論文 / 著書情報
Article / Book Information

論題(和文)	DNNに基づくインドネシア語音声認識システム
Title(English)	A DNN-Based ASR System for the Indonesian Language
著者(和文)	ホゥセン デヴィン, Price RyanWilliam, DESSIPUJI LESTARI, 篠田 浩一
Authors(English)	Devin Hoesen, Ryan Price, Puji Lestari Dessi, Koichi Shinoda
出典(和文)	日本音響学会2015年秋季研究発表会講演論文集, , , pp. 5-6
Citation(English)	Proc. ASJ 2015 Autumn Meeting, , , pp. 5-6
発行日 / Pub. date	2015, 9

A DNN-based ASR system for the Indonesian language *

Devin Hoesen[1], Ryan Price[2], Dessi Puji Lestari[1], and Koichi Shinoda[2]
[1]Institut Teknologi Bandung, [2]Tokyo Institute of Technology

1 Introduction

Indonesian is an Austronesian language, closely related to Malay [1]. It is the official language of Republic of Indonesia and also serves as a *lingua franca* in the Indonesian archipelago which has hundreds of regional languages. In Indonesian, almost every grapheme corresponds to one phoneme [6]. However, it is difficult to build an Indonesian phonetic dictionary because there are many accents (the effect of the regional languages) and much code-switching making the pronunciation of some words deviates from their standard.

The application of Hidden Markov Model combined with Gaussian Mixture Model (GMM-HMM) acoustic model (AM) in Indonesian ASR has been researched in [2, 7]. However, there has not been any research that builds the state-of-the-art Deep Neural Network HMM (DNN-HMM) AM [4] in Indonesian. The scarcity of Indonesian speech corpus (because of the difficulty to build the phonetic dictionary) is one of the problems in building the AM.

In this paper, we developed a DNN-HMM-based Indonesian speech recognizer. To build the DNN, we utilized the corpus from [2]. We also tried to overcome the speech corpus scarcity problem by incorporating English into the Indonesian DNN training in a form of Shared-Hidden-Layer DNN (SHL-DNN) [5]. The resulting DNN had two separate softmax layers for Indonesian and English. The English softmax layer could then be discarded because it was not needed for the decoding phase. The evaluation showed that the recognition accuracy of the DNN trained using only the Indonesian corpus outperformed the baseline GMM-HMM and it was improved further by using the SHL-DNN.

2 Related Works

2.1 DNN

A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs [4]. Each hidden unit maps the total input from the hidden layer below it to a scalar value that is forwarded to the layer above using a logistic (sigmoid) function. For a phoneme classification, which is a multiclass classification, an output unit of the last layer maps its total input into a class probability using the "softmax" function. The DNN can be generatively pretrained in a form of Deep Belief Network (DBN) to improve the recognition accuracy.

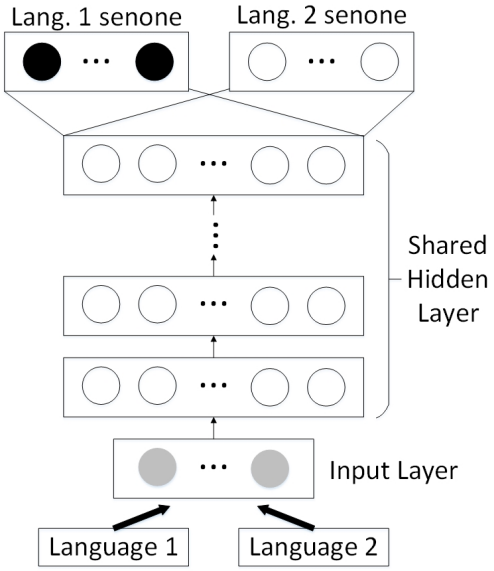

Fig. 1 Example of a SHL-DNN for multi-language learning

2.2 SHL-DNN

The idea of sharing the DNN's hidden layer across many languages was explored in [5]. Figure 1 illustrates the SHL-DNN design. The SHL-DNN was trained using the target language and one/some other language(s). It was shown that this method could reduce the WER of the target language even if the duration of the training data for the target language was not more than 10 hours. That is the major advantage of using the shared hidden layer, i.e. the low-resource language can benefit from the higher-resource language.

