Audio-Visual Speech Enhancement

Chun-Hsian Chan, Bo-Hong Cho, Cheng-Yu Lin, Chi-Sheng Yang
University of Michigan, Ann Arbor, MI, USA
{kenhchan, bhcho, chenyul, chisheny}@umich.edu

Abstract

Speech enhancement has been studied over a decade by neural science [11], signal processing [20, 28, 4, 3, 1], and machine learning [10, 12, 13]. Researchers have proposed both traditional methods and deep learning-based models to accomplish this task.

In this project, we not only reproduce several models on this problem but also propose the modified models according to the related works. We implement the traditional independent component analysis method [1], the audio-only deep neural networks [24, 33], and the audio-visual deep neural networks [9]. Moreover, we suggest three modified models based on [9, 6, 30, 29, 22].

We design several dataset scenarios to train the models and examine their performance. We utilize the Obamanet [19], Avspeech dataset [7], and the environmental sound classification(ESC-50) [26] to build up our training pools. Based on the massive experiments, we convince the advantage of a deep learning-based model and the benefit of training with additional visual input.

1. Introduction

Group$(1, 1)$ A common experience chatting under the tumult is that we can often focus on the voice that we care about and mute the noise from the crowd. Speech enhancement, an amazing ability of humans has been studied over a decade by several fields, such as neural science [11], signal processing [20, 28, 4, 3, 1], and machine learning [10, 12, 13]. In the past, several traditional optimizations methods [28, 4, 1] have been applied to extract the target signal for the mixed audio. Recently, deep neural networks have been extensively used to accomplish this task. Studies train the deep neural network to enhance the audio signal. In addition, inspired by the studies [23, 11] that humans can resolve the target speech from the noisy environment better when they gaze at the speaker, researches [18, 17] also utilize both audio and visual information to train their deep neural networks.

We take a general overview and implement several models in the speech enhancement task. Additionally, we combine several networks structure to build hybrid networks trying to improve the current results. To be specific, we reproduce three models including traditional optimization methods [1], audio-only deep neural networks [24, 33], and audio-visual deep neural networks [9]. Furthermore, we propose three hybrid networks based on the models in [9, 6, 30, 29, 22].

For the traditional method, we implement the algorithm of independent component analysis [1], which uses gradient descent to minimize the L_2 norm of the objective function. For the audio-only model, we use U-Net to build a subnetwork in [24, 33]. This model is trained only under the audio spectrogram and it separates the foreground and background audios. For the audio-visual model, we implement the model of Gabbay et al. [9]. This is an encoder-decoder style model which builds by convolution neural networks and it is trained by both audio and visual information.

Based on the networks of Gabbay et al. [9], we propose several modified models inspired by Ephrat et al. [6], Sutskever et al. [29] and Luong et al. [22]. First, we utilize the spectrogram mask to adjust our loss function. We treat the enhanced spectrogram as a mask filtering the mixed audio signal. That is, we multiply the enhanced spectrogram and the mixed spectrogram and use the product to compute the loss. Second, we reinforce the model to learn the relationship between audio and video. That is, we add a bidirectional long short-term memory(BiLSTM) layer after the concatenated layer. Besides, we apply the local attention mechanism after the BiLSTM layer, letting the features between the temporal layers and the original concatenated layer combine more properly.

To examine the performance, we design several datasets to train the models. We use the Obamanet [19], Avspeech dataset [7], and the environmental sound classification(ESC-50) [26] to build our experiment. We observe the performance difference with training the models under different mixed datasets, and this observation again highlights the advantage of training with the video information. Above all, we have several interesting observations in our test. Please refer to the section 5 for more details.

We organize our work as follows. Section 2 is the related works including traditional methods, audio-only deep neural networks, and audio-visual deep neural networks. In section 3, we introduce the detailed constructions and algorithms of the reproduced models. In section 4, we introduce the datasets that we train our model. In section 5, we describe the experiments and give a comprehensive analysis as well as comparison. Finally, we end up our works with the conclusion in section 6.

2. Related Work

Speech enhancement is one of the fundamental tasks in signal processing. In the past, several traditional methods have been developed, such as Wiener filtering [20], spectral restoration [28, 4], statistical model-based methods [3], and independent component analysis (ICA) [1]. In contrast, deep neural network methods have gained a lot of attention recently, since they generally outperformed the traditional methods in either performance and practicality.

