# THE SJTU SYSTEM FOR MULTIMODAL INFORMATION BASED SPEECH PROCESSING CHALLENGE 2021

*Wei Wang, Xun Gong, Yifei Wu, Zhikai Zhou, Chenda Li, Wangyou Zhang, Bing Han, Yanmin Qian*[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

## ABSTRACT

This paper describes the SJTU system for ICASSP **M**ulti-modal **I**nformation based **S**peech **P**rocessing Challenge (**MISP**) 2021. To solve the speech recognition problem in real complex environments where time-synchronized near- and far-field signals are available for training an enhancement frontend. We build a joint system with speech enhancement frontend and speech recognition backend. These two modules are optimized jointly by both ASR and enhancement criteria. Audio-visual fusion is explored to further boost the ASR performance. ROVER and test time augmentation techniques are used to combine recognition results from multiple systems. The final system achieves Chinese character error rates (CCER) of 34.9% on dev set and 34.0% on test set, which achieved third place in the MISP challenge. The absolute CCER reduction compared with the official baseline system is 26.9% on dev set and 28.7% on test set.

***Index Terms***— multi-modality, speech recognition, end-to-end

## 1. INTRODUCTION

Speech enhancement (SE) and automatic speech recognition (ASR) play important roles in modern speech-based human-computer interaction applications. Plenty of impressive and inspiring techniques have been developed, showing advantages in various works [1, 2, 3, 4, 5, 6]. However, complex real environment brings great challenges to the SE and ASR systems, such as channel distortion, ambient noise and reverberation, etc. In recent years, it is commonly observed that supplementary information, such as videos [7, 8, 9, 10, 11] and speaker identities [12, 13], can be utilized to assist in accomplishing SE and ASR tasks. Nevertheless, most of the works on far-field speech recognition with an enhancement frontend are conducted on simulated data. The lack of large-scale real-scenario corpora brings about the performance gap of SE and ASR systems on simulated and real data.

Thanks to the Multimodal Information Based Speech Processing (MISP) Challenge committee's work[14], a new large-scale multi-modal Chinese corpus of multi-speaker conversations in real scenario for wake word spotting and speech recognition is released. It targets the home TV scenario where several people are chatting and watching TV. This paper mainly focus on the audio-visual speech recognition task of MISP Challenge (task 2).

To accomplish the task, several features of the corpus should be taken into consideration. First, the collected audio involves not only the target speaker and the interfering speaker's speech, but also noises and human voice from a nearby television. To deal with it, a straightforward way is to perform blind source separation (BSS) on the input audio. However, the permutation problem needs to be solved after BSS. Another promising approach is to apply a beamformer as the frontend, since the speakers and the TV, are located in different directions relative to the microphone arrays. Second, although near-field audios are provided in training, there is severe mismatch between the far-field multi-speaker audios and the corresponding near-field single-speaker audios due to the differences in microphone specification and locations. Thus, it is difficult to utilize near-field audio as supervision in SE model training. Some prior works [15, 16] proposed unsupervised or semi-supervised methods to deal with the problem. Another simple yet promising approach for this task is cascading a SE frontend and an ASR backend, and using an end-to-end ASR loss to jointly train the SE and ASR modules [17, 18]. Third, although high-quality middle-field videos per speaker are available for training, only the far-field videos with speakers' faces in relatively low resolution could be utilized during inference. Super resolution techniques may be helpful according to some works on face recognition [19, 20]. However, most of the lip regions in far-field videos are too small, which poses a challenge for the above pre-trained models.

This paper describes the SJTU system for the MISP challenge. To deal with the complex multi-channel scenario, several commonly used front-end speech processing approaches are explored, including blind source separation [21], guided source separation [22], and neural beamformer [23, 24]. Advanced end-to-end architectures like Transformer and Conformer are used to build the basic ASR system with joint connectionist temporal classification (CTC)-attention multi-task training. We further adopt multi-channel input single-speaker output (MISO) [17] speech recognition, which extends the original end-to-end architecture to deal with multi-channel input. Different data augmentation technologies are investigated, including speed perturbation, SpecAugment, test time augmentation and audio-visual feature adaptation. With the Recognizer Output Voting Error Reduction (ROVER) [25] rescoring technique, our system achieved a significant improvement compared to the official baseline system.

The rest of the paper is organized as below: In Section 2, the system is explained in detail. Experimental results are presented and analyzed in Section 3. Finally, the conclusion is given in Section 4.

