Dual-channel Speech Enhancement Using Neural Network Adaptive Beamforming

Tao Jiang[1], Hongqing Liu[1], Chenhao Shuai[1], Mingtian Wang[1], Yi Zhou[1], and Lu Gan[2]

[1] School of Communication and Information Engineering Chongqing University of Posts and Telecommunications Chongqing, China

[2] College of Engineering, Design and Physical Science, Brunel University. London UB8 3PH, U.K.

s190101065@stu.cqupt.edu.cn, hongqingliu@outlook.com

Abstract. Dual-channel speech enhancement based on traditional beamforming is difficult to effectively suppress noise. In recent years, it is promising to replace beamforming with a neural network that learns spectral characteristic. This paper proposes a neural network adaptive beamforming end-to-end dual-channel model for speech enhancement task. First, the LSTM layer is used to directly process the original speech waveform to estimate the time-domain beamforming filter coefficients of each channel and convolve and sum it with the input speech. Second, we modified a fully-convolutional time-domain audio separation network (Conv-TasNet) into a network suitable for speech enhancement which is called Denoising-TasNet to further enhance the output of the beamforming. The experimental results show that the proposed method is better than convolutional recurrent network (CRN) model and several popular noise reduction methods.

Keywords: Neural network · Dual-channel · Speech enhancement.

1 Introduction

Speech enhancement algorithm is an important technology to process speech in noisy environments to improve speech quality and intelligibility [1]. In terms of number of channels available, speech enhancements are divided into multi-channel and single-channel cases. Because multi-channel methods exploit spatial information, their performances are often better than that of the singlechannel methods. Beamforming is often used in multi-channel speech enhancement.

The traditional beamformer needs a set of beamforming filters that are convolved with signals from each channel before summation. The filters are designed to enhance speech and suppress interfrence noise. The famous traditional beamforming approaches include minimum variance distortionless response (MVDR) [2] and its generalized sidelobe canceller (GSC) formulation [3], which minimizes the energy of the signal at the beamformer output. The beamformer generally requires a steering vector that points beamformer to the target

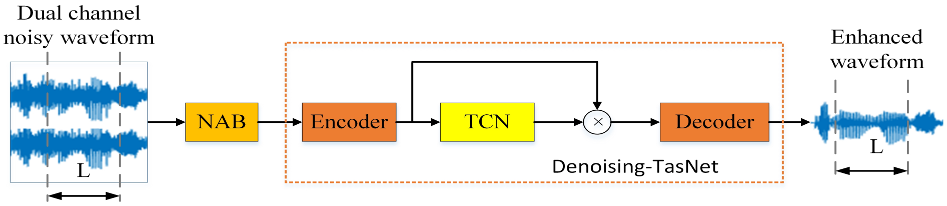

Fig. 1: NABDTN block diagram.

direction [4]. Generally speaking, the method of estimating the steering vector is direction of arrival (DOA) estimation [5]. However, the performance of the DOA algorithms is usually affected in complex acoustic environments, resulting in an inaccurate steering vector estimation, which in turn degrades the performance of beamforming.

In recent years, neural networks have been used to replace traditional beamforming methods and the results are very promising. It not only makes up for the shortcomings of beamforming that cannot be applied to complex acoustic environments, but also has a better performance than beamforming. The model based on neural network beamforming has achieved good results in multi-channel speech separation [6, 7] and speech recognition [8, 9]. Recently, beamforming based on DNNs [10] has been applied in the field of multi-channel speech enhancement. The amplitude spectrum of the short-time Fourier transform of the multi-channel speech signal is used as the input of the neural network, and the beamforming filter coefficients are calculated by estimating the covariance matrix of the target signal and the noise. The realization principle is complicated, and a simple end-to-end system is of importance.

In this work, we propose adaptive neural network beamforming (NAB) and Denoising-TasNet model (NABDTN) for speech enhancement in Fig.1. The purpose of NAB is to estimate a series of beamforming spatial-temporal filters, and then they are respectively convolved the speech signals of microphones before summing. This network mimics delay-and-sum (DS) technology [11]. However, NAB does not need to estimate the time delay of each microphone, which is difficult to accurately estimate in a complex acoustic environment. After that, the Denoising-TasNet network is developed to further enhance the performance by minimizing scale-invariant signal-to-distortion ratio (SI-SDR). This means that the time delay can be learned implicitly by the neural network. Therefore, the model has a strong applicability for different complex acoustic scenarios. The temporal convolutional network (TCN) in Denoising-TasNet has strong time-domain sequence modeling capability [12], so we use this network to further enhance the NAB output to generate clean speech.

