End-to-end indonesian speech recognition with convolutional and gated recurrent units

Rifqi Adiwidjaja, M Ivan Fanany
Faculty of Computer Science, University Of Indonesia, Kampus UI Depok, Jawa Barat 16424
E-mail: rifqi.adiwidjaja71@ui.ac.id, ivan@cs.ui.ac.id

Abstract. Automatic Speech Recognition has penetrated deeply into our life. For well-resourced language, it can be considered as solved, but that’s not the case for under-resourced language like Bahasa. Although it’s the 7th most spoken language in the world, the research of speech recognition for Bahasa was still extremely limited, specifically that supports sentence level input with variable character length with end-to-end training. We built the model using the deep learning approach, specifically utilizing the residual networks and Bi-Directional Gated Recurrent Unit (Bi-GRU). To the best of our knowledge, this is the first Indonesian ASR model that can be trained in an end-to-end manner. Our model surpassed the baseline model on all metrics and achieve competitiveness with the current best result, which used the visual modal, for the dataset even with a more difficult and prone to noise modality like sound.

1. Introduction
Automatic Speech Recognition (ASR) is the use of methodologies and technologies that enables the recognition and translation of spoken language into text by computers. ASR has penetrated deeply into our life and it can be seen from it’s usage from a home appliance like Amazon Alexa to a more sophisticated identity identification mechanism. All of this is possible because ASR can be considered solved for well-resourced language like English, but unfortunately for under-resourced languages like Bahasa and Javanese, the topic of ASR remains unsolved. Indonesian language, or Bahasa, is one of the under-resourced languages due to a limited amount of research and available datasets even though Bahasa combined with Malay is the 7th most spoken language in the world.

Traditionally the development of an ASR system was done by using several distinct models, acoustic, lexicon, and language, in tandem. However even when popular toolkit such as the Kaldi toolkit [1] have reduced the barrier required to build such models, it is still not an easy task, therefore there have been attempts to simplify the process by making it trainable in an end-to-end manner or making the 3 models into 1. For the english language there has been many and successful attempts for an end-to-end speech recognition [2] [3]. For the development of ASR in Indonesia, we have observed that it is still sticking to the traditional three models approach [4] [5] [6]. The primary goal of this paper is thus to create an Indonesian ASR model that can accept a sentence level with variable-length input and trainable in an end-to-end manner thus made it more applicable for industry.
Figure 1. outline of the ASR pipeline

The model architecture consists of these components: the feature extractor using Convolutional Neural Network (CNN) and character level sentence decoder using Gated Recurrent Unit (GRU) and using CTC as the loss function. The organization of this paper then will be continued by dataset used in the experiment, architecture of the model, experiment setup, result and discussion, and the last section is the conclusion and future works.

2. Dataset
In this section, we will describe the dataset used in the experiment which is AVID [7] dataset and the preprocessing steps we took to get the input for our model.

2.1. AVID
AVID consists of video and audio recording of 10 speakers, 5 females and 5 males, each speaking 1040 sentences. In total there are 10400 sentences and with each recording recorded with a duration of 3 seconds with a sampling rate 48kHz and bitrate of 128kbps. Sentences in AVID dataset consists of 52 different words which is distributed uniformly, and randomly generate unique sentence with the following sentence structure: command + object + color + preposition + letter + digit. Each component is composed respectively as follows: \{taruh, pukul, susun, buang\}. \{cat, tas, bibir\}. \{merah, hijau, biru, putih\}. \{di, ke, dari, dengan\}. \{A, ... , Z\}. \{nol, ... , sembilan\}. An example of sentence built with this grammar is "taruh tas hijau di A satu".

2.2. Preprocessing
The raw audio is transformed into a 96-dimensional log mel spectrograms. Each sample is normalized to have zero mean and unit variance over the training set. And lastly, every input will be padded to make each sequence has the same length.

3. Architecture
Our model was constructed using a deep learning approach, which was at the forefront of ASR research, with models reaching state-of-the-art on many datasets. Our model will use Residual Network (ResNet) [8] and Bi-Directional Gated Recurrent Unit (Bi-Gru) [9].

3.1. ResNet
To extract the spatial features from the spectrograms we trained a residual network with pre-activation identity shortcut connections [10]. ResNet was first introduced in [8] for image recognition and although we are using audio data, we used the spectrograms for our feature which is like an image thus made ResNet the appropriate choice. What made ResNet different than its predecessors is that ResNet used shortcut connections between the convolution blocks which allows a better flow of information and gradients. Due to the size of the datasets we used the ResNet-18 architecture, the details of the architecture can be seen in Table 1.

