Parallelising 2D-CNNs and Transformers: A Cognitive-based approach for Automatic Recognition of Learners' English Proficiency

Meishu Song[1,2], *Emilia Parada-Cabaleiro*[3],*Zijiang Yang*[1], *Xin Jing*[1], *Kazumasa Togami*[4], *Kun Qian*[5*], *Björn W. Schuller*[1,6], *and Yoshiharu Yamamoto*[2]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Educational Physiology Laboratory, The University of Tokyo, Japan,
[3] Institute of Computational Perception, Johannes Kepler University Linz, Austria,
[4] Educational Physiology Laboratory, The University of Tokyo, Japan
[5] School of Medical Technology, Beijing Institute of Technology, China
[6] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

ABSTRACT

Learning English as a foreign language requires an extensive use of cognitive capacity, memory, and motor skills in order to orally express one's thoughts in a clear manner. Current speech recognition intelligence focuses on recognising learners' oral proficiency from fluency, prosody, pronunciation, and grammar's perspectives. However, the capacity of clearly and naturally expressing an idea is a high-level cognitive behaviour which can hardly be represented by these detailed and segmental dimensions, which indeed do not fulfil English learners and teachers' requirements. This work aims to utilise the state-of-the-art deep learning techniques to recognise English speaking proficiency at a cognitive level, i. e., a learner's ability to clearly organise their own thoughts when expressing an idea in English as a foreign language. For this, we collected the "Oral English for Japanese Learners" Dataset (OEJL-DB), a corpus of recordings by 82 students of a Japanese high school expressing their ideas in English towards 5 different topics. Annotations concerning the clarity of learners' thoughts are given by 5 English teachers according to 2 classes: clear and unclear. In total, the dataset includes 7.6 hours of audio data with an average length for each oral English presentation of66 seconds. As initial cognitive-based method to identify learners' speaking proficiency, we propose an architecture based on the parallelization of CNNs and Transformers. With the strengthening of the CNNs in spatial feature representation and the Transformer in sequence encoding, we achieve a 89.4% accuracy and 87.6%. Unweighted Average Recall (UAR), results which outperform those from the ResNet architectures (89.2 % accuracy and 86.3 % UAR). Our promising outcomes reveal that speech intelligence can be efficiently applied to "grasp" high level cognitive behaviours, a new area of research which seems to have a great potential for further investigation.

Keywords: English Speaking, transformer, CNNs

INTRODUCTION

The ultimate purpose of learning a language is to efficiently exchange information, a process that implies, besides the comprehension of others, the ability to express the own thoughts [1]. Expressing one's ideas in English (as a foreign language) strongly depends on each learner's cognitive style, i. e., every individual's capability to process and organize both information and experiences [9]. In other words, the ability of efficiently expressing ideas refers to humans' unique way of thinking, perceiving, remembering, and solving problems [9]. Thus, "Organizing the Own Thoughts Clearly or Not" is becoming an important evaluation factor for English speaking learning [10]. From the different modalities used by people to express their ideas, speech is one of the most natural and effective. This is in part due to the inherent capability of vocal expressions to convey speakers' subjective states, e. g., emotions [5], which enhance the communication process. Even though a variety of speech-driven applications have been presented in educational contexts [17] [16] [2] [7], the automatic assessment of a student capacity to organize the own thoughts while speaking in a foreign language has not yet been performed. Hence, the application of Speech Recognition Technology on such a task will be of great benefit for both learners and teachers, since it could be considered as an automatically generated indicator of student's proficiency. Recent advances in speech technology, mainly based on Automatic Speech Recognition (ASR), have been successfully applied in the identification of learner's spoken proficiency [13]. Indeed, a variety of scoring methods for the automatic measurement of student's performance in a foreign language have been proposed. For instance, speech representations including Mel-frequency ceptral coefficients (MFCCs), Gaussian posteriograms, and English phoneme state prosteriorgrams have been investigated [12]. Pronunciation-related methods based on a variety of features, including acoustic scores from a Hidden Markov Model (HMM), du-rations of words and phones, as well as information about pauses, prosody, and syllable structure, have also been proposed [4]. Using similar features, the Stanford Research Institute (SRI) introduced EduSpeak [6], an ASR-based system for the automatic assessment of pronunciation quality. The Educational Testing Service presented also an automatic assessment system, namely Speech Rater, whose goal, beyond text reading and repetition, was also to promote spontaneous spoken communication [20]. Finally, due to the influence of speakers' affective states in their vocal expressions, an action plan aimed to alleviate foreign language speaking anxiety—hence improving speaking performance—has also been investigated [1].

