Deep convolutional neural networks-based features for Indonesian large vocabulary speech recognition

Hilman F. Pardede[1,3], Purwoko Adhi[2], Vicky Zilvan[1], Ade Ramdan[1], Dikdik Krisnandi[1]
[1]Research Center for Data and Information Sciences, National Research and Innovation Agency, Bandung, Indonesia
[2]Research Center for Telecommunication, National Research and Innovation Agency, Bandung, Indonesia
[3]Graduate School for Computer Science, Nusa Mandiri University, Jakarta, Indonesia

Article Info

Article history:

Received Apr 7, 2022
Revised Sep 22, 2022
Accepted Oct 22, 2022

Keywords:

Convolutional neural networks
Deep learning
Indonesia speech recognition
Large vocabulary speech recognition
Low resource speech

ABSTRACT

There are great interests in developing speech recognition using deep learning technologies due to their capability to model the complexity of pronunciations, syntax, and language rules of speech data better than the traditional hidden Markov model (HMM) do. But, the availability of large amount of data is necessary for deep learning-based speech recognition to be effective. While this is not a problem for mainstream languages such as English or Chinese, this is not the case for non-mainstream languages such as Indonesian. To overcome this limitation, we present deep features based on convolutional neural networks (CNN) for Indonesian large vocabulary continuous speech recognition in this paper. The CNN is trained discriminatively which is different from usual deep learning implementations where the networks are trained generatively. Our evaluations show that the proposed method on Indonesian speech data achieves 7.26% and 9.01% error reduction rates over the state-of-the-art deep belief networks-deep neural networks (DBN-DNN) for large vocabulary continuous speech recognition (LVCSR), with Mel frequency cepstral coefficients (MFCC) and filterbank (FBANK) used as features, respectively. An error reduction rate of 6.13% is achieved compared to CNN-DNN with generative training.

Corresponding Author:

Hilman F. Pardede
Research Center for Data and Information Sciences, National Research and Innovation Agency
Jl Sangkuriang, Bandung, Indonesia
Email: hilm003@brin.go.id

1. INTRODUCTION

Speech recognition, a system that converts speech signals into text, plays a very important role for spoken understanding/dialogue systems. Some of their implementations are as the man-machine interface for instance in autonomous vehicles [1], home automation [2] and smart-office [3]. The complexity of speech recognition tasks increase with the size of the vocabulary used in the systems, style of speaking (reading or spontaneous speech), or the presence of environmental distortions [4] such as noise or reverberations. These research areas are still very active.

With the emergence of deep learning technologies, there has been great progress on large vocabulary continuous speech recognition (LVCSR), i.e. speech recognition for large size of words vocabulary where the words are spoken in continuous manner. Deep learning technologies could learn the complexity of words pronounciation, syntax and language rules from data better than conventional methods such as hidden Markov model-Gaussian mixture models (HMM-GMM). Various deep learning architectures have been proposed for LVCSR. Usually, they are used to replace GMM to estimate posterior probability while determining the states of the HMM. Therefore, the networks are usually trained generatively to model the distributions of the speech units. In earlier deep learning implementations, hybrid deep belief networks and deep neural networks (DBN-DNN) are used. In these implementations, DBNs which is built with restricted Boltzmann machine (RBM) are trained generatively, layer by layer, before several layers of DNN are stacked on top of the DBNs, where backpropagation is applied to fine-tune the weights of the whole networks. In later studies, various deep learning architectures such as recurrent neural networks (RNN) [5], long short-term memory (LSTM) [6], convolutional neural networks (CNN) [7], and time delay neural networks (TDNN) [8] were proposed. Most of these networks are trained generatively. Deep learning requires more data to train than HMM-GMM does. The availablity of large amount of data for the target languages becomes one of the driving force to develop good LVCSR systems [9]. This is usually not a problem for many mainstream languages such as English [10] and Chinese [11] or languages spoken in developed countries such as French [12] and German [13]. But, this is not often the case for low resources languages such as Indonesian *(Bahasa Indonesia)*.

