PRECISE TONE GENERATION FOR VIETNAMESE TEXT-TO-SPEECH SYSTEM

Tu Trong DO, Tomio TAKARA

Department of Information Engineering,
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0213 JAPAN

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 effective and unique for Vietnamese compared to other tonal languages such as Chinese and Thai.

1. INTRODUCTION

In spite of the development of speech technology, there are very few researches on Vietnamese speech processing [1, 2, 3], particularly on speech synthesis. In this paper, a Vietnamese Text-To-Speech (VieTTS) system is proposed. 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 huyền), Broken(ngã), Curve (hỏi), Rising (sắc), and Drop (nặng). Vietnamese has more tones than Chinese (four tones) and Thai (five tones). Tones are usually considered as the time patterns of pitch and synthesized by using fundamental frequency (F0) 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 F0 patterns but also power pattern control. The synthesized tones were evaluated by a listening test.

Table 1 The six Vietnamese tones

Name	Tone mark	Example
LEVEL (ngang)	unmarked	ma – ghost
FALLING (huyền)	grave	mà – that
BROKEN (ngã)	tilde	mã – horse
CURVE (hỏi)	hook above	mả – tomb
RISING (sắc)	acute	má – cheek
DROP (nặng)	dot below	mạ – rise seedling

2. VIETNAMESE LANGUAGE'S OVERVIEW

The Vietnamese alphabet consists of 29 letters. In the alphabet, there are seven special characters with diacritic marks.

The six Vietnamese tones are shown in **Table 1**. Tone has a suprasegmental feature and affects the whole syllable. Different tones make words with the same structure of phonemes contain 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.

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.

3. VIETTS SYSTEM

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, and 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 and demisyllable parameters with suitable formats. To make system more generic, we use external definitions of interval marks, intervals, tone patterns, and a character table code.

0-7803-7663-3/03/$17.00 ©2003 IEEE

3.1 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 resolution. After that, they are down-sampled to 10 kHz for analyzing. 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 part.

Cepstrum is defined as the inverse Fourier transform of the short time log-magnitude spectrum [10].

Speech synthesis

Fig. 1 shows the structure of speech synthesis sub-system in VieTTS. For voiced sounds, excitations are 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.

3.2 System rules

Connection

A syllable is constructed from corresponding demisyllables and a tone.

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.

Tones

Tone is strongly related to fundamental frequency (F0). The six Vietnamese tones are analyzed to get F0 patterns. 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.

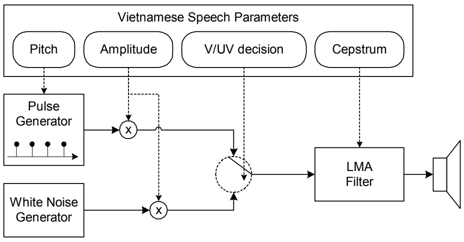

Fig. 1 VieTTS’s speech synthesis sub-system.

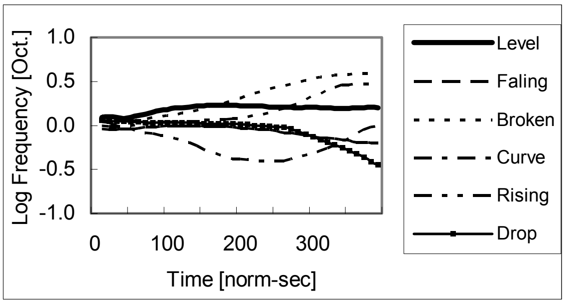

Fig. 2 The average F0 contours of the six Vietnamese tones.

After analyzing, the F0 contours of each tone are normalized to have same length, and same pitch level. **Fig.2** shows the average F0 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 glottal stop feature [2, 6, 7]. Glottal stop in a speech synthesis system has been studied by Takara [11]. By observing the waveform of Vietnamese tones, we found the changes of power at the point in which such feature occurred. From this idea, we propose the power control in VieTTS system for Broken and Drop tones. Then the hypothesis is: a tone will be applied by an F0 pattern and a power pattern control. Power control is implemented by changing the first cepstral coefficient $c[0]$. The $c[0]$ parameters of frames are weighted by some factors to make the changes of the signal's power. **Fig. 3** and **Fig. 4** explain 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. These two patterns are simple but they show very effective results when we evaluated the system by using the listening test.

