Optimizing Audio-Visual Speech Enhancement Using Multi-Level Distortion Measures for Audio-Visual Speech Recognition

Hang Chen, *Member, IEEE*, Qing Wang, Jun Du, *Senior Member, IEEE*, Bao-Cai Yin, Jia Pan, and Chin-Hui Lee, *Life Fellow, IEEE*

***Abstract*—A multi-level distortion measure (MLDM) is proposed as an objective to optimize deep neural network-based speech enhancement (SE) in both audio-only and audio-visual scenarios. The aim is to achieve simultaneous performance improvements in speech quality, intelligibility, and recognition error reductions. Moreover, a comprehensive correlation analysis shows that these three evaluation metrics exhibit high Pearson correlation coefficient (PCC) values with three commonly used optimization objectives: the mean squared error between the ideal ratio and estimated magnitude masks, scale-invariant signal-to-noise ratio, and cross-entropy-guided measure. To further improve the performance, we leverage the complementarities of the three objectives and propose another correlated multi-level distortion measure (C-MLDM) defined as a weighted combination of MLDM and an average correlation measure based on the three PCCs. Experimental results on the TCD-TIMIT corpus corrupted by additive noise demonstrate that MLDM outperforms systems optimized with each objective in both audio-visual and audio-only scenarios, offering improved performances in all three metrics: speech quality, intelligibility, and recognition performance. C-MLDM also consistently outperforms MLDM in all test cases. Finally, the generalizability of both MLDM and C-MLDM is confirmed through extensive testing across diverse datasets, SE model architectures, and linguistic conditions.**

***Index Terms*—Audio-visual, optimization objective, robust speech recognition, speech enhancement, task-generic.**

I. Introduction

SPEECH enhancement (SE) extracts clean speech from signals degraded primarily by noise [1]. SE serves various applications and goals. In human-to-human communication, the focus is on speech quality and intelligibility. In contrast, SE enhances automatic speech recognition (ASR) performance in human-to-machine communication. While task-specific SE achieves favorable results, its drawbacks include resource-intensive development, limited generalizability and increased complexity, hindering its applicability to diverse real-world challenges. To overcome these limitations and cater to a broader range of applications, developing a task-generic SE model that simultaneously enhances speech quality, intelligibility and recognition performance becomes crucial.

Conventional SE algorithms (e.g., [2], [3], [4]) often fail to track unexpected nonstationary noise in real-world conditions. In recent years, data-driven SE approaches (e.g., [5], [6], [7]) using the powerful modeling capabilities of deep neural networks (DNNs) [8], have attracted increasing attention. Intuitively, a task-generic DNN-based SE model can adopt an optimization objective, such as the mean absolute error (MAE) or mean squared error (MSE) between corresponding waveforms or spectrograms [9], [10], [11] of enhanced and clean speech. Although achieving good results, studies have shown that it is not directly related to speech quality [12], [13], [14] or intelligibility. Moreover, previous works [15], [16], [17] have noted that SE can lead to ASR performance degradation. Recent papers [14], [18], [19], [20] have addressed these challenges by exploring alternative optimization objectives. However, these objectives are limited to demonstrating effectiveness in specific evaluations, such as improving speech quality and/or intelligibility, which may inadvertently overlook ASR performance [18], [21]. Alternatively, some algorithms [20], [22] have prioritized enhancing ASR accuracy while sacrificing speech quality and intelligibility. This makes achieving task-generic SE in audio-only scenarios difficult.

The McGurk effect [23] suggests a strong influence of vision on human auditory perception. Follow-up studies (e.g., [24], [25], [26]) have shown that visual cues, such as lip movements, can help speech perception, especially in noisy environments. Recent studies [27], [28], [29] have also demonstrated that adding the visual modality can substantially enhance the speech quality and intelligibility of DNN-based SE models. Hence, there is a solid motivation to explore the potential of task-generic SE models in the audio-visual scenario, as it is more likely to yield promising outcomes. However, limited attention has

Manuscript received 22 August 2023; revised 25 February 2024; accepted 8 April 2024. Date of publication 25 April 2024; date of current version 3 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 62171427. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hakan Erdogan. *(Corresponding author: Jun Du.)*

Hang Chen, Qing Wang, and Jun Du are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230026, China (e-mail: ch199703@mail.ustc.edu.cn; qingwang2@ustc.edu.cn; jundu@ustc.edu.cn).

Bao-Cai Yin and Jia Pan are with iFLYTEK Research, Hefei 230088, China (e-mail: bcyin@iflytek.com; jiapan@iflytek.com).

Chin-Hui Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

The source codes are publicly available. https://github.com/coalboss/CMLDM.

Digital Object Identifier 10.1109/TASLP.2024.3393732

been given to investigating the impact of AVSE on audio-visual speech recognition (AVSR) performance.

In this study, our primary focus is designing optimization objectives to enhance speech with simultaneous improvements in quality, intelligibility, and recognition performance. Our main contributions can be summarized as follows:

1) conducting a comprehensive correlation analysis to demonstrate complementarities among three commonly used optimization objectives, namely, MSE between the ideal ratio and estimated magnitude masks (MSE-M), scale-invariant signal-to-noise ratio (SISNR), and cross-entropy-guided measure (CEGM), together impacting all three SE evaluation metrics, including speech quality, intelligibility and recognition performances.
2) proposing a multi-level distortion measure (MLDM) that combines MSE-M, SISNR, and CEGM in a novel sequential and weighted manner to leverage the complementarity for matching objectives to multitask evaluation metrics.
3) proposing a correlated multi-level distortion measure (C-MLDM) to augment the interactions among the three objectives by adding an additional correlation term based on the Pearson correlation coefficients to MLDM.
4) confirming the effectiveness and generalizability of MLDM and C-MLDM via a series of experiments in both audio-visual and audio-only scenarios and verifying the benefit of adding a visual modality to SE and ASR.

The rest of the paper is organized as follows. Section II introduces the related works. Section III presents the results of the correlation analysis, which motivated our research. Section IV describes our proposed methods, including MLDM and C-MLDM. Section V analyzes the experimental results. Finally, we summarize our findings in Section VI.

II. Related Works

A. Audio-Visual Speech Enhancement

AVSE has made significant progress since its inception, with early works [27], [28], [29], [30], [31], [32], [33] laying the foundation. Deep neural network-based AVSE models [34], [35], [36] have gained attention. However, they were primarily evaluated under constrained conditions, such as using fixed sets of phrases or a limited number of known speakers.

To address the challenge of unknown speakers and noise types, [37] introduced a deep AVSE model with separate magnitude and phase subnetworks. The model minimizes the MAE between the predicted magnitude spectrogram and the ground truth while maximizing the cosine similarity between the phase prediction and the ground truth. Another approach [38] directly estimated the complex spectrogram using facial embeddings of the source speaker. The optimization objective is based on MSE between estimated and clean complex spectrograms. In [39], a time-domain AVSE model was proposed using ConvTasNet [40] to estimate the waveform directly. It was trained by optimizing the SISNR between the enhanced and clean waveforms. In another work, [41] utilized phone units as the classification target, providing suitable visual embedding for time-domain AVSE. Furthermore, [42] employed audio embeddings from noisy multichannel speech to complement the visual embedding in time-domain AVSE.

Recently, [43] presented a novel multimodal embedding-aware speech enhancement (MEASE) technique that extended the visual-only pretrained embedding extractor to an audio-visual pretrained extractor. The MEASE model was optimized using MSE-M. In [44], it was reported that the visual modality can cause performance degradations at high SNR levels. To address this, a late fusion model was proposed, which combined two magnitude masks estimated by the audio and video modalities. The optimization objective in this case is still MSE-M. [45] introduced a two-stage audio-visual fusion strategy, incorporating audio-visual deep clustering to minimize the MSE between the embedding matrix and the affinity matrix of the ideal binary mask (IBM) [46]. Furthermore, [47] utilized audio-visual temporal synchronization as a direct and dominant cue to transfer knowledge from a pretrained synchronization model to a time-domain AVSE model. The model was trained using cross entropy for speaker classification and the SISNR. Lastly, [48] presented a unified framework to efficiently learn different types of audio-visual correlation evidence. The framework generates aligned audio-visual representations for time-domain AVSE and active speaker detection.

In recent advancements, self-supervised learning has emerged as a groundbreaking paradigm in AVSE. [49] uses a deep, multi-instance, multi-label learning framework to derive audiovisual object models from unlabeled video content and subsequently leverages visual context to facilitate audio source separation in novel videos. Similarly, [50] advocates cultivating an integrated multisensory representation through self-supervised means by orchestrating a neural network to determine the temporal alignment between video frames and corresponding audio segments. This novel learned representation is then used to distinguish between on-screen and off-screen audio sources. Further advancing this field, [51] employs self-supervised learning techniques to transform a video into a collection of discrete audiovisual objects. This approach introduces a model that uses attention mechanisms to localize and cluster sound sources while utilizing optical flow to assemble information across temporal dimensions, demonstrating the significant potential of self-supervised learning to enhance AVSE capabilities.

B. Optimization Objectives in Speech Enhancement

Numerous DNN-based SE models have achieved state-of-the-art performance by minimizing the MAE or MSE between enhanced and clean waveforms or spectrograms. However, they still suffer from the loss-metric mismatch problem [52]. Studies [13], [53] have indicated that MSE or MAE at the signal level exhibits a limited correlation with speech quality. [54] demonstrated that lower MSE or MAE scores do not necessarily guarantee higher perceptual evaluation of speech quality (PESQ) [55] or improved short-time objective intelligibility (STOI) [56], which are commonly used metrics to evaluate speech quality and intelligibility, respectively. Furthermore, some SE models generate unnatural-sounding speech [57]. Additionally, optimizing for MSE or MAE may not necessarily improve ASR performance, and can even increase the word error rate (WER) [9], [58], [59], [60]. This semantic gap results in inefficient model training.

