Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese

Hansjörg Mixdorff, Nguyen Hung Bach**, Hiroya Fujisaki*** and Mai Chi Luong ***

*Faculty of Computer Science, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
***Professor Emeritus, University of Tokyo, Japan
fujisaki@alum.mit.edu
**Institute of Information Technology, National Centre for Science and Technology, Vietnam
nguyenbh@netnam.org.vn; lcmai@ioit.ncst.ac.vn

Abstract

The current paper presents a preliminary study on the production and perception of syllabic tones of Vietnamese. A speech corpus consisting of fifty-two six-syllable sequences with various combinations of tones was uttered by two speakers of Standard Vietnamese, one male and one female. The corpus was labeled on the syllabic level and analyzed using the Fujisaki model. Results show that the six tone types basically fall into two categories: Level, rising, curve and falling tone can be accurately modeled by using tone commands of positive or negative polarity. The so-called drop and broken tones, however, obviously require a special control causing creaky voice and in cases a very fast drop in *F0* leading to temporary *F0* halving or even quartering. In contrast to the drop tone, the broken tone exhibits an *F0* rise and hence a positive tone command right after the creak occurs. Further observations suggest that drop and broken tone do not only differ from the other four tones with respect to their *F0* characteristics, but also as to their much tenser articulation. A perception experiment performed with natural and resynthesized stimuli shows, inter alia, that tone 4 is most prone to confusion and that tone 6 obviously requires tense articulation as well as vocal fry to be identified reliably.

1. Introduction

Vietnamese is known as a monosyllabic tone language having six different lexical tones. These are (numbers indicate the indices to be used throughout this article): Level (1), sometimes also referred to as 'mid-level', rising (2), broken (3), falling (4), curve (5), and drop (6) tones. Tones 2-6 are marked by diacritics in the Vietnamese script which uses the Latin alphabet. The widely cited description by Thompson [1] gives the following account which is also summarized in Table 1:

Tone 1 is modal and its contour is nearly level in non-final syllables not accompanied by heavy stress, although even in these cases it probably trails downward slightly. Alhough tone 1 is phonetically slightly falling, it is phonemically regarded as a level tone similar to Mandarin tone 1, but with relatively lower pitch. Tone 2 is high and rising (perhaps nearly level in rapid speech) and tense, and similar to tone 2 in Mandarin Chinese. Tone 3 is also high and rising, the *F0* contour being similar to that of tone 2, but it is accompanied by the rasping voice quality occasioned by tense glottal stricture. In careful speech such syllables are sometimes interrupted completely by a glottal stop (or a rapid series of glottal stops). Its trajectory therefore sometimes shows a characteristic break in the voicing at about half of the total duration of the syllable. Tone 4 is tense; it starts somewhat higher than tone 5 and drops rather abruptly. In final syllables, and especially in citation forms, this is followed by a sweeping rise at the end, and for this reason it is often called the 'dipping' tone. However, non-final syllables seem only to have a brief level portion at the end, and this is exceedingly elusive in rapid speech. Although tone 4 is usually described as a low falling and then rising tone, not all Vietnamese speakers have the rising part. When tone 4 consists of a falling and a rising contour, it is similar to Beijing Mandarin tone 3. Tone 5 is also lax, starts quite low and trails downward toward the bottom of the voice range. It is often accompanied by a kind of breathy voicing, reminiscent of a sigh. Tone 6 is also tense; it starts somewhat lower than tone 4. With syllables ending in a stop [p t c k] it drops only a little more sharply than tone 5, but it is never accompanied by the breathy quality of that tone. Other syllables have the same rasping voice quality as tone 3, drop very sharply and are almost immediately cut off by a strong glottal stop. Tone 6 is much shorter than other tones with a tendency to go lower.

Table 1: Description of the six syllabic tones of Vietnamese.

