Speech Recognition in Human-Computer Interactive Control

Vu Duc Lung, Phan Dinh Duy, Nguyen Vo An Phu, and Nguyen Hoang Long
University of Information technology, Vietnam National University HCMC, Vietnam
Email: {lungvd, duypd}@uit.edu.vn

Truong Nguyen Vu
Institute of Applied Mechanics and Informatics, Vietnam
Email: truong.nguyen.vu@gmail.com

***Abstract*—This paper gives an introduction of speech recognition systems for human-computer interaction using Vietnamese language. First, the paper investigates the two most common speech recognition toolkits currently used, HTK and Sphinx, and apply these tools into Vietnamese speech recognition. Then, only HTK is selected to design an application that control a web browser (Google Chrome) and an application that control a robot. The results obtained give a comparison between the uses of the two tool kits for Vietnamese speech recognition.**

***Index Terms*—information technology, speech recognition, control by speech, hmm, mfcc, htk, sphinx, vietnamese.**

I. Introduction

Human-machine interaction systems are increasingly invested these days, in which speech-based interaction methods is in high attention due to the nature of human voice. There are many toolkits for the implementation of speech recognition, but more prominent are Speech Recognition of Microsoft, HTK of Machine Intelligence Laboratory, and Sphinx of CMU.

In the world, many research groups in speech-based interaction field can be listed such as HCI group at Stanford University, The Human-Computer Interaction group at Microsoft Research Asia (MSRA HCI), The Human-Computer Interaction (HCI) group at Department of Computer Science, University of Toronto. However, the majority of research work is about English, French, Spanish, Japanese, and only a little research work about Vietnamese. In Vietnam, there are also many research groups in this field, the most typical are assoc. prof. Luong Chi Mai's group at Institute of Information Technology with the use of ANN method and CLSI tool, and assoc. prof. Vu Hai Quan's group at AILab, HCMC University of Science with the use of HMM [7] method and HTK tool. These research groups work independently and use different platforms and data sets to identify Vietnamese but lack the comparison with each other to find the most suitable tool for Vietnamese speech recognition.

From the idea to compare speech-recognition tools for Vietnamese, the group has come up to building two Vietnamese speech-recognition systems using HTK and Sphinx. These systems use the same data set for training and recognition. The recognition results are used to implement applications to control a computer software program (Google Chrome) and a peripheral device (toy tank) to demonstrate the applicability of Vietnamese speech recognition systems.

Although there is no new scientific contribution, the paper has assisted those who are embarrassed in choosing an appropriate tool for Vietnamese speech recognition. Besides, the two applications built in this paper show the potentiality of Vietnamese speech-based human-computer interaction in future.

II. Basic Theory of Speech Recognition

Fig. 1 illustrates basic steps in a speech recognition diaphragm.

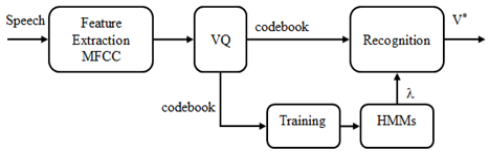

Figure 1. Diaphragm of basic speech recognition system

A speech recognition system generally consists of two primary processes: training and recognition. In the training process, the voice will go to a preprocessing to remove interference and background noise of the environment. Then, feature extraction methods will extract the most features of human voice that the computer can use as the base for voice description under a set of vectors. Feature extraction is an important stage that highly determines the accuracy of a speech recognition system. Among many feature extraction methods of voice, MFCC method [1] is more commonly used as it models the features of voice according to Mel scale, which is similar to the way human ears operate.

Manuscript received December 20, 2012; revised February 4, 2013.

doi: 10.12720/joace.1.3.222-226

Other extraction methods can be employed such as LPC, PLP, etc. The result of feature extraction stage is a set of vectors that are the most features of human voice. These vectors will be quantified again (Vector Quantization – VQ[8]) into a smaller set called codebook, which still keep the features of voice, for the convenience in storage, learning, and recognition later on. The size of this codebook depends on the size of the recognition system, the necessary accuracy, and the system resources. This codebook takes into the machine learning system to model parameters for the words to be recognized. The popular machine learning methods in speech recognition system are HMM [2], ANN, etc. in which HMM is the most commonly used. In the recognition process, the voice signals needed to be identified also go through pre-processing, feature extraction, and vector quantization process to be converted to codebook. These codebooks are included into the recognition system to calculate the likelihood with the reference models built in the training process. Together, the dictionary, language model, and grammar model in the system help to determine the word with highest similarity to be recognized.