To address this, several studies have explored optimizing the evaluation metrics directly to align model training with the final goal. For instance, some studies have adopted the STOI as an optimization objective to enhance speech intelligibility [14], [61], while others have proposed complex methods to approximate the STOI [62]. However, some evaluation metrics, such as PESQ and WER, are inherently nondifferentiable and discontinuous, making direct gradient calculation and training challenging. To address this challenge, [18], [19] explored using reinforcement learning (RL) techniques to optimize SE models with PESQ and WER as reward functions. However, RL-based methods often encounter optimization difficulties and may result in limited improvements in the target metric while potentially causing degradation in other related metrics.

Other approaches focus on addressing the loss-metric mismatch using the deep feature loss, which uses representations learned from a different task to construct similarity metrics [63]. For example, [21] trained a PESQ prediction model to optimize the SE model by improving the enhanced output. [64] introduced a novel phone-fortified perceptual loss (PFPL) for comparing enhanced and clean speech by utilizing the Wasserstein distance [65] between the latent representations extracted from the wav2vec model [66]. Ref. [67] presented a DNN-based estimator for 25 temporal acoustic parameters [68] and defined a temporal acoustic parameter (TAP) loss, minimizing the distance between estimated acoustics for clean and enhanced speech. Furthermore, [69] proposed a phonetic-aligned acoustic parameter (PAAP) loss that incorporates temporal parameters into associating acoustic parameters and phonemes based on the TAP loss. The aforementioned techniques enhanced quality but with slight ASR improvement. Ref. [22] developed two DNNs, one dedicated to SE and the other mimicking the WER derived from an ASR system. Moreover, [20] introduced a cross-entropy-guided measure (CEGM) formulated as the cross entropy of the hidden Markov model (HMM) state posteriors between the enhanced and clean outputs of the acoustic model. However, these methods improve ASR accuracy at the expense of perceptual quality.

In contrast to the aforementioned task-specific optimization objects, our proposed MLDM and C-MLDM are designed to simultaneously improve speech quality, intelligibility, and recognition performance, thus facilitating a task-generic SE model. Specifically, MLDM embodies a novel amalgamation of MSE-M, SISNR and CEGM in a sequential and weighted manner that outperforms any single target. Further advancing this methodology, C-MLDM incorporates a correlation measurement to enhance the synergistic interactions among the three basic objectives, achieving remarkable improvements.

III. Motivation

Elucidating the relationships between different optimization goals and different evaluation metrics typically requires an extensive and time-consuming training process. In an effort to bypass this laborious training phase, we first conducted a comprehensive correlation analysis among various optimization objectives and evaluation metrics in both audio-only and audio-visual scenarios. As shown in Fig. 1(a), experiments were performed on the TCD-TIMIT corpus [70] corrupted by simulated additive noises with the same process as in [43], denoted as SNTCD-TIMIT. The baseline optimization objectives are the commonly used SISNR and MSE between the ideal ratio and estimated magnitude masks (MSE-M). Moreover, we extend the audio-only CEGM to an audio-visual version as a baseline optimization objective, an approach similar to the one described in [20].

To evaluate speech quality and intelligibility, we employed PESQ and STOI, respectively. PESQ applies an auditory transform to compare the loudness spectrum of clean and enhanced speech, yielding a score ranging from -0.5 to 4.5; higher scores indicate better speech quality. In contrast, the STOI compares the temporal envelopes of clean and enhanced speech in short-time regions, providing values between 0 and 1, with higher values representing better speech intelligibility. The recognition performance is evaluated using in-domain automated speech recognition (IdASR) and in-domain audio-visual speech recognition (IdAVSR) models for audio-only and audio-visual scenarios, respectively. IdASR and IdAVSR are hybrid DNN-HMM models that share a 2-gram phone-based language model, with the main difference lying in their acoustic models. In IdASR, the acoustic model comprises an audio processing module followed by a sequence module. Additionally, the IdAVSR includes an additional video process module running in parallel with the audio process module. The outputs from these two process models are concatenated and then fed to the subsequent sequence module. See [71] for details about the model structure and training process. The phone error rate (PER) serves as a metric and is calculated as follows, out of $\mathrm{N_u}$ units being evaluated:

$$\mathrm{PER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}_u} \tag{1}$$

where S, D and I denote the number of substitution, deletion and insertion errors, respectively.

Inspired by previous studies (e.g., [20]), the Pearson correlation coefficient (PCC) [72] is adopted to identify correlations with sample pairs (y_k, z_k) of the evaluation metric and optimization objective and can be calculated as follows:

$$\rho(\{y_k\}, \{z_k\}) = \frac{\sum_{k=0}^{K-1}(y_k - \overline{y})(z_k - \overline{z})}{\sqrt{\sum_{k=0}^{K-1}(y_k - \overline{y})^2}\sqrt{\sum_{k=0}^{K-1}(z_k - \overline{z})^2}} \tag{2}$$

where K is the total number of samples. $\overline{y} = \sum_{k=0}^{K-1} y_k / K$ and $\overline{z} = \sum_{k=0}^{K-1} z_k / K$ are means of the sample points in $\{\boldsymbol{y_k}\}$ and $\{\boldsymbol{z_k}\}$, respectively. Equation (2) can be considered an expression of the ratio of how much the two datasets vary together instead of how much they vary separately. The magnitude indicates the correlation's strength and the sign indicates whether the correlation is positive or negative.

We aim to establish a monotonic relationship between the baseline optimization objectives (MSE-M, SISNR and CEGM, MAE-M) and the evaluation metrics (PESQ, STOI and PER).

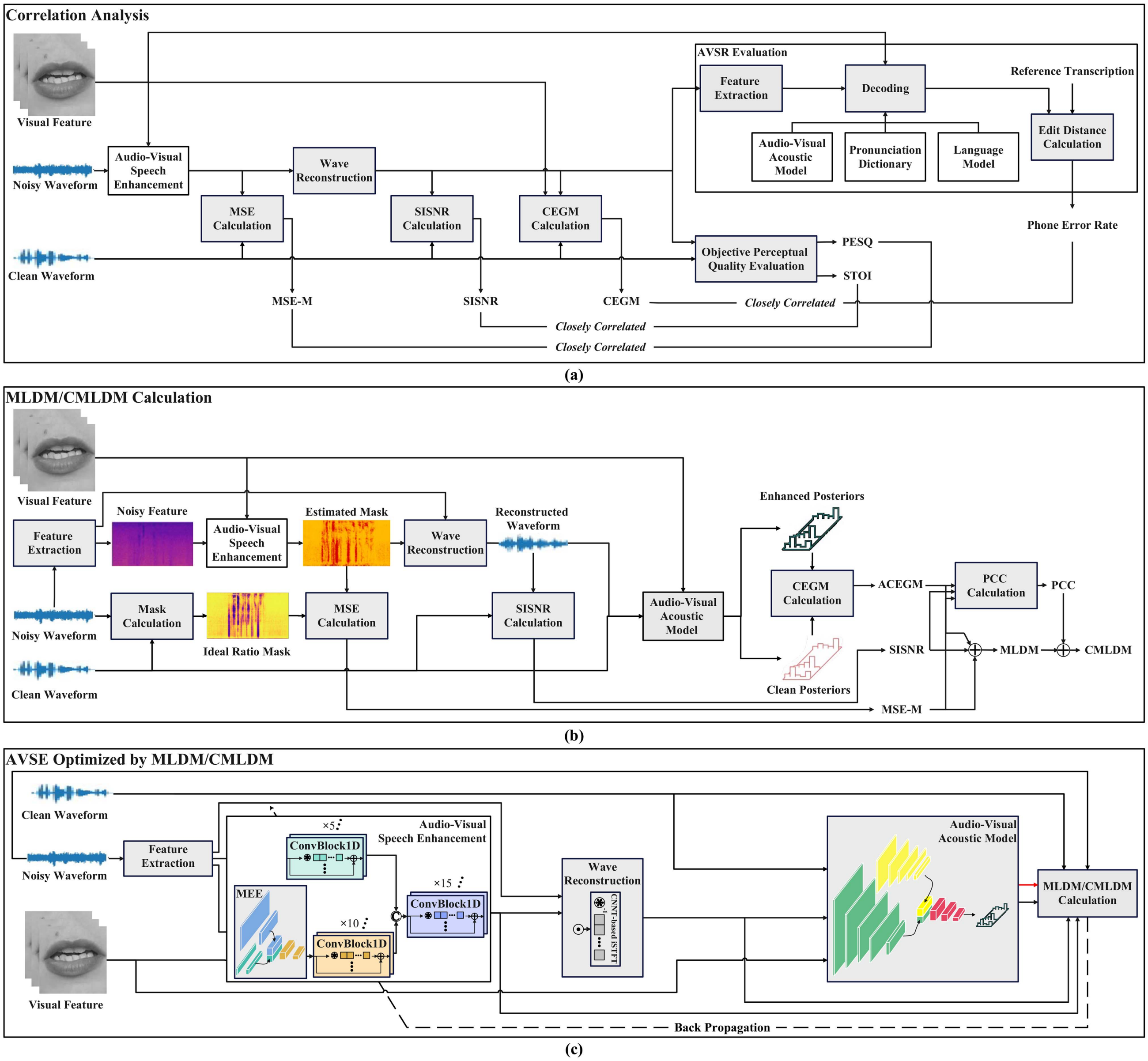

Fig. 1. Our proposed framework: (a) The correlation analysis shows the complementarity of MSE-M, SISNR, and CEGM concerning PESQ, STOI and PER; (b) a block diagram for calculating MLDM and C-MLDM; and (c) MLDM and C-MLDM are used to optimize the model parameters of DNN-based AVSE.