Indonesian is the official language in Indonesia and spoken by more than 250 million people. While most Indonesian people also use ethnic language in daily communications, they speak Indonesian language with various proficiency level, since they use it in formal communications. There have been several studies to develop Indonesian LVCSR. Two early studies of Indonesian LVCSR are reported in [14], [15]. Both of them used HMM-GMM systems. In the first, around 14 hours of speech data are used for training, while in the latter, more than 100 hours of speech data are used. However, only the data from the first study are available. As a consequence, only few later studies develop Indonesian LVCSR systems. Usually, each of the studies develops its own datasets and doesn’t make them publicly available for further studies. For instance, around 33 hours of speech data are used to develop end-to-end LVCSR using DeepSpeech [16]. Smaller datasets are developed in [17]. Small vocabulary systems such as digit and command recognition systems are proposed in [18] and [19] respectively. Unavailability of the data publicly makes the results neither reproducible nor comparable. In addition, HMM-GMM systems are also still dominant methods for small data [20].

Some approaches have been proposed for LVCSR to deal with limited data. Adding artificially generated data to the original is a simpler approach. This could be done by perturbing the original signals. In [21], the speed of the speech signals are perturbed, by making them slightly slower or faster than the original. Perturbing the vocal tract length are conducted by Jaitly and Hinton [22] to create additional data for training. Another approach is done by transferring models that have been trained with large data and tuning them to the lower resource speech data [23]. However, these methods require close relations or similarities between the transferred model and the target languages. Other approaches are using acoustic models that thrive when only small data are used for training [24]. These approaches could be done by designing networks with fewer parameters to train.

Past studies indicate DNN may not be the best option for LVCSR when limited data are used, due to their large number of parameters. In the past, other alternatives have been proposed instead of DNN. Miao *et al.* [24], used deep maxout networks (DMN) instead. CNN [25] as alternative also shows better performance for limited data scenarios. In general, using networks with less number of parameters is better when only small amount of data are available for training, since large number of parameters of DBN-DNN may prone to overfitting when only small number of data are used. In these studies, these architectures are used as acoustic models in LVCSR systems which are trained generatively. CNN has been employed for other applications such as driver conditions [26], and image segmentation [27].

In this paper, we propose the use of CNN for feature learning to deal with limited number of data. CNN has much smaller number of parameters than multi-layer perceptron (MLP) and autoencoder do, and hence, may be more suitable when only small data are available for training. We employ speed perturbation on data and then pass them to two layers of CNN that are trained discriminatively. Evaluations of the proposed method on Indonesian speech data achieves 7.26% error reduction rate over the DBN-DNN systems using mel frequency cepstral coefficients (MFCC) as features and 9.01% error reduction rate over the DBN-DNN systems using filterbank (FBANK) as features.

The remainder of the paper is organized as follow: in section 2, we biefly explain the basic of speech recognition. We explain our proposed method in section 3. We describe our experimental setup in section 4, then discuss and analyze our results in section 5. Section 6 concludes our paper.

2. THE PROPOSED METHOD

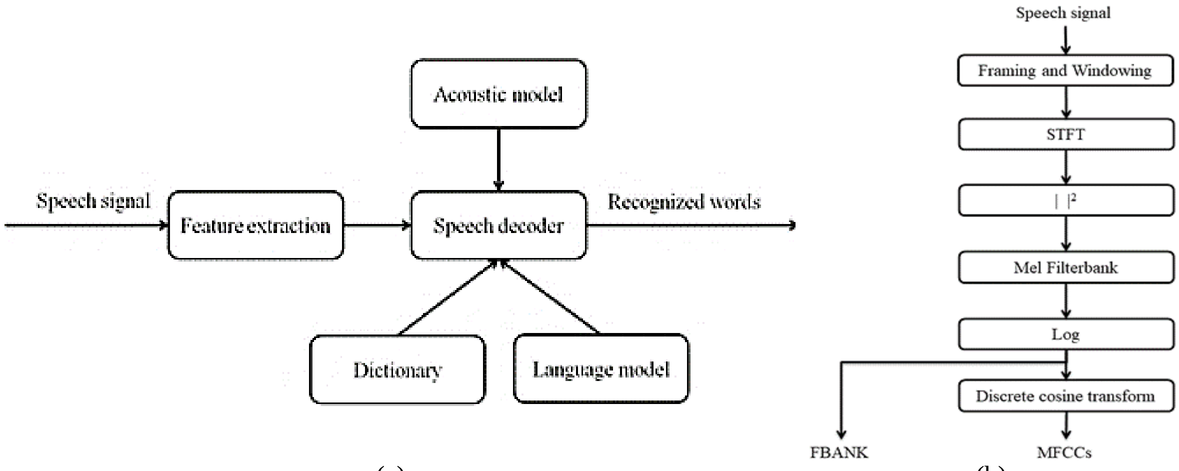
The objective of a speech recognition system is to convert a speech signal into a sequence of phonemes, syllabus, or words. The basic structure of a speech recognition system is shown in Figure 1(a). Speech signals are transformed into speech features and then, most likely speech units, such as words or p honemes, correspond to the features are computed and determined by a decoder, based on the trained acoustic and language models. Both acoustic and language models are trained using large amount of speech and text data, respectively.

