NOISE RETF ESTIMATION AND REMOVAL FOR LOW SNR SPEECH ENHANCEMENT

Lachlan Birnie, Prasanga Samarasinghe, Thushara Abhayapala, and Daniel Grixti-Cheng

Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia

ABSTRACT

A method for offline two-microphone speech enhancement in highly adverse noisy environments with signal-to-noise (SNR) ratios of -10 to -20 dB is proposed. While the topic of speech enhancement is well researched, there are very few methods developed to address such significant noise conditions. Specifically, we are interested in removing noise from unintelligible recordings such that the resulting denoised speech content is understandable to human listeners. We propose exploiting the Relative Transfer Function (ReTF), a spatial feature of the noise source in a speech enhancement algorithm. We model the noise source ReTF with a time-domain machine learning structure to estimate and subtract the noise signal from the mixture. Both a linear filtering and an autoencoder based structure are proposed. For a single interfering noise source, speech intelligibility is improved to within 9% below the Short-Time Objective Intelligibility (STOI) score of the benchmark oracle Ideal Binary Mask (IBM).

Index Terms: Relative transfer function, noise removal, speech enhancement, two-microphone, time-domain, CNN

1. INTRODUCTION

Speech enhancement, speech separation and denoising strive towards improving the intelligibility of human speakers in recordings that have been corrupted by an interfering noise [1]. The task of speech enhancement is applicable to telecommunications, hearing aids, and robust automatic speaker recognition [2]. Numerous one-microphone and multi-microphone approaches to speech enhancement have been developed. Blind Source Separation (BSS) algorithms include spectral subtraction, non-negative matrix factorization, coherence function, and more [3, 4, 5, 6]. Recently machine learning algorithms are becoming increasingly successful in speech enhancement applications. Most popular are frequency domain algorithms that estimate the Ideal Binary Mask (IBM) for the spectral magnitude of the noisy speech mixture [7]. These frequency domain algorithms, however, often experience distortions and lowered speech quality due to unprocessed phase information [8]. To address this, time domain machine learning algorithms that work directly on the noisy speech waveform have been proposed, and they are recently becoming more favored [9, 10, 11, 12]. Time domain algorithms such as WAVENET [13, 14], CONV-TASNET [15], WaveGAN [16, 17], and Wavesplit [18] have all shown impressive speech processing capabilities.

To the best of our knowledge, thus far, speech enhancement research has rarely considered scenarios of highly adverse noise. Generally most speech enhancement algorithms consider noisy scenarios of only 0 dB to -10 dB SNR [3, 5, 10, 14, 15, 17, 19, 20, 21, 22, 23, 24]. As such, scenarios of significantly noisy environments below -10 dB SNR are noticeably unexplored.

In such adverse noisy scenarios it is unreliable to use intrinsic properties of speech like pitch and sparsity for enhancement. Instead, it is more reliable to exploit properties of the noise source. We note that it is not feasible nor desirable to pre-train an algorithm to learn the spectrum of every potential noise. However, if the noise source is unmoving then spatial properties can be exploited. We propose using the Relative Transfer Function (ReTF) as a unique spatial property that describes the coupling between two microphones in response to an unmoving noise source. The ReTF property has been successfully used in source localization [25], and moderately explored for speech enhancement in beamforming algorithms [1, 26, 4]. More recently the authors of [27] have exploited ReTF in a deterministic two-microphone source separation algorithm. However, the potential of ReTF is still yet to be completely investigated in low SNR speech enhancement applications. As such, we propose exploiting the noise source ReTF in a time domain machine learning algorithm for greater speech enhancement.

