Fujisaki Model based F0 contours in Vietnamese TTS

Dung Tien Nguyen, Hansjörg Mixdorff**, Mai Chi Luong***, Huy Hoang Ngo***, Bang Kim Vu*****

*Faculty of Technology, Vietnam National University, Hanoi
nguyentiendung@hotmail.com
**Faculty of Computer Science, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
***Institute of Information Technology, Vietnamese Academy of Science and Technology
lcmai@ioit.ncst.ac.vn; nhhuy@ioit.ncst.ac.vn
****Institute of Linguistics, Vietnamese Academy of Social Sciences
bang@hn.vnn.vn

Abstract

The current paper presents preliminary work towards the integration of the Fujisaki model into the VnVoice Vietnamese TTS system, based on a set of rules to control the F0 contour. A speech corpus consisting of 20 sentences was compiled. Each of the sentences can have various meanings depending on the tone associated with a monosyllabic keyword which it contains. The corpus with a total of 46 sentences was recorded by a female speaker whose voice had also been used in the speech corpus for VnVoice, and labeled at the syllabic level. Tone contrast perception results and naturalness comparisons show that the Fujisaki model works well in modeling F0 contour of Vietnamese tones.

1. Introduction

The synthesis of near-to-natural prosodic contours is an important issue in text-to-speech. Several studies prove the strong effect of prosody on naturalness and intelligibility of synthetic speech.

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), curve (4), falling (5), and drop (6) tones. In [4] authors did a preliminary quantitative study of syllabic tones of Vietnamese with a corpus of 'nonsense' utterances. Results show that the six tones 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. The corpus used in [4] contained only utterances of voiced sounds, but no stop consonants, which are legal codas of Vietnamese syllables. In the current work based on real sentences, we found that rising and drop tones of syllables ending with stop consonants have F0 contour similar to rising and falling tone of other syllables but they rise or drop more sharply. Drop tones of syllables ending with stop consonants are never accompanied by the breathy quality of that tone observed in open syllables. Therefore, they were modeled by a positive command and a negative command, respectively.

The current study aims at:

- Examination of tone confusions in real sentences.
- Evaluation of intelligibility and naturalness of VnVoice as compared to natural speech and synthetic speech using Fujisaki model based F0 contours.

2. The Current Vietnamese TTS system

Fundamental speech units of the VnVoice system are demisyllables which are yielded by dividing a syllable into initial onset consonant and final rhyme unit consisting of nucleus and coda. Therefore each Vietnamese syllable may be considered a combination of Initial, Final and Tone components [1].

<table>
<tr><td colspan="4">Tone</td></tr>
<tr><td rowspan="2">Initial</td><td colspan="3">Final</td></tr>
<tr><td>Onset</td><td>Nucleus</td><td>Coda</td></tr>
</table>

The Initial component is always a consonant, or it may be omitted in some syllables. There are 21 Initials and 155 non-tonal Final components in Vietnamese, the Final may be decomposed into Onset, Nucleus and Coda. Final Onset and Coda are optional and may not exist in a syllable. The Onset is a [w] semivowel or a zero. The Nucleus consists of a vowel or a diphthong, and the Coda is a consonant or a semi-vowel. The Initial, Tone, Onset, Nucleus and Coda may be combined together to make a syllable; however not all combinations are possible. The Tone is associated with the syllable as a whole. For the current VnVoice system, Vietnamese syllables were recorded at a 10 kHz sampling rate and 16-bit resolution. Speech units are labeled with Initials and tonal Finals, there are totally about 900 units including 200 consonant-vowels labeled with diphones and 700 tonal Finals with monophones. TD-PSOLA technique is used for synthesizing, initials and finals are concatenated to create a sound chain, some techniques are used to smooth the concatenation point. In the current system, however, F0 and syllable duration, as well as intensity cannot be manipulated. This, of course, affects the naturalness of synthetic utterances. As a consequence, the current work aims at introducing this kind of prosodic control.

The Fujisaki model [2] has been successfully applied for decomposing F0 contours in many languages like Japanese, German, and Finnish and in some tonal languages like

Chinese, Thai, and Vietnamese [3,4]. In the current work, we are integrating the model into the current Vietnamese TTS for modeling the F0 contour of the six tones by rules in order to reduce the necessary number of speech units, as well as increase the versatility and expressiveness of the system by manipulating intonation.

