Audio-Visual Speech Recognition in MISP2021 Challenge: Dataset Release and Deep Analysis

Hang Chen[1], Jun Du[1], Yusheng Dai[1], Chin-Hui Lee[2], Sabato Marco Siniscalchi[2,4], Shinji Watanabe[3], Odette Scharenborg[6], Jingdong Chen[7], Bao-Cai Yin[5], Jia Pan[5]

[1] University of Science and Technology of China, China [2] Georgia Institute of Technology, USA
[3] Carnegie Mellon University, USA [4] Kore University of Enna, Italy [5] iFlytek, China
[6] Delft University of Technology, The Netherlands[7] Northwestern Polytechnical University, China

jundu@ustc.edu.cn

Abstract

In this paper, we present the updated Audio-Visual Speech Recognition (AVSR) corpus of MISP2021 challenge, a large-scale audio-visual Chinese conversational corpus consisting of 141h audio and video data collected by far/middle/near microphones and far/middle cameras in 34 real-home TV rooms. To our best knowledge, our corpus is the first distant multi-microphone conversational Chinese audio-visual corpus and the first large vocabulary continuous Chinese lip-reading dataset in the adverse home-tv scenario. Moreover, we make a deep analysis of the corpus and conduct a comprehensive ablation study of all audio and video data in the audio-only/video-only/audio-visual systems. Error analysis shows video modality supplement acoustic information degraded by noise to reduce deletion errors and provide discriminative information in overlapping speech to reduce substitution errors. Finally, we also design a set of experiments such as frontend, data augmentation and end-to-end models for providing the direction of potential future work. The corpus[1] and the code[2] are released to promote the research not only in speech area but also for the computer vision area and cross-disciplinary research.

Index Terms: Audio-visual, speech recognition, speech enhancement, data augmentation

1. Introduction

Modern automatic speech recognition (ASR) systems still suffer from performance degradations in real world application scenarios (e.g., home and meeting). Even with advances in technology [1, 2, 3] and the publicity of large-scale corpora recorded in real environments [4, 5, 6], state-of-the-art robust ASR system still runs into performance plateaus, e.g., the CHiME-6 [7] dinner party scenario reaches a word error rate of about 40%, which falls short of the deploy ability of the application. Many researches [8, 9, 10] have shown visual cues can help speech perception, especially in noisy environments. Inspired by these discoveries, Audio-Visual Speech Recognition (AVSR) system has been developed.

Early works on AVSR relied on handcrafted audio-visual feature extraction pipelines and statistical models [11, 12, 13]. Recently, deep neural network (DNN) brought the rapid progress to AVSR. [14] proposed a WLAS model. [15, 16] proposed an attention-based[17] AVSR model. [18] proposed an AV Align framework. [19] adopted Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU) in AVSR. [20] developed a Conformer [21]-based AVSR model. [22, 23] adopted Recurrent Neural Network Transducer in AVSR. The above AVSR systems are all end-to-end AVSR systems, a few works[24] used Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid AVSR systems.

Meanwhile, various audio-visual corpora were released for training. Conventional audio-visual corpora were collected in the controlled environment, such as GRID [25], OuluVS [26], OuluVS2 [27] and TCD-TIMIT [28], etc. These corpora are limited in duration, number of speakers and vocabulary size. More recently, an effort of the research community has been put to gather data from different sources in the wild, e.g., public media websites or TV broadcasts. One of the first audio-visual in-the-wild corpora is LRW [29]. The trend of collecting larger datasets has continued in subsequent collections which consist of materials from British television programs [15, 14], TED talks [30], [31], etc. These large-scale datasets have a vast variety of speakers, sentences, languages and visual/auditory environments, but lack sample-level information. All these corpora are in English, a few corpora [32, 33, 34] have been released to AVSR in Chinese, which is the most widely used language worldwide. Nevertheless, there is still a lack of a large-scale public audio-visual speech corpus recorded in real-world application scenarios, especially for Chinese.

