Reference-Based Speech Enhancement via Feature Alignment and Fusion Network

Huanjing Yue[1], **Wenxin Duo**[1], **Xiulian Peng**[2], **Jingyu Yang**[1*]

[1]School of Electrical and Information Engineering, Tianjin University, China
[2]Individual
{huanjing.yue, hiedean, yjy}@tju.edu.cn

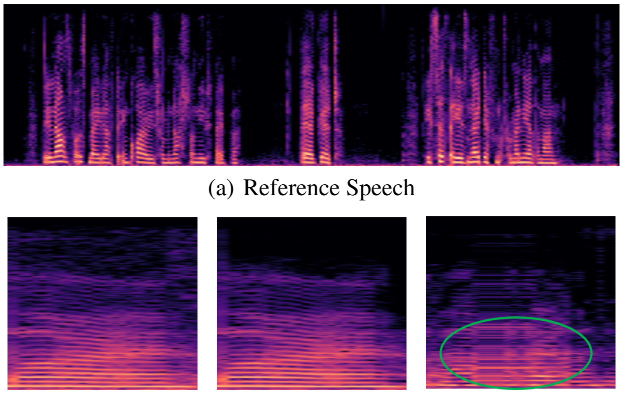
Abstract

Speech enhancement aims at recovering a clean speech from a noisy input, which can be classified into single speech enhancement and personalized speech enhancement. Personalized speech enhancement usually utilizes the speaker identity extracted from the noisy speech itself (or a clean reference speech) as a global embedding to guide the enhancement process. Different from them, we observe that the speeches of the same speaker are correlated in terms of frame-level short-time Fourier Transform (STFT) spectrogram. Therefore, we propose reference-based speech enhancement via a feature alignment and fusion network (FAF-Net). Given a noisy speech and a clean reference speech spoken by the same speaker, we first propose a feature-level alignment strategy to warp the clean reference with the noisy speech in frame level. Then, we fuse the reference feature with the noisy feature via a similarity-based fusion strategy. Finally, the fused features are skipped connected to the decoder, which generates the enhanced results. Experimental results demonstrate that the performance of the proposed FAF-Net is close to the state-of-the-art speech enhancement methods on both DNS and Voice Bank+DEMAND datasets. Our code is available at *https://github.com/HieDean/FAF-Net.*

Introduction

Collected speeches in the wild usually contain much background noise, which severely degrades the perceptual quality and intelligibility of speech. Therefore, speech enhancement, which aims to recover clean speech from a noise-corrupted speech, has attracted much attention and achieved promising results with the widely used deep-learning (DL) strategy. In this work, we focus on single-channel speech enhancement. Without specification, all the speech mentioned in this work is single-channel speech.

The DL based speech enhancement can be classified into time domain and time-frequency (T-F) domain based methods. Time domain based methods usually utilize a neural network to learn the mapping relationship between the one-dimensional (1D) waveform of noisy and clean speeches (Pascual, Bonafonte, and Serra 2017; Rethage, Pons, and Serra 2018; Koyama et al. 2020). Due to the lack of apparent geometric structure for the 1D signal, the time domain based methods are inferior to the T-F domain based methods, which operate in the 2D T-F spectrogram. The T-F domain based methods usually predict a mask to filter the noisy T-F spectrogram, generating a clean T-F spectrogram (Hu and Wang 2001; Srinivasan, Roman, and Wang 2006; Williamson, Wang, and Wang 2015). Although these methods have greatly improved the speech quality, the recovered speech still suffers from information lost.



(a) Reference Speech

(b) Noisy Speech (c) Clean Speech (d) Aligned Speech

Figure 1: STFTs Visualization of reference speech (a), noisy speech (b), clean speech (c), and aligned speech (d) using the log-spectra. The aligned speech is reconstructed by aggregating the matched frames from (a) using the matching index obtained by MFCC patch matching. However, there is obvious time discontinuity of the aligned speech, which is highlighted by a green circle.

To further improve the enhancement results, some methods introduce semantic information to assist speech enhancement (Ephrat et al. 2018; Hou et al. 2018). For example, the speaker identity (ID) is used in speech enhancement (Chuang et al. 2019; Giri et al. 2021; Shon, Tang, and Glass 2019) and separation tasks (Wang et al. 2018; Mun et al. 2020; Giri et al. 2021), proving that speaker prior derived from the noisy speech or another clean speech spoken by the same ID (*i.e.,* reference speech) is beneficial for the personalized speech enhancement and separation. However, the speaker ID is generally embedded into a global vector, and inserted into the reconstruction network for the enhancement

[*]This work was supported in part by the National Natural Science Foundation of China under Grant 62072331. *Corresponding author: Jingyu Yang.*

process. We observe that the frame-level phoneme sets of the reference speech and the target speech are shared, indicating that the reference and target speeches are correlated in frame-level. Meanwhile, the counterpart methods in image restoration (such as image denoising and super-resolution), usually utilize correlated information from the reference image in pixel (patch) level other than global feature embedding (Yue et al. 2013; Zhang et al. 2019; Yang et al. 2020; Yue et al. 2019), and have achieved significant gains compared with single image based restoration.

Based on the above observations, we propose a reference based speech enhancement (RefSE) method by exploring the local correlations between the noisy speech and reference speech. We would like to point out that this task is more challenging than that in reference based image restoration. For example, in reference based image super-resolution (RefSR), the reference can be directly aligned with the input image by pixel (patch) level matching. Directly matching the reference speech and noisy speech either on time domain or time-frequency domain, and aggregating the matched signals from the reference does not work well since this will destroy the inherent continuity of speech in time, as shown in Fig. 1 (d). Therefore, we propose a feature-level alignment strategy. Specifically, we first match the Mel Frequency Cepstrum Coefficient (MFCC) of the noisy speech and clean reference speech using several consecutive frames (termed as MFCC Patch, illustrated by the red box in Fig. 2 (a)). Then, we warp the convolutional neural network (CNN) features of reference speech with that of the noisy speech according to the matching index obtained in the MFCC matching. The warped reference feature and noisy feature are fused together via soft attention and channel attention based fusion strategy. Finally, the fused features are sent to the decoder to generate the enhanced speech. Our main contributions are summarized as follows.

