Analysis of Vietnamese Tones to Optimize Database in Speech Synthesis Using Unit Selection Method

Vu Duc Lung
Faculty of Computer Engineering
University of Information Technology –
Vietnam National University, Ho Chi Minh City
lungvd@uit.edu.vn

Nguyen Phuoc Loc
Faculty of Computer Science
University of Information Technology –
Vietnam National University, Ho Chi Minh City
locnp0209@gmail.com

Cao Van Hung
Faculty of Computer Engineering
University of Information Technology –
Vietnam National University, Ho Chi Minh City
hungcv@uit.edu.vn

Nguyen Viet Quoc
Faculty of Computer Science
University of Information Technology –
Vietnam National University, Ho Chi Minh City
viet.quoc.2569@gmail.com

Abstract *— This paper presents a novel approach to optimize data in Vietnameses speech synthesis using Unit Selection method. First, we conduct analysis of Vietnamese tone using Fujisaki model to find out the parameters of fundamental frequency contours (F0 contours) influencing on Vietnamese vowels while speech is expressed. Next, analysis, testing, and evaluation of the effects on the vowel are performed. After that, the data of unit selection speech synthesis system is optimized by recording the vowel with a level tone. As a result, when a Vietnamese word is synthesized, it will be synthesized with a level vowel first followed by being adjusted the parameter of F0 contour to create a word with appropriate tone. With this approach, recorded data can be reduced up to 64,44% while sound quality is insignificantly affected.*

Keywords – *Vietnamese speech synthesis; Fujisaki Model; F0 contours; fundamental frequency; data optimization; unit selection; level tone*

I. INTRODUCTION

Unlike many other languages like English, Italian, Spanish, etc. which are multi-syllabic language and no tone, Vietnamese is monosyllabic and tonal language. Vietnamese has 6 tones: level, falling, rising, curve, broken, and drop tone. Tone plays an important role in deciding meaning of a word. The same word with the same phoneme but different tones will result in different pronunciation and semantics. Therefore, studying on the impact of tone for a Vietnamese word has a tremendous importance for speech synthesis. This paper entails our testing process, which studies the application of the results obtained on speech synthesis system using Unit Selection method to optimize this system.

Vietnamese speech synthesis system using unit selection method requires a large amount of recorded data. This lead to the system needs a very large data storage space and a large memory. Consequently, the system works slow down and lacks flexibility. It is very hard to work in mobile platforms, embedded platforms which have a small data storage space and a small memory. It is also very difficult to operate in network environments. In order to optimize the recorded data, we firstly study and analysis Vietnamese tone. Then the diagram of simplified F0 contours movements is issued. Vietnamese words are synthesized from level words using Fujisaki model by adjusting the parameter of F0 contours based on this diagram. Theoretically, the recorded data can be reduced 5 times.

The rest of the paper is organized as followed. A comprehensive literature review on Vietnamese tone will be presented in section II. Next, the use of Fujisaki model to analyze, synthesize, and evaluate Vietnamese tone will be elaborated in section III followed by the optimization of data in speech synthesis using unit selection method in section IV. Finally, conclusions and future research directions are given.

II. EXPLORING OF VIETNAMESE TONES

The structure of a word in Vietnamese consists of 3 main components: vowels, consonants and tones. In the general form [1] [2] [3], a word in Vietnamese is denoted as follow:

	(Prosody)	
(Pre-consonant)	(Vowel)	(Post-consonant)

Thus, a word in Vietnamese will have four real cases to demonstrate:

- Case 1: it has only one vowel segment. For example: *U* (mother). This word is formed from the vowel */u/*.
- Case 2: it has Pre-consonant + vowel segment. For example: *Ba* (father). This word is formed by the Pre-consonant */b/* and the vowel */a/*.
- Case 3: it has vowel + Post-consonant segment. For example: *Anh* (brother). This word is formed by the vowel */a/* and the Post-consonant */nh/*.

