iSemantic 2017

by Andik Setyono

Submission date: 21-Jan-2020 10:42AM (UTC+0700)

Submission ID: 1244275609

File name: International_Conference_2_ISEMANTIC2017.pdf (271.27K)

Word count: 2829

Character count: 15584

Sphinx4 for Indonesian Continuous Speech Recognition System

Muljono, Askarya Qaulan Syadida, De Rosal Ignatius Moses Setiadi, Andik Setyono
Informatics Engineering Department
Dian Nuswantoro University
Semarang, Indonesia
muljono@dsn.dinus.ac.id, askaryaqaulan2@gmail.com, moses@dsn.dinus.ac.id, andik.setyono@dsn.dinus.ac.id

***Abstract*— Automatic Speech Recognition (ASR) is a technology which is capable to convert speech into text. Research in this field is growing very rapidly and is applied in multiple languages. Voice recognition of isolated word and connected word for Indonesian language has been done with various approaches. This is done to get better accuracy in recognition. However, there are still a few research about introduction of continuous speech for the Indonesian language. This paper describes the efforts made to build an Indonesian automatic speech recognition to recognize continuous speech using Sphinx4 (toolkit from CMUSphinx). There are three steps taken to build Indonesia ASR, those are preparing corpus, forming acoustic model and testing. The result of the acoustic model test which was formed showed the value of 23% word error rate and 32,8% sentence error rate. The lower the two variables, the better the introduction to the input given speech files.**

Keywords*— Indonesian Automatic Speech Recognition; acoustic model; CMUSphinx toolkit; sphinx4

I. Introduction

The advancement of technology demands ease in accessing information and delivering data. Answering the challenge, computer inputs in the form of speech is the answer to the problem. This can be realized by one condition. If the computer knows the speech of the language, the speech can be transcribed into a letter or sentence. The form of speech which is translated into mathematics will produce a probabilistic tone level because each result of tone has no limit for each unit of variables or between words. Based on these traits then the introduction of speech through a computer system will not produce 100% correct results.

In Indonesia, the development of research on Automatic Speech Recognition (ASR) has been started since 2003 until now where there are three forms of database, those are isolated digits, connected digits, limited datasets and simple dialogs [1]. Some researchers use a variety of approaches and produce mixed results. First, they use an English acoustic model to form an Indonesian speech corpus [2]. Both use the Indonesian Large Vocabulary Continuous Speech Recognition (LVCSR) development model with cross language approach [3]. Second, they use the limited dataset method which has the highest accuracy value of 86% [14]. Based on the approaches of previous researchers by using the transition from the English grammatical model or approach to generate the accuracy of Word Error Rate (WER) of great value to produce a better accuracy value, it required Indonesian database using the model and the overall approach to the order system of the corpus and the Indonesian acoustic model itself [1]. Broadly speaking to produce a good accuracy value in ASR implementation requires an approach to the rules of grammar Indonesian. This study aims to create Automatic Speech Recognition (ASR) Indonesian language by using continuous speech corpus.

II. Literatur Review

A. Automatic Speech Recognition

Some researchers propose the definition of ASR. According to Abushariah and Gunawan [8], ASR aims to create an intelligent system that can automatically translate a collection of word phonemes or phoneme strings from speech input signals. According to Hamdan Prakoso et al [14], the definition of Automatic Speech Recognition (ASR) is a technology that enables computing devices to convert human spoken words into computer readable text via microphone or telephone input. According to Anusuya and Katti [9], they define ASR as a process of converting speech signal into word order by implementing algorithm to a machine so that a system can develop and recognize speech input.

According to Kurian [10] there are advantages in applying ASR to several applications, such as public service assistance applications via telephone directory remarks, database query recognition applications, office recognition apps, speech auxiliary apps for operating activities in medicine, and translation applications automatic speech into Foreign Language.