- To our knowledge, we propose the first reference based speech enhancement (RefSE) method by exploring local correlations between noisy and clean reference speeches. To keep the time continuity of speech signals, we propose to first perform noisy-reference matching with MFCC patches and then warp the CNN feature of reference according to the matching index. This greatly improves the effect of reference speech.
- Since the reference features have different similarities with the input noisy feature and they contribute differently to the final enhancement result, we propose a soft attention and channel attention mechanism to fuse the features of reference and noisy speech together.
- Experimental results on two public datasets, *i.e.* Voice Bank + DEMAND and DNS, demonstrate the superiority of the proposed scheme. In addition, the proposed feature alignment and fusion strategy for reference outperforms global embedding by 0.22 in terms of PESQ.

Related Work

In this section, we briefly review related works in DL based single speech enhancement ("single" means the noisy speech is the only input), semantics guided speech processing, and reference based image restoration.

DL-Based Single Speech Enhancement

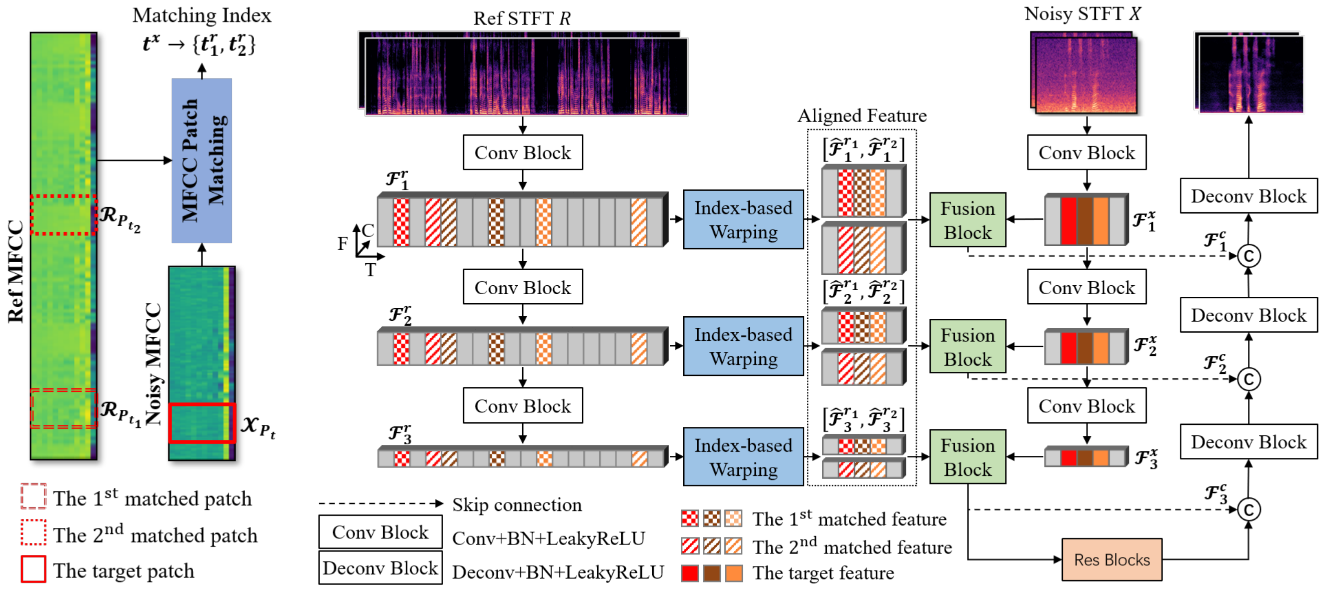
According to the processed signal domain, DL-based single speech enhancement methods can be classified into time domain based and T-F domain based.

Time domain. The time domain based methods usually directly map the noisy 1D waveform to a clean 1D waveform (Pascual, Bonafonte, and Serra 2017; Defossez, Synnaeve, and Adi 2020). A popular strategy is utilizing the encoder-decoder or UNet structure (Defossez, Synnaeve, and Adi 2020). The encoder encodes the noisy speech into feature domain, and the decoder decodes the enhanced features to time domain. The enhancement module is inserted between the encoder and decoder. Long-short term memory (LSTM) and temporal convolutional networks, which can explore the temporal correlations, are typical enhancement modules. Although these methods are effective for speech enhancement, the T-F domain based methods demonstrate that the noise patterns are more distinguishable in T-F domain than in time domain.

T-F domain. The T-F domain based methods usually predict a mask, to model the relationship between the targeted speech and noisy speech. The ideal binary mask (Hu and Wang 2001) and ideal ratio mask (Srinivasan, Roman, and Wang 2006) can only model the magnitude relationship between noisy and estimated speeches. Ignoring the phase information limits the enhancement upper bound of these methods. Hereafter, complex ideal ratio mask (cIRM) (Williamson, Wang, and Wang 2015), is proposed to model both the magnitude and phase relationships. Similarly, to take phase into account, (Yin et al. 2020) proposed a two-stream structure with mutual communication to extract features from both amplitude and phase spectrum. The work in (Choi et al. 2018) makes the network structure be aware of the phase information by extending the real-valued convolution to complex-valued convolution. Hereafter, based on (Choi et al. 2018), Hu *et al.* (Hu et al. 2020) designed a complex-valued LSTM to explore temporal correlations for phase aware speech enhancement. In this work, we also adopt a phase aware complex-valued encoder-decoder structure and learn a phase aware complexed ratio mask (Hu et al. 2020) for STFT reconstruction.

