Interspeech 2024
1-5 September 2024, Kos, Greece

TfCleanformer: A streaming, array-agnostic, full- and sub-band modeling front-end for robust ASR

Jens Heitkaemper, Joe Caroselli, Arun Narayanan, Nathan Howard

Google LLC, U.S.A.

heitkaemper@google.com

Abstract

Multiple recent publications have demonstrated the benefits of neural network based enhancement in the time-frequency domain. This paper builds on those findings to improve upon a recently published streaming, array agnostic multi-channel enhancement system called Cleanformer. The proposed streaming enhancement system achieves competitive results against a non-causal state-of-the-art model on a source separation task, outperforming Cleanformer. Additionally, the presented model improves upon Cleanformer enhancement results in multiple challenging environments without introducing further latency. A short ablation study is performed to evaluate the influence of the proposed changes on the improved performance.

Index Terms: multi-channel, speech enhancement, neural networks

1. Introduction

To this day speech enhancement is an important building block for automatic speech recognition (ASR) in challenging environments, e.g., homes, meeting rooms, cafes. This work focuses on enhancement for smart speakers which is a widely used application of automatic transcription [1]. ASR on smart speakers has two main challenges, the background noises common in a household like TV noise, music, and overlapping speakers, and the fast response time expected by the user. Therefore, a low latency is an important characteristic of an enhancement system for smart speakers. Most interactions with a smart speaker consists of three phases: a noise context, a hotword and a query, where both the hotword and query are mostly just a few seconds long. This sequence of audio signals allows the enhancement system and the acoustic model to adapt to the current scenario, e.g., by using the noise context to extract noise statistics [2, 3] or using the hotword to extract speaker information [4].

There are various enhancement approaches that have shown to significantly improve the transcription accuracy by exploiting auxiliary information, like speaker embeddings [5, 6], spatial information [7, 8] or knowledge about the noise source [9, 10]. Multiple publications have shown the benefits of using multi-stage enhancement approaches, where one part of the model uses the provided additional information, while the second stage cleans the output of the earlier enhancement model [3, 11, 12]. Such a multi-stage approach allows the first stage to focus on removing the noise signal, while the second stage finetunes the output of the initial enhancement.

One such model is the recently published Cleanformer, a streaming, array agnostic, multi-channel enhancement system for smart speakers [3]. This approach uses a statistical filter, called SpeechCleaner [13, 14], optimized on the noise context, for a first enhancement stage and a neural network (NN) as a second step to remove possible distortions introduced by the filter. The second stage network performs masking in the log-mel domain.

In this work, we propose an adaptation to the Cleanformer architecture inspired by recent publications, which show the benefits of enhancement in the time-frequency domain, especially in reverberant environments [12, 15, 16, 17, 18]. More precisely we take inspiration from the TfGridNet architecture first presented in [12], which uses recurrent neural network (RNN) layers for full- and sub-band modelling on short-time Fourier transform (STFT) input features to achieve state-of-the-art source separation performance. We propose a smaller, streaming full- and sub-band modelling architecture as preprocessing layers for Cleanformer and show the benefit of the proposed architecture on multiple databases. Furthermore, the RNN-based full- and sub-band modelling layers are replaced by a Conformer [19] architecture, leading to additional gains.

The rest of this paper is structured as follows. Section 2 gives an overview of the proposed system with a short introduction to SpeechCleaner in section 2.1 and a detailed explanation of the proposed changes to Cleanformer in section 2.2. In section 3 an overview of the chosen model parameters is given and the training and evaluation data is introduced. Section 4 offers a discussion of the performed experiments. The paper ends with a short conclusion in section 5.