MISP2021 challenge [35] presented a distant multi-microphones conversational audio-visual corpus recorded in the home TV scenario, where several people are chatting in Chinese while watching TV and interacting with a smart speaker/TV in a living room. During the challenge, the corpus was only released to the registered participants and lacked a detailed corpus analysis. In this work, we have resolved authorization and storage issues to fully release the updated AVSR corpus of MISP2021 Challenge to all researchers and make a dataset update including correcting the asynchronous sample in the training/development set and adding more data to increase the data diversity of the evaluation set. Moreover, we provide extensive experiments for various baselines and deep analysis of the corpus. Specifically, we design a set of experiments of the audio-only/video-only/audio-visual speech recognition systems with all far/middle/near audio data and far/middle video data. Error analysis is conducted to explain that video modality supplements acoustic information to reduce deletion errors in noise conditions and provide discriminative information to reduce substitution errors in speech overlap. Finally, we also conduct a set of experiments such as frontend, data augmentation and end-to-end models to provide the direction of future work.

In the following, Section 2 introduces the updated corpus. We describe the overall framework of systems in Section 3. A

[1] https://challenge.xfyun.cn/misp_dataset
[2] https://github.com/mispchallenge/MISP2021-AVSR

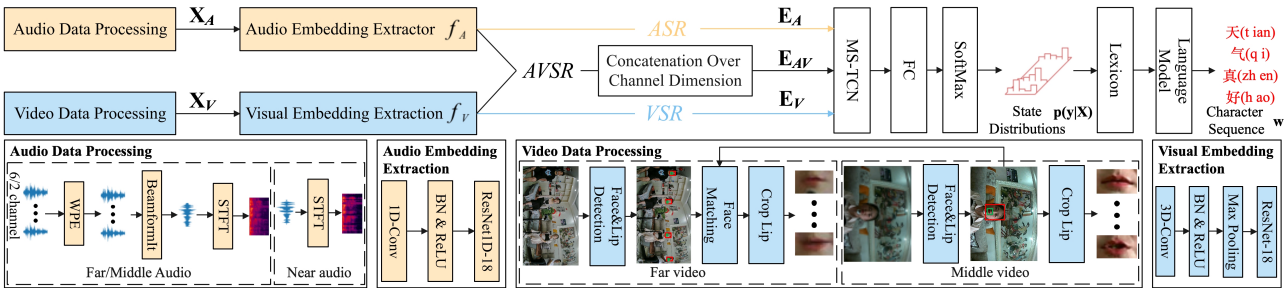

Figure 1: *Illustration of an overall framework for hybird ASR/VSR/AVSR systems with far/middle/near audio and far/middle video.*

Table 1: *Overview of the updated MISP2021-AVSR corpus.*

Dataset	Training	Dev	Eval	Total
Duration (h)	106.09	13.32	21.83	141.24
Overlapping Duration (h)	44.57	6.47	10.70	61.74
Room	21	5	8	34
Participant	200	21	42	253
Male	79	9	10	98
Female	121	12	32	165

Table 2: *Scene configurations.*

ConfigID	C1			C2			C3			C4		
Group	1						2					
TV	Off			On			Off			On		
Time	Day		Night	Day		Night	Day		Night	Day		Night
Light	on	off	on	on	off	on	on	off	on	on	off	on

deep analysis combining experimental results and examples is presented in Section 4. We conclude in Section 5

2. Dataset

As shown in Table 1, the updated MISP2021-AVSR corpus contains 141.24 hours of audio and video data from 253 native Chinese (98 males and 165 females) speaking Mandarin without strong accents which collected in 34 real real-home TV rooms. There is no overlap in speakers and recording rooms among the data in each subset. There are multiple microphone arrays and cameras used to collect far/middle/near audio and far/middle video data, respectively. [35] shows more recording details.

There are some variables that can be controlled during recording, for example, the TV/light can be turned on/off, etc. Moreover, speakers could be divided into several groups to discuss different topics, which results in higher overlap ratios. As shown in Table 2, we divide all combinations of variables into 4 categories based on noise and overlap.

