Journal of Theoretical and Applied Information Technology
10th July 2013. Vol. 53 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPHINX-4 INDONESIAN ISOLATED DIGIT SPEECH RECOGNITION

[1]IKA NOVITA DEWI, [2]FAHRI FIRDAUSILLAH, [3]CATUR SUPRIYANTO
Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
E-mail: [1]ikadewi@research.dinus.ac.id , [2]fahri@research.dinus.ac.id , [3]caturs@research.dinus.ac.id

ABSTRACT

Sphinx-4 is one of speech recognition engine which is the latest addition to Carnegie Mellon University's (CMU) repository of Sphinx speech recognition system. This study was applied Sphinx-4 for Indonesian digit speech recognition using Java^{TM} programming language. Seven men were selected in order to test the Indonesian digit speech recognition system in term of recognition rate. The result of the experiment showed that each man had different number of recognition rate caused by the difference tone, pronunciation, and speed of speech. The future works of this study will be mapping English-Indonesian pronunciation in order to increase the recognition rate and compare men-women recognition rate to identify gender aspects in speech recognition.

Keywords: *Speech Recognition; Indonesian Isolated Digit; Sphinx-4*

1. INTRODUCTION

Speech recognition that is also known as Automatic Speech Recognition (ASR) is one of the human computer interactions in interacting with machine by speaking through a microphone as an input device and the system will convert spoken words into text as an output [1]. ASR allows human to speak into the computer and faster to be used instead of using mouse and keyboard [2]. ASR has been used not to replace the overall use of keyboard and mouse but to support in entering data. ASR has been applied in many applications area, such as command recognition, dictation, interactive voice response, and dictation [3].

Indonesian language is the national language of Indonesia that widely used by Indonesian. Some research related to Indonesian speech recognition has been done by other researchers. Ferdiansyah and Purwarianti [4] investigated the use of English-based acoustic model for Indonesian speech recognition. Desi Puji *et al.* [5] has worked on Indonesian large vocabulary continuous speech recognition system and find that the developed system had low accuracy rate compared to another system for other languages.

Some researches have applied Indonesian speech recognition during recent years, however recognizing digit is needed in order to increase the number of research fields in Indonesian speech recognition. Digit recognition will identify the digit 0 to 9 that spoken in Indonesian. Digit recognition was chosen since small vocabulary and small enough to study in detail.

One of the speech recognition engines is Sphinx. Sphinx speech recognition system that currently is one of the most robust speech recognizers and freely available that developed by CMU university. This study is aimed to apply sphinx-4 for ; recognition and measuring recognition rate of the tested system. The developed system will apply isolated speech recognition in dictation mode and required to use digit from 0 to 9. The developed speech recognition was built on Java programming language and used Sphinx-4 library as a speech recognition engine.

2. RELATED WORKS

ASR is one of natural communication that aimed to give an intelligent into machine (computer) to have an interaction with human (man). ASR interacted with computer by using a voice through a microphone as an input device and resulting spoken word as an output converted by acoustic signal.

Some researchers gave the definition of automatic speech recognition system. Anusuya and Katti [6] gave the definition of ASR that is a process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program in order to develop techniques and systems for speech input to machine. According to Abushariah and Gunawan [7], ASR systems aimed to automatically extract the string of spoken words from input speech signals. Kirriemuir [8] stated that in order to build speech recognition system needed some components to be applied, such as microphone, for speaking into the system; speech recognition software, and a good quality soundcard.

Kurian [9] said that there are some benefits gotten by applying speech recognition for valuable applications, such as telephone directory assistance, spoken database querying for novice users, “hands busy“ applications in medicine, office dictation devices, and automatic voice translation into foreign languages. Kurian [9] also identified two main factors that affecting the recognition rate in speech recognition, those are acoustic variability that covers different accents, pronunciation, pitches, and volume; and temporal variability covers different speaking rates.

Benzeguiba *et al.* [10] stated that there are many factors affecting the speech realization, such as articulation and pronunciation, regional, sociolinguistic, or related to the environment or the speaker her-self covers speaker, gender, age, speaking rate, vocal effort, regional accent, speaking style, and non stationarity.