2. System

For a recording with C microphones, the vector of the frequency representations of the signal at time frame ℓ and frequency bin f, for all channels is represented by $\mathbf{Y}_{\ell,f} = [Y_{1,\ell,f}, ..., Y_{C,\ell,f}]^{\mathsf{T}}$. The signal at each microphone c can be described as follows (using the convolutive transfer function approximation [20]):

$$Y_{c,\ell,f} = H_{c,\ell,f} * S_{\ell,f} + N_{c,\ell,f} = X_{c,\ell,f} + N_{c,\ell,f}, \quad (1)$$

where $*$ is the convolution operator and $S_{\ell,f}$, $N_{c,\ell,f}$ represent the STFT coefficients of the clean target speech signal and the noise signal, respectively. $H_{c,\ell,f}$ is a vector of STFT coefficients of the transfer functions from the target speaker to the microphone c. Finally, $X_{c,\ell,f}$ represents the multi-microphone STFT coefficients of the reverberated target speech. Equation (1) in vector notation becomes $\mathbf{Y}_{\ell,f} = \mathbf{X}_{\ell,f} + \mathbf{N}_{\ell,f}$.

The building blocks of the proposed enhancement system are shown in fig. 1. It expects a noisy signal Y and at the output of the first enhancement stage $\hat{X}^{\text{CL}}$ at the input. The following sections describe the depicted building blocks and the first-stage enhancement system, called SpeechCleaner.

10.21437/Interspeech.2024-2378

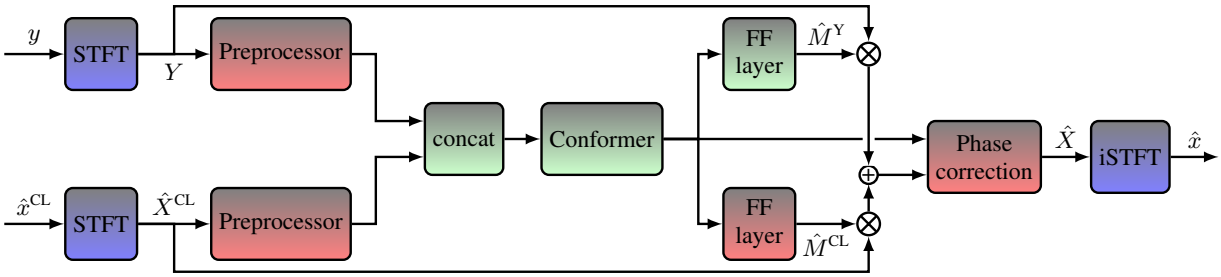


Figure 1: *Overview of the proposed enhancement system. The green blocks are (mostly) unchanged from the original Cleanformer, the blue blocks represent the feature extraction and the red blocks are the new layers added in the new TfCleanformer architecture.*

2.1. SpeechCleaner

SpeechCleaner is a multi-microphone linear filter for noise reduction [13,14]. The filter coefficients are optimized to estimate the noise signal in a microphone channel c_0 from the remaining channels $c_{\overline{0}}$. This optimization is performed on frames without target speech activity using the minimum mean squared error (MSE) criterion:

$$\hat{\boldsymbol{\omega}}_f^{\mathrm{CL}} = \underset{\boldsymbol{\omega}_f}{\operatorname{argmin}}\ \mathbb{E}\left[\left|Y_{c_0,\ell,f} - \boldsymbol{\omega}_f^{\mathsf{H}} \tilde{\mathbf{Y}}_{c_{\overline{0}},\ell,f}\right|_2^2\right], \qquad (2)$$

where $\mathbb{E}$ is the expectation operator and $|\cdot|_2$ the L2-norm. $\tilde{\mathbf{Y}}_{c_{\overline{0}},\ell,f} = [Y^{\mathsf{T}}_{c_{\overline{0}},\ell-D^{\mathrm{CL}}+1,f}, ..., Y^{\mathsf{T}}_{c_{\overline{0}},\ell,f}]^{\mathsf{T}}$ is a vector of the non-target channels with a temporal context D^{CL}. The estimated noise signal, $\hat{\mathbf{N}}_{\ell,f} = \left(\hat{\boldsymbol{\omega}}_{\ell,f}^{\mathrm{CL}}\right)^{\mathsf{H}} \tilde{\mathbf{Y}}_{c_{\overline{0}},\ell,f}$, is subtracted from $Y_{c_0,\ell,f}$ to get an estimate of $X_{\ell,f}$, $\hat{X}^{\mathrm{CL}}_{\ell,f}$. Since the algorithm has to work in a streaming fashion without perfect information, the correlation matrices for computing $\hat{\boldsymbol{\omega}}_f^{\mathrm{CL}}$ are estimated recursively with an exponential forgetting factor, α^{CL}.