III. Construction of Vietnamese Speech Recognition System on HTK and Sphinx

A. HTK

HTK is a set of tools to build speech-recognition system based on Hidden Markov Modeling (HMM). It is used as a library set, easy to expand and develop. This is an ideal tool set to study speech recognition model for every languages using HMM.

The tools in the HTK framework [6] are designed to perform different tasks of building the HMM. Building a speech recognition system on HTK requires tools to implement four stages: data preparation, training, testing, and result analysis as shown in Fig. 2.

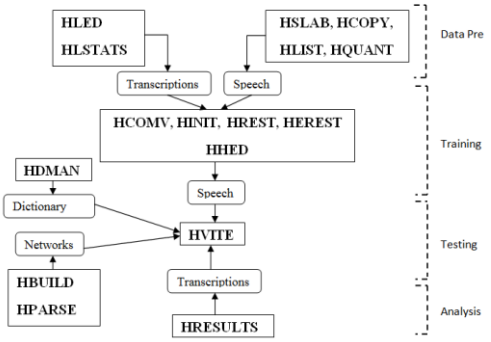

Figure 2. Stages of building a speech recognition system on HTK

In the data preparation stage, the tools HSLAB, HCOPY, HLIST, HQUANT process the voice signals obtained from the microphone into codebooks according to the feature extraction method of a speech recognition system. HDMAN creates a dictionary of words to be recognized. HLED, HLSTATS produce transcriptions to read a list of HMMs and a set of voice accents. HLSTAT calculates various statistics to analyze training phonetic data and generates language models for identification.

Next, in the training stage, HTK provides four major tools to estimate the parameters for the HMM: HCOMV, HINIT, HREST, HEREST. In which, HCOMV and HINIT are used to initialize the values of the parameters. HCOMV will calculate the expectation and variance of each Gaussian component in the HMM definition to make them nearly equal to the expectation and variance of the speech training data. HINIT will calculate the values of the new HMM by applying estimation formulas such as Viterbi algorithm. HREST will estimate the parameters on a label-assigned data segment using Baum-Welch algorithm. HEREST will concurrently train the data set based on the Baum-Welch algorithm. HHED is used to create a new HMM after parameter adjustment. The result is to create the HMMs in accordance with the dictionary of words to be recognized.

In the identification process [4], to acquire good results, this paper has developed the syntax and semantics of the Vietnamese language primarily for the structure of control commands and the recognition dictionary. There are a number of supportive tools in HTK. HPARSE is used to change a syntax file of the dictionary into a semantic network and the possibility of the words arrangement in an order. HGEN is the opposite of HPARSE tool. To identify a word, we will use HCOPY to extract features and HQUANT to convert them into a codebook. HVITE tool will apply Viterbi algorithm for continuous speech recognition based on the HMM model, the dictionary and the grammar structure constraints.

B. Sphinx

Sphinx [5] is a powerful speech recognition framework and widely used in many applications. Fig. 3 displays three basic components of Sphinx: Frontend, Decoder and Knowledge base.

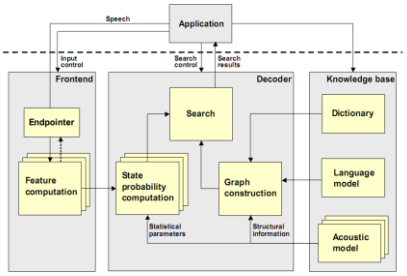

Figure 3. Components of Sphinx

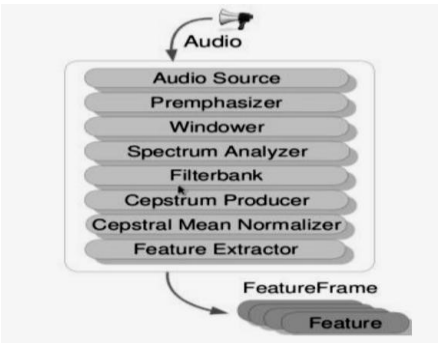

Figure 4. Stages in Frontend module

Frontend module will input voice signals from the outside, put them through a number of filters and data processors so that the result is a set of feature vectors described in Fig. 4.

