Spectrum-aware neural vocoder based on self-supervised learning for speech enhancement

[1]Yanjue Song, [2]Doyeon Kim, [2]Hong-Goo Kang, and [1]Nilesh Madhu
[1] *IDLab, Ghent University - imec, Ghent, Belgium*
[2] *Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea*

***Abstract*—Self-supervised learning (SSL) models for speech provide an efficient way to utilise raw, real-world data for acquiring versatile representations. Here, we investigate the benefits of employing such pre-trained SSL models for speech enhancement. Our approach involves customising a neural vocoder that produces enhanced speech using embeddings extracted from the noisy input by pre-trained SSL models. Specifically, we investigate the suitable incorporation of the noisy spectrogram in the network, to address possible loss of acoustic details in the embeddings. Through the exploration of different fusion techniques, we find that effectively incorporating both the SSL embeddings and noisy spectrogram into the neural vocoder results in a model that relies more on the noisy spectrogram for acoustic details and on the SSL embeddings for semantic information. Experimental results show that our proposed model yields a significant improvement of speech quality, compared to baseline models that rely solely on embeddings.**

***Index Terms*—Self-supervised learning, speech enhancement, neural vocoder, embeddings, FiLM, attention**

I. Introduction

A major obstacle in training deep learning networks using fully-supervised methods is the acquisition of sufficient, high-quality annotated data. Whereas sufficient audio data such as LibriSpeech [1] is available in the wild, they are usually unlabelled or only weakly-labelled. Self-supervised learning (SSL) models then emerge as an attractive alternative to utilise such large amounts of unlabelled data. SSL models are trained to extract versatile latent representations to enable various tasks such as automatic speech recognition, speaker identification and emotion recognition [2]. When integrating SSL models into a specific task, the final system is usually divided into an upstream component (SSL model) and a downstream component (front-end network tailored to the task).

For the task of speech enhancement, too, there has been some prior work on the use of SSLs. In [3], for example, a comparison was conducted among 13 SSL upstream models within the context of the same speech enhancement framework. A three-layer bidirectional long short-term memory network (B-LSTM) was employed to predict a complex time-frequency-domain mask from the SSL representations of noisy signals. The clean speech spectrum was then estimated by applying the mask to the noisy spectrum. The same B-LSTM network predicts a better mask when using SSL embeddings as input, compared to using the noisy spectrum.

In [4], the vocoder architecture of HiFi-GAN [5] is explored to generate the clean speech signal directly from the distorted embeddings, extracted from noisy signals. Thus, the method was termed the 'denoising vocoder'. This variant, using modified-CPC [6] embeddings as input, was shown to outperform the version using log-Mel spectrograms. Similar to [3], both studies used a *learnable, weighted* sum of all hidden states of the embedding network, instead of only using the last hidden state. As verified in [3], this allows the downstream networks to make better use of the SSL models.

The HiFi-GAN of [4] was originally designed for the text-to-speech task, where the network learns to directly convert low-resolution mel spectrograms – predicted from the text – into high-fidelity waveforms by adversarial training. By replacing the real/fake labels of the discriminator by normalised objective metrics, e.g., as in MetricGAN [7] and MetricGAN+ [8], this architecture can be further adapted towards speech enhancement task.

Successful applications of SSL models across diverse speech tasks [2] highlight the presence of essential phonetic and semantic information in the learnt representations. Directly leveraging these representations – as in the aforementioned methods – can already yield improved denoising neural vocoders, particularly at the content-related level. However, it is also reported that the acoustic cues are missing in SSL embeddings. Incorporating task-relevant *a priori* information, such as fundamental frequency and speaker embedding, into the neural vocoder then aids in generating more natural speech, compared to using SSL embeddings alone [9].