---

*DNN に基づくインドネシア語音声認識システム, デヴィン・ホゥセン (バンドン工科大学), ライアン・プライス (東京工業大学), デッシ・プジ・レスタリ (バンドン工科大学), 篠田浩一 (東京工業大学)

3 Experiments

3.1 Training Data Description

We utilized the Indonesian speech corpus developed in [2]. There were 20 native Indonesian speakers (11 males and 9 females) in the corpus, each read approximately 343 sentences taken from some Indonesian newspapers and magazines. Every recording was recorded in a clean environment, monaural, and had a sampling rate of 16kHz.

The corpus was then separated into training, development, and test set. The first 270 sentences spoken by 10 speakers (6 males and 4 females) were chosen as the training set (for a total of approximately 5.5 hours of speech). The next 30 sentences spoken by the same 10 speakers were the development set and the last 43 sentences spoken by the other 10 speakers were the test set.

For the language model (LM), we built a 3-gram LM using the text corpus from [2] which consisted of approximately 615,000 sentences. The vocabulary for building the LM consisted of all the words that appeared in the dictionary from [3] (this dictionary was also utilized as the speech dictionary).

Table 1 The WER (%) of the GMM-HMM and the DNN AMs

	Dev	Test
Speaker-adapted GMM-HMM (LDA + MLLT + SAT)	6.05	7.21
Indonesian-only DNN (CE)	1.41	5.20
Indonesian-only DNN (sMBR)	1.44	5.05
Shared-Hidden-Layer DNN	1.11	3.15

3.2 DNN Trained Using only the Indonesian Corpus

We first trained a GMM-HMM to be used as a baseline. The triphone GMM-HMM was trained using 36-dimensional MFCC. Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transformation (MLLT), and Speaker-Adaptive Training (SAT) were further applied to the GMM-HMM to further increase the recognition accuracy.

The DNN AM had 5 hidden layers of 2,048 hidden units with logistic activation function. The input to the network was 11 adjacent frames of 40-dimensional mel filterbank features. The output layer had 1,944 units derived from the context-dependent (CD) states of the baseline GMM-HMM. The DNN was first generatively pre-trained. Cross-Entropy (CE) training and 1 epoch of sequence-discriminative training in a form of state-level Minimum Bayes Risk (sMBR) were applied to train the whole DNN. We utilized the development set for early stopping of the DNN training. The evaluation results for both AMs are shown in Table 1.

3.3 SHL-DNN

The SHL-DNN was trained using the Indonesian and the English WSJ SI-284 (Wall Street Journal) corpus. The WSJ corpus was also a read speech, recorded in a clean environment, and its sentences were taken from newspapers. The SHL-DNN configuration and experiment was similar to the previous DNN. The difference was the output layer initially had 5,311 units (1,944 for Indonesian and 3,367 for English), but the English output units were then discarded because they were not needed for the decoding. The evaluation result in Table 1 shows that WER on the test set was reduced from 5.05% (the sMBR DNN) to 3.15% (37.6% relative reduction).

4 Conclusion

We have developed a DNN trained using only the Indonesian speech corpus. The DNN achieved 5.05% WER. Incorporating English into the low-resource Indonesian in the DNN training process increased the recognition accuracy for Indonesian. The SHL-DNN outperformed the DNN trained using only the Indonesian corpus by achieving only 3.15% WER.

References

[1] C.D. Soderberg and K.S. Olson, Journal of the IPA, 38(2), 209-213, 2008.

[2] D.P. Lestari, K. Iwano, and S. Furui, Proc. ISA Japan, 17-22, 2006.

[3] D. Hoesen and D.P. Lestari, Proc. ICEECS, 22-26, 2014.

[4] G. Hinton *et al.*, IEEE Signal Processing Magazine, 29(6), 82-97, 2012.

[5] J.T. Huang *et al.*, Proc. ICASSP, 7304-7308, 2013.

[6] M.J. Yap *et al.*, Behavior Research Method, 42(4), 992-1003, 2010.

[7] S. Sakti *et al.*, Proc. ASJ Autumn Meeting, 47-48, 2007.