INCORPORATING VISUAL INFORMATION RECONSTRUCTION INTO PROGRESSIVE LEARNING FOR OPTIMIZING AUDIO-VISUAL SPEECH ENHANCEMENT

Chen-Yue Zhang[1], Hang Chen[1], Jun Du[1], Bao-Cai Yin[2], Jia Pan[2], Chin-Hui Lee[3]*

[1] University of Science and Technology of China, China [2] iFlytek Research, iFlytek Co., Ltd., China
[3] Georgia Institute of Technology, USA
jundu@ustc.edu.cn

ABSTRACT

Video information has been widely introduced to speech enhancement as its contribution at low signal-to-noise ratios (SNRs). Conventional audio-visual speech enhancement networks take noisy speech and video as input and learn features of clean speech directly. To reduce the large SNR gap between the learning target and input noisy speech, we propose a novel mask-based audio-visual progressive learning speech enhancement (AVPL) framework with visual information reconstruction (VIR) to increase SNRs gradually. Each stage of AVPL takes a concatenation of pre-trained visual embedding and the previous representation as input and predicts a mask with the intermediate representation of the current stage. To extract more visual information and deal with the performance distortion, the AVPL-VIR model reconstructs the visual embedding as it is fed in for each stage. Experiment on the TCD-TIMIT dataset shows that the progressive learning method significantly outperforms direct learning for both audio-only and audio-visual models. Moreover, by reconstructing video information, the VIR module provides a more accurate and comprehensive representation of the data, which in turn improves the performance of both AVDL and AVPL.

Index Terms— Speech enhancement, progressive learning, visual embedding reconstruction

1. INTRODUCTION

Speech in noisy environments such as shopping malls, factories, streets, etc., can be severely disturbed or even drowned in noise. The aim of speech enhancement [1] is to produce enhanced speech with better quality and intelligibility by suppressing the background noise. It is commonly used as a front-end task for automatic speech recognition (ASR), hearing aids, and communications.

A number of methods have been proposed for speech enhancement. Conventional speech enhancement approaches such as spectral subtraction [2], Wiener filtering [3], and minimum mean squared error (MMSE) estimation [4] have been extensively studied. To simplify the model, these conventional algorithms are based on a series of mathematical assumptions about the noisy and clean speech. However, these conventional methods are often challenging to track non-stationary noises for real-world acoustic conditions.

In the last few years, deep neural network (DNN)-based speech enhancement algorithms have received increasing attention. DNN-based methods have been found to have a more powerful capability to model complex relationships between noisy and clean speech, in comparison to conventional algorithms [5]. According to the learning target, these DNN-based algorithms can be broadly classified into mask-based methods [6, 7] and feature mapping-based methods [8]. Recently, more and more researchers directly utilize DNN to model the relationship between noisy and clean waveforms [9, 10, 11]. Moreover, the learning process of deep structures in speech enhancement has attracted many researchers. [12] proposes a speech enhancement network based on progressive learning that decomposes direct mapping into multiple stages and gradually improves the SNR. Nian et al. [13] propose a time domain progressive learning network and prove that both the SNR-increased intermediate target and clean target can achieve better listening quality and intelligibility, as they significantly improve the performance of ASR.

Previous researches [14, 15, 16] show that visual information is beneficial to the speech perception of the corresponding speaker, especially at low SNRs, since visual information is often immune to the effects of acoustic conditions. [17] proposes an audio-visual speech enhancement network that is able to separate a speaker's voice-given lip regions in the corresponding video, by predicting both the magnitude and the phase of the target signal. Hou et al. [18] reconstruct the input mouth images for speech enhancement in order to make full use of visual information. [19] employs the phone as a classification target to extract a visual embedding with more useful information for speech enhancement. [20] proposes a new mechanism for audio-visual fusion adaptable to any feature layer of the audio and visual networks. Chen et al. [21] adopt a multi-modal embedding by fusing audio and visual embedding, which is extracted from pre-trained articulation place classification networks, to avoid performance distortion at high SNRs [22] and realize a satisfactory performance.

