Acoust. Sci. & Tech. **25**, 5 (2004)

PAPER

Vietnamese Text-To-Speech system with precise tone generation

Tu Trong Do and Tomio Takara*

Department of Information Engineering, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa, 903–0213 Japan

(*Received 16 September 2003, Accepted for publication 3 February 2004*)

Abstract: We propose a Vietnamese Text-To-Speech (**VieTTS**) system which is a parametric and rule-based speech synthesis system. Fundamental speech units of this system are demisyllables with Level tone. VieTTS uses a source-filter model for speech production and a Log Magnitude Approximation (LMA) filter as the vocal tract filter. We chose the Hanoi dialect for VieTTS. Tone synthesis of Vietnamese is implemented by using fundamental frequency (F0) patterns and power pattern control. F0 is the most important factor in Vietnamese tone synthesis and the power control strongly affects Broken and Drop tones. Applying power control for tone synthesis is unique for Vietnamese compared to other tonal languages such as Chinese and Thai. This new result is confirmed by listening tests with a reasonable listening correct rate.

Keywords: Vietnamese, Text-To-Speech, Synthesis by rule, Tone, Cepstrum

PACS number: 43.72.Ja [DOI: 10.1250/ast.25.347]

1. INTRODUCTION

In spite of the development of speech technology, there has been very little research on Vietnamese speech processing [1–3], particularly on speech synthesis. In this paper, a Vietnamese Text-To-Speech (VieTTS) system is proposed. VieTTS is a parametric and rule-based synthesis system, in which the fundamental speech units are demisyllables with Level tone. VieTTS uses a source-filter model [4] for speech production and a Log Magnitude Approximation (LMA) filter [5] as the vocal tract filter.

Vietnamese is the official language of Vietnam. We choose the Hanoi dialect for VieTTS because it is mainly used for official activities such as education and broadcast. Vietnamese is a tonal language which involves six tones: Level (Ngang), Falling (Huyen), Broken(Ngã), Curve (Hoi), Rising (Sac), and Drop (Nang). Vietnamese has more tones than standard Chinese (four tones) and Thai (five tones). Vietnamese tone synthesis becomes a very interesting theme of research. Tones are usually considered as the time patterns of pitch and synthesized by using fundamental frequency (F_0) patterns. In Vietnamese, Broken and Drop tones are accompanied by a glottal stop [2,6,7], which is different from Chinese and Thai. For this feature, we propose the power control for these two tones so that they are synthesized by using not only F_0 patterns but also power pattern control. The synthesized tones were evaluated by a listening test. The result showed that the fundamental frequency is the most important factor for all six tones and the power control is significant for Broken and Drop tones in Vietnamese tone synthesis.

*e-mail: takara@ie.u-ryukyu.ac.jp

2. VIETNAMESE LANGUAGE'S OVERVIEW

The Vietnamese alphabet consists of 29 letters as shown in Fig. 1. In the phonetic alphabet, there are seven special characters with diacritic marks (dotted-line boxes in Fig. 1).

The six Vietnamese tones are shown in Table 1. Tone has a suprasegmental feature and affects the whole syllable [6]. Different tones allow words with the same structure of phonemes to present different meanings. In the Vietnamese writing system, a tone is represented by a diacritic mark. There are a total of six tones; but when a syllable ends with an unvoiced consonant, only Rising and Drop tones occur. Figure 2 shows an example of fundamental frequency (F_0) contours of the six Vietnamese tones with syllable "ma" uttered by a male speaker.

The Level, Broken, and Rising tones belong to the high-tone group, while the Falling, Curve, and Drop tones are in the low-tone group. Among these tones, Broken and Drop tones are accompanied by a glottal stop [6,7] or by a glottal constriction [2]. This feature will be examined in this paper at the analysis and synthesis part of Vietnamese tones.

a	ă	â	b	c
d	đ	e	ê	g
h	i	k	l	m
n	o	ô	o’	p
q	r	s	t	u
u’	v	x	y	

Fig. 1 The Vietnamese alphabet.