978-1-4673-5604-6/12/$31.00 ©2012 IEEE

- Case 4: it has full 3 segment: Pre-consonant + vowel + Post-consonant. For example: *Nhanh* (fast). This word is formed by the Pre-consonant */nh/*, the vowel */a/* and the Post-consonant */nh/*.

In order to find out the rules which impact the tones of any words in Vietnamese we have experimented with four above-mentioned cases.

Similar to many other tonal languages [5] such as Mandarin, Japanese, Thai, etc, characteristics of tone in Vietnamese is also fundamental frequency F0. Vietnamese has 6 tones that are Level, Rising, Falling, Drop, Broken and Curve Tone. Each of them has very different F0 contours. Impact of these tones on a word in Vietnamese will create the accent words from which we can distinguish them. The basic parameters of tone include fundamental frequency, intensity and length. However, unlike the fundamental frequency, the intensity and the length are not importance in determining the characteristics of tone. Depending on the context and the emotions in communication language, the intensity and length can be altered. Therefore, these characteristics belong to intonation of sentence and just a phenomenon coming with tone. The main characteristic which is determinant of tone of a word is the fundamental frequency F0.

To confirm this, we explored all four real cases of words which are impacted by 6 tones for each case above in turn. We recorded and analyzed speech data from 10 people including 5 men, and 5 women with ages between 20 and 50. The recording consists of two main voices: northern and southern.

Case 1: Exploring the case that a word has only one vowel segment. Here we choose the vowel */ a /* to represent this case, thus this vowel segment will be affected by 6 Vietnamese tones which are */a/*, */á/*, */à/*, */ạ/*, */ả/*, */ã/*. We recorded these 6 words and used Speech Analyzer tool version 3.1 of SIL International *(Summer Institute of Linguistics, Inc.)* [4] to analyze the recorded signal as shown in Fig. 1. The upper part of Figure 1 shows the waveform of 6 words */a/*, */á/*, */à/*, */ạ/*, */ả/*, */ã/* and the lower is the corresponding fundamental frequency of each word.

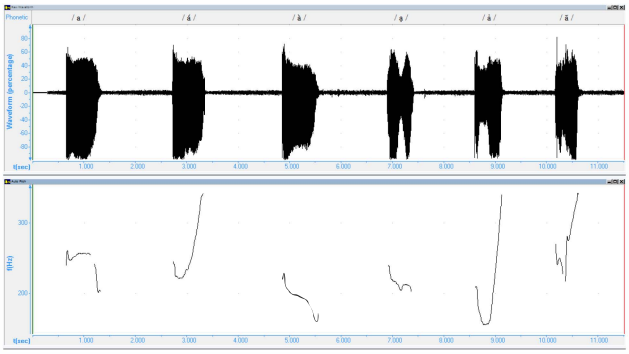

Fig. 1. Exploring on the impact of tones to 6 words: */a/*, */á/*, */à/*, */ạ/*, */ả/*, */ã/*

Analogously, selecting the words */ba/*, */an/*, and */ban/* for the cases 2, 3, and 4 respectively. We obtain similar results as in the case 1.

By above case study, we found that the tones are mainly affected by the vowel segment as follows:

- Level tone: 230Hz for the female voice and 100Hz for the male voice. F0 contours seem to be a horizontal line.
- Rising tone: 200Hz → 350Hz for female voice and 100Hz → 130Hz for a male voice. F0 contours tend to soar, performing rising tone.
- Falling Tone: about 220Hz → 200Hz for female voice and about 80Hz → 65Hz for male voice. F0 contours tend to go down, performing falling tone. Starting F0 contours of falling tone lower level tone.
- Drop tone: about 230Hz → 200Hz for female voice and about 90Hz → 60Hz for male voice. F0 contours, starting from 220Hz (90Hz for male voice), go down very deeply, performing drop tone. The time of F0 contour of drop tone is shorter than those of other tones.
- Curve Tone: about 150Hz → 330Hz for female voice and about 65Hz → 105Hz for a male voice. Original F0 contour goes down, performing falling tone and then soars high, performing rising tone.
- Broken Tone: about 220Hz→320Hz for female voice and about 80Hz → 140Hz for male voice.Starting from 220Hz (80Hz for male voice), F0 contours go up and down to perform drop tone (falling or level tone). After that sound segment is broken then soared high, performing rising tone.

