THE WHU WAKE WORD LIPREADING SYSTEM FOR THE 2024 CHAT-SCENARIO CHINESE LIPREADING CHALLENGE

Haoxu Wang[1], *Cancan Li*[1], *Fei Su*[1], *Juan Liu*[1], *Hongbin Suo*[3], *Ming Li*[1,2†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Data Science Research Center, Duke Kunshan University, Kunshan, China
[3]AI Center, OPPO, Beijing, China

ABSTRACT

The paper describes the Wake Word Lipreading system developed by the WHU team for the ChatCLR Challenge 2024. Although Lipreading and Wake Word Spotting have seen significant development, exploration of pretrained frontends for Wake Word Lipreading (WWL) remains insufficient. Our system is built upon a pretrained frontend and Transformer-liked backend architecture, incorporating Attentive Pooling and a Classifier. We investigate the effectiveness of different frontends, including Auto-AVSR and AV-Hubert, and evaluate the performance of Conformer and E-Branchformer backends. Additionally, we introduce Multi-layer Feature Aggregation to leverage features from multiple encoder block layers, demonstrating its effectiveness. Finally, we apply various fusion strategies, leading to score fusion that achieved a false reject rate of 8.21% and a false alarm rate of 8.50% along with a WWS score of 16.71% on the evaluation set, and obtain the first place in the task 1 of the ChatCLR Challenge.

Index Terms— Wake Word Lipreading, Keyword Spotting, Pretrain, Multi-layer Feature Aggregation

1. INTRODUCTION

Wake Word Lipreading (WWL) involves detecting whether a speaker says a specific wake word within a silent video stream. It is a similar task to Wake Word Spotting (WWS), Keyword Spotting (KWS) or Audio Visual Wake Word Spotting (AVWWS). WWS and AVWWS aim to detect predefined wake words in audio streams or using both audio-visual streams. As a sub-task of lipreading, it plays a significant role in video recognition tasks. In situations with complex acoustic conditions such as background noises (cheers, TV or screams), reverberations, and conversational multi-speaker interactions with a significant portion of speech overlaps) or where acoustic input is inaccessible, recognizing the speech content of the speaker in the video through visual modality becomes crucial and necessary. Additionally, in medical applications, it assists individuals with language disorders or aphasia [1] in communicating with smart devices solely through lip movements. In intelligent manufacturing and smart driving scenarios, users can use various modalities, including simple lip movements without vocalizing, to activate smart devices during conversational interactions [2].

Along with the first Multimodal Information Based Speech Processing Challenge (MISP Challenge 2021 [3]) and its data release [4], much research is being applied to AVWWS [5–9]. However, existing systems primarily concentrate on enhancing audio system performance in noisy wake-up scenarios by leveraging visual cues. Visual-only system research remains under-explored. WWL poses challenges due to visual homophemes, where phonemes have similar lip movements despite different vocalizations (e.g., 't,' 'n,' 'd' in 'tight,' 'night,' 'dight'). To improve the performance of lipreading, the 2024 Chat-scenario Chinese Lipreading Challenge (ChatCLR)[1] is launched. Task 1 WWL focuses on activating smart home devices during conversational interactions using slient far-field videos. To develop our WWL system, we first explore previous systems [8, 9] and then design a system with the architecture based on a pretrained frontend and Transformer-like backend structure, integrating Attentive Pooling and a Classifier. We investigate the effectiveness of different frontends (including Auto-AVSR [10] and AV-Hubert [11]), and different backends (including Conformer [12] and E-Branchformer [13]). Furthermore, we introduce Multi-layer Feature Aggregation (MFA) to leverage features from multiple encoder block layers. Finally, we explore various fusion strategies and achieve a false reject rate of 8.21% and a false alarm rate of 8.50% along with a WWS score of 16.71% on the evaluation set, and obtain the first place in the task 1 of the ChatCLR Challenge.