3. Speech Material and Method of Analysis

In order to examine the discriminative ability of listeners with regard to tone contrasts in real sentences, a set of 20 carrier sentences was selected. Each sentence contains a monosyllabic keyword that can convey different meanings depending on the lexical tone, with between two and four variants. We had a total number of 46 real sentences recorded by a phonetically trained native speaker of Northern Vietnamese, whose voice had also been used in VnVoice.

Here is an example of a carrier sentence: "Toi1 mua1 qua4 *dua1*| *dua2*| *dua5* nay5 ngoai5 cho6". In English, it means "I bought this *melon*| *pineapple*| *coconut* at the market". The monosyllabic keyword is "*dua*" which has three different meanings with 1, 2 and 5 tones.

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 215 Hz, while alpha and beta were set to 2 Hz and 25 Hz, respectively.

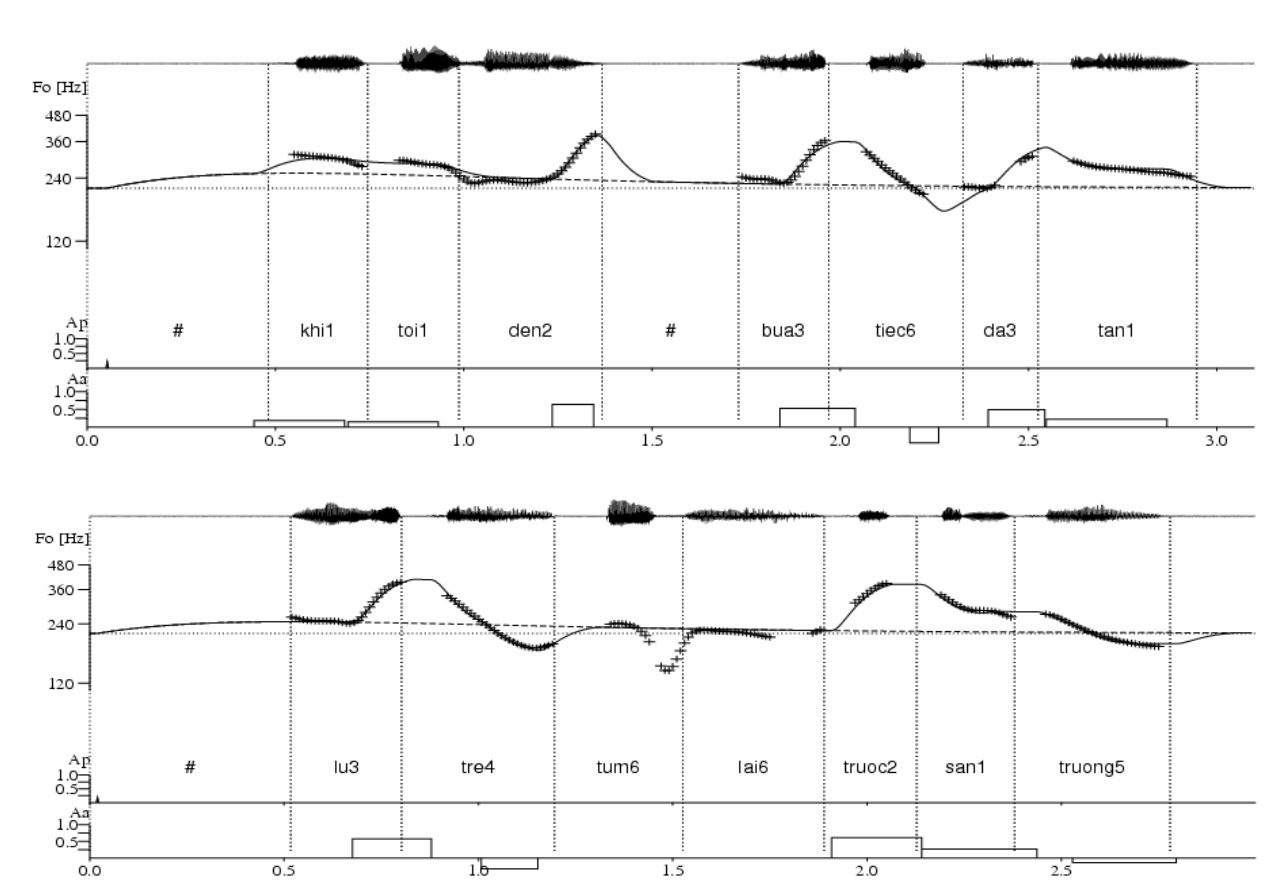

Figure 1: Examples of analysis of two utterances. Top panel: khi1 toi1 den2 bua3 tiec6 da3 tan1, bottom panel: lu3 tre4 tum6 lai6 truoc2 san1 truong5.

