SYSTEM DESCRIPTION FOR MISP CHALLENGE 2021

Qiuchen Yu[1] Ruohua zhou[1] Chenlei Hu[1]

[1]Department of Automation, Beijing University of Civil Engineering and Architecture,China

ABSTRACT

In this paper, we propose our system for task1 of MISP Challenge 2021. The task1 is aim to solve Audio-Visual Wake Word Spotting. We mainly use the audio dataset for experiment. Two kinds of neural networks we selected for wake word detection.

Index Terms— *MISP challenge 2021, wake word, Res2Net, RawNet*

1. INTRODUCTION

With the emergence of many speech-enable applications, the scenarios (e.g., home and meeting) are becoming increasingly challenging due to the factors of adverse acoustic environments (far-field audio, background noises, and reverberations) and conversational multi-speaker interactions with a large portion of speech overlaps. The state-of-the-art speech processing techniques based on the single audio modality encounter the performance bottlenecks, e.g. Motivated by this, the MISP challenge aims to tackle these problems by introducing additional modality information (such as video or text), yielding better environmental and speaker robustness in realistic applications. The MISP challenge has two speech processing tasks. We signed up for the first task: Audio-Visual Wake Word Spotting. In this paper, we use several kinds of neural networks to detect predefined wake word. The networks we used are Res2Net and RawNet2. Due to insufficient computing resources, our video detection system does not achieve good performance. Finally, video datasets are abandoned, we only use audio datasets in the experiment.

2. DATA PREPARTION

2.1.The Dataset

We use the task1 dataset including one wake word:" Xiao T Xiao T". The statistics for dataset are summarized in Table. 1

Dataset	Train		Dev		Eval
	Positive	Negative	Positive	Negative	
Duration(h)	5.67	112.86	0.62	2.77	2.87

Table. 1 Dataset statistics

2.2.Data Augmentation

Simulating reverberant and noisy data from near field speech, noise is widely adopted. We add reverberation and noise with official tools. SpecAugment [1] strategy achieved a great success in E2E speech recognition task. We apply time masking and frequency masking in this paper. For each training utterance, we randomly select 0-30 consecutive frames and set all of their mel-filter banks to zero, for time masking. For frequency masking, we randomly select 0-20 consecutive dimensions of the 256 mel-filter banks and set their values to zero for all frames of the utterance. For all the utterances in a training mini-batch, one-third of them receive only the time masking, one-third of them only the frequency masking, and the rest of them both maskings. Though this method deforms instead of doubling the original data set, it is considered augmentation as each mini-batch at different epochs is deformed differently. The training observes data with a lot more varieties than the original data amount. Hence it is considered a data-augmentation method.

3. SYSTEM DISCRIPTION

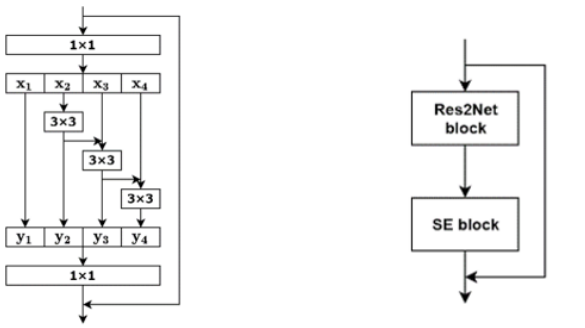
We trained two models for detection wake word with Res2Net [2] and RewNet2 [3].

3.1.Res2Net Block

The Res2Net architecture aims at improving multi-scale representation by increasing the number of available receptive fields. This is achieved by connecting smaller filter groups within one block in a hierarchical residual-like style.The Res2Net block is modified from the bottleneck block [4]. It is illustrated in Fig. 1(a)

3.2.Integration with the squeeze-and-excitation block

The SE block [5] adaptively re-calibrates channel-wise feature responses by explicitly modeling the interdependencies between channels.. This inter-dependencies modeling assigns different impact weights to channels, which improves the model's capacity to focus on channel information that is most related with spoofing cues. Motivated by this, we stack the Res2Net and SE blocks together to form the SE-Res2Net block, as shown in Fig. 1(b).



(a) Res2Net block (b) SE-Res2Net block

Fig. 1. The illustration of the Res2Net block (scale dimension s = 4, the box in each color represents the feature maps within a channel group), and SE-Res2Net block.