The filter coefficients should only be updated using non-target speech frames. Although a NN can be used to find non-target speech frames, in the considered scenario of a voice assistant invoked using a hotword, we assume that all sound prior to the hotword can be used as noise information. Therefore, the coefficients can be updated until the hotword begins. Afterwards, the coefficients are frozen and applied to the noisy input signal to get an estimate of the interfering signal, which is subtracted from $Y_{c_0,\ell,f}$ to get the cleaned signal, $\hat{X}^{\mathrm{CL}}_{\ell,f}$.

A downside of using SpeechCleaner is that the filters are optimized during interference-only segments to remove all active sources, but applied during target speech activity. This can introduce distortions due to oversuppression. Such distortions have been shown to significantly worsen ASR performance especially in high signal-to-noise ratio (SNR) conditions [3]. Therefore, Cleanformer was proposed to benefit from SpeechCleaner's strong noise suppression capabilities while removing possible distortions [3].

2.2. TFCleanformer

TfCleanformer introduces four adaptations to the Cleanformer network published in [3]. First, the input features are the STFT coefficients of the input signals instead of stacked log-Mel features. Additionally, pre-processing layers for both input signals are added and a mask is estimated to combine the pre-enhanced signal with the estimated enhanced signal. Finally, a phase correction layer is added to account for possible errors in the noisy phase. The following sections will describe each of these updates.

2.2.1. Preprocessors

These preprocessors are inspired by the results of TfGridNet for the source separation task [21]. TfGridNet utilizes a combination of sub- and full-band layers to aggregate both spectral, as well as temporal, information for a final attention-based enhancement layer. In this work, this concept is adjusted to support streaming processing and applied to Cleanformer. Figure 2 depicts the resulting preprocessor architecture.

First, the real and imaginary components of the complex-valued input signal are stacked in the channel dimension and a causal two dimensional convolutional layer with a 3×3 kernel and stride 1 is used to calculate an embedding vector of size E for each time-frequency bin. Afterwards, for each time frame the $\tilde{T}$ past frames are concatenated in the embedding dimension, leading to an embedding dimension of size $\tilde{T} \cdot E$, this process is called T-Unfold. Note, that TfGridNet uses both past and future frames. However, the proposed unfolding allows for causal processing while still benefiting from temporal context in the inter-frame layers. Layer-norm [21] is applied to each embedding vector and the resulting signal is processed by a non-causal context-aware layer (CAL) where the frequency dimension is used as the context dimension.

A CAL can be a RNN layer as in [21] or any other layer that extracts information from the neighboring bins in the context dimension, e.g., a convolutional layer. The output is processed by a deconvolution layer and added to the original embedding vectors to create a residual connection. Each of these steps is performed per time frame.

The updated embedding vectors are unfolded over the frequency dimension, where for each frequency bin $\tilde{F}$ the neighboring bins are stacked in the embedding dimension (F-Unfold). Then, a layer-norm is applied to the embedding dimension and the resulting signal is processed by a causal CAL layer. A deconvolution layer is used to undo the unfold step. Finally, the output is added to the original embedding vector and the maximum value of the embedding dimension is forwarded to the next layer. Note, that this is a departure from TfGridNet, which forwards the full embedding vector for each time-frequency bin. Forwarding only the maximum value simplifies the model and leads to better compatibility to the orignal Cleanformer.

