Improving Audio-Visual Speech Recognition by Lip-Subword Correlation Based Visual Pre-training and Cross-Modal Fusion Encoder

1st Yusheng Dai
University of Science and Technology of China
Hefei, China

2nd Hang Chen
University of Science and Technology of China
Hefei, China

3rd Jun Du*
University of Science and Technology of China
Hefei, China

4th Xiaofei Ding
Alibaba Group
Hangzhou, China

5th Ning Ding
Alibaba Group
Hangzhou, China

6th Feijun Jiang
Alibaba Group
Hangzhou, China

7th Chin-Hui Lee
Georgia Institute of Technology
Atlanta, USA

***Abstract*—In recent research, slight performance improvement is observed from automatic speech recognition systems to audio-visual speech recognition systems in end-to-end frameworks with low-quality videos. Unmatching convergence rates and specialized input representations between audio-visual modalities are considered to cause the problem. In this paper, we propose two novel techniques to improve audio-visual speech recognition (AVSR) under a pre-training and fine-tuning training framework. First, we explore the correlation between lip shapes and syllable-level subword units in Mandarin through a frame-level subword unit classification task with visual streams as input. The fine-grained subword labels guide the network to capture temporal relationships between lip shapes and result in an accurate alignment between video and audio streams. Next, we propose an audio-guided Cross-Modal Fusion Encoder (CMFE) to utilize main training parameters for multiple cross-modal attention layers to make full use of modality complementarity. Experiments on the MISP2021-AVSR data set show the effectiveness of the two proposed techniques. Together, using only a relatively small amount of training data, the final system achieves better performances than state-of-the-art systems with more complex front-ends and back-ends. The code is released at[1].**

***Index Terms*—audio-visual speech recognition, end-to-end system, GMM-HMM**

I. Introduction

Audio-visual speech recognition (AVSR) is a multi-modality application motivated by the bi-modal nature of perception in speech communication between humans [1]. It utilizes lip movement as a complementary modality to improve the performance of automatic speech recognition (ASR). In early research, handcrafted lip features were commonly extracted and added to hybrid ASR systems [2]–[5]. Recently, end-to-end AVSR systems have achieved great success due to the simplicity of end-to-end ASR system designs and an availability of a large number of public audio-visual databases [6]–[14]. Although end-to-end AVSR systems have shown their simplicity and effectiveness on many benchmarks [15]–[20], they are still far from common use. One hard piece of evidence comes from the recent MISP2021 Challenge [21].

As the largest Mandarin audio-visual corpus until now, MISP2021-AVSR is recorded in TV rooms of home environments with multiple groups chatting simultaneously. Multiple microphone arrays and cameras are used to collect far/middle/near-field audios and far/middle-field videos. In the evaluation stage, submitted systems are restricted to utilizing far-field audio and videos for the AVSR task. According to the reports from top-ranked teams [16], [22], [23], low-quality far-field videos could only slightly improve the AVSR performance over ASR systems under a common end-to-end framework. The performance degradation from a uni-modal network to a multi-modal network is also observed in [24]. Compared to the uni-modal model, it is challenging to learn an extensive integrated neural network due to unmatched convergence rates and specialized input representations between two modalities [24], [25]. Pre-training techniques [15], [26], [27] are expected to alleviate the problem which decouples the one-pass end-to-end training framework in two stages. Uni-modal networks are first pre-trained and later integrated into a fusion model following unified fine-tuning. This divide-and-conquer strategy could effectively mitigate variations in learning dynamics between modalities and promote their interactions.

