

# THE DKU POST-CHALLENGE AUDIO-VISUAL WAKE WORD SPOTTING SYSTEM FOR THE 2021 MISP CHALLENGE: DEEP ANALYSIS

Haoxu Wang<sup>1,2</sup>, Ming Cheng<sup>1,2</sup>, Qiang Fu<sup>3</sup>, Ming Li<sup>1,2†</sup>

<sup>1</sup>School of Computer Science, Wuhan University, Wuhan, China

<sup>2</sup>Data Science Research Center, Duke Kunshan University, Kunshan, China

<sup>3</sup>Alibaba Group, China

## ABSTRACT

This paper further explores our previous wake word spotting system ranked 2-nd in Track 1 of the MISP Challenge 2021. First, we investigate a robust unimodal approach based on 3D and 2D convolution and adopt the simple attention module (SimAM) for our system to improve performance. Second, we explore different combinations of data augmentation methods for better performance. Finally, we study the fusion strategies, including score-level, cascaded and neural fusion. Our proposed multimodal system leverages multimodal features and uses the complementary visual information to mitigate the performance degradation of audio-only systems in complex acoustic scenarios. Our system obtains a false reject rate of 2.15% and a false alarm rate of 3.44% in the evaluation set of the competition database, which achieves the new state-of-the-art performance by 21% relative improvement compared to previous systems. Related resource can be found at: [https://github.com/Mashiro009/DKU\\_WWS\\_MISP](https://github.com/Mashiro009/DKU_WWS_MISP).

**Index Terms**— audio-visual wake-up word spotting, simple attention module, multimodal system, complex acoustic scenarios

## 1. INTRODUCTION

With the rapid application of hand-free devices such as mobile phones and voice assistants, Wake Word Spotting (WWS), also known as a specific case of Keyword Spotting (KWS), is increasingly attractive. WWS system is designed to detect a predefined wake word or a set of wake words in the streaming audio.

Recently, many works for WWS based on deep neural network are proposed, including deep neural networks (DNN) [1], convolutional neural networks (CNN) [2], temporal convolutional neural networks [3] and Transformer [4]. These works show good performance under clean and close-talking scenarios. However, it is observed that the probability of the false alarm becomes higher under complex acoustic environments such as background noises (cheers, TV or screams), reverberations, and conversational multi-speaker interactions with a significant portion of speech overlaps, which will harm the user experience in practical applications.

Although there has been a great deal of multi-modal works in the field of audio-visual speech recognition (AVSR) [5, 6], previous works mainly focus on the audio-only WWS system, which ignores the use of visual information. Inspired by human perception, complementary visual information (e.g., lip movements) can help understand the semantics of speech in complex acoustic scenarios since the visual modal is not affected by acoustic noise. [7] uses conformer [8] architecture in AVSR, and [9] investigates an audio enhancement module

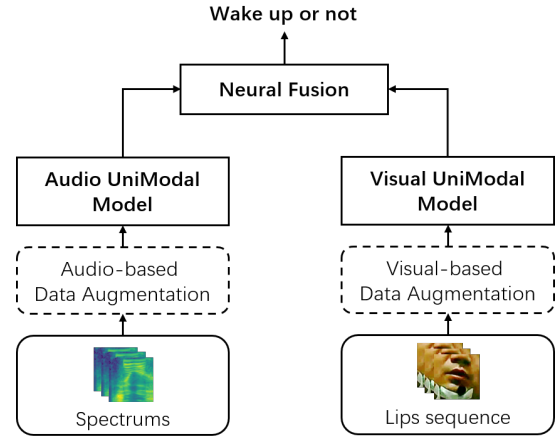


Fig. 1. Framework of our audio-visual wake word spotting system.

called V-CAFE for AVSR. In [10], the authors investigate a simple CNN-based audio-visual KWS system and demonstrate good performance for noisy audio environments. Along with the 1st Multimodal Information based Speech Processing Challenge (MISP Challenge 2021 [11]) and its data release [12], some new works are proposed for Audio-Visual Wake Word Spotting (AVWWS) including CNN-3D-based model [13] and transformer-based end-to-end model [14].