For features, Mel frequency ceptral coefficients (MFCCs) are some of the most commonly used features for automatic speech recognition (ASR). The process for extracting MFCCs are shown in Figure 1(b). It comprises of the following steps. First, speech waveform is chunked at certain length, typically around 25 ms and then windowing functions, such as Hamming windowing, are applied to each chunk of speech. After that, the power spectral components from each chunk are extracted using short-time Fourier transform (STFT) and squared of the magnitude of STFT spectra. Then, mel filterbanks are applied to give the power spectra more emphasis on low frequency. After that, the log operation is applied before taking the discrete cosine transform (DCT) to produce MFCC. Typically, only the 13 first dimensions of MFCCs are used with their first and second derivatives.

Currently, there are increasing interests in using more "raw" features such as filterbank (FBANK) instead [28], when deep learning is used [29]. The reason is, while MFCCs are good to create uncorrelated speech features, much correlation between speech components may be lost in the process. Meanwhile, FBANK may still contain local correlations between the speech components which could be modeled by deep learning. FBANK is extracted in similar way to MFCC, but without DCT as shown in Figure 1(b).

Hidden Markov model (HMM) is traditionally used for the acoustic model [30]. Each speech unit is modeled with an HMM with certain number of states, typically between 3 and 5 states for a phoneme or tiphone or between 10 and 20 states for a word. Each state of HMM is modeled with Gaussian mixture models (GMM), where the observation probability is calculated, given speech features. This system is often called HMM-GMM.

Currently, with the emergence of deep learning technologies, much effort has been done to employ them for speech recognition. One implementation of deep learning for speech recognition is by replacing GMM with deep belief networks (DBN) to compute the observation probability [31]. The block diagram of the use of DBN for ASR is shown in Figure 2. Given speech features, each layer of DBN is trained in stacked using restricted boltzmann machine (RBM). Each layer of DBN is trained separately, one layer at a time. This training method is called generative pre-training. After training several layers of DBN, deep neural networks (DNNs) are trained on top of them with soft-max activations. The output of DNN is then used as observation probability to determine the state of the HMM models. This system is often called DBN-DNN system.



(a) (b)

Figure 1. A typical structure of ASR and the feature extraction process, (a) A block diagram of an ASR system and (b) the process of feature extraction for FBANK and MFCC

Motivated by the success of DBN-DNN systems, great efforts have been done to use various deep learning architectures for speech recognition other than DBN, for examples RNN [5], CNN [32], Long-short term memory (LSTM) [6], and gated recurrent unit (GRU) [33]. In addition, for acoustic models, many studies have applied deep learning as feature learning [34]. Because of deep learning ability to model nonlinear relations between speech components, allowing it to automatically learn useful informations from data directly [35]. For language model, N-gram is the most commonly used method [36]. N-gram computes probability N sequences of words occured from text corpus.