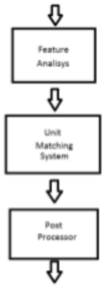

Figure 1 the ASR framework

Figure 1 explained the process of Feature Analysis which a representation of the signals explored to produce a variety of predictions. To read these signals, among others, it uses perceptual linear predictions (PLP) [11], PLP mixtures, and relative spectra (RASTA) [12], cepstrum root coefficients (RCC) [13], and Mel Cepstral (MFCC) frequency coefficients [16].

After the process of Feature Analysis, the results of predictions signals will be processed in the Unit Matching System which is known as the back-end of the ASR system. This module is responsible for recognizing features, and observed variables of speech signals by combining information and data obtained from acoustic models, language models, and lexicon.

The acoustic model or known as The Post Processor is a collection of files that describes the diversity of feature vectors [5]. In this model, the Hidden Markov Model (HMM) algorithm is used to create statistical signal models from ASR [4]. The Language Model is a set of probabilistic data formed from the sequence of transcript words [5]. This form of probabilistic can be a unigram or N-gram model. The unigram model is usually used in information retrieval. The N-gram model is used to estimate the length of word phrases or sentences and sequences which are not observed during training in the model [7]. The lexicon is often called the thesaurus, a language vocabulary consisting of all the words and expressions that will be processed to be displayed [5]. Lexicon can be interpreted as a phonetic dictionary.

B. *Continuous Speech Corpus*

Continuous Speech Corpus is a speech corpus that is recorded by pronounced naturally. The Corpus consists of words that are connected to form a sentence separated by a pause. The thing to note in forming continuous speech is the pronunciation corpus that is very dependent on the context between the words which compose the sentence. The problems encountered in establishing continuous speech are as follows:

1) Continuous Speech is unlike isolated digits and connected digits, while continuous speech has made the words overlap. As a result, the word limit becomes unclear and it becomes difficult to identify the starting point & endpoint of the word.

2) Co-articulation is a parameter that can make a natural sounding utterance, the effect of accent, and the characteristics of speech sounds clear [16]. The characteristic of this utterance can remove the limits of phonemic features, and it causes acoustic variability as the beginning & end of the spoken word in continuous speech. Hence, the errors that occur tend to be found in the recognition system.

3) Speech levels can indicate the level of speech and can impact the eloquence of the word pronunciation. The speech styles and the nature of the text to be spoken are two important properties of the speech levels context described by Mathew [15]. Two parameters based on these properties are as follows:

Mean speech rate, $\mu = f$ (Speaker)
Variance, $\sigma 2 = g$ (cognitive load)

The cognitive load can be said to be the textual nature that describes the work which can be done and the creativity required for the selection of the text to be spoken. Previous researchers found that with speech-level changes [17], and vowel duration showed more change than the consonants which were pronounced.

C. *Carnegie Mellon University Sphinx4.5-realpha*

Carnegie Mellon University (CMUSphinx) is a flexible library, consisting of pluggable modules to drive new innovations in speech recognition research with the core using the Hidden Markov Model [5] algorithm. Sphinx4 is the core framework of the modules provided by CMUSphinx. Sphinx4 provides researchers in the speech recognition field to develop a new language that has been in previous training to be implemented in a sophisticated manner [6].

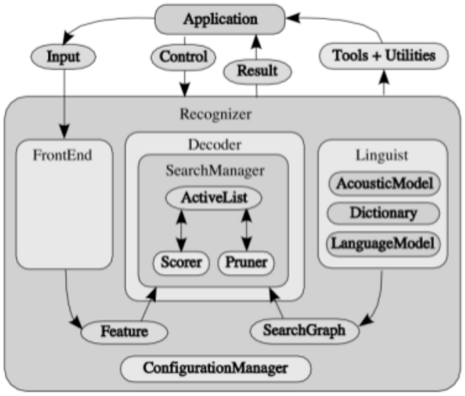

Figure 2 the Sphinx4 framework

In Fig 2, it is shown that there are three main modules in the Sphinx4, those are framework of the Front-End module, the Decoder module, and the Linguist module. The Front-End module works by taking one or more input and parameter signals into feature sequences [5]. The Linguist module works by translating all types of language models, along with pronunciation information from dictionaries and structural information from one or more sets of acoustic models, and the results can be obtained by Search-Graph. Search-Manager in Decoder works by using the result feature of the Front-End module and the Search-Graph results from the Linguist module to perform the actual decoding, the results obtained are the text or sentence (Speech to Text) [6].