In this paper, we extend our previous work [13] to an effective AVWWS system. We design the robust unimodal system by replacing previous 3D CNN with a mixture of 2D and 3D CNNs. We also adopt the simple attention module (SimAM) [15] to our WWS system without extra parameters, which has been used in automatic speaker verification (ASV) [16, 17]. To further improve the performance of our WWS system, we do experiments to explore better combination of the data augmentation methods. Finally, we study the fusion strategies for the two unimodal systems and obtain a 21% relative improvement on top of our previous second-place work. Also, we achieve the state-of-the-art (SOTA) result (5.59% WWS score) on the MISP dataset.

## 2. SYSTEM DESCRIPTION

As shown in Figure 1, we first investigate a robust model for single-modality and then study the fusion strategies for the two solid unimodal systems to get an effective audio-visual WWS system. We also investigate the data augmentation strategies and find a better combination than the baseline.

† Corresponding Author, E-mail: ming.li369@dukekunshan.edu.cn

## 2.1. Unimodal Models

### 2.1.1. 3D-ResNet34

Our previous work [13] constructs a ResNet34-like neural network based on 3D CNNs (ResNet-3D) to cope with the unimodal WWS task. The two modalities have the same network architecture, except that the dimensions of the inputs are different. The video inputs are raw image sequences with the shape of  $(T, H, W, 3)$ , in which  $T$  represents the number of frames,  $H, W$  represents the height and width of the image, and 3 represents the RGB channel. For the audio inputs, we first extract the 2D acoustic features  $X_{spec} \in \mathbb{R}^{T' \times D}$  (e.g. FBank, MFCC) from the 1D waveform, and then we use a sliding window with the shape of  $(D, D)$  to slice the 2D acoustic features along the time axis with the stride of  $\frac{T'}{D}$  because of the difference of the fps of audio and video, so we organize the audio input  $X_{audio}$  with the shape of  $(T, D, D, 1)$ . Hence, we can make the decision using the unimodal model with inputs that have the same shape of  $(T, H, W, C)$ .

### 2.1.2. 2D-ResNet34

In previous approaches for audio-only WWS tasks [2, 3, 4], the 2D spectrum  $X_{spec}$  is usually considered as a single channel image, we also design the 2D-ResNet34 network for such input. The experiments in [18] and Sec. 4.2 show that 3D-ResNet is more suitable for the temporal-spatial features compared with 2D-ResNet, and the constructed  $X_{audio}$  contains local short-time spatial information which 3D-ResNet can better utilize.

### 2.1.3. 3D-ResNet18+2D-Resnet18

[18] indicates that 3D CNNs at low-level layers focus on short-term spatial modeling, and 2D CNNs help extract semantic and temporal information at high-level layers in the Action Recognition task. Hence, we split the 3D-ResNet34 into two cascaded networks, which are 3D-ResNet18 and 2D-ResNet18. The design of converting 3D CNN to 2D CNN can reduce the number of parameters of the model. At the same time, the network with a mixture of 3D CNN and 2D CNN can collaborate on the modeling, allowing different modules to focus on different perspectives. The inputs of this network are the same as those of the 3D-ResNet34, but the feature map is averaged along the spatial axis after the 3D-ResNet18 to the shape of  $(C, T)$ . Then we consider this as a one-channel latent image with the shape of  $(1, T, C)$  and send it to the 2D-ResNet18.

### 2.1.4. 3D-ResNet18+2D-Resnet18+SimAM

The attention modules for CNN have been widely used and have achieved great performance in computer vision and speech processing. Previous channel-wise squeeze-excitation (SE) [19] obtains 1D channel-level weights and has been used in WWS [20]. Convolutional block attention module (CBAM) calculates 2D and 1D attention weights separately [21]. All of these attention modules need extra parameters.

