

ConSep: a Noise- and Reverberation-Robust Speech Separation Framework by Magnitude Conditioning

Kuan-Hsun Ho

Department of Computer Science and
Information Engineering
National Taiwan Normal University
Taipei, Taiwan
Email: jasonho610@ntnu.edu.tw

Jeih-weih Hung

Department of Electrical Engineering
National Chi Nan University
Nantou, Taiwan
Email: jwhung@ncnu.edu.tw

Berlin Chen

Department of Computer Science and
Information Engineering
National Taiwan Normal University
Taipei, Taiwan
Email: berlin@ntnu.edu.tw

Abstract—Speech separation has recently made significant progress thanks to the fine-grained vision used in time-domain methods. However, several studies have shown that adopting Short-Time Fourier Transform (STFT) for feature extraction could be beneficial when encountering harsher conditions, such as noise or reverberation. Therefore, we propose a magnitude-conditioned time-domain framework, ConSep, to inherit the beneficial characteristics. The experiment shows that ConSep promotes performance in anechoic, noisy, and reverberant settings compared to two celebrated methods, SepFormer and Bi-Sep. Furthermore, we visualize the components of ConSep to strengthen the advantages and cohere with the actualities we have found in preliminary studies.

Index Terms—speech separation, reverberation, cross-domain, multi-resolution, magnitude, conditioning

I. INTRODUCTION

SPEECH separation is a specific scenario of the source separation problem, where the targets are the overlapping speech signal sources while other irrelevant signals are treated as interferences. Recently, the field of speech separation has been revolutionized with the advent of deep learning techniques. The previous works toward an anechoic environment [1]–[7] have been fruitful and inspiring. Current systems rely, in large part, on the prestigious EMD structure, which is composed of the Encoder, Mask estimator, and Decoder. However, the assumption of an anechoic environment might be unrealistic in most speech separation studies. In practice, speech usually coincides with various interferences, and speech separation under reverberant situations is incredibly challenging [8].

Interestingly, we have observed opposing performances in [9] and [10]. In [9], the experimental results show that a smaller-sized SepFormer (SepFormer-s), i.e., the SepFormer with fewer Inter- and Intra-Transformers, could not perform well under reverberant conditions despite still giving a competitive result under an anechoic condition. After optimizations toward reverberant conditions, the resulting SepFormer is marginally superior to the optimized PIT-BLSTM. Moreover, it fails to be competitive in an anechoic condition. (See [9, Tab. 6]) Meanwhile, [10] expands the work in [7] on a regular-sized SepFormer and claims good results on multiple conditions. One explanation may be the advantage of exploiting more parameters, which enables the model to represent more

complicated functions than the one with fewer parameters. Nonetheless, this evidences that efforts to make SepFormer a more distilled yet versatile model need further investigation.

One endeavor that adapts SepFormer to multiple conditions is Bi-Sep [11]. Bi-Sep leverages two parallel encoders with different time resolutions and a Bi-projection Fusion (BPF) [12] module to integrate information from different domains. However, Bi-Sep has a potential shortcoming. Despite exhibiting better performance than existing models when facing complicated environments, Bi-Sep inevitably inherits the degradation when substituting the learnable encoder with the Short-Time Fourier Transform (STFT). Moreover, although the BPF module helps determine whether the mask estimator should attend more on a shorter or longer frame, it could cause the mask estimator to have difficulty learning from two domains when encountering a simpler environment.

Therefore, we propose a novel speech separation framework, ConSep, which exploits conditioning on magnitude spectrogram to avoid any domain mismatch or confusion. A better conditioning method could facilitate the knowledge injection instead of concatenating all the features as Bi-Sep does. ConSep likewise embraces two encoders with different time resolutions to retrieve complementary characteristics. However, we merge the respective output features into the same dominant domain, in this case, the time domain. More precisely, we modulate the time signals by magnitude spectrogram as this modulation enables the mask estimator to better distinguish speech parts and interferences during separation. Experiments show that ConSep surpasses SepFormer under an anechoic condition and prominently upgrades SepFormer under more complicated situations. This result matches our goal of enabling a model to possess the generalizability on multiple conditions.