2 Algorithm Description

2.1 Problem formulation

Let $x_c(k)$, $s_c(k)$ and $n_c(k)$ denote noisy speech, clean speech and background noise, respectively, and $c \in \{0, 1, \cdots, C-1\}$ is the index of the channel, where C is the number of channels. It should be noted that $c = 0$ is the reference channel, and its corresponding clean speech is the label during network training. The noisy speech is written as

$$x_0(k) = s_0(k) + n_0(k), \tag{1}$$

$$x_c(k) = s_c(k) + n_c(k) = s_0(k) * h_{0c}(k) + n_c(k), \tag{2}$$

where $s_0(k)$ denotes target clean speech which can be divided into overlapping segments of length L, and $k \in \{0, 1, ..., T\}$ denotes the total number of segments in the input, and * denotes the convolution operation. The impulse response of clean speech between channels is represented by $h_{0c}(k)$. In a short-distance call scenario, the distance between the main microphone and the mouth is small, and we take the clean speech on the main channel as the target clean speech.

The general finite impulse response (FIR) filter beamformer is

$$y(t) = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c[n] x_c[t - n - \tau_c], \tag{3}$$

where $h_c[n]$ is the n-th tap of the beamforming filter related to microphone c, $x_c[t]$ is the noisy speech received by microphone c at time t, and τ_c is the alignment steering delay between the speech received by another microphone and the speech received by the reference microphone, $y(t)$ is the output of beamforming speech, and N is the length of the filter.

Since the target speech arrives at each microphone at a different time, the speech in each microphone needs to be aligned with the speech of the reference microphone to perform traditional beamforming. The estimation accuracy of steering delay τ_c presents a great influence on the performance of speech enhancement. However, the proposed NABDTN model estimates the filter coefficients by minimizing the loss function of the enhanced speech and the clean speech. Therefore, steering delay estimation of τ_c is hidden in the estimated filter coefficients. The output result of the kth frame of the NAB model is

$$y(k)[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c(k)[n] x_c(k)[t - n], \tag{4}$$

where $h_c(k)[t]$ is the estimated filter coefficient of channel c in the kth frame. To estimate $h_c(k)[t]$, we jointly train the entire NABDTN network to predict the filter coefficients of each channel. The beamforming filter coefficients can be continuously adjusted according to the dataset during the training process. The more complex the dataset, the wider the adaptability of the model. This is similar to traditional adaptive beamforming, which adjusts filter coefficients according to changes in data.

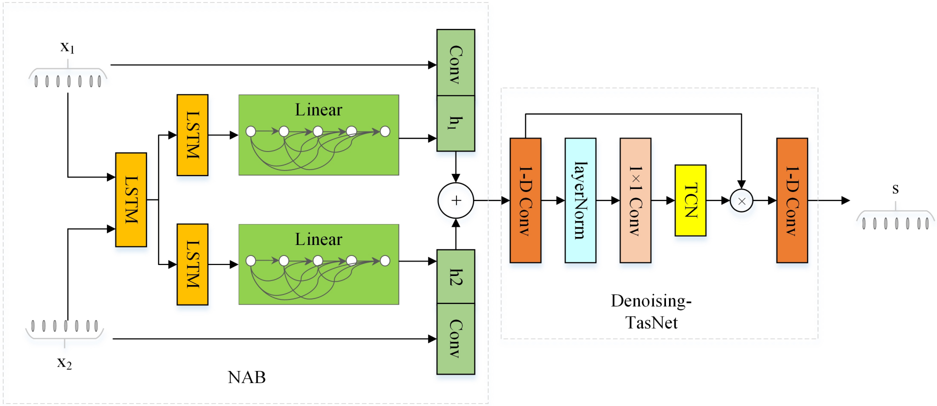

Fig. 2: Illustration of the NABDTN architecture.

2.2 NABDTN architecture

The frame structure of NABDTN containing two parts is shown in Fig.2, where two-channel case is illustrated. The NAB model consists of three LSTMs and two linear layers, where one 512-cell LSTM layer and two 256-cell LSTM layers are utilized. In our experiments, the segment length of the input noisy speech $L = 16100$, and we set the output of the linear layer as $N = 26$, which is the length of the filter coefficient. The hyperparameters are determined by measuring the PESQ metric of the trained model. To convolve the noisy speech of each frame with the filter coefficients, we use a kernel size of 1×26 to implement the convolution of $x_c(k)[t]$ and $h_c(k)[t]$ to efficiently calculate the convolution operation of multiple batches of data. It should be noted that there are two channels of noisy speech in each frame. We respectively input the speech of each channel into the 512-cell LSTM, and the two outputs are generated as the inputs for two 256-cell LSTMs.