Semantics Guided Speech Processing

To further improve the performance of deep learning-based single speech enhancement, some methods propose to incorporate semantic information to assist the enhancement process. The work in (Hou et al. 2018) introduces mouth region visual features to assist speech enhancement and verifies that there exists strong correlations between speech content and lip shapes. However, the visual features contain too much redundant information. Therefore, some work utilizes speaker identity information to help speech enhancement. Chuang *et al.* (Chuang et al. 2019) proposed to first train a speaker embedder network to generate embedding vector for a given speaker, and then utilize the embedding to enhance the target speaker. Liu *et al.* (Liu et al. 2021) pro-



(a) (b)

Figure 2: Framework of the proposed feature alignment and fusion network (FAF-Net) for RefSE. For brevity, we omit the feature map visualization for the imaginary part.

posed to utilize phoneme identities to help speech enhancement. Specifically, they first learn distribution modulation parameters from the phoneme classification vector of the noisy input, and then the distribution parameters are used to modulate the speech enhancement features. The work in (Shon, Tang, and Glass 2019) proposed new loss functions by using verification features for audio speaker verification. However, all the identity features are extracted from the noisy input itself without introducing additional reference speeches. Meanwhile, for speech separation, *i.e.,* separating the voice of a target speaker from multi-speaker signals, some works (Wang et al. 2018; Mun et al. 2020) introduce reference speeches (spoken by the same speaker as that of the target speech) in the form of global embeddings since the global embeddings of target speech and reference speech are correlated. Recently, Giri *et al.* (Giri et al. 2021) proposed to utilize the speaker identity embeddings extracted from a clean reference for both speech separation and enhancement. However, there is still no work introducing reference speeches for speech enhancement by exploring local correlations. We observe that the frame-level phoneme sets of target speech and reference speech are shared, which inspires us to explore reference based speech enhancement by exploring frame-level correlations.

Reference-Based Image Restoration

Instead of using references in global level as in speech separation (enhancement), we observe that the reference-based image restoration (which utilizes high quality reference image to assist the restoration of degraded image) usually utilizes the reference image in patch-level, and have achieved outstanding results in RefSR (Yue et al. 2013; Zhang et al. 2019; Yang et al. 2020) and Ref-denoising (Yue et al. 2015, 2019). For example, the works in (Yue et al. 2013, 2015) search similar patches in pixel domain from the high quality reference image, and utilize the similar reference patch to help restoration of the degraded query patch. With the development of deep learning, the works in (Zhang et al. 2019; Yang et al. 2020; Yan et al. 2020) search for similar patches in CNN feature domain and the feature extraction network can be set to a fixed benchmark classification network, such as VGG or ResNet. Inspired by these works, we propose to explore the correlations between noisy and reference speeches in MFCC patch level. Specifically, we search similar MFCC patches from the reference with the noisy MFCC patch as query, and then warp the deep features of the reference STFT spectrogram according to the matching index.

Approach

Overview

Suppose $X \in \mathbb{R}^{T_1 \times F \times 2}$, $R \in \mathbb{R}^{T_2 \times F \times 2}$ are the complex-valued STFT spectrograms of the noisy speech and its corresponding clean reference speech, we aim to learn a mapping function g to estimate $\hat{Y}$ from its noisy version X with the help of R, *i.e.*,

$$\hat{Y} = g(X, R). \tag{1}$$

T_1 (T_2) is the number of frames and F is the number of frequency bins. As shown in Fig. 2, there are three modules in our framework. First, given a noisy speech, we search its matched reference speech using MFCC patch, and then align the reference feature with the noisy feature according to the matching index. Then, we utilize a fusion block to fuse the reference and noisy features together. Finally, these fused features are sent to the decoder for the reconstruction of target STFT spectrogram. In the following, we give details for the three modules.

Reference Matching and Feature Alignment

In this work, the reference speeches are clean utterances spoken by the same person as that of the noisy input but having different contents. For each noisy speech clip, we randomly select several clean speech clips spoken by the same speaker and concatenate them along the time axis to construct a long reference clip of n seconds (in our experiments, n is set to 15). The reference clip usually does not contain the same words as that in the noisy clip but their frame-level phoneme sets are shared. This inspires us to utilize frame-level correlations between the reference and noisy speech to assist the enhancement process. The key problem in this task is how to align the reference frame with the noisy query frame, and this includes two steps: matching and warping.

Patch-level Matching. We observe that the MFCC features of speeches are simple and perform well in phoneme classification (Liu et al. 2021), speech recognition and speaker verification, which demonstrates that MFCC features are representative. Therefore, we propose to utilize MFCC features to search for similar speech frames. We denote the MFCC features of the noisy input and the reference speech as $\mathcal{X} \in \mathbb{R}^{T_1 \times F'}$ and $\mathcal{R} \in \mathbb{R}^{T_2 \times F'}$, where $T_1(T_2)$ is the number of frames (the same as that in X and R) and F' is the number of feature bins. For each frame $\mathcal{X}_t \in \mathbb{R}^{1 \times F'}$, we aim to find its most similar frame from $\mathcal{R}$. Since utilizing single frame for matching will ignore the temporal continuity of the frame sequence, we propose a patch based matching strategy. Specifically, three consecutive frames construct a patch, and the target patch in $\mathcal{X}$ searches its most similar patch in $\mathcal{R}$ via calculating the cosine similarity in a sliding manner. This process can be formulated as

$$d_{t,i} = < \psi(\mathcal{X}_{P_t}), \psi(\mathcal{R}_{P_{t_i}}) >, t \in [2, T_1 - 1], i \in [2, T_2 - 1] \qquad (2)$$

where $\psi(\cdot)$ represents normalization of the vector, $< \cdot, \cdot >$ represents the inner product, and $\mathcal{X}_{P_t}$ is the concatenation of $\mathcal{X}_{t-1}, \mathcal{X}_t$ and $\mathcal{X}_{t+1}$ along the feature dimension. After calculating the similarities, for each patch $\mathcal{X}_{P_t}$, we choose its top k similar patches with the highest similarity in $\mathcal{R}$ as the matched reference patches. The corresponding matching index is stored as $t^x \to \{t_1^r, t_2^r, ...t_k^r\}$. Fig. 2 (a) illustrates the matching process with $k = 2$.

