Non-uniform unit selection in Vietnamese Speech Synthesis

Thao Van Do
School of Information and Communication Technology – Hanoi University of Technology and Science
1 Dai Co Viet, Hanoi, Vietnam
+84 (0)4 38.68.25.95
thaodv.bkit@gmail.com

Do-Dat Tran
International Reseach Center MICA CNRS UMI 2954 – Hanoi University of Technology and Science
1 Dai Co Viet, Hanoi, Vietnam
+84 (0)4 38.68.30.87
Do-Dat.Tran@mica.edu.vn

Thu-Trang Thi Nguyen
School of Information and Communication Technology – Hanoi University of Technology and Science
1 Dai Co Viet, Hanoi, Vietnam
+84 (0)4 38.68.25.95
trangntt@soict.hut.edu.vn

ABSTRACT

In concatenative-based speech synthesis systems, speech is generated by concatenating acoustic units together, so, selection of units for concatenating affect directly to quality of synthetic speech. In our previous Text To Speech (TTS) system [9] , speech is synthesized by concatenating acoustic units together. These units are only one type, such as diphone, half syllable. In recent, the speed of CPU and the capacity of memory is significantly improved, we can increase the database size and perform more complex search. Non-uniform unit selection method is researched and developed. Many types of unit are used. The idea is *the longer units are, the higher quality is*. This method was applied in different ways in different languages. This paper describes the way of applying this method for Vietnamese TTS to improve the quality of speech synthesis system. This paper also present the results of perception test of uniform versus non-uniform unit selection method.

Categories and Subject Descriptors
Knowledge-based and information systems

General Terms
Algorithms, Languages.

Keywords
Non-uniform, Unit Selection, Low-level Synthesis, Text To Speech.

1. INTRODUCTION

Speech synthesis has been researched and developed from many years ago. In Vietnam, there are several TTS systems such as "*Sao Mai*" of Sao Mai Center, "*Hoa Súng"* of Mica Research Center [9] , "*Tiếng nói phương Nam"* of University Of Science Ho Chi Minh City [10] .

We can classify TTS system by the method of synthesizing speech into three type:

- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis.

Articulatory synthesis promises to deliver the best result but due to its complexity, this method is the most difficult to implement. Formant synthesis can generate speech with unlimited number of sentence but the quality of speech is not natural. The most commonly used method is concatenative synthesis.

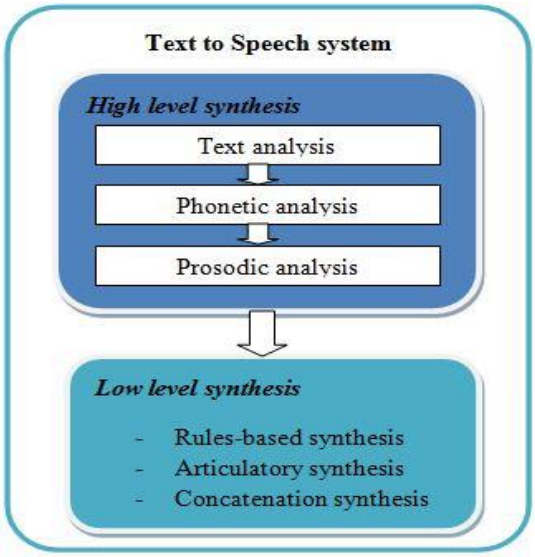


Figure 1.1 TTS system model

A concatenative-based TTS system consists of two main part: High level synthesis and low level synthesis Figure 1.1. Tasks of high level part are text analysis, phonetic analysis and intonation generation. Low level part (in concatenation synthesis) search and selection units base on parameters of high level part, carry out the concatenation of speech segments and the manipulation of acoustic parameters.

In order to improve quality of synthesized speech, we focused on low level part, especially unit selection. We increased the length of acoustic unit and the size of database to make the speech more natural. In [10] , the authors used a very large database to cover a large number of syllables in Vietnamese but we did another approach. Our database is not very large (about 80M) and we want to optimize the use of this database. Our proposed method will be described in section 3. We also implemented our method

and evaluate its results. Firstly, we will describe the backgrounds of previous approaches.

