MISP CHALLENGE 2021: MULTIMODAL WAKEUP-WORD DETECTION FOR FAR-FIELD AND NOISY ENVIRONMENT—TECHNICAL REPORT

Dianwen Ng^{1,2}, *Jin Hui Pang*^{2}, *Yang Xiao*^{2}, *Eng Siong Chng*^{2}

^{1} Alibaba Group, Singapore
^{2} School of Computer Science and Engineering, Nanyang Technological University, Singapore
dianwen.ng@alibaba-inc.com
{pang0208, yxiao009}@e.ntu.edu.sg
aseschng@ntu.edu.sg

ABSTRACT

This technical report describes the details to the strategy we adopted for Task 1 in the MISP Challenge 2021. In this work, we proposed a multimodal wake-up word detection model that handles the audio and the visual input to determine the presence of a predefined word. This involves the use of the extracted latent representation from a trained multi-channel audio keyword spotting and a video classification model with a shallow fusion to predict the outcome of our task. To allow a more robust performance under the noisy and far-field environment, curriculum learning based on the distance and the level of the augmented noise is exploited with increasing difficulty in multiple stages. Moreover, we applied the weighted prediction error (WPE) for speech dereverberation to enhance our utterance for a better learning. Lastly, we built a separate of two system with different audio model, namely with temporal-convolution (TC-ResNet) and ConvMixer and we take the average of the predictive score on two system. Our experimental result for team ALISPEECH has achieved top 6 on the leaderboard with a score of 0.109.

Index Terms— multi-channel keyword spotting, noisy far-field, video classification, multi-modality

1. INTRODUCTION

The rapid advancement of technology has allowed the interaction of humans and machines via voice command. Numerous voice assistant applications, such as the Apple Siri, are integrated into consumer electronics and household appliances. To activate such applications, a keyword spotting or wake-up word detection model is established to determine the presence of a predefined keyword in a given utterance. Formally, keyword spotting (KWS) [1] can be defined as the task of identifying keywords in the audio streams comprising speech. With the rising use of voice applications, keyword spotting task has become increasingly challenging due to the increased demand for a more robust model for unfavourable acoustic conditions (far-field sound, background noise, and reverberation) and multiple person conversations with significant voice overlap. Motivated by this, the Multimodal Information Based Speech Processing (MISP) Challenge 2021 aims to tackle these problems by introducing additional modality information (such as video or text), yielding better environmental and speaker robustness [2] in a realistic environment. Besides, human speech perception is bimodal by nature since it relies on both auditory and visual information. Keyword spotting can benefit from combining visual and audio information to enhance their performance. This is also known as the audio-visual keyword spotting [3] or multi-modality keyword spotting. In task 1, MISP2021 considers the following scenario: several people are chatting while watching TV in the living room, and they can interact with a smart speaker/TV. With the multimodality data collected from the microphones and cameras, we build an audio-visual keyword spotting (KWS) model to carry out the task.

In this work, we proposed methods for adapting our model to the background noises and reverberations, i.e. noisy and far-field environment. We were inspired by [4] and proposed a novel multi-channel ConvMixer with centroid based keywords as our wake-up word audio model for extracting features from the multi-channel audio data. In order for the model to perform well in noisy environments, we applied weighted prediction error (WPE) [5] for speech dereverberation to enhance audio data before training. In addition, we trained the model with the strategy of curriculum-based multi-condition training that surpasses the vanilla multi-condition learning. Thus, our model is more robust to noise and achieves better performance in noisy environment. On the other hand, since visual information is not affected by acoustic distortions, we propose a pre-trained visual front-end to extract feature vectors from video clips. We then pass them through Transformer layers to capture temporal information, after adding positional encoding. Finally, the extracted audio-visual data is fused [1] in order to make a determination about the presence or absence of a keyword.

The rest of the report is organized as the following. Section 2 describes the network architecture, the Weighted Prediction Error, the Keyword Centroid, and curriculum learning methods. Section 3 shows the experimental results and analysis. Finally, we conclude the work in Section 4.

2. PROPOSED METHOD

This session introduces an efficient strategy design for the task of Audio-Visual Keyword Spotting. First, in section 2.1 we present a modified version of ConvMixer and TC-ResNet for multi-channel far-field spoken keyword spotting. Second, we demonstrate how to use Transformer-based method to extract the feature from video clips in section 2.2. Finally, we present the fusion of audio and video feature for the purpose of predicting the result in section 2.3.