III. Building Indonesian Speech Recognition using Sphinx4

Before building ASR, it takes some kind of data that plays a role in building, shaping, and training some of the models required as a system requirement step.

A. Data Preparation

a) Indonesian Phoneme

The construct ASR for Indonesian language requires a list of Indonesian phonemes used to construct transcripts of sentence text and corpus speech. The phoneme is the smallest unit of sound in the form of language. The list of phonemes we used consists of 33 phonemes, namely: a, ai, au, e, ei, ê, i, u, o, oi, b, c, d, f, g, h, j, k, l, m, n, p, r, s, sy, t, w, y, z, kh, ng, ny, sil. [4]

b) Corpus Text and Corpus Speech

This study uses 407 sentences as a text corpus and 4070 speech files as a continuous speech corpus. Each text transcription and a continuous speech corpus should be appropriate. Corpus continuous speech is recorded in Microsoft Waveform Audio Format (MS WAV) format, 8 KHz sampling rate, mono channel, 16 bits. The sentence is spoken by 10 Indonesian native speakers, consisting of 6 women and 4 men with different accents.

B. Create Acoustic Model

Required data on acoustic model training are file dictionary, phone, filler, binary language model, transcription and file aids. The results obtained are called speech training databases. This database contains the information needed to extract the probabilities of a trained recording into an acoustic model.

a) Dictionary file

Dictionary file serves as a mapping of the words (grapheme) into the phoneme sequence. Every word occurrence in sentence transcripts or spoken corpus must be in the dictionary.

b) Binary language model file

The Language model in binary is used to change the language model file to N-gram language form. The language model will limit the data testing that will appear during the ASR system implementation.

c) Transcription file

Transcription files containing spoken training expressions are arranged in sequence with capital letters, punctuation and tagging symbols for initialization and ending sentences and followed by filename of speech corpus.

d) Phoneme file

Phoneme file contains phoneme list with character and pronunciation.

e) Filler file

The filler file contains a filler phoneme which is not covered by the non-linguistic language model sounds like breath, hmm or laugh.

f) File aids file

File-aids contain the PATH file. To facilitate the training and reading of PATH, the file is placed in the same folder.

C. Testing Methods

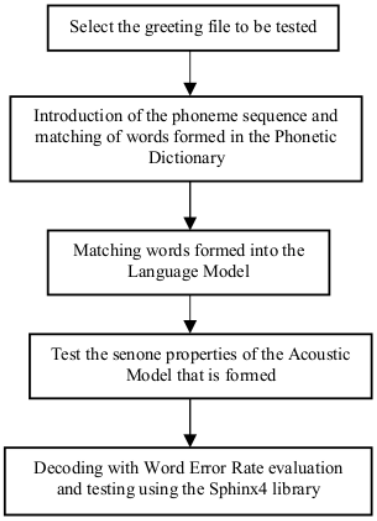
Select the greeting file to be tested
Introduction of the phoneme sequence and matching of words formed in the Phonetic Dictionary
Matching words formed into the Language Model
Test the senone properties of the Acoustic Model that is formed
Decoding with Word Error Rate evaluation and testing using the Sphinx4 library

Figure 3 testing methods

The test design of the proposed method is as follows:

1. The first step is to input in the form of speech files. The spoken word or sentence data is linear to the context of the established language model.
2. The second step is the introduction of the phonetic sequence of speech input. This process involves a dictionary of established models.
3. The third step is phonetic sequence matching into transcript of language model. This transcript acts as a dataset to be used in showing hypotheses of speech recognition.
4. The next step is to test the senone sequence of acoustic properties of speech in practice. This is done to match the phonemic nature of a word with a given input utterance.
5. The final step is the decoding of the acoustic model that is formed and evaluated based on the Word Error Rate (WER) value of the acoustic model encoding. After that, enter the testing phase of the model using the Sphinx library.