II. FINDINGS

To ascertain what fosters a model to cope with various environments, we list the optimal model configurations of relevant studies that adapt EMD models to more complicated conditions in Tab. I. We notice that an almost unanimous consensus is to employ the time-domain loss, including SI-SDR [14] and th-SDR [15]. The time-domain loss has been

TABLE I

BEST MODEL CONFIGURATION FROM REFERENCES THAT ADAPT EMD MODELS TO MULTIPLE CONDITIONS AND THE ORIGINAL SEPFORMER.

Condition(s)	Attributes		
	Loss function	Granularity (ms)	Encoder/Decoder pair
<i>Anechoic</i>			
SepFormer [7]	SI-SDR	2	Learnable
Conv-TasNet [16]	SI-SDR	0.5	Learnable ³
<i>Reverberant</i>			
SepFormer-s [9]	th-SDR	64	STFT ⁴
SepFormer [10]	SI-SDR	2	Learnable
Conv-TasNet [16]	SI-SDR	8	STFT ³
DenseUNet-TCN [18]	L1 ¹	25	STFT
WD-TCN [13]	SI-SDR	2	Learnable
Bi-Sep32 [11]	SI-SDR	2/32 ²	Both

¹ L1 Loss on real, imaginary and magnitude.

² Frame length of 2 and 32 ms for Learnable and STFT, respectively.

³ Performance reduces by less than 2 (dB) if altered to the other pair.

⁴ Performance reduces by less than 1 (dB) if altered to the other pair.

proven beneficial in numerous works, even when using STFT as an encoder [9], [16]. However, the granularity designated by each work varies over a wide range, from 0.5 ms to 64 ms. The time-domain methods usually work on short frames, whereas traditional STFT frame sizes are set to be larger (around 32 ms). As for the type of encoder-decoder pair, [16] and [17] have argued that it is not the crucial factor to success.

Furthermore, through our preliminary studies, we discover four actualities:

- 1) Time-domain methods usually perform better in SI-SDR and worse in PESQ than STFT methods [18]–[20].
- 2) A large enough window size is mandatory to avoid contravening the prerequisite of Multiplicative Transfer Function Approximation (MTFA) [9], [21]. On the contrary, recent works have gained success due to the fine-grained window size [16].
- 3) Employing STFT representation exhibits optimal performance in reverberation. However, employing the learnable encoder/decoder prevails under an anechoic condition [9], [16].
- 4) The phase becomes uninformative within a relatively large window size [22]–[24].

Although some actualities seem contradictory, a reasonable scheme to incorporate all the beneficial characteristics can compensate for those contradictions. This motivates us to propose ConSep. To encourage more instantiations in the future, we plan to provide publicly available codes used in our experiments.

III. OUR CONSEP

The high-level description of ConSep is identical to the EMD structure. Initially, the encoder transforms the mixture $x \in \mathbb{R}^T$, which contains audio from K active speakers and interferences, into a representation w that characterizes the signal. Then the mask estimator produces K masks $\{m_k\}$ for each active speaker in the mixture. Finally, the decoder reconstructs the separated K source signals in the time domain, each represented by $\hat{s}_k \in \mathbb{R}^T$.

A. Encoder

The encoder of ConSep is composed of four modules: Learnable encoder, STFT, Multi-Channel Attention (MulCA), and Modulator.

1) *Learnable encoder*: The typical method uses one-dimensional convolutional layers (Conv1d) followed by a rectified linear unit (ReLU). This encoder extracts the time representation $w_c \in \mathbb{R}^{N \times L}$ from the mixture x :

$$w_c = \text{ReLU}(\text{Conv1d}(x)). \quad (1)$$

2) *STFT with MulCA*: The conventional method obtains the time-frequency representation $X \in \mathbb{C}^{F \times L}$, a.k.a spectrogram, through STFT:

$$X[f, l] = \sum_{n=0}^{W-1} x[n + Hl]w[n]e^{-j\frac{2\pi fn}{W}}, \quad (2)$$

where f , l , and n are the frequency bin, frame, and local time indices; W is the window length, and H is the hop size. After, we extract the magnitude part $X_m \in \mathbb{R}^{F \times L}$ from the spectrogram X . Prior to the modulation, we add a MulCA block [25], [26] to weigh the magnitude spectrogram X_m . MulCA regards different frequency bins as channels, giving them different weights using Channel Attention.

