Speech Enhancement Network With Windowed Cross Attention Using Noise-Reference Microphone

Kota Suzuki*, Yosuke Sugiura† and Tetsuya Shimamura‡
* Saitama University, Japan
E-mail: k.suzuki.634@ms.saitama-u.ac.jp Tel/Fax: +8170-40725224
† Saitama University, Japan
E-mail: ysugiura@mail.saitama-u.ac.jp Tel/Fax: +8148-8583496
‡ Saitama University, Japan
E-mail: shima@mail.saitama-u.ac.jp Tel/Fax: +8148-8583496

***Abstract*—In this study, we propose an enhancement to a multi-channel speech enhancement network that leverages a noise-reference microphone. Our primary contribution is the introduction of a Windowed Cross-Attention (WCA) mechanism, designed to more effectively fuse features from a primary microphone capturing mixed speech and a secondary microphone capturing environmental noise. This paper presents a comparative evaluation of our proposed method (WCA) against several baseline architectures: a standard multi-channel DEMUCS (DEMUCS-m), our prior noise-fusion architecture (DEMUCS-prop), and a deep non-linear filter (DNLF). Experiments conducted in a simulated acoustic space assess each method in terms of objective speech quality metrics, followed by detailed analysis and discussion.**

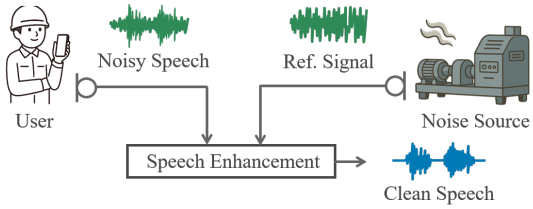


Fig. 1: Our scenario for speech enhancement using noise-reference microphone.

I. Introduction

Speech enhancement aims to improve the clarity and intelligibility of speech signals corrupted by background noise. This technique is fundamental in a wide range of applications, including automatic speech recognition (ASR), mobile communication, hearing aids, and voice-controlled systems. Recently, the increasing use of Internet of Things (IoT) devices has led to new possibilities for speech enhancement using distributed microphones [1]. Although these systems can capture sound from multiple locations, their performance often degrades in environments with high levels of noise. In this paper, we address this problem by employing a noise-reference microphone placed close to the primary noise source, as shown in Figure 1. This configuration allows the system to obtain a reference signal that is dominated by noise, which is then used to more effectively remove the noise component from the target speech.

Many conventional multi-channel speech enhancement systems assume a non-distributed, co-located microphone array to utilize spatial information. In classical beamforming, the directions of arrival (DOAs) of speech and noise must be known. More recently, neural beamformers have been developed that can automatically estimate source directions and perform optimal directional filtering [2]–[4]. These methods use attention mechanisms to compute time-frequency correlations between the inputs from two microphones to determine the optimal filter gains. However, a fundamental assumption in these approaches is the use of a fixed microphone array with a known geometry. They do not work effectively in distributed microphone environments where the distance and relative positions between microphones can change dynamically. This work targets specific industrial applications, such as suppressing noise from a particular machine on a factory floor, where the dominant noise source is identifiable and a reference microphone can be placed in close proximity. The focus of this paper is therefore on proposing a method that makes the most effective use of the reference signal under these well-defined conditions.

In contrast, some approaches focus on multi-channel enhancement without assuming a fixed array geometry. For instance, DNLF [5] was proposed to extract speech features even when microphone and source positions are unknown. It processes spatial and spectro-temporal information jointly using 3D convolutions. So it offers flexibility for varied microphone placements. This approach, however, requires a high degree of correlation between the channel inputs to be effective. Consequently, its performance diminishes when microphones are placed far apart, as is common in distributed scenarios.