## 2. METHODOLOGY

We mainly built two types of systems: Conformer-based hybrid system and Multi-channel Input Single-speaker Output (MISO) system with neural beamformer enhancement frontend and Transformer / Conformer ASR backend. We further conduct audio-visual fusion on these models for improved robustness. Both systems are trained with different setups to produce diverse models. Finally, the trained

[†]Yanmin Qian is the corresponding author.

models are combined with test time augmentation and ROVER to obtain better recognition results.

### 2.1. Conformer based Hybrid System

The speech recognition task is to find the text sequence corresponding to a given speech feature sequence. In hybrid system, it can be formulated as:

$$\begin{aligned}\mathbf{w}^* &= \arg\max_{\mathbf{w}\in\mathcal{H}} P(\mathbf{w}|\mathbf{O}) \\ &= \arg\max_{\mathbf{w}\in\mathcal{H}} \frac{P(\mathbf{O}|\mathbf{w})P(\mathbf{w})}{P(\mathbf{O})} \\ &= \arg\max_{\mathbf{w}\in\mathcal{H}} P(\mathbf{O}|\mathbf{w})P(\mathbf{w})\,,\end{aligned} \tag{1}$$

where $\mathbf{w}, \mathbf{O}, \mathcal{H}$ means words, features and hypotheses, respectively. The $P(\mathbf{O}|\mathbf{w})$ is the acoustic model and $P(\mathbf{w})$ is the language model. After that, the pronunciation lexicon $\mathbf{L}$ is introduced to model the relationship between phones and words, which bridges the acoustic model and language model.

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\in\mathcal{H}} P(\mathbf{O}|\mathbf{L})P(\mathbf{L}|\mathbf{w})P(\mathbf{w})\,. \tag{2}$$

For the deep neural network based hidden Markov model (DNN-HMM), the DNN models the probability $P(\theta|\mathbf{O})$ of pdf-ids (the clustered HMM states) $\theta$ conditioned on given features $\mathbf{O}$ , also known as senone. Different from GMM-HMMs, the DNNs are discriminative models, the transformation strategies are as follows:

$$P(\mathbf{O}|\theta) = \frac{P(\theta|\mathbf{O})P(\mathbf{O})}{P(\theta)} \approx \frac{P(\theta|\mathbf{O})}{P(\theta)}\,. \tag{3}$$

We build the hybrid ASR system using the Conformer acoustic model [26] and Kaldi [27] toolkit. We first follow the official Kaldi NN-HMM baseline to build the HCLG Weighted Finite State Transducer (WFST) and get the alignment of the whole training set. Then we train the Conformer acoustic model with the Cross Entropy (CE) criterion to predict the pdf-ids. After that, the Conformer acoustic model and the HCLG graph are cascaded for decoding. To get a relatively high quality alignment, we train and align the near-field data at the first time. Since the near-field data and far-field data are synchronized in time, we train the NN-HMM model on near-field data and apply the alignment results as the alignment for far-field data for better alignment precision .

### 2.2. Enhancement ASR joint System

The architecture of the enhancement and ASR joint system is shown in Fig. 1. In the frontend module, we adopt the mask-based minimum variance distortionless response (MVDR) [28, 29] beamformer for noise reduction and interference suppression, which is formulated as follows:

$$\mathbf{w}_f = \frac{\mathbf{\Phi}_{\mathrm{n},f}^{-1}\mathbf{\Phi}_{\mathrm{s},f}}{\mathrm{Trace}(\mathbf{\Phi}_{\mathrm{n},f}^{-1}\mathbf{\Phi}_{\mathrm{s},f})}\mathbf{u}\,, \tag{4}$$

where $\mathbf{\Phi}_{\mathrm{s},f}$ and $\mathbf{\Phi}_{\mathrm{n},f}$ represent the spatial covariance matrices of the target speech and noise signals, respectively. $\mathbf{u}$ is a one-hot vector for reference channel selection. $\mathrm{Trace}(\cdot)$ denotes the trace of a matrix. Both speech and noise covariance matrices are estimated based on the predicted masks:

$$\mathbf{\Phi}_{\alpha,f} = \frac{\sum_{t=1}^{T}\left(\sum_{c=1}^{C} M_{\alpha,t,f,c}\right)\mathbf{Y}_{t,f}\mathbf{Y}_{t,f}^{\mathsf{H}}}{\sum_{t=1}^{T}\sum_{c=1}^{C} M_{\alpha,t,f,c}}\,, \tag{5}$$

where $\alpha \in \{\mathrm{s}, \mathrm{n}\}$ is a symbol for representing speech $(s)$ or noise $(n)$. $M_{\alpha,t,f,c}$ represents the predicted mask value at the $t$-th time frame, $f$-th frequency bin, and $c$-th channel. $T$ and $C$ denote the total number of time frames and channels, respectively. $\mathbf{Y}_{t,f}$ is the observed far-field signal.