3. ISOLATED DIGIT SPEECH RECOGNITION

According to Gaikwad [11], isolated word is one of the speech recognition system type based on utterance that able to be recognized by the system. An isolated word accepts single words and requires a single utterance at a time. Isolated speech recognition play a crucial role in the future communication since speech will become the preferred medium for interacting with telecommunication services, having the argument that speech offers a very natural means of interaction with their services [12].

ASR has been applied for isolated digit recognition in widespread applications, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, data entry, voice dictation, access to information: travel, banking, commands, avionics, automobile portal, speech transcription, handicapped people (blind people) supermarket, railway reservations [12].

Speech recognition in recognizing digit had been successfully applied in English, Arabic, Malayam and Malay. This study will apply isolated digit speech recognition in order to recognize Indonesian spoken digit starting from 0 to 9. There are ten distinct Indonesian digits namely nol, satu, dua, tiga, empat, lima, enam, tujuh, delapan, and sembilan. Speakers say these ten digits in isolation mode which is presented under the isolated words speech recognition system.

4. DEVELOPING INDONESIAN DIGIT SPEECH RECOGNITION

The Indonesian isolated digit speech recognition will be built on Java programing language and using Sphinx-4 library.

4.1 Sphinx-4 Overview

Sphinx stands for Site-oriented Processor for HTML INformation eXtraction is one of speech recognition engine that developed by The Sphinx works based on Hidden Markov Model (HMM) algorithm. The Sphinx that will use in this study is Sphinx-4 that written entirely in the Java^{TM}. Sphinx-4 is created via a joint collaboration between the Sphinx group at CMU, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), and Hewlett Packard (HP), with contributions from the University of California programming language at Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT).

Sphinx recognition system is part of the CMU Sphinx, an open source technology that provides a set of speech recognizers and tools that allow the development of speech recognition system [13]. Sphinx-4 is able to recognize both discrete and continuous speech. Sphinx is generalizing front end architecture, such as pluggable implementations of preemphasis, Hamming window, FFT, Mel frequency filter bank, discrete cosine transform, cepstral mean normalization, and feature extraction of cepstra, delta cepstra, double delta cepstra features, and also capable of breadth first and word pruning searches [14].

Some advantages in using Sphinx-4 as speech recognition engine has been identified by Walker *et al.* [15]. Sphinx-4 is possible to run on a variety of platforms. Sphinx-4 consists of the rich set of platform APIs that effect to reduction of coding time, supports in multithreading and making simple to experiment with distributing decoding tasks across multiple threads, and automatic garbage collection helps developers to concentrate on algorithm development instead of memory leaks.

Fig. 1 show the Sphinx-4 architecture based on Lamere *et al.* [16].

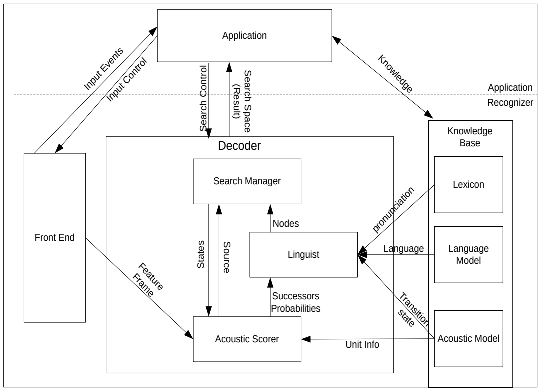

Fig. 1 Sphinx-4 Architecture

Lamere *et al.* [16] explained that Sphinx-4 architecture consists of three main blocks: FrontEnd, Decoder, and KnowledgeBase. FrontEnd has a function to parameterize an input signal such as audio into a sequence of output Features through an analysis process. The decoder block consists of three modules: search manager, linguist, and acoustic scorer. The primary function of search manager is constructing and searches a tree of possibilities for the best hypothesis. The construction of the search tree is done based on information obtained from the linguist. The linguist translates linguistic constraints provided to the system into an internal data structure, the grammar, which is usable by the search manager. Knowledge base consists of lexicon, language model, and acoustic model.