To address the limitations of existing methods in dynamic, distributed settings, we previously proposed DEMUCS with Reference-Guidance (Demucs-RG), a multi-channel speech enhancement architecture that operates effectively when microphone positions are unknown and variable [6]. This method explicitly stores noise information by fusing encoder features from the main microphone channels and the noise-reference

979-8-3315-7206-8/25/$31.00

channel. During training, the noise-reference microphone remains in a fixed position while the other microphones are moved to various locations, enabling the network to learn a generalized noise subtraction process independent of a specific spatial configuration.

In this paper, we extend our previous work, Demucs-RG, and improve its performance by introducing a Windowed Cross-Attention (WCA) framework. The WCA mechanism is conceptually similar to multi-head cross-attention (MHCA) but is specifically designed for multi-channel audio processing. In standard MHCA, attention is computed between all heads across the entire time axis. This process can be redundant and may risk performance degradation due to overfitting, as the most critical information for time-aligned audio signals typically exists between heads at the same temporal position. Based on this observation, WCA restricts the attention calculation to a limited window along the time axis. This architectural choice dramatically reduces the computational complexity compared to a standard MHA and, as we will show, generates a more effective attention mechanism for the task of noise cancellation.

The effectiveness of the proposed method was verified through objective evaluations. Experiments were conducted to compare our method with several conventional methods, using the PESQ and SI-SDR as metrics. The results show that the proposed method outperforms the other techniques, demonstrating that the WCA framework contributes to a significant improvement in speech enhancement performance.

II. Baseline Architecture

This section describes the baseline architecture used as the foundation for our proposed method. We selected Demucs, a model known for its time-domain U-Net structure, primarily for its high performance relative to its efficient use of parameters.

A. Model Selection

The proposed speech enhancement network is based on the Demucs architecture [7]. This selection is motivated by its excellent balance between a low number of parameters and high performance, which is important for building an efficient system.

As shown in Table I, Demucs provides an excellent trade-off between model size and performance compared to other related methods. While many models require substantially more parameters for high performance, Demucs achieves a competitive PESQ score of 3.07 with only 1.9M parameters, matching the performance of the much larger Uformer [8]. This high parameter efficiency makes it a suitable and robust baseline for our work.

B. Architecture Overview

Figure 2 shows the architecture of Demucs, which is based on a U-Net structure. The encoder path consists of several blocks, each containing a 1D convolution (`Conv1d`) followed by a Gated Recurrent Unit (`GRU`) to capture temporal sequences at different scales. At the bottleneck of the U-Net, two Long Short-Term Memory (`LSTM`) layers are used to process the high-level features. The decoder path mirrors the encoder, using blocks composed of a `GRU` and a transposed 1D convolution (`ConvTr1d`) for upsampling. Skip connections link corresponding encoder and decoder blocks, which helps preserve fine-grained details for high-quality signal reconstruction. This hybrid CNN-RNN architecture allows the model to effectively capture both local time-frequency features and their long-term dependencies.

TABLE I: Relationship Between Parameter Count and PESQ Values.

Method	Params. (M)	PESQ
CRN [9]	17.58	2.70
Zhong's [10]	13.00	3.70
FSN [11]	5.64	2.75
DCCRN [12]	3.67	2.13
Uformer [8]	3.34	3.07
DeepFilterNet [13]	1.8	2.81
Demucs [7]	1.9	3.07
DPCRN [14]	0.72	2.79

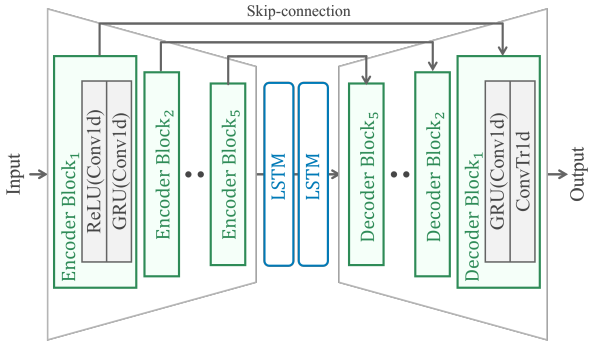


Fig. 2: Architecture of DEMUCS

III. Proposed Method

Building upon the Demucs baseline, this section introduces our proposed multi-channel speech enhancement method. Our primary goal is to effectively utilize a noise reference signal to improve performance in high-noise environments. To achieve this, we introduce two main architectural modifications: a dual-encoder branch structure to process the main and reference signals in parallel, and a Windowed Cross-Attention (WCA) module to effectively fuse the information from these two streams. The following subsections detail these components.