Table 1 The six Vietnamese tones.

Name	Tone mark	Example
Level	unmarked	ma - ghost
Falling	grave	mà - that
Broken	tilde	mã - horse
Curve	hook above	mả - tomb
Rising	acute	má - check
Drop	dot below	mạ - rise seedling

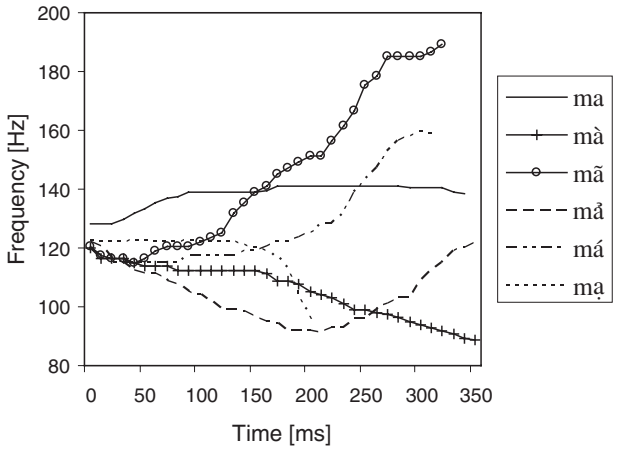

Fig. 2 An example of six tones with syllable “ma” uttered by a male.

Tone (T)			
Initial Consonant (C1)	Labialization (C2)	Vowel (V)	Final Consonant (C3)

Fig. 3 The Vietnamese syllable structure.

A Vietnamese syllable has the structure shown in Fig. 3 [8]. At position C1, there are 21 possible consonants. A labialization component makes the lip-rounding effect on the syllable and appears at position C2. A vowel is the nucleus of the Vietnamese syllable and appears at position V as one of 13 vowels or three diphthongs. There are eight phonemes at position C3: six consonants and two semi-vowels.

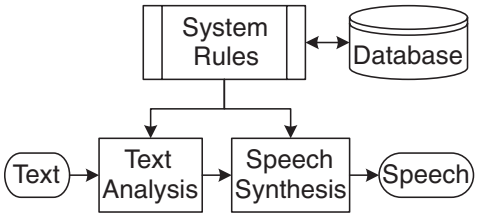

Fig. 4 The diagram of VieTTS system.

3. VIETTS SYSTEM

The design of VieTTS system is shown in Fig. 4. This design is based on the general speech synthesis system [9]. The input is Vietnamese text, and the output is synthetic speech. The text analysis sub-system converts Vietnamese text into a sequence of mapped characters, then this sequence is used to get information for synthesis. The speech synthesis sub-system generates speech from a pre-stored database under the control of synthesis rules. The database contains data for rules (character map for text analysis, tone patterns, interval length table) and demisyllable parameters with suitable formats. Each demisyllable’s parameters have a size of 2–6 KB, size of character map, tone patterns, intervals are around 10 KB. To make the system more generic, we use external definitions of interval marks, intervals, tone patterns, and a character table code.

3.1. Text Analysis

The purpose of text analysis is to extract phonetic and prosodic information. Because of the special characters and tone marks, we need to process Vietnamese text in order to get the sequence of speech unit codes (in roman characters). We choose roman characters as the target of mapped sequence because most Vietnamese letters are not changed in English display, and English is widely supported in computer operating systems (OS). With six tones and four special letter marks, we could build a mapping table using a number from zero to nine as in Table 2.

For example, the sentence “Tên tôi là Tú.,” which means “My name is Tu.,” will be mapped into a sequence of syllables as: “Te7n0 to7i0 la1 Tu4.”

Input is unrestricted Vietnamese text. At the first step, a syllable boundary is detected based on its interval mark which is defined in the external database of the Vietnamese character table code. With the rule for mapping, the sequence of mapped characters is obtained and then information is extracted. By iterating through text, we can determine interval mark, tone mark, demisyllables of each syllable and the total syllable number.