In this study, we propose a novel audio-visual progressive learning (AVPL) framework with visual information reconstruction (VIR) for speech enhancement. Specifically, the mapping between noisy and clean speech is split into multiple stages. In each stage, the intermediate representation outputted by the previous stage is concatenated with the pre-trained visual embedding to improve the SNR gradually. Moreover, the intermediate representation outputted by the current stage is also used to reconstruct the visual embedding. Experimental results show the proposed framework achieves significant performance improvements for both PESQ and STOI metrics in contrast to the baseline model. We also conduct a comprehensive ablation study and find that the progressive learning network retains more spectral details and suppresses more noise, especially in the non-speech segments.

The rest of the paper is organized as follows. Section 2 introduces our proposed AVPL-VIR model. Section 3 introduces the implementation details and experiments. Finally, we conclude our findings in Section 4.

*corresponding author

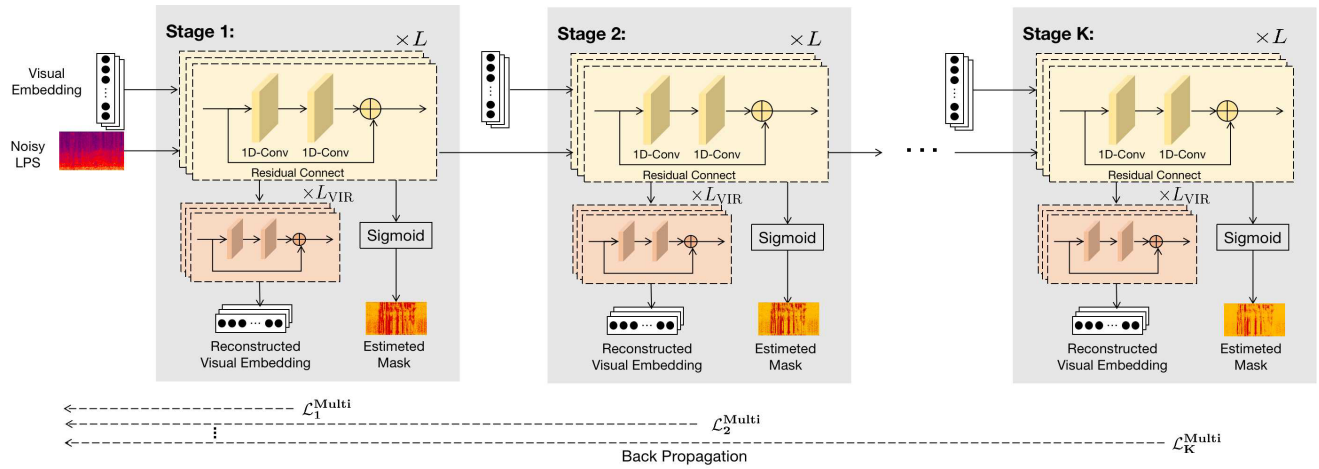


Fig. 1. The overview of proposed audio-visual progressive learning model with visual information reconstruction

2. PROPOSED AUDIO-VISUAL PROGRESSIVE LEARNING SPEECH ENHANCEMENT WITH VISUAL INFORMATION RECONSTRUCTION

This section introduces the AVPL and the VIR module in detail. We apply progressive learning in audio-visual speech enhancement, which improves SNRs gradually. Then the VIR is established to extract more visual information and prevent performance degradation with increasing convolution blocks. The overview of the proposed model is illustrated in Fig. 1.

2.1. Audio-Visual Progressive Learning Network

The proposed AVPL model takes a concatenation of noisy log power spectrum (LPS) A_{LPS} and visual embedding as input. We employ a visual embedding extractor in [21], which is pre-trained with a novel classification target, i.e. place of articulation. The lip frames pass through a 3D convolution and ResNet-18 successfully:

$$E_V = \text{ResNet-18}(\text{Conv}_{\text{3D}}(V)) \tag{1}$$

where E_V and V denote visual embedding and input lip frames respectively.

The progressive learning network divides the conventional direct learning method into multiple stages, and each of them achieves a small SNR improvement, which is easier to learn. Each stage is composed of L 1D-ConvBlocks, denoted as $\mathcal{F}_i^{\text{PL}}(\cdot)$, including a 1D convolution layer with a residual connection, a ReLU activation, and a batch normalization.

