System Description for Team CZUR (T018)

Yang Hu, Haowei Li

CZUR

{huyang, lihaowei}@czur.com

Abstract

The report describes the CZUR (T018) system for the Low-Resource Dysarthria Wake-Up Word Spotting (LRDWWS) Challenge 2024. The proposed system includes a feature extractor and a wake-up word detection (WWD) model. We adopt a pretrained automatic speech recognition (ASR) encoder as feature extractor to facilitate better feature extraction under the low resource setting of the competition. As for the WWD model, we have tested several options and our final choice is a lightweight E-Branchformer. Moreover, we employ pretraining and extra publicly-available datasets to further boost our model's performance. The WWD model is pretrained using data2vec with both normal speech data and dysarthria speech data, before it is finetuned on the dysarthria data. We expect the pretraining process helps the model learn better feature representations under the low resource setting. The proposed system achieves a score of 0.010533 in the leaderboard-B of the competition, ranked 3rd among 7 teams.

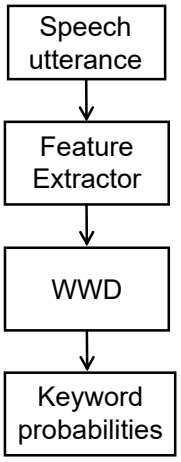

Figure 1: The CZUR system for LRDWWS 2024.

1. Model Architecture

The proposed model includes a pretrained feature extractor and a wake-up word detection (WWD) model. The feature extractor extracts features from input audios. The WWD model takes the features produced by the feature extractor as input, and it outputs 10 probabilities corresponding to the 10 wake-up words of the competition. Fig. 1 depicts the proposed system.

We adopt the encoder of the Paraformer trained on 60,000 hours of Mandarin speech data as our feature extractor [1, 2]. Given the low resource setting of the competition, we expect that a much larger model pretrained on huge general speech data outperforms a small model trained from scratch using the competition data only. In other words, we believe the advantage of model size and data amount are very likely to significantly outweigh the drawbacks of domain discrepancy in the setting of the competition.

As for the WWD model, we have tested three options: multi-scale depthwise temporal convolution (MDTC) [3, 4], squeezeformer [5, 6], E-Branchformer [7, 8]. We adjust these models' architectures to reduce the number of parameters, so that they can train well on the competition data without overfitting. This actually leads to three lightweight WWD models. Among the three models, E-Branchformer shows superior performance in our experiments, so we select this model as our final WWD model choice.

Instead of training subject-specific systems as in the competition baseline [9], we design the proposed system to be subject-independent. A single system is used to classify keywords and non-keywords for all the subjects, trained using data from all the subjects (see the sections below). We think such a subject-independent training strategy could better leverage the limited training data. And a subject-dependent strategy might need more data to realise its advantage of personalized modeling and decision.

2. Datasets

To alleviate the influence of low resource setting of the competition on the model performance, we employ two more datasets to train our system in addition to the official LRDWWS data (referred to as LRDWWS in the rest of the report). We use the Chinese Dysarthria Speech Database (CDSD) [10] as additional dysarthria data. We also select more than 2000 hours of speech segments from WenetSpeech [11] for training (referred to as WenetSpeech in the rest of the report). Despite of domain discrepancy, we expect that richer speech data helps to better learn speech structures, and these structures from general speech data could be transferable to dysarthria data.

We adopt different training data in different phases of the competition. Before the leaderboard-A period (i.e. before the leaderboard-A data is released), we use WenetSpeech, CDSD, full Train split of LRDWWS and the eval subset of the Dev split of LRDWWS as the training data. In the leaderboard-A period, we use WenetSpeech, CDSD, full Train split of LRDWWS, full Dev split of LRDWWS, and Test-A enrollment split of LRDWWS for training. In the leaderboard-B period, we use WenetSpeech, CDSD, and full Train split of LRDWWS, full Dev split of LRDWWS, full Test-A split of LRDWWS, and Test-B enrollment splits of LRDWWS for training. Note that we only use the dysarthria speech in LRDWWS. We ignore the normal speech data in this dataset.

As for validation, we use the enrollment subset of the Dev split of LRDWWS as the validation data before the leaderboard-

A period. In the leaderboard-A period, we does not retain local validation data; we tune hyper-parameters based on our submission scores on the leaderboard-A. We adopt a similar strategy for the leaderboard-B period, during which we tune several few hyper-parameters like the decision threshold based on leaderboard-B scores. Moreover, in the leaderboard-B period, we perform local evaluations of the models trained in the leaderboard-A period on the Test-A eval split of LRDWWS. This local evaluation helps us gain more insights into the effects of some hyper-parameters, providing more information for the hyper-parameter selection for our leaderboard-B submissions.