Frame-level Warping. A straightforward way for warping is utilizing the MFCC matching index to rearrange the reference frames in R. Since we have searched k similar frames from $\mathcal{R}$ for each target frame in $\mathcal{X}$, we can build k new reference STFT spectrograms $\hat{R}^1, \hat{R}^2, ..., \hat{R}^k$ by extracting frames from R, where $\hat{R}^i \in \mathbb{R}^{T_1 \times F \times 2}$ is the i^{th} similar reference of X. However, this will severely destroy the time continuity since neighboring frames in $\hat{R}^i$ are selected from different time positions in R, as shown in Fig. 1 (d). To solve this problem, we propose a convolutional feature alignment strategy. Our speech enhancement network is a complex encoder-decoder structure. By aligning the encoder features of R with those of X and fusing them together, we can then take advantage of these fused features for decoder based reconstruction. As shown in Fig. 2 (b), both the original long reference R and the noisy input X first go through the same encoder module, which is constructed by three Conv blocks. Strided convolution is used, where the stride along the frequency (time) dimension is 2 (1). In this way, we obtain three level Conv features for X and R, which are denoted as $\mathcal{F}_i^x \in \mathbb{R}^{T_1 \times F_i \times C_i \times 2}$ and $\mathcal{F}_i^r \in \mathbb{R}^{T_2 \times F_i \times C_i \times 2}$, where i is the level index. Then, we warp $\mathcal{F}_i^r$ with $\mathcal{F}_i^x$ according to the MFCC matching index, generating aligned reference feature maps $\hat{\mathcal{F}}_i^{r_j}$, where $j = \{1, 2, ..., k\}$ is the similarity index. Although neighboring frames in $\hat{\mathcal{F}}_i^{r_j}$ are still selected from different positions of $\mathcal{F}_i^r$, each frame in $\mathcal{F}_i^r$ is generated by convolution with time-continuous inputs. Therefore, each frame in $\hat{\mathcal{F}}_i^{r_j}$ perceives continuous speech frames. In other words, "feature extraction + feature alignment" is more reasonable than "spectrogram alignment + feature extraction" strategy.

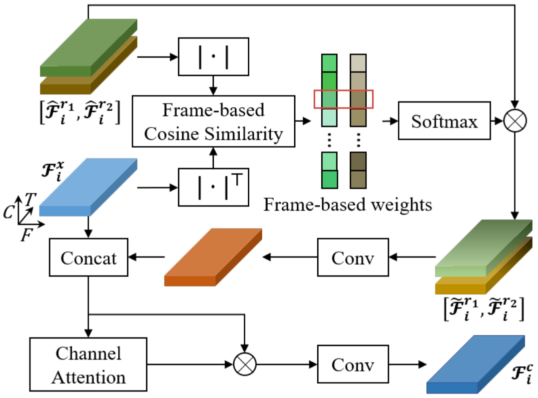

$[\hat{\mathcal{F}}_i^{r_1}, \hat{\mathcal{F}}_i^{r_2}]$
$|\cdot|$
Frame-based Cosine Similarity
$|\cdot|^{\top}$
Frame-based weights
Softmax
$\mathcal{F}_i^x$
C
T
F
Concat
Conv
$[\tilde{\mathcal{F}}_i^{r_1}, \tilde{\mathcal{F}}_i^{r_2}]$
Channel Attention
Conv
$\mathcal{F}_i^c$

Figure 3: Detailed network structure for the fusion block.

Feature Fusion Block

Since the aligned reference features have different similarities to the target speech, we propose to fuse them together via a soft attention based weighting strategy and a channel attention mechanism. Fig. 3 illustrates the fusion process for $k = 2$. We first calculate the amplitude of the complex aligned reference feature $\hat{\mathcal{F}}_i^{r_j}$ and noisy feature $\mathcal{F}_i^x$, generating $\bar{\mathcal{F}}_i^{r_j} = |\hat{\mathcal{F}}_i^{r_j}| \in \mathbb{R}^{T_1 \times F_i \times C_i}$ and $\bar{\mathcal{F}}_i^x = |\mathcal{F}_i^x| \in \mathbb{R}^{T_1 \times F_i \times C_i}$. Then, we reshape the t^{th} frame in $\bar{\mathcal{F}}_i^x$ into a vector of size $1 \times F_i C_i$ and the t^{th} frame in $\bar{\mathcal{F}}_i^{r_j}$ is reshaped into a vector of size $F_i C_i \times 1$, which are denoted as $\bar{\mathcal{F}}_{(i,t)}^x$ and $\bar{\mathcal{F}}_{(i,t)}^{r_j}$, respectively. Then, we calculate the cosine similarity between the two vectors, generating a weighting coefficients α_t^j, which can be formulated as

$$\alpha_t^j = < \psi(\bar{\mathcal{F}}_{(i,t)}^x), \psi(\bar{\mathcal{F}}_{(i,t)}^{r_j}) > . \qquad (3)$$

This process is repeated for all the frames, and finally we obtain k vectors, *i.e.*, $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, ..., \boldsymbol{\alpha}^k$. Then, we normalize these coefficients with the same frame number (highlighted by red box in Fig. 3) via softmax, generating normalized coefficients $\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\alpha}}^2, ..., \hat{\boldsymbol{\alpha}}^k$. Then, we obtain the weighted reference feature $\tilde{\mathcal{F}}_{(i,t)}^{r_j}$ for the t^{th} frame via

$$\tilde{\mathcal{F}}_{(i,t)}^{r_j} = \hat{\mathcal{F}}_{(i,t)}^{r_j} \cdot \hat{\boldsymbol{\alpha}}_t^j, j = \{1, 2, ..., k\}. \qquad (4)$$

To reduce the number of channels, the k weighted reference features are fused together via a convolution layer, followed by concatenation with the noisy feature along the channel dimension, resulting a feature map of size $T_1 \times F_i \times 2C_i \times 2$. Finally, we utilize a channel attention block (Hu, Shen, and Sun 2018) to give different weights to different channels, and then these features are fused into a new feature, denoted as $\mathcal{F}_i^c$, which has the same size as that of $\mathcal{F}_i^x$.