2. BACKGROUNDS

2.1 Concatenative Speech Synthesis

In concatenative synthesis, an acoustic unit is represented by a segment of waveform with its phonetic matching. A statement is synthesized by concatenating a set of several segments together. Each segment is natural, so we can expect a high quality output. Unfortunately, the segments are greatly affected by coarticulation [9] , and if we concatenate two speech segments which were not adjacent, it can result in spectral or prosodic discontinuities. The spectral discontinuities occur when the spectral frequency characteristics at the concatenation point are not matched. The prosodic discontinuities occur when the fundamental frequency at the concatenation point is not continuous.

According to [9] , there are four issues which we need to address to get good quality synthesized speech:

1. Select types of units. Possible units are phoneme, diphones, half syllable, initial/final part, syllable, phrases.
2. How to design the acoustic inventory from a set of recordings?
3. How to select the best sequence of speech segments for concatenating?
4. How to alter the prosody of a speech segment to best match the desired output prosody.

We focused on the first and the third problem to improve quality of the speech. To solve the fourth problem, TD-PSOLA algorithm was applied and got good results. But the discontinuities still exist. To select the best units, the discontinuities were represented by a cost function. We will describe the use of this cost function in 2.3.

2.2 Types of acoustic units

In Vietnamese, there are several types of units which can be used for synthesizing, such as phoneme, diphone, half syllable, initial/final part, syllable, and phrase. Table 2.1 show number of each of them [9] .

Table 2.1 Acoustic units in Vietnamese

Acoustic unit	Number	
	Without tone	With tone
Phoneme	40	130
Diphone	620	2976
Half syllable	590	2809
Initial/final part	22/161	22/661
Syllable	2466	7088

In these types of units, half syllable was used for Vietnamese TTS [9] [1] because its result was better than others and its database size was acceptable (below 10M); TTS program could run on PC, DSP [1] and could synthesize a large number of syllable. Recent years, with purpose of improving quality of synthesized speech, non-uniform unit selection method is researched and developed. Syllable or even phrase was used as synthesized unit. In applying this method for Vietnamese TTS, we proposed using three types of unit including half syllable, syllable and phrase.

Table 2.2 Types of using units

Types of units	Length	Number of concatenation point	Found probability
Phrase	Long ⇑	Less ⇓	Low ⇓
Syllable			
Half syllable	Short	Many	High

Table 2.2 shows the advantages and disadvantages of each type of units. Non-uniform unit selection will optimize the advantages of them: reduce number of concatenation point by using phrase and syllable, ensure the ability of synthesizing almost syllable in Vietnamese by using half syllable. But the complexity of this approach is that the problem in the use of three types of units requires a flexible process to exchange between the types of units.

2.3 Uniform unit selection

Suppose an input sentence is analyzed in a series of n acoustic units. The target for this sentence is a series of n units (t_i, i = 0…n-1) which contain the necessary prosodic information. From this target, we must select a series of n acoustic units (u_i, i = 0…n-1) in the database, allowing the system to produce a synthesized sentence with the highest possible quality.

Two cost functions were used [9] :

- The target cost $C^t(u_i,t_i)$ is the different between a unit in database u_i and the target t_i.
- The connection cost $C^c(u_{i-1},u_i)$ is an estimate of the difference between two consecutive units (u_{i-1},u_i) at the concatenation point.

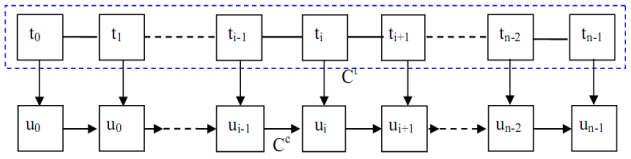
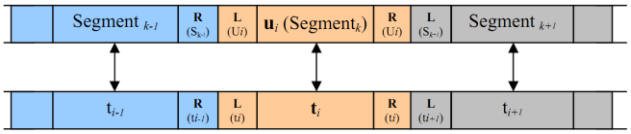


Figure 2.1 Cost function

From the specification of the target and the sequence of n units $T=(t_1, t_2,\ldots,t_n)$, the system needs to choose a set of units $U=(u_1, u_2,\ldots,u_n)$ that are closest to the target.