A. Overall Architecture

As shown in Figure 3(a), an important aspect of our method is the introduction of a dual-encoder branch structure. The main encoder processes the noisy input signal, while a parallel reference encoder processes the noise reference signal. This reference encoder has the same structure as the main encoder but does not share its parameters.

The novelty of our approach lies in the method used to fuse the information from these two streams. At each scale within the encoder, the feature map from the reference encoder is fused with the feature map from the main encoder via a Windowed Cross-Attention (WCA) module. This block-level fusion allows the main encoder to efficiently extract speech-related features while selectively suppressing noise characteristics identified from the reference branch. The output of the main encoder is then processed by the LSTM bottleneck and the decoder, following the baseline architecture, to reconstruct the enhanced speech signal.

B. Windowed Cross-Attention (WCA)

The detailed structure of the WCA module is illustrated in Figure 3(b). The module is designed to compare two fully-aligned feature segments (from the main and reference signals) at once. The WCA module operates as follows:

1) **Windowing**: The temporal dimension of both the `Input Feature` and the `Reference Feature` is first zero-padded to be a multiple of 32. The feature maps are then partitioned into 32 non-overlapping windows along this axis.
2) **Q, K, V Projection**: The Query (Q) vector is derived from each windowed `Input Feature`, while the Key (K) and Value (V) vectors are derived from the corresponding windowed `Reference Feature` using separate 1x1 convolutions.
3) **Attention**: The attention weights are calculated via a scaled-dot product of Q and K, which are then applied to the Value vector V for each window.
4) **Residual Connection**: The output of the attention mechanism is reassembled and added back to the original `Input Feature` after trimming the padding.

This WCA mechanism is distinct from standard multi-head cross-attention (MHCA). In MHCA, for instance, the feature space is divided into multiple subspaces (heads) and attention is computed across the entire time axis for each head. Our block-wise WCA is based on the property that for time-aligned audio signals, the most relevant information exists in a narrow temporal vicinity, thus avoiding the computational redundancy of MHCA.

Furthermore, this method is also distinct from Local Window Cross-Attention (Lwin-CA) [15]. Lwin-CA employs a query-wise sliding window, where each query dynamically attends to a local neighborhood of keys. In contrast, our WCA's block-wise fusion architecture first partitions the input into static windows, and then computes attention self-contained within each window. This design is better suited for comparing entire feature segments at once, rather than processing a sliding local context for every single time step.

C. Loss Function

Following the Demucs baseline, our model is trained using a composite loss function that combines a time-domain loss and a frequency-domain loss. The total loss consists of the L_1 loss on the raw waveform and a multi-resolution STFT loss on the spectrogram magnitude. Let x and y represent the clean and enhanced audio signals, respectively. The STFT loss (L_{stft}) is the sum of a spectral convergence loss (L_{sc}) and a log-magnitude loss (L_{mag}), defined as:

$$L_{\text{stft}}(x, y) = L_{\text{sc}}(x, y) + L_{\text{mag}}(x, y) \tag{1}$$

$$L_{\text{sc}}(x, y) = \frac{\| \, |\text{STFT}(x)| - |\text{STFT}(y)| \, \|_F}{\| \, |\text{STFT}(x)| \, \|_F} \tag{2}$$

$$L_{\text{mag}}(x, y) = \frac{1}{T} \| \log |\text{STFT}(x)| - \log |\text{STFT}(y)| \|_1 \tag{3}$$

Here, $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius and L_1 norms, respectively. The multi-resolution STFT loss is the sum of STFT losses computed with different analysis parameters (e.g., FFT sizes, window sizes). The final loss function is a weighted sum of the waveform loss and the multi-resolution STFT loss:

$$L(x, y) = \alpha \cdot \|x - y\|_1 + (1 - \alpha) \cdot \sum_{i=1}^{N} L_{\text{stft}}^{(i)}(x, y) \tag{4}$$

where N is the number of different STFT resolutions and α is a weighting factor.