Table 2 Rule for mapping in TextAnalysis.

Marks	Number
Level tone	0
Falling tone	1
Broken tone	2
Curve tone	3
Rising tone	4
Drop tone	5
breve	6
circumflex	7
horn	8
d bar	9

3.2. Speech Analysis and Synthesis

The fundamental speech units of VieTTS are the demisyllables which are acquired by dividing a syllable into half with the cut point at the middle of the vowel. There are about 500 demisyllables in Vietnamese. As a speech database, Vietnamese demisyllables are collected and their sounds are prepared by recording on digital audio tape (DAT) at a 48 kHz sampling rate and 16-bit resoluation. After that, they are down-sampled to 10 kHz for analyzing. The format of the speech unit is pulse code modulation (PCM) without compression. All speech units are recorded with Level tone which is a kind of natural pitch level.

Cepstral analysis

VieTTS adopts short-time cepstral analysis. In the VieTTS system, the frame length is 25.6 ms and the frame interval, or frame shifting time, is 10 ms. A time-domain Hamming window with a length of 25.6 ms is used in analysis.

The cepstrum is defined as the inverse Fourier transform of the short-time logarithm amplitude spectrum [10]. Cepstral analysis has the advantage that it could separate the spectral envelope part and the excitation part. The resulting parameters of speech unit include the number of frames and, for each frame, voiced/unvoiced (V/UV) decision, pitch period and cepstral coefficients $c[m]$, $0 \le m \le 29$.

Speech synthesis

The speech synthesis sub-system, under the control of the synthesis rules, generates speech from pre-stored parameters. The source-filter model [4] is used as the speech production model.

Figure 5 shows the structure of the speech synthesis sub-system in VieTTS. For voiced sounds, excitations are a impulse train created by the pulse generator. These impulses have an interval equal to their pitch period. For unvoiced sounds, excitations are random noises that have a flat spectrum. The voiced/unvoiced sound decision controls the switching between two kinds of excitation generators. The voiced/unvoiced sound decision parameters are obtained by comparing the power at the low frequency part of spectrum with a threshold, which is set experimentally; if the power is greater than or equal to the threshold, then speech frame is voiced, else it is unvoiced. The amplitude is obtained via the first cepstral parameter $c[0]$.

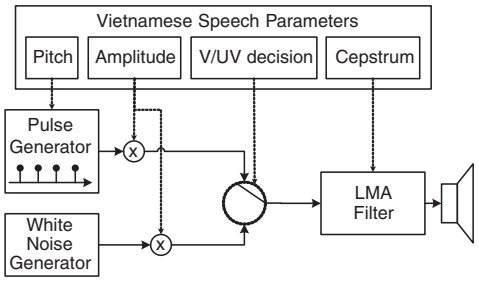

Fig. 5 VieTTS's speech synthesis sub-system.

The Log Magnitude Approximation (LMA) filter is introduced by Imai [5] to present the vocal tract characteristics. Vocal tract parameters are estimated in 30 lower-order quefrency elements. The LMA filter is a pole-zero filter that is able to efficiently represent the vocal tract features for all speech sounds. The LMA filter is constructed from a cascade connection of second order filters having impulse response $H_m(Z)$, $0 \le m \le 29$

$$H_m(Z) = \frac{\left(1 + \frac{c(m)}{2}Z^{-m} + \frac{c^2(m)}{12}Z^{-2m}\right)}{\left(1 - \frac{c(m)}{2}Z^{-m} + \frac{c^2(m)}{12}Z^{-2m}\right)}$$

where the $c(m)$ is the mth order cepstrum.

3.3. System Rules

Connection

A syllable is constructed from corresponding demisyllables and a tone. If the syllable contains two demisyllables, these two demisyllables are connected at the middle of the vowel position where acoustic features are nearly stable. Connection is implemented by interpolating cepstral coefficients.