IV. EVALUATION

a) Training Results

When the sphinxtrain identification tool has been completed, the script will automatically calculate the value of Word Error Rate (WER) and Sentence Error Rate (SER). The lower the value of the two variables, the better the model is formed. In this study, using 407 sentences of transcription data, with a phonetic dictionary of approximately 8000 words, and a speech corpus recording for approximately 1 hour consisting of 4070 files, the researcher completed an acoustic model training for 1 hour 58 minutes, decoding stage.

The results of the WER and SER values obtained are as follows:

```
SENTENCE ERROR: 32,8% (20/61)
WORD ERROR RATE: 23,0% (92/400)
```

Calculation Word Error Rate (WER)

$$WER = \frac{(I + D + S)}{N} \times 100\,\%$$

information :

I = parameter of the inserted value
S = value parameter added
D = the parameter of the deficient value
N = number of word reference parameters

b) Implementation Results

Java archive application format can run on various operating systems either windows or linux.

The ASR application provides an option to enter the name of the test file. It is needed to make sure that this file is on one PATH with the ASR project file. Here's the test file transcript:

```
"david deda"
"dengan bujuk rayunya yang manis"
"obyek vital"
"atau di kenal juga sebagai konsep dwifungsi"
"saudara agus"
"obyek vital"
"dengan bujuk rayunya yang manis"
"kendati demikian, sulistyo enggan menjelaskan mengenai alasan
yang di maksud"
"meninjau lokasi pengungsian di waihaong kecamatan nusaniwe
kodya ambon"
"membanjirnya sapu plastik moceng hingga keset baik kain atau
karet"
"hal itu dapat di pertanggung jawabkan serta memenuhi
persyaratan rencana kerja syarat"
"syahril syabirin"
"elza syarif mengaku"
"operasi terhadap nyoman akan segera di lakukan"
"itu yang disebut badai fransisca"
"baik merek"
```

The program will provide the results of the hypothesis along with the available information in example: the estimated speed of speech recognition, memory usage, N-gram tree model formed, and information about live CMN or known Mell Cepstral frequency coefficients (MFCC). Here are the results of the hypothesis given:

```
"david deda"
"dengan bujuk rayunya yang manis"
"ejek vital"
"kau di kenal juga sebagai konsep"
"dwifungsi"
"saudara agus"
"ejek vital"
"dengan bujuk rayunya yang manis"
"kendati demikian"
"sulistio enggan menjelaskan mengenai alasan yang di maksud"
"meninjau lokasi pengungsian di waihaong kecamatan nusaniwe
kodya ambon"
"membanjirnya sapu plastik moceng hingga keset baik kain atau"
"karet"
"hal itu dapat di tentang jauh pekan serta memenuhi persyaratan
rencana kerja syarat"
"syahril syabirin"
"pansus syarif mengaku"
"pengungsi terhadap nyoman akan segera di lakukan"
"pun di sebut badai fransisca"
"baik merek"
```

V. Conclusion

Based on this research, it can be concluded as follows:

1. An Indonesian-based Automatic Speech Recognition can recognize speech in the form of a given recording file, a waveform audio format file with a large bit rate of 8000 Hz. Recording data processing uses toolkit from CMUSphinx, and application development uses Sphinx4 library.
2. After the process of testing the acoustic model is formed, the value of word error rate (WER) is 23.0% and sentence error rate (SER) is 32.8%. The lower the two variables, the better the result to the input of the given speech file.

References

[1] Suyanto, "An Indonesian Phonetically Balanced Sentence Set for Collecting Speech Database," *Jurnal Teknologi Industri,* vol. XI, no. No. 1, pp. 59-68, Januari 2007.

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

[3] Sakti S, Markov K, Nakamura S., "Rapid Development of Initial Indonesian Phonemebased Speech Recognition Using The Cross-Language Approach," *Proceeding of Oriental-COSCODA,* pp. 38-43, 2005.

[4] X X Li, Y Zhao, X Pi, "Audio-visual continuous speech recognition using a coupled hidden Markov mode," *Proceedings of the 7th International Conference on Spoken Language Processing,* pp. 213-216, 2002.