4.2 Indonesian Digit Speech Recognition

Sphinx-4 has been applied for recognizing Indonesian digit spoken. The corpus of ten digits Indonesian consists of 0 to 9. Seven men speaker required to speech the digits into the system. This corpus is used for three repetitions for each speaker. The total corpus gotten is 210 tokens (10 digits, 3 times, 7 men speakers).

The use of Sphinx-4 required using acoustic model that taken from corpus file and generated online by using sphinx lm tool [17]. Table I showed the pronunciation dictionary of the Indonesian digits from 0 to 9 taken from [17].

Language model was formed in order to model the grammar of the spoken digit. Fig. 2 showed the language model used for Indonesian digit.

Table 1: Pronunciation Dictionary

Digit	Indonesian	Pronunciation
0	Nol	N aa l
1	Satu	S ae t uw
2	Dua	D y uw ah
3	Tiga	T ih g ah
4	Empat	Ih m p ae t
5	Lima	L ay m ah
5	Lima(2)	L iy m ah
6	Enam	Ah n ae m
7	Tujuh	T y uw jh uw
8	Delapan	D ah l ey p ah n
9	Sembilan	S eh m b ah l ah n

```
#JSGF V1.0;
/**
 * JSGF Grammar for Indonesian Digit Recognition
 */
grammar angka;
public <angka> = ( Nol | Satu | Dua | Tiga | Empat |
Lima | Enam | Tujuh | Delapan | Sembilan);
```

Fig. 2 Language model

5. RESULT AND DISCUSSION

The experimental phase was taken in order to measure the recognition rate of the spoken digit. The speech system was built by using Compaq Presario V3737, Windows7 Professional 32-bit, Intel(R) Core(TM) 2 Duo CPU T7250 @2.00GHz, 2.50 GB RAM, and AVF HM-338 Headset.

5.1 Experimental Result

Resulting the testing of the experiment required seven men (M) to speech into the system. The recognition rate was gained based on

$$RR = \frac{N_{Correct}}{N_{Total}} x100\% \qquad (1)$$

Where RR is recognition rate, $N_{Correct}$ is the number of correct recognition of the spoken digit, and N_{Total} is total number of samples spoken digit. Table II showed the result of each man that correctly spoken the digit into the system.

Table 2: Recognition Rate

	M1	M2	M3	M4	M5	M6	M7
Test 1	50	50	30	50	30	40	30
Test 2	50	60	20	70	40	60	20
Test 3	40	70	50	90	70	90	20
Mean Recognition Rate (%)	46.6	60	33.3	70	46.6	43.3	23.3

5.2 Discussion

The purpose of developing Indonesian digit speech recognition based on sphinx-4 is extending the research in Indonesian speech recognition. Some advantages are able to obtain by developing Indonesian digit speech recognition, such in data collection. By using speech recognition in data collection is able to reduce the time taken in entering the data. Recognition rate is needed in line with the ability of speech recognition to recognize spoken the words

The Indonesian digit speech system was built on Java^{TM} programming language and applying Sphinx-4 as a recognition engine. The first step in preparing the developed system was Indonesian digit corpus files that consist of 0 to 9. The corpus file was online generated became a language model by using Sphinx Knowledge based tool. The next step was installed the JSapi that contained on Sphinx-4 downloaded package. Then, creating a new Java application project in NetBeans and making class interface on it.

The testing result showed that each man had different recognition rate in each testing phase. Averagely, the recognition rate of experiment is not more than 50%.

In the testing phase, each man required to speech in his own way, means that each men speech in natural way like when having a conversation with others. The way of speaking into the speech system gave some effects in the recognition rate. There were three factors that influenced the recognition rate: tone, pronunciation, and speed of speech.

Tone covered how grammar is pronounced for each man based on high or low of the sounds. Pronunciation included how each man spoken the correct words and the clarity of the intonation of a word. The tone and pronunciation were performed based on the speed of speech when each man spoke into the developed system.