For speech enhancement, this raises the interesting question of whether the downstream network can learn the required, additional information directly from the noisy spectrum. Thus, we systematically investigate means to effectively combine noisy spectrograms and speech representations (derived from pre-trained SSL models) for a HiFi-GAN-architecture-based denoising vocoder, and evaluate their benefits. Lastly, in addition to common speech quality metrics to benchmark the results, we also evaluate speaker *fidelity* by comparing the enhanced speech with the clean reference using a state-of-the-art speaker verification framework. The upper bound of the proposed system, that uses embeddings extracted from the underlying clean speech is also included in the benchmark.

Sec. II introduces the proposed generative framework, with the essential design considerations investigated. The experimental setup and ablation study settings are described in Sec. III. Benchmarking results on various instrumental speech quality and intelligibility metrics are presented and analysed in Sec. IV. Key take-aways are summarised in Sec. V.

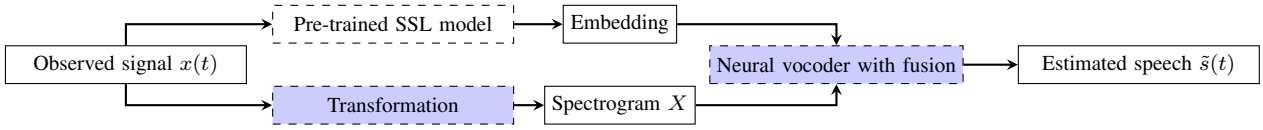


Fig. 1: The diagram of the proposed system. The blocks with blue background are the components to be investigated in our work. The pre-trained SSL model is frozen.

II. Methodology

A. Generative Model Framework

The goal of the system is to estimate the speech signal from the observed noisy and reverberant signal. Two features can be extracted from the observed signal, the spectrogram and the embedding, both of which contain distorted versions of the underlying target speech. While it has been shown in the aforementioned studies that the HiFi-GAN can estimate the underlying clean speech from either of the distorted features, here we investigate a complementary fusion of both features to achieve a better, consistent estimate. The proposed system is depicted in Fig. 1. This fusion can be achieved in various ways, and using different versions of the input features – as we discuss below.

B. Feature Extraction

For the first feature – the embedding – selecting an appropriate SSL model is important. In [10] we demonstrated that TERA [11] is robust to noise and reverberation. Such models can, therefore, extract better speech embeddings under adverse conditions - which is essential for application to speech enhancement. Thus TERA is chosen as the upstream SSL model in this work. We employ a pre-trained model, trained on 960-hours of LibriSpeech data from the s3prl toolkit[1]. The upstream model is frozen during the training of the downstream model.

The second feature – the observed signal spectrum – can be input either as magnitude or as log-magnitude. Both are examined, to see which representation allows for better denoising.

As the embedding dimensions of TERA and of the chosen spectral representation are different, additional layers are used to align the respective feature dimensionalities with our HiFi-GAN configuration. Following the approach outlined in [5], a 1D-convolutional layer is employed to transform the spectrum to the HiFi-GAN input dimension. The embeddings are projected by a fully-connected layer. The projections also allow implicit learnable weights of the two features.

C. Neural Vocoder: HiFi-GAN

HiFi-GAN consists of a generator and two discriminator components. The generator is responsible for upsampling the input to a time-domain signal using a stack of transposed 1D-convolutional layers and residual blocks at multiple scales. The discriminators are tasked with classifying the signal as either real (natural) or fake (synthesised).

We adjust the upsampling configuration to align with the frame length (25 ms) and frame shift (10 ms) of TERA. As we operate at 16 kHz, we configure the upsampling rates of the four residual blocks as $[8, 5, 2, 2]$, and the corresponding kernel sizes as $[16, 15, 4, 4]$, respectively. Other settings are identical to the original HiFi-GAN implementation[2] [5].

The total loss function for generator training consists of three components: a) the generator loss (the mean-square error (MSE) between the log-Mel spectra of the reference and the generated signals); b) the feature matching loss (MSE between the discriminator features of the reference and the generated signals); and c) the discriminator loss. Both the Multi-Scale Discriminator (MSD) and the Multi-Period Discriminator (MPD) are utilised as introduced in [5]. The training starts with a warmup stage, where only the generator is trained with the generator loss. Afterwards, all three losses are included.

