A Speech Recognition Model for Mandarin Chinese Pronunciation Evaluation

1st Jakub Kiliańczyk
Gdańsk University of Technology
Gdańsk, Poland

2nd Jakub Kwiatkowski
Gdańsk University of Technology
Gdańsk, Poland

3rd Anna Strzelecka
Gdańsk University of Technology
Gdańsk, Poland

4th Dawid Migowski
Gdańsk University of Technology
Gdańsk, Poland

5th Łukasz Smoliński
Gdańsk University of Technology
Gdańsk, Poland

***Abstract*—In recent years many Automatic Speech Recognition (ASR) solutions have been designed based on artificial intelligence algorithms. The Chinese language poses relatively new challenges when it comes to ASR or Mispronunciation Detection and Diagnosis (MDD) as the meaning of each word depends not only on pronunciation, but also on tone of speech. In order to correctly asses one's speech, both features have to be considered.**

***Index Terms*—Automatic Speech Recognition, Mispronunciation Detection and Diagnosis, Convolutional Neural Network, Transformer**

I. Introduction

In order to train new Speech Recognition models, we used AISHELL-3 [1] dataset and a self-made one consisting of our university's students' recordings. We developed and compared several approaches for the tasks of Speech Recognition and Scoring and Tone Recognition.

II. Related Work

We conducted a literature review in search of existing research and solutions for the tasks of pronunciation and tone detection for Mandarin Chinese. We assumed criteria to look for Neural Network models trained with limited data, published in years 2014-2024 in english. We based our research on the following keywords: "Speech Recognition", "Deep Learning", "Pronunciation Evaluation" and fetched a total of 878 research papers from several databases. After eliminating duplicates and assessing the titles and abstracts we chose 12 most suiting articles for further reading.

After the literature review, our first notion was the limitation of publicly available datasets, especially regarding the subject of Tone Classification (TC). [2] mentions difficulties in finding and accessing datasets for ASR in Chinese and suggests a new one for Keyword Detection.

III. Approach

We decided to first develop separate CNN models for the tasks of Pronunciation Evaluation (PE) and Tone Classification. Afterwards we attempted to solve both of these tasks using one, more complex architecture.

A. Data Processing

In order to ensure good training quality of our dataset we applied some preprocessing techniques. Firstly, we removed the incorrect recordings, for example only constituting of noise, or ones that didn't suffice the length requirements. Later, we trimmed the silent parts at the edges of the recordings and applied noise-reduction filters. Finally, we normalized the volume to achieve voice intensity consistency.

We split the dataset into training, validation and test sets and balanced representation of positive and negative cases for each class. We conducted data augmentation on the training set.

After obtaining the spectrograms required as an input for our models, we observed a significant improvement in the training.

B. CNN for Tone Recognition

We have prepared a multilayer CNN to classify the tone of speech phonemes into one of 4 defined tones (1 to 4, not counting the neutral one, denoted by 0). It consists of convolution layers, batch normalization, activation functions and pooling layers. The input layer takes two-dimensional data and passes the output to 4 convolution blocks with the number of filters ascending through values: 64, 128, 256, 512 and altering steps, which allow hierarchical feature extraction. Finally we apply Data Flattening combined with 2 Fully Connected Layers (FCLs) of 1024 units followed by a softmax layer to obtain probabilities. For this model we use the Stochastic gradient descent optimization algorithm [3] with a low learning rate and special weights applied for classes with inconsistent representation in training data. We used techniques of early stopping and saving checkpoints to avoid overfitting and ensure the best results are not lost.

C. CNN for Pronunciation Evaluation

For this task we used the numeric words from 0 to 10 and 100 from our Mandarin Chinese audio dataset with labels provided by a native Chinese speaking expert in the field. The network takes a 3-channel spectrogram as an input. Similarly to our Tone Recognition model, the architecture is based on convolution and pooling layers. The number of convolution filters grows to 128. Thanks to the MaxPooling layers, the model can be trained to focus on the meaningful parts of the signal, ignoring the noise. After extracting the features with convolutional blocks, we use a Flatten layer. Afterwards the tensors are passed to FCLs with 1024 and 128 units in this order and ReLU activation function. In order to avoid overfitting we use a Dropout layer that randomly disables half of the neurons, improving the model's generalization capability. In the end there is a 1-unit FCL with a sigmoid activation function, responsible for outputting a probability of given word, allowing binary classification. During the training we use the Adam optimizer [4] with Binary Cross-Entropy loss function.