A crucial aspect of the decoupled training framework is how to pre-train the visual frontend. A simple practice comes from [16], [22], which directly uses a pre-trained visual frontend [28] as a frozen visual embedding extractor. This approach gives only a small improvement due to domain shifts across the source and target domains. For most studies, researchers pre-train the visual frontend on an isolated word recognition task [15], [26], [29] and then fine-tune it with the AVSR model. However, these pre-training methods depend on word-level lipreading data sets that are challenging to collect on a large scale. A recent study [27] leverages self-supervised

*corresponding author

[1]https://github.com/mispchallenge/MISP-ICME-AVSR

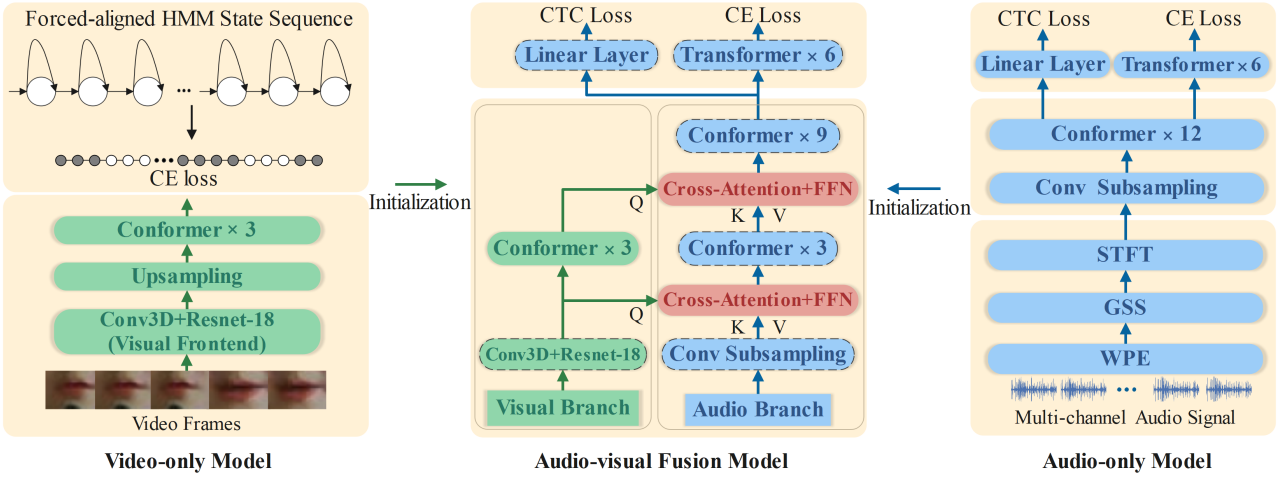

Fig. 1. Overall training framework of our AVSR system

learning on large-scale unlabeled data sets for AVSR. Although these pre-training methods improve the AVSR system performances to a certain extent, a large amount of extra labeled/unlabeled data is used.

In this paper, we propose a subword-correlated visual pre-training technique that does not need extra data or manually-labeled word boundaries. We train a set of hidden Markov models with Gaussian mixture model (GMM-HMMs) on far-field audio to produce frame-level alignment labels and pre-train the visual frontend by identifying each visual frame's corresponding syllable-related HMM states. Compared to the pre-training method based on end-to-end continuous lipreading, our method explicitly offers syllable boundaries to establish a direct frame-level mapping from lip shapes to syllables in Mandarin. These fine-grained alignment labels guide the network to focus on learning visual feature extraction of low-quality videos. On the other hand, this pre-training method could be viewed as a cross-modal conversion process that accepts video frames as inputs and generates acoustic subword sequences. It is helpful to explore potential acoustic information from lip movements and contributes to a good adaptation process with the audio stream in the fusion stage.

In the fusion stage of decoupled training, the initialized audio and visual branches already have the fundamental ability to extract uni-modal representations. Based on the straightforward assumption that the audio modality contains more linguistic information essential for ASR tasks. We propose a novel CMFE block in which the audio modality dominates and more training parameters of the network are used for modality fusion modeling. As for the modality fusion structures, motivated by the decoder architecture of the vanilla transformer [30], the layer-wise cross-attention is designed in different layers to make full use of modality complementarity.