In this work we go beyond the current state-of-the-art in the topic by proposing a cognitive-based approach for the automatic recognition of foreign language spoken proficiency. Unlike the previous works, mainly based on the assessment of features related to prosody and phonetic aspects, we aim to recognise whether the oral English communication involved a high-level cognitive performance, i. e., clearly organized thoughts. For this, we utilise Convolutional Neural Networks (CNNs) combined with Transformer deep learning techniques. Although Trans-formers have the advantage of encoding the input data as powerful features via the attention mechanism (which enables a good modelling of the global context), they show limitations in capturing details, something that can be overtaken by the use of CNNs, which are suitable to capture local information. Indeed, to benefit from the advantages of the CNNs and Transformer, efforts on combining both have already been presented, as shown, e. g., by TransUnet [3] and Transfuse [21]. Inspired by this idea, we parallelised CNNs and Transformers for recognizing learners' English speaking performance.

DATABASE: OEJL-DB

To carry out our experiments, we collected the "Oral English for Japanese Learners" Dataset (OEJL-DB) in the Communication English course in a High School from Tokyo (Japan). A total of 82 students (41 female, 41 male) with ages from15 to 17 years old and English-speaking proficiency from low to high, participated in the study. During the English classes, teachers provided 5 topics, e. g.,"Cashless Society" or "Paperless Classroom", as well as related texts to be read. After reading the provided materials, the students had 30 minutes to discuss and write notes about each topic. Subsequently, at the end of the class, each student presented the argument in a short oral presentation of 1 to 3 minutes. The presentations were recorded by other students with mobile phones and tablets.

In total, the dataset contains 7.6 hours of audio data with an average length of 66 seconds per recording. Each presentation was manually annotated by five English teachers for labelling whether the learner organising standpoints clear by using the values of 0 and 1. After gathering all teachers' annotation values, we calculate the average score for each sample as the final label. The final labels are in two classes:

(1) Learners present standpoints clearly and organise their thoughts logically;

(2) Learners illustrate chaos ideas and there is no logical relation-ships among arguments.

DEEP LEARNING MODELS

Applications of CNNs have achieved excellent results in recent studies on speech processing [22]. Despite the immense success, CNN-based methods have also shown lack of efficiency in capturing global context information [21]. Existing works obtain global information by generating very large receptive fields [22], which requires consecutively down-sampling and stacking convolutional layers until a sufficient depth is achieved. However, this procedure presents several disadvantages: (1) training deep networks can be easily affected by the diminishing feature reuse problem, i. e., a phenomenon where low-level features are washed out by consecutive multiplications; (2) the spatial resolution of very deep net-works is reduced gradually; (3) deep networks are more unstable and easily tend to overfitting.

Transformer is a sequence-to-sequence architecture originally proposed for neural machine translation in natural language processing [11]. Yet, due to its strong ability to long-range modeling, Transformers have been successfully used in a variety of domains, including speech recognition tasks [19]. The self-attention mechanism in Transformers can dynamically adjust the receptive field according to the input content; hence, being superior, w. r. t. convolutional operations, in modelling the long-range dependency [19].

Due to the individual efficiency and to some extent complementarity of CNNs and Transformers, there has been increasing interest in incorporating them both into a single architecture. For instance, TransUnet [3] was the first which utilized CNNs to extract low-level features and subsequently passed them through Transformers in order to model global interactions. Nevertheless, previous works mainly focus on replacing convolutions with transformer layers or stacking the two in a sequential manner. Differently, in order to capture global dependency and low-level spatial details in shallower networks, we propose an architecture based on the parallelization of CNNs and Transformers.