The intuition behind MulCA is that the energy of a speech utterance usually distributes non-uniformly in frequencies, and different frequency components are unequally crucial to human perception [27], [28]. For example, the lower frequency band tends to contain high energies, tonalities, and long-duration sounds; the higher frequency band may have low energies, noise, and rapidly decaying sounds. The following equations express the operations of a MulCA:

$$\begin{aligned} c_i &= \text{ReLU}(\text{AvgPool}(\text{Conv1d}(X_m; k_i))), i = 0, 1, 2, \\ c &= \text{FCN}([c_0, c_1, c_2]), c \in \mathbb{R}^F, \\ C &= \text{Broadcast}(c), C \in \mathbb{R}^{F \times L}, \\ \tilde{X}_m &= X_m \odot C. \end{aligned} \quad (3)$$

The frames in X_m are passed through three Conv1d with different kernel sizes: k_0 (small), k_1 (middle), and k_2 (large). Each is followed by an average pooling (AvgPool) and ReLU activation to deliver a weight vector c_i . Afterward, a two-layer down-up-sampled fully connected network (FCN) merges three weight vectors to create frequency-wise weights c . Then we broadcast the weights to operate element-wise multiplication with X_m to get a weighted version \tilde{X}_m —accordingly, all the frames in the magnitude spectrogram share an identical frequency-dependent emphasis.

3) *Modulator*: As mentioned earlier, we aim to explore a better conditioning method to remain in the same domain without encountering any domain conflicts. The mask estimator of SepFormer has shown its powerful ability to model sources from overcomplete features in the time domain. Hence, we build ConSep accordingly, viz., analyzing time-domain features, but additionally, the time signals are conditioned on the magnitude spectrogram. We rely on Feature-wise

Linear Modulation (FiLM) [29]. The FiLM allows adjusting the time signals to more appropriate representations based on the energy information of the given spectrogram. The residual connection is added after the FiLM layer to ensure the architecture performs well when the magnitude is relatively small. The modulating process can be formulated as:

$$w = w_c + f_1(\tilde{X}_m) \odot w_c + f_2(\tilde{X}_m), \quad (4)$$

where w denotes the modulated feature, and each f_i denotes an affine transformation.

B. Mask estimator

The mask estimator inputs modulated feature w and estimates a mask m_k . Firstly, the modulated feature w is layer-wise normalized, chunked into overlapping segments with an overlap factor of 50%, and then stacked.

Afterward, the stacked feature feeds the SepFormer blocks, which exploit the dual-path mechanism [4]. The underlying process first captures the short-term dependencies by Intra-Transformer, then extracts the long-term dependencies by Inter-Transformer, and repeats D times. The unit Transformer structure used in Intra- and Inter-Transformer includes a multi-head attention (MHA) stage and a feed-forward (FFW) block with pre-LN setting [30] and skip connections. Unlike DPTNet [6], positional encoding is applied for injecting information on the order of sequence instead of a recurrent neural network (RNN). The total number of unit Transformers employed in Intra- and Inter-Transformer is E . A linear layer further processes the output of the SepFormer block to project the feature dimension for K times deep.

Finally, the projected output is passed through an overlap-add stage, two concurrent FFWs, and a ReLU activation to obtain the non-negative mask m_k . Note that we eliminate the bottleneck projections in the original SepFormer, as our prior experiment indicates its redundancy.