The difference between the target t_i and u_i was estimated by calculating the target cost which consists of following sub-costs:

- The difference in context between the candidate and the target. In Figure 2.2, the difference is calculated by comparing the information of segments (k-1) and (k+1) against t_{i-1} and t_{i+1}.
- The prosodic difference between the candidate and target: duration, funcamental frequency, energy.



– Figure 2.2 The target cost

The connection cost $C^c(u_{i-1},u_i)$ is determined by summing the following sub-costs :

– The difference between the right segment of u_{i-1} and u_i : $d(segment_{m+1}, u_i)$.

– The difference between the left of u_i and u_{i-1}: $d(u_{i-1}, segment_{k-1})$.

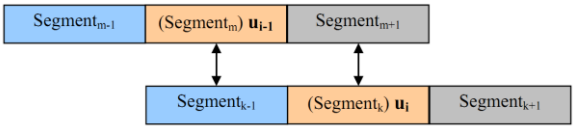


Figure 2.3 Comparison of the difference in context.

The total cost for a sequence of n units is the sum of target costs and connection costs:

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n}\sum_{j=1}^{p}w_j^t C_j^t(t_i,u_i)+\sum_{i=2}^{n}\sum_{j=1}^{q}w_j^c C_j^c(u_{i-1},u_i)+C^c(S,u_1)+C^c(u_n,S)$$

Where S denoted the silence, $C^c(S,u_1)$ and $C^c(u_n,S)$ define the conditions for the start and end data by concatenating the first and last units to the silence.

The process of selecting unit has to meet the total cost is smallest:

$$\overline{u_1^n}=\min_{u_1,\ldots,u_n} C(t_1^n,u_1^n)$$

3. NON-UNIFORM UNIT SELECTION IN VIETNAMESE SPEECH SYNTHESIS

3.1 Model of non-uniform unit selection

With purpose of reducing concatenation point, the priority of units for selecting in desending order of phrase, syllable and half syllable. Figure 3.1 shows general model of non-uniform unit selection process. Based on the selection method which was described in [9] , firstly, selection process will have one more step, that is "*Parsing sentence into phrases"*. Secondly, units will be searched in text database or half syllable database. Finally, the best units which have minimum cost will be selected for concatenating.

3.2 Separating sentences into words/phrases

In this step, the problem is how to parse a sentence into phrases and maximizing the probability of finding these phrases. If we can't select appropriate phrases, major selected units will be syllable and half syllable, efficiency of this method will be decreased. For example, there are two ways of parsing sentence "*Xin cám ơn mọi người*" into phrases:

– *Xin cám | ơn mọi | người.*
– *Xin | cám ơn | mọi người.*

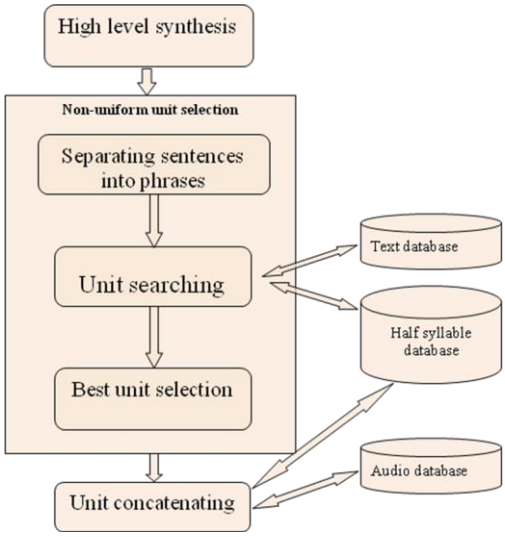


Figure 3.1 Non-uniform unit selection model

Obviously, phrases in the second line have higher probability of occurring in database than the first line. We proposed a solution to solve this problem – using parse tree of the sentence. Parse tree is the result of previous module in our TTS system. After parsing syntax phase, the sentences are divided into phrases into different levels. Example for sentence "*xin cảm ơn mọi người*" Figure 3.1:

