# Linear F0 Contour Model For Vietnamese Tones And Vietnamese Syllable Synthesis With TD-PSOLA

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

*International Research Center MICA - 1 Dai Co Viet, Hanoi, VIETNAM
(Do-Dat.Tran, Eric.Castelli, 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. Our newest results indicate that, the initial consonant of one Vietnamese syllable does not carry information of the tone, the Vietnamese tone has an effect only on the Final part of the syllable. Based on obtained results, this article proposes linear F0 contour models for the Vietnamese tones generation. These models only describe the F0 evolution of a final part of the Vietnamese syllable, and they are evaluated through perception tests.

## 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 [1][3][5][6]. However, the characterization of tonal evolutions of F0 during the production of Vietnamese syllables is still a principal issue for the development of Vietnamese speech processing. According to [1], problem of the effect of the F0 on Vietnamese syllable is now clearer; the obtained result shows that the Vietnamese tone affects only the Final part of the syllable.

In the [5][6], two Vietnamese speech synthesis systems are presented. The system of [5] 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. A disadvantage of [6] is that the F0 and duration of syllable could not be manipulated. Besides, the synthesized syllables of these systems are total voiced syllables (syllables contain all voiced phonemes), and the systems have not given a common model of F0 contour that is sufficient for generating all kinds of the Vietnamese tone.

Based on obtained results from [1] and [3], this article proposes linear F0 contour models for the Vietnamese tones. The F0 contour is controlled by applying the TD-PSOLA [9] algorithm which is chosen for building our speech synthesis system. This is a continuous work of [1] in order to build one high quality text to speech synthesis system. Besides controlling F0 contour, like result of [5], power control was implemented for synthesizing syllables with tone3 and tone6 which are different from the other Vietnamese tones and also from the Chinese tones. On the other hand, others parameters such as energy and duration of syllable which are the essential parameters to construct the Vietnamese tones, have been considered to synthesis high quality syllables.

## 2. Characteristics of Vietnamese Tones

The Vietnamese language is a monosyllabic and tonal language with 6 tones. 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 and 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">TONE (6)</td></tr>
<tr><td colspan="3">FINAL (155)</td></tr>
<tr><td>Medial (1)</td><td>Nucleus (16)</td><td>Ending (8)</td></tr>
</table>

*Figure 1: The phonological hierarchy of Vietnamese syllables with total numbers of each phonetic unit*

The F0 contours of the 6 Vietnamese tones 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 [2][3][4].
- Tone 2 - Falling tone (huyền). The low f0 at the onset gradually falls toward the end [2][3][4].
- Tone 3 - Broken tone (ngã). 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 [2][4][6].
- 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 [2][3][4].
- Tone 5 - Rising tone (sắc). Starting from high onset, the F0 gradually rises for the first two thirds of the duration. After this point, the rise becomes more rapid. [2][3][4].
- Tone 6 - Drop tone (nặng). This tone is characterized by a heavy laryngealization at the end and also by its considerably shorter duration than the other tones [2][3][4]. The main body of this tone is almost leveled or slightly falling.

These descriptions are only for isolated syllables of the Hanoi dialect, the standard dialect of Vietnamese language. They would be changed in the continuous speech and with the other dialects in the South and the Centre of Vietnam.

## 3. Vietnamese Syllable Synthesis with TD-PSOLA

The TD-PSOLA (Time Domain PSOLA) is widely used for many languages due to its computational efficiency. It has been applied successfully for speech synthesis systems of some tonal languages like Thai [6] and Chinese [7]. Vietnamese is also a tonal and monosyllabic language. It has over 2350 syllables if tone is ignored and nearly 6500 syllables if tone is considered (figure 1). It is much more than the Chinese language that has about 400 syllables without tone and over 1300 syllables with tone [7]. One of the advantage points of the TDPSOLA is that modifying both the duration and the pitch of a given segment can be performed in a straightforward way. Thus, in order to synthesize syllables with tone, firstly their base syllable either is taken from database or is synthesized by concatenating acoustic units such as diphone, half-syllable and initial/final part. Then tone will be added by controlling F0 by the TDPSOLA algorithm.