Interval

The interval rule is defined externally in database. This makes VieTTS system more generic, or easy to modify to be suitable. Currently, VieTTS has four kinds of interval marks. The proposed interval rule for Vietnamese is shown in Table 3. There is a pause between consecutive words because we analyzed and found that the short silence periods are needed for the naturalness of Vietnamese continuous speech. It needs to be studied more to improve the naturalness.

Tones

Tone is strongly related to fundamental frequency (F_0). The six Vietnamese tones are analyzed to get F_0 patterns.

Table 3 Interval rule.

Interval mark	Symbol	Interval [ms]
Space	␣	70
Comma	,	700
Question mark	?	1400
Period	.	1400

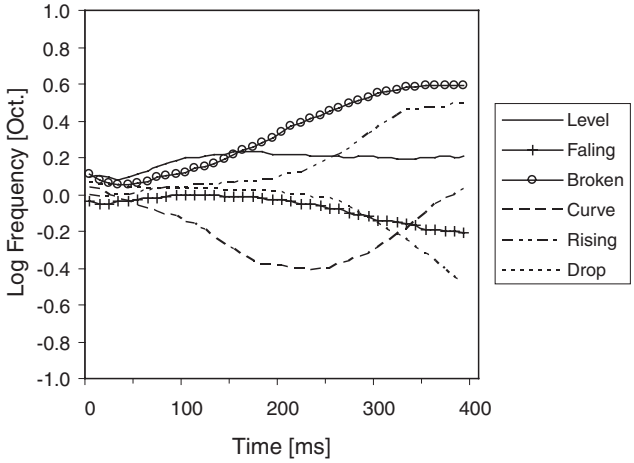

Fig. 6 The average F_0 contours of the six Vietnamese tones.

The set of words for analyzing tones is selected with the following conditions: (i) meaning words; (ii) all phonemes are voiced. We selected eight initial consonants: "b," "d," "g," "l," "m," "n," "ng"/"ngh," "v," and two vowels "a" and "i"/"y." Then we got 81 words for the analysis.

After analyzing, the F_0 contours of each tone are normalized to have same length, and same pitch level. During analysis, some errors on pitch detection occurred. We removed these problems by a program using interactive analysis-by-synthesis. In this program, users can smooth and/or fix error F_0 values and synthesize to get synthetic sound. An F_0 contour is decided when the synthetic sound is good enough; it means that the decision is manually processed by the researcher. Fig. 6 shows the average F_0 contours of the six Vietnamese tones. Zero level of log frequency here is equal to 128 Hz. As far as we know, the contours of six analyzed tones using multiple data of Vietnamese are not shown in any documentation before.

Among six tones, Broken and Drop tones have a glottal stop feature [2,6,7]. Glottal stop in a speech synthesis system has been studied by Takara [11]. By observing the recorded waveform of Vietnamese tones, we found the changes of power at the point in which such features occurred (Fig. 7). From this idea, we propose the power control in VieTTS system for Broken and Drop tones. The hypothesis is: a tone is composed of both an F_0 pattern and a power pattern control. Power control is implemented by changing the first cepstral coefficient $c[0]$. The $c[0]$