For the i-th stage in AVPL, the intermediate representation is obtained by a concatenation of visual embedding and the previous representation:

$$\begin{aligned} X_i^{\text{PL}} &= \mathcal{F}_i^{\text{PL}}([X_{i-1}^{\text{PL}}, E_V], \Lambda_i) \\ &= \mathcal{F}_i^{\text{PL}}(...[\mathcal{F}_1^{\text{PL}}([A_{\text{LPS}}, E_V], \Lambda_1), E_V]..., \Lambda_i) \end{aligned} \tag{2}$$

where Λ_i and X_i^{PL} denote the parameter set of weight matrices and bias vectors and the intermediate representation in the i-th stage, respectively. For the first stage, X_0^{PL} is equivalent to A_{LPS}.

Then a sigmoid activation is employed to obtain a mask M_i for the i-th stage:

$$M_i^{\text{PL}} = \sigma(X_i^{\text{PL}}) \tag{3}$$

where M_i^{PL} is range from 0 to 1. We expect to use the mask M_i^{PL} to obtain enhanced speech with SNR_i dB increasing in SNR to noisy input. Therefore the learning target for i-th stage [23] is denoted as M_{SNR_i}:

$$M_{\text{SNR}_i}(k,t) = \frac{|S(k,t)|^2 + |N_i(k,t)|^2}{|S(k,t)|^2 + |N(k,t)|^2} \tag{4}$$

where $S(k,t)$, $N_i(k,t)$ and $N(k,t)$ represent the short-time Fourier transform (STFT) of the clean speech, residual noise for i-th stage, and input noise at time frame t and frequency bin k, respectively. For the last learning stage, the learning target M_{SNR_K} is the same as IRM to estimate a clean speech.

To optimize the parameters, the mean square error (MSE) loss is calculated for every stage:

$$\mathcal{L}_{\text{mask}_i}^{\text{PL}} = \sum_{k,t} \left\| M_i^{\text{PL}}(k,t) - M_{\text{SNR}_i}(k,t) \right\|^2 \tag{5}$$

where $\mathcal{L}_{\text{mask}_i}^{\text{PL}}$ denotes the MSE loss for estimated mask in i-th stage. The gradient is updated as the dashed line in Fig. 1 indicated. Since multiple masks are estimated in the progressive learning network, a weighted multi-targets learning loss $\mathcal{L}_{\text{AVPL}}$ is designed as follows:

$$\mathcal{L}_{\text{AVPL}} = \frac{1}{K} \sum_{i=1}^{K} \alpha_i \mathcal{L}_{\text{mask}_i}^{\text{PL}} \tag{6}$$

where α_i is the weighting factor for i-th learning stage.

The corresponding audio-only direct learning (AODL) and audio-visual direct learning (AVDL) methods take N 1D-ConvBlocks in total the same as corresponding progressive learning, with only one target, i.e. IRM. The audio-only progressive learning (AOPL) takes the noisy LPS as input, which employs the same structure as AVPL.

2.2. AVPL with Visual Information Reconstruction

Visual information plays an essential role in improving the intelligibility of enhanced speech, especially at low SNRs. We expect more visual information can be extracted to intermediate representations. Then the VIR module is established and changes each progressive stage into a multi-target one. In addition to predicting a mask M_i^{PL}, the stage is also called for reconstructing the visual information.

The VIR module takes a similar structure as enhancement blocks, which consists of L_{VIR} 1D-ConvBlocks, denoted as $\mathcal{F}_i^{\text{VIR}}(\cdot)$. In stage i, the intermediate representation X_i^{PL} extracted by $\mathcal{F}_i^{\text{PL}}(\cdot)$ is used to reconstruct visual information, i.e. visual embedding in

this paper. The intermediate representation passes through the VIR module as follows:

$$E_i^{\mathrm{VIR}} = \mathcal{F}_i^{\mathrm{VIR}}(X_i^{\mathrm{PL}}) \tag{7}$$

where E_i^{VIR} denotes reconstructed visual embedding at i-th stage. The MSE loss for VIR $\mathcal{L}_i^{\mathrm{VIR}}$ is calculated as:

$$\mathcal{L}_i^{\mathrm{VIR}} = \sum \left\| E_i^{\mathrm{VIR}} - E_V \right\|^2 \tag{8}$$

