SLT 2024 CHALLENGE: LOW-RESOURCE DYSARTHRIA WAKE-UP WORD SPOTTING

Mingru Yang, Yongqiang Chen

School of Electronic and Information Engineering,
South China University of Technology,
Guangzhou, China

ABSTRACT

This technical report describes our proposed system for SLT 2024 Low-Resource Dysarthria Wake-Up Word Spotting (LRDWWS) Challenge. In our system, we propose Frame-Wise Normalization (FWN) Layer and Learned Feature Argumentation (LFA) Layer to replace the Batch Normalization layers in TCN. The FWN layer facilitates the network to learn dysarthria-invariant features and the LFA layer improves the network's generalization ability. To meet the complete low-resource requirements, we didn't use any additional data and the number of our system's parameter is 290,578. When evaluated on the test-a, our system outperforms the baseline by 0.1533.

Index Terms— wake-up word spotting, dysarthria, Frame-Wise Normalization, Feature Argumentation

1. INTRODUCTION

With the growing prevalence of speech-enabled applications, smart home technology has become a common feature in many households. For the general people, waking up and controlling smart devices is no longer a difficult task. However, dysarthric speakers encounter significant challenges in utilizing these technologies because they often experience pronounced difficulties in articulation, fluency, speech rate, volume, and clarity.

The primary challenges are as follows: First, individuals with dysarthria often have trouble in recording lengthy audio due to physical limitations, resulting in a limited number of speech samples available for registration. Second, inaccurate pronunciation and reduced fluency in dysarthric individuals lead to significant variability in their speech. Even among individuals with the same type of dysarthria, the pronunciation of a single word can vary under different conditions.

2. METHOD

To overcome the challenges of limited samples and high variability in dysarthric speech, we aim for our model to learn dysarthria-invariant features during the training phase. Additionally, we strive for the model to have strong generalization capabilities, achieving good test results with only a small number of registration samples from dysarthric speakers. Therefore, we proposed the Frame-Wise Normalization Layer and the Learned Feature Augmentation Layer to replace the Batch Normalization layers in TCN. Fig 1 illustrates our system diagram, encompassing the overall system architecture and the frameworks of the Frame-Wise Normalization Layer and the Learned Feature Augmentation Layer.

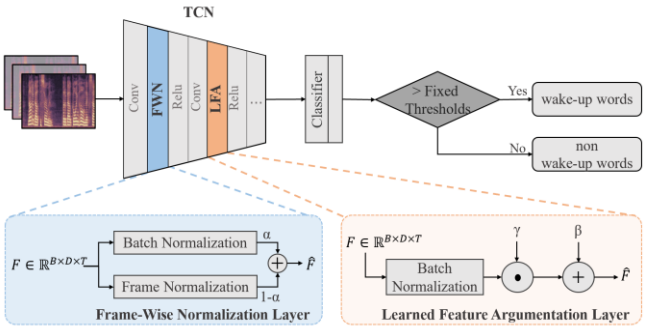

Fig 1: Diagram of Our System's Method

2.1. Frame-Wise Normalization Layer

Conventional Batch Normalization Layer constrains feature distributions to a standard space by normalizing the features of the entire batch. Let $F \in \mathbb{R}^{B\times D\times T}$ denote the input feature of certain layer for easy notation, batch normalization operation in Fig. 1 has the following form:

$$\boldsymbol{F}_{BN} = \frac{\boldsymbol{F} - \mu_{BN}}{\sqrt{\sigma_{BN}^2 + \epsilon}}$$

$$\mu_{BN} = \frac{1}{B \cdot D}\sum_{b=1}^{B}\sum_{d=1}^{D}\boldsymbol{F}_{b,d,t} \in \mathbb{R}^{T}$$

$$\sigma_{BN}^2 = \frac{1}{B \cdot D}\sum_{b=1}^{B}\sum_{d=1}^{D}(\boldsymbol{F}_{b,d,t} - \mu_{BN})^2$$

For dysarthria wake-up word spotting, it is a fact that individuals with dysarthria have substantial variability in their speech. Even among individuals with the same type of dysarthria, the pronunciation of a single word can vary under different conditions.

Consequently, we propose the Frame Normalization Layer, which normalizes each frame within a sample individually.

