A ROBUST AUDIO-VISUAL SPEECH ENHANCEMENT MODEL

Wupeng Wang[†] *Chao Xing*[†] *Dong Wang*[⋆] *Xiao Chen*[†] *Fengyu Sun*[††]

[†] Huawei Noah's Ark Lab.
[⋆] Center for Speech and Language Technology (CSLT), Tsinghua Unviersity
[††] Huawei Technologies CO. LTD

ABSTRACT

Most existing audio-visual speech enhancement (AVSE) methods work well in conditions with strong noise, however when applied to conditions with a medium SNR, serious performance degradations are often observed. These degradations can be partly attributed to the feature-fusion(early fusion etc.) architecture that tightly couples the audio information that is very strong and the visual information that is relatively weak. In this paper, we present a safe AVSE approach that can make the visual stream contribute to audio speech enhancment(ASE) safely in conditions of various SNRs by late fusion.The key novelty is two-fold: Firstly, we define power binary masks (PBMs) as a rough representation of speech signals. This rough representation admits the weakness of the visual information and so can be easily predicted from the visual stream. Secondly, we design a posterior augmentation architecture that integrate the visual-derived PBMs to the audio-derived masks via a gating network. By this architecture, the entire performance is lower-bounded by the audio-based component. Our experiments on the Grid dataset demonstrated that this new approach consistently outperforms the audio-based system in all noise conditions, confirming that it is a safe way to incorporate visual knowledge in speech enhancement.

Index Terms— speech enhancement, audio-visual, deep learning, multimodal

1. INTRODUCTION

Speech enhancement (SE) aims to remove background noise from noisy speech so that the intelligibility can be improved. Traditional algorithms include Wiener filtering [1], spectral subtraction [2], minimum mean squared error (MMSE) estimation [3]. These methods are easy to implement but not well generalizable, due to the strong assumptions they hold. Recently, deep learning based speech enhancement method gained much attention and showed significant improvement over traditional methods [4, 5, 6, 7]. The key component of these methods is a deep neural network (DNN) that learns a complicated mapping function from noisy speech to clean speech, in a way of *noisy training*. That is, construct a large mount of noisy speech by mixing noise of various types with clean speech, and the DNN is trained to recover the clean speech from the noisy speech.

In spite of the great success of deep speech enhancement, the audio-only approach is limited in some complex situations, e.g., when there are multiple speakers, the background noise is very strong especially when the noise is unknown to the audio-only methods. Visual information is valuable in these situations: it can provide knowledge of the target speaker with content information and aid to estimate the desirable time-frequency (TF) value. A number of researches have been conducted towards this audio-visual speech enhancement (AVSE). For instance, Almajai et al. [8] introduced Visually Derived Wiener Filters based on the correlation between visual signal and audio signal. Abdelaziz et al.[9] showed twin hidden Markov model to improve the performance of audio-only model when applied on low SNR. Gabbay et al. [10] proposed an encoder-decoder architecture where the encoder maps both audio (noisy speech) and visual signals (mouth images) into a shared embedding space, and the decoder produces clean speech based on the shared embedding. Ephrat et.al [11] introduced a Looking-to-Listen model, where embeddings of audio and visual signals are concatenated and are used to predict complex masks that can be used to derive speech of target speakers. Gogate [12] followed the same idea and presented an architecture called CochleaNet. Hou et.al [13] built a similar model, however the decoder reconstructs not only clean speech but also the mouth images. .

All aforementioned methods, extracting features from both visual and audio signal respectively and combining them together in fusion part, are called feature-fusion and achieved good performance in conditions with a low SNR, e.g., with multiple speakers [11] or a strong background noise [13]. However, when we tried to apply these methods to conditions with a medium SNR, serious performance degradation was observed (see in the experiments). In other words, most of existing deep AVSE methods are not consistently reliable in various real-world environments. We hypothesize that this degradation can be attributed to the weakness of the visual information. Although the correlation between visual and audio data does exist [14], the visual information is only related to the envelope of the spectrum (not source), and the mapping between visual and audio frames is not one-to-one, nor temporally aligned [15]. The weak correlation between visual and audio streams suggests that the visual stream can provide only weak information for the SE task. Unfortunately, most of existing models fuse the visual and audio streams at the feature level, which tightly couples the audio information which is essentially strong for SE and the visual information which is intrinsically weak. This tight coupling works well when audio information is unreliable (in the case of a low SNR), however when the audio information is strong (in the case of a medium and high SNR), involving the weak visual information tends to be hurtful.