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].

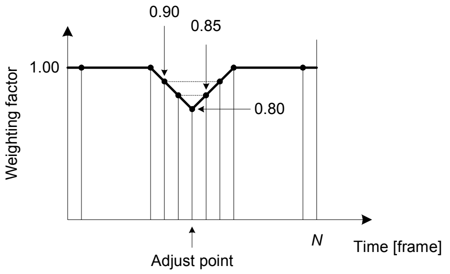

Fig. 3 Rule for power control in Broken tone.

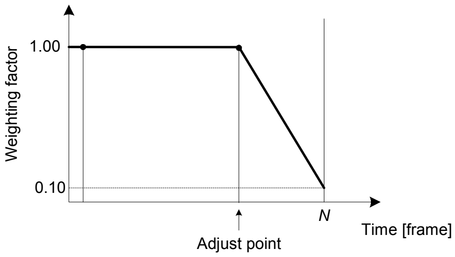

Fig. 4 Rule for power control in Drop tone.

Intonation

The intonation is implemented by applying a simple declination line in log frequency domain [14].

4. EVALUATION AND DISCUSSION

4.1 Evaluation test

Presently, the purpose of the evaluation is to test the tones' intelligibility of synthetic speech of Vietnamese syllable with generated tones and to assert the effect of power control in Vietnamese tone synthesis. 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: The average F0 pattern (Fig. 2) with power control
- Type 3: Synthetic sounds: The linear F0 pattern (Fig. 5) with power control
- Type 4: Synthetic sounds: The average F0 pattern without power control
- Type 5: Synthetic sounds: The linear F0 pattern without power control

All synthetic sounds use cepstra from speech units with Level tone. Power control here means the control of *c*[0] coefficient. Linear F0 patterns [6,7] are shown in Fig. 5, in which the horizontal axis is time in frame and the vertical axis is frequency in Hz.

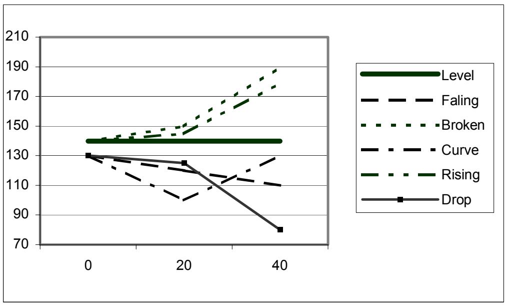

Fig. 5 Linear F0 pattern for the listening test

The word set includes two vowels “a” and “i” with initial consonants “m”, “b”, “d”, and final consonants “m”, “n”. Sixty data are chosen, in which there are ten data 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 is 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 and randomly, then the listener have to choose which word his/her had heard within a two-second period. After this period, a warning is displayed to force the user to make his/her 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.

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. The overall result of the listening tests is shown in **Fig. 6** and partly in **Table 2. Fig. 6** describes how many percentages of the synthesized tones are recognized correctly, while **Table 2** consists of the confusion matrices in this evaluation.

From **Fig. 6**, we see the correct rate as follows:

- The proposed method (type 2 - average F0 with *c*[0] control) is acceptable with around 95% correct rate.
- The analysis-synthesis (type1) sounds are 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. The average F0 with *c*[0] control (type 2) is 21% higher than the average F0 without *c*[0] control (type 4). The linear F0 with *c*[0] control (type 3) is also 17% higher than that without *c*[0] control (type 5).

- Linear F0 with *c*[0] control (type 3) is not so low intelligibility. It is only 9% less than average F0 with *c*[0] control (type 2).