[5] CMU Sphinx, "cmusphinx.github.io," [Online]. Available: https://cmusphinx.github.io/wiki/tutorialconcepts/. [Diakses 17 05 2017].

[6] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition," *SUN MICROSYSTEMS INC.,* no. SMLI TR, p. 0811, 2004.

[7] S. Cook, "Speech Recognition HOWTO," faqs.org, 13 09 2000. [Online]. Available: http://www.faqs.org/docs/Linux-HOWTO/Speech-Recognition-HOWTO.html. [Diakses 20 05 2017].

[8] Abushariah, Gunawan, and Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models," *Proceedings of International Conference Computer and Communication Engineering (ICCCE),* pp. 1-5, 2010.

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

[10] C. Kurian, BalaKrishnan, "Speech recognition of Malayalam numbers," *Nature & Biologically Inspired Computing,* pp. 1475-1479, 2009.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America,* pp. 1738-1752, 1990.

[12]Hermansky, Morgan, A. Bayya, "RASTA-PLP speech analysis technique," *Proc. International Conference on Acoustics, Speech and Signal Processing,* 1992.

[13] Lockwood, P. Alexandre and P., "Root cepstral analysis: A unified view. Application to speech processing in car noise environments,"," *Speech Communication,*, pp. 277-288, 1993.

[14] Hamdan Prakoso, Ridi Ferdiana, Rudy Hartanto, "Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset," *International Symposium on Electronics and Smart Devices (ISESD),* pp. 283-286, 29-30 November, 2016.

[15] A. S. Mathew, Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition, Pittsburgh: Carnegie Mellon University, 1995.

[16] Dang, J., Honda, M., Honda, K, "Infestigation of Co-articulation in Continuous Speech of Japanese Acoustical Science and Technology Acoustical Society Japan," vol. Volume 25, no. No. 5, 2004.

[17] Wiqas Ghai, Navdeep Singh, "Continuous Speech Recognition for Punjabi Language," *International Journal of Computer Applications (0975 – 8887),* vol. Volume 72, pp. 23-28, May 2013.

iSemantic 2017

ORIGINALITY REPORT

17%	10%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	research.ijcaonline.org Internet Source	3%
2	Veri Ferdiansyah, Ayu Purwarianti. "Indonesian automatic speech recognition system using English-based acoustic model", Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, 2011 Publication	3%
3	Submitted to Universitas Dian Nuswantoro Student Paper	2%
4	www.ijritcc.org Internet Source	2%
5	Jefri Yushendri, Alvian Rahman Hanif, Anneke Annassia Putri Siswadi, Purnawarman Musa et al. "A speech intelligence conversation bot for interactive media information", 2017 Second International Conference on Informatics and Computing (ICIC), 2017 Publication	1%
	Andik Setyono, Dewa Arga Nisadha, De Rosal	

6	Ignatius Moses Setiadi. "Analysis of 4G Network and Chat Applications to Smartphone Battery Life", 2018 International Seminar on Application for Technology of Information and Communication, 2018 Publication	1%
7	staffweb.sjp.ac.lk Internet Source	1%
8	lodel.irevues.inist.fr Internet Source	1%
9	kursorjournal.org Internet Source	<1%
10	Submitted to University of Hyderabad, Hyderabad Student Paper	<1%
11	link.springer.com Internet Source	<1%
12	Submitted to VIT University Student Paper	<1%
13	intranet.cs.man.ac.uk Internet Source	<1%
14	Submitted to Manchester Metropolitan University Student Paper	<1%
15	Submitted to University of Edinburgh	

	Student Paper	<1%
16	digital.ub.uni-paderborn.de Internet Source	<1%
17	Submitted to The University of Manchester Student Paper	<1%
18	De Rosal Ignatius Moses Setiadi, Muhammad Fadhil, Eko Hari Rachmawanto, Christy Atika Sari et al. "Secure Reversible Data Hiding in the Medical Image using Histogram Shifting and RC4 Encryption", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		