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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Comparative Analysis of HTK and Sphinx in Vietnamese Speech Recognition

Hoa Minh Dinh[1], Tan Duy Nguyen[2], Thanh Duc Pham[3]

Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology,

Ho Chi Minh City, Vietnam[1,2,3]

Abstract: This paper presents the comparative analysis of HTK and Sphinx which are the two most popular open source automatic speech recognition systems. We used many datasets to investigate the performances of these tools in Vietnamese speech recognition.

Keywords: Speech Recognition, HTK, Sphinx, Vietnamese

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 Thien Khai Tran et al. [4,7,8,9,10,11], Thang Vu and Mai Luong [6] as well as Quan Vu et al. [1,2,3] ‘s one which obtained the precision rate of over than 90% and this group successfully built many voice applications on this base. Almost these publications used HTK for Speech recognition system. In this paper, we make a comparative analysis between HTK and Sphinx in Vietnamese Speech Recognition. The rest of the paper is organized as follows. In section II, the overview is described. Inspection III, we present the comparison for HTK and Sphinx. The conclusion is in Section V.

II. OVERVIEW

A. Hidden Markov Model Toolkit (HTK)

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.

B. Sphinx – a Speech Recognition System

Sphinx is a continuous-speech, speaker-independent recognition system making use of hidden Markov acoustic models (HMMs) and an n-gram statistical language model. It was developed by Kai-Fu Lee. Sphinx featured feasibility of continuous-speech, speaker-independent large-vocabulary recognition, the possibility of which was in dispute at the time (1986). Sphinx is of historical interest only; it has been superseded in performance by subsequent versions.

Sphinx-2 is a Speech recognition system. Now, this version wasn’t supported anymore. Sphinx-2 was built based on semi-continuous Hidden Markov Model (HMM). The accurate rate of Sphinx-2 is not as high as any other Sphinx recognition systems.

Sphinx-3 is a Big vocabulary Speech recognition system (state-of-the-art) based on continuous Hidden Markov Model , Sphinx-3 used Flat encoder with high-accurate rate and Tree decoder with the fasted decode algorithm. Sphinx-4 is the most completed Speech recognition system. Sphinx-4 also used continuous Hidden Markov Model. Sphinx-4 is written using Java™.

Pocket Sphinx is known as the fastest Speech recognition system using semi-continuous Hidden Markov Model. The accuracy of Pocket Sphinx is not as high as Sphinx-3 and Sphinx-4 but it was designed for real-time applications.

III. COMPARISION FOR HTK AND SPHINX-4

Recently, we have published some research papers about PBX system integrated Vietnamese speech recognition. During the progress, we compared the performance of HTK and Sphinx4 to choose the most optimized toolkit. All of our papers used the same experimental environments:

Environment	in-door
Sampling rate	8 kHz
Quantization	16 bits
Format	PCM

The followings are the comparision results from the most 5 typical papers of us.

DOI 10.17148/IJARCCE.2015.45156

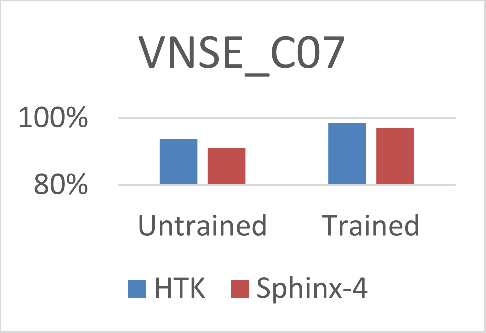
A. EDU voice - a System for Querying Academic Information via PSTN [9]

We has built an Vietnamese speech recognition system for Querying Academic Information via PSTN. There are 2429 sentences in the speech corpus. Total audio training covers 160 minutes. All speech was sampled at 8000Hz, 16bit by PCM format in a relatively quiet environment with 7 speakers.

We have comparative analysis table:

	HTK		Sphinx-4	
	Untrained	Trained	Untrained	Trained
VNSE_C01	64%	99%	62%	99%
VNSE_C05	90%	99%	89%	98%
VNSE_C07	93.73%	98.49%	91%	97%

We have a chart to compare result between HTK and Sphinx-4 with the largest capacity of corpus (VNSE_C07).

