American Journal of Signal Processing 2012, 2(4): 60-63
DOI: 10.5923/j.ajsp.20120204.01

Indonesian Automatic Speech Recognition System Using English-Based Acoustic Model

Veri Ferdiansyah*, Ayu Purwarianti

Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Bandung, 40132, Indonesia

Abstract Building an automatic speech recognizer (ASR) means that one has to provide the acoustic model, language model and lexicon for the intended language, which is also applied for Indonesian ASR. Unfortunately, providing acoustic model for a certain language is quite expensive, unlike the language model and the lexicon. This is because one has to record many utterances from several speakers to build a speaker independent ASR. In our research, we attempted to build an Indonesian ASR without providing the Indonesian acoustic model directly. Instead, we made use 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, we 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 English acoustic model size is 5,523 KB and the Indonesian language model is built from a file of 405 KB in size. By customizing Sphinx (a Hidden Markov Model based ASR tool) with Indonesian lexicon and Indonesian language model, the Indonesian ASR was built. The goal of this paper is to compare the system's accuracy with existing Indonesian ASR that use Indonesian acoustic model.

Keywords Indonesian Automatic Speech Recognition, English Acoustic Model, English-Indonesian Phoneme Mapping

1. Introduction

Speech is a natural means of communications. For two (or more) parties to communicate effectively, comprehension of a mutually understood language is required.

Speech is produced by the vocal tract. According to[1], the speech signal on which speech sound you articulate can be excited in three possible ways: voiced excitation, unvoiced excitation, and transient excitation.

Over the past five decades, automatic speech recognition and speech synthesis fields have attracted many researchers around the globe. Some of the earliest automatic speech recognition systems are Dragon[2] and Carnegie Mellon University's Harpy[3]. At that time, the Harpy system can recognize a speech from 1,011 words vocabulary with good accuracy.

To build an automatic speech recognizer (ASR), one has to provide the acoustic model, language model and lexicon for the intended language. Unfortunately, building an acoustic model for a certain language is quite expensive, unlike the language model and the lexicon. This is because one has to record many utterances from several speakers to build a speaker independent ASR.

Here, we attempted to build an Indonesian ASR without providing the Indonesian acoustic model directly. Instead, we made use an English acoustic model and mapped the English phoneme into the Indonesian one. There are 39 English phonemes and 29 Indonesian phonemes. For special Indonesian phonemes with no corresponding English phoneme, we made estimations (e.g "ny" is mapped into "n" and "y"). There are 9,509 Indonesian words equipped with corresponding English phoneme. The English acoustic model size is 5,523 KB and the Indonesian language model is built from a file of 405 KB in size.

By customizing Sphinx (a Hidden Markov Model based ASR tool) with Indonesian lexicon and Indonesian language model, the Indonesian ASR was built. The goal of this paper is to compare the system's accuracy with existing Indonesian ASR that use Indonesian acoustic model.

2. Related Work

Some researchers have built an Indonesian LVCSR such as[4] and[5]. Reference[4] has reached the accuracy of 80%, while[5] reached the accuracy of 92%. These two systems used the acoustic model from the Indonesian speech database.

Even though they both aimed to create an Indonesian LVCSR, each of them used a different language model. The differences of the LVCSRs can be seen in Table I.

* Corresponding author:
veri.ferdi@gmail.com (Veri Ferdiansyah)
Published online at http://journal.sapub.org/ajsp

Table 1. Differences Between LVCSRs in[4] and[5]

	LVCSR[4]	LVCSR[5]
Speech Corpus	20 speakers (11 males and 9 females), each of them was asked to read 328 sentences	Daily news task: 400 speakers (200 males and 200 females). Each speaker uttered 110 sentences Telephone application task: 400 speakers (200 males and 200 females). Each speaker uttered 100 sentences BTEC task: 42 speakers (20 males and 22 females). Each speaker uttered 510 BTEC sentences
Language Model	615,248 sentences	160,000 sentences
Dictionary	41,436 words	40,000 words
OOV Rate	1.7%	0.78%
Accuracy	80%	92.22%

3. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of capturing an acoustic signal by using a microphone, and then converting the signal to a set of words using computer programme[6]. A standard speech recognition system consists of three parts as shown in Figure 1.

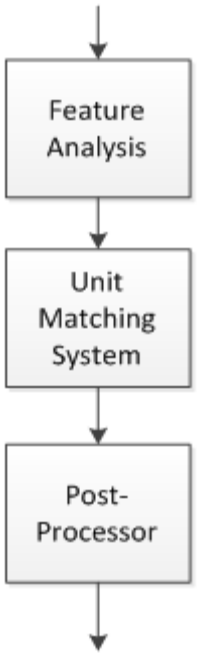

Figure 1. A speech recognition system

Feature analysis, also known as the front-end of an automatic speech recognition system, captures the speech signal and extracts the relevant feature from the captured signal. For the past few years, several feature representations have been explored such as perceptual linear prediction (PLP)[7], a combined PLP and relative spectra (RASTA)[8], root-cepstrum coefficients (RCC)[9], and Mel-frequency cepstral coefficients (MFCC)[10].