In summary, for this paper, we make the following contributions: (1) we propose a visual frontend pre-training method to correlate lip shapes with the syllabic HMM states. It does not require extra labeled/unlabeled data sets or manually-labeled word boundaries but is able to effectively utilize the visual modality; (2) we propose an audio-dominated cross-modal fusion Encoder (CMFE), in which multiple cross-modal fusions occur at different layers; (3) as a result, our AVSR system achieves a new state-of-the-art performance on the MISP2021-AVSR corpus without using extra training data and complex front-ends and back-ends.

II. Proposed Techniques

A. Overall Training Framework

As illustrated in Fig. 1, our AVSR system is trained in two stages: uni-modal pre-training and multi-modal fine-tuning. In the first stage, we pre-train a hybrid audio-only ASR CTC/Attention model [31] based on standard end-to-end ASR training as shown in the rightmost branch of Fig. 1. For the video modality as shown in the leftmost branch of Fig. 1, we explore a correlation between lip shapes and subword units as described in Section II-B to pre-train the video-only model. Then we initialize and fine-tune the audio-visual fusion model, as shown in the middle branch of Fig. 1, after the two unit-modal networks have converged. In Fig. 1 the audio-visual fusion model integrates the audio branch (blue blocks) and visual branch (green blocks) with cross-attention blocks (red blocks). The four dashed-border blocks in the middle fusion block are initialized by the pre-trained models while the solid-border blocks are initialized randomly. Both audio-only and fusion models integrate the CTC decoder with a transformer-based decoder for joint training and decoding. The loss function can be formulated as a linear combination of the logarithm of the CTC and attention posterior probabilities as shown below:

$$\mathcal{L}_{\mathrm{MTL}} = \lambda \log P_{\mathrm{ctc}}\left(Y \mid X\right) + (1-\lambda) \log P_{\mathrm{att}}\left(Y \mid X\right) \quad (1)$$

where $X = [x_1, \cdots, x_T]$ and $Y = [y_1, \cdots, y_L]$ denote the encoder output and the target sequences, respectively. T and L denote their lengths and λ is the weight factor between the CTC loss and the attention cross entropy (CE) loss.

B. Visual Pre-training by Correlating Lip Shapes with Syllable Units in Mandarin

In previous studies, some researchers applied the cold fusion method that freezes the pre-trained visual-only model and

directly combines visual embedding with audio embedding. Others commonly pre-trained the visual frontend on the isolated word classification task and fine-tuned it with the fusion model. Compared to these techniques, our visual pre-training method correlates lip shapes with frame-level syllabic sequences generated by a GMM-HMM. It offers explicit boundaries to establish a direct frame-level mapping from lip shapes to acoustic subwords and does not need extra data sets or manually-labeled word boundaries.

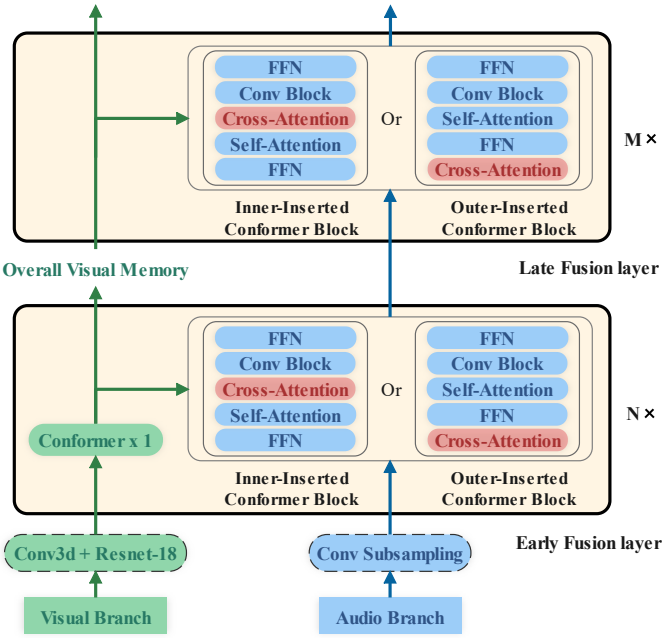

Fig. 2. Cross-modal fusion encoder.