In contrast to the challenge corpus, the updated corpus has the following features:

- During challenge, only far audio/video data in the evaluation set was released. In the updated corpus, all evaluation data has been released to support various research, such as middle video could be used for lipreading task, etc.
- The time synchronizing problem for training and development sets was reported. We have double-checked all training/development data and corrected asynchronous samples.
- 10-hour new data has been added to the evaluation set for increasing the data diversity.

3. Framework

As shown in Eq.1, the recognition process can be formulated as the Bayesian decision problem, where $\mathbf{w} = [w_0, \ldots, w_{n-1}]$, $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}]$ and $\mathcal{H}$ denotes words, input features, and hypotheses, respectively. The Language Model (LM) $\mathrm{p}(\mathbf{w})$ is the probability of an n-characters sequence $\mathbf{w}$ and can be decomposed as Eq.2. With HMM, the Acoustic Model (AM) $\mathrm{p}(\mathbf{X}|\mathbf{w})$ can be decomposed in the frame level as Eq.3 shown: $\mathbf{y} = [y_0, \ldots, y_{T-1}]$ is a clustered hidden state sequence. Each HMM with a set of states represents one tri-phone class. $\pi(y_0)$ is the initial state probability, $a_{y_{t-1}y_t}$ is the state transition probability from $t-1$ to t, $\mathrm{p}(\mathbf{x}_t|y_t)$ is the output probability, $\mathrm{p}(y_t)$ is the prior probability estimated from the training set, $\mathrm{p}(y_t|\mathbf{x}_t)$ is the posterior probability and $\mathrm{p}(\mathbf{x}_t)$ is independent of the word sequence. Gaussian Mixture Model(GMM) can be used to calculate $\mathrm{p}(\mathbf{x}_t|y_t)$ for the GMM-HMM system while DNN can be adopted to compute $\mathrm{p}(y_t|\mathbf{x}_t)$ for the DNN-HMM hybrid system.

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\in\mathcal{H}} \mathrm{p}(\mathbf{w}|\mathbf{X}) = \arg\max_{\mathbf{w}\in\mathcal{H}} \mathrm{p}(\mathbf{X}|\mathbf{w})\mathrm{p}(\mathbf{w}) \tag{1}$$

$$\mathrm{p}(\mathbf{w}) = \prod_{i=0}^{n-1} p(w_i|w_{i-1}\ldots w_0) \tag{2}$$

$$\mathrm{p}(\mathbf{X}|\mathbf{w}) = \sum_{\mathbf{y}} \left[\pi(y_0) \prod_{t=1}^{T-1} a_{y_{t-1}y_t} \prod_{t=0}^{T-1} \mathrm{p}(\mathbf{x}_t|y_t)\right]$$
$$= \sum_{\mathbf{y}} [\pi(y_0) \prod_{t=1}^{T-1} a_{y_{t-1}y_t} \prod_{t=0}^{T-1} \mathrm{p}(y_t|\mathbf{x}_t)\mathrm{p}(\mathbf{x}_t)/\mathrm{p}(y_t)] \tag{3}$$

As shown in Fig.1, all systems are DNN-HMM systems as described above. The main difference lies in $\mathbf{X}$ and $\mathrm{p}(\mathbf{X}|\mathbf{w})$, which are described in Section 3.1 and Section 3.2 respectively. Section 3.3 introduces $\mathrm{p}(\mathbf{w})$, the lexicon and the metric.

3.1. Data Processing

Mel Filter Bank (FBANK) features are adopted as audio feature $\mathbf{X}_A$. A frontend consisting of weighted prediction error (WPE) dereverberation [36] and weighted delay-and-sum beamforming (BeamformIt) [37] is applied to the far-field 6-channel speech and the middle-field 2-channel speech before the FBANK feature extraction. For the near-field monoaural speech, FBANK feature can be extracted directly from the raw audio waveform.

Lip Regions of Interest (RoIs) of the target speaker are cropped as visual feature $\mathbf{X}_V$. Because each frame contains only the target speaker in the middle-field video, an off-the-shelf face and lip detector is used to find face and lip, then a lip-centered window is cropped. All the speakers appear in each far-field video frame, we first detect all the faces and lips, then the face that best matches the mid-field video of the target speaker is selected.