No.	Vietnamese Name	English Name	*F0* contour	Diacritic used in writing	Additional features
1	Ngang	level	Trailing/falling	none	Laxness
2	Sắc	rising	Rising	Á	Tenseness
3	Ngã	broken	Rising	Ã	Glottalization
4	Hỏi	falling	Falling	Ả	Tenseness
5	Huyền	curve	Falling	À	Laxness, breathiness
6	Nặng	drop	Dropping	Ạ	Glottalization/tenseness

Hence, Vietnamese tones are not only characterized by distinct *F0* trajectories, but also by articulatory distinctions and the presence/absence of glottalization.
The current study aims at providing a preliminary quantitative description of Vietnamese tones as yielded by the analysis with the Fujisaki model of the production process of *F0*[2]. The model has already been successfully applied to tone languages such as Mandarin [3] and Thai [4]. Since, however, features such as voice quality and glottalization are not incorporated in the model, the question must be examined whether the Vietnamese tones can be reliably signaled by means of *F0* manipulation only, or whether additional control is required. To this effect, a speech corpus was recorded, analyzed and resynthesized using Fujisaki model contours. A perception experiment focusing on the potential confusion of tones was conducted.

2. Speech Material and Method of Analysis

In order to examine the realization of individual tones, as well as tone coarticulation, a set of 52 six-syllable utterances with varying combinations of tones was recorded by two phonetically trained native speakers of Northern Vietnamese, the Standard dialect of Vietnam, one male and one female. The utterances were composed of voiced sounds only for continuous *F0* contours, using only nasals and laterals as initial or final consonants in order to minimize microprosodic effects. Since it was impossible to create all desired combinations of tones with the same sequence of syllables, almost all of the utterances were of the 'nonsense' type. The segmental structure of the utterance was as follows:

nha mai lam nhan nhieu ngo (Vietnamese spelling)
[nja mai lam njan njo ngo] (rough phonetic equivalent)

Despite the somewhat artificial character of the material, the speakers were readily able to produce the sentences reading from the Vietnamese script of the syllables featuring the diacritics required for the respective tones. Recordings were auditorily checked for fluency and correct assignment of tones, and if necessary repeated.
The *F0* contours were extracted at a step of 10 ms using the PRAAT pitch estimation algorithm (© P. Boersma) and inspected visually. Especially syllables of tone types 3 and 6 exhibited extraction errors in their creaky voice parts. These syllables were checked manually period by period and the closest *F0* candidate was chosen.
Fujisaki parameters were extracted using a modified version of [5] supporting negative tone commands. *Fb* was set to 95 Hz for the male subject and to 210 Hz for the female, while *alpha* and *beta* for both speakers were set to 2 Hz and 25 Hz, respectively.

