ROBUST WAKE WORD SPOTTING WITH FRAME-LEVEL CROSS-MODAL ATTENTION BASED AUDIO-VISUAL CONFORMER

Haoxu Wang[1], *Ming Cheng*[1], *Qiang Fu*[2], *Ming Li*[1†]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China
[2]Research Center for Intelligent Robotics, Research Institute of Interdisciplinary Innovation,
Zhejiang Laboratory, Hangzhou, China

ABSTRACT

In recent years, neural network-based Wake Word Spotting achieves good performance on clean audio samples but struggles in noisy environments. Audio-Visual Wake Word Spotting (AVWWS) receives lots of attention because visual lip movement information is not affected by complex acoustic scenes. Previous works usually use simple addition or concatenation for multi-modal fusion. The inter-modal correlation remains relatively under-explored. In this paper, we propose a novel module called Frame-Level Cross-Modal Attention (FLCMA) to improve the performance of AVWWS systems. This module can help model multi-modal information at the frame-level through synchronous lip movements and speech signals. We train the end-to-end FLCMA based Audio-Visual Conformer and further improve the performance by fine-tuning pre-trained unimodal models for the AVWWS task. The proposed system achieves a new state-of-the-art result (4.57% WWS score) on the far-field MISP dataset.

Index Terms— audio-visual wake word spotting, frame-level cross-modal attention, pretrain strategy

1. INTRODUCTION

Wake word spotting (WWS), also called Keyword Spotting (KWS), is vital in speech signal processing. It aims to detect specific keywords or phrases in audio streams. It is crucial for voice-activated devices like smart speakers, mobile phones, and virtual assistants. Recently, deep neural networks have been used [1] for good performance in near-field clean speech environments, such as Convolution Neural Network (CNN) [2] and Transformer [3]. However, the performance of these systems may significantly degrade in far-field settings due to complex environments, e.g. speech overlap, background noise and etc. Some methods aim to enhance the noise robustness of WWS systems. For instance, researchers propose domain aware training systems [4], sample generation [5], and multi-look minimum variance distortion less response (MVDR) beamformers [6] to improve performance in far-field and complex scenarios.

Because visual lip movement information is not affected by acoustic noise and can serve as complementary information to the audio stream, the multi-modal audio-visual systems have become more and more popular in several fields, including automatic speech recognition (ASR) [7, 8, 9], speech separation [10], speaker verification (SV) [11], and etc. Audio-Visual systems show improved performance in high-noise environments compared to audio-only systems. [12] develops a new audio-visual KWS system based on CNN.

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

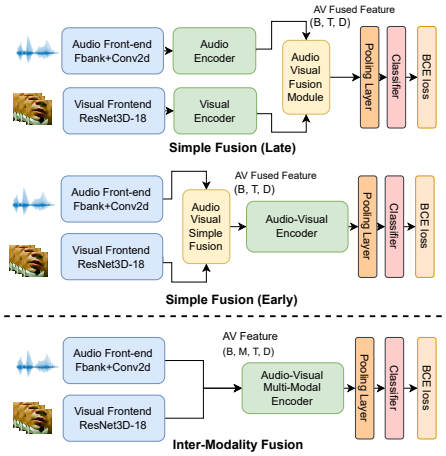

Fig. 1. Diagram of simple fusion method (Late: top, Early: middle) and our proposed inter-modality fusion method (bottom). B, T and D denote the batch size, the number of frames and the dimension of the embeddings. M represents the number of modalities.

Along with the first Multimodal Information Based Speech Processing Challenge (MISP Challenge 2021 [13]) and its data release [14], many new research works are reported targeting the Audio Visual Wake Word Spotting (AVWWS). [15] proposes a CNN-3D-based model, and [16] proposes a transformer-based model. [17] and [18] improve their systems to enhance performance further based on [15]. However, previous works [15, 17, 18] usually train two robust single-modality models and then fuse them. They do not use the end-to-end (E2E) strategy to optimize the network of two modalities simultaneously. [16] trains an audio-visual E2E Transformer model but still uses single-modality models to vote the final results. Moreover, [7, 8, 19] in the audio-visual speech recognition (AVSR) domain fuse the multi-modal information using simple addition or concatenation strategy.