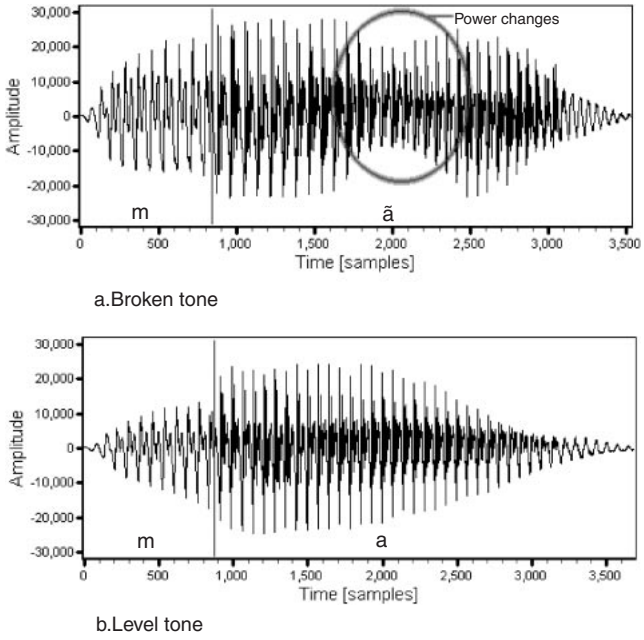

Fig. 7 An example of power changes in a word "ma" with Broken tone (a) compared with Level tone (b).

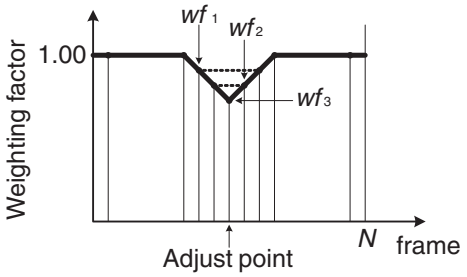

Fig. 8 Rule for power control for Broken tone.

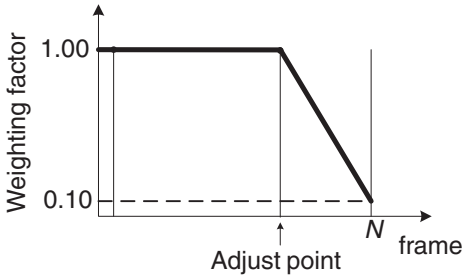

Fig. 9 Rule for power control for Drop tone.

parameters of frames are weighted to make the changes of the signal's power. Figures 8 and 9 illustrate two power patterns for the above two tones. The power control also effectively shortens the length of Drop tone words. A Drop tone word's length is usually shorter than the others (see Fig. 2). These two patterns are simple but they show very effective results when we evaluated the system using a listening test. In the Broken tone case, the adjust point wf_3

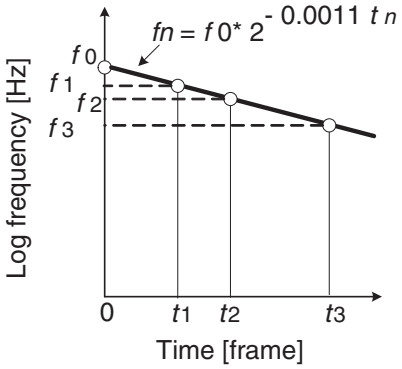

Fig. 10 Intonation rule.

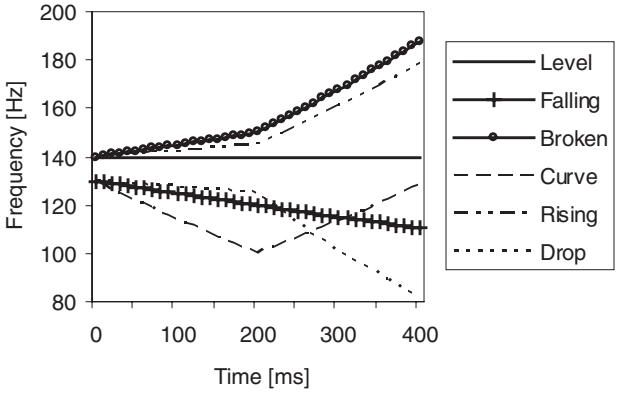

Fig. 11 Linear F_0 patterns for the listening test.

is at $3N/5$ for CV-form, at $2N/5$ for VC-form, and at the connection point for CVC-form; positions of applying wf_2, wf_1 are time point $wf_3 \pm 1$, $wf_3 \pm 2$ [frame], respectively; weighting factors wf_1, wf_2, wf_3 are experimental values, $0 < wf_i < 1$, which are chosen as 0.95, 0.90, 0.85, respectively. In the Drop tone case, the adjust point is at $3N/5$. Where N is the frame number in a syllable.