Linguist module has tools to read the structure files of a language and modeling them in the search graph to use in the recognition search. The Linguist consists of Language Model, Dictionary, Acoustic Model and the search graph [3].

Decoder module uses the feature sets from the Frontend module to associate with the search graph from the Linguist module to conduct decoding process and applying algorithms to infer recognition results.

The Sphinx installation process consists of four main steps to install: SphinxBase (base library), SphinxTrain (library of language model training), PocketSphinx (recognition library by C language), and CMUclmtk (toolkit of language modeling). From the characteristics of Vietnamese, which is monophonic and with tone, building a recognition system suitable for Vietnamese is to focus on building an appropriate phonemic dictionary. The phonemic dictionary contains the way to pronounce each word in the training data set. A Vietnamese word can be defined in the type of 3-gram or 2-gram tone. We define a phoneme accompanies a tone is one independent phoneme (e.g. " à' different from "a").

IV. Building Demo Control Application

A. Google Chrome Control Application

In this application, the users will manipulate the Google Chrome web browser using 20 control commands formed by 47 single words as shown in Table I through human voice.

TABLE I. List of Words for Computer Program Control

bản	lịch	tập
chuyển	lưu	thu
cửa	mở	thư
cuối	mới	tìm
cuộn	nghe	to
đầu	ngừng	tới
đi	nhạc	tra
đồ	nhỏ	trái
đóng	phải	trang
duyệt	phóng	trình
hãy	quép	trở
khóa	sang	trước
kiếm	sau	từ
kiểm	sổ	xóa
lại	sử	xuống
lên	tải	

The application is written in C# language combined with Julius.dll library and acoustic model trained from HTK tool. There are two major modules in the program: Recognition and Web Browser Control.

Recognition module uses functions provided by Julius.dll library to perform identifying process, using training data from HTK. Then, the recognition result will be converted to text and transferred to Control module.

Control module contains three major classes performing the control tasks for three objects: Window, Chrome, and music player WMPlayer

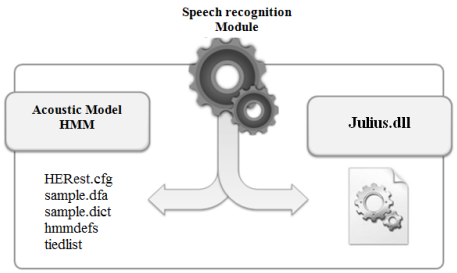

Figure 5. Speech recognition module

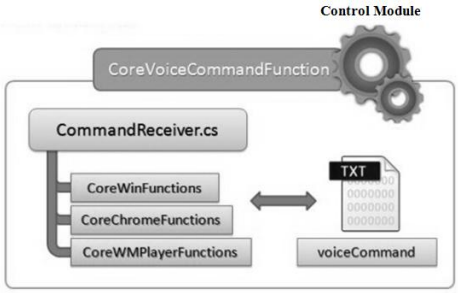

Figure 6. Control module

B. Toy Tank Control Application

This application will use Vietnamese voice to handle the operation of tank model remote control. The program supports more than 30 control commands formed by 25 single words as shown in Table II below.

Two major modules in the program are Speech Recognition module (written in Java language) and Tank Model Remote Control module (written in C# language). The two modules communicate with each other through socket. The operation of the speech recognition module is similar to the one in section 4.1.

TABLE II. List of Words for Tank Remote Control

ba	dừng	mươi	quay	tới
bắn	lại	ngừng	sang	trái
chạy	lên	nòng	sáu	trăm
đi	lui	phải	súng	vừa
độ	lui	qua	tiến	xoay

The remote control module works as a USB device driver. The computer will pass commands directly to control module and handle the tank through this module.

