INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Biophysically-inspired single-channel speech enhancement in the time domain

Chuan Wen[1], *Sarah Verhulst*[1]

Hearing Technology @ WAVES, Dept. of information technology, Ghent University, Belgium
chuan.wen@ugent.be, s.verhulst@ugent.be

Abstract

Most state-of-the-art speech enhancement (SE) methods utilize time-frequency (T-F) features or waveforms as input features and have poor generalizability at negative signal-to-noise ratios (SNR). To overcome these issues, we propose a novel network that integrates biophysical properties of the human auditory system known to perform even at negative SNRs. We generated biophysical features using CoNNear, a neural network auditory model, which were fed into a SOTA speech enhancement model AECNN. The model was trained on the INTERSPEECH 2021 DNS Challenge dataset and evaluated on mismatched noise conditions at various SNRs. The experimental results revealed that the bio-inspired approaches outperformed T-F and waveform features under positive SNRs and demonstrated stronger robustness to unseen noise at negative SNRs. We conclude that incorporating human-like features can extend the operating range of SE systems to more negative SNRs.

Index Terms: speech enhancement, biophysically-inspired feature, convolution neural network, time domain

1. Introduction

Speech enhancement is the process of improving the quality and intelligibility of speech signals that have been degraded by various factors, such as noise, reverberation, and distortion. There are numerous applications that benefit from speech enhancement, including automatic speech recognition (ASR), audio-visual conference systems and hearing aids [1, 2].

Deep-neural-network (DNN) based speech enhancement approaches have recently achieved great performance due to their complex function learning abilities. However, the SOTA speech enhancement methods were mostly formulated using time-frequency (T-F) representation obtained from the noisy signal through short-time Fourier transform (STFT). T-F-based approaches generally estimate the ideal ratio mask that provides the proportion of the clean speech in each T-F bin. T-F methods have certain limitations: Most T-F approaches only modify the magnitude of the noisy signal and disregard the phase information. The phase information is crucial, particularly at negative SNRs [3]. T-F methods that do include the phase information e.g. complex-valued spectrum modeling [4] or a trigonometric phase reconstruction [5], have improved performance, but still perform sub-optimal for negative SNRs. Secondly, a high spectral resolution is required in T-F methods to achieve successful enhancement. This requires long analysis/synthesis windows to calculate the spectrogram and results in relatively high system latency. This limits applications of these methods, e.g. for hearing aids that have high constraints on signal delay. To overcome long latency issues, several methods have been proposed such as asymmetric analysis-synthesis window pair [6] and hierarchical recurrent neural network with the use of low-frequency resolution [7], but these latency optimizations resulted in noticeable performance drops.

An alternative to traditional T-F domain approaches is to consider a time-domain formulation. This method avoids the need for frequency-domain transformation, and allows for the joint optimization of magnitude and phase information. Several studies have focused on training models that directly estimate the clean speech waveform by implicitly extracting features within the enhancement network [8, 9, 10]. Baby et al. employed biophysically-inspired human auditory features, e.g. cochlear transmission-line (TL) features, motivated by the superior performance of the human auditory system in adverse listening scenarios [11]. TL features were generated from an analytical model which simulates the processing of the human cochlear [12, 13, 14]. However, the cascade architecture of the analytical model results in high computational complexity and limits its integration into closed-loop and real-time audio processing systems. To address this obstacle, Baby et al. proposed the CoNNear, a neural network representation of a non-linear TL model which faithfully captures the properties of human cochlear processing [15]. The differential and parallel architecture of CoNNear enables integration into closed-loop audio processing systems. However, these biophysical CoNNear features have to date not been used in closed-loop SE systems, even though we expect improved robustness for operation at negative SNRs.