2.2.2. Masking adjustments

We propose two adjustments to the original masking process, which consisted of estimating a real-valued mask which is multiplied with the noisy input signal ($\hat{X}_{\ell,f} = Y_{c_0,\ell,f} \cdot \hat{M}^{\mathrm{Y}}_{\ell,f}$). First, calculating an additional mask $\hat{M}^{\mathrm{CL}}$ to mask the SpeechCleaner output and add the masked signal to output. This change allows the network to take advantage of the non-distorted parts of the

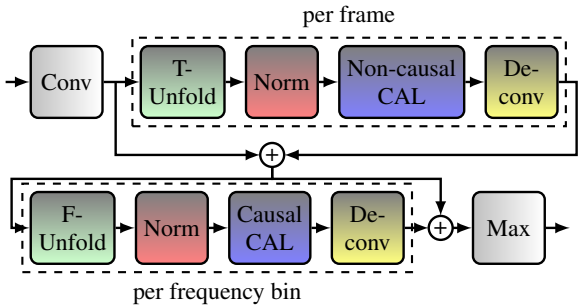


Figure 2: *Overview of the proposed preprocessing layers, where CAL stands for a context-aware layer.*

cleaned signal or further enhance the masked noisy signal.

The original masking is operating in the log-Mel feature domain, therefore the signal phase is not of concern. However, for enhancement in the STFT domain directly using the noisy phase may limit the possible performance [22]. In this work, it is assumed that the noisy phase and the clean phase are relatively close so that only small adjustments are necessary. Therefore, a phase correction layer is proposed that calculates a phase correction term $\hat{\Delta}\phi \in [-\pi, \pi]$ by processing the Conformer output with a feed forward (FF)-layer with a tanh activation and multiplying the result with π. These changes lead to the following updated output signal:

$$\hat{X}_{\ell,f} = \left(Y_{c_0,\ell,f} \cdot \hat{M}^{\mathrm{Y}}_{\ell,f} + \hat{X}^{\mathrm{CL}}_{\ell,f} \cdot \hat{M}^{\mathrm{CL}}_{\ell,f}\right) \cdot \mathrm{e}^{\mathrm{j}\hat{\Delta}\phi_{\ell,f}}. \quad (3)$$

2.3. Loss functions

Cleanformer is trained using a MSE-loss and an ASR-loss [23] on the output mask in the log-Mel feature domain. Since the new model operates in the frequency domain, a time-domain loss is applied to the reconstructed signal. In this work the scale invariant (SI)-SNR loss and the aforementioned ASR loss [23] is used. Note, that the ASR loss is applied to the time domain signal in contrast to the Cleanformer, where the ASR loss is directly applied to the enhanced Mel feature domain signal.

3. Parameters

The proposed architecture is compared to two architectures from the literature. TfGridNet, a state-of-the-art source separation model [21], is chosen as an upper bound for the performance of the presented streaming enhancement model on separation tasks. A pretrained model trained in ESPnet [24, 25] is chosen to improve comparability[1]. Although the chosen TfGridNet model was trained on a dataset different from the Cleanformer variants, both trainings are similar enough that TfGridNet as a state-of-the-art offline model should outperform the presented streaming models. The second considered model is Cleanformer as described in [3].

The TfCleanformer input STFT transformation uses a FFT size of 512, a window size of 32 ms and a shift of 16 ms. For the unfold window a size of $\tilde{F} = \tilde{T} = 4$ is chosen and the embedding vector has size $E = 64$. The main separation network is a Conformer [19] identically to the separtion layer used in Cleanformer except for the input feature dimension size.

[1] https://huggingface.co/espnet/yoshiki_wsj0_2mix_spatialized_enh_tfgridnet_waspaa2023_raw

Two different architectures are considered for the CAL, a long-short term memory (LSTM)-based architecture, called RnnCleanformer, and a Conformer-based architecture, called TfCleanformer. RnnCleanformer uses a unidirectional LSTM for the causal CAL and a bi-directional LSTM for the CAL both with one hidden layer with 64 nodes. For TfCleanformer both the causal and non-causal CAL use a single Conformer block with 8 attention heads and a kernel size of 15. The causal CAL uses the 30 prior frames as a context window while the non-causal CAL uses 30 prior and future frequency bins.

