Interspeech 2024
1-5 September 2024, Kos, Greece

DeWinder: Single-Channel Wind Noise Reduction using Ultrasound Sensing

Kuang Yuan[1*], *Shuo Han*[1*], *Swarun Kumar*[1], *Bhiksha Raj*[1]

[1]Carnegie Mellon University, United States

kuangy@cmu.edu, shuohan@andrew.cmu.edu, swarun@cmu.edu, bhiksha@cs.cmu.edu

Abstract

The quality of audio recordings in outdoor environments is often degraded by the presence of wind. Mitigating the impact of wind noise on the perceptual quality of single-channel speech remains a significant challenge due to its non-stationary characteristics. Prior work in noise suppression treats wind noise as a general background noise without explicit modeling of its characteristics. In this paper, we leverage ultrasound as an auxiliary modality to explicitly sense the airflow and characterize the wind noise. We propose a multi-modal deep-learning framework to fuse the ultrasonic Doppler features and speech signals for wind noise reduction. Our results show that *DeWinder* can significantly improve the noise reduction capabilities of state-of-the-art speech enhancement models.

Index Terms: Speech Enhancement, Sensor Fusion

1. Introduction

Wind noise is a major noise source that degrades the audio recording quality in various environments with the presence of airflow. Different from other noise sources, wind noise is generated by turbulent airflow hitting the microphone membrane instead of propagating acoustic waves [1]. Due to the non-stationary nature of turbulence, suppressing wind noise in audio remains an open challenge. A microphone windshield is used to reduce the excessive pressure from wind in professional use cases. However, such hardware solutions are not suitable for tiny microphones on embedded devices like smartphones.

To mitigate the interference of wind noise, a series of prior works [2, 3, 4, 5, 6] propose wind noise estimation and reduction algorithms relying on multi-channel microphone arrays, leveraging the assumption that wind noise is spatially uncorrelated. In contrast, for single-channel audio, spectral subtraction [7, 8, 9] and filtering [10, 11, 12] algorithms are proposed to reduce wind noise. However, such techniques are based on the stationary assumption when estimating the noise spectral distribution, the performance of which can drop significantly under real-world non-stationary wind noise. Recent speech enhancement techniques [13, 14, 15, 16, 17, 18] based on Deep Neural Networks (DNN) have shown promise in removing undesire noise in audio. However, existing speech enhancement models only treat wind noise as a general background noise without explicitly modeling its characteristics, which may cause sub-optimal performance, especially in strong wind environments with low Signal-to-Noise Ratio (SNR).

Instead of relying on pre-estimated noise distribution, we propose incorporating a new modality to sense and characterize the real-time airflow profile, and further enable more informative wind noise reduction. Specifically, for the first time, we propose to utilize ultrasound as a complementary modality to gather information about wind noise. Our design is based on a key observation: the ambient airflow not only introduces wind noise when hitting the microphone but also shapes the propagation of other acoustic signals in the air. Intuitively, if an acoustic wave is in the same direction as the airflow, it will travel faster than its original speed, such that its frequency becomes higher at the receiver because of the Doppler effect [19]. More generally, the turbulent airflow that induces wind noise contains many unsteady vortexes moving toward different directions, which shapes the frequency of the acoustic signals in a more unstructured way. We propose to use high-frequency ultrasound signals to sense and characterize the airflow, by capturing such frequency differences caused by the Doppler effect with finer granularity than audible signals, without inducing any audible disturbances.

Specifically, we implement the idea of *DeWinder* by using an ultrasound speaker co-located with the microphone. The ultrasound speaker transmits a tone signal at around 20 kHz, which is not audible but can still be captured by commodity microphones sampling at 44.1 kHz. As illustrated in Fig. 1, the signal transmitted from the speaker (Orange) would be shaped by airflow near the device before arriving at the microphone. The microphone would simultaneously capture the wind noise (Blue), as well as the ultrasound signal (Yellow) carrying information about the real-time airflow profile. We note that our proposed hardware setup can be easily extended to off-the-shelf IoT devices with co-located speakers and microphones [20, 21], such as smartphones (located at the bottom of the screen).