Combined with the progressive learning loss $\mathcal{L}_i^{PL}$, the multi-targets loss $\mathcal{L}_i^{\mathrm{Multi}}$ can be defined as follows:

$$\mathcal{L}_i^{\mathrm{Multi}} = \eta_1 \mathcal{L}_{\mathrm{mask}_i}^{\mathrm{PL}} + \eta_2 \mathcal{L}_i^{\mathrm{VIR}} \tag{9}$$

where η_1 and η_2 are the weighting factors for the estimated mask and reconstructed visual embedding in the i-th stage. To minimize the multi-target loss, each stage is required to extract more visual information to intermediate presentation X_i^{PL}, which is especially important for performance at low SNRs. The MSE loss for AVPL-VIR is defined as follows:

$$\mathcal{L}_{\text{AVPL-VIR}} = \frac{1}{K} \sum_{i=1}^{K} \alpha_i \mathcal{L}_i^{\mathrm{Multi}} \tag{10}$$

where, the weighting factor α_i is the same as AVPL. In this way, we utilize visual information not only by feeding it to each stage but also by driving the network to make full use of it. AVPL-VIR improves the SNR of input speech and reconstructs the visual information in a progressive way. It achieves this by utilizing both audio and video modalities, allowing for more accurate and detailed visual information to be extracted from the input. Overall, AVPL-VIR provides a more effective and efficient method for processing and analyzing multimodal data.

3. EXPERIMENTS

3.1. Implementation Details

Clean speech derived from TCD-TIMIT audio-visual corpus [24] is corrupted at five SNR levels (-5dB, 0dB, 5dB, 10dB, 15dB) with 100 noise types in [25] and 15 homemade noise types to build a 35-hour training set. We present the test set with four SNR levels (-5dB, 0dB, 5dB, 10dB) which includes 3 other unseen noise types from NOISEX-92 corpus [26]: Destroyer Engine, Factory1, and Speech Babble.

For AODL and AOPL models, there are 20 1D-ConvBlocks in total, i.e. $N = 20$. The visual embedding reconstruction is composed of 5 1D-ConvBlocks, i.e. $L_{\mathrm{VIR}} = 5$. The weight parameters η_1 and η_2 in multi-targets loss are set to be 1 and 0.1 respectively. The loss weight parameter α_i of each learning stage is 1. We train models with an initial learning rate set as 0.001, and it will be halved if the loss does not decrease for 3 consecutive epochs on the validation set. We adopt perceptual evaluation of speech quality (PESQ) [27] and short-time objective intelligibility (STOI) [28] to evaluate enhanced speech, both of them have higher values to indicate better performance.

3.2. Analysis on Progressive Learning and Visual Information Reconstruction

Tu et al. [12] implement 3 LSTM layers with 3 progressive targets to achieve the best performance. However, convolution blocks in this study are more complex, how many progressive targets will yield the best enhancement performance is of great interest and worth exploring. Firstly, we perform a series of experiments using AOPL with K progressive targets, where $K \in \{3, 4, 5\}$. An average PESQ and STOI comparison between AODL and AOPLs with different progressive targets over 4 SNR levels and 3 unseen noise types is shown in Table 1.

Table 1. PESQ and STOI(%) comparison on AODL and AOPL with different numbers of targets. (The value of K and its corresponding intermediate targets are set as $K = 3$: +10dB, +20dB; $K = 4$: +5dB, +10dB, +15dB; $K = 5$: +5dB, +10dB, +15dB, +20dB)

Metrics	AODL	AOPL		
		$K = 3$	$K = 4$	$K = 5$
PESQ	2.49	2.63	**2.66**	2.64
STOI(%)	74.29	76.09	**76.84**	76.21

Our analysis demonstrates that all audio-only object-oriented progressive learning (AOPL) models outperform their audio-only counterparts (AODL), indicating the effectiveness of the progressive learning approach in the complex network structure. Specifically, AOPL achieves its highest performance, with gains of 0.17 PESQ and 2.55 STOI(%) over AODL, when trained with 4 targets: 3 intermediate targets (+5dB, +10dB, +15dB) and a clean target at the end.

We also observe that increasing the number of stages can lead to overfitting and performance degradation. To balance performance and computation cost, we select a default setting of $K = 4$ for all subsequent experiments.

