論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Development of an Indonesian Large Vocabulary Continuous Speech Recognition System
著者(和文)	岩野 公司, 古井 貞熙
Authors(English)	Dessi Puji Lestari, Koji Iwano, Sadaoki Furui
出典(和文)	日本音響学会講演論文集, Vol. , No. , pp. 41-42
Citation(English)	, Vol. , No. , pp. 41-42
発行日 / Pub. date	2006, 9

Development of an Indonesian Large Vocabulary Continuous Speech Recognition System*

◎Dessi Puji Lestari, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

This paper presents our work in building an Indonesian LVCSR system. Most speech-related researchers in Indonesia focus on speech synthesizer technologies, natural language processing, and small vocabulary speech recognition systems [1]. The lack of Indonesian speech data is one of the main problems in building an Indonesian LVCSR system. Thus, we start our work by preparing a phonetically balanced text, recording speech data, preparing a text corpus, and then training acoustic models and language models.

2 Database Development

2.1 Speech Corpus

The Indonesian language called Bahasa Indonesia, is a variant of Malay language which is nowadays used in a broader area which includes Indonesia, Singapore, Brunei Darussalam, Malaysia, southern Thailand, southern Philippines, and several locations in South Africa. It was adopted as the Indonesian national language in 1928. However, the mother tongue of most Indonesian people is not Bahasa Indonesia, but rather local languages which they grew up speaking. According to the Indonesian National Survey 2000, there are 706 languages being used by different tribes of Indonesian people [2]. Thus, an ideal Indonesian speech corpus should cover not only all phones in Bahasa Indonesia, but also those all of other Indonesian dialects. However, due to the difficulties of collecting all Indonesian dialects, especially in Japan, we collected Bahasa Indonesia speech data from 20 Indonesian speakers (11 males and 9 females) from the five largest Indonesian tribes. These include Javanese, Sundanese, Madurese, Minang and Batak. Each speaker was asked to read 328 phonetically balanced sentences selected from the Information and Language System (ILPS) document collections [3]. Those document collections were taken from an Indonesian national newspaper and magazine. Speech was recorded in a quiet room on DAT tapes. Then, it was transferred to files at a 16 kHz sampling rate. After manual sentence segmentation was conducted, the total size of the speech corpus is 14.5 hours.

2.2 Text Corpus

Two document collections consisting respectively of newspaper and magazine articles collected by ILPS group were used for building the language model. All numbers appeared in the articles were changed into words, for example "103" to "seratus tiga", and all punctuation symbols were removed, except ",","."," ! " and "?" were changed into ".". Then, manual correction was conducted to split long sentences into several sentences or to merge two short grammatically incorrect sentences that appeared in the document into one correct grammatical sentence. The text corpus as described in Table 1 is obtained.

3 Experiments

Hidden Markov Model (HMM) - based acoustic models and n-gram language models were used to develop the LVCSR system for Indonesian Language.

3.1 Experimental Conditions

For conducting experiments, the speech corpus described in Subsection 2.1 was divided into training sets and testing sets. The leave-one-out method was implemented to conduct 10 experiments. For each experiment, the training set contained 18 speakers (10 males and 8 females), each uttering 293 sentences and the testing set contained 2 speakers (1 male and 1 female), each uttering 35 sentences. There was no overlap between speakers and sentences in the training set and the testing set. The 1st through 12th order Mel-Frequency Cepstral Coefficients (MFCC) were computed at every 10ms using a window with 25ms-width. Temporal differences of MFCC coefficients and energy were also incorporated. Context-dependent HMMs were trained using 32 Gaussian mixtures per state.

For training 2-gram and 3-gram language models, the training text as described in Subsection 2.2 was used. The 3-gram language model had a test-set perplexity of 86.6 and an OOV rate of 1.7%.