To implement *DeWinder*, we design a modular framework that can be adapted to different existing speech enhancement DNN models to improve the wind noise reduction performance through speech-ultrasound multi-modal fusion. As shown in Fig. 3, we first develop an ultrasound feature extraction pipeline consisting of demodulation and multi-step filtering, to convert ultrasound signal into baseband waveforms that focus on the Doppler effects induced by airflow. We demonstrate the performance of *DeWinder* by adapting the architecture of two state-of-the-art speech enhancement models: DEMUCS [14] and DCCRN [13]. Both of the networks have a convolutional encoder-decoder architecture with a unidirectional LSTM in-between for sequence modeling. For both of the architectures, we design another convolution encoder specialized for ultrasound encoding, and fuse the extracted ultrasound embedding with the speech embedding before the LSTM through a customized fusion module. While the two models process the speech in different domains (DEMUCS - waveform domain, DCCRN - spectrogram domain), our results show that the *DeWinder*'s fusion framework can improve the wind noise reduction performance in both domains, especially in low-SNR conditions.

[*]Equal Contribution

10.21437/Interspeech.2024-2180

Figure 1: DeWinder *uses ultrasound to sense and reduce wind noise.*

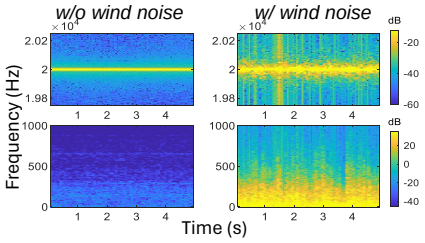

Figure 2: *The airflow induces wind noise, while shaping the ultrasound transmission.*

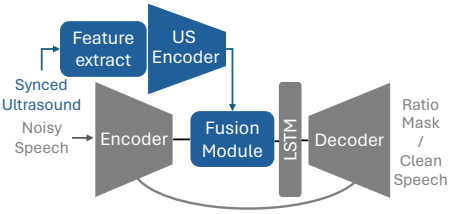

Figure 3: *Modular Design that can be adapted to existing speech enhancement models.*

2. Dataset Collection and Processing

2.1. Dataset Collection

As visualized in Fig. 1, the ultrasound speaker is placed co-located with the microphone, and towards the same direction. Specifically, we place two ultrasound speakers at the left and right side of the Rode M5 microphone, transmitting tones at 20 and 21 kHz respectively, to ensure the sensitivity of the system is symmetric to the wind from different azimuth directions (We also present the evaluation results with only a single speaker in Sec. 4). The microphone samples at 44.1 kHz and is connected to a laptop through an RME Babyface Pro sound card .

To generate consistently powerful enough airflow that can induce a significant amount of wind noise to the microphone, we use two types of high-velocity fans to generate wind. We collect the dataset in three different indoor environments to enrich the diversity of turbulence. We further place the fan along four different directions relative to the microphone in each environment at around 2-5 m away from the microphone, and allow it blow wind towards the microphone. The average wind speed in front of the microphone membrane is 1-2.5 m/s across different environments and distances, as measured by an anemometer. We collect the dataset of wind noise along with ultrasound signal transmissions for 4.2 hours, and further synthesize a 10.3-hour dataset by mixing it with clean speech utterances (detailed in Sec. 4.1).

2.2. Preprocessing and Feature Extraction

During data collection, both wind noise (mostly $<$1 kHz) and ultrasound signals (around 20 and 21 kHz) are captured by the microphone. To obtain the noise independently, we first apply a low-pass filter to extract the signal under 1.2 kHz. We further apply a high-pass filter with a cutoff frequency of 20 Hz to suppress the inaudible low-frequency artifacts caused by the wind. Finally, we resample the noise signal to 16 kHz to let it match with the speech dataset and meet the input requirements of most speech enhancement models [14, 13, 16].

For ultrasound processing, our primary goal is to convert high-frequency signals into representations that can be handled by common CNN encoders, as well as capture the frequency differences induced by the Doppler effect. We choose to down-convert the ultrasound signals into baseband waveforms by mixing them with sine and cosine waves at the corresponding center frequencies (20 or 21 kHz) and applying low-pass filtering. Through further resampling to 16 kHz, such baseband waveforms can be processed by the encoders of speech enhancement models and retain the characteristics of sample-level synchronization with noise. As a normal airflow with a wind speed of less than 8 m/s will only introduce a Doppler shift of less than 500 Hz to the 20 kHz ultrasound, we choose the low-pass filter with a cutoff frequency of 500 Hz, to let the signal focus on the features produced by the Doppler effect, as well as remove the high-frequency components introduced by mixing [22]. We further apply a highpass filter at a 10 Hz cutoff frequency to mitigate the signals reflected from nearby static objects. Through the above processing, each ultrasound signal is converted into two channels of baseband waveforms (mixed through sine and cosine waves). Thus, along with a single-channel wind noise, we obtain four channels of sample-level synchronized waveforms that help to characterize the wind noise.