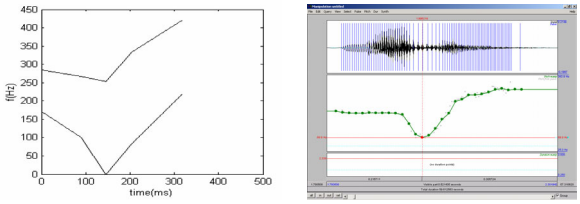



*Figure 2: Standard shape of tone3 of female subjects [3], and an example of waveform and F0 contour of a syllable /ba_3/*

In the Vietnamese 6 lexical tones, there are two glottalized tones (ton3 and ton6), they are different from the other Vietnamese tones and also the Chinese tones. According to the results of [3][4], F0 value of these tones falls greatly at the points where the glottal stop feature occurs. Value of F0 at these points could be less than half of the initial F0 value (figure 2). On the other hand, it is well known that pitch factor (Fp =T/T0) is only modified by TD-PSOLA from 0.5 to 2. Therefore, to synthesis tone3 and tone6 with TDSOLA, the minimum value of controlled F0 has to be greater than F0/2 of original signal.

## 4. Linear F0 Contour Model for Tone

In our experiments represented in [1], to evaluate the influence of F0 on Vietnamese syllable, the F0 of one syllable is controlled to obtain synthetic syllables with different F0 contours. Perception tests based on Diagnostic Rhyme Test (DRT) [8] method were carried out. Twenty listeners were asked to write down names of listened synthetic syllables and to choose the name of a tone in a list of 6 tones. The obtained results have shown that the initial consonant does not carry information of the tone, the Vietnamese tone affects only the Final part of the syllable.

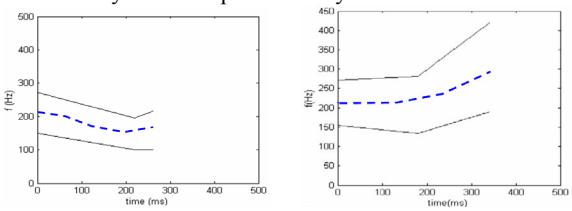



*Figure 3: Standard shape of tone4 and tone5 of female subjects [3] and linearized average F0 contours (in dashed blue)*

Based on the results of [3] about the shapes of fundamental frequency of the six Vietnamese tones, we get average values of F0 contours of the six Vietnamese tones (figure 3).

From above results, we have constructed linear F0 contour models for the Vietnamese tone (figure 4). The models only describe the F0 evolution of the final part of the Vietnamese syllable.

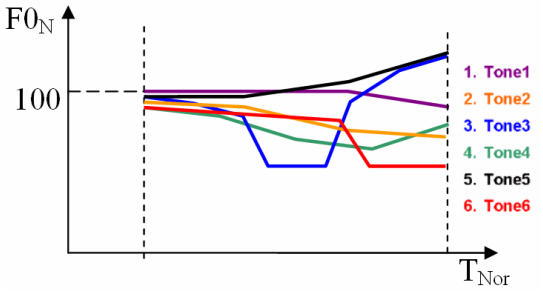



*Figure 4: Normalized linear F0 contour models for 6 Vietnamese tones*

As noting in the second section, the tolerable pitch factor is from 0.5 to 2, therefore, in the contour model of the tone3 (figure 5) and the ton6, to ensure the F0 value in an allowed range, we keep the minimum F0 equal to 55% of an initial F0.

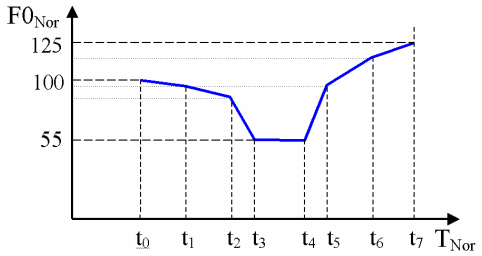



*Figure 5: Normalized linear F0 contour models for tone 3*

Besides controlling F0 contour, like result of [4], we realized that amplitude of signal decreases during the glottalized time of the tone3 and the tone6. Therefore, in order to study an influence of power variation in these tones, two power patterns are used (figure 6). Power control is implemented by changing the amplitude of signal at the points which occurs the glottal stop feature. They are weighted with weighting factor taken from power pattern.