D. *Speech-Transformer for complete ASR*

With this last model, using an implementation of [5], [6], modified to enable training with [1] and finetuning with our dataset, we attempted solving both tasks of Pronunciation Evaluation and Tone Classification. The model would originally generate transcription sequences based on processed audio input. In case of Mandarin Chinese, the transcription can be represented in traditional alphabet or in pinyin notation, which consists of syllables written with latin letters and denotes the tone of each phone. The pinyin is popular for learning purposes for foreigners who don't know the Chinese alphabet. It also enables us to simply split the phoneme into pronunciation and tone information. The model has been configured with default hyperparameter values as in [6]. Due to lack of specific transcriptions in our dataset (we were limited to pronunciation scores and tone labels) the training split for the finetuning part had to be limited to correct recordings. For the testing, however, we used balanced amount of correct and incorrect recordings for pronunciation evaluation and all the recording for tone classification.

IV. Results

A. *CNN for Pronunciation Evaluation*

This paragraph will be filled in the final version of this work.

B. *Speech-Transformer - Pronunciation Evaluation*

We measured this model with precision, recall and F1-score metrics. The results are given in the table "Tab. I". We have also plotted a confusion matrix "Fig. 1". For this task, we balanced test data to 477 correct and incorrect samples.

TABLE I
Speech-Transformer - Pronunciation Evaluation Metrics

Metric	Value
true positive	456
false positive	266
true negative	211
false negative	21
precision	0.63
recall	0.96
F1-score	0.76

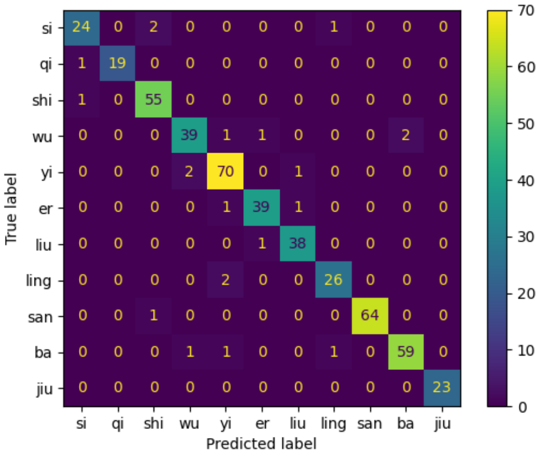

Fig. 1. Speech-Transformer - Confusion Matrix for PE

C. *Speech-Transformer - Tone Classification*

Before the finetuning, this model would achieve at most 50% accuracy, meaning half of the times it would guess one of the 4 available tones (excluding the neutral tone). Afterwards it actually dropped below that value, because it was not trained separately for this task and two tones out of four are underrepresented within the numeric words from 1 to 10, which can be observed on a confusion matrix "Fig. 2".

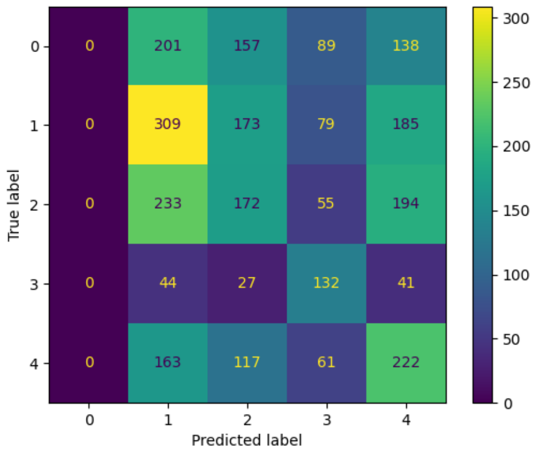

Fig. 2. Speech-Transformer - Confusion Matrix for TC

V. Discussion

This paragraph will be filled in the final version of this work.

VI. Conclusion

This paragraph will be filled in the final version of this work.

References

[1] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “Aishell-3: A multi-speaker mandarin tts corpus and the baselines,” 2021. [Online]. Available: https://arxiv.org/abs/2010.11567

[2] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” 2018. [Online]. Available: https://arxiv.org/abs/1804.03209

[3] W. Sung, I. Choi, J. Park, S. Choi, and S. Shin, “S-sgd: Symmetrical stochastic gradient descent with weight noise injection for reaching flat minima,” 2020. [Online]. Available: https://arxiv.org/abs/2009.02479

[4] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[6] L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.