For the model of the Denoising-TasNet, we employ the similar structure of Conv-TasNet [12], which is the encoder/TCN/decoder pipeline. The amplitude of each frequency in the spectrogram reflects the energy of the frequency component in the frequency domain processing. However, the amplitude of a sample point in the time domain does not provide much information, and it needs to be combined with adjacent samples to represent specific frequency components [13]. For example, if it is a high-frequency signal, and then in time, the amplitude of adjacent samples will vary greatly, and vice versa. Therefore, when processing time-domain waveforms, in order to allow the neural network to better learn the features provided by adjacent samples, we use one-dimensional (1-D) convolutional blocks with dilated convolution factors in the TCN network. As shown in Fig.3, we repeat X 1-D convolution blocks with the dilated convolution factor $d = \{1, 2, 4, ..., 2^{M-1}\}$ for R times, and the size of kernel is P. In addition, TCN estimation is no longer the original speech separation mask, but a speech enhancement mask so that the model has only one enhanced speech output. In order to avoid gradient exploding or vanish, each 1-D convolutional block in

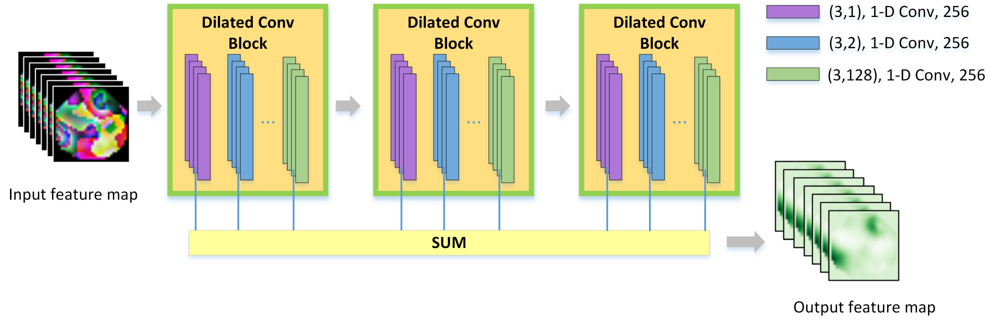

Fig. 3: Architecture of TCN. The parameters are denoted as follows: (p1,p2) Conv p3, where p1 is the kernel size, p2 is the dilated rate, and p3 is the number of filters.

TCN normalizes the input features. The normalization process is

$$gLN(F) = \frac{F - E[F]}{\sqrt{Var[F] + \epsilon}} \bigodot \gamma + \beta, \tag{5}$$

$$E[F] = \frac{1}{NT} \sum_{NT} F, \tag{6}$$

$$Var[F] = \frac{1}{NT} \sum_{NT} (F - E[F])^2, \tag{7}$$

where $F \in \mathbb{R}^{K \times T}$ is the input feature, ϵ is a small constant to prevent the denominator from being zero, $\gamma, \beta \in \mathbb{R}^{K \times 1}$ are trainable parameters, $E[F]$ and $Var[F]$ are the mean and variance of F, respectively, and $\bigodot$ denotes element-wise multiplication.

2.3 Loss functions

The studies show that, for speech enhancement in the time domain, the SI-SDR as the loss function presents a superior performance, and it is defined by [14]

$$\text{SI-SDR} = 10 \log_{10} \frac{\|\beta s\|^2}{\|\beta s - \hat{s}\|^2}, \tag{8}$$

$$\beta = \frac{\hat{s}^T s}{\|s\|^2} = \arg\min_{\beta} \|\beta s - \hat{s}\|^2, \tag{9}$$

where $s \in \mathbb{R}^{1 \times T}$and $\hat{s} \in \mathbb{R}^{1 \times T}$ are the clean speech and enhanced speech, respectively. The scaling of the target speech s ensures that the SI-SDR measure is invariant to the scale of $\hat{s}$. It is seen from Eqs. (8) and Eqs. (9), that SI-SDR is simply the signal-to-noise(SNR) ratio between the weighted clean speech signal defined as $\|\beta s\|^2$ and the residual noise defined as$\|\beta s - \hat{s}\|^2$. The network maximizes the correlation between s and $\hat{s}$ by maximizing SI-SDR, so that the model obtains a higher SNR.