3.3 Unit searching

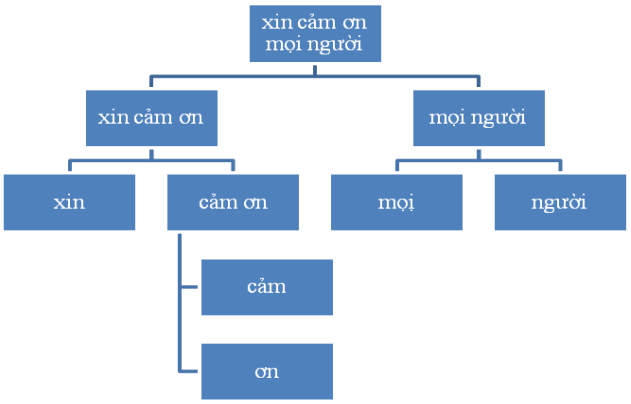


Figure 3.1 Parse tree to search.

Next step is unit searching. Searching process begins at the highest level of the parse tree (root) then goes downward to sub-nodes. If phrases at the higher level don't occur in text database, phrases in sub-nodes will be substituted for searching. Otherwise, necessary information such as found indexes, phonetic, context of units is returned to calculate cost function. If syllable at leaf is not found, it will be divided into two half syllables. They are searched in half syllable database. If not occur, that syllable will not be synthesized, but this case is very rare. Details of unit searching process are described in Figure 3.2.

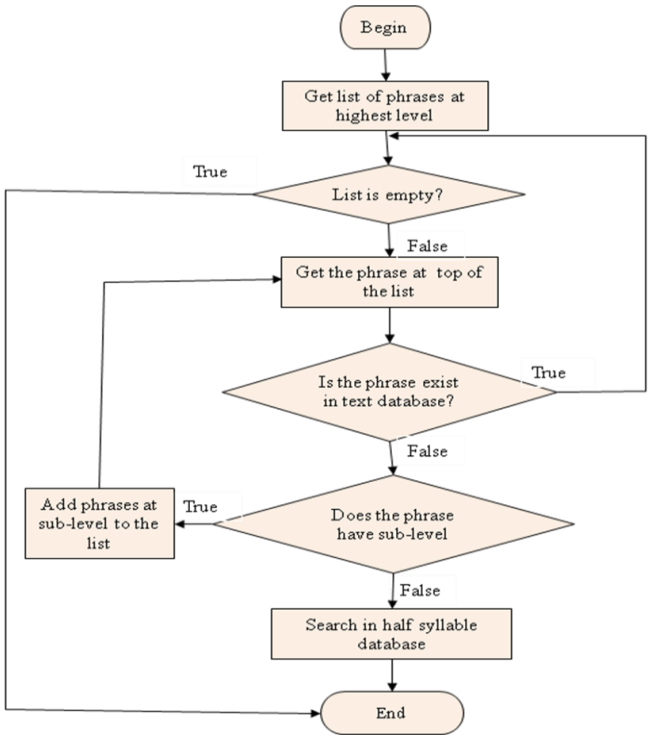


Figure 3.2 Unit searching process

3.4 Optimizing acoustic unit instances

After two previous steps, each target unit usually has several candidate units. The objective of this step is selecting candidate best match with its target unit and the context. In uniform unit selection method, this step could be carried out with the method mentioned in [9] . But in non-uniform method, there is combination of three types of units, we need another selection method. We proposed a solution that is "*optimizing local cost function*". Below are details:

Step 1: Split the sequence of candidates to subsequences in such a way that types of units in a sequence are same, only half syllable or syllable and phrase.

Step 2: Calculate cost function for subsequences and keep K-best candidates correspond to each target unit

- *Subsequences of half syllable:*
  - *Calculate target cost and connection cost for each unit*
  - *Keep K-best half syllable candidates base on target cost.*
- *Subsequences of syllable and phrase:*
  - *Calculate target cost like half syllable*
  - *Calculate connection cost. There are some changes in the way of calculating. With two consecutive candidate units, we compare phonetic and context of two pairs:*
    - *Last syllable of left unit versus left-neighbor syllable of right unit, parameters for comparision including : syllable name, final phoneme, final type and syllable tone*
    - *Right-neighbor syllable of left unit versus first syllable of right unit, parameters for comparision including: syllable name, initial phoneme, initial type and syllable tone.*
  - *Keep K-best candidates base on target cost.*

Step 3: Choose best sequence of units which have minimum total cost using Viterbi algorithm (dynamic programming) [6] . Total cost is sum of target cost and connection cost. Each cost has a weight to modify level of its affect to total cost. This weight is determined during experimental time.