To further investigate the impact of the number of 1D-ConvBlocks in each progressive stage on speech enhancement performance, we conduct additional experiments. Specifically, we compare the average PESQ and STOI performance of AVDL and AVPL models with L 1D-ConvBlocks in each progressive stage. We set L to 3, 4, or 5, and to ensure a fair comparison, we also included the results of AVDL models with $L \times K$ 1D-ConvBlocks. Our findings are presented in Table 2

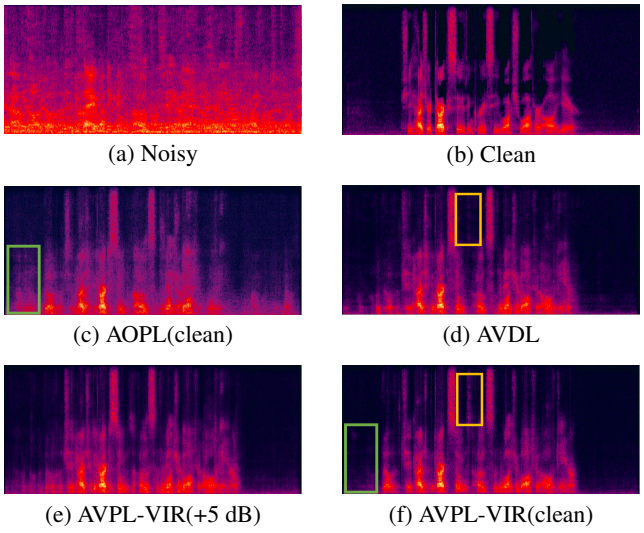
Our experiments on the impact of the number of 1D-ConvBlocks in each progressive stage show that AVPL models consistently outperform their AVDL counterparts with the same number of blocks, as indicated by the significant STOI gains of AVPL over AVDL for all SNRs. Specifically, the best performance for AVPL is achieved with 3 1D-ConvBlocks in each stage, while AVDL requires 4 1D-ConvBlocks. However, as the number of 1D-ConvBlocks increased, the STOI performance of both AVPL and AVDL declined, which could be attributed to visual information distortion. Importantly, our findings suggest that progressive learning has the potential to achieve high performance with fewer 1D-ConvBlocks, indicating its usefulness in model compression.

To confirm the effectiveness of the VIR, we conducted a comparative analysis of AVDL and AVPL models with and without VIR. When using the same number of 1D-ConvBlocks, AVDL-VIR consistently outperforms AVDL in terms of both PESQ and STOI for all SNR levels, suggesting that the VIR module effectively utilizes lip frames to improve the quality and intelligibility of the enhanced speech. In subsequent experiments with AVPL-VIR, we initially set $L = 3$ as the default, but observe performance degradation compared to AVPL. As we increase L to 5, AVPL-VIR yields notable improvements over the best AVPL system at different SNRs, suggesting that VIR helps models better leverage visual information with

Table 2. PESQ and STOI(%) comparison on AVDL, AVDL-VIR, AVPL, AVPL-VIR with different numbers of blocks in each stage and 4 targets (+5 dB, +10 dB, +15 dB, and clean) at several SNRs averaged over noise types. The *avg.* means the average score on several SNRs.

Metrics		PESQ					STOI(%)				
Model	L	-5	0	5	10	*avg.*	-5	0	5	10	*avg.*
Noisy	-	1.70	1.97	2.26	2.56	2.12	54.34	65.11	75.33	84.48	69.82
AVDL	3	2.25	2.54	2.82	3.07	2.67	66.81	75.92	83.22	89.19	78.79
	4	2.25	2.55	2.82	3.08	2.68	67.42	76.28	83.36	89.35	79.10
	5	2.25	2.55	2.81	3.07	2.67	66.95	76.00	83.35	89.31	78.90
AVDL-VIR	3	2.28	2.59	2.88	3.16	2.73	67.45	76.62	84.12	89.85	79.51
	4	2.27	2.58	2.88	3.15	2.72	67.94	77.25	84.85	90.06	80.03
	5	2.25	2.55	2.82	3.14	2.69	68.01	77.21	84.47	90.03	79.93
AVPL	3	2.31	2.63	2.93	3.21	2.77	68.62	77.65	84.77	90.14	80.30
	4	2.32	2.63	2.93	3.21	2.77	68.54	77.61	84.69	90.14	80.25
	5	2.33	2.64	2.93	3.21	2.78	68.53	77.58	84.60	90.06	80.19
AVPL-VIR	3	2.31	2.64	2.94	3.23	2.78	68.61	77.54	84.75	90.14	80.26
	5	2.34	2.65	2.95	3.22	2.79	69.28	77.91	84.87	90.13	80.55

deeper structures. The results of these experiments are presented in Table 2.