This paper thus introduces a biophysically-inspired time-domain speech enhancement model with CoNNear features (CoSE) that integrates biophysical inspiration and deep learning. The time-domain features extracted from CoNNear are fed into the noise reduction module which is an autoencoder convolutional neural network (AECNN) [16]. This SOTA SE model was employed as a generator of the generative adversarial network (GAN) model [17]. The main contributions of this paper are as follows: 1) developing a single-channel biophysically-inspired speech enhancement system, 2) comparison of SE performance using CoNNear features to other features, such as the widely-used T-F feature or the waveform, and 3) investigating the performance of CoSE-L, a low-latency configuration that utilizes overlap-save for signal reconstruction which is ideal for applications that require a minimal signal delay, such as hearing aids.

2. Biophysically-inspired framework

The aim of speech enhancement is to estimate the clean speech signal $s(t)$ from the noisy input signals $y(t) = s(t) + n(t)$, where $n(t)$ is the additive noise.

The proposed model is comprised of a biophysical feature

10.21437/Interspeech.2023-936

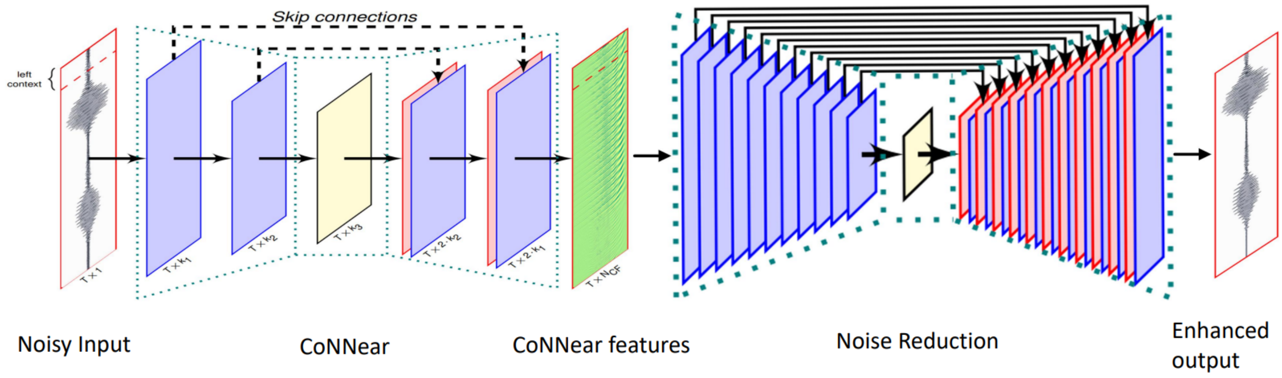

Figure 1: *Diagram of the proposed CoSE*

extractor and a noise reduction module as illustrated in Figure 1. The model utilizes a frame of noisy speech $y(t)$ as input and outputs the estimated clean speech $\hat{s}(t)$. The first stage involves the extraction of biophysical auditory features by CoNNear, converting the acoustic signals into N_{CF} representations of cochlear basilar membrane (BM) displacement waveforms with center frequencies (CF) ranging between 0.1 and 8kHz. The reason for this is that the input speech is sampled at 16kHz, which has a Nyquist frequency of 8kHz. Subsequently, AECNN will execute noise reduction and directly predict the enhanced waveform. The number of parameters is 11.5M for CoNNear and 8.9M for AECNN.

2.1. Bio-physical feature extraction module

The biophysical features of the noisy speech $y_k(t)$ are extracted using CoNNear, which is a convolutional network with an encoder-decoder architecture. The encoder component of CoNNear comprises four 1-D convolutional layers with a stride of 2, which compress the temporal dimension of the input. The decoder includes four deconvolutional layers to recover the original size of the input signals. Each (de)convolutional layer in CoNNear consists of 128 filters with a length of 64, followed by a Tanh activation function, except for the last layer of the decoder. To preserve phase information, which is crucial for speech perception [18], skip connections are implemented to carry temporal information from the encoder to the corresponding decoder layers. These skip connections not only mitigate the problem of gradients but also accelerate convergence. In order to overcome discontinuities near the frame boundaries, context information is added. We implemented a modification compared to the original model by only using historical context (N=256) to reduce overall computation latency.