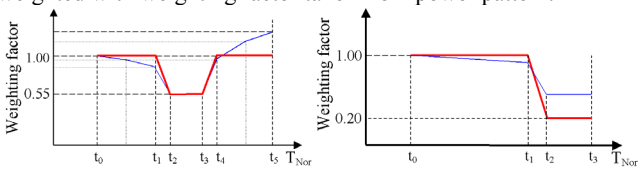



*Figure 6: Power patterns (in bold red) of the tone 3 and tone 6*

## 5. Perception Tests and Results

In order to evaluate the performance of the TD-POSLA algorithm, the F0 patterns and the influence of power control in synthesizing Vietnamese syllable, three perception tests which are based on the Diagnostic Rhythm Test (DRT) and Mean Opinion Score (MOS) test were implemented.

To implement these experiments, one corpus including 84 monosyllabic words was used. If the tone is ignored, there are 15 different base syllables (table 5.1) in the corpus. They are composed of 13 in the 16 Vietnamese nucleus vowels.

Since the F0 patterns of the Vietnamese tones are only for the Final part. Therefore, to ensure our model is usable for all kinds of Vietnamese syllables, these base syllables contain 7 syllables beginning with voiced consonants and 8 syllables beginning with unvoiced consonants. Furthermore, to take into account an

influence of ending sound, these syllables are ended by different sort of ending sounds such as voiced consonants (/zɤ̆n/, /lɯɤŋ/, /kăm/..), semi-vowels (/baw/, /kiew/) and non ending sound (/ba/, /du/ /cɤ/…).

*Table 5.1 List of 15 base syllables*

| Syllable | Transcription | Syllable | Transcription |
|---|---|---|---|
| ba | /ba_1/ | căm | /kăm_1/ |
| bao | /baw_1/ | kiêu | /kiew_1/ |
| dân | /zɤ̆n_1/ | cho | /cɔ_1/ |
| đu | /du_1/ | chơ | /cɤ_1/ |
| lê | /le_1/ | chua | /cuo_1/ |
| lương | /lɯɤŋ_1/ | he | /hɛ_1/ |
| nghi | /ŋi_1/ | hô | /ho_1/ |
| | | ria | /rie_1/ |

All of these syllables are recorded in a quiet environment at a 16 kHz sampling frequency with a 16 bit/sample precision. A female speaker of Hanoi Television, who has a high quality voice with the standard dialect and is not in the female subjects of [3], utters them. Twenty listeners (14 men and 6 women), from the North of Vietnam and from 22 to 34 years old, took part in our tests; each listener has a normal hearing ability.

In the first and the second test procedure, we have 4 groups of 84 different syllables:

- Group 1 (Natural group) includes 84 original syllables.
- Group 2 (Re-synthesis group) includes 84 syllables which are re-synthesized from the 84 original syllables by applying the TD-PSOLA algorithm.
- Group 3 is composed of 84 synthetic syllables. Fifteen base syllables in the corpus have been used as acoustic units in our speech synthesis system. Initial F0 is fixed as 230 Hz. Based on the F0 contour patterns of the 6 Vietnamese tones, 84 new syllables were synthesized.
- Group 4 is composed of 84 synthetic syllables with the power controlling. Besides applying the F0 patterns like group3, power of the syllables is controlled by using the power pattern.

Consequently, we have 84 sets of four syllables; each set includes one syllable from 4 groups.

### 5.1. The first perception test – Intelligibility of tone model

The goal of this test is to verify an effect of the constructed tone models. Based on Diagnostic Rhyme Test method, we made some changes to adapt our perception tests. The order of the 84 sets of four different syllables and the order of syllables in each set are both randomized. The listeners are asked to write down names of listened syllables and to choose the name of tone in a list of 6 tones.