C. Decoder

The decoder is a transposed convolution layer with the same stride and kernel size as the learnable encoder. The input to the decoder for active speaker k is the element-wise multiplication between the mask m_k and the modulated feature w :

$$\hat{s}_k = \text{Conv1d}^T(m_k \odot w). \quad (5)$$

IV. EXPERIMENTAL SETUP

A. Datasets

We validate the presented method on the popular WSJ0-2mix dataset [1] for the anechoic setting using the improvement of SI-SDR and SDR as the evaluation metrics. The training, validation, and test sets contain 30, 10, and 5 hours of speech data. The speech data are sampled at 8 kHz. Furthermore, we perform experiments in noisy settings. We rely on WHAM! [31] with urban noise and WHAMR! [8], which adds reverberation on top of WHAM!. These datasets are derived from WSJ0-2mix and have identical statistics.

TABLE II
SEPARATION PERFORMANCE AND SPEECH QUALITY METRIC ON MULTIPLE CONDITIONS.

Condition(s)	Metrics		
	SI-SDRi	SDRi	NB-PESQ
<i>Anechoic</i>			
SepFormer-s	16.53	17.02	3.12
Bi-Sep32	16.28	16.49	-
ConSep	16.72	17.19	3.39
<i>Noisy</i>			
SepFormer-s	13.23	14.13	2.50
Bi-Sep32	13.62	14.54	-
ConSep	13.82	14.82	2.76
<i>Noisy & Reverberant</i>			
SepFormer-s	5.90	8.95	2.12
Bi-Sep32	6.37	9.09	-
ConSep	6.50	9.07	2.30

TABLE III
ABLATION STUDY. THE THIRD AND FOURTH COLUMNS USE DIFFERENT CONDITIONING METHODS WITHOUT MULCA AS WELL.

	ConSep	w/o MulCA	w/o FiLM (concat+linear)	w/o FiLM (add)	w/o conditioning
SI-SDRi	13.82	13.72	13.59	13.34	13.23
SDRi	14.82	14.68	13.96	14.19	14.13

B. Training setup

In the case of the learnable encoder, the encoder basis N is set to 256. The input kernel size is 16 with a stride factor of 8. As for the STFT encoder, we use the Hamming window with a length of 256 (32 ms at 8 kHz), and the hop size is the same as that for the learnable encoder. For the MulCA, three kernel sizes k_0 , k_1 , and k_2 are 3, 5, and 10, respectively. Regarding the mask estimator, we follow the configuration proposed in [7], whereas E is reduced to 4. For model training, we optimize the model using the Adam optimizer, a batch size of 1, and a learning rate of 1.5e-4. Finally, the model is trained over 150 epochs with utterance-level Permutation Invariant Training (uPIT) [2] and SI-SDR losses.

V. RESULTS AND ANALYSES

A. Comparison and Ablation study

We compare the separation accuracy of ConSep with SepFormer-s and Bi-Sep as baselines. Tab. II presents the experimental results in terms of both SI-SDRi and SDRi. It demonstrates the advantages of our proposed model. For all kinds of environments, ConSep outperforms all other methods except the SDRi, which can be deceived by the loudness [14], in noisy and reverberant settings. Moreover, the fact that a better conditioning strategy not only gains stability but succeeds to the beneficial sides is proven, as ConSep surpasses Bi-Sep in non-anechoic settings and SepFormer in the anechoic setting.

Furthermore, we show the evaluation in speech quality metric. We can observe that employing the STFT features may improve narrow-band PESQ (NB-PESQ), and the same phenomenon can be observed in [18]–[20]. This adds another clue of using a larger frame time-frequency representation still presenting as a desirable feature, even if learned-domain