4. IMPLEMENTATION

4.1 General model

Our program is divided into two main modules: high level and low level module. High level module is written by Java language with main function is searching and selecting best units for synthesizing. Low level module is written by C++ language with main function is audio signal processing and concatenating units. Two modules are linked together by Java Native Interface (JNI). We used C++ because we want to integrate Hoa Súng TTS program into our program. Hoa Súng TTS program was written by C++, could synthesize speech using half syllable unit and implemented unit concatenation algorithm – TD-PSOLA. We focus on unit searching and selection by using phrase and syllable, but not on half syllable and unit concatenation.

Input: the text need to be synthesized include sentences which are parsed.

Output: sound file synthesized corresponding to input text.

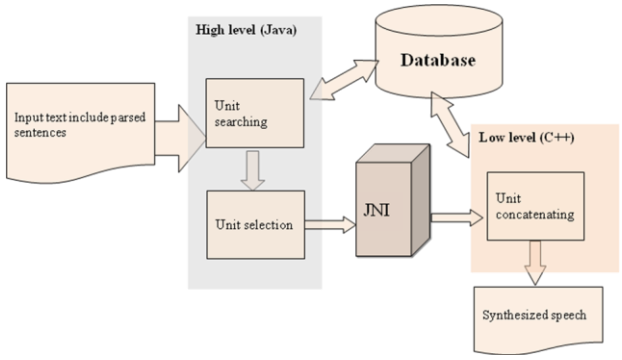


Figure 4.1 General system model

4.2 Database design

There are three sets of corpus which are used, include audio database, meta-data database and half syllable database.

4.2.1 Acoustic database

This database consists of audio files which stored in wav file and have same format. These files were recorded in a constant condition by only one speaker (female). Each file corresponds to one paragraph or one dialogue in text database. The database size is 68M, length is about 37 minutes. This is not a large database for a TTS system.

4.2.2 Meta-data database

This database consists of 250 paragraphs and dialogues of 630 sentences. There are 10852 instances of 1600 distinguish syllables. Necessary information of the syllable such as phonetic

components, tone, duration, energy, context … is prepared and stored. We used an XML file for storing text database. Structure of XML file is described in Figure 4.1:

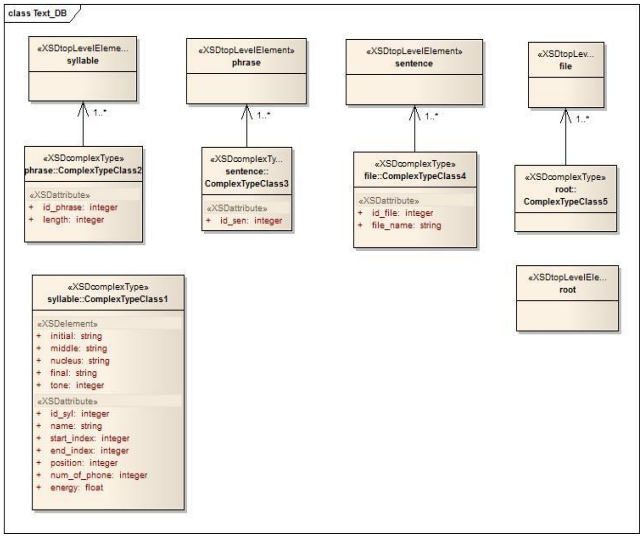


Figure 4.1 Structure of XML file.