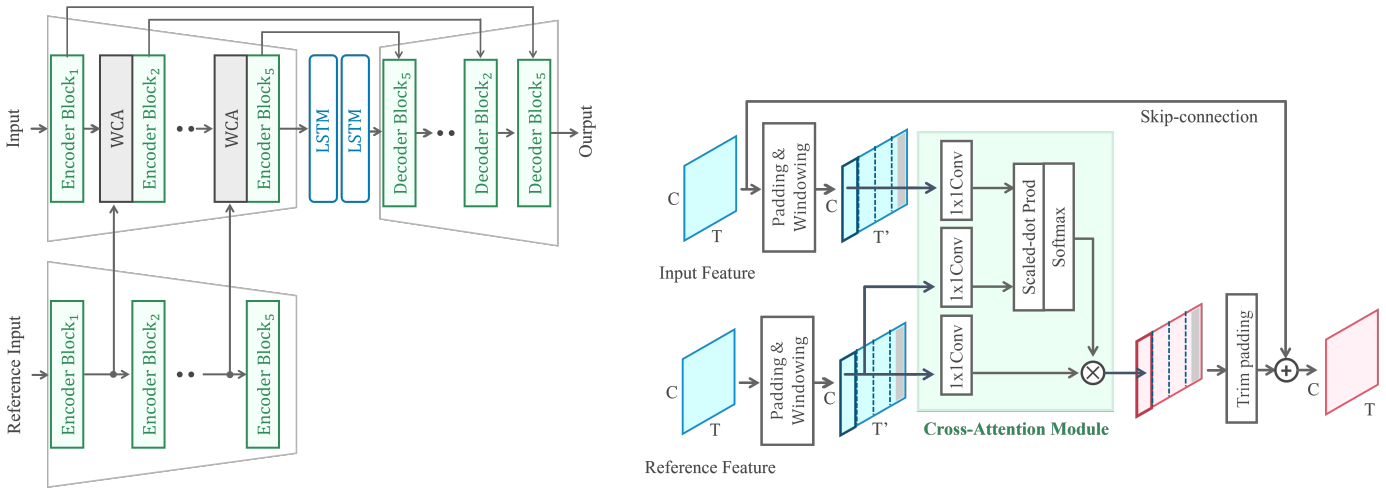
IV. Experiments

A. Experimental Setup

This section details the experiments conducted to evaluate our proposed method. We first describe the experimental setup, including the datasets, implementation details, and evaluation metrics. We then present and discuss the results, comparing our proposed model against several baseline methods to demonstrate its effectiveness.

1) Datasets and Conditions: The training and test datasets were generated by simulating various acoustic scenarios using the PyRoomAcoustics simulator **pyroomacoustics**. As illustrated in Figure 4, we defined a 15m × 15m room, and the reverberation time (RT60) was set to 0.3 seconds. All sources and microphones were placed at a height of 1.6m. The noise source and the dedicated reference microphone were placed at fixed positions of (7.5, 1.1) m and (7.5, 1.0) m, respectively. This placement is designed to simulate an ideal scenario where the dominant noise source has been identified and a reference microphone is appropriately placed in its vicinity. Such a controlled setup allows us to rule out confounding variables and specifically assess the effectiveness of the proposed Windowed Cross-Attention (WCA) mechanism in leveraging the reference noise signal. To ensure robustness to user location, the position of the user, who is assumed to be co-located with the main microphone, was randomized for each sample.