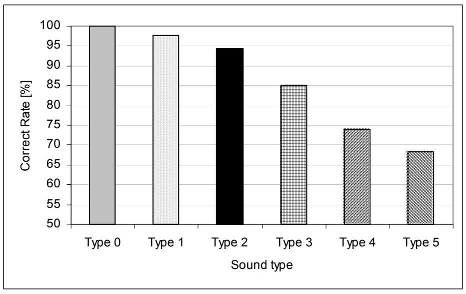

Fig. 6 Correct rate of tone synthesis.

Table 2. Confusion matrices of tone synthesis. Unit:%. Lt:Level tone, Ft:Falling tone, Bt:Broken tone, Ct:Curve:Tone, Rt:Rising tone, Dt:Drop tone.

Type		Lt	Ft	Bt	Ct	Rt	Dt
TYPE 2	Lt	100	0	0	0	0	0
	Ft	23	70	0	0	0	7
	Bt	0	0	100	0	0	0
	Ct	0	0	0	100	0	0
	Rt	0	0	3	0	97	0
	Dt	0	0	0	0	0	100
TYPE 4	Lt	100	0	0	0	0	0
	Ft	20	73	0	0	0	7
	Bt	3	0	60	0	37	0
	Ct	0	0	0	100	0	0
	Rt	0	0	20	0	80	0
	Dt	7	63	0	0	0	30

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 us that the power control makes Broken tone dramatically clear. Similarly, we can explain in the case 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 dramatic effectiveness of power control on Vietnamese tones.

. 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 speech units are demisyllables. VieTTS system could synthesize speech from Vietnamese text with precisely generated six tones. We have introduced a Text-To-Speech system for the Hanoi dialect, which 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 (F0) patterns. For synthesizing the others (Broken and Drop) tones that have glottal features, we used both F0 patterns and power pattern control. As a result, we found that F0 is the most important in Vietnamese tone synthesis, and power pattern control strongly affects Broken and Drop tones.

Acknowledgement

This study is supported by Grant-in-Aid for Scientific Research (C) of Japan Society for the Promotion of Science No. 12680419.

REFERENCES

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

[2] M. Shimizu and M. Dantsuji, "A New Proposal of Laryngeal Features for the Tonal System of Vietnamese", *Proc. of ICSLP*, 2, pp. 519-522, 2000.

[3] M. S. Han and K.-O. Kim, "Phonetic variation of Vietnamese tones in disyllabic utterances", *Jour. of Phonetics*, **2**, pp.223-232 ,1974.

[4] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", Second Edition Marcel Dekker, Inc., pp. 30-31, 2001.

[5] S. Imai, "Log Magnitude Approximation (LMA) filter", *Trans. of IECE Japan*, J63-A, 12, pp. 886—893, 1980. (in Japanese)

[6]T. T. Doan, "Vietnamese Phonetics", Hanoi National University Publishing, pp. 100-111, 1999. (in Vietnamese)

[7] B. N. Ngo, "Elementary Vietnamese", Tuttle Publishing, p. 27, 1999.

[8] B. N. Ngo, "Elementary Vietnamese", Tuttle Publishing, p. 17, 1999.

[9] T. Takara and T. Kochi, "General Speech Synthesis System for Japanese Ryukyu Dialect", *Proc. of the 7th WestPRAC*, pp. 173-176, 2000.

[10] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", Second Edition, Marcel Dekker, Inc., pp.62-66, 2001.

[11] T. Takara, "Experimental Study on Perception of the Glottal Explosive of the Japanese Ryukyu Dialect", *Proc. of 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", Jour. of Speech Communication, **35**, pp. 219-237, 2001.

[13] P. Seresangtakul and T. Takara, "Analysis of Pitch Contour of Thai Tone Using Fujisaki's Model", *Proc. of ICASSP'02*, **1**, pp. 505-508, 2002.

[14] T. Takara and J. Oshiro, "Continuous Speech Synthesis by Rule of Ryukyu Dialect", *Trans. IEE of Japan*, **108-C**, **10**, pp. 773-780, 1988. (in Japanese)