In this paper, we present a late fusion architecture to address this problem. Firstly, we define power binary masks (PBMs) as a rough representation of speech signals. The roughness of this representation matches the weakness of the visual information, and so the PBMs can be predicted from the the visual stream. Secondly, we design a posterior augmentation architecture which augments the visually-derived PBMs to the audio-derived masks via a gating network. This posterior augmentation architecture offers a loose coupling between the audio and visual information, which is different from the feature-fusion architecture taken by existing AVSE methods. By this architecture, the entire performance is dominated and lower-bounded by the audio-based system, and the visual-based system provides only auxiliary contribution. This is suited to the weakness of visual information and provides a safe way to make use of the visual knowledge for SE in general conditions. Finally, our model is based on unidirectional LSTM, which permits us performing online enhancement.

We evaluate the performance of our model in the perceptual evaluation of speech quality (PESQ) [16] under different SNR conditions. The result on the Grid database demonstrated that the OVA approach consistently outperforms the audio-based system in all experiment setups, and outperforms two state-of-the-art AVSE models in conditions of medium SNRs.

2. MODEL ARCHITECTURE

In this section, we present the OVA architecture as shown in Fig. 1, which involves the audio-based component, visual-based component and the augmentation component.

2.1. Audio-based component

The audio-based component, shown in the red box in Fig. 1, follows the design of Gao et. al. [17]. The speech feature fed to the input layer is 257-dimensional log power spectrum (LPS). Three LSTM layers are stacked, followed by a full-connection (FC) layer. Each LSTM layer contains 1024 cells, and the output of the FC layer contains 257 units, equal to the dimension of the power spectrum(PS) feature. The role of this network is to accept features of noisy speech signals and predict ideal ratio masks (IRM) that are used to reconstruct clean speech. The IRM is simply defined as the ratio between the PS of clean speech and noisy speech which is expressed as the PS of clean speech plus the PS of noise speech at each time and on each frequency bin, i.e.,

$$\text{IRM}_t[f] = \frac{\mathbf{c}_t[f]}{\mathbf{c}_t[f] + \mathbf{n}_t[f]},$$

where $\mathbf{c}$ and $\mathbf{n}$ denotes the PS features of clean and noise speech respectively, and t and f index time and frequency bins respectively.

The audio-based component is trained following the traditional noisy training recipe. The neural net is trained using the noisy speech as the input and the IRM as the target. Once the model has been trained, it will be frozen during the following training steps for the visual-based component and the augmentation component. By this frozen, we will keep a strong audio-based system and keep the performance when visual information is not available. We will denote the IRM produced by the audio-based component by aIRM and train it with Mean-Square-Error(MSE) Loss with 5e-4 learning rate.

Table 1. The CNN structure for visual feature extraction.

CNN	$conv_1$	$conv_2$	$conv_3$	$conv_4$	$conv_5$	$conv_6$
Filters	128	128	256	256	512	512
Size	(5,5)	(5,5)	(3,3)	(3,3)	(3,3)	(3,3)
Stride	(2,2)	(2,2)	(2,2)	(2,2)	(2,2)	(2,2)

2.2. Visual-based component

As mentioned already, visual information is weak and can only be used to predict rough representations of speech signals. In this study, we consider the distribution of the spectrum power over frequency bins, and convert it to power binary masks (PBMs) as the rough representation of speech. This conversion is conducted by placing a threshold on each frequency bin:

$$\boldsymbol{\xi}_t[f] = \frac{\mathbf{x}_t[f]}{\sum_{f=0}^{K} \mathbf{x}_t[f]}$$

$$\text{vPBMs}_t[f] = \begin{cases} 0.1 \quad if \;\; \boldsymbol{\xi}_t[f] \le \gamma \\ 1 \quad if \;\; \boldsymbol{\xi}_t[f] > \gamma \end{cases}$$

where $\mathbf{x}$ denotes the power spectrogram feature, and $\boldsymbol{\xi}$ denotes the power distribution. γ is a hyper-parameter and was empirically set to 10^{-5} in this study. K is the quantity of time-frequency-bins and set as 257 in experiment.Note that for the sake of clarity, we have used vPBMs to denote that it is produced by the visual component.