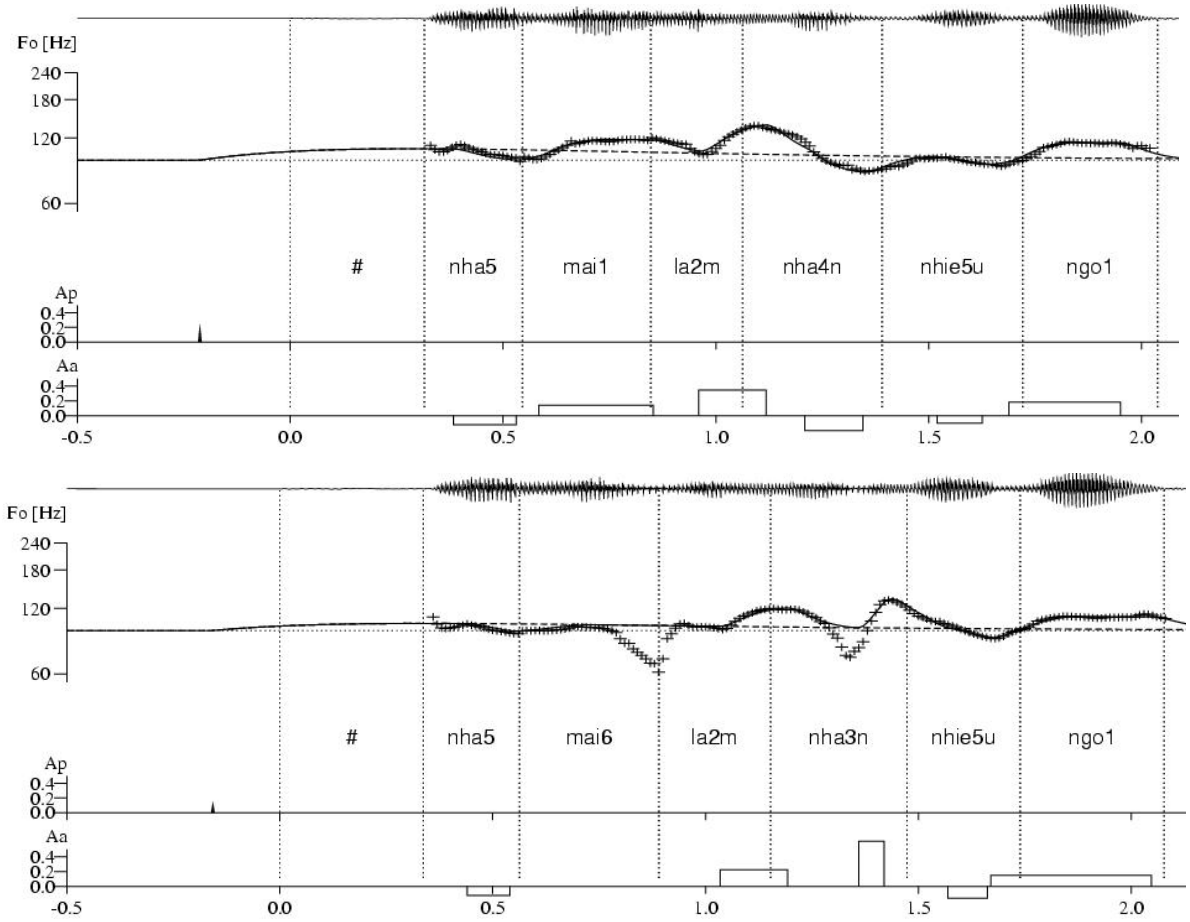

Figure 1: Examples of analysis of two utterances. Top panel: nha5 mai1 la2m nha4n nhie5u ngo1, bottom panel: nha5 mai6 la2m nha3n nhie5u ngo1. The dropping F0 contours in tones 3 and 6 associated with vocal fry is clearly visible.

3. Results of Analysis

Figure 1 shows the result of analysis for two of the utterances by the male speaker, displaying from top to bottom: The speech waveform, the *F0* contour (+signs: extracted, solid line: model-generated), the text of the utterance, and the underlying phrase and accent commands. It can be seen, that tone 1 can be modeled by using a positive tone command having roughly the duration of the syllable, whereas tone 2 requires a shorter command that begins around the middle of the syllable and has a higher amplitude. Tones 4 and 5 are modeled by using tone commands of negative polarity, the tone command amplitude for tone 5 being relatively small and often equal zero. Table 2 gives the means of amplitude and timing for the tone commands assigned to the six tones for the male and female subject, respectively. The timing is expressed relative to the syllabic duration by $T1_{rel}=(T1-t_{on})/(t_{off}-t_{on})$ and $T2_{rel}=(T2-t_{on})/(t_{off}-t_{on})$, where t_{on} and t_{off} denote the onset and offset time of the syllable, and *T1* and *T2* the tone command onset and offset time, respectively.

Table 2: Mean tone command amplitude and relative timing for the six tones, male (left) and female (right) subject.

Tone	male subject			female subject		
	Aa	$T1_{rel}$	$T2_{rel}$	*Aa*	$T1_{rel}$	$T2_{rel}$
1	.178	-.036	.892	.193	-.056	.860
2	.361	.442	1.042	.305	.358	1.072
3	.500	.456	.914	.431	.410	1.010
4	-.207	.275	.792	-.192	.221	.756
5	-.129	.294	.803	-.100	.244	.843
6	.000	-	-	.000	-	-

The table shows that tone 2 and 3 exhibit similar characteristics with respect to *F0.* As can be seen from Figure 1, bottom, Tone 3 is often connected with a voicing irregularity in the middle of the syllable where the vocal folds slacken and *F0* drops very low, often to half or even a quarter of the regular value. This phenomenon is very conspicuous with the female subject, whose vocal folds sometimes completely cease vibrating. The male speaker, however, does not always show creaky voice and an extreme drop in *F0,* but always a consistently tense manner of articulation. Tone 6 also exhibits the creaky voice that sets in after the first half of the syllable, but does not seem to present any intonational target itself, as the overall trajectory of *F0* in this tone very much depends on the type of the preceding tone, so the drop can occur from high as well as low levels (as in Figure 1, bottom) of *F0*. Different from tone 3, tone 6 has therefore not been assigned any accent command at all.