Power control is a new implementation in Vietnamese tone synthesis. This implementation is unique compared to other tonal languages such as Chinese and Thai since these languages never adopt power control in their tone synthesis; only pitch is examined [12,13].

Intonation

The intonation for a sentence is implemented by applying a simple declination line in the log frequency domain (Fig. 10) [14]. Time t_i is the initial point of a syllable, and the initial value of F_0 (circle mark) is calculated from this value. This is a simple linear line, which intends to experiment the very first step rule of intonation of Vietnamese. The sentence prosody should be studied more in the next research. We simply adopted it from the similar study [14].

4. EVALUATION AND DISCUSSION

4.1. Evaluation Test

Presently, the purpose of the evaluation is to test the tones' intelligibility for synthetic speech of Vietnamese syllable with generated tones and to determine the effect of power control in Vietnamese tone synthesis. The VieTTS system is evaluated through a listening test.

Six types of speech are prepared for the listening test:

- Type 0: Original sounds
- Type 1: Analysis - Synthesis sounds
- Type 2: Synthetic sounds: Average F_0 pattern (Fig. 6) with power control
- Type 3: Synthetic sounds: Linear F_0 pattern (Fig. 11) with power control
- Type 4: Synthetic sounds: Average F_0 pattern without power control
- Type 5: Synthetic sounds: Linear F_0 pattern without power control

All synthetic sounds use cepstra from speech units with Level tone. Average F_0 patterns are shown in Fig. 6. Power control here means the control of the $c[0]$ coefficient. Linear F0 patterns [6,7] are shown in Fig. 11, in which the vertical axis is frequency in Hz and the horizontal axis is time in frames.

The word set includes two vowels "a" and "i" with initial consonants "m," "b," "d," and final consonants "m," "n." Sixty data were selected, giving ten tokens for each tone. In this list, all sounds are utter-able, and most are meaning words. Since there are 60 sounds for each speech type, total sounds for each listening test was 360 data. The mixture of all six sound types together puts the test in a more natural situation.

In the test, each sound is played once at random, and the listener have to choose which word was heard within a two-second period. After this period, a warning is displayed to force the user to make a decision. There are five listeners, four males and one female. All listeners are from northern Vietnam and are in their twenties with a normal hearing ability.

The listening test intends to evaluate how correct the tones are, so that the test program only evaluates the confusion among tones, not among syllables. Another test for that problem might be processed separately. However, the listeners did not make any complain about the syllable confusion.

4.2. Result and Discussion

After collecting the results from the five listeners, we removed the results of the two listeners with the highest and lowest correct rate. We think that a precise result is obtained by this method. If there are listeners with extremely high or low recognition score, the averaged score is not reliable where a few listeners attend in the experiment. That can be the problem in a comparison

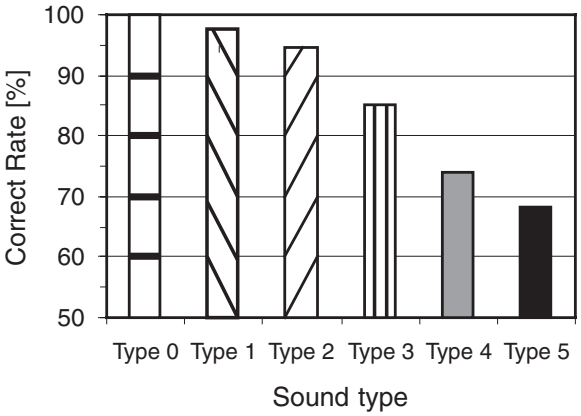

Fig. 12 Correct rate of tone synthesis.

between averaged scores of two methods. The overall result of the listening tests is shown in Fig. 12 and Table 4. Figure 12 describes how many percentages of the synthesized tones are recognized correctly, while Table 4 shows the confusion matrices from this evaluation.