3. Model

Instead of building an architecture from scratch, we design *DeWinder* as a modular framework that can be adapted to different existing speech enhancement models. We demonstrate *DeWinder* on DEMUCS and DCCRN. Both models have a convolutional encoder-decoder architecture with U-Net skip connections [23] and an LSTM in between. As shown in Fig. 3, we design another branch of an ultrasound encoder in parallel with the speech encoder, that processes the multi-channel waveforms (extracted from ultrasound, Sec. 2) into embeddings that characterize wind noise. We now note a key difference between the embeddings generated by DCCRN and DEMUCS that informs *DeWinder*'s fusion design. DEMUCS captures a speech embedding that is designed to generate clean speech output. In contrast, DCCRN's embedding captures the noise rather than the speech, to estimate a ratio mask output that helps separate noise from speech. Given that the two model embeddings have entirely different semantics, the fusion module *DeWinder* employs prior to LSTM for DEMUCS and DCCRN are customized accordingly. We detail *DeWinder*'s ultrasound encoder and fusion module for DEMUCS and DCCRN respectively below.

3.1. Dewinder - DEMUCS

DEMUCS [14] is a speech enhancement model that operates on the waveform domain. DEMUCS extracts the speech embeddings through an encoder from noisy speech and seeks to reconstruct clean speech at the decoder using the embeddings. For the ultrasound encoder, we follow a similar architectural design approach as the original speech encoder in DEMUCS. As the waveforms extracted from ultrasound have a narrower frequency range ($<$500 Hz) than speech, we reduce the number of CNN layers in the encoder to three to mitigate overfitting. Additionally, we set the base hidden channel size H to 24 instead of the default setup of 48, and the convolutional kernel size K to 10. We ensure that both the speech and ultrasound encoders maintain temporal alignment during convolutions, enabling coherent integration of the two modalities. We note that the input of the ultrasound encoder is the four-channel waveforms extracted from ultrasound, instead of a single channel.

The encoder of DEMUCS seeks to extract embeddings rep-

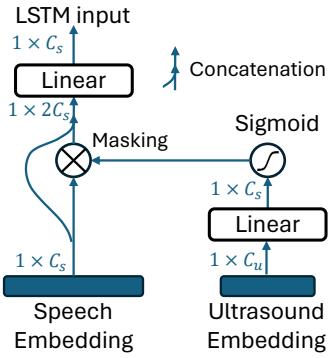

Figure 4: *Fusion Module for DEMUCS based on Masking*

resenting clean speech from noisy speech, while inevitably retaining noise information in the embedding for low-SNR inputs. Thus, to further mitigate the wind noise corruption, we design a fusion module based on masking that leverages the ultrasound embedding to *filter* out the wind noise information in the speech embedding, before feeding it into the LSTM. As illustrated in Fig. 4, we first apply a linear layer to the ultrasound embedding to let its dimension match with speech embedding. We then apply an element-wise Sigmoid activation and multiply it with the speech embedding, which is essentially an embedding mask to suppress the wind noise information. Mathematically,

$$\boldsymbol{X}_s' = \boldsymbol{X}_s \cdot \sigma(\boldsymbol{X}_u \boldsymbol{W}^T + \boldsymbol{b}) \tag{1}$$

where $\boldsymbol{X}_s, \boldsymbol{X}_u$ are the speech and ultrasound embedding respectively with dimensions illustrated in the figure. $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weights and bias of the linear layer, and $\boldsymbol{X}_s'$ is the denoised speech embedding. We further concatenate the denoised embedding with the original speech embedding and apply another linear layer to project the dimension back to the original input dimension of the LSTM.

3.2. Dewinder - DCCRN

Deep Complex Convolution Recurrent Network (DCCRN) is a speech enhancement model that operates on the Time-Frequency (TF) Domain. The encoder extracts the noise information from the complex-value spectrogram into embeddings. The LSTM and decoder process the embeddings and estimate a Complex Ratio Mask (CRM) [24], which can be applied to the original spectrogram for noise reduction.

Similarly, we design an encoder that extracts the wind noise information from ultrasound inputs by adapting the design of its original encoder. In the DCCRN-CL configuration, six layers of CNN have numbers of channels of {16,32,64,128,256,256}, with kernel size and stride of (5,2) and (2,1) respectively. To reduce the parameter size, we choose a five-layer design with fewer output channels of {16,32,64,128,128} to mitigate overfitting. To ensure time-window synchronization of the extracted embedding with the embedding extracted from speech, we modify the stride of the first layer of ultrasound encoder into (4,1).