2.2. Noise reduction network

The noise reduction is achieved through an autoencoder convolutional network (AECNN) that was earlier described in [17]. The output obtained from CoNNear serves as the input to AECNN. The encoder component of AECNN comprises nine 1-dimensional convolutional layers, which downsample the temporal axis size to $T/2^9$ with a stride of 2. The decoder component, which has a mirror structure of the encoder, restores the input length of the signal. The leaky ReLU activation function is applied after each (de)convolution layer, except for the last layer of the decoder, as the time-domain signals contain both positive and negative values. Skip connections are employed to pass the output of each encoder layer to the corresponding decoder layer. The output of the decoder from each layer is concatenated with the outputs from the corresponding symmetric layer in the encoder. Each layer has a fixed kernel size of 31 and the number of features for each layer in the encoder is 128, 128, 128, 128, 128, 128, 128, 256, 256, which are mirrored in reverse order in the decoder.

3. Experiments

3.1. Dataset

The training of CoNNear was performed using the TIMIT dataset [19], which provides an adequate representation of the acoustic diversity of speech while being phonetically balanced. A total of 2310 and 550 utterances were selected for the training and validation sets, respectively. The training target was the output of a biophysical TL model of the cochlea processing, which simulates the BM displacements corresponding to 201 center frequencies between 100Hz and 12 kHz. The input signals to the nonlinear TL model were generated by upsampling the original 16 kHz audio recordings to 100 kHz and adjusting their root mean square (RMS) energy to 70 dB SPL.

The training of the CoSE system was carried out using the INTERSPEECH 2021 DNS Challenge dataset [20, 21], where the parameters of CoNNear were frozen. This dataset comprised 11,350 speakers and over 600 noise scenarios. Three subsets of clean speech datasets were selected for the experiment, including read speech (English), emotional speech, and non-English speech (Mandarin). For the noise dataset, the Audioset [22] was selected for training, and the Freesound [23] was selected for testing. The noisy speech was created by selecting a random noise sample from the noise dataset and combining it with a randomly chosen clean utterance at a random SNR. A total of 150 hours of data was generated for training and 5 hours for validation, with SNRs ranging from -5 dB to 10 dB in 1 dB increments. The 1-hour test set was generated at each SNR of -5 dB, 0 dB, and 5 dB.

3.2. Experimental details

The training of the models was performed in two phases. First, CoNNear was trained according to [15], in which the training target was TL model simulations in response to speech utterances. Subsequently, the noise reduction model was trained using CoNNear features as inputs, where the parameters of CoNNear were frozen. Both models employed the minimization of mean absolute error (MAE) as a loss function, with a batch size of 16 at the utterance level. The negative slope of the leaky ReLu function was set to 0.3. During training, an initial learning rate of 0.0001 was utilized with the Adam optimizer [24]. The learning rate was halved if the validation loss did not decrease for two consecutive epochs. The speech utterances were segmented into frames of length 512, with an overlap of 50%. The models were trained on NVIDIA V30.

3.3. Evaluation metrics

The speech enhancement performance was evaluated by the following metrics: DNSMOS [25] and perceptual evaluation of speech quality (PESQ) [26]. DNSMOS serves as a non-intrusive measure to assess noise reduction performance, encompassing CSIG, CBAK, and COVL, exhibiting a robust association with human ratings. The COVL evaluates the overall quality of the enhanced speech signal, whereas the CSIG gauges the degree of speech distortion, and the CBAK predicts the scores of background noise distortion. A higher value of DNSMOS and PESQ signifies better performance. The models were evaluated for SNR of -5 dB, 0 dB and 5dB. For each SNR, we tested 720 utterances (1 hour) mixed with unseen noise from Freesound [23].

3.4. Comparison with other features

In our comparison study, the biophysical CoNNear features were evaluated in conjunction with two other features as baselines: log-magnitude spectrogram (T-F features) and waveform. To ensure a fair comparison, similar noise reduction modules were utilized with minor modifications between models. Specifically, for T-F features, the model was trained to predict the ideal ratio mask to enhance noisy spectrograms. Filter length of 3 was used and the non-linearity was the ReLU respectively. For the waveform feature, the filter number of the encoder was altered to 16, 32, 32, 64, 64, 128, 128, 256, 256.