In a sequence of two syllables bearing tone 1, the underlying tone commands tend to concatenate and form a longer plateau-like shape. Similar mergers tend to occur when a tone 2 syllable is followed by tone 1 and vice versa. A sequence of tone 5 syllables can be modeled by using the phrase component of the Fujisaki model only. The assignment of a positive tone command to tone 1 as well as the assignment of a negative tone command to tone 4 is consistent with the Fujisaki model formulation for Mandarin tones 1 and 3 which are, as stated above, phonetically similar to Vietnamese tones 1 and 4. Since tones 3 and 6 can not be described exclusively by means of the Fujisaki model, the question arises, how the characteristic features of these tones could be modeled adequately in speech synthesis. The large variation in the speakers' production of these tones suggests, that not an exact property, say, for instance, an *F0* halving over the course of five pitch periods, is employed for signaling tones 3 and 6, but the presence or absence of a voicing irregularity. First resynthesis experiments suggested that the irregularity required for tone 3 could be acoustically simulated by halving the *F0* for about 40 ms in the middle of the syllabic nucleus whereas that for tone 6 required a halving of *F0* over 60 ms in the second half of the syllable. These simulated irregularities will be referred to as 'vocal fry markers'.

4. Perceptual Evaluation

In order to examine whether synthesized tones could be reliably identified by native subjects, a perception experiment was conducted using stimuli by the male speaker of three different types:

1. Natural recordings from the database
2. The same utterances as in 1. resynthesized using Fujisaki parameter yielded in the analysis, but without special markers for tones 3 and 6.
3. Utterances resynthesized from a sequence of tone 5 syllables employing the mean values of *Aa* and $T1_{rel}$ and $\boldsymbol{T2_{rel}}$ for each of the tones as given in Table 2, and vocal fry markers for tones 3 and 6. The phrase component in all of these cases was kept constant.

Whereas in type 2 the original voice quality of the syllables is still preserved, while the dropping *F0* in tones 3 and 6 is not modeled, stimuli of type 3 only employ means of F0 for signaling the tones. The irregularities required for tones 3 and 6 were simulated as described in the preceding section by appropriate *F0* halving. Resynthesis was performed by using the PSOLA resynthesis capability of PRAAT. 20 native speakers of Standard Vietnamese, most of them phonetically untrained, 7 female and 13 male, listened to 60 different stimuli, 20 for each type. Since every stimulus was presented twice, the total number of decisions was 120. Subjects were provided answer sheets with a list containing the written version of the sentence, but without the diacritics indicating the tones, and were presented the stimuli in randomized order. They were asked to complete the diacritics corresponding to the tones they perceived. At each time of presentation a stimulus was played back twice, with a pause of one second between repetitions.

5. Results of Experiment

Tables 3 to 5 display confusion matrices for the three different types of stimuli. The intended tones and the perceived tones are given in the rows and columns, respectively, along with the total number of judgments *N*. As the numbers indicate, the types of tones were not evenly distributed. This is due to the fact that from utterance to utterance not all of the syllables were varied with respect to their tone. Each syllable in the sequence therefore has one tone with which it occurs most frequently, namely mai1, ngo1, la2m, nha3n, nha5 and nhie5u. The right-most column lists the percentage of correct judgments. As can be seen, for all three types of stimuli, tones

1, 2, 3 and 5 were correctly identified in well over 90% of judgments whereas tone 4 only yields a correct assessment in about half of the judgments. The drop tone, tone 6, yields relatively poor results in the resynthesized stimuli compared with the natural stimuli. Tone 4 is consistently confused with tone 5 in all three types. Tone 5 is also the most frequent counterpart of confusions concerning tone 6, as far as synthetic stimuli are concerned. These results appear plausible, as tone 4, 5 and 6 are basically all tones with a low onset of *F0*, with tone 4 falling slightly lower than tone 5. In tone 6, in addition, creaky voice is present. Simulating the creak with the vocal fry markers in stimuli of type 3 seems to work well for tone 3, but not for tone 6. Apparently tone 6 not only requires a voicing irregularity as simulated in stimuli of type 3, but also a tense voice quality, which is not present in the original tone 5 sequence, from which stimuli of type 3 were created.

Table 3: Confusion matrix, natural stimuli, rows: intended tone, columns: perceived tone. *N* denotes the total number of judgments, '% corr.' the percentage of correct votes.