In the fusion module, as the embeddings from two encoders are both designed to represent the wind noise information, we choose to combine the information from the two modalities by concatenating the embeddings. Specifically, the complex-value embedding from the speech encoder can be represented by $\boldsymbol{X} = \boldsymbol{X}_r + j\boldsymbol{X}_i$, where $\boldsymbol{X}_r, \boldsymbol{X}_i \in \mathbb{R}^{C_s \times F \times T}$, and ultrasound embedding can be represented by $\boldsymbol{Y} = \boldsymbol{Y}_r + j\boldsymbol{Y}_i$, where $\boldsymbol{Y}_r, \boldsymbol{Y}_i \in \mathbb{R}^{C_u \times F \times T}$ and $j = \sqrt{-1}$. C_s, C_u are the channel dimensions of speech embedding and ultrasound embeddings respectively. F, T are the frequency and time dimensions. We concatenate the real part and imaginary part of the two embeddings along the channel axis, before passing these into the complex LSTM.

As the first paper exploring using ultrasound for wind noise reduction, *DeWinder* shows the modular design by adapting two speech enhancement models, DCCRN and DEMUCS. The two models process the speech in different domains as well as have different forms of decoder outputs. We envision our designs have a great potential to be generalized to other speech enhancement models that share similar structures.

4. Experiment

4.1. Datasets Synthesization

In our experimental setup, we train and evaluate the proposed methods and baseline models on synthesized noisy datasets. We collect 4.2 hours of wind noise dataset along with ultrasound transmission. We randomly split the audio into 3.36 hours and 0.84 hours respectively for training and validation, ensuring no overlapping. In each set, we extract 5-second audio segments in a sliding window manner with a hop size of 2 seconds to augment the size of the dataset. We then randomly select 7420 clean utterances (each with 5 seconds) in total from LibriSpeech dataset [25] and mix them with wind noise at random low SNR between -40 dB and -20 dB. We also ensure that there is no overlap in speaker identities between the clean utterances used for training and validation. During mixing, we fix the power of wind noise to ensure it still physically matches with ultrasound, and tune the power of speech signal to generate mixed speech at different SNR. We note that the SNR values we used for synthesizing are lower than the ones used in other prior speech enhancement works [13, 14], mainly because we consider wind noise as the only noise source. We choose the range of SNR at -40 to -20 dB as it perceptually matches the audio recordings quality in real-world outdoor windy environments during our initial subjective testing.

Finally, we get a training and validation set with 5954 and 1466 5-second audio respectively (8.27 and 2.03 hours in total). Each sample includes a speech signal corrupted by wind noise with two ultrasound signals, and a corresponding clean speech signal as the ground truth. For the testing set, we randomly shuffle the wind noise audio segments and clean speech utterances used in the validation set, as well as re-generate SNR values in the range, to create a new set of mixed audio segments.

4.2. Training Setup and Baselines

We first train the original architecture of DCCRN and DEMUCS as our baseline. Specifically, we choose the configuration that with the best performance reported in the paper, namely DCCRN-CL and DEMUCS (H=48, S=4, U=4). we use the causal setup in both models where the LSTM is unidirectional. We use the AdamW optimizer [26] with a learning rate of 3e-4, a momentum of $\beta_1 = 0.9$, a denominator momentum of $\beta_2 = 0.999$, and a weight decay of 1e-3. The models are selected by early stopping. Model training are testing are performed on a NVIDIA - GeForce RTX 3090 Graphics Card.

For DEMUCS training, we use the loss presented in the paper, which is the sum of the waveform L1 loss and the multi-resolution STFT loss. Given clean signal $\boldsymbol{y}$ and estimated signal $\hat{\boldsymbol{y}}$ from the model, the loss function can be represented by:

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}||\boldsymbol{y} - \hat{\boldsymbol{y}}||_1 + L_{stft}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

	SI-SDR	PESQ	STOI
DCCRN (3.7 M)	2.685	2.265	0.653
DeWinder - Original Loss	3.581	2.374	0.671
DeWinder - Single	3.841	2.470	0.696
DeWinder (4.2 M)	**3.871**	**2.480**	**0.700**
DEMUCS (18.9 M)	6.632	2.776	0.812
DeWinder - Concat Fusion	6.057	2.805	0.820
DeWinder - Single	6.902	2.855	0.826
DeWinder (21.7 M)	**6.932**	**2.861**	**0.827**

Table 1: *Performance of* DeWinder *and ablation study*

where T is the number of samples in the waveform, and we refer the definition of STFT loss in DEMUCS paper [14].