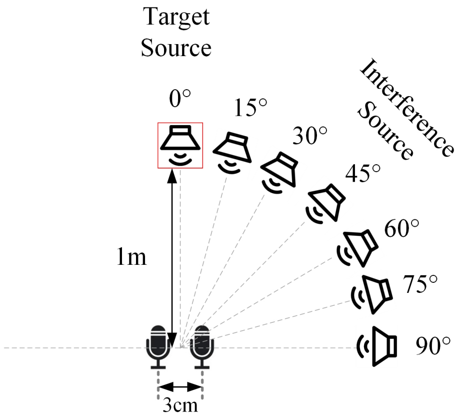

Fig. 4: Experimental settings.

3 Experimental Results

3.1 Setups

Fig.4 shows the position relationship between the microphone array and the target sound source and interference source in our experiments. In this experiment, the number of microphones is $C = 2$. The location of the target sound source is 1 m in the front of the microphone array that has a space interval of 3 cm. Based on this setting, the azimuth angle of our target source is 0 degree. We set an interference source at a distance of 1 m from the microphone array. To cover all the angles, the direction angle of interference ranges from -90 degree to 90 degree with an interval of 15 degree, and the similar setting is also utilized in [15].

To generate the dataset, we use 10,000 speech and 9,100 noise utterances from the MS-SNSD dataset, in which each speech and noise durations are 31s and 10s, respectively, and the sampling rate is 16 kHz. We randomly select 8500 speech and 8000 noise utterances from the dataset to generate a 200h training set, and then select 1500 speech and 1100 noise utterances from the remaining dataset to generate a 20h validation set, and finally the cafeteria, street and babble noises are used to generate test set. The image method [16] is used to generate a dual-channel dataset. In order to test the denoising ability of the NABDTN model and reduce the impact of the room impulse response, we use the clean speech generated by the reference microphone as our target source, and the room size is $10m \times 7m \times 3m$.

For interference sources with different azimuth angles, we also have produced training and validation sets with different signal-to-noise ratios (SNRs) ranging from -5 dB to 5 dB at an interval of 1 dB, and produced a test set of {-5dB, 0dB, 5dB}. Our training set includes 120000 mixtures that are segmented into 16100 sample points as the input of the model.

3.2 Results

In the experiments, we use two objective metrics of short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) [14] to measure the denoising performance of different models. In addition, we also use the subjective metric of deep noise suppression mean opinion score (DNSMOS) [17] to evaluate enhanced speech. We randomly select 180 speech and noise files from the test set to average the results. The interference sources of the test dataset produced by the image method include three directions of 15°, 45°, and 90° and gradually deviate from the target source, so as to avoid the same distribution with the training dataset. The STOI and PESQ of the noisy and enhanced speeches at -5 dB, 0 dB, and 5 dB are measured , respectively. As shown in Table 1, compared with the MMSE-based approch [18], dual-microphone DNN speech enhancement [19], CRN model [20] and MFMVDR model [21] the proposed method significantly improve the performance. For example, at SNR = 0 dB, the NABDTN method increased STOI by 12.7% and PESQ by 1.08, whereas the MFMVDR only improved STOI by 10.42% and PESQ by 1.02.

Table 1: STOIs and PESQs of different methods. The value is an average of 15°, 45°, and 90° interference sources.

method	STOI(%)			PESQ		
SNR	-5dB	0dB	5dB	-5dB	0dB	5dB
noisy	71.13	80.10	87.80	1.48	1.81	2.19
MMSE	70.54	79.13	84.62	1.48	1.90	2.27
DNN	79.80	84.67	90.21	2.03	2.45	2.74
CRN	81.02	89.06	91.87	2.23	2.62	2.85
MFMVDR	84.73	90.52	93.34	2.45	2.83	3.08
Prop.	**88.13**	**92.89**	**96.32**	**2.58**	**2.89**	**3.18**

As shown in Fig.5, we choose a section of noisy containing stationary noise and non-stationary noise to analyze the noise reduction performance from the spectrogram. Comparing the noisy and enhanced spectrograms, it can be found that the stationary noise and non-stationary noise marked by the red rectangular box have been removed. From the blue and white rectangles marked by enhanced and clean spectrograms, there are only slight distortions in the unvoiced audio at low and high frequencies, while the voiced distortion is even smaller.

