Chinese Audio Transcription Using Connectionist Temporal Classification

Jansen
Universitas Tarumanagara
Letjen S. Parman Street No. 1
Jakarta, Indonesia
vincentjansen16@gmail.com

Dyah E Herwindiati
Universitas Tarumanagara
Letjen S. Parman Street No. 1
Jakarta, Indonesia
dyahh@fti.untar.ac.id

Janson Hendryli
Universitas Tarumanagara
Letjen S. Parman Street No. 1
Jakarta, Indonesia
jansonh@fti.untar.ac.id

ABSTRACT
Mandarin is one of the global languages that have large users and speakers. There are several important factors for learners to be an expert in Mandarin. To be able to communicate properly, mastery in Chinese character (hànzì) and pīnyīn are required. We develop an Android-based app to help students who learn Mandarin. It can help them practice the accuracy of their pronunciation and intonation in accordance to the sentences displayed on the screen, which are taken from the HSK textbook. The system recognizes human voice and transcribes it to hànzì. The recorded voice goes through the feature extraction using the filter bank method. The feature vectors are then fed into deep learning architecture to get the pīnyīn. The architectures are the convolutional neural network, recurrent neural networks, and connectionist temporal classification. After the pīnyīn letters are generated, the Markov chain rule is used to convert it into hànzì. The best word error rate from the model is 18.919% from training and 19.922% from test data. From the user testing, we find that the error rate is 49.659±16.372%, due to background noise and user's pronunciation speed.

CCS Concepts
• **Applied computing→Education**
• **Computing methodologies→Deep learning.**

Keywords
Audio transcription; connectionist temporal classification; mandarin; word error rate.

1. INTRODUCTION
Generally, there are four main aspects that need to be mastered when learning Mandarin, namely reading (阅读), writing (书写), listening (听力), and speaking (口语). Mastering all these aspects makes studying both hànzì and pīnyīn necessary. The number of hànzì characters in Mandarin is more than 80,000 characters, although, in the published dictionary, it only discusses 20,000 characters. Normally, someone is said to be fluent in Mandarin if he/she has mastered approximately 2500 characters [1]. Meanwhile, pīnyīn letters are used to read each hànzì characters.

ICCCM'20, July 17–19, 2020, Singapore, Singapore

ACM ISBN 978-1-4503-8766-8/20/07…$15.00
https://doi.org/10.1145/3411174.3411180

Aspects that need to be considered when reading the pīnyīn are consonants (声母), vowels (韵母), and intonation (声调).

One of the difficulties when learning Mandarin is to remember hànzì characters and to be able to say each character with the correct pronunciation and intonation. Therefore, we proposed an Android-based application, the Chinese Audio Transcription (CAT), that is able to accept human voices speaking in Mandarin and transcribe it into hànzì. Considering that the app mainly aims to help Mandarin learners to speak correctly, we provide a feature where the users can learn to read hànzì from the textbook of Hànyǔ Shuǐpíng Kǎoshì, commonly abbreviated as HSK. The HSK is the standardized Chinese proficiency test for non-native speakers. The HSK levels provided in the app are from level one to level three. In addition to that, the app also has a free transcription feature. The users can use this app to correct their pronunciation or intonation when speaking Mandarin by checking the audio transcription of their voices, either by following the HSK tests as the reference, or generally, any sentences in Mandarin.

2. METHODS
The audio transcription system extracts human voices using the filter bank method. The results of the feature extraction process are in the form of feature vectors, which are then fed into the classification model. We experiment with several deep learning architectures, such as the convolutional neural network, long short-term memory network, and gated recurrent unit. After the deep learning architectures, we employ the connectionist temporal classification (CTC) model. The output of the CTC is the pīnyīn of the voice recording in Mandarin. The Markov chains method is then employed to convert the pīnyīn into hànzì characters. Finally, the metric used to evaluate the performance of our system is the word error rate.

2.1 Filter Bank
The filter bank is a feature extraction method often used in speech processing. Basically, the filter bank is similar to the mel-frequency cepstral coefficients method. The difference is that the filter bank does not use discrete cosine transform to process the sound [2]. The step in the filter bank method starts from pre-emphasis which is used for filtering the voice signal to maintain high frequency in the spectrum. This step also reduces the noise ratio of the signal and balances the sound spectrum [3].

The next step is the frame blocking process to separate the sound into several parts of the frames. Using overlapping techniques for each frame, the purpose of this step is to avoid losing the characteristics of the sound in each frame [4]. A windowing process is then applied to each frame generated after the frame blocking process. A commonly used windowing function is the Hamming window. This step aims to reduce the signal gap of each frame [5].