For DCCRN training, instead of using the SI-SNR loss [27] alone defined in the paper, we also incorporate the STFT loss presented in DEMUCS to improve the overall perceptual audio quality. The loss can be represented by:

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \lambda_1 10 \log_{10}(\frac{||\boldsymbol{y}_{target}||_2^2}{||\hat{\boldsymbol{y}} - \boldsymbol{y}||_2^2}) + \lambda_2 L_{stft}(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

where $\boldsymbol{y}_{target} =< \hat{\boldsymbol{y}}, \boldsymbol{y} > \cdot \boldsymbol{y}/||\boldsymbol{y}||_2^2$. During our training, λ_1 is set to -0.2 and λ_2 is set 1.

Once the baseline model is converged, we add the module of *DeWinder* including the ultrasound encoder and the fusion module to the network. For the original encoder, decoder, and LSTM in the model, we load the weight that was pre-trained in the baseline model as initialization. We note that the first LSTM layer in DCCRN is re-initialized as the input dimension is modified. We train the two-stream architecture of *DeWinder* using the same optimizer setup as the baseline training.

4.3. Evaluation Results

We evaluate the performance of *DeWinder* of the adaptations on DCCRN and DEMUS separately, as the two baseline models have different levels of parameter sizes (3.7 M and 18.9 M). We use the evaluation metrics including Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [28], Perceptual Evaluation of Speech Quality (PESQ) [29] (from 0.5 to 4.5), and Short-Time Objective Intelligibility (STOI) [30] (from 0 to 1).

Table 1 shows the performance of evaluation on our overall testing set with SNR values in the range of -40 to -20 dB. Besides the baseline models of DCCRN and DEMUCS and the *DeWinder*'s complete setup, we also show the performance of *DeWinder* setup but with the original SI-SNR loss used in the DCCRN paper (*DeWinder* - Original Loss). We also modify the input channel number of the ultrasound encoder and let it take only the waveforms from a single ultrasound speaker (20 or 21 kHz). We average the model performance across two frequencies (*DeWinder* - Single). For the evaluation based on DEMUCS, we also evaluated the model that employs concatenation as the fusion module (*DeWinder* - Concat Fusion) to compare with our proposed masking-based fusion.

The results show that adding *DeWinder* to the baseline models can significantly improve the wind noise reduction performance. Surprisingly, we observe that *DeWinder* based on only single ultrasound speaker can achieve almost the same level of performance compared to the full two-speaker setup, which demonstrates the potential capability of *DeWinder* to be deployed on the current hardware setup on off-she-shelf devices such as smartphones, with only a single side of speaker. Meanwhile, we report the total parameter size of the baseline models

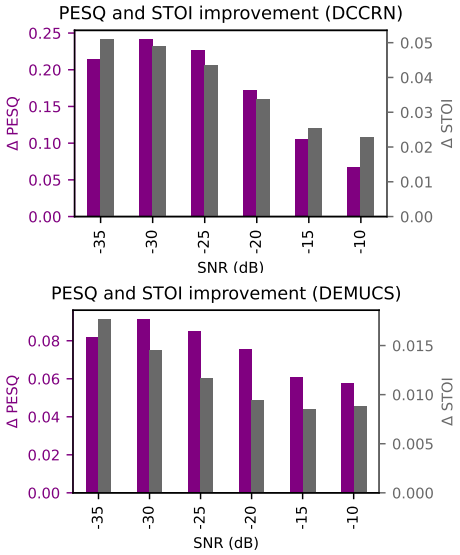

Figure 5: *Performance improvement at different SNRs*

and *DeWinder* in the table. Our design does not need to introduce a significant amount of parameters to the original network, since both the ultrasound encoder and fusion module are lightweight, and thus would not degrade the real-time interference capability of the model.

We also observe the concatenation-based fusion in DEMUCS (*DeWinder* - Concat Fusion) cannot significantly improve the performance compared to the baseline, and even obtains a lower SI-SDR value. Such a result demonstrates the necessity of our masking-based fusion module which considers the semantic meaning of the embeddings.