Table 3: *A overall comparison of CER among the DNN-HMM hybrid ASR/VSR/AVSR systems with far/middle/near audio and far/middle video over 4 recording configurations in Table 2.*

System	Audio	Video	CER (in %)			
			C1	C2	C3	C4
A1	far	/	55.47	74.70	82.19	84.21
A2	middle	/	50.51	61.79	75.05	75.83
A3	near	/	25.71	23.45	22.59	20.89
V1	/	far	95.69	95.62	95.98	95.69
V2	/	middle	89.77	88.70	89.30	87.77
AV1	far	far	50.99	66.21	71.55	73.73
AV2	far	middle	45.57	58.59	63.16	64.61
AV3	near	middle	22.64	20.44	19.97	18.58

Table 4: *A comparison of S, D and I among A1 and AV2 systems.*

System	S (in %)		D (in %)		I (in %)	
TV	Off	On	Off	On	Off	On
A1	48.16	48.87	17.71	30.96	1.3	1.43
AV2	35.05	36.88	17.69	23.44	1.15	1.18

3.2. DNN-HMM Hybrid Acoustic Model

The acoustic model consists of an embedding extractor and a followed sequence modeling model. The embedding extractor $f \in \{f_A, f_V, f_{AV}\}$ takes the audio feature $\mathbf{X}_A$ or/and the visual feature $\mathbf{X}_V$ as input to produce an embedding $E \in \{E_A, E_V, E_{AV}\}$. The sequence module is next employed to model the temporal dynamics. Finally, the posterior probability $\mathrm{p}(\mathbf{y}|\mathbf{X})$ is predicted by the ensuing full connection and SoftMax layers. Multi-Scale Temporal Convolution Network (MS-TCN) [38] is adopted as the sequential modeling module in our systems.It can be formulated as:

$$\mathrm{p}(\mathbf{y}|\mathbf{X}) = \mathrm{SoftMax}(\mathrm{FC}(\text{MS-TCN}(\mathbf{E}))) \quad (4)$$

$$\mathbf{E}_V = f_V(\mathbf{X}_V) \quad (5)$$

$$\mathbf{E}_A = f_A(\mathbf{X}_A) \quad (6)$$

$$\mathbf{E}_{AV} = \mathrm{Concat}(f^A_{AV}(\mathbf{X}_A), f^V_{AV}(\mathbf{X}_V)) \quad (7)$$

where the visual embedding extractor f_V is the same as [39] which consists of a spatiotemporal convolution followed by an 18-layer ResNet. A spatiotemporal convolution consists of a convolution layer with 64 3D-kernels, a batch normalization, a ReLU activation and a spatiotemporal max-pooling layer. The audio embedding extractor f_A has the similar structure as f_V. The 3D-kernels in spatiotemporal convolution and the 2D-kernels in ResNet-18 are replaced by 1D-kernels meanwhile the 3D-MaxPooling layer is dropped. The audio-visual embedding extractor f_{AV} consists of a visual module f^V_{AV} and an audio module f^A_{AV} which have the same structure as f_V and f_A, respectively. The concatenation is over the channel dimension and the frame mismatch between the audio and video is solved by repeating a video frame for several audio frames.

The Cross Entropy criterion $\mathcal{L}_{\mathrm{CE}}$ between $\mathrm{p}(\mathbf{y}|\mathbf{X})$ and the true distribution of state p^{GT} is calculated as:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=0}^{N_B-1} \mathrm{p}_t^{\mathrm{GT}} \log \mathrm{p}(y_t|\mathbf{X}) \quad (8)$$

where N_B is the minibatch size. $\mathcal{L}_{\mathrm{CE}}$ is minimized by using Adam optimizer [40] for 100 epochs with an initial learning rate of 0.0003 and cosine scheduler [41].The best model is selected by the highest classification accuracy on the development set.

p^{GT} is generated with the GMM-HMM system. In ASR experiments, three GMM-HMM systems with far/middle/near audio are built. Far-field and near-field GMM-HMM systems are adopted for far-field and middle-field VSR, respectively. In AVSR experiments, the GMM-HMM system with the corresponding audio is adopted.