Exploiting the ReTF requires the assumption that the noise source is unmoving with respect to the two-microphone device during a recording. However, there are many noisy applications that we are interested in where this assumption holds. Large building noise caused from heating/cooling units, data-servers, pumps, generators and manufacturing machinery are unmoving with respect to ceiling mounted microphones or security devices. Furthermore, hands-free devices in the cabins of cars/planes are at a fixed position with respect to engine noise. Finally, the noise source ReTF can be exploited in removing motor noise from drones or robots, or self-noise from smart speakers/devices that have

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

integrated microphones for human-to-device communication. In each of these applications, there is an interest in extracting the hardly perceptible speech content of human speakers when a noise source is too loud or too close, resulting in extreme (below -10 dB) SNR scenarios.

In this paper, we propose a two-microphone speech enhancement algorithm for significantly noisy (below -10 dB SNR) environments. We use a time-domain machine learning approach due to its successful dominance in recent speech processing applications. We build a structure of convolutional layers to approximately model a speech enhancement method based on exploiting the noise source ReTF. The method relies on two assumptions about the noise source. First, a segment of recording containing only the noise source (no active speech) is available and can be used to learn the noise source ReTF with machine learning. This is a common assumption for unsupervised speech enhancement algorithms [20, 23]. The noise-only recording segment can be obtained with either a human listener or an automatic Voice Activity Detection (VAD) algorithm [28]. Second, as mentioned we assume that the noise source is unmoving. For application in a reverberant environment we also assume that the two-microphone device is unmoving, as this is required for estimating the noise source ReTF.

In Section 2 we formulate the adverse (below -10 dB SNR) convolutive noisy speech mixture that we aim to enhance. Throughout this paper we formulate the case of a single interfering noise source, and at the end we briefly examine the effect of multiple noise sources. We enhance speech by exploiting the noise source ReTF in Sec. 3. In Section 4 we build a structure of unsupervised one-dimensional convolution layers to model and learn the noise source ReTF for speech enhancement. We propose two convolutional layer structures; the first based on adaptive filtering [20, 21, 22], and the second based on deep denoising autoencoders [24, 29]. In Section 5, we use real-world impulse responses to experimentally examine the proposed algorithm, and compare results to an oracle IBM.

2. PROBLEM FORMULATION

Consider a reverberant environment containing a single speech source denoted by 1, a single unmoving noise source denoted by 2, and an unmoving two-microphone device denoted by $\{a, b\}$. The device records the convolutive mixture of [30]

$$\begin{aligned} p_a(t) &= s(t) * h_{1a}(t) + n(t) * h_{2a}(t), \\ p_b(t) &= s(t) * h_{1b}(t) + n(t) * h_{2b}(t), \end{aligned} \tag{1}$$

where $*$ represents the convolution operation, p_a and p_b are the two microphone recordings, $s(t)$ is the speech source signal, $n(t)$ is the noise source signal, and $h_{\{1,2\}\{a,b\}}(t)$ are the room impulse responses (IRs) between each source $\{1, 2\}$ and microphone $\{a, b\}$. Note that we consider thermal microphone noise to be negligible.

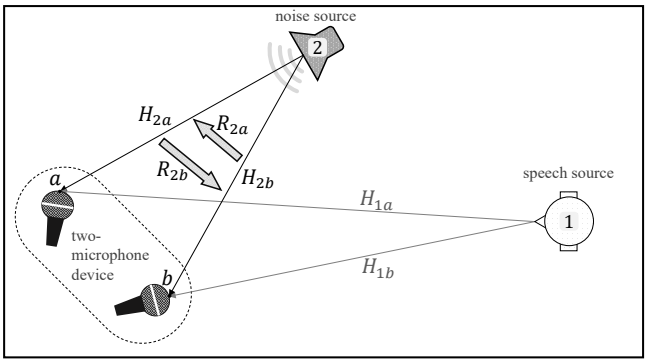

Fig. 1. Diagram of convolutional mixing system.

In frequency domain (1) is given by

$$\begin{aligned} P_a(k) &= S(k)H_{1a}(k) + N(k)H_{2a}(k), \\ P_b(k) &= S(k)H_{1b}(k) + N(k)H_{2b}(k), \end{aligned} \tag{2}$$

where k denotes the frequency bin, and $H_{\{1,2\}\{a,b\}}(k)$ are the room transfer functions between each source and microphone. We depict this noisy speech mixture in Fig. 1.