4. Result of analysis

Figure 1 shows the result of analysis for two utterances, 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 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. Tones 6 ending with stop consonants are modeled by a negative tone command. Table 1 gives the means of amplitude and timing for the tone commands assigned to the six tones. 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 1: Mean tone command amplitude and relative timing for the six tones. 2' and 6' is 2 and 6 tones ending with stop consonants.

Tone	Aa	$T1_{rel}$	$T2_{rel}$
1	.218	-.09	.86
2	.523	.61	1.04
3	.556	.53	1.11
4	-.341	.45	.91
5	-.132	.37	1.07
6	.00	-	-
6'	-.378	.42	.70
2'	.617	.16	.84

Simple duration rules were created by calculating duration means and standard deviations with respect to their tones and positions in the utterance. The results are shown in Tables 2 and 3. The duration is measured in 10 milisecond units. The statistical results show that the syllable's duration depends much more on tone than position. The syllable with tone 3, 6' and 2' are shorter than the others. It dues to tone 6' and 2' ending with stop consonants and tone 3 require a glottal stop at middle of the syllable so that it is hard to lengthen them.

Table 2: Mean duration and standard deviation of syllables with respect to their tones.

Tone	1	2	3	4	5	6	6'	2'
mean	31	33	27	33	34	31	28	27
std	7	5	4	6	6	5	3	5

Table 3: Mean duration and standard deviation of syllables with respect to their positions.

Pos	1	2	3	4	5	6	7	8	9	10
mean	27	34	33	33	31	30	30	32	31	31
std	4	7	7	6	5	5	6	4	4	3

5. Perceptual Evaluation

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

1. Natural recordings from the database
2. Utterances from VnVoice.
3. The same utterances as in 1) resynthesized using Fujisaki parameter yielded in the analysis, but without special markers for tones 3 and 6.
4. Utterances resynthesized from a sequence of level tone syllables of VnVoice employing the mean values of Aa and $T1_{rel}$ and $T2_{rel}$ for each of the tones as given in Table 1, vocal fry markers for tones 3 and 6, and the mean duration of syllables with respect to their tones given in Table 2. 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. 20 native speakers of standard Vietnamese, most of them phonetically untrained, 7 female and 13 male, listened to 104 different stimuli, 26 for each type. Subjects were show the carrier sentences without the keyword and all syllables of keyword with different tones, and were asked to choose the syllable they perceived. At each time of presentation a stimulus was played back twice.

In order to estimate the naturalness of VnVoice, and synthetic tones, and synthetic durations, a naturalness ranking was conducted on four different types of stimuli:

1. Utterances from VnVoice resynthesized using the mean duration of tones.
2. The same as type 2 in the perception experiment.
3. The same as type 3 in the perception experiment.
4. The same as type 4 in the perception experiment.

The five point scale means that stimuli was played of varying quality and randomized order to the subjects and ask whether they found them excellent (5), good (4), fair (3), poor (2), very poor (1), with respect to naturalness. 10 native speakers of Standard Vietnamese, most of them phonetically untrained, 5 female and 5 male, listened to 40 different stimuli, 10 for each type. At each time of presentation, subjects were ask to play a natural utterance and a synthetic utterance as many time as they want, then marks the point. Resynthesis was performed by using the PSOLA resynthesis capability of PRAAT.

6. Results of experiment

Tables 4 to 7 display confusion matrices for the four 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. The high correct percentages of all experiences show that we can manipulate F0 without significant loss of intelligibility.

Table 4: Confusion matrix, natural stimuli, rows: intended tone, columns: perceived tone. N denote the total number of judgments, '% Corr.' the percentage of correct votes.

T.	1	2	3	4	5	6	N	% Corr.
1	140	0	0	0	0	0	140	100.0
2	0	80	0	0	0	0	80	100.0
3	0	0	40	0	0	0	40	100.0
4	0	0	0	80	0	0	80	100.0
5	0	0	0	0	120	0	120	100.0
6	0	0	0	0	0	60	60	100.0

Table 5: Confusion matrix, resynthesized natural stimuli using averaged Fujisaki parameters.

T.	1	2	3	4	5	6	N	% Corr.
1	140	0	0	0	0	0	140	100.0
2	0	80	0	0	0	0	80	100.0
3	0	0	40	0	0	0	40	100.0
4	0	0	0	74	5	1	80	92.5
5	0	0	0	0	120	0	120	100.0
6	0	0	0	1	0	59	60	98.3

Table 6: Confusion matrix, VnVoice's stimuli.