In short, the movement of F0 contours can be visualized as in Fig. 2.

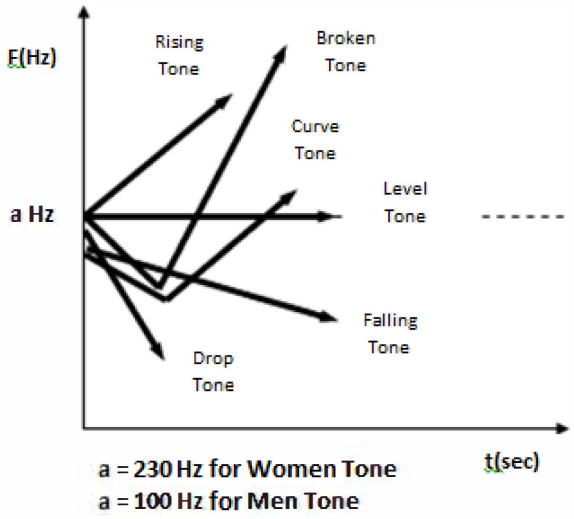

Fig. 2. Diagram of simplified F0 contours movements for female and male tones

III. USING FUJISAKI MODEL TO ANALYZE VIETNAMESE TONES

Fujisaki model [6] developed by Fujisaki and colleagues has been successfully used to analyze the Japanese intonation. By modifying some parameters on this model, it can also be used to analyze intonation and tone of the tone languages such as Chinese, Thai [5]. Mixdorff applied Fujisaki model to create MFGI (Mixdorff - Fujisaki model of German Intonation) [7] [8] to generate intonation for German Text to Speech System. Through some minor changes Fujisaki model can also be used to analyze the F0 contours in English, Spanish, and Greek [9]. The Fujisaki model can be summarized in Figure 3.

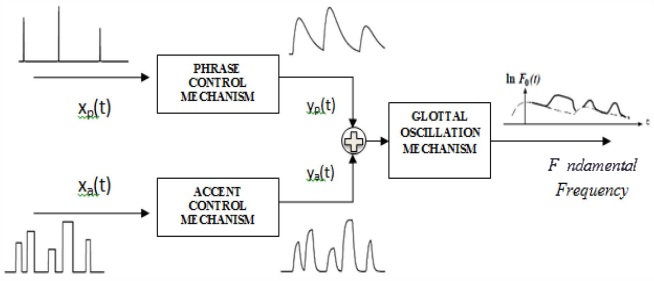

Fig. 3. Fujisaki Model generating *F0* contours.

The input of this model includes 2 commands: Phrase command composed by Dirac impulses and Accent command composed by stepwise functions. Phrase commands $x_p(t)$ are filtered by a linear processing system known as the Phrase Control Mechanism while Accent commands $x_a(t)$ are handled by linear processing system called the Accent Control Mechanism. Phrase Contribution y_p, which models the pitch baseline, accounts for speaker declination and it is characterized by a fast rise followed by a slower fall. Accent Contribution y_a modeling smaller-scale prosodic variations accounts for accent components.