Finally, a fast Fourier transform is applied to transform the signal from the time domain to the frequency domain. The purpose of applying this transformation is because signals in frequency domain tend to be invariant to loud or weak voices [6].

2.2 Convolutional Neural Network

A convolutional neural network or CNN is a type of neural network commonly applied to classify spatial data. The CNN basically consists of the convolutional layers and several pooling layers. The convolution layers convolve the input with a weight matrix and pass the output to the next layer. It can be seen as learning the best feature representation of the data. The pooling layers act as dimensional reduction, which can be a max, min, or average pooling. Several past research such as in [7] and [8] employed CNN with a filter bank.

2.3 Long Short-Term Memory

The long short-term memory network (LSTM) is a recurrent neural network (RNN) architecture that is introduced to fix the problem of long-term dependencies of vanilla RNN where the slope value is too fast to increase or decrease when updating the weights [9]. The LSTM model consists of four different gates and two states in each cell.

The gates and states in one LSTM cell work together to produce new information. The gates of the LSTM are the forget gate, input gate, candidate gate, and output gate. The forget gate is a gate that determines how important information in the previous state is to be stored. Meanwhile, the input gate and candidate gate determines the information at the current time step that needs to be saved into the cell state. The output gate determines the value of the next hidden state. The states in LSTM consists of cell state and hidden state. The cell state stores important information from the three previous gates at the current time step. Finally, the hidden state brings information from the current cell to the next cell.

2.4 Gated Recurrent Unit

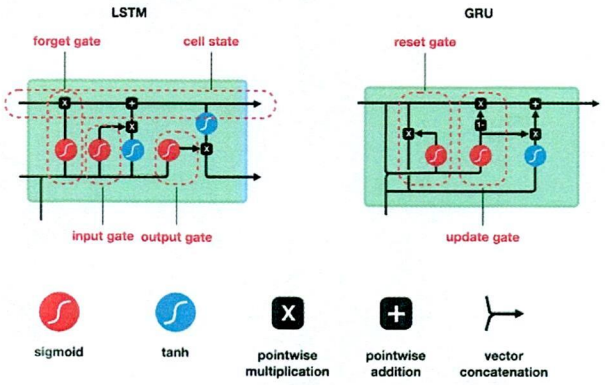

Figure 1. Comparison of LSTM and GRU model[1].

Gated recurrent unit, or usually abbreviated as GRU, is a variation of the LSTM, where it merges some existing gates in the LSTM. The advantage of combining gates is that some information deemed as unimportant can still be retained by the model [10]. The forget and input gates in the LSTM model are combined into the update gate in the GRU model. In addition to that, the cell and hidden state are also combined into the reset gate. Fig. 1 illustrates the GRU model compared to LSTM.

The update gate determines information from the previous cell that needs to be computed in the current cell. Meanwhile, the reset gate decides the amount of information that needs to be forgotten. There is also a current memory content that works together with the reset gate to save important information from the previous time step. Finally, the final memory uses the update gate to determine the important information that must be retrieved from the current memory content. This important information is continued to the next cell.

2.5 Connectionist Temporal Classification

Connectionist temporal classification (CTC) [11] is a model used for the speech recognition process. This model is widely used as an alternative to the hidden Markov model because it can directly map each frequency of sound into text. The CTC is used as the output or scoring function, usually for training recurrent neural network models, such as LSTM or GRU.

The CTC model introduces a new variable named blank token, which is a null delimiter with no meaning during the speech processing [12]. This blank token is often represented as 'ε' or '-'.

2.6 Markov Chains

So far, the output of the model is the pīnyīn representation from the human voice recording speaking Mandarin. Conversion of the pīnyīn to hànzì characters, which is called decoding, uses the Markov chain rules. The chain rules look for the highest possible value that can be arranged into a sentence [13].

2.7 Word Error Rate

The performance of the audio transcription model is measured using the word error rate (WER). Smaller WER value constitutes to better performance of the model. WER can be calculated as in Eq. 1 [14].

$$\text{WER} = \frac{I+D+S}{M} * 100\% \quad (1)$$

where M denotes the total number of spoken words, I denotes the number of inserted words even though the actual sentence does not exist, D is words not detected by the system during the transcription process, and S is the transcripted words that are the same as the actual words.