2. RELATED WORKS

Lipreading recognizes text content from the speaker's lip movements in silent videos. Early Lipreading research focuses on manually designing visual features and using statistical models [14]. The advent of large-scale Lipreading datasets like LRS2 [15] and LRS3 [16], coupled with advancements in deep learning, has improved Lipreading technology forward. Early efforts primarily concentrate on word-level recognition [17, 18], while later work shifts towards sequence-to-sequence (S2S) tasks [19]. [20] proposes a system based on Transformer [21] architectures using CTC and S2S loss. [22] future proposes a system based on Conformers [12] for Lipreading. [23] proposes an attention-based pooling mechanism to aggregate visual speech representations to improve performance.

Keyword Spotting aims 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 the deep neural networks (DNN) [24], convolutional neural networks (CNN) [25], temporal convolutional neural networks [26], text-to-speech data augmentation [27] and Transformer [28]. Because visual lip movement information is not affected by acoustic noise and can

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

Acknowledge: This research is funded in part by the National Natural Science Foundation of China (62171207) and OPPO. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

[1] https://mispchallenge.github.io/ICME2024/index.html

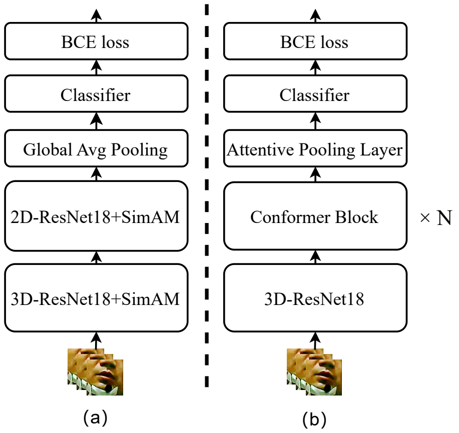

Fig. 1. (a) Framework of previous 3D-ResNet18+2D-ResNet18+SimAM system. (b) Framework of previous Visual-Conformer system.

serve as complementary information to the audio stream, some works investigate the multi-modal audio-visual systems [29, 30]. [31] develops a new audio-visual KWS system based on CNN. [5] proposes a CNN-3D-based AVWWS model, and [6] proposes a transformer-based AVWWS model. [7, 8] improve their systems to enhance performance further based on [5]. [9] proposes a Frame-Level Cross-Modal Attention (FLCMA) module to improve the performance of the AVWWS system.

Semi- and Self-Supervised Learning focuses on improving systems with untranscribed data. [32, 33] propose various self-supervised frameworks to learn audio and visual representations from large unlabelled datasets. [34] uses pre-trained multiple modalities models to teach a VSR network by using knowledge distillation. [11] learns powerful audio-visual speech representation benefiting both VSR and ASR by using mask prediction loss. [10] generates pseudo-labels for the unlabelled data using pretrained ASR models to help train a robust VSR model.

3. METHODS

3.1. Previous Systems

3.1.1. 3D-ResNet18+2D-ResNet18+SimAM

[5, 8] have shown the effectiveness of 3D convolution in previous video-only WWS. Compared to 2D convolution, which models spatial dimensions only, 3D convolution adds a temporal dimension for better performance. Using a mixture of 3D CNN and 2D CNN in a network allows 3D CNNs in lower layers to focus on short-term spatial modeling, while 2D CNNs in higher layers extract temporal information. Moreover, compared to channel-wise squeeze-excitation (SE) [35] and Convolutional block attention module (CBAM) [36], Simple Attention Module (SimAM) [37] serves as an attention mechanism that simultaneously considers both channel and spatial dimensions. It utilizes spatial suppression mechanisms to calculate attention weights, leading to improved performance. Furthermore, SimAM requires no additional model weight modules compared to SE and CBAM, as it computes attention weights for each feature map without any extra parameters, requiring only a certain increase in computational. The robust 3D-ResNet18+2D-ResNet18+SimAM [8] system achieves good performance for WWL. As shown in Fig.1 (a), the 3D-ResNet18 processes the raw RGB image sequences with the shape of (T, H, W, 3), and the feature map is averaged along the spatial axis after the 3D-ResNet18 to the shape of (T, C), and then the 2D-ResNet18 processes the hidden feature as a one-channel image with the shape of (1, T, C).