D. Fusion Methods

To fuse information from the TERA embeddings and the noisy input spectrum, we investigate three widely used strategies: addition, cross-attention transformer block [12], and the FiLM layer [13]. We depict their schematics in Fig. 2.

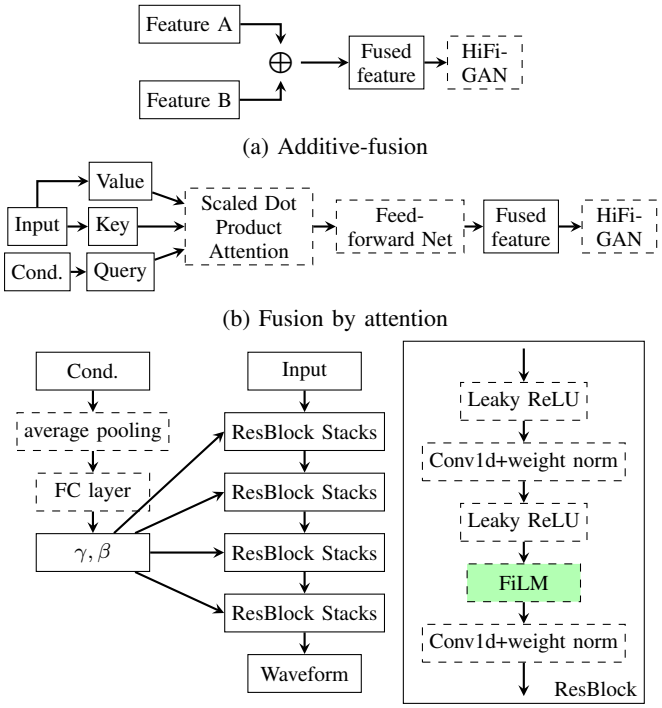
Addition. The most naïve way to combine features is by a weighted sum. Using learnable weights, the network assigns relevant importance of each feature to the output.

Cross-attention using transformers. The transformer block uses the attention mechanism to combine different features. One feature is the *input* and the other, the *conditioning*. The conditioning provides auxiliary information to highlight the vital elements in the input term. The transformer block first calculates the multi-head cross-attention between the input X and the conditioning C. The query matrix is obtained from C, and the key- and the value matrices from X. Then, we employ a feed-forward network with two fully-connected layers, with layer normalisation as suggested in [12]. Two residual connections bypass the two modules, respectively. The number of multi-head attentions is set as 8.

FiLM layer. The FiLM layer [13] introduces the auxiliary information into the network by modifying the input feature maps by affine transformations, whose parameters depend on the feature used for *conditioning*. We apply this modification of the feature maps at the residual blocks in our vocoder. A vocoder residual block consists of two convolutional layers, each preceded by an activation function (Leaky ReLU) and followed by weight normalisation. We insert the FiLM layer between the second activation function and the second convolutional layer. The conditioning feature is aggregated along both time- and feature-dimension by *mean-pooling*, and then

[1]https://github.com/s3prl/s3prl

[2]https://github.com/jik876/hifi-gan/tree/master



(a) Additive-fusion

(b) Fusion by attention

(c) Fusion by FiLM layer

Fig. 2: Different feature fusion methods, Cond.=conditioning. The feature fused by (a) addition or (b) attention block is converted into waveform by the followed neural vocoder, whereas the FiLM layer operates the feature maps of the neural vocoder by the extra conditioning information as shown in (c) (left). The architecture of the residual block with the FiLM layer is shown on the right.

linearly projected to a feature weight factor $\gamma_{i,c}$ and a bias $\beta_{i,c}$ to linearly transform the feature map from the cth channel and the ith layer of the neural vocoder.