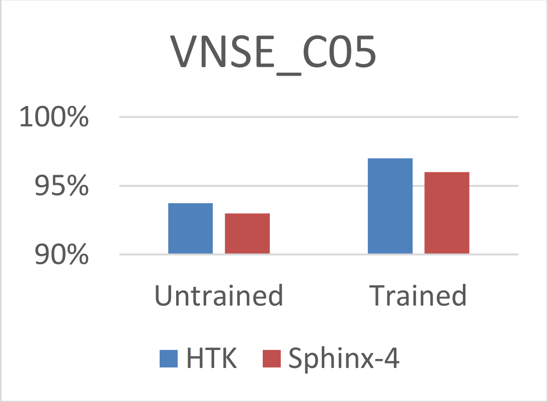


B. SentiVoice - a system for querying hotel service reviews via phone [7]

We also built an System for querying hotel service reviews via phone - SentiVoice in paper[].There are 250 sentences in the speech corpus. Total training only covers 20 minutes. All speech was sampled at 8000Hz, 16bit by PCM format in a relatively quiet environment with 5 speakers. Test result by capacity of corpus shown on the following table:

	HTK		Sphinx-4	
	Untrained	Trained	Untrained	Trained
VNSE_C01	55%	99%	53%	98%
VNSE_C03	89%	98%	88%	98%
VNSE_C05	94%	97%	93%	96%

Below is the comparisons chart of the result:

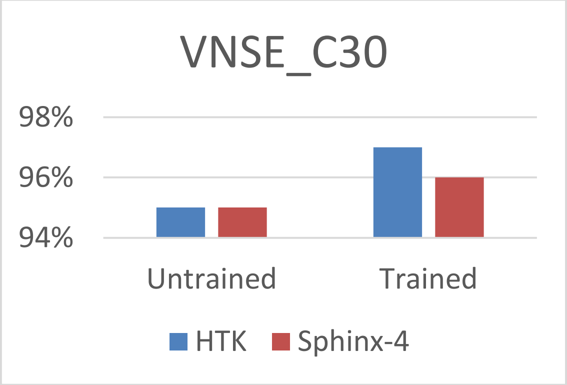


C. Edu ICR - an Intelligent Call Routing system [11]

In this paper, we built an intelligent call routing system to route all incoming calls to the most appropriate agent. There are 3500 sentences in the speech corpus. Total audio training only covers 3 hours. All speech was sampled at 8000Hz, 16bit by PCM format in a relatively quiet environment with 35 speakers. The accuracy of the system is compared between HTK and Sphinx-4 in the following table:

	HTK		Sphinx-4	
	Untrained	Trained	Untrained	Trained
VNSE_C10	93%	99%	93%	98%
VNSE_C20	94%	98%	94%	97%
VNSE_C30	95%	97%	95%	96%

Below is the comparisons chart of the result:

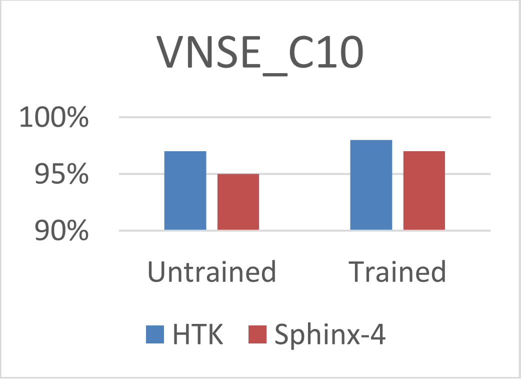


D. Admission Information Lookup System of HUFLIT using Voice – i Voice Server [10]

This system can recognite speech command of users and return the appropriate information. There are 2550 sentences in the speech corpus. Total training only covers 262 minutes. All speech was sampled at 8000Hz, 16bit by PCM format in a relatively quiet environment with 10 speakers. Test result by capacity of corpus shown on the following table:

