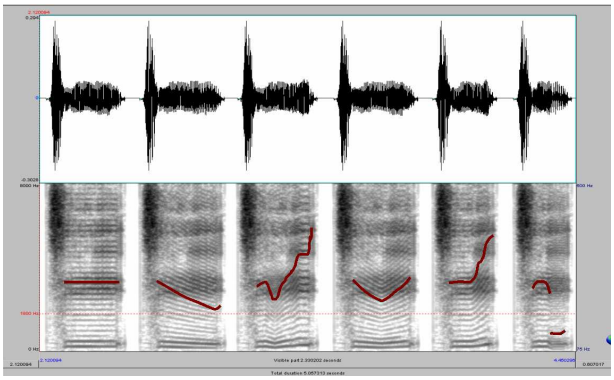


Figure 4: Syllable "chi" /ci/ with 6 different tones synthesized with half-syllable.

4. Influence of F0 on Vietnamese syllable

Previous studies found in literature [1][2][4][5], derived their experiment results from analyzing the F0 contour of limited Vietnamese syllables. They either used intonograph, kymograph machines or fundamental frequency calculating software to draw the F0 contours. The conclusions were based on subjective evaluations of analyzer, which constitutes a characteristical limitation of these methods.

In our experiment, we implemented another approach. In order to evaluate the influence of F0 on Vietnamese syllable, we used perception tests based on Diagnostic Rhyme Test (DRT) [8] method.

The F0 of one based-syllable is controlled to obtain synthetic syllables with different F0 contours. Two types of F0 manipulation are done. In the first type, F0 of syllables is only changed on the initial consonant position of the syllable: on this portion F0 is decreased or increased by about 20% of its mean value (figure 5a). In the second experiment, F0 is changed (decreased or increased) on the whole syllable (figure 5b).

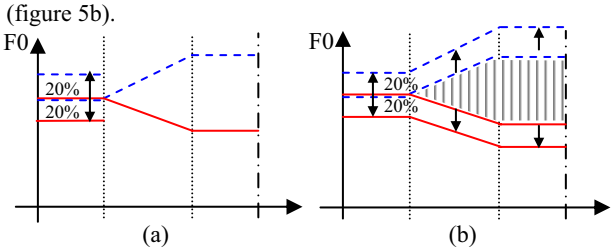


Figure 5: Two types of F0 controlling

To implement these experiments, we used one corpus including 68 mono syllabic words. These words are composed of one of the 5 Vietnamese voiced consonants /b/,/l/,/m/,/n/ and /ŋ/. The words beginning with /m/ and /n/ are combined with 12 of the 16 Vietnamese nucleus vowels. Words are recorded in a quiet environment at a 16 kHz sampling frequency with a 16bit/sample precision. They are uttered by a female speaker of Hanoi Television, who has a high quality voice pronounced with the standard dialect. Twenty listeners (10 men and 10 women), from the North and the South of Vietnam and from 19 to 40 years old, took part in our tests; each listener has a normal hearing ability.

4.1. The first perception test

In the first test procedure, the fundamental frequency F0 of the 68 syllables is only changed during the initial consonants. Thus, each syllable will provide 2 varieties, one has a low frequency initial consonant (-20%) and one has a high one (+20%), as shown on figure 6. The modification percentage is kept into the tolerance (standard shape) of F0 variation as proposed in [2] (figure 7). Thus, a corpus including 204 syllables was built and was divided into 3 groups: HighF0, LowF0 and Natural groups. Consequently, we have 68 sets of three syllables; each set includes one original syllable (natural sound) and its two modified syllables in the HighF0, LowF0 groups.

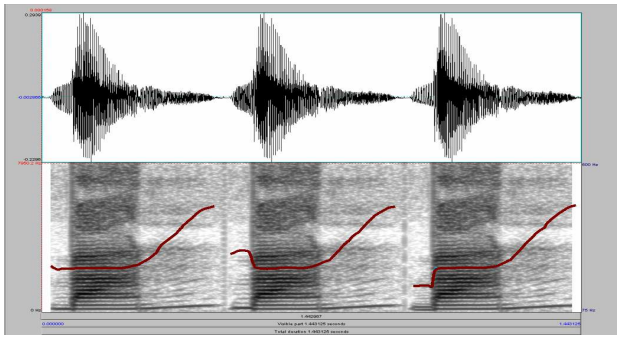


Figure 6: Waveform spectrograph and F0 contour of sound /ban_5/ with two varieties.