SpeechCleaner operates on a STFT transformed signal with FFT size 2048, window size 128 ms and shift 64 ms. The temporal context is $D^{\mathrm{CL}} = 3$ and the forgetting factor is $\alpha^{\mathrm{CL}} = 0.995$.

3.1. Training set

Both Cleanformer and TfCleanformer are trained on the simulated set described in [3] with around 50k hours of speech based on Librispeech [26], Librivox [27] and internal vendor-collected utterances. There are 281k utterances in Librispeech and 1.9M in the vendor-collected set. Librivox is segmented to create utterances that are 3 to 15 seconds long, resulting in a dataset with 18M utterances. A room simulator is used to to model reception by a 3-microphone triangular array with T60 times in the range from 0 ms to 900 ms. Noise sources are taken from internally collected sets in conditions like cafes, kitchens and cars, freely available noise sources Getty [40] and YouTube Audio Library [41]. Randomly selected speech from the training sets is used as a noise source to simulate multitalker conditions. All noise sources are reverberated to simulate a point source and added to the target signal at SNRs between -10 and 30 dB. The SpeechCleaner parameters are optimized on roughly 6 seconds of noise. Note, that similar to [3] the noise context is randomly replaced by either target speech or white noise to prevent an over reliance on the SpeechCleaner output. During training each example is split into 4 s long segments.

3.2. Evaluation set

The models are evaluated on two datasets. The first set is a new simulated multi-talker set, called LibriMird. It is created by randomly drawing two utterances from different speakers from the test-clean subset of LibriSpeech, convolving them with different real-world room impulse responses (RIRs) from the MIRD set [28] and adding the resulting reverberated speech signals at three different SNRs (-5 dB, 0 dB, 5 dB). The MIRD database consists of real RIRs from 9 different recording settings with one of three RT60 times (160 ms, 360 ms, 610 ms). For target and overlapping speech two RIRs from the same recording settings, with a minimum 15° difference in the angle to the microphone array, are selected.

The second set is a rerecorded test set introduced in [3], which is designed to mimic the voice-search use case by utilizing shorter queries (¡10 s). Here, desired speech and noise were re-recorded separately in a living-room lab using a four microphone array. 100 different queries were played through a speaker from 7 different positions 4 m from the microphone array at a height of 1.5 m at a volume of 40 dB over ambient room noise. This set includes two different types of noise, speech-based from a movie and non-speech environmental noise. Speech and noise from different locations were mixed at five different SNR levels. A noise context of at least 6 s is added to both evaluation sets.

Table 1: *WER results for the estimated signal of the the first speaker on the LibriMird set.*

System	#Params	LibriMird		
		5 dB	0 dB	−5 dB
TfGridNet	8.20 M	12.1	12.5	13.5
Cleanformer	6.50 M	13.3	14.4	16.8
RnnCleanformer	7.61 M	12.4	12.9	13.8
TfCleanformer	7.63 M	12.2	12.7	13.4

4. Experiments

The presented systems are compared in terms of word error rate (WER) improvement for ASR using the ASR model presented in [29].

4.1. Source Separation

To compare Cleanformer and the proposed adaptation to Cleanformer to the state-of-the-art, all models are evaluated on the described source separation dataset and compared to TfGridNet. Note, that only the WER of the first speaker is evaluated since the presented model performs speech extraction. For TfGridNet the first speaker estimate is found by calculating the SI-SNR between both output signals and the target speech signal and calculating the WER for the output that leads to the higher SI-SNR. The results are displayed in table 1. TfCleanformer achieves results comparable to TfGridNet without access to non-causal information. This disadvantage seems to be compensated by the additional information provided by the SpeechCleaner output signal. Cleanformer shows a slightly worse performance despite using the same auxiliary information.