3.2. Bi-GRU
Two layers of the Bidirectional Gated Recurrent Unit (Bi-GRU) will be used to get the temporal features. The first Bi-GRU layer ingests the features extracted by ResNet and the final layer
emits a character probability for every time steps. The Bi-GRU has 256 cells each, the details of the architecture can be seen in Table 1.

### 3.3. CTC

The traditional approach in handling sequence prediction involves segmenting input into several meaningful chunks, such as labeling the word boundary for the words in sentence recognition or labeling the phonemes in a word for word recognition. Such an approach, however needs to be done manually by hand, which is a time-consuming process. CTC [11] however enables training to be performed without the need for such preprocessing, by applying softmax function to the output of the network for every time step, which provides a probability for emitting each label from the dictionary of characters of output symbols at that time step. An extra `blank` output class introduced to the dictionary to denote no character at that time step. CTC computes a probability distribution over all possible sequences, for activation $y^t_l$ of unit $l$ at time step $t$ is the probability of observing class $l$ at time $t$ given an input sequence of length $T$, thus by putting together all of these probabilities we get the probability distribution over all possible sequences. Then the probability of observing a particular sequence $\pi$ given an input sequence of $x$ of length $T$ and a training set $S$ can be written in the following equation:

$$p(\pi|x, S) = \prod_{t=1}^{T} y^t_{\pi_t}$$

(1)

#### Table 1. Architecture details for the proposed model. A shortcut connection is added after every convolutional block.

| Layer Name   | Filters and Cells       |
|--------------|-------------------------|
| ConvBlock 1a | [3 x 3, 24] x 2 / [1, 1] |
| ConvBlock 1b | [3 x 3, 24] x 2 / [1, 1] |
| ConvBlock 2a | [3 x 3, 48] x 2 / [1, 2] |
| ConvBlock 2b | [3 x 3, 48] x 2 / [1, 1] |
| ConvBlock 3a | [3 x 3, 96] x 2 / [1, 2] |
| ConvBlock 3b | [3 x 3, 96] x 2 / [1, 1] |
| ConvBlock 4a | [3 x 3, 192] x 2 / [1, 2] |
| ConvBlock 4b | [3 x 3, 192] x 2 / [1, 1] |
| Bi-GRU 1     | 256 hidden units        |
| Bi-GRU 2     | 256 hidden units        |

### 4. Experiments

In this section, the evaluation will be performed for the proposed model. Lastly, experiment results will be discussed, and the conclusion will be drawn based on that. The model was trained end-to-end with log mel spectrogram of the audio as inputs and one-hot vector encoded sentences as labels.

#### 4.1. Data

Due to the size of the dataset, we have only divided it into training and testing set with a ratio of 80:20 respectively. Thus out of 10 speakers, we used 2 speakers as our testing set, 1 male speaker and 1 female speaker.
4.2. Evaluation
Evaluation was performed using three metrics: Characters Error Rate (CER), Word Error Rate (WER), and Bilingual Understudy (BLEU) [12]. The formula for CER and WER are as follows while BLEU can be found in [12]

\[
WER = \frac{S + I + D}{N} \times 100
\]

\[
CER = \frac{S + I + D}{N} \times 100
\]

where S, I, D, and N are the number of substitution, insertion, deletion, and words/characters respectively. Sentence decoding was performed using greed search and without a language model.

4.3. Augmentation
We augment the dataset using simple augmentation techniques to prevent overfitting. We used frequency mask, and time mask [13] on our spectrograms with the followings mask specification, 0 to 27 consecutive frequency will be masked and 0 to 100 consecutive timesteps will be masked for the frequency mask and time mask respectively.

4.4. Model Implementation
Our model was made of 18 layers of residual networks and 2 layers of Bi-Directional Gated Recurrent Unit (Bi-GRU), and we trained it from scratch without any pretrained models. Each convolutional layer has a filter size of $3 \times 3$ and the number of filters starts from 24 and doubled after every convolutional block and with activation function ReLu [14]. The regularization dropout [15] also applied on the ResNet with probability of 0.1. For the Bi-GRU we merge the information from the past and future by concatenating them. We trained the model using Adam optimizer [16] and learning rate $10^{-4}$.

4.5. Baseline
To evaluate our proposed model, we compare its performance with an ablation model inspired by recent state-of-the-art work that used residual networks and bidirectional-long short term memory network (Bi-LSTM) [3]. The specific architecture of the baseline model will follow our proposed model but with the Bi-GRU layers switched with Bi-LSTM layers.