To achieve this, we employ a mapping technique that accounts for the nonlinear relationship, allowing us to linearize the data to utilize the PCC for correlation evaluation. Motivated by [73], [74], [75], a logistic function is used here:

$$\mathcal{M} = f(\mathcal{L}) = \frac{c_1}{1 + \exp(c_2\mathcal{L} + c_3)} \tag{3}$$

where $\mathcal{M}$ represents the evaluation metric score (PESQ, STOI, and PER) and $\mathcal{L}$ represents the optimization objective score (MSE-M, SISNR and CEGM). The function $f(\cdot)$, with values ranging from 0 to 1, can be regarded as an estimator of $\mathcal{M}$ based on $\mathcal{L}$, and the constants c_1, c_2, and c_3 are used to balance order-of-magnitude discrepancies. They are determined through data-fitting using a least-squares method. It is worth noting that the mapping is performed with a monotonic logistic function that does not influence rankings. Subsequently, the evaluation metric's performance is represented using the PCC, applied to the mapped objective scores, $f(\mathcal{L})$. The MSE-M, SISNR, CEGM, PESQ, STOI, and PER were computed for each utterance. This procedure was utilized to calculate all the correlation coefficients in our study. We are interested in the correlation strength; thus, only the PCC magnitudes, ranging from 0 to 1, are presented in the experimental results.

Fig. 2 illustrates the average PCCs between one of the three metrics and another of the three objectives on the SNTCD-TIMIT test set. We observe varying degrees of correlation between objective-metric pairs. Specifically, MSE-M shows the highest PCC with PESQ (0.92), while SISNR demonstrates the highest PCC with STOI (0.89). CEGM exhibits the highest PCC with PER for both audio-only and audio-visual scenarios, with

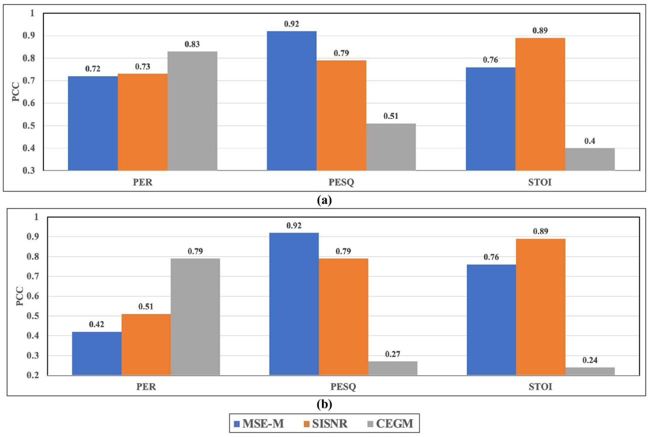

Fig. 2. Average PCC comparisons between a pair of one evaluation metric (PER, PESQ or STOI) and one optimization objective (MSE-M, SISNR or CEGM) calculated in (a) audio-only and (b) audio-visual scenarios.

values of 0.83 and 0.79, respectively. This highlights the complementary nature of MSE-M, SISNR, and CEGM concerning the three performance metrics in both scenarios. Leveraging this complementarity, we can design near-optimal objectives to achieve task-generic SE.

Nonetheless, when comparing the PCCs with PER between audio-only and audio-visual scenarios, we observed a notable decrease in PCCs for MSE-M and SISNR, with reductions of 0.3 and 0.22, respectively, when using the AVSR backend. This decrease in PCCs can be attributed to the fact that the AVSR's performance is jointly influenced by both audio and video inputs, making it less sensitive to partial audio input distortion. In contrast, CEGM incorporates both audio and video components as inputs to the optimization objective, establishing a more direct link to the final evaluation metric. The difference between the audio-only and audio-visual scenarios also illustrates that adding the visual modality enhances the complementarity of MSE-M, SISNR, and CEGM regarding all three objectives.

This analysis leads us to consider two essential questions:

$\mathcal{Q}1$ How can we leverage the complementarity to design an effective optimization objective for AVSE to improve speech quality, intelligibility, and AVSR performance?

$\mathcal{Q}2$ How will the individual characteristics of MSE-M, SISNR, and CEGM affect the optimization process?

Motivated by $\mathcal{Q}1$, we propose MLDM with more detail in Section V-C. Motivated by $\mathcal{Q}2$, we also observed a discrepancy in the convergence speeds of MSE-M, SISNR, and CEGM during training using MLDM, leading us to propose C-MLDM in Section V-D for further improvements.

IV. Proposed Techniques

Inspired by the concept of function smoothing [76], which is a commonly used approach to approximate nondifferentiable functions with differentiable functions, we adopt a similar strategy by weighting MSE-M, SISNR, and CEGM to leverage their complementarity for matching evaluation metrics from multiple tasks. As a result, we define MLDM as an iterative combination of these three selected objectives. Furthermore, our novel C-MLDM surpasses MLDM by not only considering the individual values of the three optimization objectives but also incorporating the correlations among them. Fig. 1(b) illustrates the calculation framework of our proposed MLDM and C-MLDM. In the following sections, we elaborate on MLDM and C-MLDM.

A. Multi-Level Distortion Measure

As shown in Fig. 1(b), the AVSE model takes B pairs of noisy spectrogram features $\{X_i \in \mathbb{R}^{T\times C}\}$ and the lip frame sequence $\{V_i \in \mathbb{R}^{\frac{T}{4}\times H\times W}\}$ as inputs to estimate the magnitude mask $\{\widehat{M}_i \in \mathbb{R}^{T\times C}\}$. The process is described as follows:

$$\{\widehat{M}_i\} = \mathcal{F}(\{X_i\}, \{V_i\}; W) \tag{4}$$

where $\mathcal{F}$ and W denote the AVSE model and its parameter set. $i = 0, 1, \ldots, B-1$ and B is the batch size. T and C denote the number of frames and frequency bins for the spectrogram, respectively. H and W denote the length and width of the lip frame, respectively. Here, we use $B = 32$, $C = 201$ and $H = W = 96$ by default.

We first adopted the average MSE-M $\overline{\mathcal{L}}^{\mathrm{MSE-M}}$ between $\{\widehat{M}_i\}$ and the ideal ratio mask (IRM) $\{M_i \in \mathbb{R}^{T\times C}\}$ to compare the spectral similarity between the enhanced speech and the clean speech. MSE-M can be computed as follows:

$$\begin{aligned}\overline{\mathcal{L}}^{\mathrm{MSE-M}} &= \frac{\sum_{i=0}^{B-1}\mathcal{L}_i^{\mathrm{MSE-M}}}{B}\\ &= \frac{\sum_{i=0}^{B-1}\sum_{t=0}^{T-1}\sum_{j=0}^{C-1}(\widehat{m}_{i,t,j}-m_{i,t,j})^2}{BTC}\end{aligned} \tag{5}$$

where $\mathcal{L}_i^{\mathrm{MSE-M}}$ denotes the MSE-M score of one sample. $\widehat{m}_{i,t,j}$ and $m_{i,t,j}$ are the values at the t-th frame and j-th frequency bin of $\widehat{M}_i$ and M_i, respectively.

Then, we use SISNR to measure the distortions of the enhanced speech on the waveform. In a waveform reconstruction module $\mathcal{U}$, $\widehat{M}_i$ is used to filter the noisy spectrum $X_i^{\mathrm{spec}} \in \mathbb{C}^{T\times C}$, and the filtered spectrum is fed to a 1D transposed convolution layer to reconstruct waveform $\widehat{\boldsymbol{s}}_i \in \mathbb{R}^L$. The whole reconstruction process is briefly described as follows:

$$\{\widehat{\boldsymbol{s}}_i\} = \mathcal{U}(\{X_i^{\mathrm{spec}}\}, \{\widehat{M}_i\}; W_{\mathrm{stft}}) \tag{6}$$

where W_{stft} denotes the forward weight of the STFT, which is also the parameter set of the 1D transposed convolution layer. L is the length of the waveform. Then, the average SISNR $\overline{\mathcal{L}}^{\mathrm{SISNR}}$ between $\{\widehat{\boldsymbol{s}}_i\}$ and the clean waveform $\{\boldsymbol{s}_i \in \mathbb{R}^L\}$ is calculated as follows:

$$\widetilde{s}_{i,\tau} = \widehat{s}_{i,\tau}\left(\sum_{\tau=0}^{L-1} s_{i,\tau}^2\right) \bigg/ \left(\sum_{\tau=0}^{L-1}\widehat{s}_{i,\tau}s_{i,\tau}\right)$$

$$\overline{\mathcal{L}}^{\mathrm{SISNR}} = \frac{\sum_{i=0}^{B-1}\mathcal{L}_i^{\mathrm{SISNR}}}{B}$$

$$= -\frac{10}{B}\sum_{i=0}^{B-1}\log\frac{\sum_{\tau=0}^{L-1}s_{i,\tau}^2}{\sum_{\tau=0}^{L-1}[\widetilde{s}_{i,\tau}-s_{i,\tau}]^2} \quad (7)$$

where $\mathcal{L}_i^{\mathrm{SISNR}}$ denotes the SISNR score of one sample. $\widehat{s}_{i,\tau}$ and $s_{i,\tau}$ are waveform values at the τ-th time step of estimated $\widehat{\boldsymbol{s}}_i$ and clean $\boldsymbol{s}_i$, respectively.