Audio-only deep learning-based model Previous studies on the audio-only model mainly involve single-channel speech enhancement and work well in both waveform and spectrogram input types. For example, Pascual et al. [25] train generative adversarial networks in waveform level and Lu et al. [21] use spectrogram as an input of the deep autoencoder to predicting denoising speech. Besides, some audio-only deep neural networks sever multi-speaker separation tasks. Those models [2, 15] differentiate source and noise by speech characteristics, such as spectral bands, pitches, and chirps. For more detail development in the audio-only deep-learning models can refer to the overview gave by Wang et, al. [31].

However, audio-only approaches are weak in separating similar human voices, for instance, they are not good at separating same-gender mixtures [15]. To overcome this, researchers propose the audio-visual model. This improvement leads to higher performance in both single and multi-speaker tasks and also serves well in the same-gender scenario.

Audio-visual deep learning-based model The audio-visual models [10, 18, 17, 12, 13] generally outperform the audio-only models, especially in the scenario of multi-speakers and same-gender speakers.

Recent works can be further classified according to speaker-dependence. A speaker-dependent model means that different speakers have to be labeled and train separately. Some studies of the speaker-dependent model are mention as follows. Hou et al. [14] provide a convolutional neural network inputting the frame of the speaker's lips region and the spectrogram of noisy speech. Their model then outputs an enhanced speech spectrogram as well as the reconstructed lips region. Gabbay et al. [8] use the trained speech generation network [5] with input both video frames and the spectrograms. They build the masks according to the enhanced spectrogram to filter the noise.

On the other hand, several independent works have recently take a step further to build a speaker-independent model. Ephrat et al [6] use face tracking systems to simultaneously separate different speakers in the same video. Given a user-chosen speaker they can sperate the target audio signal from the multi-speaker scenario. Owens et al. [24] use a self-supervised model. They randomly shift the audio signal to force the model to learn the temporal correlation between audio and visual streams.

3. Methods

In this section, we introduce the reproduced model's architecture, including independent component analysis method [1], deep learning-based audio-only model [24, 33], and deep learning-based audio-visual model [9]. Besides, we proposed three deep learning-based modified networks inspired by [6, 29, 22].

3.1. Traditional Method: Independent Component Analysis

We implement the algorithm of independent component analysis (ICA)[1] which is a clustering method in speech enhancement. ICA can find an optimal un-mixing matrix W given several mixed sources. By the matrix multiplication $Y = W^T X$, we can use an un-mixing matrix W to separate the mixed source X and the target source Y. The algorithm is finding a linear combination of the original mixed sources that maximize the objective signal. Intuitively, ICA obtains the statistical independence of the output signal, in terms of the probability density functions.

The algorithm of finding un-mixing matrices W has 2 steps. First, we use a contrast function $L(W)$ to measure the statistical independence of the mixed-signal X. Second, we minimize the objective function $L(W)$: $W^* = \text{argmin}_W L(W)$. Instead of obtaining an inverse matrix which is relatively computationally expensive, we can use gradient ascent to obtain matrix W. The pseudo-code is present as below:

Algorithm 1 Independent Component Analysis

procedure ICA
  Centralized signal by calculating μ
  Normalized the variance of signal by σ
  for 1 to the number of components **do**
    repeat
      $w_p = \frac{1}{n}\sum_i^n X g(w^T X) - \sum_i^n g'(w^T X) w$
      $w_{p+1} = w_p - \sum_{j=1}^{p-1}(w_p^T w_j) W_j$
      $w_{p+1} = \frac{w_{p+1}}{||w_{p+1}||}$
    until $w_p^T w_{p+1}^T = 1$

Though the ICA method can separate each mixed-signal greatly, we have to run the entire optimization process for each individual mixed-signal. Thus, it is time-consuming and fails to apply in real-time speech enhancement tasks. Another drawback is that we have to specify the number of speakers in the mixed source in this method and this is often impossible in real-world applications. The architecture of the audio-only model is constructed by U-Net [27]. Studies [24, 33] also use a similar structure on their sound-analysis sub-network. U-Net takes the spectrogram of the mixed audio and predicts both log magnitude and phase of the foreground spectrogram and the background spectrogram.