[15] proposes SimAM based on the phenomenon of spatial suppression [22] in neuroscience. The 3D weights estimate the importance of individual neurons, and spatial suppression shows that the most informative neurons usually show distinctive firing patterns from surrounding neurons. So the energy function for each target neuron output  $t$  in the 3D feature map  $X \in \mathbb{R}^{C \times H \times W}$  is defined as:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1} \sum_{i=1}^{M-1} (y_o - \hat{x}_i)^2 \quad (1)$$

where  $\hat{t} = w_t t + b_t$  and  $\hat{x}_i = w_t x_i + b_t$  are the linear transforms of  $t$  and  $x_i$ ,  $x_i$  represents the output of neurons other than  $t$  in the same

channel of the feature map.  $i$  is the index along the spatial dimension  $M = H \times W$ . The energy function attains the minimal value when  $\hat{t} = y_t$  and  $\hat{x}_i = y_o$ , and  $y_t$  should be larger than  $y_o$  since the spatial suppression. For simplicity, [15] uses the binary label to set the  $y_t = 1$  and  $y_o = -1$ , and also adds a regularizer into Equation 1:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} (-1 - (w_t x_i + b_t))^2 + \frac{1}{(1 - (w_t t + b_t))^2 + \lambda w_t^2} \quad (2)$$

It is computationally expensive to solve the above function. Luckily, Equation 2 has a fast closed-form solution with respect to  $w_t$  and  $b_t$  according to [15]. Then we can get the final minimal energy by putting  $w_t$  and  $b_t$  back to equation 2:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (3)$$

where  $\hat{\mu} = \frac{1}{M} \sum_{i=1}^M x_i$  and  $\hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^M (x_i - \hat{\mu})^2$ . The importance of each neuron can be obtained by  $\frac{1}{e_t^*}$  based on the spatial suppression. The final SimAM is defined as:

$$\tilde{X} = \sigma\left(\frac{1}{E}\right) \otimes X \quad (4)$$

where  $\otimes$  represents the element-wise multiplication,  $E$  groups all  $e_t^*$  across channel and spatial dimensions for the 3D feature map  $X$ , and  $\sigma(\cdot)$  represents the sigmoid function. SimAM calculates the 3D attention weights based on the energy function without extra parameters. Moreover, it is easy to apply the SimAM to the 4D feature map  $X_{4D} \in \mathbb{R}^{C \times T \times H \times W}$  of the 3D CNN, we calculate the energy value for each neuron along the temporal and spatial dimension and set  $M = T \times H \times W$ .

## 2.2. Fusion Strategy

### 2.2.1. Score-Level Fusion

To fusing the audio-visual systems together, the official baseline makes the final decision just by weighted summation from the separate unimodal system.

$$P_{av} = \alpha \times P_a + \beta \times P_v \quad (5)$$

where  $P_a$  and  $P_v$  are the posterior probabilities generated by the audio- and video-only systems. However, the posterior probability distributions of the audio- and video-only systems are pretty different. Hence, we need to investigate a more suitable fusion strategy.

### 2.2.2. Cascaded Fusion

Since there is a performance gap between the two systems, we can use a cascade scheme to let one system with poorer performance detect the wake-up words samples with low confidence  $th_l$  first and then let the other system with better performance give the final result with a higher threshold  $th_h$ . Here we choose the video-only system as the first stage and the audio-only system as the second stage.

### 2.2.3. Neural Fusion

The latent features extracted from the unimodal system are rich in terms of semantic information and can be used to determine whether they are wake-up word samples. Hence, we train a simple neural network to use these latent features further.

We use the Hierarchical Modality Aggregation (HMA) approach in our previous work [13]. We apply global average pooling to the latent feature maps from the hierarchical residual CNN blocks in each

unimodal neural network at different layer levels. We concatenate the obtained embedding vectors at the same level from two modalities to get multimodal embedding vectors  $c_l$ , where  $l$  represents layer level. Then we use the HMA to aggregate all the  $c_l$  and give the final multimodal results.

### 2.3. Data Augmentation

#### 2.3.1. Audio Stream

For the audio stream, we have tried some audio-base data augmentation methods according to [9, 13, 11], including offline noise/reverberation adding, beamforming enhancement, negative sub-segmentation, speed perturbation, volume perturbation, trimming slightly and SpecAugment [23].