4.2.3 Half syllable database

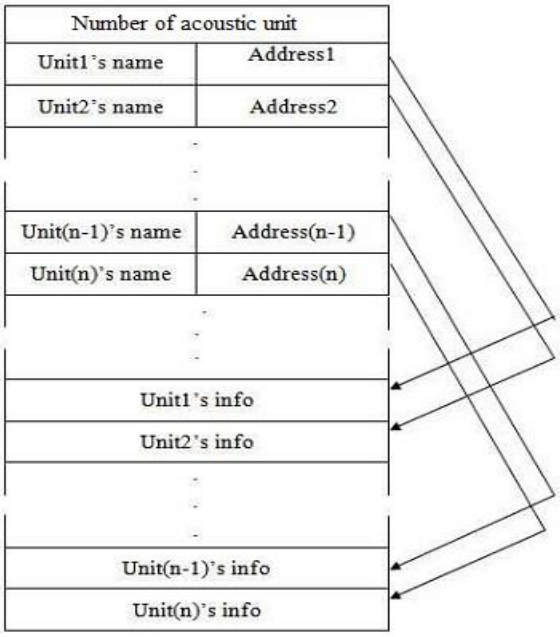


Figure 4.2 Structure of half syllable database

This database was used in Hoa Súng TTS system and its structure consists of following parts Figure 4.2:

- Two first bytes contain number of all acoustic unit in database
- Next is header part, size is 50000 bytes. The header is divided into blocks; each block has 8 bytes size. Four first bytes hold acoustic unit's name, four remaining bytes hold address of that acoustic unit.
- Final is data part. This part contains data of all acoustic units. Each unit includes elements described in Table 4-1:

Table 4-1 Elements of a unit

Element	Size (byte)	Description
bDeleted	1	Is this unit exist in database
nTranPoint	2	Transition point between voiced and unvoiced phoneme
dwUnitLen	4	Length of unit
unitType	1	Type of unit
bTone	1	Tone of unit
bleftTone	1	Tone of left-neighbor unit
brightTone	1	Tone of right-neighbor unit
dwLowFEnergy	4	Energy of low frequency part of unit
dwHighEnergy	4	Energy of high frequency part of unit
leftUnitName	4	Name of left-neighbor unit
rightUnitName	4	Name of right-neighbor unit
Reserved	1	Reserved byte
Signal data		Data of acoustic unit, data size depends on each unit
Pitchmark i	4	M pitch mark values
MFCCi	4	12 MFC coefficients of acoustic unit

5. EXPERIMENTAL RESULTS & EVALUATION

We developed a program that could synthesize speech using phrase and syllable. To assess quality of generated speech, we prepared data to do perception test.

5.1 Experimental corpus

In this test, a text corpus of 7 sentences or paragraphs of Vietnamese was used. This corpus is random chosen from web. They were put into two TTS system for comparison and evaluation. The first is *Hoa Súng* TTS system; another is the system which we developed. So, the speech corpus contained 2 groups of 7 paragraphs.

The experiment was carried out in studio room of MICA center. There was 8 persons took part in this test. After listening, the listeners were asked to rate the quality of speech according to 2 criterions. Two samples of the same paragraph would be placed adjacent but the order of them is random to make an objective assessment.

5.2 Perception test

To assess quality of generated speech, we based on two criterions: *clearness in pronunciation* and *naturalness of speech*. Each criterion will be assessed in range from 1 to 5 as Table 5.1:

Table 5.1 Criterions for assessing

Criterion	Level and description

Clearness in pronunciation	1. *Indistinguishable* 2. *Not clear* 3. *Proximate clear* 4. *Enough clear* 5. *Very clear*
Naturalness of speech	1. *Not natural absolutely* 2. *Not natural* 3. *Proximate natural* 4. *Enough natural* 5. *Very natural*

The task of assessing person is to hear the speech and score them according to above criterions.

5.3 Results & Evaluation

After handling data, we got Figure 5.1 and Figure 5.2. According to statistical results, the average score of system 2 is better than system 1. Almost score of paragraphs synthesized in system 2 was higher than system 1. The reason is units in system 2 are phrase and syllable while units in system 1 are half syllable. System 1 was a big system with nearly completed modules of a TTS system, but system 2 included only unit searching and selection module and a simple unit concatenation module. Therefore, score of naturalness of paragraph 5 of system 2 was lower than system 1. For system 2, the score of clearness was 4.0, this was considered a high level in the scale of 5; while the naturalness's score was lower - 3.64. To improve naturalness of system 2, we need to change the prosody of generated speech, calculate spectral distance at concatenation point between units.