The enhanced signal $\hat{\mathbf{X}}_f$ is finally obtained by applying the estimated time-invariant beamformer filter $\mathbf{w}_f$ to the input signal $\mathbf{Y}_f$:

$$\hat{\mathbf{X}}_f = \mathbf{w}_f^{\mathsf{H}}\mathbf{Y}_f\,. \tag{6}$$

The enhanced spectrum is then fed into a differentiable feature extraction module to calculate the 80-dim log Mel-filterbank feature $\mathbf{O}$ for speech recognition:

$$\mathbf{O} = \mathrm{LMF}(\hat{\mathbf{X}}_f)\,, \tag{7}$$

where $\mathrm{LMF}(\cdot)$ represents the feature extraction module.

In conventional speech enhancement tasks, we train models with simulated noisy signals and their corresponding clean reference signals. Although parallel near-field signals are also provided in this task, they cannot be directly used for training due to the sample shift and channel mismatch caused by various factors such as reverberation and device difference.

To resolve the power mismatch between near-field signals and far-field signals. We introduce trainable power adaptive filters $H$ to resolve the scale mismatch between far-field signals and near-field signals in each frequency band. That is, we rescale the enhanced signal obtained from Eq. (6) per frequency band with $H$ as:

$$\hat{\mathbf{X}}_f := H_f\hat{\mathbf{X}}_f\,. \tag{8}$$

To mitigate the sample shift and channel mismatch, $L_2$ loss is calculated between Mel-filterbank features of far-field signals and their corresponding near-field signals for enhancement frontend training:

$$\mathcal{L}_{\text{enh}} = \left\|\mathbf{O} - \mathbf{O}_{\text{near}}\right\|_2\,. \tag{9}$$

where $\mathbf{O}_{\text{near}}$ is Mel-filterbank feature of the near-field signals. To stabilize the training process, $\mathbf{O}_{\text{near}}$ is also fed into the ASR module to train the ASR backend.

We follow the fully end-to-end training scheme proposed in [17, 18, 30] with enhancement loss to jointly optimize the neural beamformer and end-to-end ASR systems. That is, the final ASR criterion is used to optimize both frontend and backend modules:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{enh}} + \lambda_2\mathcal{L}_{\text{ctc}} + (1-\lambda_1-\lambda_2)\mathcal{L}_{\text{att-dec}}\,, \tag{10}$$

where $\lambda_1$ and $\lambda_2$ are interpolation factors for multi-task learning. $\mathcal{L}_{\text{enh}}$ denotes the L2 enhancement frontend loss. $\mathcal{L}_{\text{ctc}}$ and $\mathcal{L}_{\text{att-dec}}$ denote the CTC loss and attention-based decoder CE loss. In addition, following [31], several techniques, such as diagonal loading, mask flooring, and double precision, are used to improve the numerical stability of the end-to-end system during training.