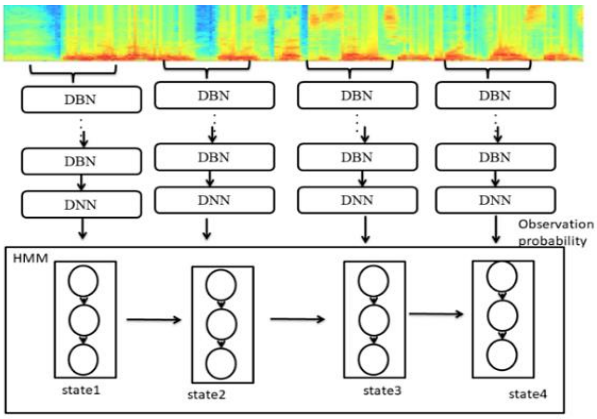
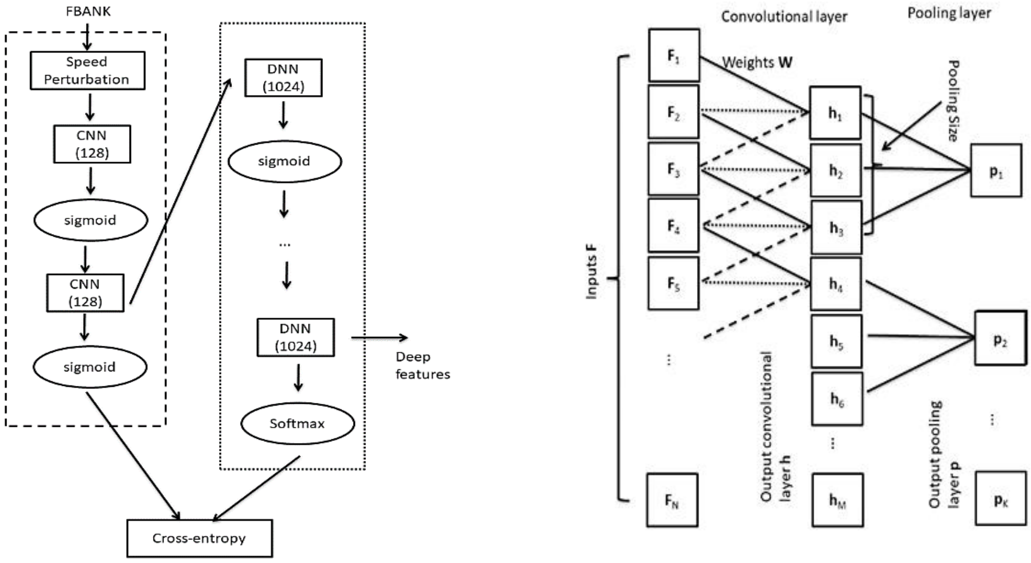


Figure 2. The use of DBN-DNN for speech recognition

In this paper, we propose deep features for ASR using CNN. The architecture of the proposed features is shown in Figure 3. In Figure 3(a), the block diagram of the system is explained. First, speed perturbation is applied to training data, with perturbation factors of 0.9 and 1.1. This factor is found effective in previous study to improve the performance of speech recognition [21]. The detail architecture of CNN is shown in Figure 3(b). The architecture for the deep features consists of two layers of CNN followed by several DNN layers. We vary the number of the DNN layers from two to ten to find the optimum number of layers. For both CNN and DNN, we apply sigmoid as activation function. The numbers of output nodes are set to 128 for CNN and 1024 for DNN.



(a) (b)

Figure 3. The proposed method; (a) The block diagram of the structure of the proposed CNN-DNN system, and (b) detailed archicture of a convolutional layer of a CNN and a max-pooling layer

Each CNN layer is built with convolutional and max-pooling layers as illustrated in Figure 3. This is a typical structure for CNN. As shown in Figure 4, the main difference between a CNN and an MLP or a fully connected network as in DNN is the conection between the input nodes. In a fully connected network, each activation $\mathbf{h}_m$ is connected to the entire input **F** and is computed by applying weighted sum between **W** and **F**. In a convolutional layer, there are only small number of local input (for instance [$\mathbf{F}_1$; $\mathbf{F}_2$; $\mathbf{F}_3$] for $\mathbf{h}_1$). By doing so, the number of shared parameter **W** would be reduced. After computing **h** in the convolutional layer, maxpooling is applied to local **h** input (for instance [$\mathbf{h}_1$; $\mathbf{h}_2$; $\mathbf{h}_3$] for $\mathbf{p}_1$). In max pooling, the maximum value of the local **h** is passed while the others are dropped. This is effective to reduce feature variability that may exist due to difference in speaking style and noise. In this paper, we set 3 as the pooling size.