For the clean speech source, we used the dataset from Valentini et al. [16], which contains speakers from the same accent region (England). The training set utilizes speech from 14 male and 14 female speakers, while the test set uses 1 male and 1 female speaker unseen during training. For the noise source, we used 10 types of noise for training (2 artificial, 8 from the DEMAND database [17]) and 5 different, unseen noises from the DEMAND database for testing. All audio signals were downsampled to 16 kHz.



(a) Overall dual-encoder architecture.

(b) Detail of the Windowed Cross-Attention (WCA).

Fig. 3: Architecture of Proposed method.

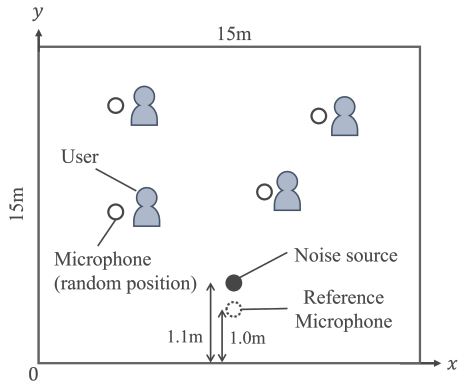


Fig. 4: Simulated experimental environment.

TABLE II: Details of the Original Dataset.

	Training Set	Test Set
Speaker	14 males, 14 females	1 male, 1 female
Number of speech data	11,572	824
Type of noise	10	5
SNR level [dB]	$-10 \sim 20$ dB	
Sampling rate	16 kHz	
Frame size	512	
Overlap	256	

For each generated sample, room impulse responses (RIRs) were calculated for the paths from the random user position and the fixed noise position to both the main and reference microphones. The amplitude of the noise source was then adjusted such that the signal-to-noise ratio (SNR) at the main microphone was between -10 dB and 20 dB. The final composition of the dataset is summarized in Table II.

2) Implementation Details: The network operates on audio sampled at 16 kHz, using a frame size of 512 (32 ms) and a frame shift of 256 (16 ms), resulting in a 50% overlap. All models were trained for 100 epochs with a batch size of 28. The Adam optimizer was used for training all models. The learning rates were set to 3e-5 for our proposed DEMUCS+WCA, 3e-4 for the baseline DEMUCS, 1e-2 for DNLF, and 3e-5 for our previously proposed fused DEMUCS.

We employed several data augmentation techniques used in the original Demucs implementation, including random shifts, Remix, BandMask, and Revecho. Random shift is applied to all datasets. Remix creates new noise samples by shuffling noises within a batch. BandMask applies a band-stop filter to remove 20% of frequencies sampled on the Mel scale. Revecho adds decaying echoes to the noise signal.

3) Evaluation Metrics: The performance of all models was evaluated using several standard objective metrics: Perceptual Evaluation of Speech Quality (PESQ) [18], Short-Time Objective Intelligibility (STOI) [19], and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [20]. We also report the composite metrics CSIG, CBAK, and COVL [21] for a comprehensive analysis.

B. Results and Discussion

1) Experimental Results: To quantitatively evaluate our proposal, a comprehensive comparative experiment was conducted. Our proposed method, which integrates Windowed Cross-Attention (WCA) into the Demucs baseline, is hereafter referred to as Demucs-WCA. The performance of Demucs-WCA was benchmarked against three competing methods: the baseline Demucs [7], our previously proposed Demucs with Reference-Guidance (Demucs-RG) [6], and a deep non-linear filter (DNLF) [5].

The experimental results validate the effectiveness of the proposed method under the specific condition that a single dominant noise source is present and the reference microphone is placed in its proximity. It is important to acknowledge, however, that in practical scenarios involving multiple or non-stationary noise sources, or where the reference microphone

is misplaced, the performance is expected to degrade. This underlying assumption thus constitutes a key limitation of the current work.