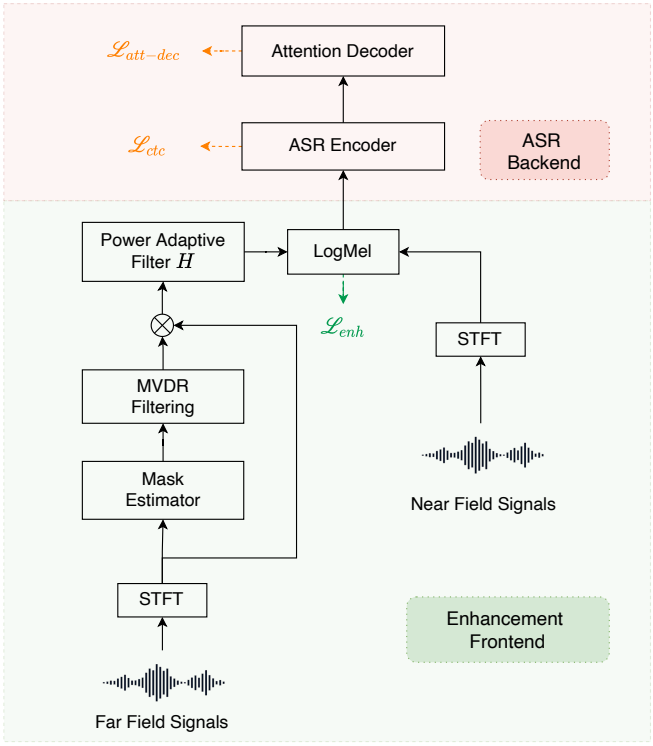



**Fig. 1**: End-to-End Training of Enhancement and ASR system

### 2.3. Visual Feature Preparation

We build a lipreading model with the same structure as that in [32]. The lipreading model is firstly pre-trained with CAS-VSR-W1k (the original LRW1000) [33] dataset. The pre-training is conducted on a 1000-word classification task, and then it is fine-tuned with the videos in MISP data.

In the MISP data fine-tuning stage, we firstly use the near-field audio to train a GMM-HMM model following Kaldi AIShell recipe, then the syllable-level alignment of step `tri4` is extracted for near-field audio. The videos are segmented with the corresponding audio alignment, to ensure each segment contains only 1 syllable. Then the lip-reading model is fine-tuned with the segmented video on the syllable classification task. The last fully-connected (FC) layer of the pre-trained model is replaced with a random initialized FC layer for syllable classification. This fine-tune procedure is firstly performed for the MISP middle-field, and then performed for far-field videos.

Finally, the 3D-ResNet frontend of the fine-tuned lipreading model is used as the visual feature extractor. By processing the videos with the visual feature extractor, each 2-D picture frame in the MISP video will be converted into a 1-D vector.

### 2.4. Test time Augmentation

Test time augmentation is a useful technique validated in image classification [34] and accent identification [35]. Instead of predicting the label of the test audio itself, our model takes multiple augmented versions of the test audio as input, and the predicted results are then aggregated to obtain the final result. Concretely, speed perturbation [36] with ratios 0.9, 1.0 and 1.1 is applied to the test audios. The decoding results on augmented test sets are rescored by the ROVER technique.

**Table 1**: The performance on dev-far set of Conformer based hybrid models

| System | CCER(%) |
|---|---|
| TDNN-nnet3-combine | 76.1 |
| Conformer-nnet3-far | 69.7 |
| Neural BF + Conformer-nnet3-far | 60.0 |

## 3. RESULTS

All experiments are carried out with ESPnet [37] and Kaldi toolkits. SpecAug [38] and speed perturbation with ratio 0.9, 1.0 and 1.1 are applied for all training data. Details on MISP dataset specifications can be found at `https://mispchallenge.github.io/task2_data.html`.

### 3.1. Conformer based Hybrid System

We adopted 12 layers of Conformer multi-headed self-attention blocks with 256 hidden dims and feedforward layers of 2048 hidden units as our acoustic model. We also tested Kaldi's time delay neural network (TDNN) models. The 1024-dim visual features are projected to 40-dim features through principal component analysis (PCA) and concatenated with the 80-dim FBANK features before being fed into the model. For the HCLG in our system, we used the DaCiDian in Kaldi's AIShell-2 [39] recipe as our pronunciation lexicon. The 4-gram language model was trained using the kenLM toolkit. All data were adopted without any enhancement technique.

Table 1 shows performance of Conformer based hybrid models on dev far set. 'combine' means the model is trained with data from a combination of near-field middle-field and far-field data. The neural beamformer in the third row is jointly optimized by CE loss with Conformer ASR backend. Conformer based hybrid models showed limited improvement even if a neural beamformer enhancement frontend is applied.