Unit matching system is the back-end of an automatic speech recognition system. This module is responsible for recognizing the observed feature of the speech signal by combining information from the acoustic model, the language model, and the lexicon.

Acoustic model is a collection of files that describes a variability of feature vectors[6]. A widely used acoustic model is a Gaussian mixture Hidden Markov Model (HMM). HMMs are used to create a statistical model of the speech signals.

Language model is a set of probabilistic data assigned to a sequence of words[11]. Estimating the probability of words sequence can be done by using unigram or N-gram models. Unigram models are usually used in information retrieval. N-gram models are used to approximate long phrases or sentences and sequences that are not observed during data training.

Lexicon is the language vocabulary consisting of words and expressions. Lexicon can also be called a thesaurus.

4. CMU Sphinx

CMU Sphinx, also called Sphinx in short, is an open source toolkit for speech recognition, created via a joint collaboration between the Sphinx group at Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), and Hewlett Packard (HP), with contributions from the University of California at Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT). Sphinx consists of a series of speech recognizers (Sphinx 2 – 4), speech recognizers for embedded systems (PocketSphinx), and an acoustic model trainer (SphinxTrain).

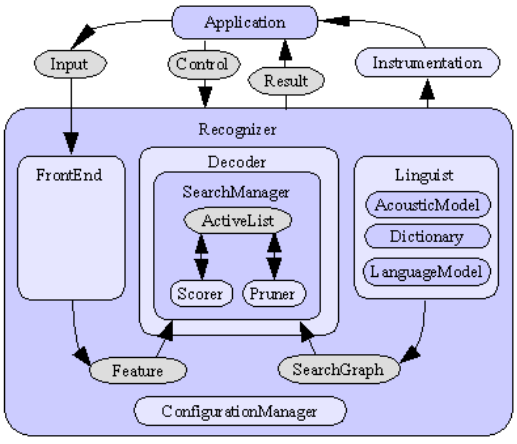

Figure 2. Sphinx architecture

Figure 2 shows the architecture of the Sphinx system. Each element in the Figure 2 can be replaced to match the researcher's needs. That means they can customize some features without worrying about the other features.

In this paper, we only change the dictionary and the language model in the Linguist block. We change the dictionary into a set of Indonesian words and then we create the language model based on that dictionary.

5. Experiments

5.1. Indonesian ASR Implementation

The language models that we create are based on Indonesian lexicon previously created by Hari Bagus Firdaus [12]. We modified the lexicon to incorporate additional phoneme translations. The results of the translations are shown in Table 2 and 3.

Table 2. List of the Phoneme Translations

Indonesian Phoneme	Translated Phoneme	Indonesian Phoneme	Translated Phoneme
/a/	/ah/	/o/	/ow/
/b/	/b/	/p/	/p/
/c/	/ch/	/q/	/k/
/d/	/d/	/r/	/r/
/E/	/ah/	/s/	/s/
/e/	/eh/	/t/	/t/
/f/	/f/	/u/	/uh/
/g/	/g/	/w/	/w/
/h/	/hh/	/y/	/y/
/i/	/ih/	/z/	/z/
/j/	/jh/	/ai/	/ah ih/
/k/	/k/	/kh/	/k hh/
/l/	/l/	/ny/	/n y/
/m/	/m/	/sy/	/s y/
/n/	/n/		

Table 3. A Section of the Indonesian Dictionary With Word, Original Base Form, And The After Translations

Indonesian Word	Phoneme Sequence Base Form	Phoneme Sequence After Translations
ABSOLUTISME	/a b s o l u t i s m E/	/ah b s ow l uh t ih s m ah/
AKHMAD	/a kh m a d/	/ah k hh m ah d/
ANEKA	/a n e k a/	/ah n eh k ah/
ANTARANYA	/a n t a r a n y a/	/ah n t ah r ah n y ah/
BERLANTAI	/b E r l a n t ai/	/b ah r l ah n t ah ih/
BERMASYARAKAT	/b E r m a sy a r a k a t/	/b ah r m ah s y ah r ah k ah t/

Table 2 shows the translated phoneme and Table 3 shows some words with their base phoneme and the after translations phoneme.

The phoneme translations are required in order for the Sphinx system to recognize Indonesian words. This technique is similar to those in[13]. Reference[13] translated English phonemes into Indonesian phoneme to increase the IR system accuracy. In contrast, we translated Indonesian phoneme into English ones. The translated phonemes are based on the phoneme that being used in the Sphinx. There are no specific rules that we used to translate the phoneme. The phoneme translations are based on the similarity of the pronunciation of each letter in a word. Figure 3 shows how input was processed in Sphinx system.