To train the networks, discriminative training recipe as reported in [35] is performed. No pretraining as in DBN-DNN system are conducted. For each iteration of training, the cross-entropy losses are computed on a held-out data, which are randomly selected from 10% of training data. The stopping condition for training is when there is no longer significant reduction of the loss from previous iterations.

3. RESULTS AND DISCUSSION

3.1. Setup

We use Indonesian speech dataset as reported in [14] to build the acoustic model. Called Tokyo Institute of Technology Multilingual speech corpus for Indonesian language (TITML-IDN) dataset, it is freely available from speech resources consortium-National Institute of informatics (SRC-NII) for research purpose. It comprises of recordings from 20 Indonesian speakers (11 males and 9 females) where each speaker read up to 343 phonetically balanced sentences. It needs to be noted that several speakers do not have complete recordings of 343 sentences. For training, we use recordings of 16 speakers, while the rest are used for testing. We use recordings of indexes 1 to 293 of the sentences for each speaker for training and the recordings of indexes 294-343 for testings.

We compare the proposed method with conventional HMM-GMM systems and deep learning based system. For HMM-GMM system, HMM is used to model monophones (mono) and triphones. Then, we also apply three types of feature transformations. They are linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) (notated as LDA + MLLT), LDA + MLLT with speaker adaptation transformation (SAT), and maximum mutual information (MMI). For reference methods, DBN-DNN, we apply the system reported in [31]. Typically, four layers DBN with layers of DNN are used in literature. we vary DBN layers from two to ten to find the optimum settings for TITML-IDN. We use two layers of DNN afterwards. We also develop CNN-DNN systems describe in [32]. FBANK features are used as inputs for CNN-DNN. CNN-DNN comprises of CNN layers followed by DBN-DNN system. Two layers of DBN followed by two layers of DNN are usually used.

For the features, MFCC and FBANK are used. For MFCC, 13 dimensions with first and second derivatives are used to make up 39 dimensions. For FBANK, 40 dimensions of FBANK with first and second derivatives are used to obtain 120 dimensions of features. For the language model, a trigram (N-gram with N=3) is trained using Stanford Research Institute language model (SRILM) Language modelling Toolkit. The text corpora to train the language model are taken from Tala [14] and set of Leipzig corpora for Indonesian which are taken from Wikipedia data [37]. We evaluate the performance of our LVCSR with word error rate (WER) and for comparison between LVCSR systems, we use error reduction rate (ERR).

3.2. Results and analysis

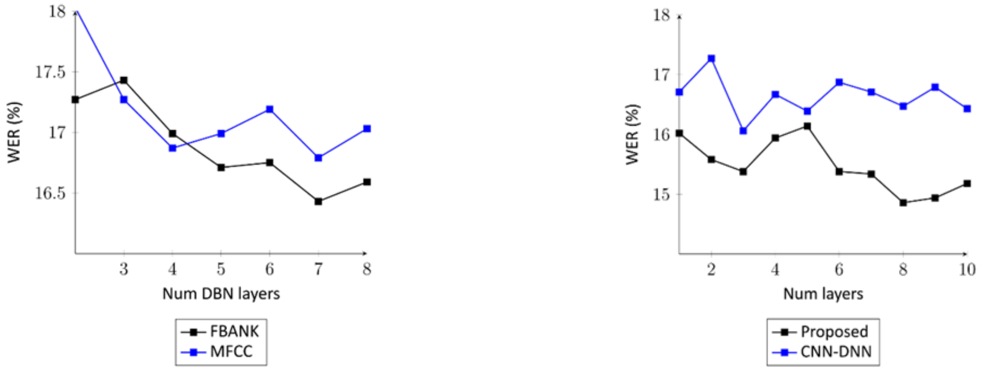
The performances of TITML-IDN for various acoustic models is shown in Table 1. From the results, we can observe that MFCC is better than FBANK for HMM-GMM systems. This is not surprising since GMM for HMM-GMM system is built to have zero covariance with the exception of its diagonal values assuming that the features are uncorrelated, to reduce the computation. Hence, only diagonal parts of the covariance matrices of GMM is considered. For MFCC, this assumption is true since DCT is good to decorrelate features. Contrarily, components of FBANK are still highly correlated and cannot be modeled only with the diagonal parts of the covariance matrices. However, we notice that applying feature transformations proves effective for FBANK. It is generally better than MFCC. We find that HMM-GMM with LDA+MLLT achieves the best performance for FBANK and subsequently decide to use this model for back-end of our deep learning system.