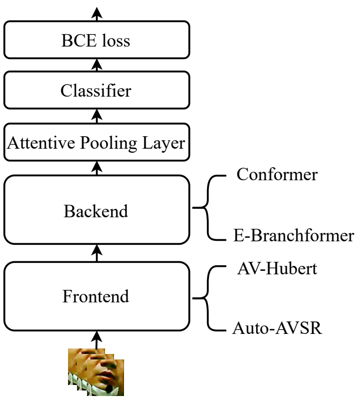

Fig. 2. Framework of our proposed system with frontend and backend.

3.1.2. Visual-Conformer

In prior work [6], an AV-Transformer model for the WWL task is proposed, inspired by Visual Transformer (ViT) [38]. The model operates on video feature sequences $X \in \mathbb{R}^{T\times D}$, with the addition of a special <cls>token $X_{cls} \in \mathbb{R}^{1\times D}$ at the sequence start, leveraging attention mechanisms for information exchange. Subsequently, this <cls>token is used as the final classification vector. In contrast, [9] introduces a Visual-only WWS System based on Transformer and Conformer, while presenting an End-to-End AVWWS system. Illustrated in Fig. 1 (b), this system includes a 3D-ResNet18 frontend, followed by an Encoder based on Transformer or Conformer, and subsequent Attentive Pooling and Classifier components. The 3D-ResNet18 frontend transforms input video frames of shape (T, H, W, C) into temporal features (T, D) via global average pooling along the spatial dimension. A linear layer then projects these temporal features to match the encoder block's dimension. Unlike [6], this system does not use a <cls>token for classification at the sequence beginning. Instead, it utilizes Attentive Pooling to aggregate all feature sequences into a single vector for classification. We implement a visual-only WWS system based on Conformer, referred to as Visual-Conformer.

3.2. Proposed Systems

With the advancements in Lipreading [19, 20, 22, 39], Transformer-like models have gradually become a paradigm. As shown in Fig. 2, we can abstract the aforementioned model Visual-Conformer in Sec. 3.1.2 into a frontend semantic feature extraction network and a backend semantic feature analysis network. The frontend and backend network of visual-Conformer are the 3D-ResNet18 and the Conformer blocks. The frontend feature extraction network processes the 4-dimensional raw video frames (T, H, W, C) to 2-dimensional (T, D) features $X_{video} \in \mathbb{R}^{T\times D}$ of T frames, similar to the speech-based features $X_{spec} \in \mathbb{R}^{T\times D}$ (MFCC, FBank, etc.). It transforms the original video into a sequence of features, standardizing the input format for subsequent backend Transformer-like networks.

Past studies have demonstrated that utilizing more data can improve model performance, and fine-tuning additional simple network structures on pre-trained models (such as BERT [40]) can achieve good results in downstream tasks. In order to further enhance the effectiveness of the frontend model, we explore pretrained Auto-AVSR [10] and AV-Hubert [11] models to replace our original 3D-ResNet18 frontend model.

3.2.1. *Pretrained Frontend*

Auto-AVSR: The model adopts the off-the-shelf architecture presented in [22], which achieves good performance on LRS2 and LRS3 without utilizing any additional training data. The system consists of a visual frontend, visual encoder, and visual decoder. It employs a modified ResNet18 as its frontend for lip movements, followed by an encoder with stacked Conformer blocks and a Transformer-based decoder, ultimately trained using CTC/S2S attention loss for VSR tasks. The advantage of Auto-AVSR lies in leveraging a large amount of additional data generated by efficient ASR models, producing pseudo-labels for the audios from AVSpeech [41] and VoxCeleb2 [42]. Experimental results demonstrate that a large amount of pseudo-labeled data can enhance the performance of the VSR model. Finally, it combines these pseudo-labels from AVSpeech and VoxCeleb2 and oracle transcription labels from the training sets of LRS2 and LRS3 to train its Conformer-based VSR model, achieving SOTA performance on LRS2 and LRS3. We remove its Transformer Decoder and adopt its ResNet18 and Conformer Encoder as our pretrained frontend.