Finally, we notice that low-level acoustic features such as spectrum and waveform are not directly correlated with AVSR accuracies. Inspired by CEGM, we adopt a DNN-HMM audio-visual acoustic model for extracting high-level representations derived from low-level acoustic and visual features. By utilizing valuable acoustic knowledge from the backend AVSR model, we believe that the high-level representations can better assess the AVSR performances. Given $\{\widehat{\boldsymbol{s}}_i\}$ and $\{V_i\}$, the audio-visual acoustic model outputs the clustered HMM state posterior probabilities $\{\mathrm{p}(\widehat{H}_i|\widehat{\boldsymbol{s}}_i, V_i) \in \mathbb{R}^{T\times I}\}$. The extraction process can be summarized as follows:

$$\{\mathrm{p}(\widehat{H}_i|\widehat{\boldsymbol{s}}_i, V_i)\} = \mathcal{G}(\{\widehat{\boldsymbol{s}}_i\}, \{V_i\}; W_{\mathrm{am}}) \quad (8)$$

where $\mathcal{G}$ and W_{am} denote the audio-visual acoustic model and the corresponding parameter set, respectively, $\mathrm{p}(\widehat{H}_i|\boldsymbol{s}_i, V_i)$ denotes the state posteriors of one sample, $\widehat{H}_i = [\widehat{h}_{i,0}, \widehat{h}_{i,1}, \ldots, \widehat{h}_{i,T-1}]$ is a random process of length T, and $\widehat{h}_{i,t}$ is a random variable whose values range over all clustered HMM states $\{0, 1, \ldots, I\}$. The acoustic model also maps the clean waves $\boldsymbol{s}_i$ and V_i to the high-level label $\mathrm{p}(H_i|\boldsymbol{s}_i, V_i)$.

Next, we adopt an average cross entropy to measure the similarity between enhanced and clean high-level features:

$$\overline{\mathcal{L}}^{\mathrm{ACEGM}} = \frac{\sum_{i=0}^{B-1}\mathcal{L}_i^{\mathrm{ACEGM}}}{B}$$
$$= -\sum_{i=0}^{B-1}\sum_{t=0}^{T-1}\sum_{j=1}^{I-1}\frac{\mathrm{p}_t(h_{i,t}=j|\boldsymbol{s}_i, V_i)\log \mathrm{p}_t(\widehat{h}_{i,t}=j|\widehat{\boldsymbol{s}}_i, V_i)}{TB}. \quad (9)$$

Notably, unlike [20], which solely outputs FBANK features desired by the backend, our method allows reconstructing waveforms from the mask outputted by the SE front-end, and coupling with the backend is realized through an online DNN-based feature extractor. To distinguish our method from previous approaches, we refer to it as the audible cross-entropy-guided measure (ACEGM). $\mathcal{L}_i^{\mathrm{ACEGM}}$ denotes the CEGM score of one sample. There is a total of I clustered HMM states and j denotes the j-th state.

However, the raw values of MSE-M, SISNR, and ACEGM exhibit significant differences in their order of magnitude. To address this disparity, we perform a normalization as follows:

$$\widetilde{\mathcal{L}}^{\mathrm{MSE-M}} = 10^{\lfloor \log_{10}\frac{1}{|\overline{\mathcal{L}}^{\mathrm{MSE-M}}|}\rfloor}\overline{\mathcal{L}}^{\mathrm{MSE-M}} \quad (10)$$

where $\lfloor\cdot\rfloor$ is the floor function. The normalization factor $c^{\mathrm{MSE-M}}$ is treated as a constant when computing the gradient. c^{SISNR} and c^{ACEGM} are also calculated in the same way for normalizing $\mathcal{L}^{\mathrm{SISNR}}$ and $\mathcal{L}^{\mathrm{ACEGM}}$ to $\widetilde{\mathcal{L}}^{\mathrm{SISNR}}$ and $\widetilde{\mathcal{L}}^{\mathrm{ACEGM}}$, respectively. The normalization operation ensures the magnitude of $\widetilde{\mathcal{L}}^{\mathrm{MSE-M}}$, $\widetilde{\mathcal{L}}^{\mathrm{SISNR}}$ and $\widetilde{\mathcal{L}}^{\mathrm{ACEGM}}$ in the range of 0 to 1. Then, MLDM can be calculated as follows:

$$\mathcal{L}^{\mathrm{MLDM}} = \alpha\widetilde{\mathcal{L}}^{\mathrm{MSE-M}} + \beta\widetilde{\mathcal{L}}^{\mathrm{SISNR}} + (1-\alpha-\beta)\widetilde{\mathcal{L}}^{\mathrm{ACEGM}} \quad (11)$$

where the weights α and β are determined as hyperparameters, which are discussed in Section V.

MLDM is differentiable and thus can be easily used as the objective function to optimize DNN-based AVSE. MLDM focuses on distortions contained in the magnitude spectrum, degraded speech waveform, and in the high-level representation extracted by the audio-visual acoustic model. Clearly, in contrast to baselines, MLDM provides a more comprehensive similarity measure between enhanced and clean speech.

The MLDM framework for optimizing the DNN-based AVSE is also shown in Fig. 1(c). The model is trained with gradient descent by back-propagation [77].

B. Correlated Multi-Level Distortion Measure

The critical contribution of C-MLDM lies in its incorporation of the values of the three similarity measures and explicit modeling of their correlations. This motivation stems primarily from our observation of the experimental results obtained from MLDM. While the MLDM-optimized AVSE model demonstrated consistent improvements across the three evaluation metrics regarding overall average results, the sample-level improvements displayed an inconsistent trend. In particular, certain samples demonstrated significant PER reduction but showed less noticeable improvements in PESQ and STOI. Conversely, other samples exhibited the opposite pattern.

Accordingly, we propose a correlation measure (CM) between three basic optimization objects in the MLDM. During the training stage, it is imperative not only to minimize the values of these three basic optimization objects but also to ensure their synchronized variation. To achieve this, we use the data in a batch to calculate the correlation coefficient between any pair of the basic optimization objects. By averaging these three coefficients, we derive the final correlation measure.

As outlined in Section III, we initially employ the PCC to quantify the correlation between two basic optimization objectives and a logistic function to capture the nonlinear relationship and linearize the data. The calculation process can be briefly described as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{CM}} = \frac{1}{3}[&\rho(\{f(\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}})\}, \{f(\widetilde{\mathcal{L}}_i^{\mathrm{SISNR}})\})\\ &+ \rho(\{f(\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}})\}, \{f(\widetilde{\mathcal{L}}_i^{\mathrm{ACEGM}})\})\\ &+ \rho(\{f(\widetilde{\mathcal{L}}_i^{\mathrm{SISNR}})\}, \{f(\widetilde{\mathcal{L}}_i^{\mathrm{ACEGM}})\})]\end{aligned} \quad (12)$$

where $\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}}$, $\widetilde{\mathcal{L}}_i^{\mathrm{SISNR}}$ and $\widetilde{\mathcal{L}}_i^{\mathrm{ACEGM}}$ represent the normalized versions of $\mathcal{L}_i^{\mathrm{MSE-M}}$, $\mathcal{L}_i^{\mathrm{SISNR}}$ and $\mathcal{L}_i^{\mathrm{ACEGM}}$, respectively. The normalization and $\rho(\cdot)$ are the same as in (10) and (2), respectively. $f(\cdot)$ is the same as in (3) with $c_1 = c_2 = c_3 = 1$.

We notice that all three PCCs are positive and would like to make $(1 - \mathcal{L}^{\mathrm{CM}})$ as small as possible. Accordingly, C-MLDM is defined as follows:

$$\mathcal{L}^{\mathrm{C-MLDM}} = (1-\gamma)\mathcal{L}^{\mathrm{MLDM}} + \gamma(1 - \mathcal{L}^{\mathrm{CM}}) \quad (13)$$

where γ is a hyperparameter to control the correlation measure weight; discussed in Section V-C2.

C-MLDM emphasizes the correlation between each basic objective in MLDM, explicitly demanding a similar changing trend among them. This prevents the model from converging to a minimum point that is solely associated with a specific objective. The framework of C-MLDM for guiding the front-end DNN-based AVSE is also illustrated in Fig. 1(c). Similar to previous approaches, the model is trained with gradient descent by back-propagation.

V. Experimental and Results Analyses

A. Implementation Detail

We first performed a series of experiments on the SNTCD-TIMIT. MEASE and its audio-only version, no embedding-aware speech enhancement (NoEASE) [43] were adopted as SE models. As shown in Fig. 1(c), the MEASE model consists of a pretrained multimodal embedding extractor (MEE) module and three stacks of ConvBlock1Ds. Each ConvBlock1D includes a 1D convolution layer with a residual connection, a ReLU activation, and a batch normalization, as in [37]. The MEE module combines the noisy filter bank (FBANK) feature and the lip frames to generate a multimodal embedding. This embedding is processed by the orange stack consisting of 10 ConvBlock1Ds. The noisy log power spectra (LPS) feature is processed by the green stack consisting of 5 ConvBlock1Ds. The outputs of these stacks are concatenated along the channel dimension and fed into the top stack (blue-violet), which consists of 15 ConvBlock1Ds, to obtain a magnitude mask. In comparison, the NoEASE model lacks the pre-trained MEE module and the orange ConvBlock1Ds stack. For training, we used the Adam [78] optimizer for 100 epochs, implementing early stopping if there was no improvement in the validation loss for 10 consecutive epochs. The initial learning rate was set to 0.0003 and halved during training if there is no improvement for 3 epochs in the validation loss. The best model was selected with the lowest validation loss.

B. Complementarity Analysis

To further validate the complementarity of MSE-M, SISNR, and CEGM regarding speech quality, intelligibility and recognition errors, as discussed in Section III, we first compare the average PER (in %), PESQ, and STOI (in %) among the unprocessed system ("noisy") and the SE models optimized using MSE-M, SISNR and CEGM on the SNTCD-TIMIT test set in audio-only and audio-visual scenarios. The results are depicted in Fig. 3.