Data Preprocessing To reproduce the experiments in [24], we used the spectrogram with 64 ms window length and 16 ms step size, which gives 128×1025 spectrograms. Training samples are created by randomly pairing the audios from the training set and the noise set. We normalized each waveform by their root mean squared amplitude before the mixing.

Training details As the approach in [24], we use the permutation invariant loss proposed by [32].

$$L_p(x_F, x_B, \hat{x_1}, \hat{x_2}) = \min(L(\hat{x_1}, \hat{x_2}), L(\hat{x_2}, \hat{x_1})), \quad (1)$$

where both $L(x_i, x_j) = ||x_i - x_F||_1 + ||x_j - x_B||_1$ and $(\hat{x_1}, \hat{x_2})$ are predictions, x_F is the foreground spectrogram, and x_B is the background spectrogram. The loss of the phase is scaled by 0.01 since it is empirically less important.

3.2. Audio-Visual Model

We implement the model of Gabbay et al. [9]. The model separately encodes the audio and visual signal with encoder built by the convolution neural networks, then it concatenates two encoders with a shared embedding layer, and finally, the audio decoder extracts the output spectrogram from the networks. Detail construction in each step is as follows:

Video Preprocessing Every input video is first resampled to 25 fps and each 1-seconds visual signal will be divided into 5 individual grayscale frames of size 128×128 with the lip's region only. The cropped process is according to the mouth landmarks in [16].

Audio Preprocessing The corresponding 1-seconds audio signal will first be resampled to 16 kHz, and Short-Time-Fourier-Transform(STFT) will turn the waveform information into a spectrogram. Instead of inputting both phase and magnitude into the audio encoder. We only use STFT magnitude as the input of the encoder. The spectrogram will be divided into 5 segments per second, to match the 5 frames. Each segment of the spectrogram has a size of 80×20, which processes 20 temporal samples and 80 frequency bins. We remark that the STFT phase information will be kept to reconstruct the enhanced waveform.

Video Encoder The video encoder has 6 consecutive convolution layers with each layer built by Batch Normalization, Leaky-ReLU for non-linearity, max pooling, and Dropout of 0.25.

Audio Encoder The audio encoder consists of 5 convolution layers with each layer including Batch Normalization and Leaky-ReLU for non- linearity. Audio encoder use the stridden convolutions to maintain the temporal order.

Shared Representation After encoding the videos and the audios separately, we concatenate the feature vector of two encoders together. The shared embedding representing feature vector has 5,248 values (2,048 values for video feature vector and 3,200 values for audio feature vector). The concatenate vector than fed into 3 consecutive fully-connected layers, of sizes 1,312, 1,312, and 3,200, respectively.

Audio Decoder After all, we put the output vector to the audio decoder which has 5 transposed convolution layers(transposed structure of the audio encoder) to reconstruct the enhanced audio spectrogram.

Optimization The loss is measure by the mean square error(l2-norm) between the enhanced audio spectrogram and the target audio spectrogram. Initially, Adam optimizer has a learning rate $5e-4$, and if the validation error does not improve in 5 epochs, the learning rate will decrease by one half.

3.3. Proposed Method

Spectrogram Mask Based on the studies of [6, 30], we know that if we predict a spectrogram mask according to the magnitude part of the enhanced spectrogram, then we can use such mask as a filter to separate the noise for the mixed audio with higher quality. In practice, we use the ratio mask, a matrix indicates the ratio between the magnitudes of the enhanced spectrogram and the mixed spectrogram. After creating a ratio mask by the enhanced spectrogram, we compute the filtered result by element-wise multiplication between the mask and the mixed spectrogram, and we use such a product to compute the loss function and recover the enhanced speech.

Temporal Models The original model of Gabbay et al. [9] uses three fully connected (fc) layers in their shared representation. To further realize the temporal relationship between audio and video, we added a bidirectional long short-term memory(BiLSTM) layer to reference the model to learn the temporal feature of the concatenated layer. In addition, inspired by the temporal features extraction in the sequence-to-sequence[29] models such as attention mechanism[22]. We applied the local attention mechanism after the BiLSTM layer to combine the features between the temporal layers and the concatenated layer.