3.3. Decoding and Scoring

We use the DaCiDian[3] dictionary as the pronunciation dictionary. A 3-gram language model is trained by the maximum entropy modeling method implemented in the SRILM toolkit.

We adopt Character Error Rate (CER) as the metric. It is represented with Eq. 9:

$$CER = S + D + I \quad (9)$$

where S, D, I are substitution, deletion, insertion error rates, respectively. The lower the CER value (with 0 being a perfect score), the better the recognition performance. For speech overlap segments, we calculate all the substitution, deletion and insertion errors based on the recognition results and the ground truth for each speaker based on the oracle speaker diarization.

4. Experiments

Three ASR systems with far/middle/near audio and two VSR systems with far/middle video have been built as baselines for various research. For AVSR system, we considered three combinations: the most challenging combination of far audio and far video, the most ideal combination of near audio and middle video and the compromise of far audio and middle video. Far video to middle video could be achieved with the help of a zoom lens in real-world application. Table 3 shows a comparison of CER among all systems over 4 recording configurations in Table 2. Three ASR results show the performance degradation of ASR results from far-field channel distortion, ambient noises and speech overlap. Two VSR results show the performance of VSR is robust to noise and speech overlap but still very limited. The challenges mainly come from two aspects. Unfavorable lighting and variable head posture make it difficult for the far camera to capture lip changes, which is improved in the middle video. Chinese is a tone language and the tone is the use of pitch in language to distinguish lexical or grammatical meaning, which leads more words to look the same on the lip when pronounce.

4.1. Analysis of errors in noise condition

Based on the presence of noise, we divide the whole evaluation set into two categories and list the corresponding S/D/I in Table 4. Table 4 reflects two trends. With the presence of TV noise, the D of the A1 system rose sharply from 17.71 to 30.96, while neither S nor I rose much. AV2 system shows improvements over A1 across all evaluation metrics, specifically, there is a stable gain in S whether noise's presence or absence and larger improvement of D in the noise condition. Due to the segmentation according to the oracle speaker diarization information during preprocessing, I is very rare. Accordingly, we propose that video modality supplement acoustic information degraded by noise to reduce deletion errors.

In Fig.2, we show a noisy example selected from the evaluation set randomly and the comparison between the outputs of A1,V2,AV2 systems. In the second half of the far spectrum, the noise component almost completely drowns out the target speech component, which causes four deletion errors in the results of the A1 system. As mentioned above, the challenges

[3]https://github.com/aishell-foundation/DaCiDian

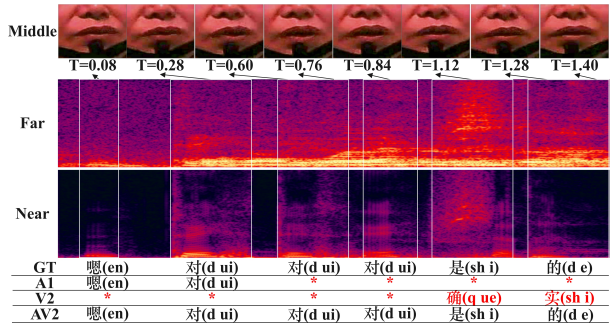

Figure 2: *A noisy sample and comparison between the outputs of A1,V2,AV2 systems. GT means Ground Truth.*

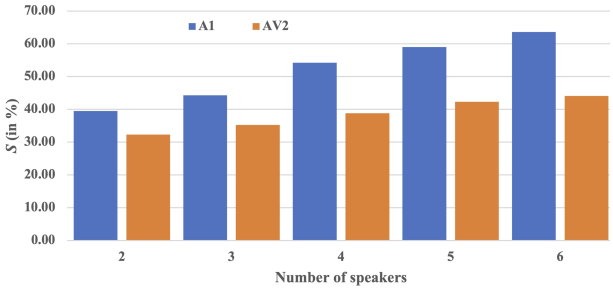

Figure 3: *A comparison of the substitution error among A1 and AV2 systems on different categories with the same number of speakers participating in the conversation.*

from the environment and the tone language make the performance of V2 very limited. But the AV2 system utilizes the acoustic information in the change of lip shape to predict the last four characters correctly.