AV-Hubert: This recent and advanced unsupervised audio-visual model, known as AVHubert, utilizes audio and visual data to predict cluster labels, achieving effective unsupervised training. AV-Hubert builds upon the Audio Hubert framework [33], which is a self-supervised framework for training audio-based models. Hubert's training involves two stages: feature clustering and masked prediction. Predicting cluster labels for masked regions allows the model to leverage unmasked areas to learn local representations and long-range temporal dependencies among latent features. Iteratively, these stages enhance both the quality of clustering and the feature representation capability.

AV-Hubert extends the Audio Hubert framework by incorporating ResNet18 for video processing in the frontend and fully-connected layers for audio processing and downsampling. This ensures that both modalities achieve consistent feature dimensions (B, T, D). Features from both modalities are concatenated into (B, T, 2D) and fed into an Encoder with stacked Transformer blocks for information exchange and cluster label prediction. The system uses Modality Dropout to maintain robust feature representation capabilities, even with only the input of one modality. After pretraining with unlabeled data, additional decoders or CTC classification layers can be added for supervised training, yielding good performance in VSR tasks. We use its ResNet18 and Transformer Encoder as our pretrained frontend.

3.2.2. *Backend*

Conformer: We investigate the Conformer encoder's effectiveness, known for its effectiveness in ASR [12], AVWWS [9], and VSR [22] tasks. The Conformer block includes a multi-head self-attention (MHSA) module, a convolution (CONV) module, and a pair of feed-forward network (FFN) modules in the Macaron-Net style. This block combines CONV and MHSA to capture both local and global information from the visual features.

E-Branchformer: We investigate the performance of the E-Branchfor mer [13] architectures. The E-Branchformer is an improved version of Branchformer [43]. The Branchformer encoder integrates two parallel branches to capture diverse contextual ranges. While one branch utilizes self-attention mechanisms to grasp long-range dependencies, the other branch employs a multi-layer perceptron module with convolutional gating (cgMLP) to extract intricate local correlations concurrently. Additionally, [13] augments the Branchformer by incorporating a depth-wise convolution-based merging module and integrating an extra pointwise feedforward module, thereby introducing the E-Branchformer.

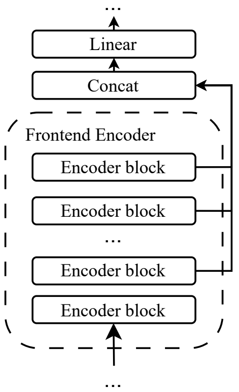

Fig. 3. Illustration of Multi-layer Feature Aggregation on the results of the penultimate layers of the encoder.

3.2.3. *Multi-layer Feature Aggregation*

As shown in Fig. 3, we propose the Multi-layer Feature Aggregation by utilizing features from multiple layers of the pretrained frontend instead of solely relying on the output of the final layer, inspired by MFA-Conformer [44]. The approach involves concatenating the output feature maps from these last N selected encoder blocks and passing them through a linear layer to ensure compatibility with the input dimensions of the backend.

$$\begin{aligned} H^{'} &= Concat(h_L, h_{L-1}, ..., h_{L-N+1}) \\ H &= Linear(H^{'}) \end{aligned} \tag{1}$$

where L represents the number of encoder layers in frontend, $h_{L-n}, n \in [0, N]$ represents the feature from the last n encoder layer. If the dimension of the input of the backend is D, then the $H \in \mathbb{R}^{T \times D}$ represents the output feature of the MFA module.

3.2.4. *Attentive Pooling and Classifier*

After we get the output feature from the backend, we send it to an attentive pooling layer, a technique commonly employed in Speaker Verification (SV), to capture the weighted summation of feature sequences and get a more robust classification vector. Subsequently, this vector is sent to a sequence of fully-connected linear layers with a final sigmoid function, facilitating the generation of the wake word probability output.