3. Model Training

We adopt a three-step training process for our system. Note that we freeze the feature extractor throughout our training process, and we only update the WWD model parameters. Also, we would like to state that this section only describes what datasets are used for training. The actual splits of dataset used in training vary in different competition periods and follow our description in Section 2.

First, we use WenetSpeech, CDSD and LRDWWS to pretrain our model in a self-supervised way. We choose data2vec [12, 13] as our pretraining algorithm. The reason is two-fold. On the one hand, data2vec is a recently proposed algorithm that achieves competitive performance in various speech-related tasks [14]. On the other hand, a previous work [13] has explored data2vec to pretrain keyword spotting systems. This work makes data2vec a promising foundation for our experiments, easing the effort of hyper-parameter selection.

Next, we finetune the WWD model trained in the first step on CDSD and LRDWWS data, still using self-supervised data2vec algorithm. We think this step could transfer the model learned on general speech data to a model specific to dysarthria speech structures. Due to the limited amount of dysarthria data, we believe a model can better learn dysarthria speech structures if it is sufficiently trained to explore general speech structure before it focuses on dysarthria speech.

Finally, we finetune the pretrained model in the second step on LRDWWS for keyword spotting, using the keyword labels in this dataset in a fully supervised manner. In this step, we follow the training pipeline of WeKws [4] with its alignment-free max-pooling loss. Different from the original WeKws implementation, we adopt a cosine annealing learning rate decay with warmup. And we bypass the last few iterations of an epoch, as the batch size of these iterations can be small with the WeKws dataloading pipeline. We think these small batches might lead to noisy gradients and harm the model convergence, especially when the training data is limited and the number of iteration is small in each epoch.

Note that our model is trained with all the three steps in the leaderboard-A period of the competition. As for the leaderboard-B period, we only do the third step due to limited training time. Specifically, we obtain multiple data2vec pretrained models in the second step in the leaderboard-A period. We select the one with the best leaderboard-A scores and finetune it on labeled LRDWWS data for leaderboard-B keyword classification.

4. Results

The proposed system has attained a leaderboard-B score of 0.010533, with a FAR of 0.003033 and a FRR of 0.007500. It is ranked the third place among the seven teams in the leaderboard B. Besides, the proposed system is rank 1st place among the seven teams in the leaderboard-A, with a FAR of 0.0042, a FRR of 0.0225 and a final score of 0.0267.

5. References

[1] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *Proc. Interspeech 2022*, 2022, pp. 2063–2067.

[2] https://github.com/modelscope/FunASR.

[3] J. Hou, L. Xie, and S. Zhang, "Two-stage streaming keyword detection and localization with multi-scale depthwise temporal convolution," *Neural Networks*, vol. 150, pp. 28–42, 2022.

[4] J. Wang, M. Xu, J. Hou, B. Zhang, Z. X., L. Xie, and P. F., "Wekws: A production first small-footprint end-to-end keyword spotting toolkit," in *Proc. ICASSP*, 2023, https://github.com/wenet-e2e/wekws.

[5] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 9361–9373.

[6] https://github.com/wenet-e2e/wenet.

[7] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. Han, and S. Watanabe, "E-Branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. IEEE SLT Workshop*, 2022.

[8] https://github.com/espnet/espnet.

[9] https://github.com/greeeenmouth/LRDWWS.

[10] M. Sun, M. Gao, X. Kang, S. Wang, J. Du, D. Yao, and S. Wang, "CDSD: Chinese dysarthria speech database," *arXiv:2310.15930*, 2023.

[11] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *Proc. ICASSP*, 2022.

[12] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 1298–1312.

[13] H. Bovbjerg and Z. Tan, "Improving label-deficient keyword spotting through self-supervised pretraining," in *Proc. ICASSP Workshop*, 2023, https://github.com/HolgerBovbjerg/data2vec-KWS.

[14] S. Yang, H. Chang, Z. Huang, A. Liu, C. Lai, H. Wu, J. Shi, X. Chang, H. Tsai, W. Huang, T. Feng, P. Chi, Y. Lin, Y. Chuang, T. Huang, W. Tseng, K. Lakhotia, S. Li, A. Mohamed, S. Watanabe, and H. Lee, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–16, 2024.