T.	1	2	3	4	5	6	N	% Corr.
1	140	0	0	0	0	0	140	100.0
2	0	80	0	0	0	0	80	100.0
3	0	0	40	0	0	0	40	100.0
4	0	0	0	76	4	0	80	95.0
5	0	0	0	0	120	0	120	100.0
6	0	0	0	0	0	60	60	100.0

Table 7: Confusion matrix, stimuli resynthesized from VnVoice's level tone sequence using averaged Fujisaki parameters, and average mean of tones.

T.	1	2	3	4	5	6	N	% Corr.
1	140	0	0	0	0	0	140	100.0
2	0	80	0	0	0	0	80	100.0
3	0	0	40	0	0	0	40	100.0
4	0	0	0	76	4	0	80	95.0
5	0	0	0	0	120	0	120	100.0
6	0	1	0	0	0	59	60	98.3

Figure 2 shows the naturalness scores of four different types of stimuli compared with natural utterances. From left to right: Utterances resynthesized from natural utterances using Fujisaki model rules (Nat-P), VnVoice's utterances (VnVoice), utterances resynthesized with modified durations from VnVoice (VnVoice-D), utterances resynthesized from mean duration level tone syllables of VnVoice and averaged Fujisaki parameter (VnVoiceLT-PD). The result show that the F0 contour generated by the Fujisaki model generally work well. The fact that VnVoice with modified durations (VnVoice-D) is rated worse than the original VnVoice can be explained by the fact that durations in fluent speech are generally shorter than in the inventory of VnVoice which is based on syllables uttered in isolation. When these units are compressed, the underlying natural F0 contours sound extremely exaggerated, leading to a reduced perceived naturalness. When in turn only level tone units are used (VnVoiceLT-PD) and the F0 contours are modified depending on the rules, we observe a comparable degradation in naturalness. This, however, as careful examination shows, is rather due to the segmental naturalness degradation caused by the PSOLA method, and not caused by unnaturally sounding F0 contours. Furthermore, the duration rules are too simple since they are based on too few instances. Therefore, the naturalness does not to increase when we use duration rules. Results suggest that a future implementation of VnVoice should be based on units taken from continuous utterances in order to facilitate prosodic manipulation.

7. Conclusions

The current paper presented preliminary work towards the integration of the Fujisaki model into the VnVoice Vietnamese TTS system. Although the approach is currently based on a simple set of rules, the tone contrast perception results and naturalness ranking shows that the Fujisaki model can be successfully employed. Consequently, we can reduce the necessary number of speech units, as well as increase the versatility and expressiveness of the system by manipulating intonation. Our results indicate, however, that level tone syllables might not be the best building blocks for a future system, since they are usually uttered with a relatively high F0 in the upper part of a speaker's range. PSOLA manipulation of these segments in order to create low tones might lead to reduced segmental quality. As an alternative, syllables uttered at an F0 in the middle of a speaker's range might be an option. Our duration rules were too simple to bring improvement. In further research, we will build a large database to analyze Vietnamese prosody in various contexts in order to be able to build statistical models for predicting F0 contours, durations, as well as intensity.

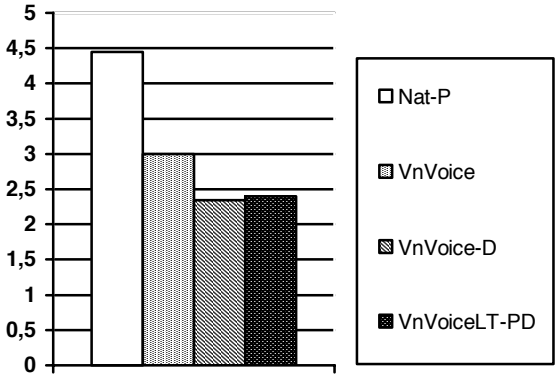

Figure 2: Mean naturalness scores of synthetic utterances compared with natural utterances.

8. References

[1] Đoàn Thiện Thuật. *Ngữ âm tiếng Việt*. Nhà xuất bản đại học quốc gia Hà Nội, In lần thứ 2, 2003.

[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] C. Wang, H. Fujisaki, S. Ohno, T. Kodama, "Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command-response model", *Proceeding of the 6th Eurospech*, vol. 4, pp. 1655-1658, 1999.

[4] Mixdorff, H., Hung, N. et al., "Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese". *In Proceedings of Eurospeech2003, Geneva*, 2003.

[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.