Specifically, we utilize Dacidian Dictionary [32] as basic pronunciations to map Chinese characters to Pinyin-based syllables with tones. Then we follow the Kaldi AIShell recipe to train a triphone GMM-HMM model on far-field audio. Next, an HMM-based Viterbi forced alignment is applied to obtain the frame-level state boundaries of the clustered HMM states. As shown on the leftmost branch of Fig. 1, the pre-trained video-only model consists of a conv3d+resnet18 block, an up-sampling block and a 3-layer conformer block. The conv3d+resnet18 block has the same architecture as [33]. Due to the mismatch in sampling rate between the video frames (25fps) and alignments (100fps), two deconvolution layers are adopted to up-sample video embedding by four times. We avoid sub-sampling the tied-triphone state alignments because it could destroy a complete HMM transition. Moreover, in video recordings, a sampling process for continuous lip movement could be considered as a naturally masked operation that drops extra visual information between two frames. The up-sampling operation is intended to reconstruct dropped video frames, which helps explore potential temporal associations between two consecutive video frames.

We pre-train the video-only model using the frame-level clustered HMM state boundaries obtained in the forced alignment process described earlier. A CE criterion $\mathcal{L}_{\text{CE}}$ between the output prediction posterior $P\left(Y \mid X\right)$ and the ground truth posterior of state P^{GT} is computed as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=0}^{T-1} P_t^{\text{GT}} \log P\left(y_t \mid X\right) \tag{2}$$

where T is the length of the ground truth and y_t is the posterior probability of corresponding HMM state classification on the tth frame.

C. Cross-Modal Fusion Encoder

Attention-based fusion has shown its advantages in recent studies [18], [34]. Unlike direct concatenation, attention-based fusion is not constrained to frame rate discrepancies and audio-visual asynchrony. Most attention-based fusion network follows a symmetric dual-branch structure without considering a modality priority. Within the two-stream framework as shown in Fig. 1, the pre-trained audio/visual branches are already able to extract uni-modal representations at the beginning of the fusion stage, so more learning parameters are considered for modal fusion rather than uni-modal representation learning. Since speech contains more linguistic and semantic information, we reduce the depth of the visual branch, and more parameters are used for multiple modal cross-attention in different layers to make full use of modality complementarity. As a result, we degenerate the classical dual-branch structure into the audio-dominate cross-modal fusion encoder (CMFE).

As shown in Fig. 2, the backbone of the CMFE is composed of N early fusion layers and M late fusion layers with $N + M = 12$ and $N \in [1, 2, 3]$. Compared to conventional symmetric dual-branch structures, only the early fusion layer includes a conformer block in the visual branch in our fusion model. In each fusion layer, one cross-attention block is inner/outer inserted into the conformer of the audio branch. Following the design of the decoder block of vanilla transformer, inner insertion means inserting the cross-attention between the self-attention block and the convolution block of the conformer block. Outer insertion means inserting a cross-attention layer in the front of the conformer block, which does not break the structure of a complete conformer block. For the n^{th} early fusion layer, video embeddings produced by the conformer block X_V^n are considered as a query (Q) to conduct cross-attention operation with audio embedding X_A^n as a key (K) and a value (V) in the same layer. We consider the visual modality as the query, intended to use its robustness against acoustic signal corruption, to match target the audio components in noises. After the n^{th} fusion layer, all video embedding elements from each early fusion layer are concatenated over the channel dimension and projected into a 512-dimension overall visual memory X_V^O for late fusion, as formulated in the following:

$$X_V^O = \text{FC}(\text{Concat}(X_V^1, \cdots, X_V^N)) \tag{3}$$

Motivated by the decoder architecture of the vanilla transformer, the same overall visual memory is directly integrated

TABLE I
COMPARISON OF DIFFERENT PRE-TRAINING METHODS USING TWO STREAMS. ISOLATED DENOTES ISOLATED WORD RECOGNITION AND CONTINUOUS MEANS CONTINUOUS LIPREADING RECOGNITION.