- Offline noise/reverberation adding (NR): We generate room impulse response to the original near-field data to simulate mid and far-field data by using the pyroomacoustic tool and mix noises provided by the officials with a random SNR from -15 to 15 dB referring to the official codes.
- Beamforming enhancement (BE): We implement beamforming (MVDR [24]) to multi-channel audio signals to exhaust the potential of microphone arrays and incorporate these beamforming-enhanced audios into the training data.
- Negative Sub-segmentation (NS): It is possible for the model to detect the wake-up word only according to the length of the audio with low accuracy since the duration distribution of the positive and negative samples are quite different. Hence, we refer to [25] sub-segment the negative samples during training.
- Volume perturbation (VP): Each audio is randomly selected to change its volume. The volume changes in the range [0.125, 2].
- Speed perturbation (SP): The speed of the audio is randomly changed to be faster or slower, with the ratio in the range [0.9, 1.1].
- Trimming slightly (TS): Each audio is randomly selected to be trimmed slightly at the beginning or end, with the ratio of 0.95.
- SpecAugment (SA): we use the frequency masking and the time masking for each randomly selected audio.

#### 2.3.2. Video Stream

For the visual stream, we adopt multiple video-based data augmentation methods, including speed perturbation, frame-wise rotation, horizontal flip, frame-level cropping, color jitters and gray scaling. Exact details can be found in [13].

## 3. EXPERIMENTAL SETUP

### 3.1. Dataset and Evaluation Metrics

We use the dataset provided by the MISP Challenge 2021. The target of the challenge is to detect the wake word 'Xiao T, Xiao T' spoken by the participants in the far-field smart-home scenarios.

The released database has two subsets: the training set (47k+ negative samples and 5K+ positive samples) and the development set (dev: 2k+ negative samples and 600+ positive samples), which are in the near, middle and far fields. Moreover, an evaluation set (eval: 8K+) without annotations is provided to competition participants, which is only in the far field. After the competition, we get the annotations from the MISP committee to compare our results with other teams fairly.

In the AVWWS task, the positive class represents the existence of the wake word in a given sample, and the negative class indicates the opposite. Following the requirements of the evaluation plan, we use False Reject Rate (FRR), False Alarm Rate (FAR), and the Score of WWS as the evaluation criteria. Let  $N_{wake}$  denote the number of samples that contain the wake word, and  $N_{non.wake}$  represent the number of samples without the wake word. The FRR and FAR are defined as follows:

$$FRR = \frac{N_{FRR}}{N_{wake}}, \quad FAR = \frac{N_{FAR}}{N_{non.wake}} \quad (6)$$

where  $N_{FRR}$  denotes the number of samples containing the wake word while not recognized by the system.  $N_{FAR}$  denotes the number of samples containing no wake words while predicted to be positive by the system. Hence, the final score of Wake Word Spotting (WWS) is defined as:

$$Score^{WWS} = FRR + FAR \quad (7)$$

### 3.2. Data Preprocess

For the visual stream, we only use the RGB lip region images of the video. We employ a face detector (RetinaFace [26]) to get the face images and facial landmarks from the video. And then, we deploy a face recognizer (ArcFace [27]) to select the target speaker in the far-field video. Finally, we crop the lip regions of the target speaker based on the detected facial landmarks. The details can be found in [13]. Each extracted lip-region video is resized to have a resolution of  $112 \times 112$  with 3 RGB channels. The dimension  $T$  is set to 64, which means each video is sampled to contain 64 frames. Therefore, the shape of the video sample becomes (64, 112, 112, 3).

For the audio stream, we extract the 80-dim log-mel filterbank features (FBank) from the waveform with 25ms long and 10ms shift. The dimension  $T'$  is set to 256, which means each audio is sampled to contain 64 frames. Therefore, the shape of the audio sample becomes (256, 80). For 3D-ResNet input, we use a sliding window with the shape of (80, 80) to slice the features with the stride of 4. Hence, we organize the audio samples with the shape of (64, 80, 80, 1).