Our problem considers the noise source to be extremely disruptive, such that the noisy speech recording $p(t)$ has an SNR below -10 dB. Moreover, we consider the speech content $s(t)$ to be unintelligible to humans when listening directly to the recording. Our objective is to enhance the recording enough to make the content of $s(t)$ understandable to human listeners. We assume that it is unfitting to rely on properties of the speech signals for enhancement due to the high noise levels. Therefore, we propose exploiting the spatial properties of the noise source for speech enhancement in the next section.

3. SPEECH ENHANCEMENT WITH NOISE RETF

In this section, we briefly introduce the ReTF of the noise source. We then show how a two-microphone recording of noisy speech can be enhanced when the noise source ReTF is known.

3.1. Noise source relative transfer function

The ReTF describes the coupling between two microphones in response to a given sound source. For the system in Fig. 1, the ReTF between the two microphones $\{a, b\}$ in response to the noise source is given by the transfer function ratio of [31]

$$R_{2a}(k) = \frac{H_{2a}(k)}{H_{2b}(k)}, \qquad R_{2b}(k) = \frac{H_{2b}(k)}{H_{2a}(k)}. \tag{3}$$

Note that microphone-b is taken as the reference for R_{2a}, and microphone-a is the reference for R_{2b}. In another sense, R_{2b} can be seen as a mapping between the noise signal received at microphone-a to the noise signal at microphone-b, as shown

by

$$N(k)H_{2a}(k) \times R_{2b}(k) = N(k)H_{2b}(k). \quad (4)$$

Note that R_{2a} maps the noise signal from microphone b-to-a.

Typically, it is difficult to estimate (3) without directly measuring the noise source transfer functions $H_{\{2\}\{a,b\}}(k)$. Some ReTF estimation techniques exist for both noise and speech signals [32]. Alternatively, the ReTF (3) can be modeled through a machine learning algorithm, which we propose in Sec. 4.

3.2. Speech enhancement using noise source ReTF

We propose that the noisy speech recording, $P_a(k)$, and the noise source ReTF, $R_{2b}(k)$, can be used together to approximately estimate the noise signal at microphone-b. Following the ReTF mapping in (4), we propose that

$$P_a(k) \times R_{2b}(k) \approx N(k)H_{2b}(k). \quad (5)$$

Assuming that the above approximation is reasonable, we can subtract (5) from the noisy speech recording $P_b(k)$ to recover/enhance the speech, such that

$$P_b(k) - (P_a(k) \times R_{2b}(k)) \approx S(k)H_{1b}(k). \quad (6)$$

In practice the approximation of (5) is not accurate, however, expanding the left side of (6) shows that the actual recovered signal is a filtered version of the speech, expressed as

$$\begin{aligned} P_b - P_a R_{2b} &= (SH_{1b} + NH_{2b}) - (SH_{1a} + NH_{2a})R_{2b}, \\ &= SH_{1b} + NH_{2b} - SH_{1a}R_{2b} - NH_{2a}\frac{H_{2b}}{H_{2a}}, \\ &= SH_{1b} + NH_{2b} - SH_{1a}R_{2b} - NH_{2b}, \\ &= SH_{1b} - SH_{1a}R_{2b}, \\ &= S(H_{1b} - H_{1a}R_{2b}). \end{aligned} \quad (7)$$

Note that we have dropped the notation of k for brevity. We consider the resulting filtered speech in (7) to be enhanced, as the noise signal has been removed from the ReTF filtering and subtraction. Furthermore, we observe that the distortion remaining in the enhanced speech signal is only given by the room transfer functions of the noise and speech source. Therefore, we assume that the intelligibility of the resulting speech signal is negligibly impacted by these room transfer function distortions.