Inspired by [20, 21, 22], which enhance multi-channel speaker diarization and ASR using Channel-Level Cross-Channel Attention (CLCCA) for frame-level correlation modeling of multi-channel speech signals, we propose the Frame-Level Cross-Modal Attention (FLCMA) module to improve the performance of the AVWWS system. Instead of simple addition or concatenation fusion, FLCMA models multi-modal semantic information at the frame level. We adopt an E2E training strategy, training an AVWWS Conformer for two modalities simultaneously to enhance system performance. Additionally, we utilize a pretrain strategy, pre-training robust unimodal models, transferring their parameters to the multi-modal

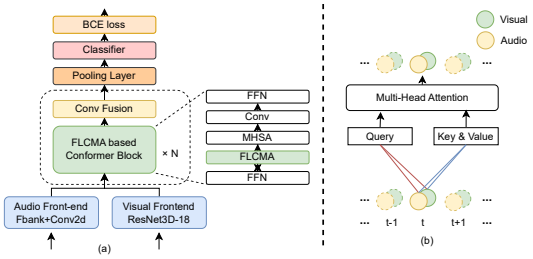

Fig. 2. (a) Framework of our FLCMA based audio-visual conformer wake word spotting system. (b) Diagram of FLCMA module. Each circle means audio or visual feature embedding at each time step.

model, and performing fine-tuning. This approach achieves a new state-of-the-art (SOTA) result (4.57% WWS score) on the MISP dataset, showcasing the effectiveness of our E2E AVWWS system.

2. METHODS

In this section, we introduce the Frame-Level Cross-Modal Attention module. As shown in Fig.1, assumed that we extract audio feature $X_{spec} \in \mathbf{R}^{T\times D}$ and visual feature $X_{lip} \in \mathbf{R}^{T\times D}$ of T frames, previous works usually get the audio-visual early fused feature by concatenating along feature dimension or adding, in which $X_{fused} = [X_{spec}||X_{lip}]W_{fc}$ or $X_{fused} = X_{spec} + X_{lip}$, where $||$ means concatenation and $W_{fc} \in \mathbf{R}^{2D\times D}$ means the fully-connected projection layer. These usually are called Simple Fusion (Early). Moreover, some works late fuse the multi-modal features after the single-modality encoder, usually called Simple Fusion (Late). These approaches only involve a simple fusion of multi-modal features and thus do not effectively model the correlations among multi-modal information. Thus, we propose the FLCMA module. We expand a new modal dimension for each single modality feature and define the multi-modal feature as $X_{av} = [X_{spec}, X_{lip}] \in \mathbf{R}^{M\times T\times D}, M = 2$ by concatenating along the new modal dimension.

2.1. Frame-Level Cross-Modal Attention

FLCMA focuses on modeling the inter-modality information on each time frame. As the Figure 2(b) shows, the attention module is defined as:

$$\begin{aligned} Q_i^{cm} &= X_{av}W_i^{cm,q} + (b_i^{cm,q})^T \in \mathbf{R}^{T\times M\times \frac{D}{h}}, \\ K_i^{cm} &= X_{av}W_i^{cm,k} + (b_i^{cm,k})^T \in \mathbf{R}^{T\times M\times \frac{D}{h}}, \\ V_i^{cm} &= X_{av}W_i^{cm,v} + (b_i^{cm,v})^T \in \mathbf{R}^{T\times M\times \frac{D}{h}}, \\ O_i^{cm} &= softmax(\frac{Q_i^{cm}(K_i^{cm})^T}{\sqrt{\frac{D}{h}}})V_i^{cm} \in \mathbf{R}^{T\times M\times \frac{D}{h}} \\ O^{cm} &= [O_0^{cm}, O_1^{cm}, ..., O_h^{cm}] \in \mathbf{R}^{T\times M\times D} \end{aligned} \tag{1}$$

where $Q_i^{cm}, K_i^{cm}, V_i^{cm}$ are the i-th head of Query, Key, Value of this attention module, $W_i^{cm,*}, b_i^{cm,*}$ are learnable parameters, and h is the number of attention heads. This FLCMA module can help capture inter-modality correlations at the frame level through the high synchronous lip movements and speech signal. By leveraging multi-modal information, this module provides complimentary lip movement for noisy acoustic information and reduces phoneme confusion for lip movement according to the speech signal.