3.3. Fusion Strategy

Integrating the results of multiple systems mitigates the risk of overfitting in individual systems, enhances the overall model's generalization ability, and improves system robustness. Thus, multiple-system fusion not only improves the performance of WWL systems but also enhances their stability and reliability in practical applications.

Vote Fusion: If we have multiple systems (Sys_1, Sys_2, ..., Sys_n), we can adopt a majority vote method to integrate these models and obtain the final result. Each model determines the current result as either 0 or 1 based on its threshold, and then the mode of all model results is taken as the final result.

Score Fusion: If we have multiple systems (Sys_1, Sys_2, ..., Sys_n), we can average the output score of each model to obtain a fused score. This fused score is then compared with a threshold to determine whether it corresponds to 0 or 1.

Table 1. *Performance of various systems in the development (Dev) and evaluation (Eval) sets. MFA represents the Multi-layer Feature Aggregation. The $*$ represents retraining the previous systems using the current dataset. The ♡ represents unfreezing the pretrained frontend for finetuning. The ♣ represents finetuning with freezing the pretrained frontend.*

ID	Model	Params[M]	Dev[%]				Eval[%]			
			AUC	FRR	FAR	WWS	AUC	FRR	FAR	WWS
1	3D-ResNet18+2D-ResNet18+SimAM+Pretrain [8]	11.10	98.246	6.60	5.20	11.80	94.416	11.66	12.70	24.36
2	3D-ResNet18+2D-ResNet18+SimAM+Pretrain*	11.10	97.731	8.09	4.43	12.52	94.608	15.22	8.70	23.92
3	Visual-Conformer [9]	17.64	98.136	8.72	4.61	13.33	93.544	17.45	9.46	26.91
4	Visual-Conformer*	17.64	98.383	10.43	4.19	14.62	93.805	17.99	8.43	26.42
5	Auto-AVSR+Conformer(S)	191.48	96.970	8.94	7.92	16.86	91.767	18.06	13.52	31.58
6	AV-Hubert(a)+Conformer(S)	334.54	98.296	4.26	5.49	9.75	95.294	8.94	12.78	21.72
7	AV-Hubert(a)+E-Branchformer(S)	336.99	98.209	6.60	5.49	12.09	96.096	8.09	12.31	20.40
8	AV-Hubert(b)+Conformer(L)+MFA(1L)	472.13	98.619	3.83	5.32	9.15	95.532	7.19	13.77	20.96
9	AV-Hubert(b)+Conformer(L)+MFA(3L)	475.28	98.603	4.89	3.73	8.62	95.816	9.60	11.30	20.90
10	AV-Hubert(b)+Conformer(L)+MFA(6L)	478.43	98.683	4.47	5.20	9.67	95.745	7.13	13.64	20.77
11	AV-Hubert(c)+Conformer(L)+MFA(1L)	472.13	98.516	5.96	5.08	11.04	94.642	10.81	12.42	23.23
12	AV-Hubert(c)+Conformer(L)+MFA(3L)	475.28	98.693	6.17	4.08	10.25	95.784	9.00	11.63	20.63
13	AV-Hubert(c)+Conformer(L)+MFA(6L)	478.43	98.785	3.62	5.02	8.64	95.632	10.45	11.83	22.28
14	AV-Hubert(a)+Conformer(L)+MFA(1L)	472.13	98.434	3.62	5.38	9.00	95.723	6.64	13.30	19.94
15	AV-Hubert(a)+Conformer(L)+MFA(3L)	475.28	98.658	2.77	5.08	7.85	95.705	8.45	11.87	20.32
16	AV-Hubert(a)+Conformer(L)+MFA(6L)	478.43	98.760	2.98	5.26	8.24	96.016	7.13	12.81	19.94
17	AV-Hubert(a)+Conformer(L)+MFA(1L)$^\heartsuit$	472.13	98.972	5.53	3.37	8.90	96.525	8.28	10.79	19.07
18	AV-Hubert(a)+Conformer(L)+MFA(1L)$^\clubsuit$	472.13	98.447	5.96	4.14	10.10	96.118	8.58	9.76	18.34
19	AV-Hubert(a)+Conformer(L)+MFA(3L)$^\clubsuit$	475.28	98.616	5.11	3.96	9.07	95.976	9.66	9.92	19.58
20	AV-Hubert(a)+Conformer(L)+MFA(6L)$^\clubsuit$	478.43	98.840	3.62	3.78	7.40	96.301	8.82	9.37	18.19
21	VoteFusion(ID17 + ID18 + ID20)	-	-	-	-	-	-	8.09	8.89	16.98
22	ScoreFusion(ID17 + ID18 + ID20)	-	-	-	-	-	96.655	8.21	8.50	16.71