	HTK		Sphinx-4	
	Untrained	Trained	Untrained	Trained
VNSE_C01	44%	99%	43%	98%
VNSE_C05	86%	98%	86%	98%
VNSE_C10	97%	98%	95%	97%

Below is the comparision chart of the result:



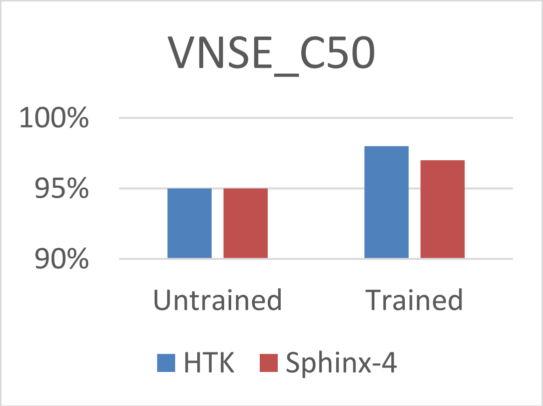
DOI 10.17148/IJARCCE.2015.45156

E. Vietnamese Speech Processing and Synthesis in VNS Expenses System [4]

This system helps users to manage their personal expenses by Vietnames Speech. 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. Test result by capacity of corpus shown on the following table:

	HTK		Sphinx-4	
	Untrained	Trained	Untrained	Trained
VNSE_C01	40%	100%	38%	100%
VNSE_C010	75%	100%	74%	100%
VNSE_C25	85%	99%	85%	98%
VNSE_C50	95%	98%	95%	97%

Below is the comparision chart of the result:



IV. CONCLUSION

In this paper, the experiments on spoken Vietnamese for the speech recognition between HTK and Sphinx were described. The 2 methods HTK and Sphinx-4 that we have mentioned in this paper have similar results. HTK can be installed simply and be executed quickly but Sphinx can easily be integrated into application systems by using support from Java™.

ACKNOWLEDGMENTS

We would like to thank Doctor Candidate Thien Khai Tran for his supporting by recommending and supplying corpus for this research.

REFERENCES

[1]. 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.

[2]. 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.

[3]. 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.

[4]. Quoc The Van, Nguyen B. P. Nguyen, Anh K. V. Nguyen, Hien Thanh Vu, Thien Khai Tran “Vietnamese Speech Processing and Synthesis in VNSExpenses System”. International Journal of Advanced Research in Computer and Communication Engineering. Vol. 3, Issue 4, 2014.

[5]. 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].

[6]. Thang Vu and Mai Luong, (2012). “The Development of Vietnamese Corpora Toward SpeechTranslation System”. RIVF-VLSP 2012. Ho Chi Minh City, Viet Nam.

[7]. Thien Khai Tran (2015), “SentiVoice - a system for querying hotel service reviews via phone”, The 11th IEEE-RIVF International Conferenceon Computing and Communication Technologies (RIVF 2015), Cantho, 2015.

[8]. 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.

[9]. Thien Khai Tran, Tien Cat Khai Tran, Tho Anh Mai, Nhat Minh H. Nguyen and Hien Thanh Vu (2014), “EDUVoice - a system for querying academic information via PSTN”, The Third Asian Conference on Information Systems (ACIS 2014). Nha Trang, 2014.

[10]. Trần Khải Thiện, Văn Thế Quốc, Nguyễn Phạm Bảo Nguyên, Nguyễn Vũ Kiều Anh, Vũ Thanh Hiền (2014), "Hệ thống Tra cứu thông tin tuyển sinh trường đại học Huflit qua mạng điện thoại", Kỷ yếu Hội nghị quốc gia lần thứ VII "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin", (FAIR 2014), Thái Nguyên, 2014

[11]. Thien Khai Tran, Hoa Minh Dinh, Phuong Hong Vo, Tan Duy Nguyen, Dung Minh Pham, Binh Van Huynh, “EduICR - an Intelligent Call Routing system,” The second NAFOSTED Conference on Information and Computer Science 2015, Viet Nam.

DOI 10.17148/IJARCCE.2015.45156