Table 1: Text Corpus Statistics

Attributes	Training Set
Number of Sentences	615,248
Number of Words	9,853,517
Vocabulary Sizes	129,919
Average Sent. Length (words)	16.02

* インドネシア語大語彙連続音声認識システムの構築
デッシ プジ レスタリ，岩野公司，古井貞熙（東工大）

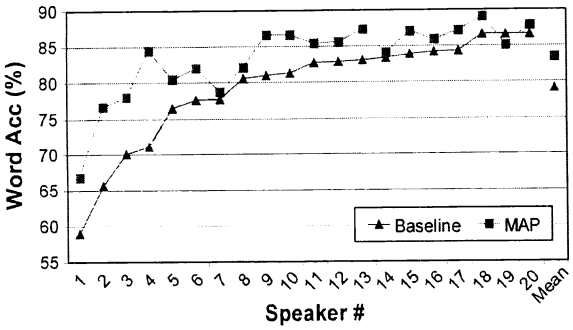

Figure 1: Evaluation Results

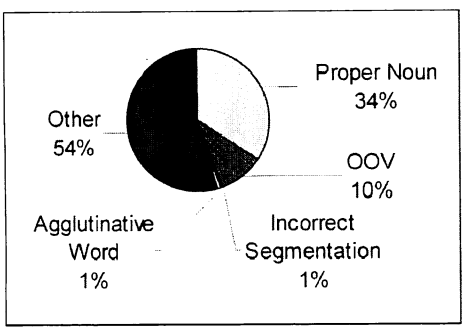

Figure 2: Error Statistics

For building the dictionary, 41,436 words that occur in the training set for more than 3 times were taken. An automatic transcription tool was made and employed to add pronunciation to this dictionary.

3.2 Evaluation

Using Julius [4] as speech decoder, the word accuracy for all 20 speakers and the average of all speakers can be seen in Figure 1, labeled as "Baseline". The result is sorted from the speaker with worst result to the best result. In order to improve the recognition result, 293 utterances of each speaker which were not taken for training and testing were used to conduct supervised adaptation by the MAP speaker adaptation technique [5]. This technique increased the accuracy by 4.1% on average. The results can be seen in Figure 1, labeled as "MAP".

The evaluation results of the system show relatively low accuracy. To investigate the problems, error analysis of the baseline results was conducted. The error statistics can be seen in Figure 2. The errors were caused by proper nouns, OOV words, incorrect word segmentation, for example the word "ke arah" was misrecognized as "kearah". Though not significant, the agglutinative properties of Bahasa Indonesia also raise the difficulty of recognition, for example the word "pe-nerima-an" was misrecognized as "me-nerima".

Since major errors were caused by proper nouns, the effect of proper nouns on the recognition results was investigated. We removed all proper nouns from the recognition results and calculate the word accuracy without counting proper nouns. For each speaker there were 68 proper nouns (17.2%). The testing set without proper nouns has word accuracy of 86.3% on average which is 7.1% higher than the average word accuracy for the baseline system. We assume that the difficulty in name recognition arises because there are many names which are acoustically confusable, for example, the word "budiono" which was misrecognized as "mudjiono" or "sutiono" and there are also mismatched pronunciations because some names are very difficult to pronounce, for example foreign names.

Some speakers have relatively lower accuracy than other speakers. Speaker #1 speaks slowly with short pauses between syllables. Speaker #2 makes various noises in the middle of utterances, like breath sounds and coughing. For Speakers #3 and #4, unclear pronunciation seems to be the reason why their recognition results do not compare favorable with other speakers. The effect of Indonesian dialect variability on the recognition results could not be determined in this experiment due to the lack of speech data.

4 Conclusions and Ongoing Work

We have described our work in developing an initial Indonesian LVCSR system. Evaluation results reveal that one of major sources of errors was proper nouns. Thus, proper nouns adaptation is now being implemented. Language model improvement is also needed in order to avoid the errors caused by incorrect word segmentations and agglutinative words.

Acknowledgements

The authors would like to thank ILPS, University of Amsterdam for giving us the Kompas and Majalah Tempo collections.

References

[1] S. Sakti, P. Hutagaol, A. A. Arman, and S.Nakamura, "Indonesian Speech Recognition for Hearing and Speaking Impaired People", Proc. Interspeech /ICSLP, 2, 1037-1040, 2004.

[2] B. K Purwo, "Pelestarian Bahasa Ibu Sebaiknya dari Keluarga", Kompas Cyber Media, 13 February 2003.

[3] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia", M.Sc. Thesis, the Information and Language System (ILPS) Group, Informatics Institute, University of Amsterdam, 2003.

[4] http://julius.sourceforge.jp/en/julius.html

[5] C. H. Lee and J. L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters", Proc. ICASSP, 2, 558-561, 1993.