Decoder Module

The fused features at different levels are all sent to the decoder module via skip connection. Note that the features at the bottom level, *i.e.*, $\mathcal{F}_3^c$, further go thorough four residual blocks to enhance the fused features. The residual block is constructed by two convolutional layers with a short connection. The decoder module is symmetrical with the encoder module. It contains three deconvolution blocks, which includes deconvolution+BN+LeakyRelu layers, and the last layer outputs a complex mask to recover the enhanced STFT spectrogram directly. Similar to the encoder, the deconvolution layer only performs upsampling along the frequency dimension and the time dimension is invariant. Note that, to enlarge the receptive field along the time-dimension, all the (de)convolution layers in the Conv Block and Deconv Block are dilated convolution with dilation rate (s_F, s_T), which are setting to (1,1), (1,2), and (1,4) from top to bottom layers, respectively. The filter size $h_F \times h_T$ is set to 7×5, 6×5, and 6×5, respectively, where h_F (h_T) is the size along the frequency (time) dimension.

We utilize the phone-fortified perceptual loss (PFPL) proposed in (Hsieh et al. 2021), which is beneficial for perceptual evaluation, as our loss function. It is formulated as

$$\hat{y} = \text{ISTFT}(X_{rea} \cdot M_{rea} + jX_{img} \cdot M_{img}), \quad (5)$$

$$\mathcal{L} = ||y - \hat{y}||_1 + \text{sup}_{q \in Q} \mathbb{E}_\mu[q(e)] - \mathbb{E}_v[q(\hat{e})], \quad (6)$$

where M_{rea} and M_{img} represent the real and imaginary parts of the estimated complex mask M, "$\cdot$" is element-wise multiplication, and ISTFT represents the inverse STFT. $e = \Phi_{\text{wav2vec}}(y)$ and $\hat{e} = \Phi_{\text{wav2vec}}(\hat{y})$ are the features of the clean speech y and recovered speech $\hat{y}$. μ and v are the densities of e and v in the latent space. For more information, please refer to (Hsieh et al. 2021).

Two-Stage Strategy

Considering that the presence of noise severely degrades the MFCC matching precision, we further propose a two stage based matching and fusion scheme, where the second stage is used to improve the matching results with the first stage enhanced result as query input, and consequently improve the final enhancement results. We denote the enhancement result obtained in the first stage as $\hat{Y}^1$. We utilize it as the query input of the second stage MFCC patch matching, and realign the reference features in $\mathcal{F}_i^r$ according to the second stage matching index. For the feature fusion block, the soft weighting coefficients are calculated between the features of the new reference and $\hat{Y}^1$, and then the fused reference features are concatenated with the features of the noisy input and $\hat{Y}^1$. After going through the channel attention module, we obtain the fused features. Finally, these fused features go through the decoder module, which is the same as that in the first stage, to reconstruct the second stage enhancement result $\hat{Y}^2$.

Experiments

Datasets

Two benchmark datasets, *i.e.* DNS Challenge and Voice Bank+DEMAND, are used in our experiments.

Voice Bank + DEMAND Although Voice Bank + DEMAND dataset (Valentini-Botinhao et al. 2016) is not quite large, it is widely used in evaluating speech enhancement methods. The clean speeches in this dataset are from the the Voice Bank corpus (Veaux, Yamagishi, and King 2013) and the noise clips are from Diverse Environments Multichannel Acoustic Noise Database (DEMAND) (Thiemann, Ito, and Vincent 2013). There are totally 30 speakers in the clean speech: 28 are included in the training set and the other 2 are in the test set. For training, 40 different noise conditions are mixed with the clean speech from training set, generating 11572 noisy-clean speech pairs. For testing, 20 noise conditions are mixed with the clean speech from test set, generating 824 noisy-clean speech pairs. Neither speakers nor noise conditions in the test set exist in the training set.

DNS Challenge The Interspeech 2020 DNS challenge dataset (Reddy et al. 2020) is a large speech enhancement dataset. The clean speech are collected from Librivox and totally includes 500 hours utterances from 2150 speakers. The noise clips are from Audioset and Freesound, including 60000 noise clips with 150 classes. Following (Zheng et al. 2020), we synthesize 500 hours noisy clips with SNR levels of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB for training. For evaluation, we use another 150 noisy clips from the test set without reverberation. The testing SNR levels are randomly distributed in the range from 0 dB to 20 dB.

Training Details

The proposed method is implemented in Pytorch. The batch size is set to 32 and an Adam optimizer is used. The frequency bins F_i $(i = 1, 2, 3)$ is set to 128, 64, and 32, respectively. The channel number C_i $(i = 1, 2, 3)$ is set to 32, 64, and 128, respectively. All audios are resampled to 16kHz. STFT is calculated using Hann window with a window length of 512 samples, and the hop length is 128 samples. The FFT size is also set to 512 samples. The MFCC features are calculated with the same window length and hop length. In this way, the MFCC features and STFT spectrograms are frame-level matched along the time dimension. For VoiceBank-DEMAND, we train the first stage for 100 epochs and then we train the second stage for another 100 epochs with the first stage model fixed. The learning rate is initially set to $1e$-4, and reduced to $1e$-5 after 80 epochs for both stage training. Considering the training cost for two stage model, for DNS, we only present its one stage result which is generated by training for 50 epochs.