T.	1	2	3	4	5	6	*N*	% corr.
1	1305	1	2	3	6	3	1320	98.9
2	6	910	35	0	8	1	960	94.8
3	5	2	662	41	4	6	720	91.9
4	5	4	1	86	51	13	160	53.8
5	35	1	5	23	1370	6	1440	95.1
6	0	0	9	13	7	171	200	85.1

Table 4: Confusion matrix, resynthesized natural stimuli using Fujisaki parameters.

T.	1	2	3	4	5	6	*N*	% corr.
1	1300	2	2	2	9	5	1320	98.5
2	8	907	35	0	7	1	960	94.5
3	9	16	685	0	2	8	720	95.1
4	1	2	4	85	51	17	160	53.1
5	30	1	2	54	1350	3	1440	93.8
6	4	1	6	20	57	112	200	56.0

Table 5: Confusion matrix, stimuli resynthesized from tone 5 sequence using averaged Fujisaki parameters.

T.	1	2	3	4	5	6	*N*	% corr.
1	1308	1	5	0	5	1	1320	99.1
2	13	890	42	1	14	0	960	92.7
3	2	14	702	0	1	1	720	97.5
4	1	6	9	71	35	10	160	44.4
5	30	6	1	3	1395	5	1440	96.9
6	8	0	8	3	77	104	200	52.0

Since tones 4 and 6 occurred much less frequently in the material than the other tones, there was also a possibility that their poor rate of recognition was partly connected with some kind of adaptation effect in favor of the remaining tones, though this would not explain the deterioration in the identification of tone 6 from natural to resynthesized utterances. In order to test the adaptation hypothesis, a confusion matrix was calculated on a subset of syllables which did not bear the 'majority tone', that is, for instance, all syllables 'mai' which did not carry tone 1. The resulting matrix for the natural stimuli is displayed in Table 6. As becomes clear, results are similar to that in Table 3, with the exception of tone 2. The frequent confusion with tone 3 occurs in the syllable 'nhan' which has the 'majority tone' 3 being similar to tone 2 with respect to rising *F0.* This indeed shows a certain adaptation effect to be avoided in future experiments. There is also a slight possibility of an unbalanced vote for tone 1, since it is the only one that did not require adding a diacritic to the respective syllable on the answer sheet. However, except for tone 5, tone 1 is not the most confused counterpart of any of the other tones, and even in tone 5 the contribution is as little as 2.4%.

Table 6: Confusion matrix, natural stimuli, syllable subset.

T.	1	2	3	4	5	6	*N*	% corr.
1	77	0	0	0	3	0	80	96.3
2	2	93	23	0	1	1	120	77.5
3	2	1	112	1	4	0	120	93.3
4	2	2	2	81	29	4	120	67.5
5	2	0	3	0	74	1	80	92.5
6	0	0	6	8	0	106	120	88.3

6. Discussion and Conclusions

The current paper reported on a preliminary quantitative study of the syllabic tones of Vietnamese. A corpus of 'nonsense' utterances was produced in order to examine the realization of various combinations of tones while keeping the segmental context constant. It became clear that the tones of Vietnamese exhibit a bundle of coexisting features such as a specific manner of articulation and glottalization which need to be taken into account in addition to the Fujisaki parameters describing the macroprosodic *F0* contour. With respect to speech synthesis segmental irregularities can be modeled partly by introducing standard cues such as *F0* halving over a couple of periods. In the case of tone 6, however, additional articulatory cues might be necessary. Some of the tonal confusions observed, especially that for tone 4, might also be due to the 'nonsense' character of utterances employed and the resulting cognitive load. Future research will therefore, inter alia, concern tonal confusion occurring in 'real' sentences.

7. References

[1] Thompson, Laurence. *A Vietnamese Reference Grammar.* Hawaii: University of Hawaii. 1987.

[2] Fujisaki, H.; Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, **5**(4), 233-241, 1984.

[3] Fujisaki, H., Hallé, P. and Lei, H., "Application of F_0 contour command-response model to Chinese tones," *Reports of Autumn Meeting, Acoustical Society of Japan,* **1***:* 197-198, 1987.

[4] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki, and P. Charnvivit, "Perception of tone and vowel quantity in Thai," in *Proceedings of ICSLP2002*, pp. 753-756, Denver, USA, 2002.

[5] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proceedings ICASSP 2000*, vol. 1, 1281-1284, Istanbul, Turkey, 2000.