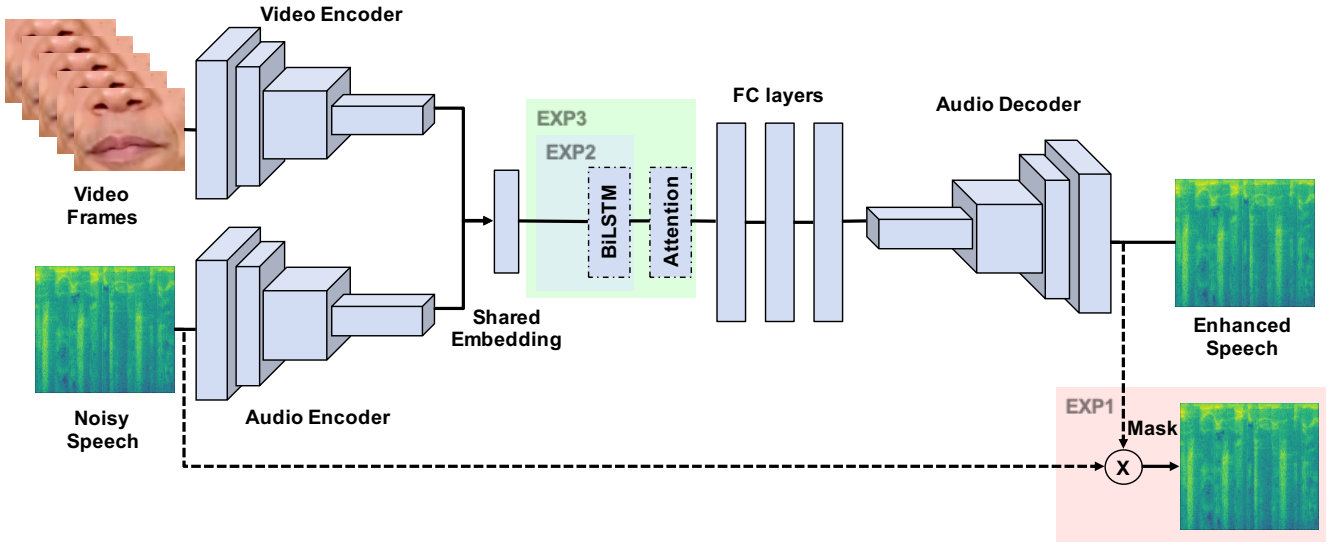

Figure 1. The architecture of the audio-visual model. **Baseline:** 5 video frames centered on the lip region and the spectrogram of noisy speech are fed into the convolutional encoder separately and the enhanced spectrogram is recovered by audio decoder after the concatenation. **Exp1:** Rather than directly generating the spectrogram, the model generates the spectrogram mask and take the multiplication between the mask and the input spectrogram to get the prediction. **Exp2:** we use the biLSTM layer after the shared embedding to get the temporal information. **Exp3:** We use the attention layer after the biLSTM layer to enhance the temporal information.

4. Datasets

We utilize several datasets to build up our training pool. All the models are trained under the 2100 training samples, 250 validation samples, and 250 testing samples. Each input is 3-second mixed audio(or mixed audio + video). We build noisy audio samples by mixing the target source(Obamanet [19]) and the noise source(Avspeech [7] or ESC-50 [26]). In this section, we have an overview of each dataset and in section 5, we describe the detailed combination between the target source and the noise source that we use in our experiment.

Obamanet This dataset includes 300 videos and each video consists of 2-3 minutes of the talk given by Barack Obama. Those videos are diverse in background, lighting, and vocabulary. To build our target source, we take one-third of the dataset and divide those videos into 2600 different 3-second videos.

Avspeech Dataset The Avspeech dataset is introduced in [6], and it contents enormous diversity. Especially, the dataset has approximately 150,000 distinct speakers, spanning a wide variation of people, languages, and the environment. We randomly pick 1000 distinct speakers as our noise source.

Environmental Sound Classification The Environmental Sound Classification (ESC-50) dataset is a collection of environmental recordings that build by the Freesound.org project. The audios record from the public field and content 2000 environmental recordings. We randomly pick 1000 distinct recordings as our noise source.