During training, we randomly select 1s noisy samples as the target input and construct a corresponding 15s reference speech with different contents but from the same speaker. Fortunately, Voice Bank+DEMAND includes multiple speakers and each speaker has multiple content-

Methods	PESQ-WB	CSIG	CBAK	COVL
Baseline (w/o ref.)	2.89	4.90	3.22	3.48
Baseline-two stage (w/o ref.)	3.00	3.89	3.20	3.43
w/o Patch-level matching	2.97	4.15	3.28	3.56
w/o k neighbors	3.04	3.94	3.31	3.49
w/o Feature warping	3.06	4.05	3.29	3.56
w/o Soft weighting	3.06	**4.20**	3.32	3.63
Ours-one stage	3.15	4.13	**3.39**	3.65
Ours-two stage	**3.19**	4.13	3.38	**3.66**

Table 1: Ablation Study on Voice Bank + DEMAND dataset.

different speeches. For Voice Bank+DEMAND, there are no overlapped contents in each noisy-reference pair. For DNS dataset, since its clean samples are randomly selected from Librivox, there may exist overlapped contents for some noisy-reference pairs in the training set. During test, we manually remove overlapped contents when constructing the reference speech. Besides, DNS testset (without reverberation) does not include the speaker identity information, we manually label the 150 clean speeches into 19 speakers.

Evaluation Metrics

We use the following metrics to evaluate the proposed method. For all these metrics, higher values mean better results.

- PESQ (Rec 2005): Perceptual evaluation of speech quality, using the wide-band version recommended in ITU-TP.862.2 (from -0.5 to 4.5).
- CSIG (Hu and Loizou 2007): Mean opinion score (MOS) prediction of the signal distortion (from 1 to 5).
- CBAK (Hu and Loizou 2007): MOS prediction of the intrusiveness of background noises (from 1 to 5).
- COVL (Hu and Loizou 2007): MOS prediction of the overall speech quality (from 1 to 5).
- STOI (Taal et al. 2010): Short-Time Objective Intelligibility.

Ablation Study

In this section, we evaluate the effectiveness of the proposed matching, warping, and fusion strategies by replacing them with other straightforward strategies respectively. Without specific clarification, all the variants are evaluated based on our one stage module. First, we give the results of our baseline network, namely an encoder-decoder structure with skip connections and channel attention module. There is no reference and fusion block in the baseline network. As shown in Table 1, the baseline result is much worse than our full model. Second, to evaluate the effectiveness of the proposed MFCC patch based matching strategy, we replace it with a single MFCC frame matching strategy, which will introduce discontinuity in aligned feature and make precise matching more difficult. It can be observed that the PESQ value for this variant is 0.18 less than our one stage result. We also evaluate the matching scheme by utilizing the first neighbor as the reference instead of using k neighbors as references. The result of this variant is also inferior to the one stage result on three metrics. Third, to evaluate the feature warping strategy, we give the result of directly warping on the STFT spectrograms. The aligned STFT spectrogram is generated by selecting matched STFT frames according to the MFCC patch matching index, and averaging the overlapped STFT frames. However, time discontinuity is introduced for this variant. Its result is worse than the one-stage result on all the four metrics. Fourth, to evaluate the proposed fusion strategy, we give the result without soft weighting. Its result is still inferior to our full model since different candidates have different contributions. Finally, we give our results using two-stage enhancement, which is slightly better than that of the one stage enhancement result because better matching leads to better enhancement results. Note that, for two stage scheme, our complete model outperforms the two-stage baseline result by 0.19 on PESQ-WB, which demonstrates the effectiveness of the proposed reference based speech enhancement strategy.

Methods	PESQ-WB	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
SEGAN	2.16	3.48	2.94	2.8
DFL	-	3.86	3.33	3.22
MetricGAN	2.86	3.99	3.18	3.42
CTS-Net	2.92	4.25	3.46	3.59
PHASEN	2.99	4.21	3.55	3.62
T-GSA	3.06	4.18	3.59	3.62
DEMUCS	3.07	**4.31**	3.40	3.63
MetricGAN+	3.15	4.14	3.16	3.64
PFPL	3.15	4.18	**3.60**	**3.67**
FAF-Baseline(w/o Ref.)	2.89	4.09	3.22	3.48
FAF-Net (Ours)	**3.19**	4.13	3.38	3.66

Table 2: Comparison with state-of-the-arts on Voice Bank + DEMAND dataset. The best results are highlighted in bold.

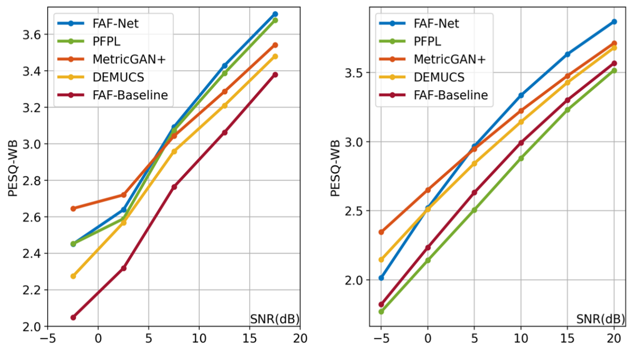
Methods	PESQ-WB	PESQ-NB	STOI
Noisy	1.58	2.45	91.52
NSNet	2.15	2.87	94.47
DTLN	-	3.04	94.76
Conv-TasNet	2.73	-	-
DCCRN-E	-	3.27	-
PoCoNet	2.75	-	-
FullSubNet	2.78	3.31	96.11
CTS-Net	**2.94**	3.42	**96.66**
FAF-Net(Ours)	2.93	**3.44**	96.37

Table 3: Comparison with state-of-the-arts on DNS dataset. The best results are highlighted in bold.