4. CONVOLUTIONAL STRUCTURES

In this section, we first propose approximating the speech enhancement (7) by modeling the noise source ReTF with a unsupervised machine learning algorithm. We then propose two convolutional layer structures for modeling the noise source ReTF. Both models require a recording segment of inactive speech and only noise. The first structure is based on BSS adaptive filtering speech enhancement [20, 21, 22]. Here we train a single convolutional layer offline as a substitution of an adaptive filter, which instead models the noise ReTF as a linear filter. The second structure is based on deep denoising autoencoders [24, 29]. Here the autoencoder input is the noise at microphone-a and the output is an estimation of the noise at microphone-b.

4.1. Modeling noise source ReTF with machine learning

We cannot estimate the noise source ReTF (3) directly because we assume that the noise source room transfer functions are unknown. We instead propose modeling the noise source ReTF with a machine learning algorithm that is trained to learn the mapping of (4). Furthermore, we are interested in utilizing a time-domain convolutional layer structure. We use a machine learning algorithm denoted as $f_{2b}(\cdot)$ to approximate the mapping of (4) in the time-domain, such that

$$f_{2b}(n(t) * h_{2a}(t)) \equiv n(t) * h_{2b}(t). \quad (8)$$

We train $f_{2b}(\cdot)$ such that it is equivalent to $R_{2b}(k)$ in a time-domain application.

Once the machine has learned the mapping of (8), we can use it to approximate the speech enhancement method of (7). Therefore, the noisy speech recording of microphone-a is used to estimate and subtract the noise signal from microphone-b, expressed in the time-domain as

$$\begin{aligned} p_b(t) - f_{2b}(p_a(t)) &= s(t) * (h_{1b}(t) - h_{1a}(t)r_{2b}(t)), \\ &= \hat{s}(t), \end{aligned} \quad (9)$$

where $r_{2b}(t)$ denotes the time-domain relative impulse response of the noise source, and $\hat{s}(t)$ is the recovered/enhanced speech signal. Next, we propose two convolutional filter structures for the modeling of $f_{2b}(\cdot)$.

4.2. Convolutional filter

We use a single one-dimensional convolutional layer to model the coupling between two-microphones as a linear filter. Figure 2(a) depicts this structure. We note that this structure is linear as there are no activation functions used. The convolutional filter is trained on a segment of the recording where only the noise source is active, and no speech is present. We input overlapping windowed (2048 samples) time-frames of the noise recording at microphone-a and -b, and train to estimate the noise at microphone-b. The inputs are zero padded such that the output is equal in length (2048 samples) after passing through the convolutional layer. Note that the window length should be longer than the room's reverberation time to satisfy the multiplicative transfer function [33]. A mean squared error loss function between the pure-noise signal, $n_b = (n(t) * h_{2b}(t))$ (given by (1) when $s(t) = 0$) and

the estimated noise, denoted as $\hat{n}_b = f_{2b}(n_a)$, is optimized with a 0.001 learning rate;

$$L(n_b, \hat{n}_b) = \frac{1}{M} \sum_{i=0}^{M} (n_b - \hat{n}_b)^2 , \tag{10}$$

where i is the sample index and M is the window-length. Ideally the microphone-a-to-b convolutional filter will model the noise source ReTF on its own. However, we have found that including an additional microphone-b-to-b convolutional filter improves the result. We can add artificial noise to the microphone-b input to dissuade the trivial identity mapping [29], however, we have omitted this here for simplicity. After training on pure-noise segments we fix the filter weights. The noise speech mixture, $p_a(t)$, is processed offline to estimate $\hat{n}_b$. We then subtract this estimated noise signal from $p_b(t)$ to extract an enhanced speech signal $\hat{s}(t)$ given in (9).