2.2. Pretrain Strategy

The Pretrain strategy, widely employed in speech recognition [23] and WWS tasks [15, 18], has demonstrated its effectiveness. It involves pretraining an unsupervised model on an unlabeled database and fine-tuning it on a labeled database. Some approaches are to pretrain on one labeled database and fine-tune on another labeled database, known as transfer learning. In this work, we adopt the pretrain strategy by training single-modal models with a similar architecture to the multi-modal model. We transfer the parameters from the corresponding single-modal models to the multi-modal model, allowing it to leverage the knowledge acquired from the single-modal models. This approach provides a strong starting point for the multi-modal model and enhances its performance.

2.3. E2E Model Architecture

Here, we present our AVWWS system, the FLCMA based Audio-Visual Transformer or Conformer. As shown in Figure 2(a), our system consists of 6 modules: audio and visual frontend, Transformer or Conformer encoder with FLCMA module, convolution fusion module, attentive pooling layer and the final fully-connected classifier.

2.3.1. Front-end

For the visual stream, a ResNet-18 with 3D convolution layers is employed to transform the input video frames into temporal features. The visual inputs have a shape of (T, H, W, C), which is then squeezed along the spatial dimension using global average pooling to obtain a shape of (T, D). A linear layer is then used to project these temporal features to the dimension of the conformer encoders.

For the audio stream, a subsampling module comprising two 2D convolution layers is used to convert the extracted acoustic features into temporal features. A linear layer is then employed to project these features to the dimension of the encoders. As the video's sampling rate is generally lower than that of audio, the audio features are downsampled in the time dimension to $\frac{T'}{rate}$ after standard feature extraction, aligning the time dimension of audio and video ($\frac{T'}{rate} = T$). Here we get the audio and visual features X_{spec}, X_{lip}.

2.3.2. Encoder

We use the modified transformer or conformer block as our encoder layer. FLCMA based Transformer block includes a FLCMA module, a standard multi-head self-attention (MHSA) module, and a feed-forward (FFN) module. FLCMA based Conformer block includes a FLCMA module, a MHSA module, a convolution (CONV) module, and a pair of FFN modules in the Macaron-Net style. Standard conformer block combines CONV and MHSA to capture local and global correction in the single modality. Together with the FLCMA module, this modified conformer block can further leverage the inter-modality information at the frame level.

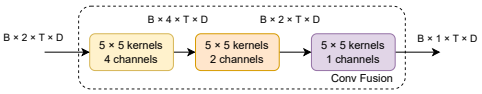

Fig. 3. Framework of convolution fusion module.

2.3.3. Convolution fusion

Inspired by [22], we use a convolution fusion module to fuse the audio-visual feature after the conformer encoder blocks instead of averaging or concatenation, which is usually used in previous works [7, 16]. As shown in Fig 3, We use a multi-layer convolution module to help reduce the corruption caused by the direct fusion of multi-modal features, which consists of a series 2D convolution module with the channel of 4, 2, 1.

2.3.4. Attentive Pooling and Classifier

Additionally, we incorporate an attentive pooling layer, commonly utilized in SV, to capture the significance of each frame and extract a more robust classification vector. This vector is then passed through a series of fully-connected linear layers with a sigmoid function to output the probability of the wake word.

3. EXPERIMENTAL SETUP

3.1. Dataset and Evaluation Metrics

We evaluate our proposed system on the AVWWS dataset from the 1st MISP challenge 2021 [13]. This dataset is utilized to detect the wake word 'Xiao T, Xiao T' spoken in far-field home scenarios. The released database contains about 125 hours and has two subsets: training set (47k+ negative samples and 5K+ positive samples) and development (Dev) set (2k+ negative samples and 600+ positive samples). Audio samples include single-channel near-field audio, 2-channel middle-field audio, and 6-channel far-field audio; video samples include single-person high-definition middle-field and multi-person far-field video. Moreover, an evaluation set (8K +) without annotations is provided to competition participants, which is only in the far-field. We obtain the annotations from the MISP committee to ensure a fair comparison of our results with those of other teams.

