**EMITTER International Journal of Engineering Technology**
**Vol.2, No.2, December 2014**
**ISSN: 2443-1168**

# Indonesian Automatic Speech Recognition For Command Speech Controller Multimedia Player

**Vivien Arief Wardhany, Sritrusta Sukaridhoto, Amang Sudarsono**

Electronics Engineering Polytechnic Institute of Surabaya
Jl. Raya ITS – Kampus ITS Sukolilo Surabaya 601111, Indonesia
Telp : 6231 5947280 Fax : 6231 5946114
E-mail: vivienarief@pasca.student.pens.ac.id, dhoto@pens.ac.id, amang@pens.ac.id

**Abstract**

The purpose of multimedia devices development is controlling through voice. Nowdays voice that can be recognized only in English. To overcome the issue, then recognition using Indonesian language model and accousticc model and dictionary. Automatic Speech Recognizier is build using engine CMU Sphinx with modified english language to Indonesian Language database and XBMC used as the multimedia player. The experiment is using 10 volunteers testing items based on 7 commands. The volunteers is classifiedd by the genders, 5 Male & 5 female. 10 samples is taken in each command, continue with each volunteer perform 10 testing command. Each volunteer also have to try all 7 command that already provided. Based on percentage clarification table, the word "Kanan" had the most recognize with percentage 83% while "pilih" is the lowest one. The word which had the most wrong clarification is "kembali" with percentagee 67%, while the word "kanan" is the lowest one. From the result of Recognition Rate by male there are several command such as "Kembali", "Utama", "Atas " and "Bawah" has the low Recognition Rate. Especially for "kembali" cannot be recognized as the command in the female voices but in male voice that command has 4% of RR this is because the command doesn't have similar word in english near to "kembali" so the system unrecognize the command. Also for the command "Pilih" using the female voice has 80% of RR but for the male voice has only 4% of RR. This problem is mostly because of the different voice characteristic between adult male and female which male has lower voice frequencies (from 85 to 180 Hz) than woman (165 to 255 Hz).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. For further work needs to be done in order to improving the accouracy of the Indonesian Automatic Speech Recognition system

**Keywords**: Automatic Speech Recognizer, Indonesian Acoustic Model, CMU Sphinx, indonesian Language Model, Recognition Rate, XBMC.

## 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. [5]Most of the available means of information access like print media are useful only for the literate members of the society. Other modes like television and radio are non-interactive and computers, although being interactive are not suitable for the major portion of the society that is either not familiar with its interface and usage or does not have access to it at all. As we can see there is no Indonesian language can be recognize as command in multimedia player. One solution to this problem isbuild Indonesian Automatic Speech Recognition to utilize as interface for human-computer interaction.

## 2. RELATED WORKS

"Sphinx-4 Indonesian Isolated Digit speech recognition".

Sphinx-4 Indonesian Isolated Digit speech recognition, This study was applied Sphinx-4 for Indonesian digit speech recognition using JavaTM 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.[2]

"Indonesian Automatic Speech Recognition System Using English-Based Acoustic Model."

In this journal the author build an automatic speech recognizer (ASR) without providing the acoustic model, language model and lexicon, but build an Indonesian ASR without providing the Indonesian acoustic model directly. Instead, using English acoustic model and mapped English phoneme into Indonesian one. There are 39 English phonemes and 29 Indonesian phonemes. For special Indonesian phoneme with no corresponding English phoneme, tried to make estimation such as "ny" is mapped into "n" and "y". There are 9,509 Indonesian words equipped with corresponding English phoneme. The goal of this paper is to compare the system's accuracy with existing Indonesian ASR that use Indonesian acoustic model [3].

## 3. ORIGINALITY

The automatic speech recognition nowdays became popular as aplication human computer interaction we use the voice command to control the multimedia player. The command using indonesian languange which this this mean we create the acoustic and language model. We build the indonesian dictionary using the Imtools which provided by CMU Sphinx.

## 4. SYSTEM DESIGN

### 4.1 Automatic Speech Recognizer

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.