5. Experiments and Analysis

In this section, we set up several experiments to compare different models' performance. We design several datasets to support our analysis. The dataset $\boldsymbol{Obama/ESC}$ is built by mixing the target source of 2600 audios from Obamanet with the noisy source of 1000 distinct environment audios from ESC-50. The dataset $\boldsymbol{Obama/self}$ is generated by mixing the target source with the noisy source of itself. We call a dataset as $\boldsymbol{Obama/AV_n}$ if the data set is created by the target source and the noisy source of n distinct speaker from Avspeech. In our experiments, we have $\boldsymbol{Obama/AV_1}$, $\boldsymbol{Obama/AV_{10}}$, $\boldsymbol{Obama/AV_{100}}$, and $\boldsymbol{Obama/AV_{1000}}$.

We have 1 traditional model as well as 5 deep neural network models. The model $\boldsymbol{M}_{ICA}$ denotes the traditional model made by the ICA algorithm. The model $\boldsymbol{M}_{U}$ is the audio-only model introduced in [24, 33]. The model $\boldsymbol{M}_{AV}$ is the audio-visual model introduced by Gabbay et al. [9]. For the proposed models based on $\boldsymbol{M}_{AV}$, we use $\boldsymbol{M}_{Mask}$ to denote the spectrogram mask model, the model $\boldsymbol{M}_{LSTM}$ is the model adding the BiLSTM layers, and the model $\boldsymbol{M}_{Atten}$ is the model adding an attention layer. We remark that all the deep neural network models are trained under 2100 training samples, 250 validation samples, and 250 testing samples.

There are two main experiments. First, we compare the performance between those 6 models under the datasets $\boldsymbol{Obama/AV_{1000}}$, $\boldsymbol{Obama/ESC}$, and $\boldsymbol{Obama/self}$. Second, we compare the performance of the deep

Model	Obama/AV$_{1000}$		Obama/ESC		Obama/self	
	PESQ	SDR	PESQ	SDR	PESQ	SDR
M_{AV}	**3.43**	8.89	**3.59**	10.21	**3.14**	7.08
M_{Mask}	3.12	**10.18**	3.15	**11.53**	2.85	**8.36**
M_{LSTM}	3.34	7.78	3.49	8.80	3.03	6.06
M_{Atten}	3.33	7.69	3.48	8.65	2.99	5.97
M_U	2.00	2.41	1.93	2.52	3.12	6.54
M_{ICA}	3.38	42.71	2.62	46.91	3.77	42.58

Table 1. The evaluation of all methods with datasets $\boldsymbol{Obama/AV_{1000}}$, $\boldsymbol{Obama/ESC}$, and $\boldsymbol{Obama/self}$.

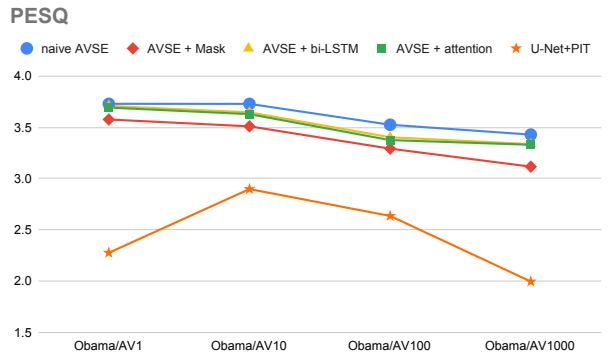

Figure 2. Comparison of PESQ among all the deep learning-based methods on the datasets $\boldsymbol{Obama/AV_1}$, $\boldsymbol{Obama/AV_{10}}$, $\boldsymbol{Obama/AV_{100}}$, and $\boldsymbol{Obama/AV_{1000}}$.

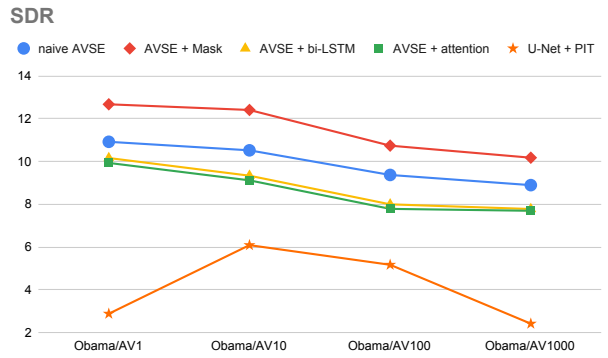

Figure 3. Comparison of SDR among all the deep learning-based methods on the datasets $\boldsymbol{Obama/AV_1}$, $\boldsymbol{Obama/AV_{10}}$, $\boldsymbol{Obama/AV_{100}}$, and $\boldsymbol{Obama/AV_{1000}}$.

learning-based models under different noise complexity. That is, we compare those models under the datasets $\boldsymbol{Obama/AV_1}$, $\boldsymbol{Obama/AV_{10}}$, $\boldsymbol{Obama/AV_{100}}$, and $\boldsymbol{Obama/AV_{1000}}$. Due to the limitation of the traditional model, we exclude it from the second comparison.