Table 1. Baselines results (WER) with MFCC and FBANK features for HMM-GMM and DBN-DNN systems

Acoustic Models	MFCC	FBANK
HMM-GMM (Mono)	27.79	68.15
HMM-GMM (triphones)	27.27	85.42
HMM-GMM with LDA+MLLT	26.14	22.85
HMM-GMM LDA+MLLT+SAT	24.54	23.73
HMM-GMM +MMI	24.26	23.53
DBN-DNN	16.67	16.43
CNN-DNN	17.15	16.47
PROPOSED		15.46

As expected, better performance is achieved using deep learning. For DBN-DNN system error reduction rate of 35.77% is achieved for MFCC, while 28.10% is achieved for FBANK. We notice that the performance of DBN-DNN is affected by the number of the DBN layers as illustrated in Figure 4. This fact is more observable for FBANK. The best performance is achieved when seven layers of DBN are used.

Figure 4 shown comparisons of the porposed method with several baselines systems. In the proposed method, we notice that the variation of DNN layer also affects the performance. The performance improves with the increase in number of DNN layers. The best performance is achieved when 8 layers of DNN layers. WER of 15.46 could be achieved. We notice, adding more layers further is not effective in improving the accuracy. This can be seen in Figure 4(a). In addition, as shown in Figure 4(b), compared to CNN-DBN systems, it is clear that the use of CNN as feature learning is more effective than their use in generative training. The proposed method is consistently better for all variations of DNN layers confirming the effectiveness of our method.



(a) (b)

Figure 4. The Performance of systems when the number of layers is varied for, (a) The effect of the depth of DBN to the performance of DBN-DNN; DBN-DNN systems using two different features of MFCC and FBANK, and (b) The performance comparison of the proposed method and CNN-DNN systems when number of DNN layers are varied; CNN-DNN and Proposed systems

From our experiments as shown in Table 1, it is clear that the proposed method is superior to other methods. word error rate (WER) of 7.25% and 9.85% are achieved compared to DBN-DNN and CNN-DNN, respectively, when MFCC is used, while improvements of 9.01% and 6.13% of EER are achieved for FBANK with DBN-DNN and CNN-DNN, respectively.

4. CONCLUSION

In this paper, we have explored the development of Indonesian LVCSR. The limited amount of data was one major challenge in the field. We have developed conventional HMM-GMM system for this data as well as deep learning based method such as state-of-the-art DBN-DNN and CNN-DNN systems. Both DBN-DNN and CNN-DNN systems showed significantly better performance than HMM-GMM. We also found that setting the depth of DBN and/or DNN layers could slightly improve the performances of our LVCSR. In addition, we also have proposed CNN based features for deep features of our ASR system. Evaluation on TITML-IDN dataset showed that the propose method improved the performance of LVCSR by 9.01% and 6.13% compared to state-of-the-art DBN-DNN and CNN-DNN systems, respectively. We noticed that the best performance ~~is~~ was achieved when we used 2 CNN and 8 DNN layers.

REFERENCES

[1] M. Zhou, Z. Qin, X. Lin, S. Hu, Q. Wang, and K. Ren, “Hidden voice commands: Attacks and defenses on the VCS of autonomous driving cars,” *IEEE Wireless Communications*, vol. 26, no. 5, pp. 128–133, 2019, doi: 10.1109/MWC.2019.1800477.

[2] C. J. Baby, N. Munshi, A. Malik, K. Dogra, and R. Rajesh, “Home automation using web application and speech recognition,” *2017 International Conference on Microelectronic Devices, Circuits and Systems, ICMDCS 2017*, vol. 2017-Janua, pp. 1–6, 2017, doi: 10.1109/ICMDCS.2017.8211543.