### 3.3. Model Details

For SimAM, the hyper-parameter  $\lambda$  is set to 0.001. For Score-Level Fusion,  $\alpha$  and  $\beta$  are set to 0.5.  $th_l$  is set to 0.1 and  $th_h$  is set to 0.4 for Cascaded Fusion. All models are trained on 1 GPU 3090, and the batch size is 64. The learning rate is set to 0.001 while training unimodal model and 0.0001 while training HMA fusion model by the Adam optimizer. And we adopt the weighted BinaryCrossEntropy (BCE) Loss (negative:positive=5:1) to tackle the imbalance between positive and negative samples.

## 4. RESULTS AND DISCUSSION

### 4.1. Ablation experiments for data augmentation

We do the ablation experiments for the audio-based data augmentation to find the best combination of these methods. Table 1 reports the details results on the far-field development and evaluation set. For convenience, we use 2D-ResNet34 for our experiments here. We find that the performance of the system degrades without any data augmentation. We first apply NR and BE to the audio-only system, and there was some improvement because of increase in the number of training samples. Based on this, SP, TS and SA bring more significant boosts. We find that VP and NS will slightly degrade the performance of the model. However, the combination of NS and other methods

**Table 1.** Performance of different data augmentation methods based on audio-only 2D-ResNet34 for the far-field data. NR, BE, NS, VP, SP, TS and SA represent offline noise/reverberation adding, beamforming enhancement, negative sub-segmentation, volume perturbation, speed perturbation, trimming slightly and specaugment in Sec. 2.3.1.

ID	DA						Dev[%]			Eval[%]		
	NR+BE	NS	VP	SP	TS	SA	FRR	FAR	WWS	FRR	FAR	WWS
D1	-	-	-	-	-	-	12.5	3.42	15.92	16.62	4.75	21.36
D2	✓	-	-	-	-	-	10.42	5.68	16.09	13.24	6.77	20.01
D3	✓	✓	-	-	-	-	9.78	5.44	15.21	12.08	8.12	20.2
D4	✓	✓	✓	-	-	-	8.33	6.35	14.68	13.12	7.42	20.54
D5	✓	✓	-	✓	-	-	7.53	4.62	12.15	7.66	8.59	16.25
D6	✓	✓	-	-	✓	-	8.65	4.71	13.37	10.79	6.35	17.14
D7	✓	✓	-	-	-	✓	5.45	7.12	12.57	7.05	10.15	17.2
D8	✓	✓	-	✓	✓	✓	5.45	4.76	10.21	5.64	6.05	<b>11.69</b>
D9	✓	-	-	✓	✓	✓	6.09	4.43	10.51	4.97	7.33	12.3

**Table 2.** Comparison of different audio-only systems. 3D-18+2D-18 means 3D-ResNet18+2D-ResNet18.

ID	Model	Field	Dev[%]			Eval[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
A1	2D-ResNet34	Far	5.45	4.76	10.21	5.64	6.05	11.69
A2	3D-ResNet34	Far	5.77	4.28	10.05	5.46	5.88	11.34
A3	3D-18+2D-18	Far	6.76	2.98	9.71	7.05	4.19	11.24
A4	A3+SimAM	Far	5.93	3.61	9.54	6.38	4.65	<b>11.03</b>

**Table 3.** Comparison of different video-only systems. 3D-18+2D-18 means 3D-ResNet18+2D-ResNet18.

ID	Model	Field	Dev[%]			Eval[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
V1	3D-ResNet34	Far	8.81	8.03	16.85	18.52	9.03	27.54
V2	3D-18+2D-18	Far	9.78	6.64	16.41	14.41	9.62	24.03
V3	V2+Pretrain	Far	10.1	6.3	16.3	11.83	9.51	21.34
V4	V3+SimAM	Far	6.89	9.09	15.98	9.56	13.37	22.93
V5	V4+Pretrain	Far	9.13	6.25	15.39	8.03	11.1	<b>19.13</b>

will assist in performance improvement. We finally conclude the set of data augmentation methods, achieving a performance of 11.69% WWS on the eval set.