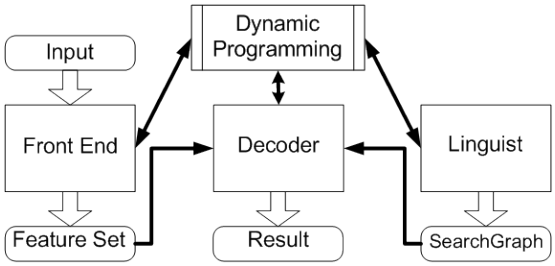

Figure 3. Input processing in Sphinx

5.1. Evaluation

Table 4, 5 and 6 shows the average accuracy of the ASR system that we built for 50, 100, and 200 words language model. We divide the accuracy measurement into 2 categories, discrete words and continuous words. Discrete word is a single spoken word and a continuous word is a phrase or sentence containing 2 or more words.

Table 4. ASR Accuracies For 50 Words Language Model

Speech Input	Word Accuracy (%)
Discrete Words	70%
Continuous Words	73%

Table 5. ASR Accuracies For 100 Words Language Model

Speech Input	Word Accuracy (%)
Discrete Words	63.33%
Continuous Words	49.32%

Table 6. ASR Accuracies For 200 Words Language Model

Speech Input	Word Accuracy (%)
Discrete Words	61.9%
Continuous Words	32.353%

The variations of the accuracy in words recognition by the ASR system can be affected by an inaccurate acoustic and language modelling and also a different pronunciation of words for each person. In this paper, the sources of the variations in the accuracy can also be affected by the wrong Indonesian-to-English phoneme translations.

Another factor that can affect the accuracy is noise. Sometimes, the noise captured by the speaker was translated into a word and affecting the accuracy. For example, some words like "saya" or "quo" were often added at the beginning or the end of the word pronounced even though we did not pronounce them (examples in Table 7).

Table 7. Word Insertion

Spoken Words	Recognized Words
universitas oxford	/quo universitas oxford/
xanana enggak emosi	/saya xanana enggak emosi/
kawasan reboisasi	/kawasan saya reboisasi/
nyonya janda kawasan	/nyonya janda quo kawasan saya/

Word substitution also took part even though the effects were less significant. Some examples of word substitutions

are “saya” substituted with “kaya” or “daya” and vice versa shown in Table 8.

Table 8. Word Substitutions

Spoken Words	Recognized Words
system qolbu	/system quo/
saya orang remaja	/kaya orang remaja/
jenazah janda kaya	/jenazah janda daya/

Word substitution in the system was more difficult to handle because the word pronunciation must be perfect. For a few continuous words such as “jenazah janda kaya”, the pronunciations of phoneme “da” in “janda” followed by “ka” in “kaya” was often blurred and created misrecognition.

6. Conclusions

This paper presented our solution to create a new automatic speech recognition system by using an existing acoustic model. The evaluation result of the system shows relatively low accuracy than existing ASR systems for Indonesian language such as[4]. Therefore, further work needs to be done in order to improve the accuracy of the Indonesian ASR system.

ACKNOWLEDGEMENTS

The authors would like to thank Aswin Juari for giving us the corpus needed in this paper.

REFERENCES

[1] B. Plannerer, An Introduction to Speech Recognition, Germany, 2005.

[2] J. K. Baker, “The Dragon system – an overview,” *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 23, pp. 24-29, Feb. 1975.

[3] B. Lowerre, “The Harpy speech understanding system,” Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 1976.

[4] D. P. Lestari and S. Furui, “A large vocabulary continuous speech recognition system for Indonesian language,” in *15th Indonesian Scientific Conference in Japan Proceeding*, p. 17-22.

[5] S. Sakti, et al, “Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project,” in *Proc. Technology and Corpora for Asia Pacific Speech Translation*.

[6] V. Stouten, *Robust Automatic Speech Recognition in Time-Varying Environments*, Kasteelpark Arenberg 10, B—3001 Leuven, Netherlands: Katholieke Universiteit Leuven, 2006.

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

[8] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, “RASTA-PLP speech analysis technique,” in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1992.

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

[10] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognitions in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing,* pp. 357-366, August. 1980.

[11] D. Hiemstra, “Language models,” *Encyclopedia of Database Systems*, 2009.

[12] H. B. Firdaus, “ASR Bahasa Indonesia dengan metode HMM untuk pendiktean digital secara real-time,” B. Eng. thesis, Institut Teknologi Bandung, Bandung, Indonesia, 2010.

[13] D. P. Lestari and S. Furui, “Adaptation to pronunciation variations in Indonesian spoken query-based information retrieval,” *IEICE Transactions on Information and Systems*, vol. E93-D, pp. 2388-2396, Sept. 2010.