ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

Vietnamese Speech Processing and Synthesis in VNSExpenses System

Quoc The Van[1], Nguyen B. P. Nguyen[2], Anh K. V. Nguyen[3], Hien Thanh Vu[4], Thien Khai Tran[5]

Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology, Ho Chi Minh City, Vietnam [1, 2, 3, 4, 5]

Abstract: This paper presents the Vietnamese speech processing task in VNSExpenses system, a tool able to understand users' voice commands, would help users with managing and querying their personal expenses by Vietnamese speech. In this system, we use HTK toolkit for speech recognition and Unit Selection method for synthesis operations. Having been built and tested in PC environment, our system proves its accuracy attaining more than 95%.

Keywords: Speech Recognition, Speech Synthesis, Vietnamese, VNSExpenses

I. INTRODUCTION

In the field of applying Vietnamese speech processing techniques to build speech-based human- computer interaction systems, at now, we know some newest remarkable publications of some research groups in Vietnam such as Institute of Information Technology (Vietnamese Academy of Science and Technology) and University of Science (VNU-HCM), such as Thang Vu and Mai Luong [1] as well as Quan Vu et al. [2,3,4,5] 's one which obtained the precision rate of over than 90% and this group successfully built many voice applications on this base. However, it does have some limitations in Vietnamese speech recognition: There is no high precision recognition system with a very large vocabulary; There is no publicly available corpus. The major approaches to speech recognition for Vietnamese that is based on statistical pattern recognition. In this research, we deal with a Vietnamese speech recognition task by using HTK (Hidden Markov Model Toolkit) [6] and speech synthesis operations by using Unit Selection method.

II. SYSTEM ARCHITECTURE

Our system is designed to carry out these functions as bellow:

1. Add: add the expenditure into the VNSExpenses by Vietnamese speech.
2. Delete: remove the expenditure out of the VNSExpenses by Vietnamese speech.
3. Edit expenses: edit the expenditure from the VNSExpenses by Vietnamese speech.
4. Query expenses: query the expenditure from the VNSExpenses by Vietnamese speech.

VNSExpenses meets these above functions in observing the further scenario:

The interaction between users and system can be presented in brief as following steps:

Step 0 Listening stage

Step 1 User says to the system by Vietnamese.

Step 2 The speech sentence is converted into the Vietnamese text sentence thanks to the Speech Recognizer.

Step 3 The system analyses the syntax structure and gets the key information of the text sentence.

(3.1) If the input sentence is a command:

If it is an add command:

- The system adds the associated expenditure to the database and confirms the result to user by voice.
- Return Step 0.

If it is a delete command:

- The system deletes the associated expenditure from the database and confirms the result to user by voice.
- Return Step 0. If it is an edit command:
- The system deletes the expenditure needing to edit and adds the associated expenditure to the database and then confirms the result to user by voice.
- Return Step 0.

(3.2) If the input sentence is a query

- The system executes the query, searches information in the database and shows the result to user by voice.
- Return Step 0.

(3.3) In case the syntax is incorrect, the system will inform user of it and user can take another

To realize the functions in observing the above scenario, the system must be composed of following components:

1. Automatic speech recognizer (ASR): identify words that user speaks, then convert them into written text.
2. Vietnamese language processor: resolve the syntax and semantic representations of all the command sentences or query sentences of user.
3. Central processor:

- Transform the semantic representations of the command / query sentences into the SQL commands and execute it.
- Filter, organize, and return the results to user.

4. Database: store schedule information
5. Synthesizer: convert text to speech

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

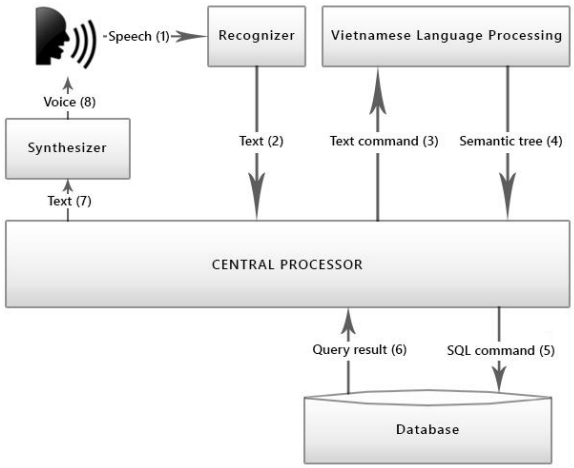

Fig. 1. Architecture of VNSExpenses

III. AUTOMATIC SPEECH RECOGNIZER

In VNSExpenses system, we have used HTK (Hidden Markov Model Toolkit) [12] to build the Automatic Speech Recognition component.

Hidden Markov Model (HMM) is a statistical model in which the system being modelled assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from an observation parameters. In speech recognition process, after our voice is recorded, it will be divided into many frames that we need to process in order to generate the sentence in text form. Each frame is represented as state, group of some states is represented as phoneme, and group of some phonemes is represented as word that we need to recognize. In database known as linguist model, we store the reference value of state, phoneme, and word in order to compare with the observed data (voice).