4. EXPERIMENTS

4.1. Dataset and Evaluation Metrics

The dataset is provided by the 2024 ChatCLR Challenge Task 1. The dataset is also similar to the previous database in the First MISP Challenge [3]. The organizers double-check all data and fix it, then release all the database [4][2] and regrade it as the dataset in this 2024 ChatCLR Challenge Task 1. The dataset covers audio and video related data, and we basically only use video data in this WWL task. This dataset is utilized to detect the wake word 'Xiao T, Xiao T' spoken in far-field home scenarios in the silent video.

The released database has two subsets: training set (45k+ negative samples and 5K+ positive samples) and development (Dev) set (1.6k+ negative samples and 470 positive samples). Video samples include single-person high-definition middle-field and multi-person far-field video. Moreover, a new evaluation (Eval) set (8K+) without annotations is provided to competition participants, which is only in the far-field. The ChatCLR committee releases the annotations to all the teams after the challenge to report detailed results.

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:

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

To further represent the models' performance, we also evaluate the models by calculating the area under the receiver operating characteristic curve (AUC).

4.2. Preprocessing

Region of Interest (RoI): Our system focuses solely on the lip region as input rather than the entire face. We use the RetinaFace model [45] for facial detection, extracting all faces and their five facial landmarks. In wide-angle videos with multiple faces, we reference official oracle results and use the nearest distance principle to identify specific faces for testing. Following [8], we crop the lip region using facial landmarks to generate lip movement videos. All lip movement videos are resized to 112×112 with 3 RGB channels.

Data Augmentation: We use the same video-based data augmentation methods referred to the [8], including speed perturbation, frame-wise rotation, horizontal flip, frame-level cropping, color jitters, gray scaling and histogram equalization. The probability for all data augmentation strategies is set to 0.5.

Model Training: For Conformer(S) and E-Branchformer(S) structures, we use 6 self-attention blocks with 4 heads, a 256-dimensional hidden size, and a feed-forward layer of 1,024 dimensions ($D = 256, h = 4, N = 6$). For Conformer(L), we use $D = 1024, h = 4, N = 6$. The batch size is 16 with a learning rate of 0.002, warmed up for the first 2,000 steps using the Adam optimizer. We use weighted

[2] https://challenge.xfyun.cn/misp_dataset

BinaryCrossEntropy (BCE) Loss (negative:positive=1:5) to address sample imbalance. Previous systems sampled videos with 64 frames, resulting in a shape of (64, 112, 112, 3). For Auto-AVSR, lip videos are resized to 96×96 with 1 channel, and for AV-Hubert, to 88×88 with 1 channel.