E. Research Questions

Based on the above overview and the framework of Fig. 1, the following questions now require to be addressed: **Q1.** How does the spectrum representation (i.e., magnitude or log-magnitude) affect the system performance? **Q2.** What is the relative contribution of representations extracted from each hidden state of TERA to the overall performance of the neural vocoder? **Q3.** How does the neural vocoder performance differ with the fusion method? **Q4.** For the cross-attention block and the FiLM layer, the input feature and conditioning feature play different roles in the system. Which feature is best suited to which role (i.e., spectrum as input and embedding as conditioning or vice-versa)?

III. Experiment Setup

For training and validation, we use the DNS 2021 challenge dataset [14] to synthesise 140 hours of noisy and reverberant audio. The signal-to-noise ratios (SNRs) are sampled from a uniform distribution ranging between -5 and 20 dB at 21 levels. Reverberation is simulated using synthetic room impulse responses (RIRs) SLR26 and SLR28 [15], where the reverberation time (RT60) is limited between 0.3 s to 1.3 s. To evaluate the performance of the proposed systems, we create a fully unseen test set using the CSTR VCTK speech corpus [16] and the NOISEX92 noise database [17], along with recorded RIRs from the MIT RIR database [18]. The SNRs are evenly distributed among $(-7, 0, 5, 10, 15)$ dB.

The network is trained by the AdamW optimiser with a learning rate of 0.0002 and $betas = [0.8, 0.99]$. Random cropping is adopted to further augment the training data: 10 second-utterances are cropped to 3 second segments from a randomly selected starting point after the warmup stage. The warmup stage continues for 20 epochs, and the system is further trained for a total of 200 epochs thereafter.

A. Metrics

Intrusive metrics have been reported to be less appropriate for assessing quality of generative models [19]. Therefore, we employ two non-intrusive metrics, DNSMOS [20] and NISQAv2 [21], to evaluate the quality of the enhanced speech. Yet, with the reference signal available, one advantage of intrusive metrics is their sensitivity to 'hallucinated' input (a prominent artefact with generative models). Therefore, we also use the STOI metric [22], which is sensitive to perturbations of the speech envelope. Degraded STOI could, therefore, indicate compromised intelligibility.

Aside from speech quality metrics, we also evaluate how effectively the generative model preserves speaker information. For this we use the ECAPA-TDNN [23] speaker verification framework. Specifically, we compute the cosine similarity ($\in [0, 1]$) between the *speaker* embedding extracted from the clean reference and that extracted from the enhanced signal. A higher score indicates greater similarity between the two embeddings, and, consequently, speech synthesis with a more faithful capture of speaker characteristics.

B. Ablation Studies

Our *reference* proposed system employs the transformer block for feature fusion, followed by the neural vocoder. The cross-attention block takes the log-magnitude spectrum as the input feature and the (learnable) weighted sum of embeddings from all four TERA layers as the condition feature. As *baseline*, we train the neural vocoder to generate speech directly from the distorted TERA embeddings. This *Denoising Vocoder* network configuration is identical to [4].

To answer the questions listed in Sec. II-E, we investigate each component individually *always* on the basis of the proposed *reference* system, where only one component is replaced each study. Each variant is independently trained on the same dataset with training settings identical to the reference system. The variants are listed in Tab. I, and briefly described below. To address **Q1**, the log-magnitude spectrum input is replaced by the magnitude spectrum in variant 2. To benchmark the benefit of using information from multiple layers of the SSL embedding (**Q2**), variant 3 is trained using only the embedding from the *last* hidden state of TERA.

TABLE I: Evaluation results on the noisy, reverberant test set. The highest score of each metric is highlighted in bold, and the lowest by underline.