[3] Z. Y. Chan and P. Shum, “Smart office-A voice-controlled workplace for everyone,” *ACM International Conference Proceeding Series*, 2018, doi: 10.1145/3284557.3284712.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, “Robust automatic speech recognition: a bridge to practical applications,”

Academic Press, 2015.

[5] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, pp. 6645–6649, 2013, doi: 10.1109/ICASSP.2013.6638947.

[6] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schluter, and H. Ney, “A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, pp. 2462–2466, 2017, doi: 10.1109/ICASSP.2017.7952599.

[7] P. Swietojanski, A. Ghoshal, and S. Renals, “Convolutional neural networks for distant speech recognition,” *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014, doi: 10.1109/LSP.2014.2325781.

[8] B. Liu, W. Zhang, X. Xu, and D. Chen, “Time Delay Recurrent Neural Network for Speech Recognition,” *Journal of Physics: Conference Series*, vol. 1229, no. 1, 2019, doi: 10.1088/1742-6596/1229/1/012078.

[9] D. F. Campbell, C. Mcdonnell, M. Meinardi, and B. Richardson, “The need for a speech corpus,” *ReCALL*, vol. 19, no. 1, pp. 3–20, 2007, doi: 10.1017/S0958344007000213.

[10] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[11] W. Zou, D. Jiang, S. Zhao, G. Yang, and X. Li, “A comparable study of modeling units for end-to-end Mandarin speech recognition,” *2018 11th International Symposium on Chinese Spoken Language Processing, ISCSLP 2018-Proceedings*, pp. 369–373, 2018, doi: 10.1109/ISCSLP.2018.8706661.

[12] M. Neumann and N. G. Thang Vu, “CRoss-lingual and Multilingual Speech Emotion Recognition on English and French,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol. 2018-April, pp. 5769–5773, 2018, doi: 10.1109/ICASSP.2018.8462162.

[13] J. Xu, K. Matta, S. Islam, and A. Nürnberger, “German Speech Recognition System using DeepSpeech,” *ACM International Conference Proceeding Series*, pp. 102–106, 2020, doi: 10.1145/3443279.3443313.

[14] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, and K. Hosseinzadeh, “A large vocabulary continuous speech recognition system for persian language,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, pp. 1–12, 2011, doi: 10.1186/1687-4722-2011-426795.

[15] S. Sakti, E. Kelana, H. Riza, and S. Sakai, “Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project,” *Tcast*, pp. 19–24, 2008.

[16] S. Suyanto, A. Arifianto, A. Sirwan, and A. P. Rizaendra, “End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language,” *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166346.

[17] A. Winursito, R. Hidayat, and A. Bejo, “Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition,” *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, vol. 2018-Janua, pp. 379–383, 2018, doi: 10.1109/ICOIACT.2018.8350748.

[18] E. R. Swedia, A. B. Mutiara, M. Subali, and Ernastuti, “Deep learning long-short term memory (LSTM) for Indonesian speech digit recognition using LPC and MFCC Feature,” *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*, 2018, doi: 10.1109/IAC.2018.8780566.

[19] M. H. Tambunan, Martin, H. Fakhruroja, Riyanto, and C. Machbub, “Indonesian speech recognition grammar using Kinect 2.0 for controlling humanoid robot,” *2018 International Conference on Signals and Systems, ICSigSys 2018-Proceedings*, pp. 59–63, 2018, doi: 10.1109/ICSIGSYS.2018.8373568.

[20] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, “Speech recognition of Moroccan dialect using hidden Markov models,” *Procedia Computer Science*, vol. 151, pp. 985–991, 2019, doi: 10.1016/j.procs.2019.04.138.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, pp. 3586–3589, 2015, doi: 10.21437/interspeech.2015-711.

[22] N. Jaitly and G. E. Hinton, “Vocal Tract Length Perturbation (VTLP) improves speech recognition,” *Proceedings of the 30 th International Conference on Machine Learning*, vol. 90, pp. 42–51, 2013, [Online]. Available: http://www.iarpa.gov/.

[23] M. C. Stoian, S. Bansal, and S. Goldwater, “Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol. 2020-May, pp. 7909–7913, 2020, doi: 10.1109/ICASSP40776.2020.9053847.