4.2. Noise reduction

To further evaluate the presented model it is compared to Cleanformer on the re-recorded data set for 5 different SNR settings for both speech and environmental noise. The results are presented in table 2.

All models achieve similar performance to the unprocessed signal (3.9 dB) on the Clean set (∞), which shows that they remove the distortions introduced by SpeechCleaner. The proposed TfCleanformer variant using Conformer CALs achieve the strongest results outperforming both Cleanformer as well as RNNCleanformer. While RNNCleanformer outperforms the original Cleanformer for most scenarios it results in an increased WER on the signal recorded at -12 dB. Therefore, the following discussion is focusing on the Conformer-based architecture instead of the original RNN-based preprocessors.

Overall, the TfCleanformer model achieves on average a relative improvement of 15.9 % over Cleanformer on signals distorted by speech-like noise and a relative improvement of 5.9 % in case of environmental noise. These results imply that the main advantages of TfCleanformer over Cleanformer is its improved performance on speech-like noise. One possible explanation for this difference is the sub-band processing module in the preprocessors. Operating on each frequency individually could benefit the separation of sparse speech source signals more than the more global environmental noise.

4.3. Ablation study

To evaluate which of the changes to Cleanformer have the greatest influence on the performance a short ablation study is performed on the re-recorded evaluation set. The results are presented in table 3.

Table 2: *Comparison of enhancement systems on the re-recorded evaluation set at different SNR. The first five rows depict the WER results on speech-like noise while the last five rows represents the WER results in scenarios with environmental noise. The best result per scenario is shown in gray.*

Model	SNR in dB					
	∞	-12	-6	0	6	12
-	3.9	97.9	89.6	52.7	19.2	7.5
SpeechCleaner	5.0	17.2	10.1	6.8	6.0	5.5
Cleanformer [3]	4.0	14.9	7.2	5.2	4.7	4.6
RnnCleanformer	3.9	15.2	7.1	5.0	4.3	4.3
TfCleanformer	3.8	12.4	5.8	4.3	4.1	4.0
-	3.9	52.8	40.5	24.4	11.8	6.3
SpeechCleaner	5.0	17.2	10.8	8.0	7.1	6.0
Cleanformer [3]	4.0	9.9	6.1	4.9	4.5	4.3
RnnCleanformer	3.9	11.6	6.0	4.5	4.2	4.0
TfCleanformer	3.8	9.7	5.9	4.5	4.2	3.9

While both the phase correction and the additional mask have a positive influence on the performance they alone do not lead to a decrease in WER compared to Cleanformer. Only after adding the preprocessing layers does the frequency domain model outperform the log-mel domain Cleanformer.

Note, that the base frequency domain model without a second mask or phase correction shows a similar behavior as Cleanformer, where the results on environmental noise is a lot better than on speech-like noise. Adding the masked phase correction reverses this behavior and leads to a better performance on the speech-like noise than on the environmental noise. Suggesting that the phase correction is especially beneficial in case of overlapping speech.

To compare the models processing speed we report the averaged real-time factor (RTF) of 1000 runs on a single core of a CPU, Intel(R) Xeon(R) CPU @ 2.20GHz. Here, we define RTF as processing time divided by the audio length. Although all considered models are able to perform real-time processing, the added preprocessing layers do significantly increase processing time, increasing the RTF from 0.010 to 0.389. Removing the preprocessors reduces the RTF to 0.014, which is only slightly higher than the Cleanformer RTF.

5. Conclusion

This paper introduces a novel streaming enhancement system based on Cleanformer, taking inspiration from TfGridNet. The presented model achieves state of the art performance on a challenging source separation dataset and outperforms Cleanformer on all considered datasets with an SNR above -12 dB. Further investigation into the propossed model architecture is necessary to improve processing speed and evaluate the influence of the preprocessor sub-modules on the enhancement performance.