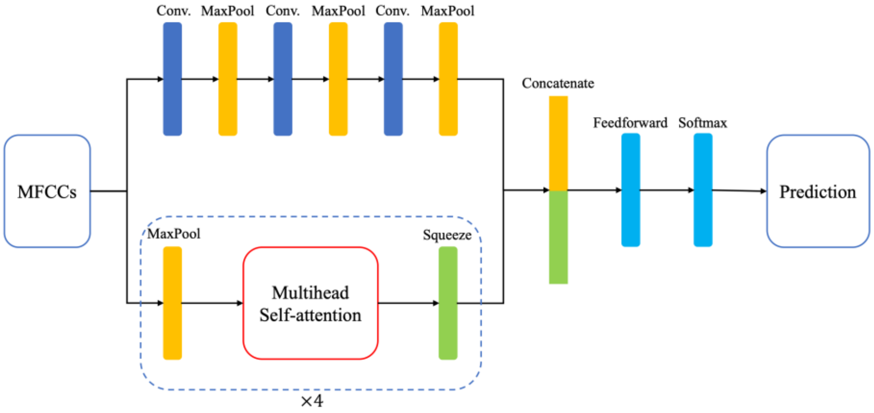

Figure 1. A graphical overview of our framework

The architecture of the proposed model (cf. Fig. 1) comprises four components:(i) extraction of MFCCs features in grayscale as input layer; (ii) 3 layers of CNNs embeddings with maxpooling layers after each convolutional layer; (iii) 4 layers of Transformer embeddings; (iv) combination of all embeddings into a softmax layer aimed to make the final prediction.

The CNN-Embeddings contain 3 layers of 2D blocks. Following VGGNet [14],in order to achieve better performance, we used fixed sized kernels (3X3) through-out deeply stacked CNN layers. In addition, in the first layer of each convolutional block, we used a maxpool kernel size of stride 2. Each layer was followed by batch normalization and ReLU activation.

The design of Transformer-Embeddings follows the typical transformer encoder architecture [18], i. e., the encoders is composed with identical layers with multi-head self-attention mechanism. We firstly maxpool the input MFCC map to the transformer block to considerably reduce the number of parameters the network needs to learn. Then, we define four transformer encoder layers by ap-plying the four multi-head self-attention layers of the transformer enable the network to look at multiple previous time steps when predicting the next.

RESULTS

To carry out the experiments, we performed an independent seven-fold Leave-One-Speaker-Out (LOSO) strategy, as outlined in previous work [8]. Due to the relatively low amount of speakers in the OEJL-DB dataset, we kept only 12 for testing and used the remaining 70 in the training and validation sets. In each fold, all instances of 10 speakers were kept for evaluation. To evaluate the prediction results, standard metrics will be reported: Unweighted Average Recall (UAR) and accuracy. In order to create a baseline for comparison, we also supply results obtained from classical Residual Networks, i. e., ResNet-18, ResNet-34, ResNet-50 and WideResNet-50, chose as they have shown good performance in speech-driven applications on a similar user group [15].

Our experimental results indicate that the proposed approach outperforms classical ResNet architectures (cf. Table 1). From the baseline approaches, ResNet-50 achieved best Accuracy (89.2%) and UAR (86.3%), which indicates that with the deeper layers of ResNet architectures, the models achieved better efficiency in presenting suitable data representations. However, when the feature map's width of ResNet-50 was doubled (WideResNet-50), there was a detriment in the performance, which is probably due to the fact that a wider feature map increases the model complexity, hence yielding to overfitting. From the proposed models, i. e., the paralized architectures, considering two CNNs blocks combined with Transformer achieved the best Accuracy (89.4%) and UAR (87.6%).

The main reason to explain why the proposed combined model outperforms the classical ResNet architectures is that the long average length of the samples makes the transformer block (missing in the ResNet architectures) an essential component to capture the global information of each sample. Furthermore, we interpret the superiority of using two blocks of CNNs w. r. t. one, to the more detailed in-formation from MFCCs capture by the former w. r. t. the latter. These results confirm the potential of combined architectures and demonstrate at the same time that they can be successfully used to predict English proficiency from a cognitive perspective.

Table 1: Evaluation Accuracy and Unweighted Average Recall (UAR) [%] for the proposed models: combination of CNN and Transformer; combination of two CNNs and Transformer. As a baseline for comparison, we also supplied results from classical CNNs architecures, i. e., ResNet-18, ResNet-34, ResNet-50, and WideResNet-50. Results are presented for the LOSO evaluation and Test sets; the best results in test sets are highlighted in bold.