By applying HMM, we construct a statistical model on each phone that its states are assigned specific possibilities in comparison with reference value. The possibility of each state depends on itself and the previous one. The goal of speech recognition system is to find out the sequence of states that has the maximum probability.

Same to the approach of Quan Vu et al. [2, 3, 4], we have applied the context- dependent model based on triphone to recognize words. Besides, we have defined the tied rules for the grammar.

Steps to build the Automatic Speech Recognizer

A. Training Data

The speech corpus has 9000 sentences. Total audio training covers 540 minutes. All speech was sampled at 16000Hz, 16bit by PCM format in a relatively quiet environment with 50 speakers.

The lexical comprises of 79 keywords as shown in Table I. For our application a part of the grammar is as follows:

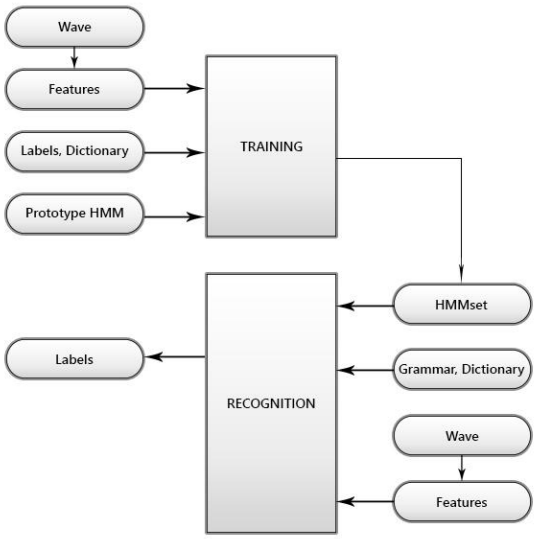

Fig. 2. Steps to build the Automatic Speech Recognizer [7]

TABLE I
LISTS OF WORDS

áo	ăn	ba	bạn	bao	bảy	bằng	bệnh
bốn	bò	cả	chiều	chín	cho	chồng	chơi
chợ	chủ	con	đi	điện	đinh	đồ	đổ
đóng	gia	hai	hết	học	hôm	khám	kia
lăm	linh	mất	mai	một	mốt	mua	mục
mười	mươi	này	năm	nay	ngàn	ngày	nhật
nhà	nhiêu	qua	quần	sài	sách	sáng	Sáu
sắm	sửa	tất	tám	thành	thay	thêm	thứ
tiền	tiêu	tinh	tối	tới	trăm	triệu	trưa
trước	tuần	tư	vào	vợ	xăng	xóa	

B. Grammar Rules / Constraints

A grammar is a kind of constraint that defines the set of phrases that a speech recognition engine can use to match speech input. We can either provide the speech recognition engine with the predefined grammars that are included with custom grammars that we create.

HTK also provides a grammar definition format, and an associated HParse tool, that can be used to build this word network automatically. We write the grammar definition in a file called gram.txt (or gram).

For our application a part of the grammar is as follows:

```
$day = ([VAFO] NGAFY $ngay) | ([VAFO] THUWS $thu) | ([NGAFY] MAI) | ([NGAFY] MOOST) | ([NGAFY] HOOM NAY);
…
$v_spend = (TIEEU | SAFI | MAAST) HEEST | HEEST | TAAST CAR;
$ask_sen = $ask [$v_spend] BAO NHIEEU;
$result = $add_sen | $del_sen | $upd_sen | $ask_sen;
(SENT-START $result SENT-END)
```

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

IV. TEXT TO SPEECH

Text-to-Speech system is a system that converts free text into speech. This is a process that a computer reads out the text for people. There is a wide range of application for text-to-speech system.

A typical text-to-speech system consists of three main parts, which are text analysis, prosody generation and speech synthesis. The text analysis part understands the text and determines the sound of each word. The prosody generation part generates some parameters that control the variability of the speech. The speech synthesis part generates the speech utterance based on the pronunciation and prosody requirements.

In recent years, many approaches have been used to synthesize speech for Vietnamese such as "Sao Mai" of Sao Mai Center, "Hoa Súng" of Mica Research Center [8] , "Tiếng nói phương Nam" of University Of Science Ho Chi Minh City [5]. The main approaches can be classified into two main categories, i.e. rule-based formant synthesis and concatenation synthesis. Formant synthesis generates speech using a set of rules. The rules are usually accumulated during a long process of experiments. This approach needs small computer memory. But the speech quality is not very good. Concatenation synthesis, however, uses some pre-recorded speech units as templates. During synthesis, the speech units are usually modified using signal processing techniques, and then concatenated together to form an utterance. This approach usually needs a larger memory. But the speech quality is relatively better.

Our system has adopted the Unit selection-based speech synthesis method [9]. Its blueprint can be seen in Figure 3.