To evaluate our system's performance, we follow the guidelines provided by the competition committee. We utilize the False Reject Rate (FRR), False Alarm Rate (FAR), and the WWS Score. Let N_{wake} represent the number of samples containing the wake word, and N_{non_wake} represent the number of samples without the wake word. The FRR and FAR are defined as:

$$FRR = \frac{N_{FR}}{N_{wake}}, \quad FAR = \frac{N_{FA}}{N_{non\_wake}} \tag{2}$$

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

$$Score^{WWS} = FRR + FAR \tag{3}$$

3.2. Experimental Setup

Preprocess: For the audio stream, we extract log-mel filterbank features (FBank) with a dimension of 80 from the waveform using a window length of 25ms and a shift of 10ms. Each audio sample is sampled to contain 256 frames, which means that $T^{'}$ is set to 256. Furthermore, after the audio front-end module, the time dimension of the audio features is downsampled to a quarter of its original size and is aligned to the lip movements.

For the visual stream, we refer to [15], focus solely on the RGB lip region images of the video. We use a face detector (RetinaFace [24]) and a face recognizer (ArcFace [25]) to get the reference face in the far-field video and the face landmarks. We crop the lip regions of the target speaker from the facial images based on the detected facial landmarks according to [15, 26]. The lip-region videos are extracted and resized to a resolution of 112×112 with 3 RGB channels. The videos are sampled to contain 64 frames, resulting in a shape of $(64, 112, 112, 3)$ and meaning T is set to 64. Additionally, each pixel value in the video is normalized to the range of $[0, 1]$.

Data Augmentation: For the audio stream, we use various techniques inspired by [13, 15], including negative sub-segmentation, speed perturbation, slight trimming, and SpecAugment [27]. We also perform several additional steps on the original near-field audio to simulate middle and far-field audio. The pyroomacoustic tool is used to generate room impulse responses. We also incorporate noises provided by official sources, randomly adjusting the signal-to-noise ratio (SNR) within the range of -15 to 15 dB. The MVDR method is also used to add beamforming-enhanced audio into our training data.

For the visual stream, we also use the same video-based data augmentation methods referred to the [15], including speed perturbation, frame-wise rotation, horizontal flip, frame-level cropping, color jitters and gray scaling. Additionally, we incorporate random histogram equalization to augment the data further.

Model Training: For the FLCMA module based transformer or conformer structure, we use 6 self-attention blocks, each with 4 heads, 256-dimensional hidden size, and 1,024 dimensions for the feed-forward layer, which means that $D = 256, h = 4, N = 6$. The batch size is set to 48. The learning rate is set to 0.001 and warmed up at the first 10,000 steps by the Adam optimizer. And we adopt the weighted BinaryCrossEntropy (BCE) Loss (negative:positive=1:5) to tackle the imbalance between positive and negative samples. For the Pretrain strategy, we first train two uni-modal models, which consist of a frontend module, uni-modal encoder, pool layer and classifier using the above setup, then initialize our FLCMA based multi-modal model with the parameters of the uni-modal models, finally fine-tune the multi-modal model with the learning rate of 0.0001.

Table 1. *Ablation study results of our proposed FLCMA module and Pretrain strategy. For the sake of simplicity of presentation, all AUCs have 99.0 as the 0.0 baseline, i.e. 0.636 means 99.636. L means Late, and E means Early.*

Methods	Dev[%]				Eval[%]			
	AUC	FRR	FAR	WWS	AUC	FRR	FAR	WWS
AV-Transformer(L)	0.585	3.69	2.02	5.71	0.421	4.23	2.44	6.67
AV-Transformer(E)	0.232	2.72	4.76	7.48	0.083	1.66	7.29	8.95
+ FLCMA	0.453	1.76	4.18	5.94	0.518	1.04	6.02	7.06
+ Pretrain	0.645	2.24	2.21	4.45	0.576	2.14	3.33	5.47
AV-Conformer(L)	0.510	5.45	1.68	7.13	0.416	4.54	2.16	6.70
AV-Conformer(E)	0.456	4.97	1.92	6.89	0.231	4.78	2.99	7.77
+ FLCMA	0.617	2.24	2.31	4.55	0.541	1.59	3.91	5.50
+ Pretrain	**0.699**	2.08	1.78	**3.86**	**0.636**	2.02	2.55	**4.57**