Based on these experiments, we discuss our observation and analysis in the following subsections.

5.1. Traditional Model vs Deep learning-based Model

Though the traditional ICA method gives remarkable results, it has several limitations that fail to apply to the general scenario. That is, we need to run over entire the algorithm for separating every mixed sample, and we have to specify the number of the speaker in the mixed audio when we execute the algorithm. Due to those limitations, we calculate the SDR and PESQ of M_{ICA} by averaging 10 sample's results take from the datasets $\boldsymbol{Obama/AV_{1000}}$, $\boldsymbol{Obama/ESC}$, and $\boldsymbol{Obama/self}$. As shown in Table 1 the ICA method can deal perfectly with a human speech given the number of speakers(no matter the noises are self-noise or others). However, the ICA method is weak to separate the environment noise and the target speech as it has relatively low PESQ. We think the ICA method could not find a proper linear separation to cluster the speech since the noise probability distribution in the environment is too extensive.

Compared to the ICA algorithm, deep learning-based models can sever under diverse datasets setup, as shown in Table 1. Due to those essential differences, we separate the ICA method with others in our analysis and we detailly compare the deep learning-based models in subsection 5.2, 5.3, and 5.4.

5.2. Audio-only vs Audio-video

By the results shown in Table 1, we can see that the visual information essentially helps in the audio separation task. The audio-visual models outperform the audio-only model according to PESQ and SDR. That is, the audio-visual models on average have 1.4 times larger PESQ and 2.5 times larger SDR than the audio-only model. We further highlight some observation in Table 1.

The Audio-only model has better PESQ and SDR in $\boldsymbol{Obama/self}$ than other datasets. However, after manual inspection, we can not tell apart between the target and the noise. The high score in PESQ and SDR might due to the similarity between self-noise and target speech.

For the audio-visual method, all the models perform well in both $\boldsymbol{Obama/AV_{1000}}$ and $\boldsymbol{Obama/ESC}$. We note that self-noise has similar features with the target audio. Thus, all the models have the worst PESQ and SDR in the dataset $\boldsymbol{Obama/self}$.

5.3. Effect on Noise Complexity

In the noise complexity experience, we train all the deep learing-based models (M_U M_{AV} M_{Mask} M_{LSTM} M_{Atten}) under the datasets with different noise complexity $\boldsymbol{Obama/AV_1}$, $\boldsymbol{Obama/AV_{10}}$, $\boldsymbol{Obama/AV_{100}}$, and $\boldsymbol{Obama/AV_{1000}}$). All the audio-visual models(M_{AV} M_{Mask} M_{LSTM} M_{Atten}) perform as expected. The accuracy and quality decrease as the noise complexity increases.

The models are suffered from separating the noise as the noise pool becomes larger.

However, the audio-only model($\boldsymbol{M}_U$) has low performance in the dataset $\boldsymbol{Obama/AV_1}$ comparing to the datasets $\boldsymbol{Obama/AV_{10}}$ and $\boldsymbol{Obama/AV_{100}}$. This result is due to the following reason. In the single noise scenario, $\boldsymbol{M}_U$ tends to overfit the nose, since the noise source is more simple than the target source. Thus, $\boldsymbol{M}_U$ predicts the background(noise) spectrograms better than the foreground(target) spectrograms. However, we only evaluate the PESQ and SDR on the foreground audios, so we get a bad evaluation. As the noise complexity increase, $\boldsymbol{M}_U$ will no longer tend to fit the noise source. If this speculate is true, from the model aspect, it is quite interesting that the diversity of 10 distinct speaker noise from the Avspeech dataset is larger than the diversity of more than 100 audio from the Obamanet. Finally, we remark that audio-visual models do not face this problem since the additional video information can help the model distinct between the target and the noise.