3.5. Low latency configuration

To make our model suitable for future applications with high constraints on signal delay, e.g. hearing aids that require a total latency lower than 10ms. We also developed a low-latency system (CoSE-L). In this system, we used the overlap-save method to reconstruct the signal using a future context of 2ms and a frame shift of 4ms, which results in a total system delay of 6ms. Additionally, to minimize model size and computational complexity, we reduced the filter length to 15 and decreased the number of layers in the encoder or decoder to 7, consisting of 128, 128, 128, 128, 128, 256, 256. We compared the performance of CoSE-L to the full CoSE system to evaluate whether latency optimization still yielded satisfactory SE performance on all considered metrics.

4. Results and discussions

We compared noise reduction performance with the various input features on mismatched noise conditions. Table 1 shows that the biophysical-inspired CoNNear features performed the best among the evaluated features, especially at negative SNR conditions. CoSE demonstrated an 0.02 improvement in COVL over waveform features and 0.24 over T-F features at an SNR of 5 dB. Additionally, CoSE achieved a boost of 0.13 over waveform features and 0.34 over T-F features at an SNR of -5 dB. It can be observed that T-F features demonstrated commendable results at positive SNRs, and the CoNNear feature displayed slightly superior performance. As the SNR decreased, the T-F feature exhibited the most significant decrease in performance. T-F features lack phase information and are therefore sensitive to SNR. The other two tested features can leverage phase information to enhance performance and were more noise robust. The waveform feature had a comparable performance to CoSE at SNR of 5 dB, obtaining a COVL of 3.18, while CoSE obtained a COVL of 3.20. As the SNR decreased to negative, the gap between the waveform and CoSE increased, with CoSE showing an improvement of 0.13 COVL over the waveform feature. This suggests that the CoNNear feature can benefit speech enhancement systems in adverse noise scenarios. Despite the added computation, it is still beneficial to use CoNNear features over the raw waveform itself. This can also be appreciated in the samples we provided in `github.com/JasonCC001/CoSE`.

Table 1: *Comparison of different features in terms of various objective metrics. Higher values indicate superior performance and bold fonts highlight the best performance.*

Metrics	Approaches	SNRs (dB)			
		-5	0	5	Avg.
PESQ	noisy	1.12	1.35	1.54	1.34
	CoNNear-L	1.25	1.42	1.73	1.47
	CoNNear	**1.46**	**1.72**	**2.25**	**1.81**
	T-F	1.34	1.54	2.01	1.63
	waveform	1.35	1.55	2.04	1.65
COVL	noisy	1.41	1.60	1.92	1.64
	CoNNear-L	2.06	2.33	2.58	2.32
	CoNNear	**2.84**	3.08	**3.20**	**3.04**
	T-F	2.50	2.76	2.96	2.74
	waveform	2.71	**3.09**	3.18	2.99
CSIG	noisy	1.84	2.31	2.94	2.36
	CoNNear-L	2.56	2.84	3.02	2.81
	CoNNear	**3.13**	**3.34**	**3.45**	**3.31**
	T-F	2.87	3.11	3.29	3.09
	waveform	2.91	3.16	3.30	3.13
CBAK	noisy	1.43	1.57	1.85	1.62
	CoNNear-L	2.99	3.24	3.52	3.25
	CoNNear	**3.94**	**4.04**	**4.07**	**4.02**
	T-F	3.63	3.78	3.88	3.76
	waveform	3.93	4.02	4.06	4.00

To accommodate future applications, we also investigated a low-latency configuration of CoSE with a minimum system latency of only 6 ms, known as CoSE-L. While CoSE-L did not demonstrate comparable performance to the other three approaches, it significantly reduced overall latency while achieving competitive results as seen in Table 1. This highlights its potential for use in applications that require strict constraints on system delay, such as hearing aids.