The superposition of a Phrase component and an Accent component resulting in *F0* contours based on the following equation:

$$ln(F0) = ln(F_b) + \sum_{k=1}^{N_p} A_{p,k} \cdot g_p(t - T_{p,k}) +$$

$$+ \sum_{k=1}^{N_a} A_{a,k} \cdot [g_a(t - T'_{a,k}) - g_a(t - T''_{a,k})] \quad (1)$$

which :

$$g_p(t) = \begin{cases} \alpha^2 t.exp(-\alpha t), & \forall t \geq 0 \\ 0, & \forall t \geq 0 \end{cases} \quad (2)$$

and

$$g_a(t) = \begin{cases} 1 - (1 + \beta t).exp(-\beta t), & \forall t \geq 0 \\ 0, & \forall t \geq 0 \end{cases} \quad (3)$$

where $g_p(t)$ is the impulse-response function of the phrase control mechanism and $g_a(t)$ is the step-response function of the accent control mechanism. The symbols used in Equation (1), (2), (3) are entailed as following:

F_b – asymptotic value of the fundamental frequency in absence of accent-command;

N_p – number of phrase-commands;

N_a – number of accent-commands;

$A_{p,k}$ – magnitude of the k[th] phrase-commands;

$A_{a,k}$ – magnitude of the k[th] accent-commands;

$T_{p,k}$ – timing of the k[th] phrase-commands;

$T'_{a,k}$ – onset of the k[th] accent-commands;

$T''_{a,k}$ – end of the k[th] accent-commands;

α – Natural angular frequency of the phrase control mechanism to the phrase-commands;

β – Natural angular frequency of the phrase control mechanism to the accent-commands;

In the previous section we explored the shape and issued the movement of *F0* contours of six tones in Vietnamese language. In order to confirm that the fundamental frequency *F0* is the main feature of the tone we use FujiParaEditor of Mixdorff to adjust the *F0* contours and re-synthesize the sound. This tool which was developed by Mixdorff and colleagues in 2010 automatically extracted *F0* contours based on the Fujisaki model [7][8]. At first, we record the words with the level tone. After that, we adjust these words based on the three expressions of the Fujisaki model by inputting the appropriate Fujisaki parameters in the FujiParaEditor tool so that they are consistent with the diagram of simplified *F0* contours in Figure 2. Finally, after changing the parameters and re-synthesizing, we obtain the waveform and *F0* contours of the words of the remaining five tones from the word with level tone.

We recorded words at level tone for 4 cases mentioned in Section II. They are tabulated in Table 1.

TABLE 1. LIST OF THE LEVEL TONE WORDS RECORDED FOR TEST.

No	Case 1		Case 2		Case 3		Case 4	
	Word	Phonetic	Word	Phonetic	Word	Phonetic	Word	Phonetic
1	***a***	/a:/, /a/	***an***	/a:n/	***ba***	/ba:/	***ban***	/ba:n/
2	***i, y***	/i/, /j/	***in***	/in/	***vi***	/vi/	***tin***	/tin/
3	***ê***	/e/, /ɜ/	***êm***	/em/	***lê***	/le/	***nên***	
4	***u***	/u/, /w/	***ung***	/uŋ/	***vu***	/vu/	***tung***	/tuŋ/
5	***ư***	/ɨ/	***ưng***	/ɨŋ/	***tư***	/tɨ/	***dưng***	/dɨŋ/
6	***o***	/ɔ/, /w/, /aw/	***ong***	/ɔŋ/	***co***	/kɔ/	***bon***	/bɔn/
7	***ô***	/o/, /ɜ/, /ɜw/	***ôn***	/on/	***tô***	/to/	***công***	/coŋ/
8	***ơ***	/ə:/, /ɜ/	***ơn***	/ə:n/	***mơ***	/mə:/	***đơn***	/də:n/
9	***e***	/ɛ/	***em***	/ɛm/	***ke***	/kɛ/	***ben***	/bɛn/
10	***ui***	/uj/			***tui***	/tuj/		
11	***oa***	/ɔa:/	***oan***	/ɔa:n/	***loa***	/lɔa:/	***đoan***	/dɔa:n/
12	***iu***	/iw/			***liu***	/liw/		
13	***ôi***	/oj/			***tôi***	/toj/		
14	***eo***	/ɛw/			***leo***	/lɛw/		
15			***uông***	/uɜŋ/			***muôn***	/muɜn/
16	***ua***	/uɜ/			***đua***	/duɜ/		
17	***ây***	/ɜj/			***dây***	/zɜj/		
18	***oai***	/ɔa:j/			***hoai***	/hɔa:j/		
19	***ươi***	/ɨɜj/			***cươi***	/kɨɜj/		
20	***iêu, yêu***	/iɜw/			***tiêu***	/tiɜw/		