2.1. Audio Stage

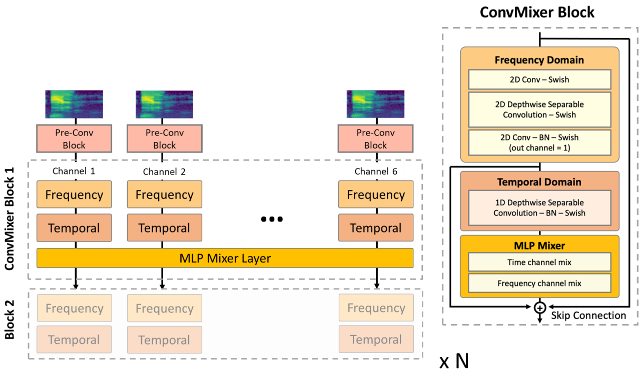

Fig. 1. Multi-channel ConvMixer

2.1.1. Multi-channel ConvMixer

Our multi-channel ConvMixer networks expands from previous work in [4]. For a single-channel ConvMixer, the architecture is divided into three main sections, namely the pre-convolutional block, the convolution-mixer block, and the post-convolutional block. We designed our pre- and post-convolutional blocks using the same neural layers of a 1-dimensional DWS, batch normalization followed by the swish activation. To preserve the dimension from the previous time frame, each of the following blocks is convolved with a different kernel size.

The ConvMixer block takes the previous channel × time feature and passes it through the 2D convolutional sub-block for frequency domain extraction. This creates a third dimension that expresses the rich information from the frequency domain. To maintain the shape from the previous input, we employed a pointwise convolution to compress it back to the original shape. Then, we implemented the temporal domain feature extraction with a 1-dimensional DWS block. The product of these two operations will result in frequency and temporal rich embeddings. Following that, we constructed a mixer layer to enable information to flow over the global feature channel. Here, we added an additional audio channel mixing on top of the previous single-channel model. Lastly, we added skip connections from the previous output and the 2D feature connecting to the output of the block.

2.1.2. TC-ResNet

We adopt TC-ResNet [6] as a secondary KWS system. It is built based on ResNet, one of the most widely used CNN architecture but utilize 3×1 kernels for first layer and 9×1 kernels for the other layers. Temporal Convolution which is 1D convolution along the temporal dimension is applied on the network. It is proven to increase the effective receptive field in comparison to the original ResNet that computes with strided convolutions and without dilated convolutions. This audio model will be later fused with the video model similar to our ConvMixer audio-visual model to create our second KWS system. Subsequently, we will perform an average ensembling to obtain our final predictive score.

2.1.3. Keyword Centroid

We propose training the network to learn the spatial relationship of the posterior latent space where all input utterances are projected. When a target input is given, an embedding vector can be extracted with a trained model as in figure 2.1.3. By computing distances between the embedding vector and each centroid vectors, we can classify the given utterance with the posterior space distance and the representation features. Our loss function is the sum of the normal classification loss function and the centroid loss function.

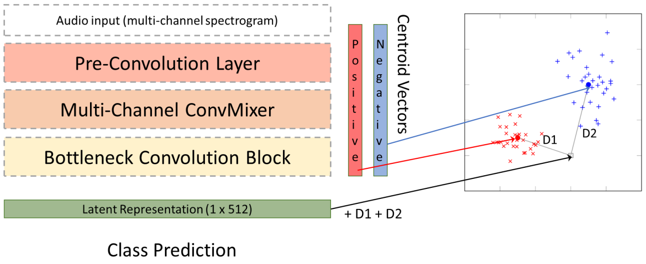

Fig. 2. Keyword Centroid

2.1.4. Weighted Prediction Error(WPE)

Background noise and signal reverberation caused by enclosure reflections are the two primary impairments in acoustic signal processing and far-field speech recognition. This work addresses signal dereverberation techniques based on WPE [5] for speech recognition and other far-field applications. WPE is a compelling algorithm to blindly dereverberate acoustic signals based on long-term linear prediction. We used WPE [5] and SpecAugment [7] to pre-process our data.