No.	Model Description	STOI	DNSMOS			NISQAv2					Speaker embedding cosine similarity
			OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	3.086	3.379	**4.043**	3.097	3.601	3.325	2.953	3.726	0.551
1	Proposed reference	0.811	**3.054**	**3.405**	3.892	3.691	3.526	3.998	**3.494**	**3.992**	0.529
2	Magnitude spectrum feature	**0.819**	2.999	3.374	3.835	3.566	3.406	3.997	3.433	3.906	0.552
3	TERA - last hidden state	0.798	2.955	3.303	3.876	3.605	3.609	3.955	3.351	3.870	0.524
4	Additive-fusion	0.814	3.017	3.306	3.997	**3.768**	**3.932**	**4.032**	3.465	3.984	**0.584**
5	FiLM	0.739	2.696	3.005	3.827	2.828	3.409	3.408	2.614	3.434	0.387
6	Attention conditioned by spectrum	0.811	2.966	3.261	3.968	3.522	3.602	3.862	3.276	3.876	0.524
7	FiLM conditioned by spectrum	0.777	2.814	3.149	3.825	3.122	3.315	3.679	3.017	3.659	0.539
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-

Next, we compare the attention block with the other two fusion methods (**Q3**): Variant 4 performs additive-fusion of the two features, and variant 5 uses FiLM layers. In attention blocks and FiLM layers, the input term and conditioning term serve different purposes by architecture design. To determine which feature is best suited to which role (**Q4**), we further train two networks. Variant 6 uses log-magnitude spectrum as the conditioning feature in the transformer block and variant 7 does the same for the FiLM.

Finally, to benchmark the upper bound of the proposed system, we train the neural vocoder for the ideal case, where the embeddings are extracted from the clean speech and the log-magnitude spectrum from the noisy and reverberant input.

IV. Results and Discussions

The evaluation results of all variants are summarised in Table I. Compared to the distorted signals and the baseline, the proposed *reference* model and most of its variants show significant improvements in speech quality metrics, with some variants scoring higher in STOI as well.

Comparing variants 1 & 2, using the magnitude spectrum leads to a degradation in speech quality, especially in terms of background noise (BAK-DNSMOS / NOIS-NISQAv2). However, this change seems to help the neural vocoder in capturing the speech envelope more effectively, resulting in the highest score on STOI. This could be because the spectral nulls would be more strongly indicated in log scale, leading to better noise suppression and finer harmonics, but the stronger suppression could also lead to more aggressive attenuation of weak signal components (affecting STOI).

Comparing variants 1 & 3, by reducing the information included from the SSL model, all the metrics apart from DIS-NISQAv2 degrade. While it is commonly assumed that the final layer feature map contains the majority of speech-related information, this ablation study indicates that *other layers still contain some valuable information relevant to speech reconstruction*. We now examine the contribution of each hidden state of TERA. Tab. II shows the learned weights used to combine different TERA hidden states for the different variants. Although there are differences in scale, the trend indicating the importance of the last two layers, particularly the last one, remains consistent. This differs from the trend reported in [3], [4] that the first few layers are given greater weight across all tested models. However, in [3], [4], the neural vocoder decodes only the distorted SSL embeddings. Thus, *we attribute this difference to the acoustic cues provided by the additional spectrum information in our method.* It is commonly assumed that the earlier layers contain more detailed phonetic information [3], while the final layer tends to encode more semantic information [2]. With the incorporation of auxiliary information, the model appears to prioritise the SSL embeddings for semantic information, because the detailed phonetic information can be deduced from the spectrogram. This is in line with the findings in [9]: embeddings alone are insufficient to capture all phonetic details necessary for authentic speech reconstruction.

TABLE II: Combination weights for TERA hidden state layers

Variant	Layer1	Layer2	Layer3	Layer4
1	-0.002	-0.011	0.036	0.098
2	0.003	0.016	-0.105	-0.248
4	0.017	0.025	-0.479	-1.229
5	0.015	0.116	-0.586	-1.495
6	0.015	0.105	-0.524	-1.275
7	-0.004	-0.166	0.037	0.157
8	-0.068	-0.048	-0.041	0.115

Regarding the fusion methods: additive-fusion (variant 4) of the projected spectrum and the embedding performs similarly to, or even better than the reference in terms of STOI, BAK-DNSMOS, and NOIS.-NISQAv2. Generally there is more residual noise in speech inactive frames in speech signal generated by the reference approach, which could be introduced by the attention block. However, listening to the samples, we observe that the additive-fusion method generates less content in the high-frequency range. As a consequence, the output signals sometimes sound less pleasant[3]. This is also reflected by the lower score on COL.-NISQAv2 - which indicates this colouration.