After recording the words in Table 1 with all of the level tone and adjusting *F0* contours to produce the words with other tones, we have 10 volunteers to help whether they can recognize exactly the resulting words or not. The results are presented in Table 2 in term of recognition accuracy:

TABLE 2. ACCURACY OF RECOGNIZING THE WORDS OBTAINED BY MODIFYING THE WORDS IN TABLE 1 BY 10 VOLUNTEERING LISTENERS

Case		Rising Tone	Falling Tone	Curve Tone	Broken Tone	Drop Tone
1	Vowel	97%	91%	78%	84,4%	79,89 %
2	Consonant + vowel (monophthong + diphthong)	94,3%	87,75 %	76,7%	73,8%	83,3 %
	Consonant + triphthong					
3	Vowel (monophthong + diphthong) + Consonant	93,3%	79,4%	85,3%	82,3%	77,5 %
	Triphthong + Consonant					
4	Consonant + monophthong + Consonant	98,6%	86,3%	92%	83,3%	88,8 %
	Consonant + diphthong + Consonant					
	Consonant + triphthong + Consonant					
Accuracy of recognizing the words at at Level Tone		95,8%	86,1%	83%	80,95 %	82,4 %
The average accuracy		85,65%				

From Table 2, the average recognition accuracy of words generated from the original words at level tone is of 85.65%.

IV. DATA OPTIMIZATION IN SPEECH SYNTHESIS MODEL BASED ON UNIT SELECTION METHOD:

In speech synthesis system using unit selection method, pieces of recorded speech that are stored in a database are concatenated to make a speech sentence. The recorded data may be one or all of the following sound segments: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences [14] [12] [13] [11]. To synthesize the desired target utterance the system will select the most appropriate unit from the database to concatenate followed by smoothing using many different smoothing methods.

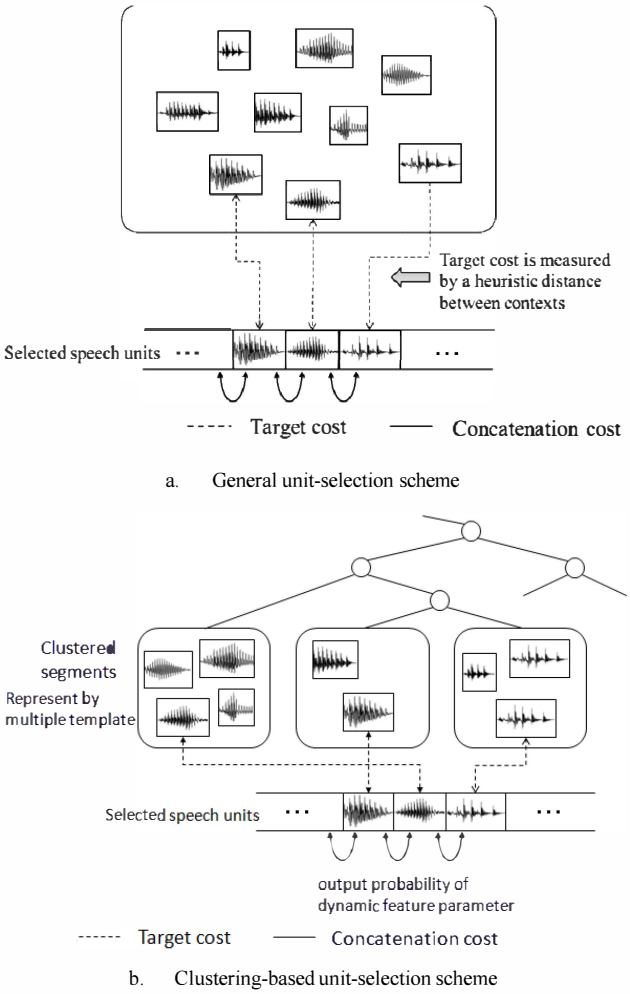

Fig. 4. An overview of general unit-selection scheme and clustering-based unit-selection scheme. Solid lines represent concatenation costs and dashed lines represent target costs.