2.1.5. Curriculum Based Multi-condition Training

To enhance the noise robustness of our model, the curriculum learning based on the distance level and noisy environment is employed as a training strategy. To execute the training process, we divide it into three progressively harder steps. Noise is added using the provided official dataset, which can be found in the noise folder. This involves the following procedures:

- Stage 1: Training with near-field audio + official noise (clean, 5 dB)
- Stage 2: Training with middle-field audio + official noise (clean, 0 dB)
- Stage 3: Training with far-field audio + official noise (clean, -5 dB)

2.2. Video Stage

In this section, we will introduce our proposed method with its overview as the figure 2.2. Following that, we will describe each stage of the pipeline for keyword spotting in lip videos. It comprises a video front-end, transformers, and prediction heads for multi-layer perceptrons (MLP).

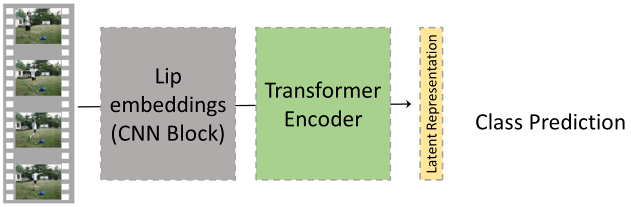

Fig. 3. Video model architecture

2.2.1. Model Architecture

Our model is fed with video clips in which we used to spot the keyword ‘Xiao T Xiao T’. With the input video frames, the visual front-end will extract low-level visual features. We then pass them through Transformer layers to capture the temporal information, after adding positional encoding. To predict the probability of the keyword being present in the video derived from the output feature, the two multi-layer perceptrons (MLP) [8] heads are used for binary classification and localization of the keyword that is shared throughout all the video output states from the Transformer respectively.

2.2.2. Data preprocessing

To facilitate the extraction of the content from the video, input video data was converted into sequential frames whereas the massive amount of images makes it difficult to employ video information for further processing. To eliminate redundant information, images were cropped and resized to a fixed size of 96×96 which only focused on the region of interest around the mouth. The frames are then transformed to grayscale and normalized with respect to the mean and variance of the entire dataset.

2.2.3. Video front-end

By applying the processed frames, the pre-trained visual front-end is implemented to extract feature vectors. It utilizes the spatio-temporal convolutional layers that include a convolutional layer with 64 filter maps of 3D convolution with a kernel size of $5 \times 7 \times 7$, Batch Normalization (BN) [9], Rectified Linear Units (ReLU) and a spatio-temporal maxpooling layer which extracted feature maps pass to reduce their spatial size. It is vital with the benefit of capturing the short-term dynamics of the region of interest. Followed by the video front-end, the feature maps then pass through the residual network (ResNet) [10] with 18-layer identity mapping version, the ResNet gradually reduces the spatial dimension of its output until each time step produces a single dimensional tensor.

2.2.4. Transformer

Instead of using convolutional neural networks (CNNs) or long short-term memory (LSTM), a Vision-Transformer [11] model was selected as the back-end of our model, since the Transformer model has been applied widely in the areas of natural language processing [12, 13] and visual learning [14, 15] with its relatively simple structure but capable of achieving competitive performance.

2.3. Fusion Stage

We propose a feature-level fusion model. Specifically, speech and visual features are concatenated before their joint classification using a neural network model. As in the figure, both latent representations input to the full connected layers as shallow fusion, then output the class prediction.

3. EXPERIMENTS

3.1. Experiment Setting

Datasets and metrics. We evaluate the proposed method on the datasets provided by MISP challenge 2021, it includes three datasets: training, development, and evaluation. We have a total of 124.79 hours of data. There are more than 300

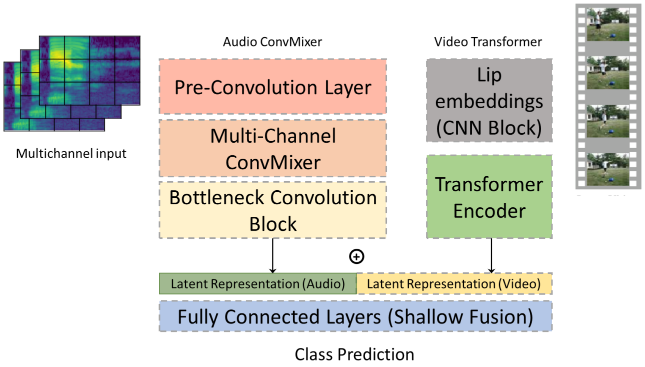
Multichannel input
Audio ConvMixer
Pre-Convolution Layer
Multi-Channel ConvMixer
Bottleneck Convolution Block
Video Transformer
Lip embeddings (CNN Block)
Transformer Encoder
⊕
Latent Representation (Audio)
Latent Representation (Video)
Fully Connected Layers (Shallow Fusion)
Class Prediction