(a) Noisy (b) Clean

(c) AOPL(clean) (d) AVDL

(e) AVPL-VIR(+5 dB) (f) AVPL-VIR(clean)

Fig. 2. Spectrograms of a speech corrupted by babble noise at -5 dB SNR.(a)Noisy speech, (b) clean speech, (c) clean output of AOPL, (d) clean output of AVDL,(e) +5dB intermediate output of AVPL-VIR, (f) clean output of AVPL-VIR.

3.3. Comparison with others' methods

Table 3 compares the performance of our proposed models (the best AOPL, AVPL, and AVPL-VIR) with the state-of-the-art MEASE module [21] on the TCD-TIMIT dataset. Our proposed AVPL-VIR model outperforms AOPL by 3.71 STOI gains and 0.13 PESQ gains, indicating the essential role of visual information in speech enhancement. Compared to AVDL, AVPL-VIR shows improvements of 0.11 PESQ and 1.45 STOI(%), demonstrating the effectiveness of incorporating VIR into AVPL.

In summary, AVPL-VIR achieves even better performance than the MEASE module, leveraging the benefits of progressive learning and visual embedding reconstruction. By breaking down the learning process into smaller, more manageable steps, AVPL-VIR reduces the difficulty of each learning phase. The VIR module allows for more comprehensive usage of the visual information, resulting in superior performance. Overall, our results demonstrate the effectiveness of incorporating visual information and utilizing progressive learning for audio-visual speech enhancement.

Table 3. PESQ and STOI comparison of best AOPL, AVDL, and AVPL-VIR averaged on all SNR levels and noise types.

Model	Noisy	AOPL	AVDL	AVPL-VIR	MEASE [21]
PESQ	2.12	2.66	2.68	2.79	2.73
STOI(%)	69.82	76.84	79.10	80.55	80.29

In Fig. 2, an example of utterance from the test set corrupted by babble noise at -5 dB SNR is selected to intuitively compare the performance of AVDL, AOPL, and AVPL-VIR. The region of the green box in Fig. 2(c) and Fig. 2(f) illustrates that AVPL-VIR can do better in suppressing noise in the non-speech segments, as it can extract and utilize visual information adequately. The region of the yellow box in Fig. 2(d) and Fig. 2(f) indicates that progressive learning can reduce speech distortion at high frequency, remaining more details of clean speech. Combining progressive learning and VIR, we increase the SNR of noisy speech gradually with multiple intermediate targets and take advantage of visual information, which outperforms the baseline model.

4. CONCLUSION

In this paper, we propose an AVPL-VIR model, which splits the conventional mapping between noisy and clean speech into multiple stages to achieve SNR improvement gradually and establishes a visual information reconstruction module to make better use of visual information. The experiment shows that progressive learning has a significant improvement on PESQ and STOI for both audio-only and audio-visual speech enhancement. In addition, the VIR can help models make better use of visual information with deep structures, which brings further improvement to AVDL and AVPL.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[2] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Jae Lim and Alan Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[7] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5220–5224.

[8] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[9] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[10] Ashutosh Pandey and DeLiang Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 1270–1279, 2021.

[11] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[12] Yan-Hui Tu, Jun Du, Tian Gao, and Chin-Hui Lee, "A multi-target snr-progressive learning approach to regression based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1608–1619, 2020.

[13] Zhaoxu Nian, Jun Du, Yu Ting Yeung, and Renyu Wang, "A time domain progressive learning approach with snr constriction for single-channel speech enhancement and recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6277–6281.

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[15] W. H. Sumby et al., "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[16] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[18] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[19] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Time domain audio visual speech separation," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 667–673.

[20] Michael L Iuzzolino and Kazuhito Koishida, "Av (se) 2: Audio-visual squeeze-excite speech enhancement," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 7539–7543.

[21] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Bao-Cai Yin, and Chin-Hui Lee, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, vol. 143, pp. 171–182, 2021.

[22] Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun, "A robust audio-visual speech enhancement model," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 7529–7533.

[23] Nan Zhou, Jun Du, Yan-Hui Tu, Tian Gao, and Chin-Hui Lee, "A speech enhancement neural network architecture with snr-progressive multi-target learning for robust speech recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 873–877.

[24] Naomi Harte and Eoin Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[25] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[26] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001*. IEEE, 2001, vol. 2, pp. 749–752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.