4.2. Analysis of errors in overlap condition

We think substitution errors are more related to speech overlap. So based on the number of speakers participating in the conversation, we divide the whole evaluation set into six categories and plot the corresponding S in Fig.3. More speakers participating in the conversation means more overlapping speech. A sharp rise of S follows to the number of speakers participating in the conversation is shown in Fig.3. AV2 system shows improvements over A1 across all six categories and larger improvement is observed in the category with more speakers.

In Fig.4, we show an overlapping example selected from the evaluation set randomly and the comparison between the outputs of A1,V2,AV2 systems. The target speech overlaps with the interfering speech almost completely, consequently, the output of A1 contained four substitution errors and a deletion error. V2 excels on the last character due to unambiguous pronunciation actions. AV2 corrected these five errors by utilizing the discriminative acoustic information in the change of lip shape.

4.3. Analysis on advanced techniques

Lastly, we explored some advanced techniques for the most challenging combinations of far-field audio and far-field video. Guided source separation (GSS) [42] is adopted as frontend and SpecAug [43] is applied during the acoustic model training, denoted as AV4 in 5. Further, we build three end-to-end AVSR systems which consist of a hybrid CTC/Attention based

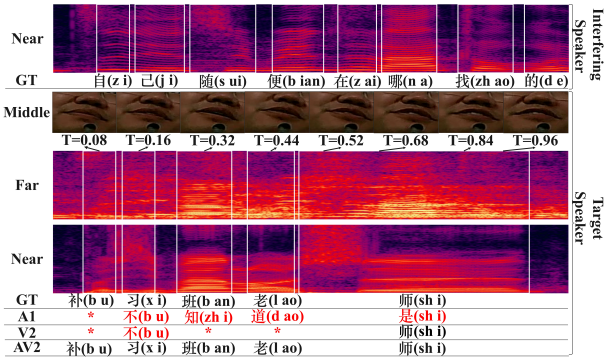

Figure 4: *An overlapping sample and comparison between the outputs of A1,V2,AV2 systems. GT means Ground Truth.*

Table 5: *A comparison of CER between DNN-HMM hybrid AVSR systems and end-to-end AVSR systems with different frontends, data augmentation schemes and acoustic models. Embedding extractors in AMs are omitted. DA: data augmentation.*

System	Frontend	DA	AM	CER (in %)
AV1	BeamfomIt	/	MS-TCN	64.37
AV4	GSS	SpecAug	MS-TCN	48.78
E2E1	GSS	SpecAug	MS-TCN	76.46
E2E2	GSS	SpecAug	Conformer	60.15
E2E3	GSS	WavAug	Conformer*	51.04

acoustic model [20] and a transformer-based language model, denoted as E2E1, E2E2 and E2E3 respectively. E2E1 and E2E2 are developed on a three-layer MS-TCN and a six-layer conformer based encoder backbone respectively using FBANK as audio input. E2E3 has a more flexible conformer-based encoder backbone with skip-connection and multiple fusion. The audio input is raw waveform and WavAug [44] is adopted during training. More details can be found in our source code.

As shown in Table 5, GSS and SpecAug yield significant performance improvements in the DNN-HMM hybrid AVSR system. The end-to-end AVSR system can achieve a comparable performance by simply changing. Besides, it is worth mentioning that many native end-to-end AVSR systems have been proposed on MISP2021-AVSR corpora in the MISP2021 challenge [35]. It is worth exploring the powerful frontend, various data augmentation schemes, and novel sequence model.

5. Conclusions

The updated AVSR corpus of MISP2021 Challenge is fully released. The corpus is the first distant multi-microphone conversational Chinese audio-visual corpus and the first large vocabulary continuous lip-reading Chinese corpus in the adverse home-tv scenario which promotes the research in the speech area, the computer vision area and cross-disciplinary area. Extensive experiments and error analysis show video supplement acoustic information to reduce deletion errors in noise condition and provide discriminative information in overlapping speech to reduce substitution errors. Finally, a set of experiments with advanced techniques provide the direction of the future work.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

7. References

[1] H. Erdogan, J. R. Hershey, S. Watanabe *et al.*, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech 2016*, 2016, pp. 1981–1985.