Fig. 4. Fusion Stage

speakers with different Mandarin accents, and all data was collected in over 30 real-world rooms. If the wake-up word is included in the sample, it is considered as a positive sample, otherwise it will be regarded as a negative sample. Additionally, performance is evaluated by the sum of the false alarm rate (FAR) and the false reject rate (FRR). This score will be used as the judging criteria.

Implementation Details. For audio input features, we use 40-dimensional log Mel spectrograms with length of 200ms. Audio will be padded with zeros if the time length is shorter than 200ms or it will be trimmed if it exceeded 200ms. During training, we augment data to get a more generalized model. In the time dimension, we randomly shift each input feature in the range of -10 to 10. We also use Specaugment [7] with two frequency masks and two temporal masks with mask parameters of 25 and 7, respectively. The number of epochs is 30 for the first training stage, 50 for the second training stage, and 50 for the third training stage using AdamW optimizer with weight decay of 5e-7 and mini-batch size of 64.

For video input information, we adopt the pre-trained feature front-end using transformers for lip reading, to obtain faster training time, the pre-computed visual features for each backbone has been applied. Based on it, we remove the fully-connected classification layer, and add transformers and two prediction heads to classify and localize the keyword. The total epoch number is 50 with mini-batch size of 4 for the input dimension of 256. We use ADAM as the optimizer with the constant learning rate of 8e-6 and weight decay of 5e-4.

Result. We have achieved the top 6 in the leaderboard with the score of 0.109.

4. CONCLUSIONS

In this work, we designed a system to achieve two goals; 1) Adapting the noisy far-field environment and 2) Fusion the visual information. To design an efficient Keyword Spotting model, we suggest a multi-channel version of ConvMixer [4] with Keyword Centroid Loss. Besides, we further enhance the data by WPE [5] and Specaugment [7]. Additionally, to improve the noise robustness of our model, curriculum learning based on the distance level and noisy environment are employed as a training strategy. We set a visual feature pipeline comprises of a video front-end, transformers, and prediction heads for multi-layer perceptrons (MLP). Finally, we deploy a fusion stage to combine features with audio and video to predict. We evaluate our proposed system on the MISP2021 dataset. We are the TOP 6 group in the leaderboard. And we get the score of 0.109.

5. ACKNOWLEDGEMENTS

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore.

6. REFERENCES

[1] Iván López-Espejo, Zheng-Hua Tan, John Hansen, and Jesper Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, 2021.

[2] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu, "Audio-visual recognition of overlapped speech for the lrs2 dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6984–6988.

[3] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman, "Seeing wake words: Audio-visual keyword spotting," *arXiv preprint arXiv:2009.01225*, 2020.

[4] Dianwen Ng, Yunqi Chen, Biao Tian, Qiang Fu, and Eng Siong Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," 2022.

[5] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[6] Seungwoo Choi, Seokjun Seo, Beomjun Shin, Hyeongmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, and Sungjoo Ha, "Temporal convolution for real-time keyword spotting on mobile devices," *arXiv preprint arXiv:1904.03814*, 2019.

[7] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *Interspeech 2019*, Sep 2019.

[8] PA Castillo, J Carpio, JJ Merelo, Alberto Prieto, V Rivas, and Gustavo Romero, “Evolving multilayer perceptrons,” *Neural Processing Letters*, vol. 12, no. 2, pp. 115–128, 2000.

[9] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[10] Sasha Targ, Diogo Almeida, and Kevin Lyman, “Resnet in resnet: Generalizing residual architectures,” *arXiv preprint arXiv:1603.08029*, 2016.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[13] Scott E Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Christopher Miller, et al., “Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2021, vol. 43.

[14] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al., “A survey on visual transformer,” *arXiv e-prints*, pp. arXiv–2012, 2020.

[15] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, “Is space-time attention all you need for video understanding?,” 2021.