Model	Pre-training Method	Unit(number)	CER(in %)
A0	No Pre-training	\	35.53
AV1	No Pre-training	\	37.81
AV2	Only Pre-train A	\	34.97
AV3	A&V(Isolated)	Word(500)	34.49
AV4	A&V(Isolated)	Word(1000)	29.84
AV5	A&V(Continuous)	Char(3385)	29.22
AV6	A&V(Continuous)	Char(3385)	30.13
AV7	A&V(Proposed)	Senone(3168)	**28.66**
AV8	A&V(Proposed)	Senone(6272)	28.90

into the audio in the late fusion layers. This multiple-fusion design is aimed at reducing forgetfulness and making full use of modality complementarity in different layers.

III. EXPERIMENTS AND RESULT ANALYSIS

A. Experimental Setup

1) Data Sets and Preprocessing: Most experiments are evaluated on the updated version of the MISP2021-AVSR corpus [8], denoted as $MISP_{update}$. For fairness, we experiment with the original version of the MISP2021-AVSR corpus [21] released in the MISP2021 Challenge, denoted as $MISP_{original}$, to compare our proposed system with the state-of-the-art systems. These two data sets share the same train set, while $MISP_{update}$ adds 10 hours of new data to the evaluation set to increase the data diversity. We use far-field audio and far/middle-field videos in the training stage and evaluate systems' performance on far-field audio and videos. Conventional signal processing algorithms, including weighted prediction error (WPE) [35] and guided source separation (GSS) [36], are applied on far-field audio for dereverberation and source separation. Then 80-dimensional filterbank feature vectors are extracted and utterance normalization is applied. We adopt speed perturbation, SpecAug [37] and continuous segments splicing for data augmentation when training ASR model and only use SpecAug for AVSR. For video, we follow [8] to obtain gray-scale lip ROI with 88×88 pixels.

2) Implementation Detail: All conformers in Fig. 1 use the same set of hyper-parameters (n_{head} = 8, d_{model} = 512, d_{ffn} = 2048, CNN_{kernel} = 5). The attention decoder branch of the audio-only and audio-visual fusion models in Fig. 2 consists of the 6-layer transformer (n_{head} = 8, d_{model} = 512, d_{ffn} = 2048). We train audio-only and audio-visual fusion models with a joint CTC loss weight of λ = 0.3. All models are optimized using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the learning rate of 6.0×10^{-4}. The learning rate is warmed up linearly in the first 6000 steps and decreases proportionally to the inverse square root of the step number. A 6-layer transformer-based language model trained on the transcription of the training set is applied when decoding with a weight of 0.2.

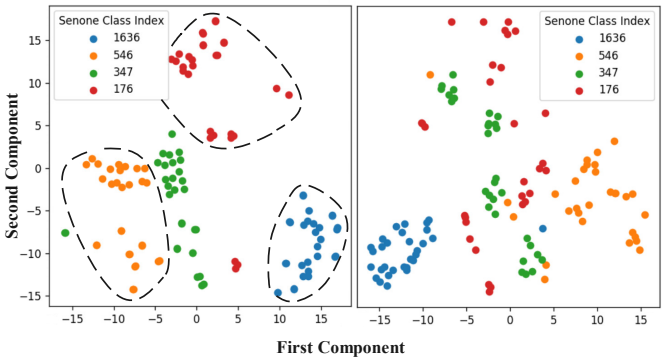

Fig. 3. Resnet output embedding projection through t-SNE from AV7 (left) and AV5 (right)

B. Experiment Results

1) Comparison of Visual Frontend Pre-training: In Table I, we investigate the performance of different pre-training methods following the decoupled training framework with far-field audio and far+middle field videos. Character error rate (CER) is used for all evaluations. We first train the audio-only network A0 and audio-visual fusion network AV1 from scratch, resulting in CERs of 35.53% and 37.81% respectively, which confirms the performance degradation method in subsection I. We initialize the audio branch of AV2-AV8 with A0 and initialize the visual frontend of AV3-AV8 with different pre-training methods. For AV1-AV8, they have the same architecture of visual frontend as shown in the middle of Fig. 1. A tendency is observed that audio branch initialization inhibits performance degradation, and audio-visual branch initialization achieves further performance improvement, which implies that decoupled training framework effectively mitigates variations in learning dynamics between modalities.