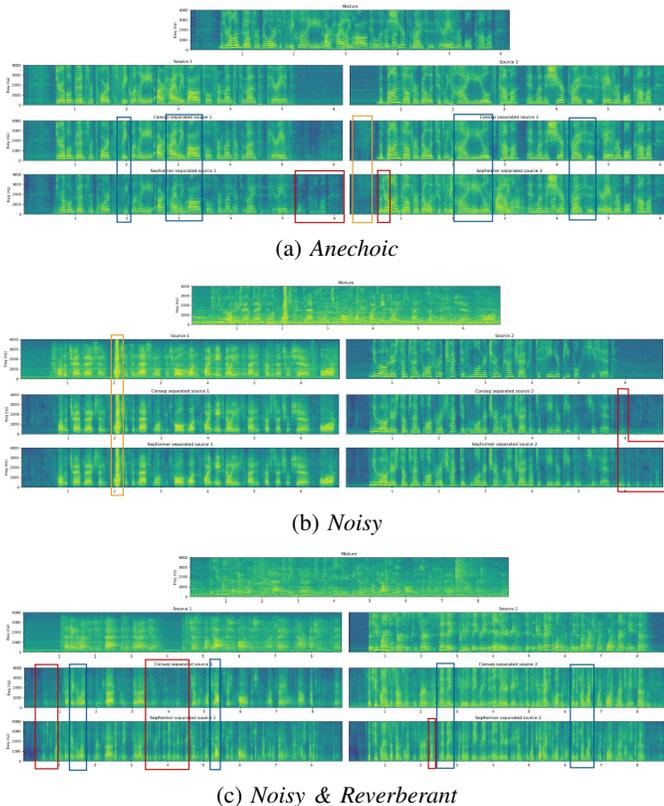


Fig. 1. Case studies. Generally, the rows indicate the spectrogram of mixture, sources, ConSep output, and SepFormer output from top to bottom. The two columns indicate the first and second sources from left to right. Also, red and blue boxes denote false alarm issues and spectral/harmony clarity. For (a) and (b), the non-speech signals cropped in the orange box are the sounds of inhaling and microphone pop, respectively.

models have more freedom to adapt to the SI-SDR training loss.

To validate the effectiveness of ConSep, we also conduct an ablation study in noisy environments, as shown in Tab. III. We validate the performance without MulCA and magnitude conditioning. Additionally, we experiment with various conditioning methods, including simply adding or concatenation followed by a down-sizing linear layer. The result shows that with attention to magnitudes, the separation performance improves, as "concat+linear" performs closely to Bi-Sep. However, employing FiLM brings the most apparent improvement among other methods. (see the entry "w/o MulCA")

B. Visualization

We visualize the mixture, clean sources, and separated outputs from ConSep and SepFormer, as shown in Fig. 1. The mixture in Fig. 1(a) consists of two women speaking with a similar pitch range. We can see that the frequency contour is more prominent in ConSep-separated sources and that SepFormer tends to produce false alarms. This implies that merely analyzing time signals from an overcomplete set of encoder bases renders the model confusing when facing speakers with similar pitch identities. As for Fig. 1(b), the

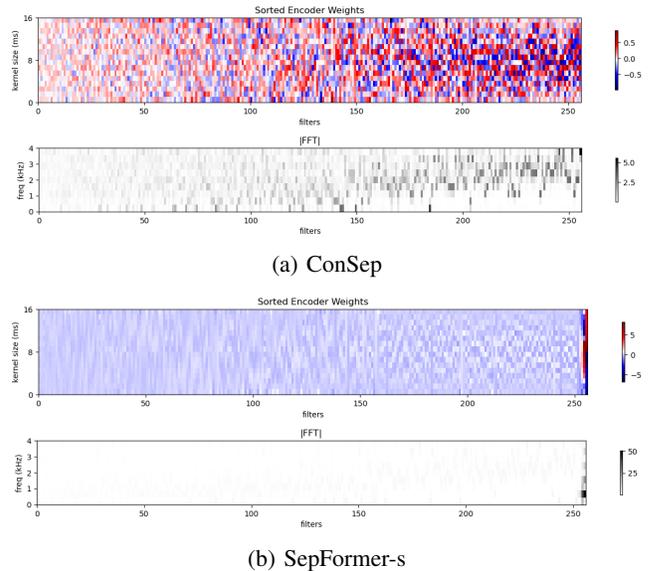


Fig. 2. For each sub-figure, the upper and lower panel depict the encoder bases sorted by Euclidean similarity and their frequency response, respectively.

mixture consists of a woman, a man speaking, and cafeteria noises. We can see that both models denoise well, but likewise, some speech-related false alarms occur. The same observation can also be pointed out in Fig. 1(c), whose mixture consists of the reverberant version of two men speaking with poll hall noises. Regardless, the superiority of a more apparent contour in ConSep remains, probably due to the enhancement by attending the spectrogram that better presents the harmonics. This concludes that the essential components of speech signals are better captured in ConSep.