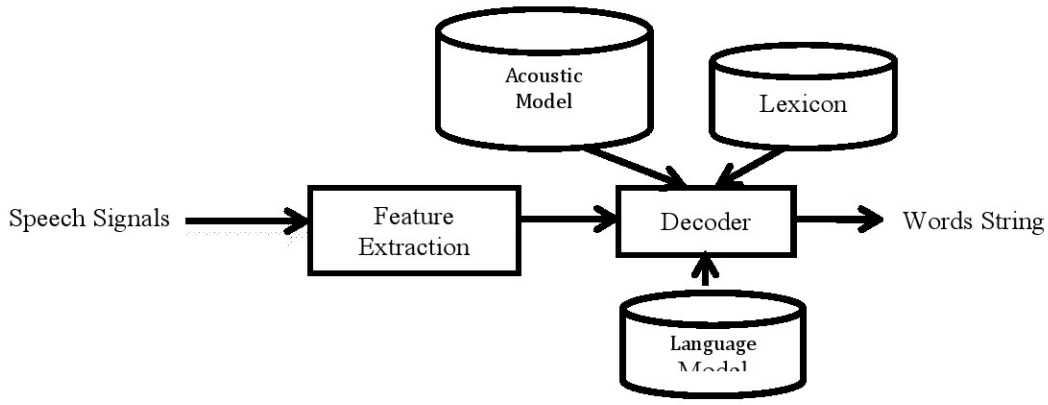


**Figure 1.** Diagram of Automatic Speech Recognition System

The basic standard speech recognition system consist of two major important part : The Front end and Decoder. The front end block extracts spectrum representation of the speech waveform. The most widely used features are Mel Frequency Cepstral Coefficients (MFCC). The decoder block searches the best match of word sequences for the input acoustic features based on acoustic model, lexicon, and language model.Lexicons and grammars may range from simple command and control and digit recognition of word vocabularies. [1]

Applications interface with the decoder to get recognition resultsthat may be used to adapt other components in the system. Acoustic models include therepresentation of knowledge about acoustics, phonetics, microphone and environment variability,gender and dialect differences among speakers, etc. Language models refer to a system'sknowledge of what constitutes a possible word, what words are likely to co-occur, andin what sequence. The semantics and functions related to an operation a user may wish toperform may also be necessary for the language model. Many uncertainties exist in these areas, associated with speaker characteristics, speech style and rate, recognition of basic speech segments, possible words, likely words, unknown words, grammatical variation, noise interference, nonnative accents, and confidence scoring of results.

The speech signal is processed in the signal processing module that extracts salient feature vectors for the decoder. The decoder uses both acoustic and language models to generate the word sequence that has the maximum posterior probability for the input feature vectors. It can also provide information needed for the adaptation component to modify either

the acoustic or language models so that improved performance can be obtained. [4]

### 4.2 Sphinx-4

The Sphinx4 speech recognition system is the latest addition to Carnegie Mellon University's repository of the Sphinx speech recognition systems. It has been jointly designed by Carnegie Mellon University, Sun Microsystems laboratories, Mitsubishi Electric Research Labs, and Hewlett-Packard's Cambridge Research Lab. Sphinx 4 is different from the earlier CMUSphinx systems in terms of modularity, flexibility and algorithmic aspects. It uses newer search strategies, is universal in its acceptance of various kinds of grammars and language models, types of acoustic models and feature streams.

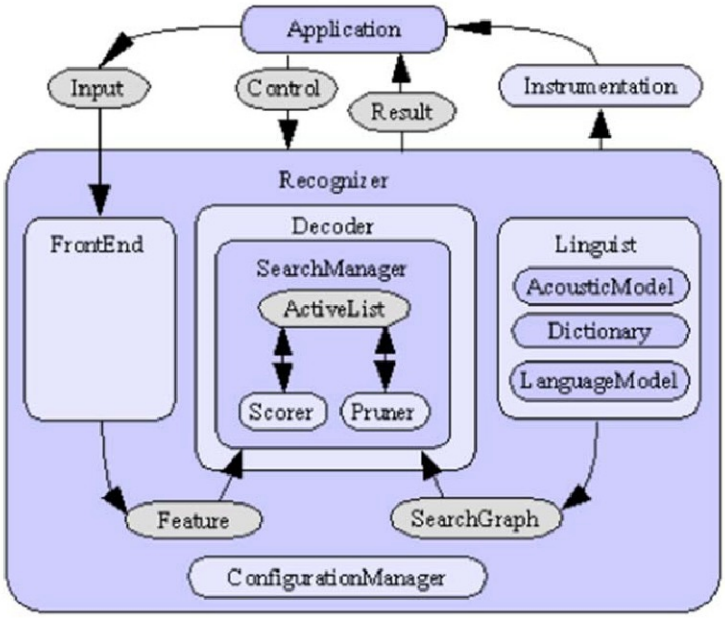


**Figure 2.** Sphinx Architecture

Figure 2 shows the architecture of the Sphinx system. Each of the element in the Figure 2 can be replaced to match the researcher's needs. And when we change a few part of the element system will not change about another feature. In this research wes change the language model and dictionary using the indonesian language. The front end block gets a single several input signals and computes them then it will created a sequence of features. Then the liguist generated searchgraph by translating the language model and lexicon that give helps about the information of pronounciation called Dictionary and acoustic model store set of the structural information. The Decoder block contain of Search manager that uses the feature as well as the search graph in order to realize the decoding, then result is produce. [2]