### 3.2. Enhancement ASR joint System

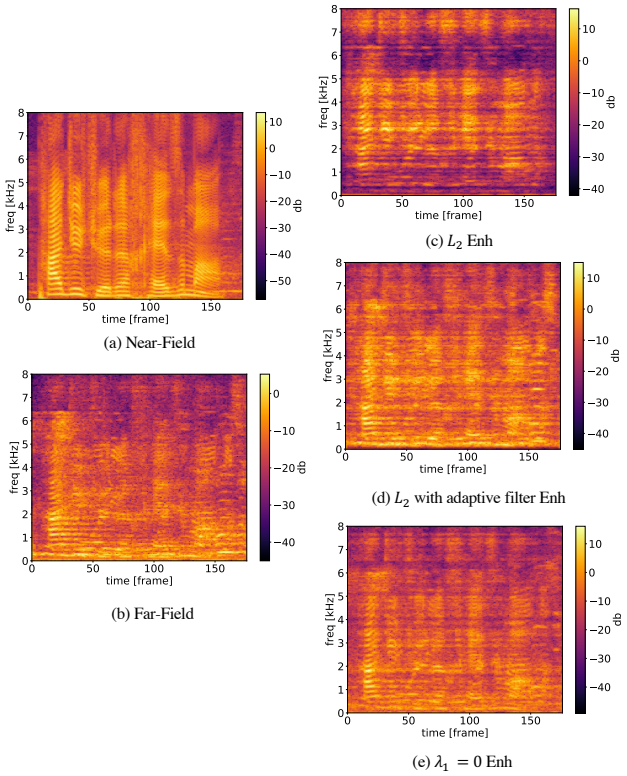
For neural beamformer-based enhancement frontend, the mask estimation network is a 3-layer bidirectional long-short term memory with projection (BLSTMP) network with 512 cells in each direction. For Transformer-based ASR backend, we adopted 12 layers of encoder and 6 layers of decoder with 2048 hidden units. Each layer is a Transformer block with 8 heads of 64 dimension self-attention layer. For Conformer-based ASR backend, we replaced the Transformer blocks with Conformer blocks and adopted the same configurations for the number of layers, hidden units and attention heads. The $\lambda_2$ for multi-task learning (MTL) in Eq. (10) is set to 0.3 during training. For all systems, the 1024-dim visual features are projected to 256-dim through a feedforward layer and concatenated with the output of the first convolutional layer after the neural beamformer.

Fig 2 shows the spectrum of the far-field signal enhanced with different frontend training objectives. When the power adaptive filter $H$ in Fig 1 is not applied, we observe vanishing of multiple frequency bands on the spectrum of enhanced signal as shown in Fig 2 (c). A possible reason is that the attenuation of speech signal during propagation is frequency-dependent, and the relative power scale among different frequency bands differs in far-field signals and their corresponding near-fields signals. Therefore, a trainable power adaptive filter $H$ is appended after the neural beamformer to simulate the signal attenuation during propagation and adjust the relative power scale of the enhanced far-field signals on different frequency band as shown in Fig 2 (d). Although no explicit constraint

is enforced on the enhancement frontend during training as shown in Fig 2 (e), the intermediate signals after the neural beamformer still shows good quality in terms of audibility.

**Table 2**: The performance on dev far set of Enhancement ASR joint models

| System | $\lambda_1$ | CCER(%) |
|---|---|---|
| Conformer | N/A | 49.9 |
| Neural BF + Transformer | 0 | 41.0 |
| Neural BF + Transformer | 0.1 | 41.4 |
| Neural BF + Conformer | 0 | **39.8** |
| Neural BF + Conformer | 0.1 | 40.7 |




(a) Near-Field

(b) Far-Field

(c) $L_2$ Enh

(d) $L_2$ with adaptive filter Enh

(e) $\lambda_1 = 0$ Enh

**Fig. 2**: Example of enhanced spectra after neural beamformer

Table 2 shows the performance of end-to-end trained enhancement and ASR joint models. The first row shows an end-to-end ASR Conformer baseline with no enhancement frontend. A significant 20% relative CCER reduction can be observed by adding a neural beamformer frontend which is jointly optimized with ASR backend module. Although the enhanced far-fields signal shows better quality in terms of audibility by enforcing explicit constraints on enhancement frontend ($\lambda_1 = 0.1$), the ASR performance of the joint system suffers degradation.

### 3.3. Audio-Visual Fusion

Since the frame rate of the voice feature was 100Hz and the video feature was 24Hz, we had upsampled the video feature to the same level as the voice feature. After that, we explored a frame-level feature fusion strategy.

**Table 3**: The performance on dev far set of Enhancement ASR joint models with audio-visual fusion

| System | CCER(%) |
|---|---|
| Neural BF + Conformer | 39.8 |
| Neural BF + Conformer + V | (96.6) |
| Neural BF + Conformer + V + init | **39.3** |

Our video features used the hidden representation of the lip-reading neural network, so the dimension 1024 is much larger than the 80 dimensions of speech features. Due to the success of PCA in the field of visual features [40], we first used PCA to reduce the dimension of visual features to 80.

We estimated the PCA transformation matrix on our training set. We sampled feature vectors from our training set instead of using all features. This is because using all feature frames is costly for estimating the transformation matrix. In addition, the 80-dimensional visual features were concatenated with the log Mel-filterbank and follows a projection layer to the Conformer encoder.

