THE XIAOMI-NPU-ASLP SYSTEM FOR SLT2024 LOW-RESOURCE DYSARTHRIA WAKE-UP WORD SPOTTING CHALLENGE

Shuiyun Liu[1], Yuxiang Kong[2*], Pengcheng Guo[1], Weiji Zhuang[2], Peng Gao[2], Yujun Wang[2], Lei Xie[1]*

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Speech Group, AI Lab, Xiaomi Inc., China

ABSTRACT

This paper presents an end-to-end Pretrain-based Dual-filter Dysarthria Wake-up word Spotting (PD-DWS) system developed by the Xiaomi-NPU-ASLP team (T011) for the SLT 2024 Low-Resource Dysarthria Wake-Up Word Spotting Challenge. Specifically, our system improves performance from two perspectives: audio modeling and dual-filter strategy. For audio modeling, we propose an innovative 2branch-d2v2 model based on the pre-trained data2vec2 (d2v2), which can simultaneously model automatic speech recognition (ASR) and wake-up word spotting (WWS) tasks through a unified multi-task fine-tuning paradigm. Additionally, a dual-filter strategy is introduced to reduce the false accept rate (FAR) while maintaining the same false reject rate (FRR). Experimental results demonstrate that our PD-DWS system achieves a FAR of 0.003210 and an FRR of 0.005, with a total score of 0.00821 on the test-B eval set, securing first place in the track.

***Index Terms*—** LRDWWS, 2brach-d2v2, dual-filter

1. INTRODUCTION

Speech has become a natural user interface with broad applications. However, individuals with dysarthria face difficulties due to the variability in their speech. To address this, the IEEE SLT 2024 Workshop has initiated the Low Resource Dysarthric Wake Word Spotting (LRDWWS) Challenge.

This study describes our participation in the LRDWWS challenge, focusing on the development of a dysarthric wake word system named Pretrain-based Dual-filter Dysarthria Wake-up word Spotting (PD-DWS). Our efforts encompass two directions: audio modeling and a dual-filter strategy. Firstly, in the audio modeling part, we introduce an innovative 2branch-d2v2 model by fine-tuning the pre-trained data2vec2 (d2v2) [1] model within a multi-task framework, which simultaneously models both automatic speech recognition (ASR) and wake-up word spotting (WWS). Next, a dual-filter module is proposed to process the model outputs. Specifically, the output from the WWS branch is sent to the threshold filter, while the ASR branch output is sent to the ASR filter for filtering. The threshold filter performs initial filtering on the wake-up word probabilities, preliminarily determining the audio's predicted label. The ASR filter then performs secondary filtering using the ASR output from the model as well as ASR results obtained from the fine-tuned Paraformer. Combining the aforementioned strategies, our proposed PD-DWS achieves a false accept rate (FAR) of 0.00321 and a false reject rate (FRR) of 0.005, with total scores of 0.00821 on the test-B eval set in this Challenge.

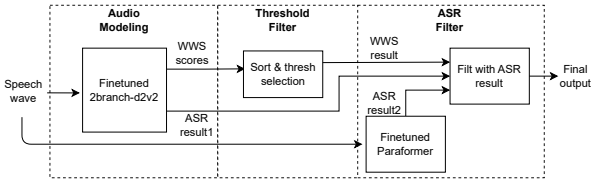

Fig. 1. An overview of our proposed PD-DWS system.

2. PROPOSED SYSTEM

Fig. 1 overviews our proposed PD-DWS system which comprises an audio modeling and a dual-filter, which includes a threshold filter and an ASR filter.

2.1. Audio Modeling

In the audio modeling part, we experiment with both the Conformer encoder and a novel 2branch-d2v2 encoder. Here, we primarily introduce the 2branch-d2v2 approach. As shown in Fig 2, the 2branch-d2v2 encoder is initialized with a pre-trained d2v2 model and fine-tuned with a multi-task framework. The output of d2v2 is directed into two branches: one for ASR and the other for WWS. The WWS branch employs the official max pooling loss, $\mathcal{L}_{\text{WWS}}$, while the ASR branch adopts the CTC loss, $\mathcal{L}_{\text{CTC}}$. The final training loss is computed as $\mathcal{L} = 0.5 \cdot \mathcal{L}_{\text{CTC}} + \mathcal{L}_{\text{WWS}}$. Besides, dynamic augmentation techniques are applied, including volume augmentation, MUSAN [1] noise addition, and speed perturbation. We follow the official training data flow [2], but in the third step, we use all enrollment sets instead of separate sets for each individual.