Next, we compare different visual branch pre-training methods with our proposed method. We utilize the conv3d+resnet18 modules pre-trained on isolated word recognition tasks with LRW [7] and LRW-1000 [38] offered by [39] as the visual frontend of AV3 and AV4. AV4 performs better than AV3 because LRW-1000 is a Mandarin corpus and the visual frontend of AV4 can be easily adapted to our evaluation task in MISP2021 challenge. For AV5 and AV6, their visual frontends are pre-trained on an end-to-end continuous lipreading recognition task with a CTC loss and an attention loss, respectively. Specifically, their video-only models have a similar encoder architecture to the one shown in Fig. 1 on the leftmost branch but without the up-sampling block. The video-only decoder of AV6 consists of a 6-layer transformer. As shown in Table I, the visual front-ends of AV7 and AV8 are pre-trained on our proposed method. The only difference is the number of clustered HMM states. AV7 and AV8 perform better than other pre-training methods in AV3-AV6, and AV7 model trained with fewer senone units achieves a slightly better CER than AV8 (28.66% vs. 28.90%). The results show an advantage of the fine-grained alignment labels that offer frame-level syllable boundaries to guide visual feature extraction. Moreover, we

TABLE 3
COMPARISON OF TRAINING DATA, FRONTEND AND BACKEND TO SOTA SYSTEMS ON $MISP_{original}$.

System	Training Data		Frontend	Backend Encoder	CER(in %)
	A	V			
NIO	3300 hours	LRW-1000	WPE+GSS	CAE (Seven Stream)	25.07
XIAOMI	3000 hours	LRW-1000	WPE+GSS+SPEx+	AV-Encoder (Dual Stream)	27.17
Proposed	500 hours	w/o extra data	WPE+GSS	CMFE (Dual Stream)	**24.58**

TABLE 2
COMPARISON OF DIFFERENT AUDIO-VISUAL MODAL FUSION STRATEGIES.

Model	Fusion	P_{insert}	N_{vblock}	CER(in %)
AV9	TM-CTC	Outer	3	28.98
AV10	TM-Seq	Outer	3	34.70
AV7	Baseline	Outer	3	28.66
AV11	CMFE	Outer	3	28.00
AV12	CMFE	Outer	2	**27.90**
AV13	CMFE	Outer	1	28.13
AV14	CMFE	Inner	3	28.32
AV15	CMFE	Inner	2	28.09
AV16	CMFE	Inner	1	28.15

use t-distributed Stochastic Neighbor Embedding to visualize the output embedding from the pre-trained visual frontend of AV5 and AV7 in Fig 3 and observe that the embedding projection of AV7 shows a better clustering on syllable units than AV5. It indicates that using the fine-grained frame-level syllable labels enables the visual frontend to explore potential acoustic information from lip movements and contributes to a better adaptation with the audio stream in the fusion stage.

2) Comparison of Audio-visual Fusion Strategies: The results of different fusion strategies on $MISP_{update}$ are shown in Table 2. All models are trained following the decoupled training framework. P_{insert} denotes that the cross-attention block is inner/outer inserted in the conformer block as shown in Fig. 2. N_{vblock} is the number of conformer blocks in the visual branch of the fusion model. TM-CTC [17] and TM-Seq [17] are two classic attention-based fusion structures in which audio and visual streams are integrated in the encoder/decoder respectively. The architecture of the Baseline is shown in the middle branch of Fig. 1. Compared with TM-CTC, TM-Seq and Baseline, the proposed CMFE (AV11-AV16) performs better with multiple fusions in different layers. And the best model (AV12) achieves an CER of 27.90% on $MISP_{update}$ (more difficult than $MISP_{original}$ with an CER of 26.21%). We gradually decrease the number of conformer blocks in the visual branch, and no obvious performance drop is observed. It indicates that more training parameters can be used for modality fusion within the decoupled training framework attributed to visual frontend initialization. Finally, we compare two methods of inserting cross-attention blocks into the original conformer layer. Outer insertion (AV11-AV13) slightly outperforms inner insertion (AV14-AV16) as it does not break the complete conformer block structure.