The main role of the visual-based component is to predict vPBMs of an audio stream given the associated visual stream. This is implemented as a CNN-LSTM network, where the CNN layers extract visual features from the input image and the LSTM layers deal with the audio-visual alignment and dynamic smoothing. More specially, the mouth image is firstly cropped from the input image, and the visual features are extracted by a CNN network whose architecture is the same as the one used in [10] and has been shown in Table 1. The extracted visual features are propagated to an LSTM network with five layers, where each layer contains 1024 LSTM cells. The recurrence speed of the first layer is the same as the sampling rate of the visual frames, while this speed is 4x times for the 2-4 layers, in order to meet the sampling rate of the audio stream. The output of the 4th LSTM layer is converted to a 257-dimensional vPBMs vector by a FC layer. Note that due to the memorial capability of LSTM, this architecture can deal with the time-shift problem between the visual and audio streams, and can smooth the vPBMs prediction by learning the dynamic property of speech signals. The Visual-based component is trained with video stream and vPBM generated by clean speech with Cross-Entropy loss with 5e-4 learning rate.

2.3. Augmentation component

Once the rough speech patterns vPBMs have been predicted by the visual-based component, we need augment them to the output of the audio-based output (aIRM). As shown in Fig. 1,

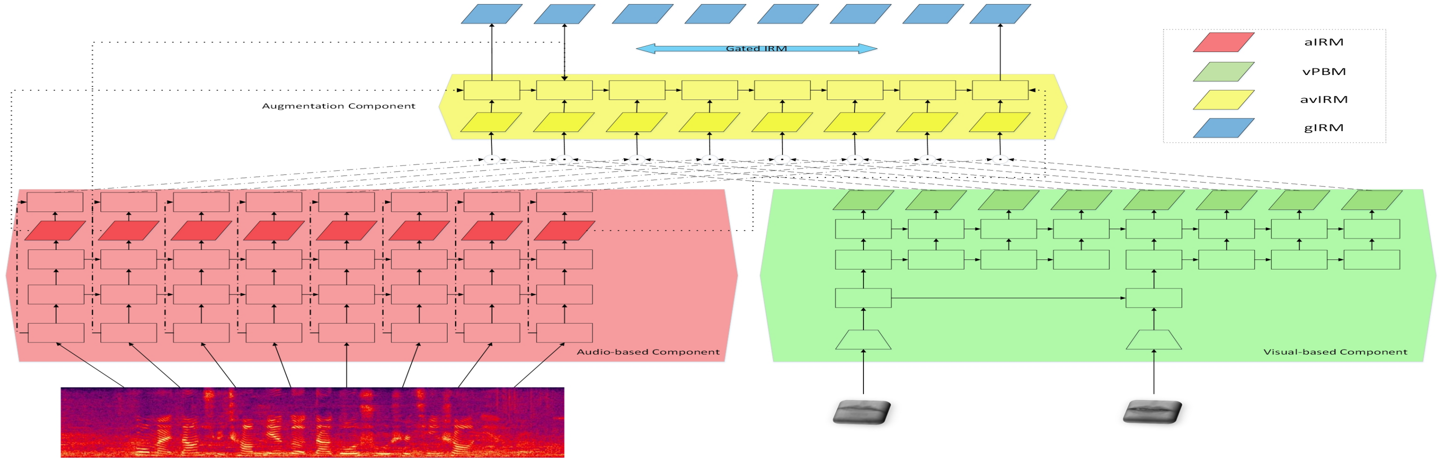


Fig. 1. The online visual augmented (OVA) SE architecture. The red-box represents the audio-based component, which produces audio-based IRM (aIRM); the green box represents the visual-based component, which produces visual-based PBMs (vPBMs). The yellow box is the augmentation component, which augments vPBMs to audio features to form avIRM, and outputs gated IRM (gIRM) via a gating network.

the vPBMs multiplied in element-wise by audio features produced by a stacked FC layer that has two dense layers from concatenated aIRM and noisy LPS(nLPS), and the resulting avIRM is fed to the gating network together with aIRM, producing the gate λ. This gate is used to integrate avIRM and aIRM and produce gIRM where formally written by:

$$\begin{aligned} \text{avIRM}_t &= f([\text{aIRM}_t; \text{nLPS}_t]) \odot \text{vPBMs}_t \\ \lambda_t &= g(\text{avIRM}_t) \\ \text{gIRM}_t &= \lambda_t \times \text{aIRM}_t + (1-\lambda_t) \times \text{avIRM}_t \end{aligned}$$

where f is postprocessing network to enhance $aIRM$ prediction with noisy speech signal features, which involving 2 stacked 1024 cells LSTM layers and 257 cells dense layer output, $\odot$ is element-wise multiply and g is the gating network that consists of an LSTM layer and a FC layer. The reason why we use gating network instead of feed $aIRM$ and $avIRM$ together into postprocessing network is that when visual signal is unavailable the $vPBM$ switches to 1-matrix and $gIRM$ becomes $aIRM$ and keeps model performance lower-bounded by audio-based component. The Augmentation component is trained with MSE Loss with 5e-4 learning rate after other two components converge.