Table 2. Experiment results.

| Method  | CER   | WER   | BLEU  |
|---------|-------|-------|-------|
| Baseline| 7.7%  | 15.6% | 87.5% |
| Proposed| 6.6%  | 14.1% | 88.9% |

5. Result And Discussion
The result of our experiments can be seen in Table 2. As it clearly is shown that our proposed model outstrip the performance of the baseline model on all metrics. We attribute this due to the Bi-GRU layers in our proposed model more suited to the AVID dataset, which has small amount of data consistent with the findings of Chung et al. (2014) [17]. Lastly due to how the dataset was made, with unique sentences, we believe that our model have also learnt on the word level.
6. Conclusion
In this paper, we have built an Indonesian ASR model, which is to the best of our knowledge the first sentence-level end-to-end Indonesian ASR model which can handle sentence level input with a variable length of characters, to make it more applicable to real-world settings such as the industry. The model consists of a spatiotemporal feature extractor and character-level sentence decoder, which was trained using CTC, to allow end-to-end training. Our proposed model have achieved a competitive result with [18] despite using the more difficult sound modality. Although it’s worth to note that we suspect that the ambient noise prevalent in the audio dataset might hinder our model, thus in the future we will try to apply speech enhancement, transfer learning, and language models to further improve our result.

References
[1] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G and Vesely K 2011 IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (IEEE Signal Processing Society) iEEE Catalog No.: CFP11SRW-USB
[2] Graves A and Jaitly N 2014 Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 ICML’14 (JMLR.org) pp II–1764–II–1772 URL http://dl.acm.org/citation.cfm?id=3044805.3045089
[3] Saon G, Kurata G, Sercu T, Audhkhasi K, Thomas S, Dimitriadis D, Cui X, Ramabhadran B, Picheny M, Lim L, Roomi B and Hall P 2017 CoRR abs/1703.02136 (Preprint 1703.02136) URL http://arxiv.org/abs/1703.02136
[4] Prakoso H, Ferdiana R and Hartanto R 2016 2016 International Symposium on Electronics and Smart Devices (ISESD) pp 283–286
[5] Ferdiansyah V and Purwarianti A 2011 Proceedings of the 2011 International Conference on Electrical Engineering and Informatics pp 1–4
[6] Muljono, Syadida A Q, Setiadi D R I M and Setyono A 2017 2017 International Seminar on Application for Technology of Information and Communication (iSemantic) pp 264–267
[7] Maulana M R A R and Fanany M I 2017 2017 International Conference on Advanced Computer Science and Information Systems (ICACISIS) pp 381–386
[8] He K, Zhang X, Ren S and Sun J 2016 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 770–778
[9] Goodfellow I, Bengio Y and Courville A 2016 Deep Learning (The MIT Press) ISBN 0262035618, 9780262035613
[10] He K, Zhang X, Ren S and Sun J 2016 vol 9908 pp 630–645 ISBN 978-3-319-46492-3
[11] Graves A, Fernández S, Gomez F and Schmidhuber J 2006 Proceedings of the 23rd International Conference on Machine Learning ICML ’06 (New York, NY, USA: ACM) pp 369–376 ISBN 1-59593-383-2 URL http://doi.acm.org/10.1145/1143844.1143891
[12] Papineni K, Roukos S, Ward T and Zhu W J 2002 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL ’02 (Stroudsburg, PA, USA: Association for Computational Linguistics) pp 311–318 URL https://doi.org/10.3115/1073083.1073135
[13] Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D and Le Q V 2019 Spec augment: A simple data augmentation method for automatic speech recognition (Preprint 1904.08779)
[14] Glorot X, Bordes A and Bengio Y 2011 Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research vol 15) ed Gordon G, Dunson D and Dudik M (Fort Lauderdale, FL, USA: PMLR) pp 315–323 URL http://proceedings.mlr.press/v15/glorot11a.html
[15] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 J. Mach. Learn. Res. 15 1929–1956 ISSN 1532-4435 URL http://dl.acm.org/citation.cfm?id=2627435.2670313
[16] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization (Preprint 1412.6980)
[17] Chung J, Güçlüler C, Cho K and Bengio Y 2014 CoRR abs/1412.3555 (Preprint 1412.3555) URL http://arxiv.org/abs/1412.3555
[18] Maulana M R A R and Fanany M I 2017 2017 International Conference on Advanced Computer Science and Information Systems (ICACISIS) pp 375–380