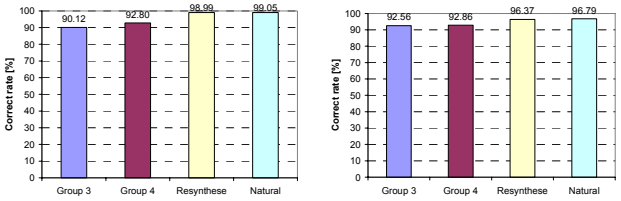



*Figure 7: Correct rate of tone name (a) and syllable name (b) recognition of 4 groups*

Figures 7&8 successively show results of the perception tests. Correct rates of the name of syllable recognition (a) and of the name of tone recognition (b) are presented in the figure 7. We can clearly note that recognition correct rates of the both are quite high, above 90%. From the figure 7, we can find that the natural group has the highest correct rates. The next is the re-synthesis group which has syllables re-synthesized by applying the TD-PSOLA algorithm. This is a foreseeable result.

In the figure 7a, the correct rate of the tone's name recognition of the group 4 is a little bit higher than the group 3's. Error rates of the 6 tones of the four groups are more clearly shown in a figure 8a. We can found that, all groups have a high correct recognition rate at tone 1, tone 2 and tone 3. Correct rates of tone 4 of the groups 3&4 are lower than other groups. However, their correct rates are above 90%, so it is a quite good and acceptable result. The result of the tone 3 shows that, power control for this tone does not give a significant change in recognizing this tone.

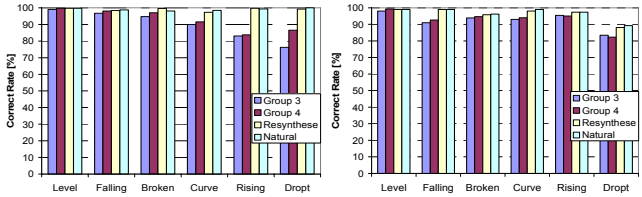



*Figure 8: Correct rate of tone name (a) and syllable name (b) recognition of six tones of 4 groups*

*Table 5.2 Confusion matrix of group 4*

| | Level | Falling | Broken | Curve | Rising | Drop |
|---|---|---|---|---|---|---|
| Level | 99.67 | 0.33 | 0 | 0 | 0 | 0 |
| Falling | 0.3 | 98.00 | 0 | 1.00 | 0.7 | 0 |
| Broken | 0 | 0 | 95.30 | 3.50 | 0.80 | 0.40 |
| Curve | 0.80 | 5.40 | 0 | 91.90 | 0 | 1.90 |
| Rising | 13.00 | 1.00 | 2.70 | 0.30 | 83.00 | 0 |
| Dropt | 3.50 | 1.90 | 2.70 | 5.00 | 0.40 | 86.50 |

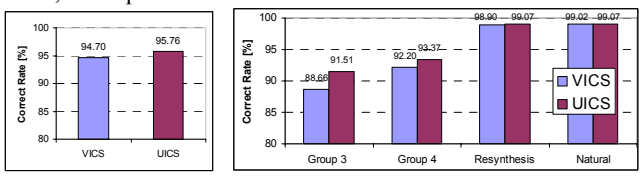
On the other hand, we pay more attention to the results of the tone 5 and tone 6. As usually, syllables of the natural group and the re-synthesis group have the highest correct rate (higher than 99%). However, with these tones, the distance between these groups in comparison with two remaining groups is quite large. The tone 6 of the group 3 has the lowest correct rate, about 76%. The group 4 which is composed power controlled syllables has the better result. Its correct rate is 87%, it is higher than group 3 of about 11%. It means that, power control might contribute to synthesize Vietnamese syllable with tone 6 better.

Besides, we can see that, the tone 5 has the lowest correct rate among the 6 tones. In fact, syllables with tone 5 of two groups 3 & 4 are identical. Consequently, an approximation of their correct rates is logical. The error recognition rate of these groups is about 15%. In the confusion matrix (table 5.2), a major reason for this error is that, instead of selecting tone 5, listeners select tone 1. It means that, the variation of F0 contour pattern is not enough and it need to be changed.

The correct rate of the syllable name recognition is high (figure 7b). An error rate of each tone is presented in detail in figure 8b. In this figure, a distribution of error rates of the 4 groups is similar. The obtained results show that, our patterns do not have an effect on the syllable name recognition.

The result of correct rate of two kinds of syllable 1) syllables beginning with voiced initial consonants (VICS) and 2) syllables

beginning with unvoiced initial consonant (UICS) in the figure 9a&b, could prove that our tone model works well for the both.