3. EXPERIMENTS

3.1. Experimental Setup

3.1.1. Data Collection

The experiment is conducted with the Grid dataset, which contains speech signals associated with the video of the front face of the speaker. The full dataset contains 32 speakers, each contributing 1000 video segments. We divide these videos into a **Training** set which contains 30 speakers and 900 videos per speaker; a **Test (S)** set which contains the same 30 speakers as the training set and 100 videos per speaker not in the training set; and an **Test (U)** set which contains 2 speakers that are not in the training set, each with 1000 videos.

The noise signals are from two datasets: the CHiME background noise and the AudioSet noise. The noise in CHiME is categorized into 4 types: Cafe, Street, Bus and Pedestrian. For each type, part of the noise signals (80%) will be used to corrupt the training data and the rest are used to corrupt the test data. The Audioset involves 18,000 human speech segments, all of which are used to corrupt the test data. In other words, the CHiME noise can be regarded as *known* while the Audioset human noise is *unkown*.

3.1.2. Visual data preparation

All the videos are re-sampled into 25 FPS and are cutted into short segments of 3 seconds. The mouth anchors are detected by the MTCNN[18] and is cropped from the image. This mouth image is then resized to an image of 160x160 pixels, which is used as the input of the visual-based component. This process is identical for both the training set and the two test sets.

3.1.3. Audio data preparation

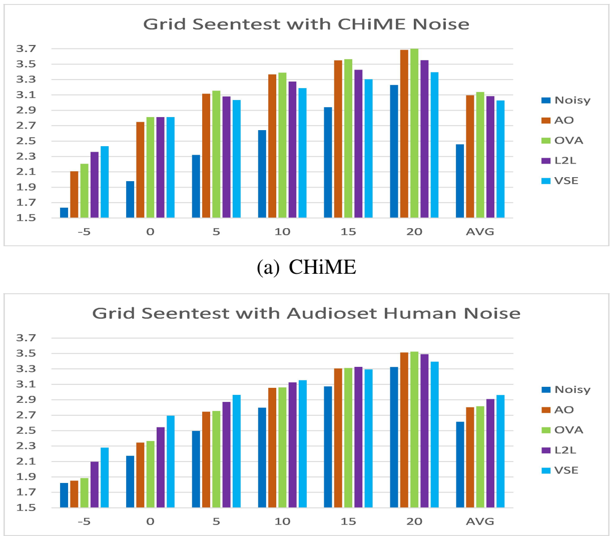
We follow the pipeline used in the previous study [17, 11] to prepare the audio data. It involves mixing various types of noise into clean speech to produce noisy speech. Before the mixing, noise signals is firstly clipped to 5-second segments, and then a random truncation is applied to produce 3-second segments to meet the length of the clean speech segments.

For the training set, the noise is from CHiME, and the mixing is conducted at 3 SNR levels (-5,0,5), leading to 243,000 utterances in total. For the two test sets, the noise is from the either CHiME or AudioSet, and the mixing is conducted at 6 SNR levels (-5,0,5,10,15,20). This leads to 18,000 utterances for Test(S) and 12,000 utterance for Test(U) after mixing.

The speech signals are re-sampled to 16 kHz, and LPS features are extracted using the Scipy and Librosa packages, with the window length set to 400, hop size to 160, and the FFT length to 512. The LPS features are used as the input of the audio-based component.

3.2. Experimental Results

We compare our model (OVA) with two state-of-the-art AVSE models: the visual speech enhancement (VSE) [10] and the Looking-to-Listen (L2L) model [11]. All these methods showed good performance in low-SNR conditions, especially in condition with human noise. We use the code published by the original authors to reproduce VSE, and implement the L2L model following the original paper [11]. For a better comparison, we also report the results with the original noisy speech (Org) and that of our audio-based component, i.e., an audio-only (OA) baseline. The metrics used for performance valuation is PESQ.