* stand for equal contribution

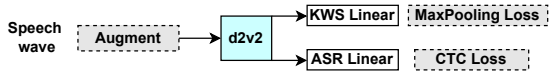

Fig. 2. Details of the 2branch-d2v2 encoder.

2.2. Dual-Filter: Threshold Filter

The threshold filter module receives the probabilities of the ten wake-up words from the WWS branch. For each audio, the highest probability is used as the temporal score, and the corresponding label is assigned as the temporal label. The audios are then sorted in descending order based on their temporal scores. The module iterates through all the ranks of temporal scores to establish provisional thresholds. For each audio with a score lower than the threshold, the temporal label is changed to a filter label, while the labels for the other audio remain unchanged. Experiments on the test-A eval set have shown that selecting the score at 60th rank as the threshold yields the best performance.

2.3. Dual-Filter: ASR Filter

The ASR filter module corrects the WWS results from the previous step based on the ASR output. In this module, we utilize two ASR results for comparison. One is obtained through the beam search decoding of the ASR branch, and the other is derived from the ASR results of the open-source Paraformer large model [3] that fine-tuned with competition data and TTS-synthesized speech. In response to the WWS results from the previous step, if a wake-up word's length matches the length of any ASR result, we keep this result. Conversely, if the wake-up word's length does not match any ASR results, we consider the final result to be filler. The TTS data used for training Paraformer is generated by VITS [4], which is trained with the LRDWWS train set. Table 1 details the datasets used in training each module.

Table 1. Details of the training data used for each module.

Module	Training data
pretrained_d2v2	LibriHeavy, GigaSpeech, WenetSpeech, Aishell [5] {1,2}, ACAV100M, OpenSLR [6] {38,47,68,82,87,111, 119,123,124,133}, CommonVoice, LRDWWS training set
finetuned_2branch-d2v2	LRDWWS training set LRDWWS enrollment set
TTS-generator	LRDWWS training set

[1]https://www.openslr.org/17/
[2]https://github.com/greeeenmouth/LRDWWS
[3]https://github.com/modelscope/FunASR
[4]https://github.com/jaywalnut310/vits
[5]https://www.aishelltech.com/kysjcp
[6]https://www.openslr.org/resources.php

Table 2. Performance of different systems on test-B-eval set.

System	Thresh rank	Score	FAR	FRR
baseline	–	0.130306	0.028639	0.101667
ASR result2	–	0.021101	**0.001601**	0.019500
2branch-d2v2	55	0.014287	0.004287	0.010000
+ASR filter	60	0.010822	0.003322	0.007500
+test-A eval set	60	**0.008210**	0.003210	0.005000

3. EXPERIMENT

3.1. Configuration

The model configurations used in the experiments are as follows. The Conformer encoder has 12 layers, 4 attention heads, 256 hidden dimensions, and approximately 31M parameters. For the d2v2 pertaining, we use the same configuration as the d2v2 large model [7], which has about 300M parameters. Gradient accumulation is set to 6 and training is conducted on 8 A800 GPUs. For our proposed 2branch-d2v2 fine-tuning, the model is trained on 2 A800 GPUs with a dynamic batch size, counting about 300 seconds. The learning rate for the d2v2 part reaches a maximum of 0.00005 with 1600 steps, while the other parts follow a schedule of 0.001 with 450 steps. The vocabulary used in the ASR branch is derived from the text in the LRDWWS training set, using characters for Chinese and letters for English, totaling 451 units, including <blank>, <unk>, <sos/eos>.

3.2. Experiments with PD-DWS

Firstly, we conduct experiments with the baseline models on the test-A set. The results show that 2branch-d2v2 achieves a FAR of 0.0053 and an FRR of 0.0525, significantly outperforming the conformer, which has a FAR of 0.0183 and an FRR of 0.0825. Then, Table 2 presents the results of our PD-DWS system on the test-B eval set. In the second row of the table, we directly input the audio into the Paraformer. Based on its output ASR result2, if the result matches a word in the wake-up word list, we select the corresponding label; otherwise, we consider it a filler. This approach results in a low FAR and a relatively high FRR. With the help of the ASR filter module, our system achieves optimal performance. The last row shows that incorporating the test-A eval dataset into the second phase of fine-tuning yields the best results.

4. CONCLUSION

This paper introduces the PD-DWS system developed by Xiaomi-NPU-ASLP team (T011) for the 2024 LRDWWS Challenge. Our system achieves a FAR of 0.003210 and an FRR of 0.005 in the evaluation set of this challenge, ranking first in the competition.

5. REFERENCES

[1] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.

[7]https://github.com/facebookresearch/fairseq