4.3. Convolutional autoencoder

We utilize multiple convolutional layers in a denoising autoencoder fashion, as shown in Fig. 2(b). We input 15 windowed (128 sample) time-frames of microphone-a recording segments that contain pure-noise, with no speech present. This input size is selected such that the convolutional autoencoder has a receptive field longer than the room's reverberation time. For output, we train an estimation of the microphone-b noise segment corresponding to the middle (8^{th}) time-frame using (10). Typically denoising autoencoders are trained to produce the input signal with subtracted noise. For our application, we train the autoencoder to produce the input noise segment filtered by the noise source ReTF, which is equivalent to the noise segment at the other microphone. Therefore, the autoencoder learns the mapping of noise signal from microphone-a-to-b, as described in (4). We use an initial learning rate of 0.001 that is multiplicatively reduced by 0.9 every 600 time-frames. After training, we fix the filter weights and filter $p_a(t)$ to subtract the estimated noise to get the enhanced speech in (9).

5. EXPERIMENTAL VALIDATION

We generate noisy speech mixture recordings (1) with IR measurements 1 m from a KEMAR dummy head [34] which portrays our arbitrary two-microphone device. The IRs are performed inside a carpeted room with a dropped ceiling and some partial acoustic treatment ($T_{60} = 181$ ms). We convolve the noise source IRs with air-conditioner noise, and the speech source IRs with female speech utterances from the TIMIT data-set [35]. The reverberant speech and noise signals are down sampled to 16 kHz and leveled from -10 to -20 dB SNR before summing (1). We use an additional 60 second segment of reverberant air-conditioner noise (not part of the mixture) for training. We note again that we assume this pure-noise recording segment is obtainable in practice

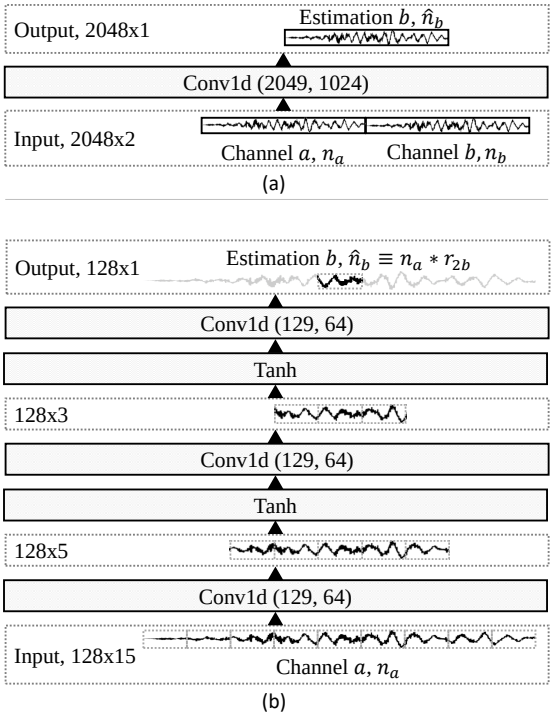

Fig. 2. Convolutional structures that model the noise source ReTF: (a) convolutional filter, (b) convolutional autoencoder. (ν, μ) denotes kernel size ν, and zero-padding μ. These sizes are selected for a 16 kHz sampling frequency.

by either a VAD algorithm or a human listener identifying periods of inactive-speech.

The two convolutional structures (Sec. 4.2 / 4.3) are implemented in PyTorch [36]. For training, we use 6 epochs of the 60 second pure-noise segment. Filter weights are fixed and we process the noisy speech recordings with (9) to extract the enhanced speech $\hat{s}(t)$. For analysis we use the Perceptual Evaluation of Speech Quality (PESQ) score [37] to evaluate speech quality from -0.5 to 4.5, and the STOI score [38] to evaluate intelligibility from 0% to 100%. We compare results to an oracle IBM enhanced speech that uses 2048 sample Hann windowed STFT frames with 50% overlap and a -4.77 dB local criterion [39], given as

$$\text{IBM}(t,k) = \begin{cases} P(t,k), & \text{if SNR}(t,k) \geq \text{-4.77 dB}, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where $P(t,k)$ is the STFT of the recording $p_{\{a,b\}}(t)$.