From results of variants 5 & 7, the FiLM layer is less effective in terms of feature fusion for the speech enhancement task. When the neural vocoder is *conditioned* by spectral

[3]The audio samples can be found at https://aspireugent.github.io/EUSIPCO2024YS/.

information (variant 7), the performance is even worse than the baseline denoising vocoder. The performance degrades further when spectral information is taken as input (variant 5).

For the attention layer, the speech quality degrades when using the noisy spectrum information as the conditioning of the attention block, despite their similar performance in terms of the speech intelligibility. This indicates that the acoustic deatils cannot be captured by conditioning.

In terms of preserving speaker characteristics, feature fusion by addition performs best. These scores seem correlated with the STOI values.

Considering the overall performance, the proposed reference system and the additive fusion system exhibit respective advantages. Additive-fusion demands slightly lower computational resources and generates cleaner speech, whereas the attention-based system provides a speech signal with more details and fine-structure. The FiLM layers, however, degrade the neural vocoder performance. It should be noted that there is still a noticeable delta between the upper bound (variant 8) and the optimal systems across all metrics, particularly in the STOI score. Since the upper-bound system uses speech embeddings extracted from the clean speech, this supports the hypothesis that the network primarily relies on the speech embedding to obtain semantic information.

V. Conclusions

We proposed a method to enhance noisy and reverberant speech by a neural vocoder that incorporates both noisy spectrogram and the distorted embeddings extracted from the noisy input using pre-trained SSL models. Experiments show that introducing *relevant* additional information improves the quality of the speech generated by neural vocoder. Among the three fusion methods investigated, ablation studies demonstrate that addition or cross attention block conditioned on SSL embedings are effective. FiLM layers, however, degrade the neural vocoder performance no matter which feature is chosen as the condition. We also examine the reconstruction fidelity in terms of preserved speaker characteristics with the help of a speaker verification system. The score indicates that additive-fusion best captures speaker information. Additionally, by analysing the learnable weights used to combine different layers of embeddings, it becomes evident that the network extracts phonetic cues from the spectrogram and semantic information from the embeddings. Compared to the upper bound, which utilises SSL embeddings from clean signals, there is still a gap in the performance, highlighting potential for further research.

References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[2] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.

[3] Z. Huang, S. Watanabe, S.-W. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 6837–6841.

[4] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 17 022–17 033, 2020.

[6] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[7] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Intl. Conf. on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[8] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli *et al.*, "MetricGAN+: An improved version of metricgan for speech enhancement," in *Proc. INTERSPEECH*, 2021, pp. 201–205.

[9] A. Polyak, Y. Adi, J. Copet, E. Kharitonov *et al.*, "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. INTERSPEECH*, 2021, pp. 3615–3619.

[10] Y. Song, D. Kim, N. Madhu, and H.-G. Kang, "On the disentanglement and robustness of self-supervised speech representations," in *Intl. Conf. on Electron., Inf. and Commun. (ICEIC)*. IEEE, 2024, pp. 662–665.

[11] A. T. Liu, S.-W. Li, and H.-y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 29, pp. 2351–2366, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention is all you need," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 30, 2017.

[13] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI conf. on artificial intelligence*, vol. 32, no. 1, 2018.

[14] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler *et al.*, "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2021, pp. 6623–6627.

[15] T. Ko, V. Peddinti, D. Povey *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 5220–5224.

[16] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research*, 2017.

[17] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[18] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[19] D. de Oliveira, J. Richter, J.-M. Lemercier, T. Peer, and T. Gerkmann, "On the behavior of intrusive and non-intrusive speech enhancement metrics in predictive and generative settings," in *Speech Communication; 15th ITG Conference*. VDE, 2023, pp. 260–264.

[20] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2010, pp. 4214–4217.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. INTERSPEECH*, 2020, pp. 3830–3834.