6. CONCLUSION

The Indonesian digit speech recognition system was built by using Java Programming language and using Sphinx-4 library in order to identify the speech recognition spoken in Indonesian digit from zero to nine. Seven men were selected in order to test the developed system. The resulted recognition rate averagely is not more than 50% and influenced by three factors, tone, pronunciation, and speed of speech.

There were some other works that should be conducted to improve the Indonesian digit speech recognition. Mapping English - Indonesian pronunciation can be used to gain the higher recognition rate and comparing men and women recognition rate in order to expand the Indonesian digit speech recognition.

REFERENCES:

[1] I. N. Dewi, "Recording Approach for Patient Health Record: A Comparison between Speech Recognition and Text Input Using Computer Keyboard," MS. Thesis, Universiti Teknikal Malaysia Melaka, 2012.

[2] G. Sreenu, P. N. Girija, M. N. Prasad, and M. Nagamani., "A human machine speaker dependent speech interactive system," in Proc. of the IEEE INDICON, 2004, pp. 349–351.

[3] H. Satori, M. Harti, and N. Chenfour, ”Introduction to Arabic Speech Recognition Using CMUSphinx System," Information and Communication Technologies International Symposium, pp. 2-5, 2007.

[4] V. Ferdiansyah, and A. Purwarianti, "Indonesian automatic speech recognition system using English-based acoustic model," in Proc. of International Conference Electrical Engineering and Informatics (ICEEI), 2011, pp. 1-4.

[5] D. P. Lestari, K. Iwano, and S. Furui, "A large vocabulary continuous speech recognition system for Indonesian language," in Proc of 15th Indonesian Scientific Conference in Japan (ISA-Japan), 2006, pp.17–22.

[6] M. A. Anusuya, and S. K. Katti, "Speech Recognition by Machine: A Review," International Journal of Computer Science and Information Security, vol.6, no.3, pp. 181-205, 2009.

[7] M. A. M. Abushariah, T. S. Gunawan, and O. O. Khalifa, “English Digits Speech Recognition System Based on Hidden Markov Models,” in Proceedings of International Conference Computer and Communication Engineering (ICCCE), 2010, pp. 1-5.

[8] John Kirriemuir, 2003. Speech recognition technologies. JISC. [Online] May 2012. Available: http://www.jisc.ac.uk/uploaded_documents/tsw_03-03.pdf

[9] C. Kurian, and K. BalaKrishnan, “Speech recognition of Malayalam numbers,” in Nature & Biologically Inspired Computing, pp. 1475-1479, 2009.

[10] M. Benzeguiba, R. De Mori, O. Deroo, and S. Dupont, “Automatic speech recognition and speech variability: A review,” Speech Communication, vol.11, pp.10-11, 2007.

[11] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A Review on Speech Recognition Technique,” International Journal of Computer Applications, vol.10, no.3, pp.16-24, 2010.

[12] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “Human Computer Interaction Using Isolated-Words Speech Recognition Technology,” in Proc. of International Conference on Intelligent and Advanced System, pp.1173-1178, 2007.

[13] I. F. C. Cerón, and A. G. G. Badillo, “A Keyword Based Interactive Speech Recognition System for Embedded Applications,” MS Thesis, School of Innovation, Design and Engineering-Malardalen University, 2011.

[14] Sphinx-4, a speech recognizer written entirely in the Java^{TM} programming language. [Online] October 2011. Available at: http://cmusphinx.sourceforge.net/sphinx4/

[15] P. Lamere, P. Kwok, E. B. Gouvêa, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, and J. Woelfel,., 2004. Available at:http://cmusphinx.sourceforge.net/sphinx4/doc/Sphinx4Whitepaper.pdf

[16] P. Lamere, P. Kwok, E. B. Gouvêa, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, “The CMU SPHINX-4 speech recognition system,” in Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 2-5,2003.

[17] Sphinx Knowledge Base Tool.

http://www.speech.cs.cmu.edu/tools/lmtool-new.html