Comparison with State-of-the-arts

Since the competing methods for the two datasets are different, we introduce the compared methods according to the datasets on which they are evaluated.

DNS Challenge Dataset. For this dataset, we directly compare our one stage model with seven state-of-the-art methods. Among them, two are LSTM based methods, *i.e.,* FullSubNet (Hao et al. 2021) and DTLN (Westhausen and Meyer 2020), and two are U-Net based methods, *i.e.,* DCCRN (Hu et al. 2020) and PocoNet (Isik et al. 2020). Conv-TasNet (Koyama et al. 2020) is a time-domain network and NSNet (Xia et al. 2020) utilizes weighted speech distortion losses. CTS-Net (Li et al. 2021) is a two-stage complex spectral mapping method. The results of NSNet are quoted from



(a) (b)

Figure 4: Enhancement results with the inputs setting to different SNR levels. (a) Evaluated on the original VBD test set, (b) Rebuilding the VBD test set by re-adding noises.

FullSubNet, and others are directly quoted from their papers.

VoiceBank+DEMAND Dataset. For this dataset, we compare our two-stage model with eight state-of-the-art methods. There are three GAN based methods, *i.e.,* SEGAN (Pascual, Bonafonte, and Serra 2017), MetricGAN (Fu et al. 2019), and MetricGAN+ (Fu et al. 2021). PHASEN (Yin et al. 2020) is a two branch based network. DFL (Germain, Chen, and Koltun 2018) and PFPL (Hsieh et al. 2021) utilize perceptual losses. T-GSA (Kim, El-Khamy, and Lee 2020) is a complex-valued transformer network and CTS-Net (Li et al. 2021) mentioned above. The results for the eight methods are quoted from their papers.

For DNS dataset, as shown in Table 3, our method achieves the best result in terms of PESQ-NB, and is very close to the best method (CTS-Net) in terms of PESQ-WB.

Table 2 lists the comparison results on Voice Bank + DEMAND dataset. The proposed FAF-Net achieves the best results on this dataset in terms of PESQ-WB, outperforming the second best method PFPL, whose FLOPs are two times of our model. For CSIG and CBAK, our method is not the best. The main reason is that the introduced reference is beneficial for perceptual level reconstruction other than distortion level reconstruction since the reference is not similar with the ground truth in terms of SNR values. Note that, our baseline (w/o reference) is much worse than the compared methods indicating that the reference plays an important role in our framework. Fortunately, our feature alignment and fusion strategy can be plugged in other single speech processing baselines and our results can be further improved by using a better baseline.

Reference Evaluation

In this section, we evaluate the effectiveness of the references in three aspects on the Voice Bank + DEMAND (VBD) dataset.

First, we evaluate the enhancement results at different SNR values. The results in Fig. 4 (a) are produced on the original VBD test set. We classify its noise level (SNR ranging from -5 to 20 dB) into five groups and give the average PESQ result for each group. Since the speeches in different groups are different, we further generate a new VBD test set by adding noise with different SNR values to all the clean speeches in the test set. It can be observed that our scheme (FAF) generates the best results when the SNR values are larger than 5 dB for the two conditions. In addition, the proposed FAF-Net greatly outperforms FAF-Baseline, which does not utilize reference speech, for all the SNR ranges. The results of FAF-Net are worse than other methods for lower SNR values because our baseline network is much worse for lower SNR values. Note that, the result of PFPL in (b) is inferior to that in (a), which demonstrates that PFPL does not generalize well to new noise conditions.

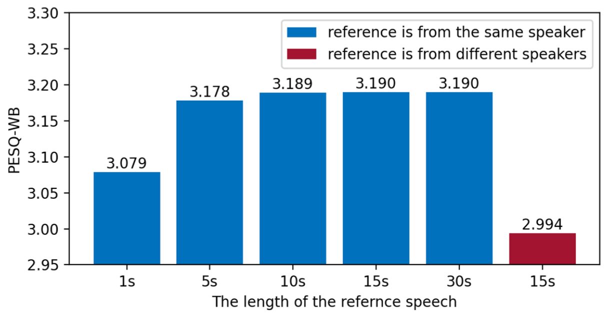


Figure 5: The speech enhancement results (in terms of PESQ) using references with different time lengths.

Second, we evaluate the enhancement results using references with different lengths. Generally, longer references generate better results. Considering the trade off between computing complexity and performance, we utilize the reference with 15s in our experiments. If we change the references to the speeches spoken by different speakers, the results will be heavily degraded. This indicates that high correlations between the reference and the noisy speeches are beneficial for the speech enhancement.

Third, we give results by utilizing the reference in the way of global embedding, similar to (Giri et al. 2021; Wang et al. 2018). Specifically, we extract the speaker's identity features from the reference speech with a pretrained speaker encoder (Jia et al. 2018), and then embed the global vector into our encoder features by concatenating along the feature dimension. The PESQ value of this variant is 2.97, which is worse than that obtained by feature alignment and fusion strategy. This demonstrates that utilizing the reference speech via exploring local correlations is better than exploring global correlations.

Conclusion

In this paper, we have proposed a novel RefSE method by exploring the frame-level correlations between the reference and noisy speeches. To solve the time discontinuity problem of the aligned speeches, we have proposed a MFCC patch matching and matching index based encoder feature warping strategy. The aligned reference feature and the noisy feature are fused together via soft attention and channel attention based fusion strategy. The ablation study and comparison with state-of-the-arts demonstrate the effectiveness of the proposed RefSE method. We believe our work will inspire more works for other reference based speech processing tasks, such as reference-based speech super-resolution and echo removal.

References

Choi, H.-S.; Kim, J.-H.; Huh, J.; Kim, A.; Ha, J.-W.; and Lee, K. 2018. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*.