Figure 2 illustrates the distribution of COVL scores for the three features considered at different SNR levels. CoSE exhibited the least amount of variation across the range of SNRs, and also demonstrated a more gradual decrease in performance compared to the other two features as the SNR decreased. This resilience of CoSE in challenging noise conditions may be attributed to its ability to exploit sharply tuned filter properties, which have been previously demonstrated to enhance the SNR by 4-5 dB when using a nonlinear TL model of cochlear processing that captures longitudinal coupling [27]. Regarding computational complexity, the average processing time per

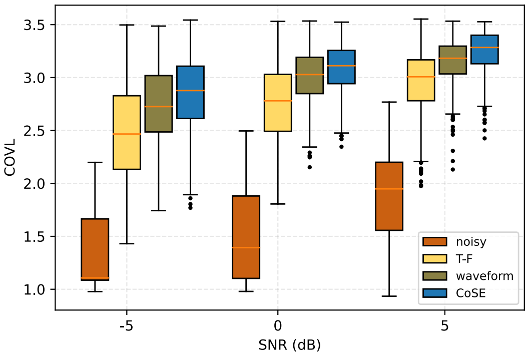

Figure 2: *Distribution of COVL scores for different features under various SNR conditions. Kruskal-Wallis test followed by Dunn's test with Bonferonni correction was used to calculate the significance for each SNR. All the pairwise comparisons within each SNR demonstrated statistical significance with a p-value <0.001.*

frame is 10.81 ms tested on an Intel i7-1265U PC.

We conducted additional experiments to compare our model with top-ranking models such as DCCRN [4], SDD-Net [28] and the baseline model NSNet2 from the DNS challenge, and present these results and the model complexity in table 2. Our tests were conducted on real recordings from the dev set of DNS challenge 3. The results indicate that our model performs comparably to NSNet2 (while one-third of the participants performed worse than the baseline), but that there is still room for improvement to achieve SOTA performance.

Table 2: *Comparison to other SOTA SE model.*

	Para. (M)	GMACs	DNSMOS
noisy	-	-	2.91
NSNet2	2.8	-	3.24
DCCRN	3.7	14.36	3.37
SDD-Net	6.38	6.0	3.51
CoSE	18.3	15.85	3.19

4.1. Spectrogram analysis

We conducted a spectrogram analysis of enhanced speech samples for different input features in Figure 3. An example of speech contaminated by music noise at an SNR of -5 dB is presented in panel (a). T-F features exhibit noticeable residual noise and an indistinct representation of harmonic components in Figure 3a. It is also noteworthy that T-F features maintain relatively complete components of speech. This difficulty arises from the challenge of distinguishing between music noise and speech, given the similarity between the spectra of music noise and the harmonic components of speech as seen in Figure 3a. In contrast, waveform features produced a more lucid spectrogram with less residual noise but with the loss of some speech components compared with T-F features (Figure 3c). Notably, CoSE demonstrated superior preservation of clear and abundant harmonic components, while minimizing residual noise (Figure 3d). This suggests that CoSE combines the strengths of both waveform and T-F features. It could be attributed to the fact that CoNNear acts as a time domain frequency analyzer with its capacity to sharply amplify the tone at center frequencies over surrounding noise components [27]. Lastly, we also investigated the spectrogram of enhanced speech by CoSE-L. CoSE-L preserved more residual noise compared to the other three approaches and manifested spectral leakage as seen in Figure 3e. This can be attributed to CoSE-L failing to track rapid changes in noise power. Our future research will explore a new architecture that can address these issues.