Figure 12 shows the correct rate as follows:

- The proposed method (type 2 - average F_0 with $c[0]$ control) is acceptable with around 95% correct rate.
- The analysis-synthesis (type 1) sounds is 2% lower than that of the original sounds (type 0). This is thought to be caused by noises during analysis and/or synthesis procedures.
- Control of $c[0]$, or power control, is effective in Vietnamese tone synthesis. Comparing to average F_0 without $c[0]$ control (type 4), average F_0 with $c[0]$ control (type 2) are 21% higher. Linear F_0 with $c[0]$ control (type 3) is also 17% higher than that without $c[0]$ control (type 5).
- Linear F_0 with $c[0]$ control (type 3) is not so low in intelligibility. It is only 9% less than average F_0 with $c[0]$ control (type 2). The linear F_0 patterns are very simple patterns, easier for coding, and adopted in a basic Vietnamese learning books [6,7]. The values of those patterns are obtained by manually testing some synthetic sounds, a kind of experimental value. The average patterns are the results from many data, so that it should be better, which was proved by the evaluation.

From the confusion matrices, the error rates of the Broken tone recognized as Rising tone were 0%, 3%, 37%, 50% for type 2, type 3, type 4 and type 5, respectively. It shows that the power control makes Broken tone more clear. Similarly, we see that the error rates of Drop tone recognized as Falling tone were 0%, 0%, 63%, 70% for type 2, type 3, type 4 and type 5, respectively. These affirm the effectiveness of power control on Vietnamese tones.

Interestingly, by using speech units of Level tone and F_0 patterns with power pattern control, we can make

Table 4 Confusion matrices of tone synthesis. Unit: %.

		Lt	Ft	Bt	Ct	Rt	Dt
T	Lt	100	0	0	0	0	0
Y	Ft	0	100	0	0	0	0
P	Bt	0	0	100	0	0	0
E	Ct	0	0	0	100	0	0
0	Rt	0	0	0	0	100	0
	Dt	0	0	0	0	0	100
T	Lt	93	7	0	0	0	0
Y	Ft	0	100	0	0	0	0
P	Bt	0	0	100	0	0	0
E	Ct	0	0	0	100	0	0
1	Rt	0	0	7	0	93	0
	Dt	0	0	0	0	0	100
T	Lt	100	7	0	0	0	0
Y	Ft	23	70	0	0	0	7
P	Bt	0	0	100	0	0	0
E	Ct	0	0	0	100	0	0
2	Rt	0	0	3	0	97	0
	Dt	0	0	0	0	0	100
T	Lt	100	0	0	0	0	0
Y	Ft	40	60	0	0	0	0
P	Bt	0	0	97	0	3	0
E	Ct	0	0	0	100	0	0
3	Rt	27	0	20	0	53	0
	Dt	0	0	0	0	0	100
T	Lt	100	0	0	0	0	0
Y	Ft	20	73	0	0	0	7
P	Bt	3	0	60	0	37	0
E	Ct	0	0	0	100	0	0
4	Rt	0	0	20	0	80	0
	Dt	7	63	0	0	0	30
T	Lt	100	0	0	0	0	0
Y	Ft	10	83	0	0	0	7
P	Bt	13	0	37	0	50	0
E	Ct	0	0	0	97	3	0
5	Rt	17	0	20	0	63	0
	Dt	0	70	0	0	0	30

Lt: Level tone, Ft: Falling tone, Bt: Broken tone, Ct: Curve tone, Rt: Rising tone, Dt: Drop tone.

syllables with different tones. F_0 patterns affect all tones while power control is effective for Broken and Drop tones. Therefore, we may conclude that F_0 is most important factor and power control is significant (for Broken and Drop tones) in Vietnamese tone synthesis.

5. CONCLUSION

We have introduced a Text-To-Speech system for Vietnamese, VieTTS, which is a rule-based synthesis system using a cepstral method with demisyllable speech units. The VieTTS system could synthesize speech from Vietnamese text with six precisely generated tones. The Hanoi dialect is used as the standard Vietnamese in this system.