There are several step need to prepared the experiment first is preparing the data The trainer learns the parameters of the models of the

sound units using a set of sample speech signals, this is called Database. The database contains information required to extract statistics from the speech in form of the acoustic model.The trainer needs to be told which sound units that want to learn the parameters of, and at least the sequence in which they occur in every speech signal in your training database. This information is provided to the trainer through a file called the *transcript file.* The trainer then looks into a *dictionary* which maps every word to a sequence of sound units, to derive the sequence of sound units associated with each signal.After training, it's mandatory to run the decoder to check training results. The Decoder takes a model, tests part of the database and reference transcriptions and estimates the quality (WER) of the model. During the testing stage we use the *language model* with the description of the order of words in the language.The file structure for the database is:

**etc**

your_db.dic - Phonetic dictionary
your_db.phone - Phoneset file
your_db.lm.DMP - Language model
your_db.filler - List of fillers
your_db_train.fileids - List of files for training
your_db_train.transcription - Transcription for training
your_db_test.fileids - List of files for testing
your_db_test.transcription - Transcription for testing

**wav**

speaker_1
file_1.wav - Recording of speech utterance
speaker_2
file_2.wav

The second step is Setting up the training script. There are several task in this stage of setting up the training script first we need to Setup the format of database audio using .wav, then configure the model type and model parameter. Then continue to configure sound feature parameters, the default for sound files used in Sphinx is a rate of 16 thousand samples per second (16KHz). If this is the case, the etc/feat.params file will be automatically generated with the recommended values.And then configure decoding parameter.

The third step is Training the data, It will take a few minutes to train. On large databases, training could take a month. During the stages, the most important stage is the first one which checks that everything is configured correctly and your input data is consistent.for each training utterance, a sequence of 13-dimensional vectors (feature vectors) consisting of the Mel-frequency cepstral coefficients (MFCCs). And the last step is testing. In this stage we need to trained database in order to select best parameters, understand how application performs and optimize performance. To do that,

a test decoding step is needed. The decoding is now a last stage of the training process.

### 4.3 XBMC Multimedia Player

XBMC supports most common audio, video, and image formats, playlists, audio visualizations, slideshows, weather forecasts reporting, and third-party plugins. It is network-capable (internet and home network shares). XBMC includes full internationalization and localization support with translations to many different languages by default, with its language files translated to over 60 languages to date. XBMC source code is distributed as open source under the GNU General Public License (GPL), it is sponsored via the tax-exempt registered non-profit US organization, XBMC Foundation, and is developed by a global free software community of unpaid volunteers.

### 4.4 Shell Script with JSON-RPC

JSON-RPC is a stateless, light-weight remote procedure call (RPC) protocol. Primarily this specification defines several data structures and the rules around their processing. It is transport agnostic in that the concepts can be used within the same process, over sockets, over http, or in many various message passing environments.JSON-RPC works by sending a request to a server implementing this protocol. During the lifetime of a connection, peers may invoke methods provided by the other peer. To invoke a remote method, a request is sent. Unless the request is a notification it must be replied to with a response.