A key finding is the strong agreement between the correlation analysis and the optimization results in both audio-only and audio-visual scenarios. Specifically, the optimization objective that exhibits a higher correlation with a specific metric tends to yield a greater improvement for that metric. For instance, MSE-M gives the highest PCC of 0.92 for PESQ, as shown in the middle of Fig. 2(a) and (b), and achieves the highest PESQ gain of 0.37 and 0.60, respectively, in the middle of

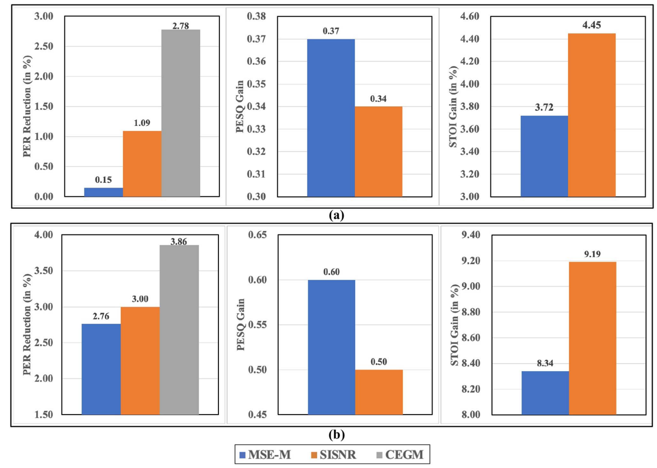

Fig. 3. Comparison of average PER, PESQ and STOI among noisy and SE models optimized by MSE-M, SISNR and CEGM on the SNTCD-TIMIT test set in audio-only (a) and audio-visual scenarios (b). Note that CEGM solely outputs FBANK features desired by the back end, which cannot perfectly reconstruct the waveform for calculating PESQ and STOI.

Fig. 3(a) and (b) for both audio-only and audio-visual scenarios. Similarly, SISNR shows the highest PCC of 0.89 in the right part of Fig. 2(a) and (b) and obtains the highest STOI gains of 4.45% and 9.19%, respectively, in the right part of Fig. 3(a) and (b) for both scenarios. Moreover, CEGM exhibits the highest PCCs of 0.83 and 0.79, with PER shown on the left of Fig. 2(a) and (b), which also achieved the highest PER reductions of 2.78% and 3.86%, respectively, on the left of Fig. 3(a) and (b) for both cases.

Moreover, the inclusion of the visual modality leads to great improvements in speech quality, intelligibility and recognition accuracies. Specifically, MSE-M and SISNR demonstrate superior performances to the noisy baseline across all SNR levels, with average extra reductions in PER of 2.61% (from 0.15% in Fig. 3(a) to 1.91% in Fig. 3(b)) and 3.01% (from 1.09% in Fig. 3(a) to 3.00% in Fig. 3(b)), respectively. This indicates that including unprocessed visual input helps mitigate the data mismatch between training and testing for the backend AVSR model. While CEGM consistently achieves PER reductions from the noisy baseline at all SNR levels, it can be inferred that the visual modality amplifies the advantages of MSE-M, SISNR, and CEGM and their complementarity.

C. Performance Analysis of MLDM

1) Overall Comparisons: To evaluate the effectiveness of our proposed MLDM, we present a comparison of average PER (in %), PESQ, and STOI (in %) among noisy, three baseline objectives (MSE-M, SISNR, and CEGM) and our proposed MLDM in both audio-only and audio-visual scenarios, as shown in Table I. From the results, we can observe that the MLDM-optimized AVSE model effectively combines the advantages of the three baselines, resulting in top performances across all evaluation metrics in both scenarios. Additionally, we also calculate the average PCCs between the three evaluation metrics and MLDM

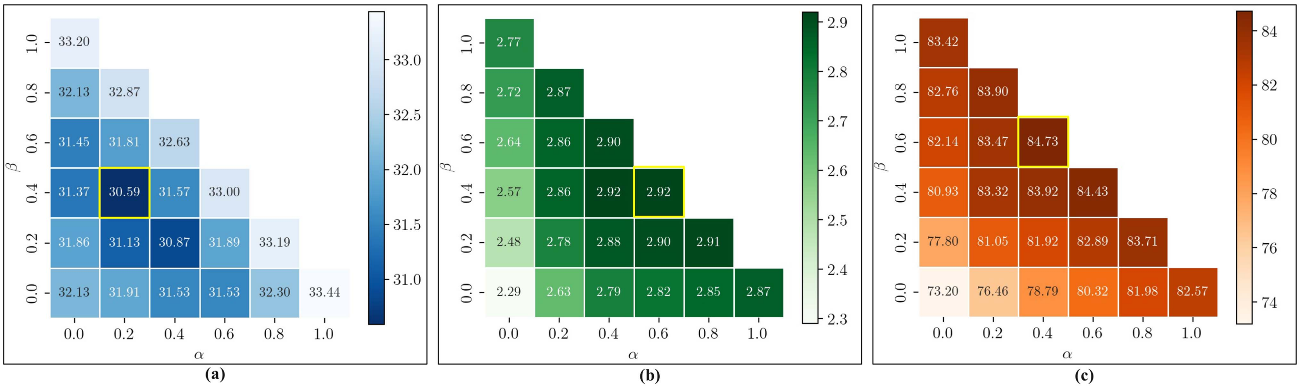

Fig. 4. Average performance comparisons of MLDMs with different α and β parameter values in (11) evaluated on the SNTCD-TIMIT test set in the audio-visual scenario for (a) PER (in %), (b) PESQ and (C) STOI (in %).

TABLE I
COMPARISON OF AVERAGE PER, PESQ AND STOI AMONG NOISY AND SE MODELS OPTIMIZED BY MSE-M, SISNR, CEGM, MLDM AND C-MLDM ON THE SNTCD-TIMIT TEST SET IN AUDIO-ONLY AND AUDIO-VISUAL SCENARIOS

Metric	**PER** (in %) ↓		**PESQ** ↑		**STOI** (in %) ↑	
Scenario	**A**	**AV**	**A**	**AV**	**A**	**AV**
Noisy	50.23	36.20	2.27		74.23	
MSE-M	50.08	33.44	2.64	2.87	77.94	82.57
SISNR	49.13	33.20	2.61	2.77	78.68	83.42
CEGM	47.45	32.34	\		\	
MLDM	44.66	30.74	2.70	2.91	79.57	83.77
C-MLDM	40.47	28.09	2.79	3.02	80.64	84.91
MLDM's PCC	0.87	0.80	0.94	0.93	0.91	0.90

A: audio-only, AV: audio-visual.

on the SNTCD-TIMIT test set and list them in the bottom row of Table I. Notably, when compared with the PCC results in Fig. 2, MLDM consistently exhibits the highest PCCs for all evaluation metrics in both scenarios.

The strong alignment between the performance of MLDM-optimized SE models and the results of correlation analysis highlights the robust alignment of MLDM with quality, intelligibility and recognition performance. Further, it demonstrates the effectiveness of MLDM as an optimization objective. These consistent results provide additional support for the effectiveness of MLDM as an optimization objective in various evaluation scenarios.

2) An Ablation Study on Hyperparameter Setting: We also investigate the impact of the hyperparameters α and β in (11) on the AVSE performance. Fig. 4(a), (b), and (c) present the average PER, PESQ, and STOI of MLDM values with different hyperparameter settings on the SNTCD-TIMIT test set, respectively. The hyperparameters α and β are constrained to satisfy $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1 - \alpha$. We systematically vary the values of α and β with a step size of 0.2, resulting in a total of 21 experimental configurations.

An important observation is the varying complementarity among the three components of the MLDM. As the weights α and β increase, both PESQ and STOI consistently exhibit an upward trend. The highest PESQ score is achieved at $\alpha = 0.6$ and $\beta = 0.4$, while the highest STOI score is obtained at $\alpha = 0.4$ and $\beta = 0.6$, closely aligning with the trends observed in the correlation analysis. Specifically, MSE-M demonstrates a higher correlation with PESQ, whereas SISNR shows a stronger correlation with STOI. Conversely, as the weight of CEGM increases, the perceptual quality degrades. We hypothesize that the deep feature extraction process for calculating CEGM is irreversible. Reducing the distortion of high-level audio-visual representations does not necessarily imply reducing the distortion of low-level acoustic features.

The lowest PER is achieved when $\alpha = 0.2$ and $\beta = 0.4$, leading us to conjecture that CEGM primarily reduces distortion in high-level audio-visual representations. Additionally, MSE-M and SISNR focus on minimizing distortion in low-level acoustic features, such as spectrum and waveform. This combination helps alleviate the mismatch between auditory and visual inputs to the AVSR backend, ultimately reducing the distortion of high-level audio-visual representations. Therefore, a complementary relationship exists among MSE-M, SISNR, and CEGM concerning PER. These insightful findings shed light on the intricate relationships among the MLDM components and their impact on the overall AVSE performance. Understanding the individual MSE-M, SISNR, and CEGM contributions in shaping the system's effectiveness provides valuable insights into optimizing hyperparameters α and β to achieve top audio-visual speech enhancement results.