We further present the evaluation result of the performance improvement at different SNRs. We synthesize the testing sets using fixed pairs of wind noise and speech, but repeat the mixing at different SNR values. We evaluate the PESQ and STOI improvements at SNRs from -35 to -10 dB compared with baseline models. We observe similar trends in both DCCRN and DEMUCS. The performance improves significantly in low-SNR scenarios under -25 dB, while still outperforming the baseline in higher SNR cases. We observe that the adaptation on DCCRN achieves a higher value of performance improvements than the adaptation on DEMUCS. We attribute this to the larger parameter size of DEMUCS, and better capability to reduce non-stationary noise of the waveform-domain models. Thus the baseline DEMUCS model can achieve better performance in wind noise reduction and leave less room for improvement.

5. Conclusion

In this work, we present *DeWinder*, which for the first time, utilizes ultrasound as a complementary modality to sense the wind noise and perform noise reduction. We design a modular framework that can be adapted to different speech enhancement models without introducing significant computational overhead. We collect a wind noise dataset along with ultrasound transmissions and demonstrate *DeWinder* can significantly improve the wind noise reduction capability of two state-of-the-art speech enhancement models: DEMUCS and DCCRN. We leave the design of more complex transmitted ultrasound signals, and exploring other multi-modal fusion mechanisms for further work.

6. Acknowledgement

We acknowledge support from the NSF (2106921, 2030154, 2007786, 1942902, 2111751), ONR, AFRETEC, MFI, CISCO, Safety21 and CyLab-Enterprise. We thank the anonymous reviewers for their constructive feedback.

7. References

[1] G. W. Lyons, C. R. Hart, and R. Raspet, "As the wind blows: Turbulent noise on outdoor microphones," *Acoustics Today*, vol. 17, no. 4, pp. 20–28, 2021.

[2] D. Mirabilii and E. A. Habets, "Multi-channel wind noise reduction using the corcos model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 646–650.

[3] S. Grimm and J. Freudenberger, "Wind noise reduction for a closely spaced microphone array in a car environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–9, 2018.

[4] D. Mirabilii and E. A. Habets, "Spatial coherence-aware multichannel wind noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1974–1987, 2020.

[5] P. Thuene and G. Enzner, "Maximum-likelihood approach to adaptive multichannel-wiener postfiltering for wind-noise reduction," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[6] A. Leśniak, T. Danek, and M. Wojdyła, "Application of kalman filter to noise reduction in multichannel data," *Schedae Informaticae*, vol. 17, no. 18, pp. 63–73, 2009.

[7] S. Kuroiwa, Y. Mori, S. Tsuge, M. Takashina, and F. Ren, "Wind noise reduction method for speech recording using multiple noise templates and observed spectrum fine structure," in *2006 International Conference on Communication Technology*, 2006, pp. 1–5.

[8] O. Eldwaik and F. F. Li, "Mitigating wind induced noise in outdoor microphone signals using a singular spectral subspace method," *Technologies*, vol. 6, no. 1, p. 19, 2018.

[9] B. Kotnik, Z. Kacic, and B. Horvat, "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[10] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 177–180.

[11] C. M. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary, "Single microphone wind noise reduction using techniques of artificial bandwidth extension," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2328–2332.

[12] N. Chatlani and J. J. Soraghan, "Emd-based filtering (emdf) of low-frequency noise for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1158–1166, 2011.

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

[14] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[15] X. Hao, X. Su, R. Horaud, and X. Li, "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)*, 2021, pp. 6633–6637.

[16] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1329–1341, Oct. 2022.

[17] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A mask free neural network for monaural speech enhancement," 2023.

[18] Q. Li, F. Gao, H. Guan, and K. Ma, "Real-time monaural speech enhancement with short-time discrete cosine transform," 2021.

[19] E. U. H. N. M. Command, *Principles and Applications of Underwater Sound: Originially Issued as Summany Technical Report of Division G, NDRC Volume 7, Washington, DC. 1946*, ser. Navmat P-9674. Head quarters Naval Material Command, 1968.

[20] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *Proceedings of the 27th annual international conference on mobile computing and networking*, 2021, pp. 160–173.

[21] G. Cao, K. Yuan, J. Xiong, P. Yang, Y. Yan, H. Zhou, and X.-Y. Li, "Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. New York, NY, USA: Association for Computing Machinery, 2020, p. 95–108.

[22] M. Löhning, T. Hentschel, and G. Fettweis, "Digital down conversion in software radio terminals," in *2000 10th European Signal Processing Conference*. IEEE, 2000, pp. 1–4.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[24] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[27] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," 2018.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" *CoRR*, vol. abs/1811.02508, 2018. [Online]. Available: http://arxiv.org/abs/1811.02508

[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 2, 2001, pp. 749–752 vol.2.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.