Chuang, F.-K.; Wang, S.-S.; Hung, J.-w.; Tsao, Y.; and Fang, S.-H. 2019. Speaker-Aware Deep Denoising Autoencoder with Embedded Speaker Identity for Speech Enhancement. In *Interspeech*, 3173–3177.

Defossez, A.; Synnaeve, G.; and Adi, Y. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.

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

Fu, S.-W.; Liao, C.-F.; Tsao, Y.; and Lin, S.-D. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, 2031–2041. PMLR.

Fu, S.-W.; Yu, C.; Hsieh, T.-A.; Plantinga, P.; Ravanelli, M.; Lu, X.; and Tsao, Y. 2021. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. *arXiv preprint arXiv:2104.03538*.

Germain, F. G.; Chen, Q.; and Koltun, V. 2018. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*.

Giri, R.; Venkataramani, S.; Valin, J.-M.; Isik, U.; and Krishnaswamy, A. 2021. Personalized PercepNet: Real-time, Low-complexity Target Voice Separation and Enhancement. *arXiv preprint arXiv:2106.04129*.

Hao, X.; Su, X.; Horaud, R.; and Li, X. 2021. FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6633–6637. IEEE.

Hou, J.-C.; Wang, S.-S.; Lai, Y.-H.; Tsao, Y.; Chang, H.-W.; and Wang, H.-M. 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2): 117–128.

Hsieh, T.-A.; Yu, C.; Fu, S.-W.; Lu, X.; and Tsao, Y. 2021. Improving Perceptual Quality by Phone-Fortified Perceptual Loss using Wasserstein Distance for Speech Enhancement. *Proc. Interspeech 2021*, 196–200.

Hu, G.; and Wang, D. 2001. Speech segregation based on pitch tracking and amplitude modulation. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, 79–82. IEEE.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, Y.; Liu, Y.; Lv, S.; Xing, M.; Zhang, S.; Fu, Y.; Wu, J.; Zhang, B.; and Xie, L. 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*.

Hu, Y.; and Loizou, P. C. 2007. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1): 229–238.

Isik, U.; Giri, R.; Phansalkar, N.; Valin, J.-M.; Helwani, K.; and Krishnaswamy, A. 2020. Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. *arXiv preprint arXiv:2008.04470*.

Jia, Y.; Zhang, Y.; Weiss, R. J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I. L.; et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*.

Kim, J.; El-Khamy, M.; and Lee, J. 2020. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6649–6653. IEEE.

Koyama, Y.; Vuong, T.; Uhlich, S.; and Raj, B. 2020. Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks. *arXiv preprint arXiv:2005.11611*.

Li, A.; Liu, W.; Zheng, C.; Fan, C.; and Li, X. 2021. Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1829–1843.

Liu, Y.; Peng, X.; Xiong, Z.; and Lu, Y. 2021. Phoneme-Based Distribution Regularization for Speech Enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 726–730. IEEE.

Mun, S.; Choe, S.; Huh, J.; and Chung, J. S. 2020. The sound of my voice: Speaker representation loss for target voice separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7289–7293. IEEE.

Pascual, S.; Bonafonte, A.; and Serra, J. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Rec, I. 2005. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*.

Reddy, C. K.; Beyrami, E.; Dubey, H.; Gopal, V.; Cheng, R.; Cutler, R.; Matusevych, S.; Aichner, R.; Aazami, A.; Braun, S.; et al. 2020. The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework. *arXiv preprint arXiv:2001.08662*.

Rethage, D.; Pons, J.; and Serra, X. 2018. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5069–5073. IEEE.

Shon, S.; Tang, H.; and Glass, J. 2019. Voiceid loss: Speech enhancement for speaker verification. *arXiv preprint arXiv:1904.03601.*

Srinivasan, S.; Roman, N.; and Wang, D. 2006. Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 48(11): 1486–1501.

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, 4214–4217. IEEE.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, 035081. Acoustical Society of America.

Valentini-Botinhao, C.; Wang, X.; Takaki, S.; and Yamagishi, J. 2016. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. In *SSW*, 146–152.

Veaux, C.; Yamagishi, J.; and King, S. 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, 1–4. IEEE.

Wang, Q.; Muckenhirn, H.; Wilson, K.; Sridhar, P.; Wu, Z.; Hershey, J.; Saurous, R. A.; Weiss, R. J.; Jia, Y.; and Moreno, I. L. 2018. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. *arXiv preprint arXiv:1810.04826.*

Westhausen, N. L.; and Meyer, B. T. 2020. Dual-signal transformation lstm network for real-time noise suppression. *arXiv preprint arXiv:2005.07551.*

Williamson, D. S.; Wang, Y.; and Wang, D. 2015. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(3): 483–492.

Xia, Y.; Braun, S.; Reddy, C. K.; Dubey, H.; Cutler, R.; and Tashev, I. 2020. Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 871–875. IEEE.

Yan, X.; Zhao, W.; Yuan, K.; Zhang, R.; Li, Z.; and Cui, S. 2020. Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 52–68. Springer.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5791–5800.

Yin, D.; Luo, C.; Xiong, Z.; and Zeng, W. 2020. PHASEN: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9458–9465.

Yue, H.; Liu, J.; Yang, J.; Sun, X.; Nguyen, T. Q.; and Wu, F. 2019. Ienet: Internal and external patch matching convnet for web image guided denoising. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 3928–3942.

Yue, H.; Sun, X.; Yang, J.; and Wu, F. 2013. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 22(12): 4865–4878.

Yue, H.; Sun, X.; Yang, J.; and Wu, F. 2015. Image denoising by exploring external and internal correlations. *IEEE Transactions on Image Processing*, 24(6): 1967–1982.

Zhang, Z.; Wang, Z.; Lin, Z.; and Qi, H. 2019. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7982–7991.

Zheng, C.; Peng, X.; Zhang, Y.; Srinivasan, S.; and Lu, Y. 2020. Interactive speech and noise modeling for speech enhancement. *arXiv preprint arXiv:2012.09408.*