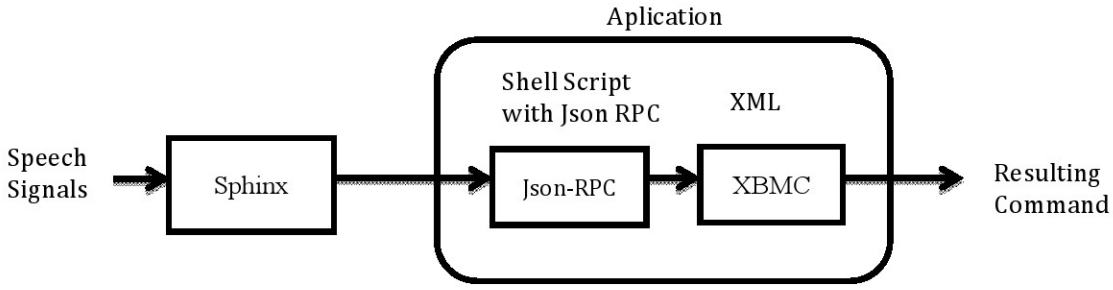


**Figure 3.** Indonesian ASR with Application

As we can see in the figure 3. Sphinx has function as the Automatic Speech recognition and the XBMC multimedia player using Json RPC to sends the request to XBMC Multimedia player. Nowdays modern shell script also provide many features which usually only can be found in language programming for complex purpose, for example construction control, variable, comment arraay, subroutine, etc. With this feature, it's possible to write sophisticated application as shell script, in this case JSON-RPC programming can be run through shell script & applicable as a command. The 7 commandd in Indonesian Bahasa like table1 was integrated into voice command which is speech to text become direct command to XBMC multimedia player.

## 5. EXPERIMENT AND ANALYSIS

There are several indonesian words use as commands in multimedia player. The7commands are Atas (Up), Bawah(Down), Kanan(Right), Kiri(Left), Pilih (Select), Utama(Home) and Kembali(back). Using the Imtools we generates the Dictionary of command speech. Then integrated the file into the Sphinx as language model and dictionary in cisampa format.

**Table 1.** Indonesian command versus XBMC commandin JSON-RPC script plus Dictionary in CISAMPA format

| Indonesian Command | Phoneme | XBMC Command Line |
|---|---|---|
| Pilih | P IH L IY | Input.Select |
| Kembali | K EH M B AH L IY | Input.Back |
| Kanan | K EY N AH N | Input.Right |
| Kiri | K IH R IY | Input.Left |
| Atas | AE T AH Z | Input.Up |
| Bawah | B AO AH | Input.Down |
| Utama | Y UW T AH M AH | Input.Home |

For the experimental intereset we use 5 male voices and 5 female voices, each male and female speak the 7 commands 10 times for each commands to interact with the system.

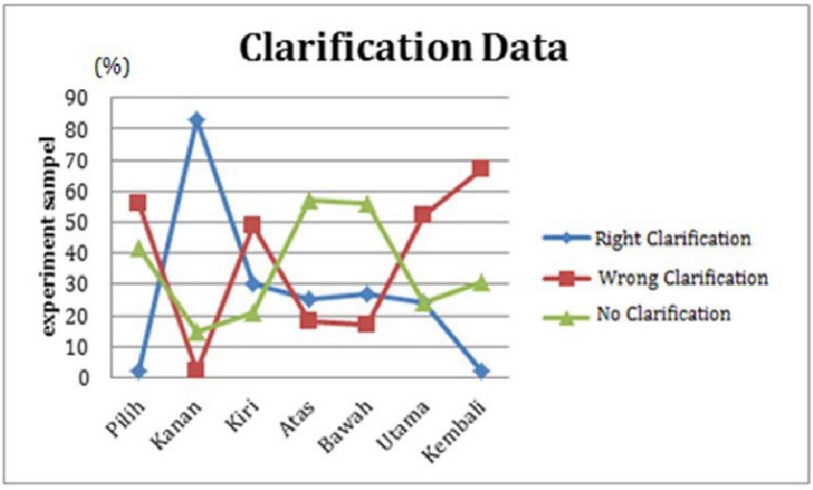


**Figure 4.** Clarification Graph

Based on the test result of Data clarification we can see the value of successful for each command. The high value of command is for command "kanan" has 83% of the correct command as "kanan". And the lowest procentage of succesfull command is "pilih and "kembali" has 2%. Based on the result of wrong clarification the command "kanan" is is the most emerge

command when the volunteers speak the other command, this mean the "kanan" command is the easiest command to recognize but when it became easy to recognize it caused interfence another command. For the command "pilih" and "kiri" which has the same vocal letter "i" when the command run together verry dificult to recognize because they have resemble pronounciation caused by the vocal "i", this case happend when the volunteers speak the command very quickly.

**Recognition Rate (RR)**

Resulting the testing of the experiment required five male (M) and FiveFemale(F) to speech into the system of each command. The recognition rate was gained based on :

$$RR = \frac{N_{Correct}}{N_{Total}} \times 100\% \quad (1)$$

Where RR is recognition rate, $N_{correct}$ is the correct recognition words and $N_{total}$ is the total words of the spoken command. And table 4 and 5 showed the result of the each commandsthat correctlty recognized.The test result from measurement quizioner which involved 5 men & 5 women which control the system using voice with 10 times command with the average speed for each command are pilih=0,4645 sec, kanan=0,5465 sec, kiri= 0,634 sec, atas=0,5195 sec, bawah=0,4905 sec, utama=0,5855 sec and kembali=0,407 sec. Which in the audiobook are recommended to be 150–160 words per minute, which is the range that people comfortably hear and vocalize words. The result from experiment is word clarification graph and recognition rate for each men & women.