In Table 1, the results are an average of interferences at three directions. However, interference at each angle may contribute to the system performance differently. To demonstrate this effect, we also listed the denoised performance of the NABDTN model at different azimuths. As shown in Table 2, the SNRs of the mixed speech in each direction include -5 dB, 0 dB, and 5 dB, which indicates STOI and PESQ are an average of different SNRs. We found that system performance increases first and then decreases as the azimuth of the interference source

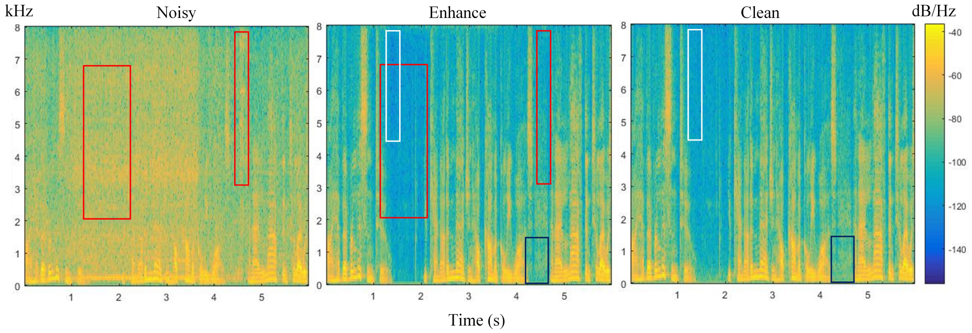

Fig. 5: Example of noise suppression for stationary noise and non-stationary noise at 5 dB SNR. The ordinate and abscissa represent frequency and time respectively. The color bar on the right represents power/frequency.

Table 2: The performance of the proposed method when interference sources at different directions, where STOI and PESQ are the average of -5 dB, 0 dB, and 5 dB.

metrics	Interference direction					
	15°	30°	45°	60°	75°	90°
noisy STOI	80.52					
noisy PESQ	1.86					
noisy DNSMOS	2.76					
STOI(%)	91.35	94.86	98.69	95.64	96.82	92.65
PESQ	2.67	3.15	3.68	3.20	3.35	2.74
DNSMOS	3.12	3.23	3.50	3.34	3.51	2.90

varies. Nonetheless, the average PESQ and STOI of noisy have been greatly improved, which means the NABDTN model is not greatly affected by the azimuth of the interference source. Therefore, the proposed model is more adaptable to complex acoustic scenes. In addition, since the objective measurement indicators STOI and PESQ cannot fully reflect the quality of human auditory perception, we also use DNSMOS provided by DNS-Challenge as a performance measure as shown in the last row of Table 2.

4 Conclusion

In this study, we proposed a neural network adaptive beamforming (NAB) and developed Conv-TasNet structure for dual-channel speech noise reduction. The Conv-TasNet was originally tasked of processing time-domain speech separation, but we produced a denoising mask instead to achieve single-channel speech denoising. The NAB takes multiple inputs to perform beamforming in suppressing the interferences, and after that, Denoising-TasNet is utilized to finally output the denoised signals. Experimental results show that our proposed method is superior to DNN and recently proposed CRN model. Additionally, we found that the NABDTN model is resistant to interferences from different directions.

References

1. Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "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)*, 2020, pp. 871–875.
2. B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.
3. O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on signal processing*, vol. 47, no. 10, pp. 2677–2684, 1999.
4. L. Pfeifenberger, M. Zohrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2162–2172, 2019.
5. L. Pfeifenberger and F. Pernkopf, "Blind source extraction based on a direction-dependent a-priori snr," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
6. Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.
7. B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech 2016*, 2016, pp. 1976–1980.
8. T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.
9. Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 260–267.
10. Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.
11. J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing.* Springer Science & Business Media, 2008, vol. 1.
12. Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
13. S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).* IEEE, 2017, pp. 006–012.

14. M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.
15. N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder." in *INTERSPEECH*, 2019, pp. 86–90.
16. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
17. C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv e-prints*, pp. arXiv–2010, 2020.
18. R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4266–4269.
19. I. López-Espejo, J. A. González, Á. M. Gómez, and A. M. Peinado, "A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: application to noise-robust speech recognition," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014, pp. 119–128.
20. K. Tan, X. Zhang, and D. L. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
21. M. Tammen and S. Doclo, "Deep multi-frame mvdr filtering for single-microphone speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8443–8447.