5.4. Proposed Audio-Visual Models

We propose several adjustments based on the model $\boldsymbol{M}_{AV}$. To compare the effect on these modifications, we use the model $\boldsymbol{M}_{AV}$ as the control group and the experimental group including $\boldsymbol{M}_{Mask}$, $\boldsymbol{M}_{LSTM}$, and $\boldsymbol{M}_{Atten}$.

By adding the spectrogram mask, the model $\boldsymbol{M}_{Mask}$ has a better performance in SDR, but weaker performance in PESQ. Higher SDR indicates the spectrogram mask method has a lower phase difference. However, it remains unclear for the lower PESQ in the model $\boldsymbol{M}_{Mask}$.

By modifying the temporal layers, the performance of the models $\boldsymbol{M}_{LSTM}$ and $\boldsymbol{M}_{Atten}$ are both slightly weaker than the naive model $\boldsymbol{M}_{AV}$. These results occur because we do not use large enough dataset to train the models $\boldsymbol{M}_{LSTM}$ and $\boldsymbol{M}_{Atten}$. As the adjusted temporal layer cause the network's complexity significantly increases(increase 50% in total parameters), the model should have large enough dataset to generalize its massive parameters.

6. Conclusions

In this paper, we have presented several methods on the speech enhancement task, including conventional and deep learning-based methods. Deep learning-based methods can sever on a more generalized task compared to the conventional method and visual information significantly improves the performance of the deep learning-based models.

Based on our results, we propose several future works.

Spectrogram mask We need to further confirm the idea of spectrogram mask proposed by [6] will improve the enhancement speech quality.

Temporal model We need to further confirm the effect on adding the BiLSTM layer and the attention layer. Adding the BiLSTM layer and attention layer on shared representation give similar audio quality if we evaluated the audio by our ears, but they have bad SDR and PESQ scores. This result might come from the monotonous video clips in Obamanet, which has less temporal information.

Increase Computational Power We need to have more computational power and larger datasets to train the general model. In reproducing the study of Ephrat et al. [6] and Owens et al. [24] we have used the Avspeech dataset to our train the models. However, we are unable to deliver satisfying results because of the time constraint and the computing power. All of our reproduced models and preliminary results are put on the GitHub, though not all of them are presented in this paper.

References

[1] Lucas C. Parra Barak A. Pearlmutter. Maximum likelihood blind source separation: A context-sensitive generalization of ica. *NIPS*, 1997.

[2] Zhuo Chen. *Single Channel auditory source separation with neural network*. PhD thesis, Columbia University, 2017.

[3] Yariv Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526–1555, 1992.

[4] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[5] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462, 2017.

[6] 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.

[7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[8] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through noise: Speaker separation and enhancement using visually-derived speech. *arXiv preprint arXiv:1708.06767*, 4(11), 2017.

[9] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017.

[10] Laurent Girin, Jean-Luc Schwartz, and Gang Feng. Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America*, 109(6):3007–3020, 2001.

[11] Elana Zion Golumbic, Gregory B Cogan, Charles E Schroeder, and David Poeppel. Visual input enhances selec-

tive speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, 33(4):1417–1426, 2013.

[12] John Hershey, Hagai Attias, Nebojsa Jojic, and Trausti Kristjansson. Audio-visual graphical models for speech processing. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–649. IEEE, 2004.

[13] John R Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. In *Advances in Neural Information Processing Systems*, pages 1173–1180, 2002.

[14] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and HM Wan. Audio-visual speech enhancement based on multimodal deep convolutional neural network. *arXiv preprint arXiv:1703.10893*, 2017.

[15] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey. Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*, 2016.

[16] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[17] Faheem Khan. *Audio-visual speaker separation*. PhD thesis, University of East Anglia, 2016.

[18] Faheem Khan and Ben Milner. Speaker separation using visually-derived binary masks. In *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[19] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *CoRR*, abs/1801.01442, 2018.

[20] Jae Lim and Alan Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(3):197–210, 1978.

[21] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[22] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[23] Wei Ji Ma, Xiang Zhou, Lars A Ross, John J Foxe, and Lucas C Parra. Lip-reading aids word recognition most in moderate noise: a bayesian explanation using high-dimensional feature space. *PLoS One*, 4(3), 2009.

[24] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[25] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[26] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. 2015.

[27] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[28] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE, 1996.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[30] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *CoRR*, 2017.

[31] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[32] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245, 2017.

[33] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018.