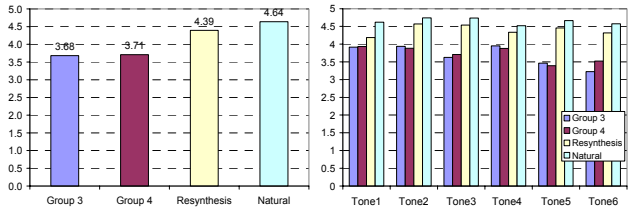



(a) (b)

*Figure 9: Average correct rates of tone recognition of Voiced/Unvoiced initial consonant syllables (a) in all four groups (b) and in each group.*

### 5.2. The second perception test - Naturalness of tone model

The same corpus of the first test was used for this test. A perception test based on the MOS test was implemented. The listeners were asked to rate the speech quality of four groups on a scale 1-5, where 1 is bad and 5 is completely natural.




(a) (b)

*Figure 10: Average MOS of 4 groups (a) and of 6 tones (b)*

Figure 10 shows the naturalness scores of four groups of stimuli compared with natural group. Like the result of the tone name recognition, the natural has the highest score. The next is the re-synthesis group. Groups 4 and 3, which have an average score value of '3.7', consequently take the third and the fourth places. The results in the figure 10b show that, by controlling power, the naturalness score of tone 3 syllables and tone 6 syllables is a little bit higher than syllables on which are only applied the F0 contour pattern.

### 5.3. The third perception test.

Besides implementing the perception test for monosyllables, we begin applying the TD-PSOLA algorithm and our tone models for sentence synthesis. Firstly, acoustic units such as diphones, half-syllables, initial/final parts are extracted from phrase corpus of VNSpeechCorpus. Secondly, with each acoustic unit type, 32 sentences are manually synthesized. Values of set of 3 parameters including duration, energy and F0 of synthetic syllables are manipulated by one expert, and the same sentences of 3 types of acoustic unit have an identical set of 3 parameters. Thus, we have 3 groups of 32 synthesized sentences.

- Diphone group includes the sentences synthesized by using the diphone as concatenated acoustic unit.
- Half-syllable group includes the sentences synthesized by using the half-syllable as concatenated acoustic unit.
- And initial/final group.

Thirdly, the F0 contour patterns and the power patterns were used to create a Tone model group. The synthesized sentences of this group are taken from the half-syllable group. However, besides using the same duration and energy values likes all of three groups, instead of giving directly F0 values for synthetic sentences, 32 sentences of the tone model group are synthesized by applying the tone models with initial F0 value as 230 Hz.

Consequently, one corpus which is composed of 4 groups of 32 synthesized sentences was built, and it was used in the test MOS for evaluating the naturalness of each group.

Figure 11 represents results of MOS test for 4 groups. In general, the scores of all groups are higher than 3. This is a quite good and satisfactory result. Furthermore, the score of the Tone model group shows that our Tone model could be applied for Vietnamese sentence synthesis. Though these results are only for a small number of synthetic sentences, but it will help us to develop our speech synthesis system.

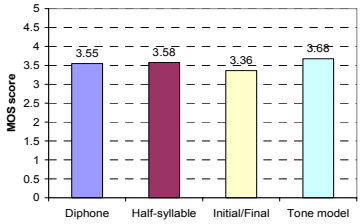



*Figure 11: Average Mean Opinion Score of 4 groups*

## 6. Conclusions

Through 5 parts of this paper, based on the perception tests, we can see that, like other tonal language, the TD-PSOLA can be used for Vietnamese speech synthesis. We could also affirm the conclusion in [1] of the influence of fundamental frequency on the Vietnamese tone. The initial consonant does not carry information of the tone; the Vietnamese tone affects only the Final part of the syllable. Furthermore, the obtained results prove that the linear F0 contour model works well for Vietnamese tone generation. These results will help us to construct one high quality TTS system in the future, but also could be very useful to improve automatic speech recognition systems.

## 7. References


[1] Tran D.D, Castelli E., Serignat JF., Le X.H., Trinh V.L., "Influence of F0 on Vietnamese syllable perception", *Proc. of Interspeech2005,* Lisbon, pp. 1697-1700, 2005.

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

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

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

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

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

[7] 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.

[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 Publics,* 326 pp,1996