(a) CHiME

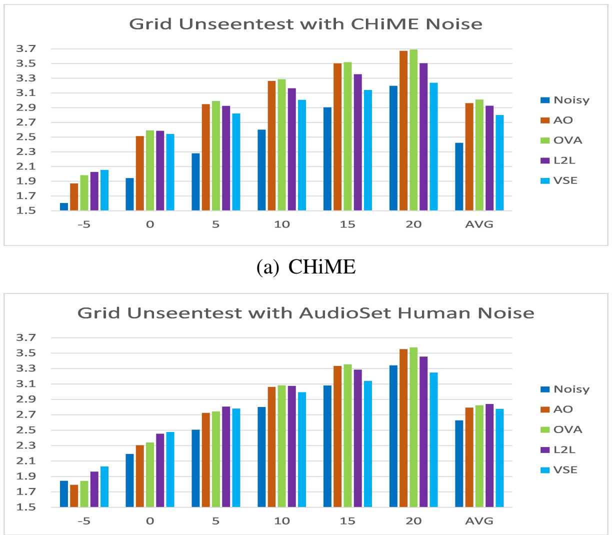
(b) AudioSetHuman

Fig. 2. PESQ results on seen speakers with CHiME noise (top) and AudioSet noise (bottom).

We first investigate performance on Test (S), where the speakers have been seen during model training. The results are shown in Fig. 2, where the results with the CHiME noise (known) and the AudioSet human noise (unknown) are reported separately. It can be seen that in all the noise conditions, the OVA approach outperforms the AO baseline, which means that OVA is a safe way to make use of the weak but useful information within the visual stream. In contrast, both L2L and VSE contribute only in conditions with a low SNR. This is more clear in the case with the CHiME noise, where L2L and VSE provide significant performance gains when SNR = -5.

Compared to L2L, VSE seems more visual-dependent, i.e., when the visual information is the most useful (i.e., SNR is low), it offers the most benefit, however when the visual stream is less useful (i.e., SNR is high), it causes the most damage (compared to the OA baseline). This indicates that the audio-visual information coupling with VSE is the strongest among all the three audio-visual enhancement methods.

Comparing the two noise types (CHiME and AudioSet), it can be seen that the AVSE approaches provide more significant gains in conditions with AudioSet human noise. This is expected as the human noise possesses similar spectral patterns as the target speech, thus difficult for the audio-only system to identify and remove.



(a) CHiME

(b) AudioSetHuman

Fig. 3. PESQ results on unseen speakers with CHiME noise (top) and AudioSet noise (bottom).

In the second experiment, we test the performance on Test (U), where the speakers are not in the training set. This is a more practical but more difficult scenario. The PESQ results are reported in Fig. 3. Once again, the OVA approach outperforms the AO baseline in all the test conditions, and it beats L2L and VSE on average. A particular observation is that the VSE approach is less effective in this unknown-speaker condition. For example, with the AudioSet human noise, it performs worse than L2L at SNR=5, however in the known-speaker condition, it performs the best even with SNR=10. This may imply that VSE tends to overfit to seen speakers.

4. CONCLUSION AND FUTURE WORK

We presented an online visual Audio (OVA) SE model, which predicts rough speech patterns from visual stream and the prediction is augmented to the output of an audio-based system, leading to a conservative but robust utilization of the weak information embedded in the visual stream. Our experiments demonstrated that the proposed method can consistently outperform the audio-only baseline at different SNR levels. This is contrast to the conventional stream-fusion methods that work well in low-SNR conditions but miserably fail when the SNR is medium or high. Future work will investigate combing the OVA approach and the stream-fusion approach so that we can boost the performance on low-SNR conditions.

5. REFERENCES

[1] Jae Soo Lim and Alan V Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[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] 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.

[4] 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.

[5] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[6] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[7] Mengyuan Zhao, Wang Dong, Zhiyong Zhang, and X-uewei Zhang, "Music removal by convolutional denoising autoencoder in speech recognition," in *Signal & Information Processing Association Summit & Conference*, 2016.

[8] Ibrahim Almajai and Ben Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2010.

[9] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, "Twin-hmm-based audio-visual speech enhancement," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3726–3730.

[10] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement using noise-invariant training," *arXiv preprint arXiv:1711.08789*, 2017.

[11] 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.

[12] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain, "Cochleanet: A robust language-independent audio-visual model for speech enhancement," *arXiv preprint arXiv:1909.10407*, 2019.

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

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

[15] L. Girin, J. L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007, 2001.

[16] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[17] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (I-CASSP)*. IEEE, 2018, pp. 5054–5058.

[18] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.