Table 2. *The information of different pretrained frontend checkpoints. T represents Type. P represents the parameters used in our frontend. TS represents the training strategy. Pse represents the Pseudo-label data. Unl represents the Unlabeled data. Lab represents the Labeled data. NA represents the Noise-Augmented. WER represents the result of the checkpoint on LRS3.*

Pretrain Frontend	T	P[M]	TS	Pse (h)	Unl (h)	Lab (h)	WER[%]
Auto-AVSR	-	182.03	-	2630	-	818	19.1
AV-Hubert	a	325.03	-	1326	1759	433	26.9
	b	325.03	-	-	1759	-	-
	c	325.03	NA	-	1759	-	-

4.3. Results

4.3.1. Previous systems

We first evaluate the performance on the current Eval set using checkpoints from previous studies [8] and [9], yielding results corresponding to ID1 and ID3 in Table 1. Furthermore, we retrain the 3D-ResNet18+2D-ResNet18+SimAM+Pretrain and Visual-Conformer model with the revised training dataset from the current competition, presenting results as ID2 and ID4. Compared to previous checkpoints, the performance improves. Analysis revealed that the MISP2021 competition training set lacked oracle labels to select the correct speaker in multi-person far-field videos, requiring a face recognition tool for identification. This leads to errors in speaker selection during lip movement training. Conversely, the updated training set includes oracle target person labels, preventing the selection of incorrect far-field speaker lip movements and reducing error accumulation.

4.3.2. Results of proposed systems

Pretrained Frontend: As described in Sec 3.2.1, we utilize pretrained Auto-AVSR and AV-Hubert models as our frontend. Details are provided in Table 2. For AutoAVSR, the selected checkpoint[3] achieves a 19.1% Word Error Rate (WER) on LRS3. AV-Hubert is utilized with three checkpoints: (a) a checkpoint[4] pretrained on LRS3 + VoxCeleb2(En) and finetuned with Self-Training and annotated data, achieving a 26.9% WER on LRS3; (b) an original unsupervised checkpoint[5] pretrained on LRS3 + VoxCeleb2(En); and (c) an unsupervised checkpoint[6] with Noise-Augmentation applied to the audio part during pretraining. As shown in Table 1 comparing ID5 with ID6, Auto-AVSR outperforms AV-Hubert(a) on LRS3 but underperforms for the current task. This could be attributed to Auto-AVSR not incorporating audio information. Additionally, unsupervised models may exhibit better robustness than supervised models in downstream tasks.

Backend: We also explore the use of E-Branchformer as our backend. As shown in Table 1 comparing ID6 with ID7, E-Branchformer exhibits slightly lower performance on the Dev set compared to Conformer, but outperforms Conformer on the Eval set. Due to time constraints during the challenge, we focus our further exploration only on the downstream aspects of the Conformer structure.

Parameters: Comparing ID6 with ID14 in Table 1, we employ a larger Conformer model (L), which performs better than the Conformer (S). Additionally, the dimension of Conformer (L) is consistent with the frontend of AV-Hubert, potentially resulting in fewer information losses during feature input.

AV-Hubert: From Table 1 (ID8, ID11, ID14); (ID9, ID12, ID15); (ID10, ID13, ID16), we can compare the performance of different AV-Hubert frontends. AV-Hubert(b) and AV-Hubert(c) show similar performance, while AV-Hubert(a) exhibits relatively better performance. This could be attributed to AV-Hubert(a) leveraging more labeled information compared to (b) and (c), including finetuning on the pseudo-label generation from VoxCeleb2(En) and oracle labels from LRS3.

Multi-layer Feature Aggregation: From Table 1 (ID8-10); (ID11-13); (ID14-16), we compare the performance of Multi-layer Feature Aggregation. The notation '1L' indicates using $N = 1$, where only the output of the final layer of the frontend is utilized without employing the Multi-layer Feature Aggregation mechanism. '3L' signifies using $N = 3$, incorporating features from the last three hidden layers, while '6L' denotes $N = 6$. It can be observed that the performance of the 3L models outperforms the 1L models on some metrics. Furthermore, deeper utilization of more layers yields additional benefits in downstream WWL tasks, demonstrating the effectiveness of the Multi-layer Feature Aggregation mechanism.