Tone synthesis of Vietnamese is implemented. Four

tones (Level, Falling, Curve and Rising) are synthesized by using fundamental frequency (F_0) patterns. For synthesizing the others (Broken and Drop) tones that have glottal features, we used both F_0 patterns and power pattern control. As a result, we found that F_0 is most important in Vietnamese tone synthesis, and power pattern control strongly affects Broken and Drop tones. Applying power control for tone synthesis may seem strange since tones are normally considered as the patterns of pitch variation, but it is effective in Vietnamese tone synthesis. This new result was confirmed by listening tests with reasonable correct rate.

Some areas remain as future work, such as: to improve of speech naturalness, to control sentence prosody using Fujisaki's model, and to treat the variation of tones in continuous speech.

REFERENCES

[1] T. T. Doan, *Vietnamese Phonetics* (Hanoi National University Publishing, Hanoi, 1999).

[2] M. Shimizu and M. Dantsuji, "A new proposal of laryngeal features for the tonal system of Vietnames," *Proc. ICSLP 2000*, Vol. 2, pp. 519–522 (2000).

[3] M. S. Han and K.-O. Kim, "Phonetic variation of Vietnamese tones in disyllabic utterances," *J. Phonet.*, **2**, 223–232 (1974).

[4] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd Ed. (Marcel Dekker, New York•Basel, 2001), pp. 30–31.

[5] S. Imai, "Log Magnitude Approximation (LMA) filter," *Trans. IECE Jpn.*, **J63-A**, 886–893 (1980).

[6] T. T. Doan, *Vietnamese Phonetics* (Hanoi National University Publishing, Hanoi, 1999), pp. 100–111.

[7] B. N. Ngo, *Elementary Vietnamese* (Tuttle Publishing, Boston•Tokyo•Singapore, 1999), p. 27.

[8] B. N. Ngo, *Elementary Vietnamese* (Tuttle Publishing, Boston•Tokyo•Singapore, 1999), p. 17.

[9] T. Takara and T. Kochi, "General speech synthesis system for Japanese Ryukyu dialect," *Proc. 7th WestPRAC*, pp. 173–176 (2000).

[10] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd Ed., (Marcel Dekker, 2001), pp. 62–66.

[11] T. Takara, "Experimental study on perception of the glottal explosive of the Japanese Ryukyu dialect," *Proc. EuroSpeech '95*, pp. 953–956 (1995).

[12] C.-H. Wu and J.-H. Chen, "Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis," *J. Speech Commun.*, **35**, 219–237 (2001).

[13] P. Seresangtakul and T. Takara, "Analysis of pitch contour of Thai tone using Fujisaki's model," *Proc. ICASSP '02*, Vol. 1, pp. 505–508 (2002).

[14] T. Takara and J. Oshiro, "Continuous speech synthesis by rule of Ryukyu dialect," *Trans. IEE Jpn.*, **108-C**, 773–780 (1988).

Tu Trong Do received a B.E. degree in Electronics and Telecommunication from Hanoi University of Technology, Vietnam, in 1999 and an M.E. degree in Information Engineering from University of the Ryukyus, Japan, in 2002. He joined Hanoi University of Technology, Vietnam from 1999 as a teaching and research assistant and from 2003 as a lecturer. From 2000 to 2002, he has studied as a graduate student in the Department of Information Engineering, University of the Ryukyus, Japan.

Tomio Takara received a B.S degree in Physics from Kagoshima University, Japan, in 1976 and an M.E. and a Dr. Eng. degree in Information Processing from Tokyo Institute of Technology, Japan, in 1979 and 1983, respectively. During 1991–1992, he studied at Carnegie Mellon University as a visiting scientist. He has been a Professor in the Department of Information Engineering and the director of the Computer and Networking Center, University of the Ryukyus, Japan, since 1995 and 2002, respectively. He is the recepient of the 1990 Okinawa Society Award for Encouragement of study on Okinawa. He is presently interested in spoken language processing and machine intelligence.