3) Overall Comparison with State-of-the-art Systems: In Table 3 we present an overall comparison of our proposed system, NIO system [16] and XIAOMI system [22] (the 1st and 2nd place in MISP2021 Challenge). We compare these systems in terms of audio-visual training data, frontend, and backend encoders. For training, NIO and XIAOMI adopted all far/middle/near-field audio and applied a series of simulation and augmentation methods to extend the training set to 3300 and 3000 hours, respectively. Both of them initialized their visual branch on isolated word tasks with extra word-level data sets, LRW-1000. In comparison, our audio-only model is trained on a 500-hour data set and we do not use any extra data to pre-train the visual branch. In terms of frontend and backend, XIAOMI trained an extra neural signal separator (SPEx+ [40]) for source separation. NIO had a large-size backend encoder denoted as CAE to process all original 6-channel signals, the enhanced channel, and the visual features. In contrast, our system is simple yet effective with the frontend system consisting of GSS and WPE and a dual-stream backend. We then use the recognizer output voting error reduction (ROVER) [41] procedure to rescore the output transcripts of A0, AV7, AV12, AV15 models in Tables I-2. As a result, our system attains a state-of-the-art CER of 24.58% and outperforms the NIO system [16] by an absolute CER reduction of 0.5%. On the more difficult $MISP_{update}$ test, our proposed ROVER system also gives a good CER of 25.96%.

IV. CONCLUSION

In this paper, we decouple one-pass end-to-end AVSR training into two stages to mitigate modality variations. Furthermore, we propose a visual pre-training framework by correlating lip shapes with syllables to establish good frame-level syllable boundaries from lip shapes. Moreover, a novel CMFE block is introduced to model multiple cross-modal attentions in the fusion stage and make full use of multi-modal complementarities. Compared to the currently top-performance systems in MISP2021-AVSP Challenge, our proposed system is simple yet effective and achieves a new state-of-the-art performance without using extra training data and complex front-ends and back-ends. In the future, more types of subword units, such as visemes and phonemes, will be explored to improve correlation-based visual pre-training and cross-modal fusion encoder.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No.XDC08050200.

REFERENCES

[1] Julien Besle, Alexandra Fort, Claude Delpuech, et al., "Bimodal speech: early suppressive visual effects in human auditory cortex," *European journal of Neuroscience*, vol. 20, no. 8, pp. 2225–2234, 2004.

[2] J.N. Gowdy, A. Subramanya, C. Bartels, et al., "Dbn based multi-stream models for audio-visual speech recognition," in *Proc. ICASSP 2004*, 2004, pp. I–993.

[3] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, et al., "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 423–435, 2009.

[4] G. Potamianos, C. Neti, G. Gravier, et al., "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, pp. 1306–1326, 2003.

[5] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals, et al., "Lip reading sentences in the wild," in *Proc. CVPR 2017*, 2017, pp. 3444–3453.

[7] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Proc. ACCV 2016*, 2016.

[8] Hang Chen, Jun Du, Yusheng Dai, et al., "Audio-visual speech recognition in MISP2021 challenge: Dataset release and deep analysis," in *Proc. Interspeech 2022*, 2022, vol. 2022, pp. 1766–1770.

[9] Martin Cooke, Jon Barker, Stuart Cunningham, et al., "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, pp. 2421–2424, 2006.

[10] Iryna Anina, Ziheng Zhou, Guoying Zhao, et al., "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. FG 2015*, 2015, pp. 1–5.

[11] Guoying Zhao, Mark Barnard, and Matti Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, pp. 1254–1265, 2009.