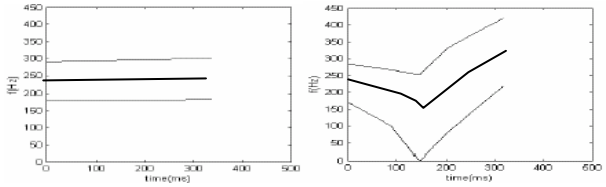


Figure 7: Standard shape of ton1 and ton3 of female subjects [2]

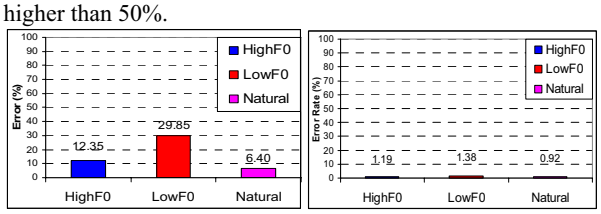
Based on Diagnostic Rhyme Test method, we made some changes to adapt our perception tests. The order of the 68 sets of three different syllables and the order of syllables in one set are both randomized. The listeners are asked to write down names of listened syllables (mono-word without tone) and to choose the name of a tone in a list of 6 tones.

4.2. Results

Figures 8, 9 & 10 successively show results of the perception tests. Error rates of the name of syllable recognition (a) and of the name of tone recognition (b) are presented in the figure 8. We can clearly note that the tone name recognition error rates are low, about 1%. The distance between the highest and lowest error rate is about 0.5%. *These first results seem to show that the changing of F0 in the initial consonant does not affect tone recognition.*

However, we pay more attention to the results in the figure 8a: error rate of the LowF0 group is quite high. By analyzing the results, we found that many modified syllables are recognized as other syllables which have different initial consonants such as “mía”, “mi”, “ni”… (see table 2).

Error rates of 68 syllables in three groups are more clearly shown in the figures 9 and 10. In these figures, we can see that, some syllables are not influenced by changing F0, (/ba/, /baw/, /ŋa/ for example). On the contrary, some others are more influenced by this modification, especially syllables modified by reducing F0 at the initial consonant such as /muo_1/, /mie_5/ niŋ_1/ and /ŋi_1/, which have an error rate higher than 50%.



(a) *(b)*

Figure 8: Syllable name (a) and tone name (b) recognition error rates of 3 groups

Besides, there are some of natural syllables having quite high error rates; they are /mi_5/, /mi_2/, /mi_3/, /mi_6/. Most of these errors are caused by a similarity between two labial stop Vietnamese consonants /b/ and /m/. However, the error rates of these syllables in the LowF0 group are much higher.

Table 3: Some varieties of listened syllables

Original syllable	Transcription	Listened syllable	Transcription
mỹ	/mi_3/	bĩ	/bi_3/
mía	/mie_5/	bía	/bie_5/
mua	/muo_1/	bua	/buo_1/
nỉ	/ni_4/	đỉ or tỉ	/di_4/ or /ti_4/
no	/no_1/	đo	/do_1/
nuôi	/nuoj_1/	đuôi	/duoj_1/

Therefore, in order to distinguish the error rates of the three groups, we subtracted the error rates of Natural group from the LowF0 group’s and HighF0 group’s (figure 10). Through this figure, it is easy to realize that the error rate of the LowF0 group is much higher than that of the two remaining groups. It means that the perception of the syllables could be changed when the F0 of their initial consonant is reduced.

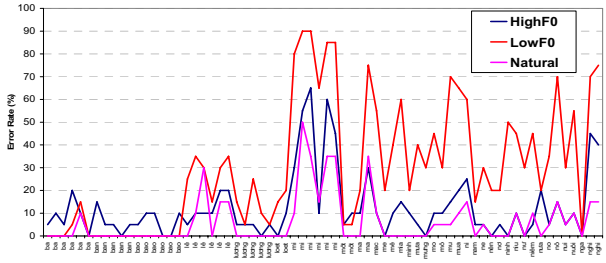


Figure 9: Error rates of the 68 syllables in alphabetical order

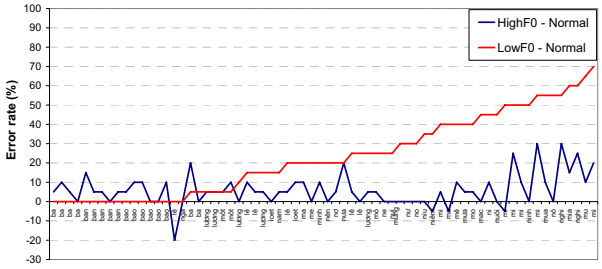


Figure 10: Error rates of the syllables after subtraction

4.3. Second perception test

In this test, we chose 16 syllables in a list of syllables which have the highest error rate. We carried out the re-synthesis of these syllables with our synthesis system in three ways (figure 11):

- Re-synthesize of a complete syllable with F0 parameters extracted from original syllable.
- Re-synthesize of a complete syllable with value of F0 parameters extracted from original syllable minus 20% value of extracted F0 average.
- Re-synthesize a Final part with F0 parameters extracted from original syllable, and 80% value of extracted F0 at the initial consonant position. A transition of F0 between the initial consonant and the nucleus vowel of final part is smoother than the transition of syllables in LowF0 group in the (§4.1).