Table 2. *Comparisons of recent released uni-modal systems.*

Methods	Dev[%]			Eval[%]		
	FRR	FAR	WWS	FRR	FAR	WWS
V-LSTM [14]	38.4	8.7	47.1	31.7	26.7	58.4
V-ResNet3D [15]	8.65	8.41	17.06	-	-	21.7
V-ResNet3D [17]	6.73	6.68	**13.41**	18.39	7.67	26.01
V-SimAM [18]	9.13	6.25	15.39	8.03	11.1	**19.13**
V-Transformer (Ours)	12.18	5.29	17.47	15.63	7.45	23.08
V-Conformer (Ours)	10.41	5.58	15.99	13.36	7.94	21.30
A-LSTM [14]	10.4	6.0	16.4	14.7	11.5	26.2
A-ResNet3D [15]	6.41	6.01	12.42	-	-	12.2
A-Conformer [16]	-	-	10.5	-	-	11.6
A-ResNet3D [17]	6.73	3.12	9.85	5.82	3.95	**9.78**
A-SimAM [18]	5.93	3.61	**9.54**	6.38	4.65	11.03
A-Transformer (Ours)	4.81	10.05	14.86	3.55	9.04	12.59
A-Conformer (Ours)	5.28	6.88	12.16	5.88	5.45	11.33

4. RESULTS AND DISCUSSIONS

4.1. Ablation Study

Table 1 shows the ablation study results of our proposed FLCMA module and Pretrain strategy. To further represent the models' performance, we also evaluate the models by calculating the area under the receiver operating characteristic curve (AUC). The AV-Transformer/Conformer(E) uses the standard transformer/conformer encoder block without the FLCMA module. These models consist of a simple concatenation fusion module before the encoder blocks, which is described as X_{fused} in Section 2. The AV-Transformer/Conformer(L) uses the two split uni-modal transformer/conformer encoders without the FLCMA module, with a simple concatenation fusion module after the encoder blocks. AV-Transformer(E) and AV-Conformer(E) achieve a performance of

Table 3. *Comparisons between pervious state-of-the-art audio-visual systems and ours. ∗ means the model with pretrain strategy.*

Model	Dev[%]			Eval[%]		
	FRR	FAR	WWS	FRR	FAR	WWS
Official [14]	7.3	6.8	14.1	10.1	15	25.1
Xu et al. [16]	-	-	-	-	-	9.1
Cheng et al. [15]	3.85	3.42	7.27	-	-	7.1
MISP 2021 1st [13]	-	-	4.1	-	-	5.8
Zhang et al. [17]	1.60	2.02	**3.62**	2.79	2.95	5.74
Wang et al. [18]	3.04	2.55	5.59	2.15	3.44	5.59
FLCMA-Transformer	1.76	4.18	5.94	1.04	6.02	7.06
FLCMA-Transformer*	2.24	2.21	4.45	2.14	3.33	5.47
FLCMA-Conformer	2.24	2.31	4.55	1.59	3.91	5.50
FLCMA-Conformer*	2.08	1.78	3.86	2.02	2.55	**4.57**

99.083% and 99.231% AUC, 8.95% and 7.77% WWS on the eval set, which shows that training from scratch using E2E training strategy can leverage the audio-visual multi-modal features at the same time and achieves good performance. AV-Transformer(L) and AV-Conformer(L) achieve a performance of 99.421% and 99.416% AUC, 6.67% and 6.70% WWS on the eval set. This suggests that fusing deep multi-modal latent features from the encoders is superior to fusing shallow features from the basic frontend module.