2) Objective Evaluation: A summary of the final scores at 100 epochs is presented in Table III. The proposed Demucs-WCA method consistently demonstrates superior performance across most metrics. It achieved the highest scores for PESQ (1.98), STOI (0.94), CBAK (2.75), COVL (3.70), and SI-SDR (11.00 dB). The most significant improvement is seen in background noise evaluation (CBAK), where the score of 2.75 indicates the most effective noise suppression. This directly contributes to its leading score in overall quality (COVL). While Demucs-RG obtained a marginally higher CSIG score, the proposed Demucs-WCA shows the best overall signal fidelity, with an SI-SDR of 11.00 dB. These objective results strongly indicate that the WCA module effectively utilizes the noise reference signal to improve speech quality and suppress noise.

V. Conclusions

In this paper, we proposed Demucs-WCA, a dual-encoder speech enhancement architecture that utilizes a noise reference signal via a novel Windowed Cross-Attention (WCA) module, designed for scenarios with unknown user positions. Experimental results demonstrated that our method significantly outperforms strong baselines, including standard Demucs and our prior work, Demucs-RG, across multiple objective metrics such as PESQ and SI-SDR. This confirms that the block-wise WCA mechanism is an effective strategy for integrating a reference signal to improve speech quality.

Future work will focus on evaluating the method with real-world recordings and investigating the optimal placement of the reference microphone for various noise conditions.

References

[1] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A framework for speech enhancement with ad hoc microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1038–1051, 2016.

[2] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 836–840.

[3] X. Ren, X. Zhang, L. Chen, *et al.*, "A causal u-net based neural beamforming network for real-time multi-channel speech enhancement.," in *Interspeech*, 2021, pp. 1832–1836.

[4] J. Bai, H. Li, X. Zhang, and F. Chen, "Attention-based beamformer for multi-channel speech enhancement," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[5] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023. DOI: 10.1109/TASLP.2022.3230302.

[6] K. Suzuki, Y. Sugiura, and T. Shimamura, "Multi-channel speech enhancement network using noise-reference microphone," in *International Conference on Genetic and Evolutionary Computing*, Springer, 2024, pp. 301–311.

[7] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[8] X. Xu and J. Hao, "U-former: Improving monaural speech enhancement with multi-head self and cross attention," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 663–669.

[9] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[10] Z.-Q. Wang *et al.*, "Complex spectral mapping for single- and multi-channel speech enhancement and robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.

[11] 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)*, IEEE, 2021, pp. 6633–6637.

[12] Y. Hu, Y. Liu, S. Lv, *et al.*, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[13] H. Schroter *et al.*, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7407–7411.

[14] X. Le, H. Chen, K. Chen, and J. Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," *arXiv preprint arXiv:2107.05429*, 2021.

[15] M. A. Rahman and S. A. Fattah, "Dwinformer: Dual window transformers for end-to-end monocular depth estimation," *IEEE Sensors Journal*, vol. 23, no. 18, pp. 21 443–21 451, 2023.

[16] C. Valentini-Botinhao *et al.*, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. of Interspeech*, 2016.

[17] J. Thiemann *et al.*, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. of Meetings on Acoustics ICA*, 2013.

TABLE III: Objective scores for all methods at 100 Epochs. The best performance is highlighted in bold.

Method	PESQ	STOI	CSIG	CBAK	COVL	SI-SDR [dB]
Noisy	1.11	0.70	4.79	2.13	2.87	-3.68
DNLF [5]	1.65	0.90	4.90	2.48	3.40	10.50
Demucs [7]	1.90	0.93	4.96	2.69	3.60	10.55
Demucs-RG [6]	1.90	0.93	**4.97**	2.70	3.68	10.65
Demucs-WCA	**1.98**	**0.94**	4.96	**2.75**	**3.70**	**11.00**

[18] ITU-T, "P.862: Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Tech. Rep., 2001.

[19] 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*, IEEE, 2010, pp. 4214–4217.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.