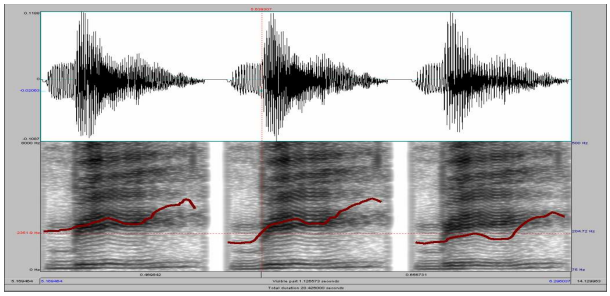
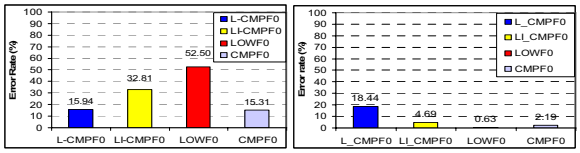


Figure 11: Three type of re-synthesis of syllable /mi_3/.

Same perception tests were carried out with 4 groups of syllables:

- CMPF0 group includes the syllables of type (1);
- L-CMPF0 group includes type (2) syllables;
- LI-CMPF0 group consists in syllables of type (3);
- And 16 syllables are taken from LOWF0 group of the first experiment.



(a) (b)

Figure 12: Syllable name (a) and tone name (b) recognition error rates of the 4 groups

From the figure 12a, we can find that the LowF0 group still has the highest error rate. The next is the LI-CMPF0 group which has syllables re-synthesized by type (3). The syllables which are from L-CMPF0 group have an error rate lower than the LowF0 group and LI-CMPF0 group and it nears to result of CMPF0 group. This is a foreseeable result. On the contrary, in figure 12b, the syllable name recognition of the L-CMPF0 group has the highest error rate value. By analyzing obtained results, we understood that this is a systematical error. There are two kinds of error.

- An error caused by dialect problems: in our tests, the listeners are from two dialect regions, the North and the South of Vietnam. The tone3 et tone4 of the south dialect are quite similar [1][2], so the listeners from the South usually made a mistake when choosing the name of these tones.
- An error caused by difference of F0 (60% of the error rate): In one set of 4 syllables, F0 of the syllable which is from L-CMPF0 group has the lowest value. The error usually occurs with the tone1 syllables. In fact, listeners selected the tone2 for the lowest F0 syllable instead of choosing ton1. Because the F0 contour of the tone2 is similar to tone1 but lower than it.

However, this error rate is not high. It is much smaller when we reject these systematical errors (it is decreased from 18% to 8%). From the error rates of L-CMPF0 group, we can see that the name of syllable and the name of tone are less influenced when decreasing F0 value on whole syllable.

5. Conclusions

Through 5 parts of this paper, based on the controlling values of F0 parameter of the Vietnamese syllable, we have an objective conclusion of the influence of fundamental frequency, which is one of the most important parameters, on the Vietnamese tones. ***The initial consonant does not carry information of the tone, it does not participate to construct the tone of the Vietnamese syllable, it only contributes to create the syllable. The Vietnamese tone affects only on the Final part of the syllable***. These results will help us to construct one high quality TTS system in the future, but also could be very useful to improving automatic speech recognition systems.

6. Acknowledgments

This study was done in the framework of the MAE CORUS international co-operation programme.

7. References

[1] Doan, T.T., "Ngữ âm tiếng Việt" (Vietnamese Phonetics), Hanoi National University Publishing House, pp. 99-148, 1999.

[2] Nguyen, Q.C, "Reconnaissance de la parole en langue Vietnamienne", *PhD. thesis INP- Grenoble, France,* June 2002 .

[3] Mieko S.Han and Kong-On K. "Phonetic variation of Vietnamese tones in disyllabic utterances", *Journal of Phonetics April 1974*, pp.223-232, 1974.

[4] Do, T.T, Takara, T., "Precise tone generation for Vietnamese Text-to-Speech system", *Proc. of ICASSP'03,* I, pp. 504 – 507, 2003.

[5] Nguyen, D.T., Mixdorff, H., et al., "Fujisaki Model based F0 contours in Vietnamese TTS", *ICSLP2004, Korea,* pp. 1429-1432, 2004 .

[6] Xu Y., Araki M., and Niimi Y., "A chinese speech synthetic system based on TD-PSOLA", *Proc. of Int'l Conf. on Chinese Computing,* pp. 171-175, 2001.

[7] Sin-Horng C., Shaw-Hwa H., Yih-Ru W., "A Mandarin Text-to-Speech System", *Computational Linguistic and Chinese Language Processing Vol.1, no.1, August 1996*, pp 87 -100, 1996.

[8] Donovan R.E., "Trainable Speech Synthesis", *PhD. Thesis, Cambridge University Engineering Department,* 1996.

[9] Dutoit T., "An introduction to text-to-speech Synthesis", *Kluwer Academic Publishers,* 326 pp,1996.