Furthermore, we plot the sorted bases of the learnable encoder trained in anechoic condition and their frequency response, as shown in Fig. 2. Resembling [3], most filters are tuned to lower frequencies. This suggests an essential role for low-frequency speech features such as pitch to achieve better performance. However, we notice a few weird filters on the right side of Fig. 2(b), which can translate to the high-valued low-pass filter at the end of the frequency response. This may be the reason why SepFormer could not perform well when attention to high-frequency information is required, such as the circumstances when facing similar-pitch speakers.

VI. CONCLUSIONS

In this study, we propose a noise- and reverberation-robust speech separation framework, ConSep, by means of conditioning the time signals by magnitude spectrogram. The goal of generalizability has been fulfilled as this framework upgrades an existing model to fit various environments. Furthermore, we analyze and visualize the results to get a better picture of the advantages of ConSep, which as well demonstrates phenomena coherent with the actualities found through preliminary studies.

REFERENCES

- [1] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *INTER-SPEECH*, 2016.
- [2] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
- [4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP*, 2020.
- [5] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End speech separation by speaker clustering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2840–2849, 2021.
- [6] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *INTERSPEECH*, 2020.
- [7] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *ICASSP*, 2021.
- [8] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *ICASSP*, 2020.
- [9] T. Cord-Landwehr, C. Boeddeker, T. v. Neumann, C. Zorila, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," *arXiv preprint arXiv:2111.07578v2*, 2022.
- [10] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "On using transformers for speech-separation," *arXiv preprint arXiv:2202.02884*, 2022.
- [11] K. Ho, J. Hung, and B. Chen, "Bi-Sep: A Multi-Resolution Cross-Domain Monaural Speech Separation Framework," in *TAAI*, 2022.
- [12] F. A. Chao, J. W. Hung, and B. Chen, "Cross-Domain Single-Channel Speech Enhancement Model with BI-Projection Fusion Module for Noise-Robust ASR," in *ICME*, 2021.
- [13] W. Ravenscroft, S. Goetze, and T. Hain, "Utterance Weighted Multi-Dilation Temporal Convolutional Networks for Monaural Speech Dereverberation," in *IWAENC*, 2022.
- [14] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP*, 2019.
- [15] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems*, 2020.
- [16] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *ICASSP*, 2020.
- [17] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *ICASSP*, 2020.
- [18] Z. Q. Wang, G. Wichern, and J. L. Roux, "On the Compensation Between Magnitude and Phase in Speech Separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.
- [19] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the Best Loss Function for DNN-Based Low-latency Speech Enhancement with Temporal Convolutional Networks," *arXiv preprint arXiv:2005.11611v3*, 2020.
- [20] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient Transformer-based Speech Enhancement Using Long Frames and STFT Magnitudes," in *INTERSPEECH*, 2022.
- [21] Y. Averbach 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.
- [22] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.
- [23] T. Peer and T. Gerkmann, "Intelligibility Prediction of Speech Reconstructed From Its Magnitude or Phase," in *ITG Conference on Speech Communication*, 2021.
- [24] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *arXiv preprint arXiv:2203.16222*, 2022.
- [25] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullSubNet+: Channel Attention Fullsubnet with Complex Spectrograms for Speech Enhancement," in *ICASSP*, 2022.
- [26] Y. Tsao, K. Ho, J. Hung, and B. Chen, "Adaptive-FSN: Integrating full-band extraction and adaptive sub-band encoding for monaural speech enhancement," in *SLT*, 2022.
- [27] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *WASPAA*, 2017.
- [28] Rong Chao, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao, "Perceptual contrast stretching on target feature for speech enhancement," in *INTERSPEECH*, 2022.
- [29] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI-18*, 2018.
- [30] R. Xiong et al., "On layer normalization in the transformer architecture," in *ICML*, 2020.
- [31] G. Wichern et al., "Wham!: Extending speech separation to noisy environments," in *INTERSPEECH*, 2019.