[24] Y. Miao, F. Metze, and S. Rawat, “Deep maxout networks for low-resource speech recognition,” *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013-Proceedings*, pp. 398–403, 2013, doi: 10.1109/ASRU.2013.6707763.

[25] H. F. Pardede, A. R. Yuliani, and R. Sustika, “Convolutional Neural Network and Feature Transformation for Distant Speech Recognition,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, p. 5381, 2018, doi: 10.11591/ijece.v8i6.pp5381-5388.

[26] T. Faisal, I. Negassi, G. Goitom, M. Yassin, A. Bashir, and M. Awawdeh, “Systematic development of real-time driver drowsiness detection system using deep learning,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 148–160, 2022, doi: 10.11591/ijai.v11.i1.pp148-160.

[27] F. Taher and N. Prakash, “Automatic cerebrovascular segmentation methods-A review,” *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 576–583, 2021, doi: 10.11591/ijai.v10.i3.pp576-583.

[28] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, “End-to-end speech recognition from the raw waveform,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, pp. 781–785, 2018, doi: 10.21437/Interspeech.2018-2414.

[29] H. Seki, K. Yamamoto, and S. Nakagawa, “A deep neural network integrated with filterbank learning for speech recognition,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, pp. 5480–5484, 2017, doi: 10.1109/ICASSP.2017.7953204.

[30] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998, doi: 10.1006/csla.1998.0043.

[31] A. R. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012, doi: 10.1109/TASL.2011.2109382.

[32] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014, doi: 10.1109/TASLP.2014.2339736.

[33] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018, doi: 10.1109/TETCI.2017.2762739.

[34] A. M. Badshah *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, 2019, doi: 10.1007/s11042-017-5292-7.

[35] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, pp. 4153–4156, 2012, doi: 10.1109/ICASSP.2012.6288833.

[36] T. R. Niesler and P. C. Woodland, "Variable-length category-based n-gram language model," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol. 1, pp. 164–167, 1996, doi: 10.1109/icassp.1996.540316.

[37] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages," *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, vol. 29, pp. 759–765, 2012.

BIOGRAPHIES OF AUTHORS

Hilman F. Pardede received his bachelor degree in electrical engineering from University of Indonesia. His master degree is from The University of Western Australia and his doctoral degree is obtained from Tokyo Institute of Technology in Computer Science. He is currently a research professor at Research Center for Data and Information, The National Research and Innovation Agency. His research interests include machine learning, multimedia signal processing and pattern recognition, artificial intelligence. He can be contacted at hilm003@brin.go.id.

Purwoko Adhi received his engineering degree in electrical engineering from ESIGELEC and his master degree in robotics from Pierre and Marie Curie University, Paris 6, France. He received his PhD in signal processing from Curtin University of Technology, Western Australia. He is currently a senior researcher at Research Center for Telecommunication, National Research and Innovation Agency, Indonesia. His research interests include signal processing, image processing, machine learning, and wireless communications. He can be contacted at email: purwoko.adhi@brin.go.id.

Vicky Zilvan is currently a researcher in the Research Center for Data and Information Sciences, National Research and Innovation Agency Republic of Indonesia. He received his undergraduate degree in computer science from Institut Pertanian Bogor (IPB)-Indonesia, and his Master's degree in informatics from Institut Teknologi Bandung (ITB)-Indonesia. His areas of interest include artificial intelligence, machine learning, pattern recognition, speech processing, and data mining. He can be contacted at email: vicky.zilvan@brin.go.id.

Ade Ramdan obtained his bachelor degree from Bandung Institute of Technology in 2008. He is currently a researcher at the Research Center for Data and Information Sciences, National Research and Innovation Agency Republic of Indonesia. His interests include Machine Learning, Pattern Recognition dan Embedded System. He can be contacted at email: ader004@brin.go.id.

Dikdik Krisnandi received his M. Eng. (2005) in Electrical Engineering from the Bandung Institute of Technology, with a focus on Information Technology. In 2008, he joined the Indonesian Institute of Sciences (It is now known as the National Research and Innovation Agency-Indonesia) as a researcher. His research interests are focused on Machine Learning and on the Internet of Things applications. He can be contacted at email: dikd003@lipi.go.id.