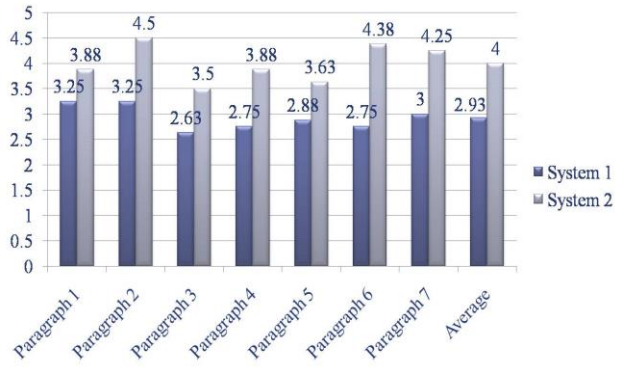


Figure 5.1 Result of clearness

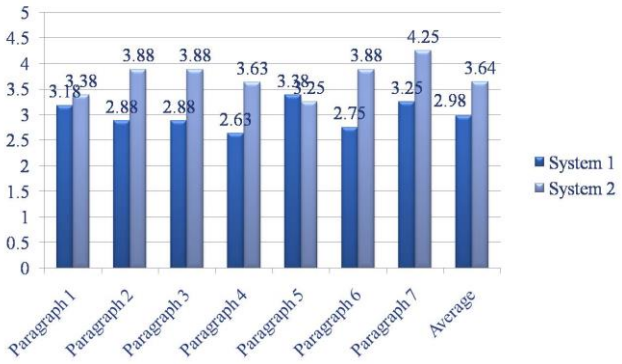


Figure 5.2 Result of naturalness

The text corpus included only statements, so the prosody of synthesized speech in system 2 was similar natural. But if input is question, exclamatory or imperative, the prosody of output will not good because of insufficient of our program.

6. CONCLUSIONS & FUTURE WORKS

This paper described the way we applied non-uniform unit selection method for Vietnamese TTS. We also organized text and audio database for searching and expanding easily. A TTS program was written by Java and C++ to evaluate of our proposed method. Although the initial result was relatively good but our program was only a small part of a TTS system.

We are trying to write a complete TTS program, make the program have ability in combining three types of units - including phrase, syllable and half syllable – as we proposed. Intonation models will be applied to improve quality of synthesized speech.

7. ACKNOWLEDGMENTS

We would like to thank Research Center MICA for helping us with mechanisms and rooms for the research. We also want to thank MICA staffs and our friends who willingly participated in our tests and experiment.

8. REFERENCES

[1] Lại Hoàng Nam, Quách Đại Quang, "*Xây dựng chương trình tổng hợp tiếng nói trên DSP*", đồ án tốt nghiệp K49, ĐH Bách Khoa Hà Nội, 2009.

[2] Lukas Latacz, Yuk On Kong, Werner Verhelst, "*Unit Selection Synthesis Using Long Non-Uniform Units and Phonemic Identity Matching*", Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, 2007.

[3] Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, Stefano Sandri, "*Choose the best to modigy the least: a new generation concatenative synthesis system*", CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.A., Torino, Italy.

[4] Mark Tatham, Katherine Morton, "*Development in Speech Synthesis*", Wiley, 2005.

[5] Min Chu, Hu Peng, Hong-yun Yang, Eric Chang, "*Selecting non-uniform units from a very large corpus for concatenative speech synthesizer*", Microsoft Research China, Beijing.

[6] Minghui Dong, Kim-Teng Lua, Haizhou Li, "*A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese*", Institute for Infocomm Research.

[7] Paul Taylor, "*Text-to-Speech Synthesis",* University of Cambridge, Cambridge University Press, 2006.

[8] Tian-Swee Tan and Sh-Hussain, "*Implementation of Phonetic Context Variable Length Unit Selection Module for Malay Text to Speech",* Faculty of Biomedical Engineering and Health Science, University Teknologi Malaysia, Malaysia, 2008.

[9] Trần Đỗ Đạt, "*Synthèse de la parole a partir du texte en langue Vietnamienne*", Ph.D. Thesis, Thèse en cotutelle international MICA, Hanoi, 2007.

[10] Vũ Hải Quân, Cao Xuân Nam, "*Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ*". Tập V-1, Số 1, tháng 04/2009

[11] Xuedong Huang, Alejandro Acero, Hsiao-Wuen Hon, "*Spoken language processing*", Prentice Hall, 2001.