[12] Hong Liu, Zhengyan Chen, and Wei Shi, "Robust audio-visual mandarin speech recognition based on adaptive decision fusion and tone features," in *Proc. ICIP 2020*, 2020.

[13] Jun Yu, Rongfeng Su, Lan Wang, et al., "A multi-channel/multi-speaker interactive 3d audio-visual speech corpus in mandarin," in *Proc. ISCSLP 2016*, 2016, pp. 1–5.

[14] Ya Zhao, Rui Xu, and Mingli Song, "A cascade sequence-to-sequence model for chinese mandarin lip reading," in *Proceedings of the ACM Multimedia Asia*, 2019.

[15] Pingchuan Ma, Stavros Petridis, et al., "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP 2022*. IEEE, 2021, pp. 7613–7617.

[16] Gaopeng Xu, Song Yang, Wei Li, et al., "Channel-wise av-fusion attention for multi-channel audio-visual speech recognition," in *Proc. ICASSP 2022*. IEEE, 2022, pp. 9251–9255.

[17] Triantafyllos Afouras, Joon Son Chung, et al., "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[18] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6565–6569.

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

[20] 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 *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 905–912.

[21] Hang Chen, Hengshun Zhou, Jun Du, et al., "The first multimodal information based speech processing (MISP) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*. IEEE, 2022, pp. 9266–9270.

[22] Quandong Wang, Xinyu Cai, Weij Zhuang, et al., "The XIAOMI-TALKFREELY system for audio-visual speech recognition in MISP challenge 2021," https://MISPchallenge.github.io/papers/task2/Xiaomi_tr_task2.pdf.

[23] Wei Wang, Xun Gong, Yifei Wu, Zhikai Zhou, et al., "The Sjtu system for multimodal information based speech processing challenge 2021," in *Proc. ICASSP 2022*. IEEE, 2022, pp. 9261–9265.

[24] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *Proc. CVPR 2020*, 2020, pp. 12695–12705.

[25] Arsha Nagrani, Shan Yang, Anurag Arnab, et al., "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14200–14213, 2021.

[26] Xingxuan Zhang, Feng Cheng, et al., "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proc. CVPR 2019*, October 2019.

[27] Xichen Pan, Peiyu Chen, Yichen Gong, et al., "Leveraging uni-modal self-supervised learning for multimodal audio-visual speech recognition," *arXiv preprint arXiv:2203.07996*, 2022.

[28] Brais Martinez, Pingchuan Ma, Stavros Petridis, et al., "Lipreading using temporal convolutional networks," in *Proc. ICASSP 2022*. IEEE, 2020, pp. 6319–6323.

[29] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, et al., "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *Proc. SLT 2018*. IEEE, 2018, pp. 513–520.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] Shinji Watanabe, Takaaki Hori, Suyoun Kim, et al., "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[32] Jiayu Du, Xingyu Na, Xuechen Liu, et al., "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[33] Hang Chen, Jun Du, Yu Hu, et al., "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, vol. 143, pp. 171–182, 2021.

[34] Yong-Hyeok Lee, Dong-Won Jang, Jae-Bin Kim, Rae-Hong Park, and Hyung-Min Park, "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model," *Applied Sciences*, vol. 10, no. 20, pp. 7263, 2020.

[35] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[36] Christoph Boeddecker, Jens Heitkaemper, Joerg Schmalenstroeer, et al., "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME 2018*, 2018, pp. 35–40.

[37] Daniel S. Park, William Chan, Yu Zhang, et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[38] Shuang Yang, Yuanhang Zhang, Dalu Feng, et al., "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. FG 2019)*. IEEE, 2019, pp. 1–8.

[39] Dalu Feng, Shuang Yang, and Shiguang Shan, "An efficient software for building lip reading models without pains," in *Proc. ICMEW 2021*. IEEE, 2021, pp. 1–2.

[40] Meng Ge, Chenglin Xu, Longbiao Wang, et al., "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.

[41] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. asrU 1997*. IEEE, 1997, pp. 347–354.