Table 3: *Ablation study of the changes to Cleanformer on rerecorded speech with environmental (Env) and speech-like noise (Speech). The presented results are the average of WER at all 5 considered SNRs per noise type. RTF is the real-time factor of each model for a single core of a CPU, Intel(R) Xeon(R) CPU @ 2.20GHz.*

Model	Clean	Speech	Env	RTF
Cleanformer	4.0	7.3	5.9	0.010
TfCleanformer	3.8	6.1	5.6	0.389
- Preprocessors	3.9	7.9	7.5	0.014
- Second mask	3.9	10.9	12.2	0.014
- Phase correction	3.9	19.7	14.6	0.012

6. References

[1] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[2] A. Narayanan, C.-C. Chiu, T. O'Malley, Q. Wang, and Y. He, "Cross-attention conformer for context modeling in speech enhancement for ASR," in *Proc IEEE ASRU*, 2021.

[3] J. Caroselli, A. Narayanan, N. Howard, and T. O'Malley, "Cleanformer: A Multichannel Array Configuration-Invariant Neural Enhancement Frontend for ASR in Smart Speakers," in *Proc. IEEE ICASSP*, 2023.

[4] R. Maas, S. H. K. Parthasarathi, B. King, R. Huang, and B. Hoffmeister, "Anchored speech detection," in *Proc. InterSpeech*, 2016.

[5] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[6] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, 2019.

[7] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE ASRU*, 2015.

[8] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE ASRU*, 2015.

[9] R. Rehr and T. Gerkmann, "An analysis of noise-aware features in combination with the size and diversity of training data for DNN-based speech enhancement," in *Proc. IEEE ICASSP*, 2019.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[11] Y. Luo, "A time-domain real-valued generalized wiener filter for multi-channel neural separation systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3008–3019, 2022.

[12] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. IEEE ICASSP*, 2023.

[13] Y. Huang, T. Z. Shabestary, and A. Gruenstein, "Hotword Cleaner: Dual-microphone Adaptive Noise Cancellation with Deferred Filter Coefficients for Robust Keyword Spotting," in *Proc. IEEE ICASSP*, 2019.

[14] Y. Huang, T. Z. Shabestary, A. Gruenstein, and L. Wan, "Multi-Microphone Adaptive Noise Cancellation for Robust Hotword Detection," in *Proc. InterSpeech*, 2019.

[15] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *Proc IWAENC*, 2022.

[16] 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 *Proc. IEEE ICASSP*, 2021.

[17] J. Heitkaemper, A. Narayanan, T. Z. Shabestary, S. Panchapagesan, J. Walker, B. Gajare, S. Regev, A. Dudani, and A. Gruenstein, "Improving Acoustic Echo Cancellation for Voice Assistants using Neural Echo Suppression and Multi-Microphone Noise Reduction," in *Proc. IEEE ICASSP*, 2024.

[18] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A Dissecting Approach," in *Proc. IEEE ICASSP*, 2020.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. InterSpeech*, 2020.

[20] R. Talmon, I. Cohen, and S. Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[21] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

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

[23] N. Howard, A. Park, T. Z. Shabestary, A. Gruenstein, and R. Prabhavalkar, "A Neural Acoustic Echo Canceller Optimized Using An Automatic Speech Recognizer and Large Scale Synthetic Data," in *Proc. IEEE ICASSP*, 2021.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Yalta, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. InterSpeech*, 2018.

[25] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, . Hayashi, C. B"oddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT Workshop*, 2021.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015.

[27] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.

[28] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IWAENC*, 2014.

[29] T. N. Sainath, Y. R. He, A. Narayanan, R. Botros, R. Pang, D. J. Rybach, C. Allauzen, E. Variani, J. Qin, A. Gruenstein *et al.*, "An Efficient Streaming Non-Recurrent On-Device End-to-End Model with Improvements to Rare-Word Modeling," in *Proc. Interspeech*, 2021.