As shown in Table 3, the training process of an audio-visual enhancement ASR joint model from scratch could not converge in our experiments. Initializing the audio-visual joint model with audio-only joint model stablized the training process and yielded further improvement.

### 3.4. System Combination

Finally, we combine several systems through ROVER to reach the best performance as in Table 4. These systems include enhancement ASR joint model with neural beamformer frontend and conformer backend, enhancement ASR joint model with neural beamformer frontend and transformer backend and end-to-end conformer model. The final system reaches 34.9% CCER on dev far set and 34.0% CCER on the official test set.

**Table 4**: The performance on dev far set with ROVER and test time augmentation

| System | CCER(%) |
|---|---|
| Chain-TDNN-AV* (official baseline) | 61.8 |
| Neural BF + Conformer + V + init | 39.3 |
| ROVER | 36.0 |
| ROVER + Test Time Augmentation | **34.9** |

## 4. CONCLUSIONS

In this paper, we present the SJTU system for MISP challenge. Multi-channel in and single speaker output (MISO) end-to-end speech recognition system are firstly explored in real far-field scenario. Different data augmentation schemes like speed perturbation, SpecAugment and test-time augmentation are also applied to improve the system performance. Audio-visual fusion is later used to improve ASR system. Overall, with ROVER scheme, our system achieve 34.9% CCER on dev set and 34.0% CCER on test set, which are 26.9% and 28.7% absolute CCER reduction over the official dev set and test set baseline respectively.

## 5. ACKNOWLEDGMENTS

The research in this paper used the LRW-1000 database collected by Institute of Computing Technology Chinese Academy of Sciences.

## 6. REFERENCES


[1] J. R. Hershey *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[2] D. Yu *et al.*, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[3] N. Morgan and H. Bourlard, "Continuous speech recognition using multilayer perceptrons with hidden Markov models," in *Proc. IEEE ICASSP*, 1990, pp. 413–416.

[4] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*. PMLR, 2014, pp. 1764–1772.

[5] W. Chan *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 4960–4964.

[6] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE ICASSP*, 2017, pp. 4835–4839.

[7] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[9] C. Li and Y. Qian, "Listen, watch and understand at the cocktail party: Audio-visual-contextual speech separation," in *Proc. Interspeech*, 2020, pp. 1426–1430.

[10] S. Petridis *et al.*, "End-to-end audiovisual speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 6548–6552.

[11] T. Afouras *et al.*, "Deep audio-visual speech recognition," *IEEE Trans. PAMI.*, 2018.

[12] M. Delcroix *et al.*, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE ICASSP*, 2018, pp. 5554–5558.

[13] Q. Wang *et al.*, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

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

[15] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[16] K. Shimada *et al.*, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 27, no. 5, pp. 960–971, 2019.

[17] T. Ochiai *et al.*, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, pp. 2632–2641.

[18] X. Chang *et al.*, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. IEEE ASRU*, 2019, pp. 237–244.

[19] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *ACCV*. Springer, 2018, pp. 605–621.

[20] P. Li *et al.*, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. IFS*, vol. 14, no. 8, pp. 2000–2012, 2019.

[21] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[22] N. Kanda *et al.*, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR," in *Proc. Interspeech*, 2019, pp. 1248–1252.

[23] H. Erdogan *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[24] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, 3 2016, pp. 196–200.

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

[26] M. Zeineldeen *et al.*, "Conformer-based hybrid ASR system for Switchboard dataset," *arXiv preprint arXiv:2111.03442*, 2021.

[27] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[28] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[29] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[30] W. Zhang *et al.*, "End-to-end far-field speech recognition with unified dereverberation and beamforming," in *Proc. Interspeech*, 2020, pp. 324–328.

[31] ——, "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *Proc. IEEE ICASSP*, 2021, pp. 6898–6902.

[32] B. Martinez *et al.*, "Lipreading using temporal convolutional networks," in *Proc. IEEE ICASSP*, 2020, pp. 6319–6323.

[33] Y. Shuang *et al.*, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[35] H. Huang *et al.*, "Aispeech-SJTU accent identification system for the accented English speech recognition challenge," in *Proc. IEEE ICASSP*, 2021, pp. 6254–6258.

[36] T. Ko *et al.*, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015.

[37] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[38] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[39] J. Du *et al.*, "Aishell-2: Transforming Mandarin ASR research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[40] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE CVPR*, vol. 2. IEEE, 2004, pp. II–II.