3) Optimization Differences Between Audio-Only and Audio-Visual Scenarios: To compare the optimization performance of MLDM between audio-only and audio-visual scenarios, we visualized the learning curves of the MLDM and its three components (MSE-M, SISNR, and CEGM) on the development set, as illustrated in Fig. 5(a), (b), (c), and (d), respectively. Remarkably, the inclusion of visual modalities consistently results in lower MSE-M, higher SISNR, and lower CEGM values across all epochs, leading to lower MLDM values throughout the training process. The optimization process and the final results, as shown

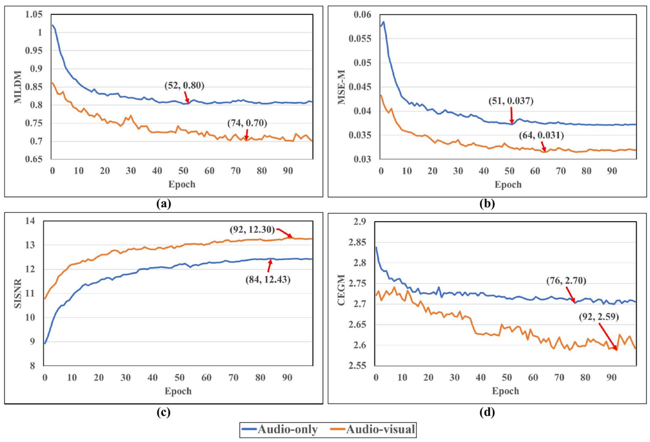

Fig. 5. Learning curve comparisons in audio-only and audio-visual cases for models optimized by (a) MLDM, (b) MSE-M, (c) SISIR, and (d) CEGM. Red arrows denote convergence points.

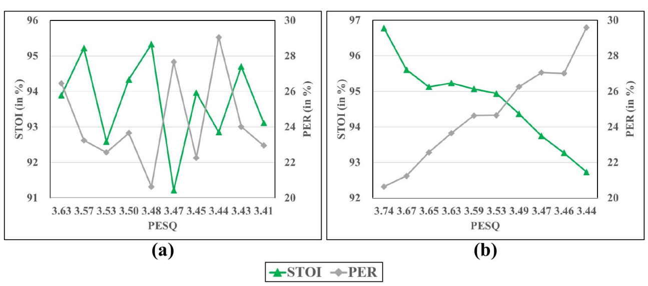

Fig. 6. A comparison of the average PER, PESQ and STOI among the top 10 out of 100 PESQ groups on the SNTCD-TIMIT test set enhanced by the (a) MLDM-optimized and (b) C-MLDM-optimized AVSE models. PESQ scores in the x-axis range from high to low, right to left.

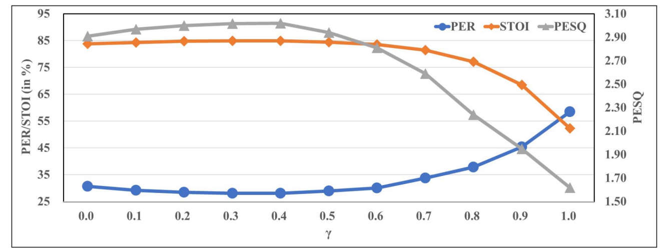

Fig. 7. Average PER, PESQ and STOI comparisons of C-MLDMs with various γ values on the SNTCD-TIMIT test set in the audio-visual scenario.

in Table I, strongly support the superiority of MLDM over all other evaluation metrics.

Upon further analysis of the learning curves, it becomes evident that there is a discrepancy in the convergence speed of MSE-M, SISNR, and CEGM. Specifically, MSE-M reaches its lowest point at the 64-th epoch in the audio-visual scenarios, while SISNR and CEGM converge after 92 epochs. This discrepancy in convergence speech may potentially impact the final MLDM model performance.

D. Performance Analysis of C-MLDM

1) Comparisons of MLDM and C-MLDM Results: First, we explore the impact of the discrepancy in the convergence speed of MSE-M, SISNR, and CEGM on performance. We first divide all samples in the test set into 100 groups based on their PESQ scores, ranging from high to low. We then select the top 10 PESQ groups to calculate their average PER and STOI scores within each group. Fig. 6(a) illustrates a comparison of the average PER and STOI among the top 10 groups on the SNTCD-TIMIT test set for the MLDM-optimized AVSE model. Interestingly, as PESQ declines in the x-axis, the changing trend of STOI and PER becomes chaotic. Consequently, we propose C-MLDM for explicitly enforcing the correlation among MSE-M, SISNR and CEGM and display the changing trends of PER and STOI scores for the C-MLDM-optimized model in Fig. 6(b). Both STOI and PER trends are negatively correlated with PESQ. As the PESQ score decreases, STOI decreases, while PER increases.

To assess the effectiveness of our proposed C-MLDM-optimized model, shown in the result row below MLDM in Table I, we discuss the average PER, PESQ, and STOI values on the SNTCD-TIMIT test set, covering both audio-only and audio-visual scenarios. In the audio-visual setting, C-MLDM consistently outperforms MLDM across all evaluation metrics. Notably, C-MLDM achieves a good average PER reduction of 2.65%, with an improved average PESQ score of 0.11 and an increased average STOI of 1.14% compared to MLDM. These improvements are also observed in the audio-only scenario, with C-MLDM exhibiting an average PER reduction of 4.19%, an improved average PESQ score of 0.09, and an increased average STOI of 1.07% over MLDM.

2) An Ablation Study on Hyperparameter Setting: Next, we study the impact of γ in (13) on the AVSE performance. Fig. 7 illustrates the average PER, PESQ, and STOI of C-MLDM on the SNTCD-TIMIT test set. γ varies from 0 to 1 in increments of 0.1 to obtain 11 sets of results.

Our analysis reveals that as the hyperparameter γ increases, the enhanced speech shows a mixed trend in the three evaluation metrics. Specifically, the SE performance improves initially and then deteriorates with increasing γ. Notably, when γ reaches 1, indicating that only CM is used for optimization, the enhanced speech performs worse than unprocessed speech across PER, PESQ, and STOI. This intriguing finding suggests that while incorporating the correlation-based objective can initially improve AVSE performance, an excessive emphasis on this objective might lead to suboptimal results because CM does not provide constraints on the rise or fall of the optimization objectives. We select $\gamma = 0.4$ in our proposed algorithm based on these observations.

3) Perceptual Analysis: In addition to PESQ and STOI, we evaluated the subjective quality of the enhanced speech through a carefully designed psychophysical experiment. In this experiment, 10 subjects with normal hearing were asked to rate the auditory quality of the enhanced speech sounds. Due to the inherent limitations of human psychophysical experiments, 25 samples were randomly selected from the SNTCD TIMIT test set to ensure a distribution of 5 samples in each SNR level. Participants were instructed to rate the quality of the noisy utterances

TABLE II
COMPARISONS OF MEAN OPINION SCORE (MOS ↑) AMONG NOISY AND MEASE MODELS OPTIMIZED BY MSE-M, SISNR AND CEGM, MLDM AND C-MLDM ON THE 25 SELECTED UTTERANCES FROM THE SNTCD-TIMIT TEST SET

Method	SNR (in dB)					Avg.
	−5	0	5	10	15	
Noisy	1.64	2.40	3.17	3.31	3.56	2.82
MSE-M	2.43	3.29	3.76	3.80	4.10	3.48
SISNR	2.37	3.24	3.76	3.85	4.12	3.47
MLDM	2.52	**3.53**	3.64	3.76	4.10	3.51
C-MLDM	**2.54**	3.49	**3.89**	**3.95**	**4.35**	**3.64**

along with those processed using different MEASE models, including MSE-M, SISNR, CEGM, MLDM, and C-MLDM. The evaluation was performed on a scale from 1 (indicating "poor") to 5 ("excellent"), with a pristine utterance first provided as a benchmark to represent the maximum achievable score, i.e., a score of 5. Subsequently, the samples processed by MSE-M, SISNR, CEGM, MLDM, and C-MLDM were presented to the participants in a randomized order. The mean opinion score (MOS) for each of the 25 utterances was calculated by averaging the ratings provided by the 10 subjects.

A comprehensive comparison of the MOS between the Noisy and MEASE models optimized by MSE-M, SISNR, CEGM, MLDM, and C-MLDM on the 25 carefully selected samples is systematically described in Table II. It is evident from the analysis that the MLDM model significantly outperforms the MSE-M and SISNR models in subjective quality, manifesting absolute enhancements of 0.03 and 0.04, respectively. Moreover, the MOS for the C-MLDM model is markedly higher than that for the MLDM model, with this superiority manifesting consistently across most SNR levels. This trend aligns with the comparative outcomes observed for the PESQ and STOI, further substantiating the efficacy of the MLDM and C-MLDM models in enhancing speech quality.

And in Fig. 8, we also present an illustrative comparison of the results of the SE models optimized by MSE-M, SISNR, and C-MLDM in both audio-only and audio-visual scenarios. An example utterance was randomly selected from the SNTCD-TIMIT test set, and all spectral features were subjected to utterance-level mean normalization. Notably, the MSE-M-enhanced speech in Fig. 8(c) and (d) shows a lack of detail in the non-silence segment, while the SISNR-enhanced speech in Fig. 8(e) and (f) retains broadband noise in the silence segments at the beginning and end. Conversely, the C-MLDM-enhanced speech in Fig. 8(g) and (h) not only preserves finer details but also significantly reduces high-frequency noise at the ends of the utterance, resulting in a spectral structure very similar to that of the clean speech in Fig. 8(a).