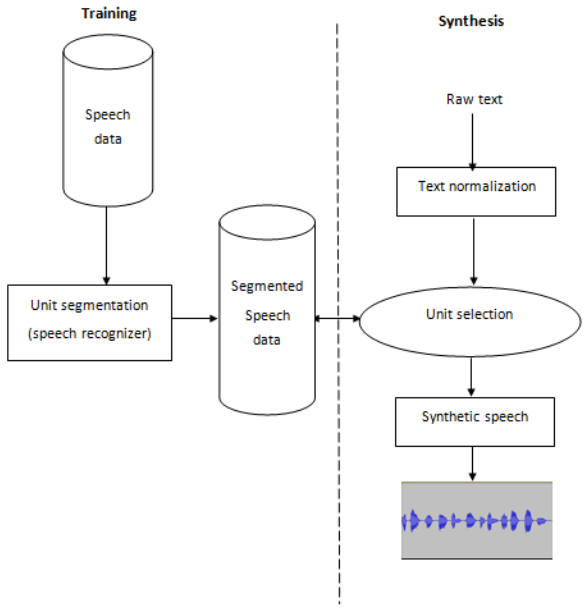

Fig. 3. Speech Synthesizer blueprint

V. EXPERIMENTS AND EVALUATION

We have carried out test the Automatic Speech Recognizer component and have conducted an investigation on how people response to the VNSExpenses.

A. Speech recognizer

1. Evaluation Score

The speech recognition performance is typically evaluated in terms of Word Error Rate (WER), which can then be computed as: WER= (S + D + I) / N x 100%, where N is the total number of words in the testing data, S denotes the total number of substitution errors, D is the total number of deletion errors and I is the total number of insertion errors.

We make use of Word Accuracy (WA) instead, which is computed as WA = (1 – (S + D + I)/ N) x 100%, to report performance of the speech recognizer.

2. Performance

TABLE II
Regression Test Result by Area

Model	Descriptions	Results (accuracy)		
		North	Center	South
VNSE_ A1	Train only corpus of south speakers	85%	65%	95%

TABLE III
Regression Test Result by Gender

Model	Descriptions	Result (accuracy)	
		Female	Male
VNSE_ G1	Train only corpus of male speakers	80%	95%

TABLE IV
Regression Test Result by Age

Model	Descriptions	Result (accuracy)	
		18-30	Others
VNSE_G1	Train only corpus of speakers are from 18-30 years old	97%	90%

TABLE V
Regression Test Result by Capacity of Corpus

Model	Descriptions	Result (accuracy)	
		Trained users	Untrained users
VNSE_C1	Train only corpus of 1 speakers	100%	40%
VNSE _C10	Train only corpus of 10 speakers	100%	75%
VNSE _C25	Train only corpus of 25 speakers	99%	85%
VNSE _C50	Train only corpus of 50 speakers	98%	95%

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

B. Investigation

We have made a survey on the VNSExpenses system by asking users' views with the question: "Is the VNSExpensese easy to use?" Table VI shows the users responses to the system.

Table VI
Show the users responses to the system

Very comfort	Fairly comfort	A bit comfort	Not comfort
26%	27%	24%	23%

VI. CONCLUSION

This paper has presented the architectural model of VNSExpenses system as well as our approach to build the speech processing components. In next steps, our jobs to be accomplished have wide vocabulary to realize the application as well as to develop similar applications based on this research background.

REFERENCES

[1] Thang Vu and Mai Luong, (2012). "The Development of Vietnamese Corpora Toward Speech Translation System". RIVF-VLSP 2012. Ho Chi Minh City, Viet Nam.
[2] Duong Dau, Minh Le, Cuong Le and Quan Vu, (2012). "A Robust Vietnamese Voice Server for Automated Directory Assistance Application". RIVF-VLSP 2012. Ho Chi Minh City, Viet Nam.
[3] Hue Nguyen, Truong Tran, Nhi Le, Nhut Pham and Quan Vu, (2012). "iSago: The Vietnamese Mobile Speech Assistant for Food-court and Restaurant Location". RIVF-VLSP 2012. Ho Chi Minh City, Viet Nam.
[4] Quan Vu et al., (2012). "Nghiên cứu xây dựng hệ thống Voice Server và ứng dụng cho các dịch vụ trả lời tự động qua điện thoại". Technical report, Research project, HCM City Department of Science and Technology, Viet Nam.
[5] Vũ Hải Quân, Cao Xuân Nam (2009), "Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ". Technical report, HCM City Department of Science and Technology, Viet Nam.
[6] Steve Young et al., (2006). The HTK Book (version 3.4). [On-line]. Available: www.htk.eng.cam.ac.uk/docs/docs.shtml [Nov. 1, 2012].
[7] Thien Khai Tran, Dang Tuan Nguyen (2013). "Semantic Processing Mechanism for Listening and Comprehension in VNSCalendar System". International Journal on Natural Language Computing (IJNLC) Vol. 2, No.2, April 2013.
[8] Trần Đỗ Đạt, "Synthèse de la parole a partir du texte en langue Vietnamienne", Ph.D. Thesis, Thèse en cotutelle international MICA, Hanoi, 2007.
[9] A. Hunt, A. Black and W. Alan, "Unit selection in a concatenative speech synthesis system using a large speech database," Pro c. ICASSP-96, 1, pp. 373{376 (1996).