3. EXPERIMENTS

3.1 Data

The data set used in this research is the THCHS-30 [15] from the Center for Speech and Language Technology at Tsinghua University and the ST-CMDS-20170001_1 Free ST Chinese Mandarin Corpus made available by Surfingtech[2]. The first data set contains 13,388 speech data recorded from a microphone. Meanwhile, the second data set is recorded using a cellphone in a silent indoor environment with 855 speakers. Each speaker records 120 utterances. The transcriptions are included in both data set. We split the first data set 75%-5%-20% for training, validation, and testing, respectively. Likewise, the second data set is split 95%-2%-3%. Since the second data set is recorded using a cellphone, which is similar to our target application, we use more training data from the second data set. It should also be noted that all of the recordings do not include any noises.

[1] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[2] http://www.surfing.ai

3.2 Implementations

Initially, we experiment on three different deep learning architectures for the transcription of a human voice speaking in Mandarin to pīnyīn. The first model (CAT 1) uses the convolutional neural network followed by the connectionist temporal classification. The second model (CAT 2) consists of the convolutional neural network followed by long short-term memory network and connectionist temporal classification. And finally, the third model (CAT 3) uses the convolutional neural network, gated recurrent unit, and connectionist temporal classification. Summarily, all of the models start with a convolutional neural network and end with connectionist temporal classification. We try to take advantage of sequential models, such as the long short-term memory network and the gated recurrent unit, to improve the performance of the model by inserting it between convolutional neural network and connectionist temporal classification as in CAT 2 and CAT 3 models.

The architecture of each model can be seen in Fig. 2, 3, and 4. The models are implemented using Python with Tensorflow and Keras library. They are trained using Adam optimizer with default Keras parameters and batch size of 16. Additionally, the feature extractions are performed using the python-speech-features library.

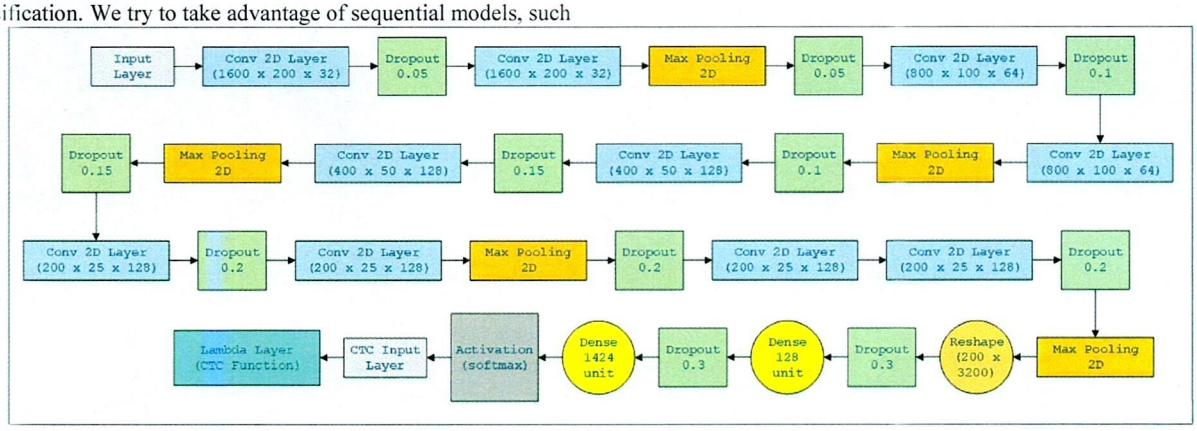

Figure 2. The architecture of CAT 1 model.

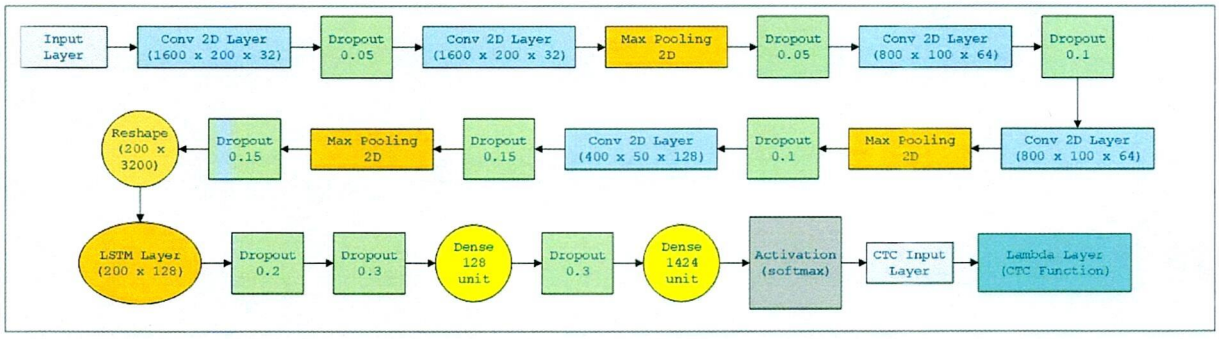

Figure 3. The architecture of CAT 2 model.