A consistent pattern emerges when comparing audio-only and audio-visual scenarios across different optimization targets. Significant changes in lip movements are evident within the non-silence segments, highlighting the role of visual acoustic

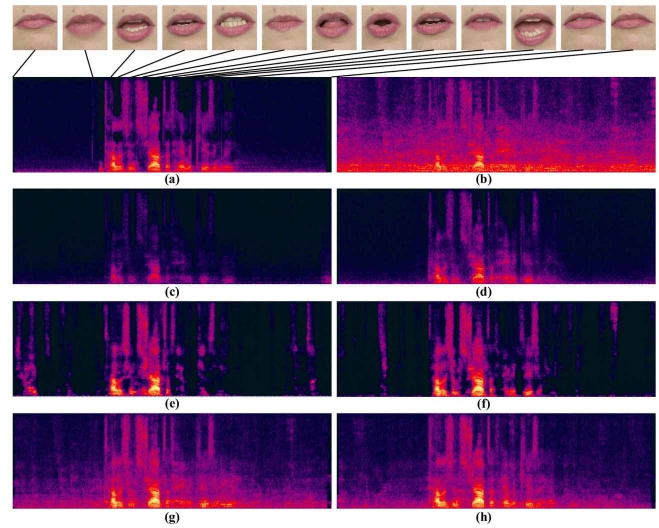

Fig. 8. An utterance example comparing the outputs of different optimization objects, including (a) clean spectrum features; (b) noisy spectrum features; MSE-M-enhanced spectrum features in (c) audio-only and (d) audio-visual scenarios; SISNR-enhanced spectrum features in (e) audio-only and (f) audio-visual scenarios; and C-MLDM-enhanced spectrum features in (g) audio-only and (h) audio-visual scenarios.

information in enriching articulation. Conversely, during silence intervals, the lips remain closed, highlighting the ability of the visual modality to provide distinctive cues that reduce residual noise and improve overall speech quality.

E. Generalizability of MLDM and C-MLDM

1) Dataset Diversity and Performance Impact: In a comprehensive effort to assess the generalizability of the proposed MLDM and C-MLDM models across different datasets, we extended our evaluation scope to include the prestigious Oxford-BBC Lip Reading Sentences 2 (LRS2) benchmark [79]. The LRS2 dataset, taken from BBC broadcasts, contains 144,482 video clips. It is systematically organized into pre-training, training, validation, and test sets, with allocations of 96,318 (195 hours), 45,839 (28 hours), 1,082 (0.6 hours), and 1,243 (0.5 hours) video clips, respectively. For this experiment, the pre-training and training segments were combined to formulate a comprehensive training dataset. Following the established simulation protocol applied to the SNTCD-TIMIT dataset, this resulted in a training set of approximately 1115 hours, supported by a validation set of 9 hours and a test set of 7.5 hours. This specially constructed, noisy version of the dataset was, therefore, named the SN-LRS2 dataset.

Following the training process and the best configure above, we retrained the ASR/AVSR backends and all SE models utilizing the SN-LRS2 dataset. Table III systematically presents a comparison of average WER, PESQ, and STOI among Noisy and MEASE models optimized by MSE-M, SISNR, and CEGM, MLDM, and C-MLDM on the SN-LRS2 test set in both audio-only and audio-visual scenarios.

TABLE III
COMPARISONS OF AVERAGE WER, PESQ AND STOI AMONG NOISY AND MEASE MODELS OPTIMIZED BY MSE-M, SISNR AND CEGM, MLDM AND C-MLDM ON THE SN-LRS2 TEST SET IN AUDIO-ONLY AND AUDIO-VISUAL SCENARIOS

Metric	WER (in %) ↓		PESQ ↑		STOI (in %) ↑	
Scenario	A	AV	A	AV	A	AV
Noisy	34.90	21.85	2.13		77.34	
MSE-M	34.77	20.44	2.45	2.67	81.18	85.94
SISNR	34.41	20.16	2.38	2.58	82.15	86.73
CEGM	32.94	19.05	\		\	
MLDM	30.92	18.06	2.50	2.74	83.20	87.45
C-MLDM	28.12	16.97	2.62	2.82	84.11	88.72

A: audio-only, AV: audio-visual.

TABLE IV
COMPARISONS OF AVERAGE PER, PESQ AND STOI AMONG NOISY AND CONV-FAVSNET MODELS OPTIMIZED BY MSE-M, SISNR AND CEGM, MLDM AND C-MLDM ON THE SNTCD-TIMIT TEST SET IN AUDIO-ONLY AND AUDIO-VISUAL SCENARIOS

Metric	PER (in %) ↓		PESQ ↑		STOI (in %) ↑	
Scenario	A	AV	A	AV	A	AV
Noisy	50.23	36.20	2.27		74.23	
MSE-M	49.82	32.85	2.70	2.95	79.44	84.78
SISNR	49.70	32.72	2.66	2.86	80.55	85.08
CEGM	47.01	32.00	\		\	
MLDM	43.80	30.28	2.77	3.01	81.29	85.73
C-MLDM	40.15	27.71	2.88	3.12	82.50	87.19

A: audio-only, AV: audio-visual.

The results are in concordance with the outcomes derived from our SNTCD-TIMIT experiments, wherein MLDM consistently outperforms the three baseline objectives, with the C-MLDM model achieving even more significant improvements across all metrics. Notably, within the audio-visual evaluation framework, MLDM realizes an average WER reduction of 0.99%, alongside gains of 0.07 in PESQ and improvements of 0.72% in STOI compared to the best baseline. The C-MLDM model further elevates these metrics, manifesting additional WER reductions of 1.09%, augmented PESQ improvements of 0.08, and average STOI enhancements of 1.27%. This consistent trend of superiority is replicated in the audio-only scenario, reinforcing the outstanding effectiveness and adaptability of MLDM and C-MLDM across diverse evaluation benchmarks.

2) Model Architecture Sensitivity Analysis: We further evaluated the robustness of our two advocated optimization goals, MLDM and C-MLDM, across various model architectures using the SNTCD-TIMIT dataset. Specifically, we instantiated a classic Conv-FavsNet [41], conceptually rooted in the Conv-TasNet framework described in [40]. The architecture of Conv-TasNet is built around three core elements: 1D convolution and deconvolution to encode audio waveforms and decode masked coded sequences, while a stack of 3×8 temporally dilated convolutional blocks tasked with estimating masks for isolating the target speech. As an extension of the Conv-TasNet model, Conv-FavsNet incorporates a pre-trained video encoder similar in structure to the MEE above but differs in its training target, which focuses on the classification of phonemes.

Regarding training, we applied the optimal hyperparameters outlined in [41] and ensured that the training process was consistent with that detailed in Section V-A. A comparative analysis of the average PER (in %), PESQ and STOI (in %) among the Noisy and Conv-FavsNet models, optimized by MSE-M, SISNR, CEGM, MLDM, and C-MLDM, on the SNTCD-TIMIT test set, is systematically presented in Table IV.

As evident from the tabulated results, MLDM and C-MLDM demonstrate superior performance over the three baseline objects, with C-MLDM consistently showing more excellent benefits across all metrics. Within the context of audiovisual evaluations, MLDM secures notable improvements, achieving reductions in PER of 1.72%, improvements in PESQ of 0.06, and increases in STOI of 0.65%, when compared to the baseline results. The C-MLDM model extends these gains, with further reductions in PER of 2.57%, additional improvements in PESQ of 0.11, and increases in STOI of 1.46%. This consistent pattern of performance improvement is also evident in the audio-only evaluations, underscoring the exceptional robustness of both MLDM and C-MLDM to different model architectures.

3) Cross-Linguistic Robustness Evaluation: We further assess and confirm the generalizability of the proposed MLDM and C-MLDM against the cross-linguistic scenario by evaluating on an extensive in-house audio-visual Mandarin corpus called SN-Mandarin. The SN-Mandarin corpus consists of 7,081 videos recorded by various speakers in everyday environments using mobile phones. For training, we randomly select 6,900 utterances, while 85 utterances were used for validation and an additional 96 utterances for testing. Real noise data from bathrooms, kitchens, balconies, and living rooms are adopted to create noisy-clean pairs. This results in approximately 81 hours of training data, 3 hours for validation, and 3 hours for testing. Importantly, there is no overlap in terms of speakers or noise recording rooms among the training, validation, and test subsets. Five SNR levels, 15, 10, 5, 0 and -5 dBs, are used to evaluate the performances of the models. A notable aspect to emphasize is that we have made the SN-Mandarin corpus publicly accessible[1]. to ensure transparency and reproducibility.

We employ high-performance ASR (HpASR) and high-performance AVSR (HpAVSR) models for training and evaluating recognition performances using the character error rate (CER). HpAVSR is a hybrid DNN-HMM AVSR model consisting of a deeper audio-visual acoustic model, a 4-gram word-based language model, and an extensive pronunciation dictionary containing over 600,000 Chinese words. The audio-visual acoustic model consists of 20 ResBlocks [80] and a visual encoder, followed by a 12-layer transformer. The visual encoder includes a deep spatiotemporal convolution, ResNet18 (identity mapping version [81]), and a 6-layer transformer. On the other hand, HpASR consists of an acoustic model with 20 ResBlocks, a 12-layer transformer and the same language model as the

[1][Online]. Available: https://github.com/coalboss/CMLDM/data

TABLE V
COMPARISONS OF AVERAGE CER, PESQ AND STOI AMONG NOISY AND MEASE MODELS OPTIMIZED BY MSE-M, SISNR AND CEGM, MLDM AND C-MLDM ON THE SN-MANDARIN TEST SET IN AUDIO-ONLY AND AUDIO-VISUAL SCENARIOS

Metric	**CER** (in %) ↓		**PESQ** ↑		**STOI** (in %) ↑	
Scenario	**A**	**AV**	**A**	**AV**	**A**	**AV**
Noisy	13.05	10.06	2.41		74.88	
MSE-M	12.78	8.47	2.56	2.78	78.02	81.61
SISNR	12.82	8.36	2.53	2.74	79.42	81.96
CEGM	15.00	13.03	\		\	
MLDM	10.65	7.69	2.59	2.81	79.71	82.55
C-MLDM	9.75	6.87	2.65	2.89	80.37	83.36

A: audio-only, AV: audio-visual.