There are two basic techniques used in speech synthesis system by unit selection methods: General and clustering-based techniques which are illustrated in Fig. 4a and 4b

respectively. General selection had existed in the ATR v-talk system since 1992 (Sagisaka, et al) and was introduced by Hunt and Black 1996 [11]. The clustering-based technique was introduced in 1995 by Donovan and Woodland [12]. Theoretically, two techniques do not differ significantly [13]. However, the clustering-based technique performs clustering contexts in advance before selecting each unit from a cluster.

Since a speech synthesis system based on Unit Selection method requires a large amount of recorded data, data optimizing is very essential to help the system to reduce data size, to process faster and to be more flexible. The overview of this system is showed in Fig. 5. First, Text Analysis Module analyzes text input, and assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme. Next, recorded data is reduced by adjusting the *F0* contours of level tone words to give out the other tone words using Fujisaki Model. After selecting the most appropriate units to concatenate, the target utterance is generated by Speech Generation Module. The accuracy of system is affected slightly but remained at 85,65%. The system becomes more flexible thanks to a smaller database which can be implemented on mobile and embedded devices with low storage memory. It can also operate in the network environments and the cloud services, etc.

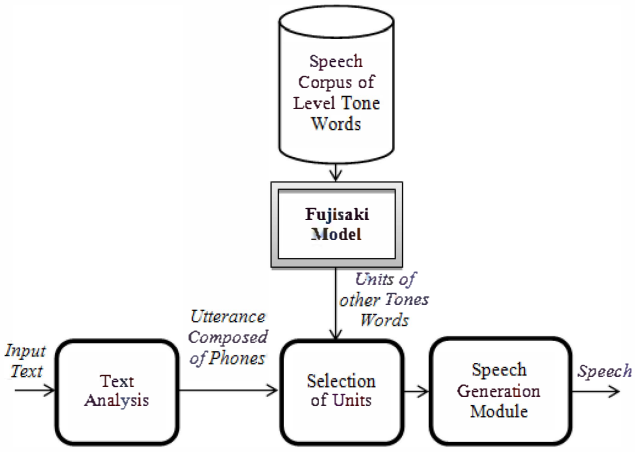

Fig. 5. The overview of Vietnamese speech synthesis system using Fujiaki Model

Vietnamese has 22 pre-consonants, 16 vowels (including 13 monophthong and 3 diphthong), and 8 post-consonants [1][2][3][15]. Because tone only affects on the vowel segment, we have to collect every recording vowel in 6 cases with these vowels, which is very labour-intensive. For example, with / a / we have to record six cases which are */a/*, */á/*, */à/*, */ạ/*, */å/*, */ã/*. Instead, we can record / a / only, then adjust and re-synthesize words with different tones. Using this method, the amount of data can be reduced by 5 times in theory.