The FLCMA module enhances the performance of the E2E AV-Transformer/Conformer system. The FLCMA-based AV-Transformer achieves 99.518% AUC and 7.06% WWS on the eval set, outperforming the baseline AV-Transformer(E). Similarly, the FLCMA-based AV-Conformer achieves better 99.541% AUC and 5.50% WWS on the eval set. The result shows that the FLCMA module can model the correlation between modalities at the frame level through the high synchronous lip movements and speech signal. FLCMA based AV-Transformer is slightly worse than AV-Transformer(L) at the WWS score but is better at the AUC on the eval set, we find that the selection of thresholds easily influences the scores of WWS, and the classification ability reflected by AUC can be used as another reference. Compared to AV-Transformer/Conformer(L), FLCMA based AV-Transformer/Conformer achieves better AUC performance (+0.097/+0.125) on the eval set and shows the effectiveness of the FLCMA module when using a single parameter-less audio-visual multi-modal encoder instead of two split uni-modal encoders.

Finally, combined with the Pretrain strategy, both FLCMA based AV-Transformer/Conformer models improve their performance. Our final system (FLCMA-based AV-Conformer with Pretrain strategy) has a further 17% reduction on WWS score, eventually reaching the WWS of 4.57%.

4.2. Results of the Pretrained single-modality model

Table 2 shows the performance of recent uni-modal systems on the MISP dataset. Our system outperforms the official baseline [14] and slightly underperforms the recent unimodal systems. We find that even though the uni-modal system for the multi-modal pretrained system does not achieve the best, it can further improve the performance of our multi-modal E2E system.

4.3. Results compared with previous works

Table 3 compares the performance of our system with previous multi-modal approaches. [18, 17] first train two robust single-modality models and then fuse these two frozen models together. Our FLCMA based Transformer with pretrain strategy and our FLCMA based Conformer achieve 5.47% and 5.50% WWS on the eval set, which are better compared with the previous works. Our model slightly underperforms [17] on the dev set but outperforms it on the eval set. It does not overfit on the dev set, demonstrating better generalization performance. Finally, our FLCMA based Conformer with pretrain strategy achieves the SOTA audio-visual result (4.57% WWS).

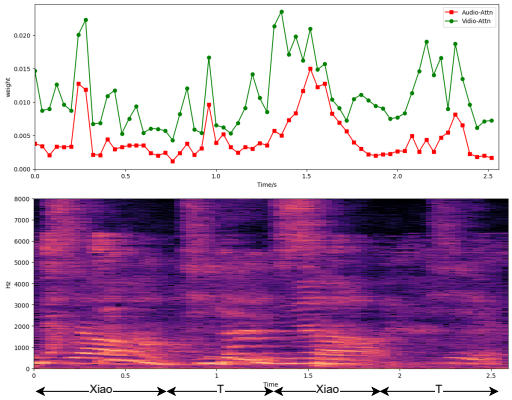

Fig. 4. Visualization of the attention weights of the FLCMA based Audio-Visual Conformer model.

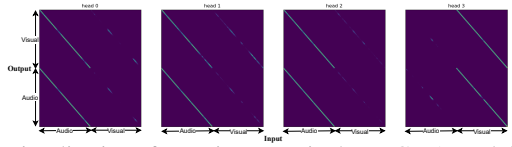

Fig. 5. Visualization of attention maps in the FLCMA module of the last encoder conformer block. The head means the four head in the FLCMA module.

4.4. Visualization of the FLCMA module

We use the attention rollout [28] to visualize the attention weights of the FLCMA based Audio-Visual Conformer model. Fig. 4 illustrates higher attention weights around 0.25s, 1.0s, 1.5s, and 2.3s, corresponding to identifiable features of the wake word. We also visualize the attention maps in the FLCMA module of the last encoder conformer block. As shown in Fig. 5, the features in the same modality have different attention weights at different frames. The model can customize attention weights at different frames, enabling it to capture inter-modality correlations. If one modality lacks confidence, the model can fuse information from multiple modalities for enhanced performance.