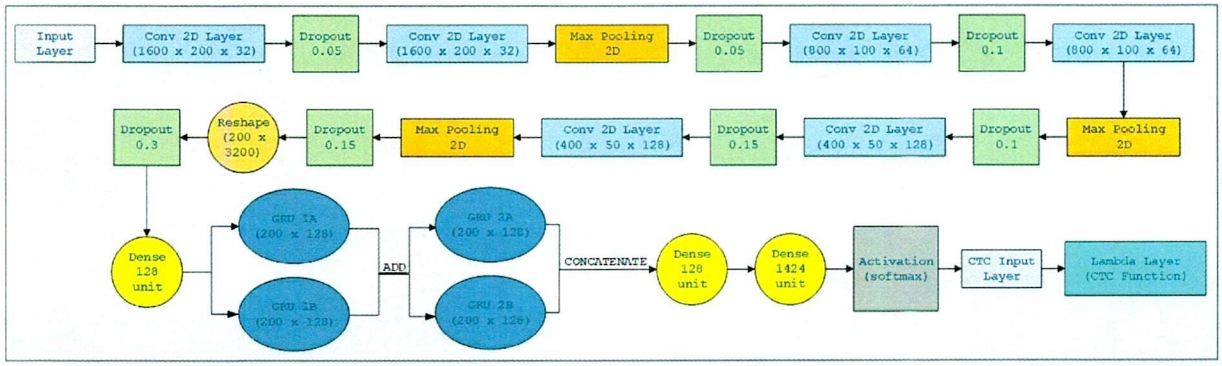

Figure 4. The architecture of CAT 3 model.

4. RESULTS

The summary of the experimental results on each CAT 1, CAT 2, and CAT 3 model can be seen in Fig. 5. Overall, the CAT 1 and CAT 2 model has good performance as can be seen from the low word error rates. However, the CAT 2 has a higher error when tested with the test data. From analyzing the output of the CAT 2 model, the bad performance is due to the fact that words are often lost during the transcription process using the model. In this case, the deletion factor, or D in Eq. 1, increases, so that the error also increases.

The CAT 3 model fails in conducting the audio transcription. The model is almost unable to predict any spoken words with word error rate close to 100%. This model can only recognize 15 pīnyīn words: hǎo, huáng, jiā, jīn, jiù, nǐ, shàng, shè, shí, shì, tā, wǒ, xià, yè, and zài.

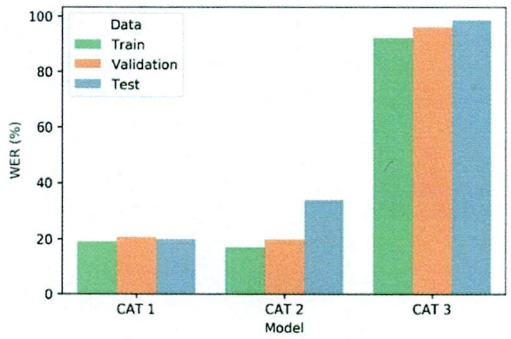

Figure 5. The results of each CAT model for training, validation, and testing data.

From the WER results above, we select CAT 1 model, with a training error of 18.919% and a test error of 19.922%, to be used for the Chinese audio transcription app. Fig. 6 shows the user interface of the app. Furthermore, 27 users test the free and HSK transcription feature of the app. The average WER of each test can be seen in Fig. 7.

Figure 6. The user interface of the Chinese audio transcription app.

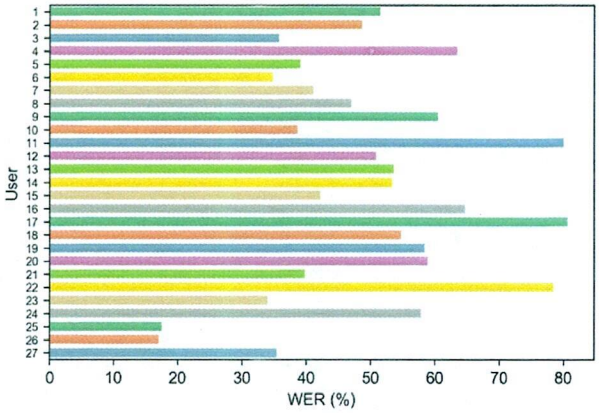

Figure 7. The average WER of each user testing the free and HSK transcription feature of the app.

Out of 27 users, there are 13 users with WER below 50%. Overall, the average WER is 49.659±16.372%. Some factors affecting the high error rate in this testing scenario are the variation in background noise, incorrect pronunciation or intonation from the users, and variation in speaking speed, particularly with users who are speaking slowly.