## 4.2. Results for different systems

Table 2 shows the results for audio-only systems of different architectures. 3D-ResNet34 performance is better than 2D-ResNet34 because of the modeling of the short temporal feature. 3D-ResNet18+2D-ResNet18 achieves better results according to the mixture of 3D and 2D CNN. Finally, the 3D-ResNet18+2D-ResNet18+SimAM achieves the best result (11.03% WWS) without introducing extra parameters.

Table 3 shows the results for video-only systems of different architectures. 3D-ResNet18+2D-ResNet18 achieves better results according to the mixture of 3D and 2D CNN. The 3D-ResNet18+2D-ResNet18+SimAM achieves (22.93% WWS) without introducing extra parameters. In addition, we follow [13] to pretrain the backbone model on a lip-reading database named CAS-VSR-W1k database [28], then have it fine-tuned on the MISP database. Finally, the 3D-ResNet18+2D-ResNet18+SimAM+Pretrain achieves the best result (19.13% WWS).

The results for the audio-visual fusion systems are shown in table 4. Moreover, we also compare our system with the previous SOTA result. We combine the best models from the respective modalities

**Table 4.** Comparison of different audio-visual systems.

ID	Model	Field	Dev[%]			Eval[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
VA1	Official [12]	Far	7.3	6.8	14.1	10.1	15	25.1
VA2	Xu et al. [14]	Far	-	-	-	-	-	9.1
VA3	Cheng et al. [13]	Far	3.85	3.42	7.27	-	-	7.1
VA4	MISP 2021 1st [11]	Far	-	-	4.1	-	-	5.8
VA5	Score-Level	Far	1.92	3.37	5.29	1.54	4.82	6.36
VA6	Cascaded Fusion	Far	5.29	2.3	7.59	7.29	3.5	10.79
VA7	HMA	Far	3.04	2.55	5.59	2.15	3.44	<b>5.59</b>

(A4+V5) and test multiple fusion strategies. The cascade system is not good enough compared to the score-level fusion because the distribution of scores of the two modalities is inconsistent, which plays a limited screening effect and requires fine threshold adjustment. We find that Score-Level fusion and Cascaded fusion do not fully utilize the information of the visual modality. In contrast, HMA can further utilize the features of the respective modality to give multimodal results finally. And, we obtain our final model by averaging the top-3 best models which have a lower loss on the dev set. Finally, we obtain a 21% relative reduction WSS compared with our previous work [13] and achieve the SOTA audio-visual result (5.59% WWS).

## 5. CONCLUSION

In this paper, we extend our previous work [13] to investigate robust audio-visual WWS system. We use a hybrid 3D and 2D convolution network to model the low-level spatial and high-level semantic information, respectively, while introducing the SimAM without additional parameters to improve the performance of the unimodal network further. We also explore different multimodal fusion schemes and find that HMA can take full advantage of the information from both modalities to give the best results. The performance of the system has been significantly improved compared with our previous work and achieved SOTA result (5.59 % WWS) on the MISP dataset.

## 6. ACKNOWLEDGMENTS

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011), Science and Technology Program of Suzhou City (SYC2022051), and Alibaba Air Fund. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