Methods	Accuracy		UAR	
	LOSO	Test	LOSO	Test
ResNet-18	85.2	86.1	69.8	68.4
ResNet-34	88.9	88.7	79.2	75.5
ResNet-50	90.7	89.2	87.5	86.3
WideResNet-50	88.9	86.2	79.1	78.8
CNN+Transformer	88.9	87.1	71.8	74.5
CNN+CNN+Transformer	92.6	**89.4**	88.5	**87.6**

CONCLUSIONS

In this paper we proposed a combined CNNs and Transformer architecture forautomatic recognition of learners’ English proficiency. Compared to traditional ResNet, our proposed architecture takes full advantage of long-range and local information (captured through the paralellization of CNNs and Transformers),which yields to a superior performance. For future studies, combining ResNet and Transformer as one architecture is deserved to be explored. We also plan to investigate to which extent different time length of features of CNNs architectures might impact the achieved results.

REFERENCES

Alkan, H., B¨umen, N.T.: An action research on developing english speaking skills through asynchronous online learning. International Journal of Curriculum andInstruction12(2), 127-148 (2020)

Bahreini, K., Nadolski, R., Westera, W.: Towards real-time speech emotion recognition for affective e-learning. Education and information technologies21(5), 1367-1386 (2016)

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou,Y.: Transunet: Transformers make strong encoders for medical image segmentation.arXiv preprint arXiv:2102.04306 (2021)

Cucchiarini, C., Strik, H., Boves, L.: Automatic evaluation of dutch pronunciation by using speech recognition technology. In: Proc. workshop on automatic speech recognition and understanding proceedings. pp. 622-629 (1997)

Darwin, C.: The expression of the emotions in man and animals. John Murray,London, UK (1872)

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier,R., Cesari, F.: The sri eduspeaktm system: Recognition and pronunciation scoringfor language learning. STILL2000, 123-128 (2000)

Han, T., Zhang, J., Zhang, Z., Sun, G., Ye, L., Ferdinando, H., Alasaarela, E.,Sepp¨anen, T., Yu, X., Yang, S.: Emotion recognition and school violence detec-tion from children speech. EURASIP Journal on Wireless Communications andNetworking2018(1), 1-10 (2018)

Hantke, S., Schmitt, M., Tzirakis, P., Schuller, B.: Eat- the icmi 2018 eating anal-ysis and tracking challenge. In: Proc. International Conference on Multimodal Interaction. pp. 559-563 (2018)

Ho, S.C., Hsieh, S.W., Sun, P.C., Chen, C.M.: To activate english learning: Listennand speak in real life context with an ar featured u-learning system. Journal ofnEducational Technology & Society20(2), 176-187 (2017)

Ihsan, M.D.: Students’ motivation in speaking english. Journal of English Educators Society1(1) (2016)

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M.,Soplin, N.E.Y., Yamamoto, R., Wang, X., et al.: A comparative study on trans-former vs rnn in speech applications. In: Proc. Automatic Speech Recognition and Understanding Workshop. pp. 449-456 (2019)

Lee, A., Glass, J.: Pronunciation assessment via a comparison-based system. In:Speech and Language Technology in Education (2013)

Luo, D., Xia, L., Zhang, C., Wang, L.: Automatic pronunciation evaluation in high-states english speaking tests based on deep neural network models. In: Proc. In-ternational Conference on Artificial Intelligence and Big Data. pp. 124–128 (2019)

Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K.: Going deeper in spiking neuralnetworks: Vgg and residual architectures. Frontiers in neuroscience13, 95 (2019)

Song, M., Mallol-Ragolta, A., Parada-Cabaleiro, E., Yang, Z., Liu, S., Ren, Z.,Zhao, Z., Schuller, B.W.: Frustration recognition from speech during game inter-action using wide residual networks. Virtual Reality & Intelligent Hardware3(1),76–86 (2021)

Song, M., Yang, Z., Baird, A., Parada-Cabaleiro, E., Zhang, Z., Zhao, Z., Schuller,B.: Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database. In: Proc. Affective Computing and Intelligent Interaction. pp. 517–523. Cambridge, the UK (2019)

Song, Z.: English speech recognition based on deep learning with multiple features.Computing102(3), 663–682 (2020)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. Advances in neural informa-tion processing systems. pp. 5998–6008 (2017)

Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv preprint arXiv:2103.03024 (2021)

Zechner, K., Higgins, D., Xi, X., Williamson, D.M.: Automatic scoring of non-native spontaneous speech in tests of spoken english. Speech Communication51(10), 883–895 (2009)

Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005 (2021)

Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., Schuller, B.W.:Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. Neural Networks141, 52–60 (2021)