[2] L. Chai, J. Du, Q.-F. Liu *et al.*, "A cross-entropy-guided measure (cegm) for assessing speech recognition performance and optimizing dnn-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 106–117, 2021.

[3] S. Watanabe, M. Delcroix, F. Metze *et al.*, *New Era for Robust Speech Recognition - Exploiting Deep Learning*. Springer, 2017.

[4] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Proc. ASRU 2007*, 2007, pp. 238–247.

[5] A. Janin, D. Baron, J. Edwards *et al.*, "The icsi meeting corpus," in *Proc. ICASSP 2003*, 2003.

[6] W. Rao, Y.-H. Fu, Y.-X. Hu *et al.*, "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *Proc. ASRU 2021*, 2021.

[7] S. Watanabe, M. Mandel, J. Barker *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. CHiME 2020*, 2020, pp. 1–7.

[8] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, 1976.

[9] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," *Current Directions in Psychological Science*, pp. 405–409, 2008.

[10] D. W. Massaro and J. A. Simpson, *Speech perception by ear and eye: A paradigm for psychological inquiry*. Psychology Press, 2014.

[11] J. 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.

[12] G. Papandreou, A. Katsamanis, V. 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.

[13] 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.

[14] J. S. Chung, A. Senior, O. Vinyals *et al.*, "Lip reading sentences in the wild," in *Proc. CVPR 2017*, 2017, pp. 3444–3453.

[15] T. Afouras, J. S. Chung, A. Senior *et al.*, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *Proc. SLT 2018*, 2018, pp. 513–520.

[17] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[18] G. Sterpu, C. Saam, and N. Harte, "How to teach dnns to pay attention to the visual modality in speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1052–1064, 2020.

[19] B. Xu, C. Lu, Y. Guo *et al.*, "Discriminative multi-modality speech recognition," in *Proc. CVPR 2020*, 2020, pp. 14 433–14 442.

[20] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP 2021*, 2021, pp. 7613–7617.

[21] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[22] T. Makino, H. Liao, Y. Assael *et al.*, "Recurrent neural network transducer for audio-visual speech recognition," in *Proc. ASRU 2019*, 2019, pp. 905–912.

[23] O. Braga, T. Makino, O. Siohan *et al.*, "End-to-end multi-person audio/visual automatic speech recognition," in *Proc. ICASSP 2020*, 2020, pp. 6994–6998.

[24] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1290–1302, 2018.

[25] M. Cooke, J. Barker, S. 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.

[26] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, pp. 1254–1265, 2009.

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

[28] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, pp. 603–615, 2015.

[29] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV 2016*, 2016.

[30] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[31] A. Ephrat, I. Mosseri, O. Lang *et al.*, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation." *ACM Trans. Graph.*, 2018.

[32] H. Liu, Z. Chen, and W. Shi, "Robust audio-visual mandarin speech recognition based on adaptive decision fusion and tone features," in *Proc. ICIP 2020*, 2020.

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

[34] Y. Zhao, R. Xu, and M. Song, "A cascade sequence-to-sequence model for chinese mandarin lip reading," in *Proceedings of the ACM Multimedia Asia*, 2019.

[35] H. Chen, H. Zhou, J. Du *et al.*, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.

[36] L. Drude, J. Heymann, C. Boeddeker *et al.*, "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*, 2018, pp. 1–5.

[37] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2011–2022, 2007.

[38] B. Martinez, P.-C. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. ICASSP 2020*, 2020, pp. 6319–6323.

[39] H. Chen, J. Du, Y. Hu *et al.*, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Network*, p. 171–182, 2021.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR 2015*, 2015.

[41] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. ICLR 2017*, 2017.

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

[43] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[44] E. Kharitonov, M. Rivière, G. Synnaeve *et al.*, "Data augmenting contrastive learning of speech representations in the time domain," in *Proc. SLT 2021*, 2021, pp. 215–222.