Totally, Vietnamese language consists of about 7218 single words [1] [2] [3] [15] [16]. We have listed and arranged each single word with its appropriate case (pronounceable and mean) as the same as in Section II. To optimize the recording data for system, we just need to record the words at level tone. Then, we use the algorithm in Section III to synthesize the words at other tones before feeding them into the system for speech synthesis. As a result, the amount of data need to be recorded is significantly reduced, which help to save a lot of effort. Percentages of data reduction using above optimization method are tabulated in Table 3:

TABLE 3. PERCENTAGE OF DATA REDUCTION USING OPTIMIZATION METHOD

Case		The number of single words need to be recorded		Reduction percentage
		No optimization	Optimization	
1	Vowel	116	43	62.93%
2	Consonant + vowel (monophthong + diphthong)	2458	635	74.17%
	Consonant + triphthong	229	98	57.21%
3	Vowel (monophthong + diphthong) + Consonant	171	87	49.12%
	Triphthong + Consonant	1	1	0%
4	Consonant + monophthong + Consonant	3350	1269	62.12%
	Consonant + diphthong + Consonant	847	415	51.00%
	Consonant + triphthong + Consonant	46	19	58.70%
	Total	7218	2567	64.44%

From Table 3, using data optimization method, the amount of recorded data required for the system is reduced up to 64.44%.

V. CONCLUSION

In this paper, we conducted experiments to discover the effect of *F0* contours on tone of Vietnamese words. Consequently, from a word original level tone, we can synthesize the words with other tones based on the Fujisaki model. We have shown that the synthesized words are similar to their respective real recorded data with up to 85.65% recognition accuracy. Using this finding, we proposed an optimization method to reduce the amount of recorded data required for a speech synthesis system with up to 64.44%. For future work, we will focus on improving quality of the synthesized to minimize their affect on performance of the system.

ACKNOWLEDGMENT

We are very grateful to the Advanced Program of the University of Information Technology, Vietnam National University – HCMC, for its valuable grant to create this article.

REFERENCES

[1] Nguyễn Thiện Giáp. Vietnamese Lexical, University and technical secondary schools of Hanoi Publishers, 1985 (in Vietnamese)

[2] Mai Ngọc Chừ. Basis of linguistic and Vietnamese, Education Publishers, H. 1997, pp 91–105 (in Vietnamese)

[3] Mai Ngọc Chừ, Vũ Đức Nghiệu, Hoàng Trọng Phiến. Basis of linguistic and Vietnamese, Education Publishers, H. 1997 (in Vietnamses)

[4] http://www.sil.org/computing/sa/

[5] Suphattharachai Chomphan. Fujisaki's Model of Fundamental Frequency Contours for Thai Dialects. Journal of Computer Science 6 (11): 1263-1271, 2010

[6] Fujisaki, H. and H. Sudo. A model for the generation of fundamental frequency contours of Japanese word accent. J. Acoust. Soc. Jap., 57: 445-452, 1971

[7] Mixdorff, H. and H. Fujisaki. Automated quantitative analysis of F0 contours of utterances from a German ToBI-labeled speech database. Proceeding of the Eurospeech, Sept. 22-25, ISCA, Rhodes, Greece, pp: 187-190, 1997

[8] 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.

[9] H. Fujisaki, and S. Ohno, The Use of a Generative Model of F0 Contours for Multilingual Speech Synthesis. Fourth International Conference on Signal Processing, Vol. 1, pp. 714–717, 1998.

[10] Kota Yoshizato, Hirokazu Kameoka, Daisuke Saito, Shigeki Sagayama. Statistical approach to Fujisaki-Model parameter estimation from speech signals and its quantitative evaluation. Speech Prosody, 6th International Conference

[11] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In ICASSP 1996, pp.373–376.

[12] R. Donovan and P. Woodland, "Improvements in an HMM-based speech synthesiser," in Eurospeech, 1995, pp. 573–576.

[13] Heiga Zena, Keiichi Tokuda, Alan W. Black. Statistical Parametric Speech Synthesis. Speech Communication, Volume 51, Issue 11, November 2009, Pages 1039–1064.

[14] K. Tokuda, H. Zen, A.W. Black, An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sept. 2002.

[15] Đoàn Thiện Thuật. Vietnamese Phonetic, Hanoi National University Publishers, 1999 (in Vietnamese)

[16] Hoàng Phê. Vietnamese Dictionary, Encyclopedia Dictionary Publishers, 2000 (in Vietnamese)