5. CONCLUSION

In this work, we introduce the Frame-Level Cross-Modal Attention (FLCMA) module to enhance the performance of the Audio-Visual Wake Word Spotting system. This module enables the modeling of multi-modal semantic information at the frame level by leveraging synchronous lip movements and speech signals. We train the end-to-end FLCMA based Audio-Visual Conformer and further improve the performance by fine-tuning pre-trained uni-modal models. The proposed system achieves a new state-of-the-art result (4.57% WWS score) on the MISP dataset, demonstrating the effectiveness of the approach.

6. ACKNOWLEDGMENTS

This research is funded in part by the National Natural Science Foundation of China (62171207, 52105128), Science and Technology Program of Suzhou City (SYC2022051) and Youth Foundation Project of Zhejiang Lab (K2023BA0AA03). 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, pp. 1478–1482.

[3] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, "Wake Word Detection with Streaming Transformers," in *Proc. ICASSP*, 2021, pp. 5864–5868.

[4] Haiwei Wu, Yan Jia, Yuanfei Nie, and Ming Li, "Domain Aware Training for Far-Field Small-Footprint Keyword Spotting," in *Proc. Interspeech*, 2020, pp. 2562–2566.

[5] Haoxu Wang, Yan Jia, Zeqing Zhao, Xuyang Wang, Junjie Wang, and Ming Li, "Generating TTS Based Adversarial Samples for Training Wake-Up Word Detection Systems Against Confusing Words," in *Proc. Odyssey*, 2022, pp. 402–406.

[6] Yueyue Na, Ziteng Wang, Liang Wang, and Qiang Fu, "Joint Ego-Noise Suppression and Keyword Spotting on Sweeping Robots," in *Proc. ICASSP*, 2022, pp. 7547–7551.

[7] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," in *Proc. ICASSP*, 2021, pp. 7613–7617.

[8] Maxime Burchi and Radu Timofte, "Audio-Visual Efficient Conformer for Robust Speech Recognition," in *Proc. WACV*, 2023, pp. 2257–2266.

[9] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2118–2122.

[10] Ruohan Gao and Kristen Grauman, "VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency," in *Proc. CVPR*, 2021, pp. 15490–15500.

[11] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf, "A Multi-View Approach To Audio-Visual Speaker Verification," in *Proc. ICASSP*, 2021, pp. 6194–6198.

[12] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman, "Seeing Wake Words: Audio-Visual Keyword Spotting," in *Proc. BMVC*, 2020.

[13] 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*, 2022, pp. 9266–9270.

[14] 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.

[15] 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*, 2022, pp. 9256–9260.

[16] 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*, 2022, pp. 9246–9250.

[17] Ao Zhang, He Wang, Pengcheng Guo, Yihui Fu, Lei Xie, Yingying Gao, Shilei Zhang, and Junlan Feng, "VE-KWS: Visual Modality Enhanced End-to-End Keyword Spotting," in *Proc. ICASSP*, 2023, pp. 1–5.

[18] Haoxu Wang, Ming Cheng, Qiang Fu, and Ming Li, "The DKU Post-Challenge Audio-Visual Wake Word Spotting System for the 2021 MISP Challenge: Deep Analysis," in *Proc. ICASSP*, 2023, pp. 1–5.

[19] 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.

[20] Weiqing Wang, Xiaoyi Qin, and Ming Li, "Cross-Channel Attention-Based Target Speaker Voice Activity Detection: Experimental Results for M2MeT Challenge," in *Proc. ICASSP*, 2022, pp. 9171–9175.

[21] Shota Horiguchi, Yuki Takashima, Paola Garcia, Shinji Watanabe, and Yohei Kawaguchi, "Multi-Channel End-to-End Neural Diarization with Distributed Microphones," in *Proc. ICASSP*, 2022, pp. 7332–7336.

[22] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yuhao Liang, Zhihao Du, Yuxiao Lin, and Lei Xie, "MFCCA:Multi-Frame Cross-Channel attention for multi-speaker ASR in Multi-party meeting scenario," in *Proc. SLT*, 2023, pp. 144–151.

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proc. CVPR*, June 2020.

[25] 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.

[26] 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, pp. 1–8.

[27] 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, pp. 2613–2617.

[28] Samira Abnar and Willem Zuidema, "Quantifying Attention Flow in Transformers," in *Proc. ACL*, 2020, pp. 4190–4197.