**Table 4.**Recognition Rate using male voices

| | Pilih | Kanan | Kiri | Atas | Bawah | Utama | Kembali |
|---|---|---|---|---|---|---|---|
| M1 | 0 | 90 | 100 | 10 | 90 | 90 | 0 |
| M2 | 0 | 100 | 90 | 30 | 20 | 90 | 0 |
| M3 | 10 | 90 | 10 | 50 | 0 | 0 | 0 |
| M4 | 0 | 50 | 0 | 0 | 0 | 0 | 10 |
| M5 | 10 | 100 | 20 | 100 | 0 | 30 | 10 |
| Mean recognition rate (0%) | 4 | 86 | 44 | 38 | 22 | 42 | 4 |

From the result of Recognition Rate by male there are several command such as "Kembali", "Utama", "Atas " and "Bawah" has the low Recognition Rate. Especially for "kembali" cannot be recognized as the command in the female voices but in male voice that command has 4% of RR.

**Table 5.** Recognition Rate using Female voices

| | Pilih | Kanan | Kiri | Atas | Bawah | Utama | Kembali |
|---|---|---|---|---|---|---|---|
| F1 | 90 | 90 | 10 | 0 | 10 | 10 | 0 |
| F2 | 20 | 20 | 30 | 10 | 70 | 0 | 0 |
| F3 | 100 | 100 | 10 | 20 | 40 | 10 | 0 |
| F4 | 90 | 90 | 0 | 10 | 40 | 20 | 0 |
| F5 | 100 | 100 | 30 | 20 | 0 | 0 | 0 |
| Mean recognition rate (0%) | 80 | 80 | 16 | 12 | 32 | 8 | 0 |

Also for the command "Pilih" using the female voice has 80% of RR but for the male voice has only 4% of RR. This problem is mostly because of the different voice characteristic between adult male and femalewhich male has lower voice frequencies (from 85 to 180 Hz) than woman (165 to 255 Hz)and so for command "pilih" in cisampa format dictionary "P IH L IY" has two "i" vocal letter which "i" has the higher frequency than "e" and "a" in the word "kembali" which indictionary "K EH M B AH L IY" and so the female voice characteristic is higher than male that's why it become the most influence factor when we tested using the same hardware and the same environtment. Especially for the command "kembali" has the lowest value because of the deficiency of the acoustic model which is provide by CMU based on english not from the original indonesian speech corpus and the words kembali doesn't have the similar word in english, so it caused the command words unrecognized by the system.

## 6. CONCLUSION

This paper is presented our solution to create automatic speech recognizer to utilize as interfce for human-computer interaction. This automatic speech recognizer build using sphinx-4. 5 male and 5 female were selected in order to test the system using 7 command. especially for the command "kembali" had 4% for male Recognition rate , 0% for fermale Recognition rate and had 67% of wrong clarification as the other command. 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. For further work needs to be done in order to improving the accouracy of the Indonesian Automatic Speech Recognition system.

## REFERENCES

[1] Ehsani Farzad and Knodt Sehda, **Speech Technology in Computer-Aided Language Leraning: Strengtgs and Limitation of new Call Paradigm,** LLT Journal: Speech Technology in Computer-Aided Language Learning**,** Vol 2, No 1. Pp 54-73. 1998.

[2] IkaNovitaDewi, FahriFirdausillah, CaturSupriyanto, **Sphinx-4 Indonesian Isolated Digit Speech Recognition,** *Journal of Theoretical and Applied Information Technology,* Vol. 53 No.1, E-ISSN: 1817-3195**,** July 2013.

[3] V. Ferdiansyah, and A. Purwarianti, **Indonesian automatic speech recognition system using English-based acoustic model**, *American Journal of Signal Processing* 2(4): 60-63, 2012.

[4] Huang Xuedong, Acero alex, Hon Hsiao-Wuen. **Spoken Language** Processing**:a guide to theoty, algorithm, and system development**. Prentice Hall (New Jersey), Ed , pp 05-06, 2001.

[5] Raza, Agha Ali. **Design and Development of an Automatic Speech Recognition System for Urdu**. Thesis, FAST-National University of Computer and Emerging Sciences, Lahore Pakistan, 2009.