One noise source results: Figure 3 shows the speech enhancement performance averaged over five minutes of TIMIT utterances spaced by one second pauses. Both the convolutional filter and autoencoder structure display similar performance, with the filter structure exhibiting slightly better speech quality (PESQ). We observe that the IBM and the proposed methods have the same trend in performance degradation with worsening SNR. As a result, on average the STOI score is 8.7% below the IBM for the convolutional filter, and

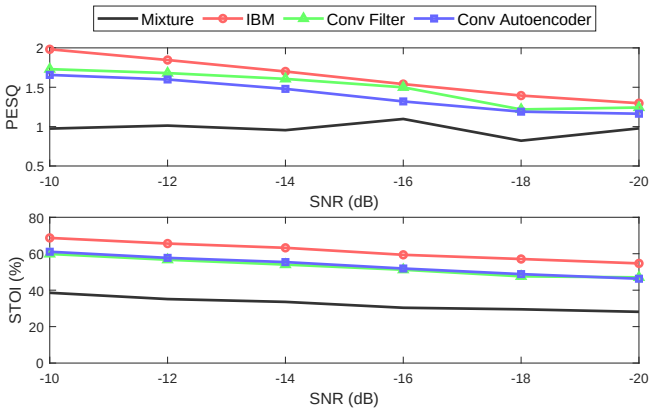

Fig. 3. PESQ (a) and STOI (b) scores for one noise source.

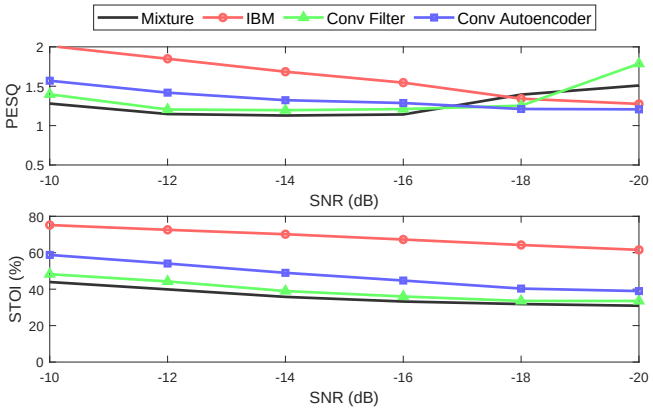

Fig. 4. PESQ (a) and STOI (b) scores for two noise sources.

7.9% below for the autoencoder, over the -10 dB to -20 dB SNR range. From informal listening, we find that both proposed methods are able to greatly increase the understandably of the speech content.

Two noise source results: We repeat the experiment with an additional vacuum-cleaner noise source in Fig. 4. For these two continuously overlapping noises, we observe that the performance between the proposed methods and the IBM has widened compared to the single noise case. The convolutional filter's average STOI has decreased to 29.4% below the IBM, while the autoencoder shows better intelligibility with average STOI of 20.9% below IBM. From informal listening we find that the autoencoder's enhanced speech remains to have significant residual noise. Whereas, we find the convolutional filter has better noise reduction but at the cost of destroyed speech content. The degraded speech content may also explain the abnormal PESQ results at -20 dB SNR. Overall, both enhancement methods produce similarly poor intelligibility for the two interfering noise source case. As such, more complex convolutional structures designed for multiple noise sources are still to be investigated.

6. CONCLUSION

We have proposed an offline two-microphone speech enhancement algorithm for adverse (below -10 dB SNR) noisy environments. The algorithm aims the exploit the spatial properties of the noise source by modeling the noise ReTF with a time-domain convolutional structure. We proposed two convolutional structures based on an adaptive filter and an autoencoder, where both aim to estimate the noise source waveform from a noisy speech recording. Strong speech enhancement was achieved for a single interfering noise source. Each convolutional structure approached the intelligibility of the oracle IBM without the requirement of pre-training on large speech data-sets, and instead only requiring a 60 second recording segment of noise. Informal listening showed the method to be capable of extracting understandable speech content from hardly perceptible recordings. Poorer enhancement was attained for two interfering noise sources. To address this, further convolutional structures designed to model multiple noise source ReTFs should be investigated. Furthermore, integrating a front-end VAD algorithm for below -10 dB SNR environments is left for future work.

7. REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, pp. 1702–1726, 2018.

[3] F. Kallel, M. Frikha, M. Ghorbel, A. Hamida, and C. Berger-Vachon, "Dual-channel spectral subtraction algorithms based speech enhancement dedicated to a bilateral cochlear implant," *Applied Acoustics*, vol. 73, no. 1, pp. 12–20, 2012.

[4] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ica-based blocking matrix for improved noise estimation," *EURASIP J. Adv. Signal Process.*, vol. 2014, p. 26, 2014.

[5] W. Nabi, M. Nasr, N. Aloui, and A. Cherif, "A dual-channel noise reduction algorithm based on the coherence function and the bionic wavelet," *Applied Acoustics*, vol. 131, pp. 186–191, 2018.

[6] S. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with gcc-nmf," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, pp. 745–755, 2017.

[7] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[8] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[9] C. Liu, S. Fu, Y. Li, J. Huang, H. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1888–1900, 2020.

[10] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1570–1584, 2018.

[11] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6875–6879.

[12] ——, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1179–1188, 2019.

[13] A. Oord *et al.*, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[14] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5069–5073.

[15] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ByMVTsR5KQ

[17] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1428

[18] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[19] N. Saleem, M. Khattak, and M. Shafi, "Unsupervised speech enhancement in low snr environments via sparseness and temporal gradient regularization," *Applied Acoustics*, vol. 141, pp. 333–347, 2018.

[20] R. Bendoumia and M. Djendi, "Acoustic noise reduction by new two-channel proportionate forward symmetric adaptive decorrelating algorithms in sparse systems," *Applied Acoustics*, vol. 137, pp. 69–81, 2018.

[21] P. Thaitangam, R. Laishram, K. Devi, R. Khwairakpam, M. Singh, and C. Oinam, "Speech enhancement using adaptive filter with bat algorithm," in *IEEE Int. Conf. Comput. Intell. Comput. Research*, 2018, pp. 1–5.

[22] R. Henni, M. Djendi, and M. Djebari, "A new efficient two-channel fast transversal adaptive filtering algorithm for blind speech enhancement and acoustic noise reduction," *Comput. & Elect. Eng.*, vol. 73, pp. 349–368, 2019.

[23] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by noisy2noisy signal mapping," *Applied Acoustics*, vol. 172, 2021.

[24] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1759–1763.

[25] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 8, pp. 1393–1407, 2016.

[26] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[27] A. Bates, D. Grixti-Cheng, P. Samarasinghe, and T. Abhayapala, "Investigating the use of the relative transfer function for source separation on two-channel recordings," in *Asia Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.* IEEE, 2020.

[28] Z. Tan, A. Sarkar, and N. Dehak, "rvad: An unsupervised segment-based robust voice activity detection method," *Comput. Speech & Language*, vol. 59, pp. 1–21, 2020.

[29] D. Grozdić, S. Jovičić, and M. Subotić, "Whispered speech recognition using deep denoising autoencoder," *Eng. Appl. of Artificial Intell.*, vol. 59, pp. 15–22, 2017.

[30] E. Weinstein, M. Feder, and A. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 405–413, 1993.

[31] O. Shalvi and E. Weinstein, "System identification using non-stationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, 1996.

[32] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Eur. Signal Process. Conf.* IEEE, 2018, pp. 2499–2503.

[33] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, 2007.

[34] M. Burkhard and R. Sachs, "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Amer.*, vol. 58, no. 1, pp. 214–222, 1975.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, vol. 10, no. 5, p. 0, 1993.

[36] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752.

[38] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[39] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, 2010.