5. CONCLUSIONS

Learning to speak Mandarin fluently includes the study of both hànzì and pīnyīn, not only in writing but also how to speak it with correct pronunciation and intonation. To help learners study Mandarin, we develop a Chinese audio transcription app, utilizing the deep learning models to automatically recognize the user utterances in Mandarin and transcribe it into text. The app provides examples from the HSK textbook and also has the free transcription feature where the users can check their pronunciation and intonation when speaking any Mandarin. Each audio recording of the users is first going through the feature extraction

process where the filter bank method is used. From the experiments, we find that the best model consists of the convolutional neural network and the connectionist temporal classification. The user testing finds that the model is still affected a lot by background noise and variation in speaking speed.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to NVIDIA Corporation for the Titan X and Titan Xp GPU grant used in this research.

7. REFERENCES

[1] Olmanson, J. and Liu, X. 2017. The challenge of Chinese character acquisition: leveraging modality in overcoming a centuries-old problem. *Emerging Learning Design Journal* 4 (May 2017), 1-9.

[2] Fayek, H. 2016. *Speech processing for machine learning: filter banks, mel-frequency cepstral coefficients, and what's in-between*. Available from https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html, accessed on Aug. 18, 2019.

[3] Putra, D. and Resmawan, A. 2011. Verifikasi biometrika suara menggunakan metode MFCC dan DTW. *Lontar Komputer* 2, 1 (Jun. 2011), 8-21.

[4] Nasution, T. 2012. Metoda mel frequency cepstrum coefficients untuk mengenali ucapan pada bahasa Indonesia. *Jurnal Sains dan Teknologi Informasi* 1, 1 (Jun. 2012), 22-31. DOI=https://doi.org/10.33372/stn.v1i1.309.

[5] Permana, I. S., Nurhasanah, Y. I. and Zulkarnain, A. 2018. Implementasi metode MFCC dan DTW untuk pengenalan jenis suara pria dan wanita. *Mind Journal* 3, 1 (Jun. 2018), 49-63. DOI=https://doi.org/10.26760/mindjournal.

[6] Helmiyah, S., Fadlil, A. and Yudhana, A. 2018. Pengenalan pola emosi manusia berdasarkan ucapan menggunakan ekstraksi fitur mel-frequency cepstral coefficients. *CogITo Smart Journal* 4, 2 (Dec. 2018), 372-381.

[7] Golik, P., Tüske, Z., Schlüter, R. and Ney, H. 2015. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In *Proceedings of the Annual Conference of the International Speech Communication Association* (Dresden, Germany, Sept. 6-10, 2015). INTERSPEECH '15, 26-30.

[8] Cakir, E., Ozan, E.C. and Virtanen, T. 2016. Filterbank learning for deep neural network based polyphonic sound event detection. In *Proceedings of the International Joint Conference on Neural Networks* (Vancouver, Canada, Jul. 24-29, 2016). IJCNN '16, IEEE, 3399-3406, DOI=https://doi.org/10.1109/IJCNN.2016.7727634.

[9] Kamath, U., Liu, J. and Whitaker, J. 2019. *Deep Learning for NLP and Speech Recognition*. Mannheim: Springer Nature, Switzerland AG.

[10] Le, T., Kim, J. and Kim, H. 2016. Classification performance using gated recurrent disaggregation. In *Proceedings of the International Conference on Machine Learning and Cybernetics* (Jeju, South Korea, July 10-13, 2016). IEEE, 105-110, DOI=https://doi.org/10.1109/ICMLC.2016.7860885.

[11] Graves, A., Fernández, S. and Gomez, F. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, ICML '06, 369–376, DOI=https://doi.org/10.1145/1143844.1143891.

[12] Graves, A. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. New York: Springer.

[13] Zhou, X. Hu, X., Zhang, X. and Shen, X. 2017. A segment-based hidden Markov model for real-setting pinyin-to-Chinese conversion. In *Proceedings of ACM Conference on Information and Knowledge Management* (CIKM '07). ACM, New York, NY, 1027-1030. DOI=https://doi.org/10.1145/1321440.1321602.

[14] Ali, A. and Renals, S. 2008. Word error rate estimation for speech recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 2 (Melbourne, Australia, July 15-20, 2008). ACL '08, 20-24.

[15] Wang, D., Zhang, X. and Zhang, Z. 2015. *THCHS-30: A free Chinese speech corpus*. Available from http://arxiv.org/abs/1512.01882, accessed on Jan. 31, 2020.