3.3.Res2Net architecture

This work chooses SE-Res2Net50, and an overview of its architecture is shown in Table. 2.

Stage	SE-Res2Net50
Conv1	[conv2d, 3 × 3, 16, *stride* = 1] ×3
Conv2	[SE-Res2Net BLK, 16] ×3
Conv3	[SE-Res2Net BLK, 32] ×4
Conv4	[SE-Res2Net BLK, 64] ×6
Conv5	[SE-Res2Net BLK, 128] ×3
global average pool, 2-d fc, softmax	

Table. 2 The overall model architecture of SE-Res2Net50. The type of a residual block and the number of channels is specified inside the brackets, while the repeat times of each block on one stage are specified outside the brackets. "2-d fc" denotes a fully connected layer with 2 output units.

3.4.RawNet2

RawNet2 is a end-to-end network that operate on the raw speech waveform. RawNet2, proposed in 2020, combines the merits of the original RawNet approach (RawNet1 [6]) with those of SincNet. The first layer of RawNet2 is essentially the same as that of SincNet [7], whereas the upper layers consist of the same residual blocks and GRU layer as RawNet1. New to RawNet2 is the application of filter-wise feature map scaling (FMS) using a sigmoid function applied to residual block outputs as in [8]. FMS acts as an attention mechanism and has the goal of deriving more discriminative representations. The embedding dimension for RawNet2 is also greatly increased, from 128 for RawNet1 to 1024 for RawNet2. Last, whereas RawNet1 obtains better results with a DNN-based back-end classifier, RawNet2 gives better results with a cosine similarity score.

The RawNet2 architecture used for detection wake word is modified by [9] and shown in Table3.

Layer	Input:59049 samples	Output shape
Sinc conv	SINC(256,1,128) MaxPool(3) BN LeakyReLU	(19683,128)
Res block	{ BN LeakyReLU Conv(3,1,128) BN LeakyReLU Conv(3,1,128) MaxPool(3) FMS } ×2	(2187,128)
Res block	{ BN LeakyReLU Conv(3,1,256) BN LeakyReLU Conv(3,1,256) MaxPool(3) FMS } ×4	(27,256)
GRU	GRU(1024)	(1024,)
Speaker embedding	FC(1024)	(1024,)

Table. 3 The overall model architecture of RawNet2

4. EXPERIMENTS

4.1.Experimental configurations

We extracted 256-dimensional mel-filter banks features with 25ms frame length and 10ms frame shift are extracted as input features for training SE-Res2Net50. And we used raw waveforms with pre-emphasis applied as input to the RawNet2. Here, we modified the duration of the input waveforms to 59049 samples ($\approx$3.69 s). The other settings are consistent with the original paper.

4.2.Proposed systems

SE-Res2Net50 and RawNet2 are mainly used by us because their performance is better than other systems. However, we found that if we mix several networks, we will get better results. Finally, there are four systems we used to cooperative detection, SE-Res2Net50, RawNet2, SqueezeNet [10], Inc-TSSDNets [11].

4.3.Results

Table. 4 shows results in terms of score for SE-Rs2Net50 and RawNet2. The score is the sum of false alarm rate and false reject rate.

System	Dev	Eval
SE-Res2Net50	-	0.179
RawNet2	0.064	0.216

Table. 4 Experimental results of SE-Rs2Net50 and RawNet2

4.4.Fusion

We get the result of the two systems fusion. Table. 5 Shows the result. We also tried to integrate some other networks,

SqueezeNet and Inc-TSSDNets. These two systems do not perform well, but they can improve the final fusion system..

System	Eval
SE-Res2Net50+ RawNet2	0.141
Submitted system	0.123

Table. 5 Performance for the task1 evaluation partition in terms of pooled score for two systems.

5. REFERENCES

[1] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.

[2] X. Li et al., "Replay and synthetic speech detection with res2net architecture," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: IEEE, pp. 6354-6358.

[3] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," arXiv preprint arXiv:2004.00526, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[6] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," arXiv preprint arXiv:1904.08104, 2019.

[7] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018: IEEE, pp. 1021-1028.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.

[9] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6-11 June 2021 2021, pp. 6369-6373, doi: 10.1109/ICASSP39728.2021.9414234.

[10] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," arXiv preprint arXiv:1602.07360, 2016.

[11] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," IEEE Signal Processing Letters, vol. 28, pp. 1265-1269, 2021.