HpAVSR model. For training, we first train the acoustic model with over 100,000 hours of Mandarin audio data collected in real-world conditions. Then, we fine-tune the audio-visual acoustic model using approximately 5,000 hours of Mandarin audio-visual data. The language model is trained with over 500 million sentences. The extensive coverage of diverse acoustic environments in the training data significantly improves the noise robustness of HpASR and HpAVSR. However, due to the achieved robustness through the extensive training data, further improvements in the recognition performances of enhanced speech become challenging.

Table V lists a comparison of average CER (in $\%$), PESQ, and STOI (in $\%$) among noisy, MSE-M, SISNR and CEGM, MLDM and C-MLDM on the SN-Mandarin test set in both audio-only and audio-visual scenarios. Remarkably, MLDM consistently outperforms the three baseline objectives, and C-MLDM consistently exceeds MLDM. Specifically, in the audio-visual scenario, MLDM achieves an average CER reductions of 1.49%, PESQ gains of 0.03, and STOI gains of 0.59% compared to the three baseline objectives. C-MLDM further improves over MLDM, achieving additional CER reductions of 0.82%, higher PESQ gains of 0.08, and average STOI gains of 0.81%. The same trend is observed in the audio-only scenario, reaffirming the remarkable effectiveness and generalizability of MLDM and C-MLDM in both evaluation scenarios. Interestingly, CEGM leads to an average CER increase of 2.97% when compared to noisy across all SNR levels. Similarly, an average CER increase of 1.96% in the audio-only scenario is observed. We hypothesize that the deep architecture of the high-performance backend model makes the gradients prone to vary freely, resulting in instability in AVSE model training.

VI. CONCLUSION

In this study, we develop effective optimization objectives, MLDM and C-MLDM, for AVSE that simultaneously improve speech quality, intelligibility and recognition performance. A comprehensive correlation analysis shows a complementarity among the MSE-M, SISNR and CEGM objectives. Accordingly, MLDM iteratively combines MSE-M, SISNR, and CEGM to match evaluation metrics from multiple tasks. C-MLDM further enhances their interactions by adding an additional correlation measure based on the Pearson correlation coefficient on top of MLDM. Experimental results demonstrated MLDM's superior performance over the three individual objectives in both audio-visual and audio-only scenarios. Moreover, C-MLDM consistently outperforms MLDM, highlighting the effectiveness of the additional correlation measures. Integrating the visual modality also amplifies the benefits of MSE-M, SISNR, and CEGM, enhancing their complementarity. These observations support our proposed MLDM and C-MLDM, which effectively improve the performance of the SE models across all evaluation metrics.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.
[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[6] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[9] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5210–5214.
[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 196–200.
[11] H. Erdogan et al., "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.
[12] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6493–6497.
[13] P. Manocha et al., "A differentiable perceptual audio metric learned from just noticeable differences," in *Proc. Interspeech*, 2020, pp. 2852–2856.
[14] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1780–1792, Oct. 2018.
[15] S.-J. Chen et al., "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Proc. Interspeech*, 2018, pp. 1571–1575.
[16] T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6660–6664.
[17] Y.-H. Tu et al., "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," in *Proc. Interspeech*, 2017, pp. 394–398.
[18] S.-W. Fu et al., "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2031–2041.

[19] Y.-L. Shen, C-Y. Huang, S-S. Wang, Y. Tsao, H-M. Wang, and T-S. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6750–6754.
[20] L. Chai, J. Du, Q-F. Liu, and C-H. Lee, "A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 106–117, 2021.
[21] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Process. Lett.*, vol. 27, pp. 26–30, 2020.
[22] R. Sawata, Y. Kashiwagi, and S. Takahashi, "Improving character error rate is not equal to having clean speech: Speech enhancement for ASR systems with black-box acoustic models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 991–995.
[23] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[24] L. E. Bernstein and C. Benoit, "For speech perception by humans or machines, three senses are better than one," in *Proc. IEEE 4th Int. Conf. Spoken Lang. Process.*, 1996, pp. 1477–1480.
[25] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," *Curr. Directions Psychol. Sci.*, vol. 17, pp. 405–409, 2008.
[26] D. W. Massaro and J. A. Simpson, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. London, U.K.: Psychol. Press, 2014.
[27] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1995, pp. 1559–1562.
[28] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Amer.*, vol. 109, no. 6, pp. 3007–3020, 2001.
[29] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3726–3730.
[30] J. W. Fisher III et al., "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 772–778.
[31] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization)," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, 2002, pp. 68–71.
[32] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2002, pp. 2025–2028.
[33] J. Hershey and M. Casey, "Audio-visual sound separation via hidden Markov models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 1173–1180.
[34] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 3051–3055.
[35] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 1170–1174.
[36] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[37] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.
[38] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.
[39] E. Ideli, B. Sharpe, I. V. Bajić, and R. G. Vaughan, "Visually assisted time-domain speech enhancement," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2019, pp. 1–5.
[40] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[41] J. Wu et al., "Time domain audio visual speech separation," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2019, pp. 667–673.
[42] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.
[43] H. Chen et al., "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Netw.*, vol. 143, pp. 171–182, 2021.
[44] W. P. Wang et al., "A robust audio-visual speech enhancement model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7529–7533.
[45] R. Lu, Z. Duan, and C. Zhang, "Audio-visual deep clustering for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1697–1712, Nov. 2019.
[46] D.-L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Boston, MA, USA: Springer, 2005, pp. 181–197.
[47] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.
[48] J. Xiong, Y. Zhou, P. Zhang, L. Xie, W. Huang, and Y. Zha, "Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement," *IEEE Trans. Multimedia*, vol. 25, pp. 5800–5812, 2023.
[49] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–53.
[50] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.
[51] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.
[52] C. Huang et al., "Addressing the loss-metric mismatch with adaptive loss alignment," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2891–2900.
[53] S.-W. Fu et al., "MetricGAN: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.
[54] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.
[55] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.
[56] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[57] C. K. A. Reddy et al., "A scalable noisy speech dataset and online subjective test framework," in *Proc. Interspeech*, 2019, pp. 1816–1820.
[58] H.-J. Hsieh, B. Chen, and J.-W. Hung, "Employing median filtering to enhance the complex-valued acoustic spectrograms in modulation domain for noise-robust speech recognition," in *Proc. IEEE Int. Symp. Chin. Spoken Lang. Process.*, 2016, pp. 1–5.
[59] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on Mel-frequency Cepstra for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* 2008, pp. 4041–4044.
[60] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7398–7402.
[61] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
[62] H. Zhang, X. Zhang, and G.-L. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5374–5378.
[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[64] T.-A. Hsieh et al., "Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement," in *Proc. Interspeech*, 2021, pp. 196–200.
[65] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra Appl.*, vol. 48, pp. 257–263, 1982.
[66] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[67] M. Yang et al., "PAAPLoss: A phonetic-aligned acoustic parameter loss for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[68] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.
[69] Y.-Y. Zeng et al., "TapLoss: A temporal acoustic parameter loss for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[70] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, pp. 603–615, 2015.
[71] H. Chen et al., "Audio-visual speech recognition in MISP2021 challenge: Dataset release and deep analysis," in *Proc. Interspeech*, 2022, pp. 1766–1770.
[72] P. Y. Chen, M. Smithson, and P. M. Popovich, *Correlation: Parametric and Nonparametric Measures*. Newbury Park, CA, USA: Sage, 2002 .
[73] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006.
[74] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[75] T. H. Falk et al., "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.
[76] J. Kreimer and R. Y. Rubinstein, "Nondifferentiable optimization via smooth approximation: General analytical approach," *Ann. Operations Res.*, vol. 39, no. 1, pp. 97–119, 1992.
[77] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[79] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6447–6456.
[80] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[81] K. M. He, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

Hang Chen (Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2018 and 2024, respectively. He is currently a Postdoctoral Researcher with USTC. His research focuses on audio-visual speech enhancement and recognition.

Qing Wang received the B.S. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2012 and 2018, respectively. From 2018 to 2020, she was with Tencent working on single-channel speech enhancement. She is currently a Postdoctor with USTC. Her research interests include speech enhancement, robust speech recognition, audio-visual scene classification, and sound event localization and detection.

Jun Du (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was a Team Leader with iFLYTEK Research working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an associate Researcher, working on handwriting recognition and OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an Associate Editor for IEEE/ACM TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING and a Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team won several champions of the CHiME-4/CHiME-5/CHiME-6 Challenge, SELD Task of 2020 DCASE Challenge, and DIHARD-III Challenge.

Bao-Cai Yin received the B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2012 and 2022, respectively. He is currently a Researcher with iFLYTEK, Hefei. In 2015, he joined iFLYTEK and has been engaged in artificial intelligence research. His research interests include optical character recognition, medical image analysis, and multimodality learning. He won first place in many computer vision challenges, including Lung Nodule Analysis, IDRiD Diabetic Retinopathy Segmentation, and Grading Challenge.

Jia Pan received the B.S., M.S., and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2006, 2009 and 2021, respectively. Since 2009, he has been with iFLYTEK Research, Hefei, working on speech recognition and spoken dialog systems. His research interests include speech recognition and machine learning. He won first place in many speech challenges, including Open Automatic Speech Recognition (OpenASR) Challenge 2021 and Blizzard Challenge.

Chin-Hui Lee (Life Fellow, IEEE) is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he has 20 years of industrial experience, ending with Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff and the Director of the Dialog Systems Research Department. He has authored or coauthored more than 600 papers and 30 patents and has been cited more than 80000 times for his original contributions, with an H-index of 80 on Google Scholar. He was the recipient of numerous awards, including the Bell Labs President's Gold Award in 1998, and SPS's 2006 Technical Achievement Award for Exceptional Contributions to the Field of Automatic Speech Recognition. In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was also the recipient of the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition. He is also a Fellow of ISCA.