In Table 1, ID17-20 represent the second-stage finetuning of previous models with a lower learning rate of 0.0002. The symbol ♡ represents unfreezing the pretrained frontend for finetuning. We find that unfreezing and finetuning the pretrained frontend to obtain ID17 resulted in a similar performance to ID18, possibly due to overfitting caused by the large model parameters. The symbol ♣ represents finetuning without unfreezing the pretrained frontend, leading to further performance improvement of the model.

Fusion Strategy: Finally, we also test different fusion mechanisms. Multi-system fusion can enhance model robustness. Ultimately, using score fusion, we achieve a final performance of 16.71 WWS on the Eval set, ranking first on the leaderboard.

5. CONCLUSION

The paper presents the Wake Word lipreading system developed by the WHU team for the ChatCLR Challenge 2024. We design a system based on a pretrained frontend and Transformer-like backend structure, incorporating Attentive Pooling and a Classifier. We explored the effectiveness of various frontends from Auto-AVSR and AV-Hubert, as well as investigated the performance of Conformer and E-Branchformer backends. Additionally, we proposed Multi-layer Feature Aggregation by leveraging features from multiple layers of the encoder block and demonstrated its effectiveness. Finally, we employ various fusion strategies, culminating in score fusion, which achieves a false reject rate of 8.21% and a false alarm rate of 8.50% along with a WWS score of 16.71% on the evaluation set and obtain the first place in the task 1 of the ChatCLR Challenge.

6. REFERENCES

[1] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas, "Large-Scale Visual Speech Recognition," in *Proc. Interspeech*, 2019, pp. 4135–4139.

[3] https://drive.google.com/file/d/19GA5SqDjAkI5S88Jt5neJRG-q5RUi5wi/view?usp=sharing

[4] https://dl.fbaipublicfiles.com/avhubert/model/lrs3_vox/vsr/self_large_vox_433h.pt

[5] https://dl.fbaipublicfiles.com/avhubert/model/lrs3_vox/clean-pretrain/large_vox_iter5.pt

[6] https://dl.fbaipublicfiles.com/avhubert/model/lrs3_vox/noise-pretrain/large_vox_iter5.pt

[2] Liqiang Nie, Mengzhao Jia, Xuemeng Song, Ganglu Wu, Harry Cheng, and Jian Gu, "Multimodal activation: Awakening dialog robots without wake words," in *Proc. SIGIR*, 2021, pp. 491–500.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Haoxu Wang, Ming Cheng, Qiang Fu, and Ming Li, "Robust Wake Word Spotting With Frame-Level Cross-Modal Attention Based Audio-Visual Conformer," in *Proc. ICASSP*, 2024, pp. 11556–11560.

[10] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *Proc. ICASSP*, 2023, pp. 1–5.

[11] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proc ICLR*, 2022.

[12] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[13] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*, 2023, pp. 84–91.

[14] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen, "A review of recent advances in visual speech decoding," *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.

[15] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017, pp. 6447–6456.

[16] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[17] 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.

[18] Themos Stafylakis and Georgios Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. Interspeech*, 2017, pp. 3652–3656.

[19] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[20] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2022.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Proc. NIPS 2017*, 2017, vol. 30.

[22] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP*, 2021, pp. 7613–7617.

[23] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman, "Sub-word level lip reading with visual attention," in *Proc. CVPR*, 2022, pp. 5162–5172.

[24] 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.

[25] Tara N Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. Interspeech*, 2015.

[26] 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," in *Proc. Interspeech*, 2019, pp. 3372–3376.

[27] 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.

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

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

[30] Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li, "SlideSpeech: A Large Scale Slide-Enriched Audio-Visual Corpus," in *Proc. ICASSP*, 2024, pp. 11076–11080.

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

[32] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, 2020, pp. 12449–12460.

[33] 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.

[34] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. CVPR*, 2021, pp. 13325–13333.

[35] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[37] 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.

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.

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

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds. 2019, pp. 4171–4186, Association for Computational Linguistics.

[41] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112, 2018.

[42] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[43] Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*. PMLR, 2022, pp. 17627–17643.

[44] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Proc. Interspeech*, 2022, pp. 306–310.

[45] 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.