## 7. REFERENCES

- [1] Ming Sun, D. Snyder, Yixin Gao, Varun K. Nagaraja, Mike Rodehorst, S. Panchapagesan, N. Strom, Spyridon Matsoukas, and Shiv Vitaladevuni, “Compressed time delay neural network for small-footprint keyword spotting,” in *Proc. Interspeech*, 2017, pp. 3607–3611.
- [2] Tara N Sainath and Carolina Parada, “Convolutional neural networks for small-footprint keyword spotting,” in *Proc. Interspeech*, 2015.
- [3] Seungwoo Choi, Seokjun Seo, Beomjun Shin, Hyeonmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, and Sungjoo Ha, “Temporal Convolution for Real-Time Keyword Spotting on Mobile Devices,” in *Proc. Interspeech*, 2019, pp. 3372–3376.
- [4] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, “Wake word detection with streaming transformers,” in *Proc. ICASSP*, 2021, pp. 5864–5868, IEEE.
- [5] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Senior, “Deep audio-visual speech recognition,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [6] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan, “Recurrent neural network transducer for audio-visual speech recognition,” in *Proc. ASRU*, IEEE, 2019, pp. 905–912.
- [7] Pingchuan Ma, Stavros Petridis, and Maja Pantic, “End-to-end audio-visual speech recognition with conformers,” in *Proc. ICASSP*, IEEE, 2021, pp. 7613–7617.
- [8] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [9] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro, “Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition,” in *Proc. Interspeech*, 2022, pp. 2838–2842.
- [10] Liliane Momeni, Triantafyllos Afouras, Themis Stafylakis, Samuel Albanie, and Andrew Senior, “Seeing wake words: Audio-visual keyword spotting,” in *Proc. BMVC*, 2020.
- [11] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, et al., “The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results,” in *Proc. ICASSP*, IEEE, 2022, pp. 9266–9270.
- [12] Hengshun Zhou, Jun Du, Gongzhen Zou, Zhaoxu Nian, Chin-Hui Lee, Sabato Marco Siniscalchi, Shinji Watanabe, Odette Scharenborg, Jingdong Chen, Shifu Xiong, and Jian-Qing Gao, “Audio-Visual Wake Word Spotting in MISP2021 Challenge: Dataset Release and Deep Analysis,” in *Proc. Interspeech*, 2022, pp. 1111–1115.
- [13] Ming Cheng, Haoxu Wang, Yechen Wang, and Ming Li, “The dku audio-visual wake word spotting system for the 2021 misp challenge,” in *Proc. ICASSP*, IEEE, 2022, pp. 9256–9260.
- [14] Yanguang Xu, Jianwei Sun, Yang Han, Shuaijiang Zhao, Chaoyang Mei, Tingwei Guo, Shuran Zhou, Chuandong Xie, Wei Zou, and Xiangang Li, “Audio-visual wake word spotting system for misp challenge 2021,” in *Proc. ICASSP*, IEEE, 2022, pp. 9246–9250.
- [15] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie, “Simam: A simple, parameter-free attention module for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2021, pp. 11863–11874.
- [16] Xiaoyi Qin, Na Li, Chao Weng, Dan Su, and Ming Li, “Simple attention module based speaker verification with iterative noisy label detection,” in *Proc. ICASSP*, IEEE, 2022, pp. 6722–6726.
- [17] Xiaoyi Qin, Danwei Cai, and Ming Li, “Robust multi-channel far-field speaker verification under different in-domain data availability scenarios,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.
- [18] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proc. CVPR*, 2018, pp. 6450–6459.
- [19] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” in *Proc. CVPR*, 2018, pp. 7132–7141.
- [20] Menglong Xu and Xiao-Lei Zhang, “Depthwise Separable Convolutional ResNet with Squeeze-and-Excitation Blocks for Small-Footprint Keyword Spotting,” in *Proc. Interspeech*, 2020, pp. 2547–2551.
- [21] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, “Cbam: Convolutional block attention module,” in *Proc. ECCV*, 2018, pp. 3–19.
- [22] Ben S Webb, Neel T Dhruv, Samuel G Solomon, Chris Tailby, and Peter Lennie, “Early and late mechanisms of surround suppression in striate cortex of macaque,” *Journal of Neuroscience*, vol. 25, no. 50, pp. 11666–11675, 2005.
- [23] 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,” in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.
- [24] Mehrez Souden, Jacob Benesty, and Sofiène Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.
- [25] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, “Wake Word Detection with Alignment-Free Lattice-Free MMI,” in *Proc. Interspeech*, 2020, pp. 4258–4262.
- [26] Jiankang Deng, J. Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” *ArXiv*, vol. abs/1905.00641, 2019.
- [27] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proc. CVPR*, 2019, pp. 4690–4699.
- [28] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen, “Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,” in *Proc. FG 2019*, 2019, pp. 1–8.