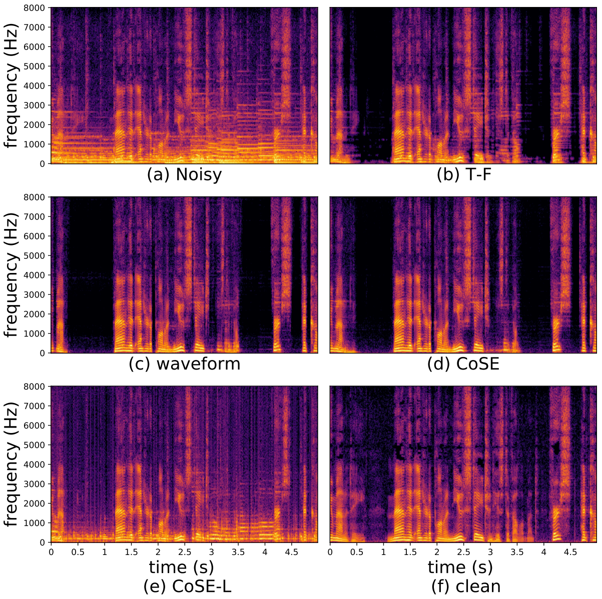

Figure 3: *Spectrograms of noisy speech corrupted with music noise at -5 dB SNR and enhanced speech from various approaches.*

5. Conclusions

In this paper, we proposed a novel biophysically-inspired end-to-end speech enhancement system. The system integrates a differentiable feature extractor, CoNNear, which simulates cochlear processing of the human auditory system. The objective evaluations indicate that compared to T-F and waveform features, CoSE demonstrated superior generalizability and stronger robustness across diverse SNRs. Additionally, CoSE combines the benefits of T-F and waveform features, resulting in clearer and more harmonic speech components. Moreover, we explored the low-latency configuration of CoSE, namely CoSE-L, which targets applications with strict system delay requirements. CoSE-L significantly reduces the system delay to 6 ms but lags behind CoSE's performance. In our future research, we will focus on developing novel approaches to improve the performance of CoSE-L and reduce its model size, thereby making it more compatible with embedded systems with limited power consumption and low-latency constraints.

6. Acknowledgements

This work was supported by G063821N FWO Machine Hearing 2.0 and EU-EIC/EARDITECH 101058278.

7. References

[1] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*. Springer, 2015, pp. 91–99.

[2] G. R. Popelka, B. C. Moore, R. R. Fay, and A. N. Popper, *Hearing aids*. Springer, 2016, vol. 56.

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

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

[5] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 71–75.

[6] S. Wang, G. Naithani, A. Politis, and T. Virtanen, "Deep neural network based low-latency speech separation with asymmetric analysis-synthesis window pair," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 301–305.

[7] H. Schröter, T. Rosenkranz, P. Zobel, A. Maier *et al.*, "Lightweight online noise reduction on embedded devices using hierarchical recurrent neural networks," *arXiv preprint arXiv:2006.13067*, 2020.

[8] S.-W. Fu, T.-W. 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[9] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[10] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7098–7102.

[11] D. Baby and S. Verhulst, "Biophysically-inspired features improve the generalizability of neural network-based speech enhancement systems," in *19th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2018)*. ISCA, 2018, pp. 3264–3268.

[12] S. Verhulst, T. Dau, and C. A. Shera, "Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3842–3848, 2012.

[13] S. Verhulst, A. Altoe, and V. Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hearing research*, vol. 360, pp. 55–75, 2018.

[14] S. Verhulst, H. M. Bharadwaj, G. Mehraei, C. A. Shera, and B. G. Shinn-Cunningham, "Functional modeling of the human auditory brainstem response to broadband stimulation," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1637–1659, 2015.

[15] D. Baby, A. Van Den Broucke, and S. Verhulst, "A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications," *Nature machine intelligence*, vol. 3, no. 2, pp. 134–143, 2021.

[16] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[17] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain." in *Interspeech*, 2018, pp. 1136–1140.

[18] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. C. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18 866–18 869, 2006.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[20] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[21] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP*, 2021.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[23] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[27] A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, and S. Verhulst, "A comparative study of seven human cochlear filter models," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1618–1634, 2016.

[28] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A simultaneous denoising and dereverberation framework with target decoupling," *arXiv preprint arXiv:2106.12743*, 2021.