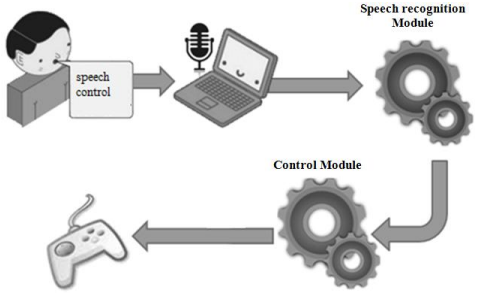

Figure 7. System of tank remote control using human voice

V. Results

A. Comparision Results of HTK and Sphinx

The training data has recording duration approximate 15 hours, with 5300 training sentences including control command sentences for computer program and tank model. The total of training words are 120 single words constructed from 72 phonemes. Testing data are 1000 sentences with recording duration approximate 2 hours. The comparison results are presented in Table III.

TABLE III. Comparison Result of Htk and Sphinx

	Percentage of correct sentences (%)	Percentage of correct words (%)	The accuracy of words (%)
HTK	41.60	99.97	94.38
SPHINX	68	98.2	96.7

In the result, percentage of correct sentences is equal to the total number of sentences that all words are recognized correctly divided by the total number of testing data (1000 sentences). Percentage of correct words is equal to the number of words recognized correctly divided by the total number of training words (120 words). The accuracy of words is equal to the total number of correct recognition times divided by the total number of testing times.

B. Result of Control Applications

In the design of Google Chrome control application under the environment not too noisy, using the training and testing data set as described in section 5.1, with 30 testing times for each command sentence (20 sentences in total), the accuracy is 90%. Fig. 8 shows the control program interface.

Figure 8. Control program interface

In the design of tank remote control application, using the same training and testing data set, with 30 testing times for each command sentence (30 sentences in total), the accuracy is about 87% due to the engine noise.

Figure 9. Tank remote control model

VI. Conclusion

From the testing results, the group has general evaluation for the two frameworks as follow:

- The ability/capability of recognizing correct words of the two frameworks is very high (more than 98%), and HTK is somewhat better.
- The decoding time of Sphinx is significantly shorter than HTK.
- HTK produces more insertion errors than Sphinx, resulting to reduced accuracy of sentence recognition.

Although the outcomes were simply tested, they demonstrate the ability to employ the two tool kits HTK and Sphinx into Vietnamese speech recognition and many natural humane-computer interactions in future.

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] Hossan, Memon, Gregory, "A novel approach for MFCC feature extraction," *Signal Processing and Communication System*, 4th 2010.
[2] R. Dugad and U. B. Desai, "A tutorial on hidden Markov models. Signal Processing and Artificial Neural Networks Laboratory," Department of Electrical Engineering, Indian Institute of Technology, Bombay Powai, India, May 1996.
[3] Training Acoustic Model for CMUSphinx. (July 2012). Carnegie Mellon University. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorialam
[4] Recording the Test Data. (July 2012) [Online]. Available: http://www.voxforge.org/home/dev/acousticmodels/windows/test/htk--julius/data-prep/step-3
[5] Sphinx-4 Application Programmer's Guide. (July 2012). Carnegie Mellon University, [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorialsphinx4
[6] S. Young, G. Evermann, M. Gales, T. Hain, and D. Kershaw, et al., *HTK Book*, Cambridge University Engineering Department, 2009.
[7] L. Rabiner, A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition, 1989.
[8] A. Gersho and R. M. Gray, *Vector Quantization And Signal Compression*, Kluwer Academic Publishers Group, Ninth Printing, 2003.

Phan Dinh Duy was born on October 26, 1988 in Binh Dinh province, Vietnam. He obtained his B.S. degree in Computer Engineering from the University of Information Technology where he is working on Circuit Design and machine learning.

Vu Duc Lung received the B.S. and M.S. degree in computer engineering from Saint Petersburg State Polytechnical University in 1998 and 2000, respectively. He got the Ph.D. degree in computer science from Saint Petersburg Electrotechnical University in 2006. From 2006 until now, he works at the University of Information Technology, VNU HCMC as a lecturer. His research interests include machine learning, human-computer interaction and FPGA technology.

He is a member of IEEE, ACOMP 2011 and Publication Chair of ICCAIS 2012.