This layer adjusts the sample's feature style without being influenced by other sample's frames in the batch, facilitating the network in finding and learning dysarthria-invariant features. Frame normalization operation in Fig. 1 has the following form:

$$\boldsymbol{F}_{FN} = \frac{\boldsymbol{F} - \mu_{FN}}{\sqrt{\sigma_{FN}^2 + \epsilon}}$$

$$\mu_{FN} = \frac{1}{D}\sum_{d=1}^{D} \boldsymbol{F}_{b,d,t} \in \mathbb{R}^{B\times T}$$

$$\sigma_{BN}^2 = \frac{1}{D}\sum_{d=1}^{D} (\boldsymbol{F}_{b,d,t} - \mu_{BN})^2$$

In summary, the Batch Normalization Layer constrains feature distributions to a standard space, reducing internal covariate shift. The Frame Normalization Layer adjusts the style of individual sample frames, facilitating the network in finding dysarthria-invariant features. To balance the effects of both, we introduce a learnable balance weight α. The output is as follows:

$$\widehat{\boldsymbol{F}} = \alpha\, \boldsymbol{F}_{BN} + (1-\alpha)\, \boldsymbol{F}_{FN}$$

2.2. Learned Feature Argumentation Layer

Considering that the target dysarthric speaker samples are scarce during the training stage, and most of the training samples are from non-dysarthric speakers or other dysarthric speakers, we propose the Learned Feature Augmentation Layer to prevent overfitting on the training samples. This layer integrates feature transformations into the Batch Normalization Layer to enhance intermediate feature activations. Intuitively, this operation can generate more diverse feature distributions, thereby improving the network's generalization ability.

As shown in Fig 1, we insert the feature transformation operation after the Batch Normalization Layer, the learnable parameters $\theta_\gamma \in \mathbb{R}^{D\times 1}$ and $\theta_\beta \in \mathbb{R}^{D\times 1}$ indicate the standard deviations of the Gaussian distributions for sampling the affine argumentation parameters. Given an intermediate feature activation map Z in the feature encoder with the dimension of $D \times T$, we first sample the scaling term γ and bias term β from Gaussian distributions:

$$\gamma \sim N\big(1, softplus(\theta_\gamma)\big)$$

$$\beta \sim N\big(1, softplus(\theta_\beta)\big)$$

We then compute the modulated activation $\hat{Z}$ as:

$$\widehat{\boldsymbol{Z}} = \gamma \times \boldsymbol{Z} + \beta$$

3. EXPERIMENT

3.1. Experimental Setup

The training is divided into three stages: control train, uncontrol train, and enrollment. In each stage, we replace the BN layers in the baseline model with FWN and LFA layers for model training. And each stage is trained 80 epochs with the batch size of 128. We used the original parameter optimization and learning rate adjustment strategies.

To meet the complete low-resource requirements, we didn't use any additional data and the number of our system's parameter is 290,578.

3.2. Experimental Results

In the enrollment stage, we sequentially enrolled and tested 10 dysarthric speakers from test-a. Considering the practical application convenience of the wake-up word spotting system, when submitting the challenge results, we use a fixed threshold to distinguish keywords and non-keywords. A fixed threshold of 0.002 was used during testing. The test results are shown in Table 1.

Table 1: results of our method on test-a with fixed threshold

Test set-a	Intelligibility	FAR	FRR	Score
DF0023	49.91	0.1860	0.1000	0.2860
DF0026	77.53	0.0907	0.0000	0.0907
DF0028	91.10	0.0895	0.1000	0.1895
DF0030	90.5	0.0560	0.0000	0.0560
DM0022	57.58	0.2294	0.1000	0.3294
DM0024	38.90	0.0722	0.4500	0.5222
DM0025	78.13	0.0982	0.0250	0.1232
DM0027	67.40	0.1502	0.0750	0.2252
DM0029	45.80	0.2423	0.0500	0.2923
DM0031	89.73	0.0976	0.0000	0.0976
average		0.1312	0.0900	0.2212 **(+0.09)**

In Table 2, we present the test results using the optimal thresholds. The last column shows the corresponding optimal thresholds. Compared to the baseline, the final test results show an improvement of 0.1533.

Table 2: results of our method on test-a with optimal threshold

Test set-a	FAR	FRR	Score	Thr
DF0023	0.1860	0.1000	0.2860	0.002
DF0026	0.0056	0.0000	0.0056	0.693
DF0028	0.0293	0.1250	0.1543	0.016
DF0030	0.0020	0.0000	0.0020	0.694
DM0022	0.1093	0.1500	0.2593	0.009
DM0024	0.1106	0.3500	0.4606	0.001
DM0025	0.0226	0.0500	0.0726	0.05
DM0027	0.0615	0.0750	0.1365	0.007
DM0029	0.1169	0.0750	0.1919	0.01
DM0031	0.0097	0.0000	0.0097	0.188
average	0.0653	0.0925	0.1578 **(+0.1533)**	