Supervised Speech Enhancement using Transformers

Anirudh Dash EE21BTECH11002
Gaurang Dahad EE21BTECH11014
Satvik Bejugam EE21BTECH11051
Sohal Malviya EE21BTECH11053

Abstract

This paper proposes a transformer-based approach to speech enhancement, leveraging attention to address the limitations of traditional methods involving standard auto-encoder setups using CNNs. We first train a baseline vanilla auto-encoder setup for speech enhancement. We improve this model using modified activation functions, batch normalization, and residual connections. Finally, we incorporate the transformer into this setup and compare the results. We show that a naive transformer approach performs worse than our improved CNN-only-based approach, indicating there is a lot of scope for improvement whilst trying to capture global dependencies.

INTRODUCTION

Speech enhancement is a crucial task in signal processing and machine learning. The objective of speech enhancement is to improve the quality of speech signals degraded by noise. The motivation behind supervised speech enhancement lies in its wide range of real-world applications. This includes improving telecommunication, assistive devices such as hearing aids, and automatic speech recognition systems. In real-world scenarios, speech signals are often contaminated by various types of noise, such as background chatter, traffic, or other sounds, making it challenging to extract clean speech. Supervised learning methods, where models are trained on pairs of noisy and clean speech signals, have become a popular approach to tackle this problem. This problem is very well studied, and several standard techniques exist. Early approaches were based on statistical techniques, such as spectral subtraction [1], [2], Weiner filtering [3], [4], Kalman filtering [5], [6], [7], and non-negative matrix factorization. These traditional methods operate either in the time or time-frequency (T-F) domain by applying filters to the raw noisy speech signal or its spectrogram. However, in the T-F domain, most of these techniques deal with solely the magnitude spectrum of the speech, considering it to be the primary factor determining speech quality.

Despite attempts to address phase-related issues, traditional methods struggle to fully incorporate this information. These limitations, along with the rise of machine learning techniques capable of learning complex patterns directly from data, prompted a shift towards deep learning-based approaches that can adaptively learn from noisy speech and provide more robust solutions for a diverse set of noisy signals.

Recent advances in deep learning have enabled the development of powerful neural network models to denoise speech signals. These models have significantly outperformed traditional speech enhancement methods in terms of both objective and subjective speech quality metrics. Several deep models deal directly with the raw time domain signals, such as Temporal Convolutional Neural Networks (TCNN) [8], Wave-U-Net [9], and ConvTasNet [10]. Others continued to use the transform domain- dealing with the magnitude spectrum for training and improving performance. However, there was much scope for improvement, considering the phase spectrum was completely neglected. Eventually, using ideas from image processing, the spectrogram was separated into two channels, one accounting for the real part and one for the imaginary part of the short-time Fourier Transform (STFT) of the speech signals. Considering the excellent performance of CNNs [11] for colored images, which are decomposed to the R, G, and B channels, attempts were made to use CNNs for speech enhancement [12], [13]. More recent advancements have allowed the use of complex numbers to train these neural networks, allowing for combining the real and imaginary parts into a single spectrogram.

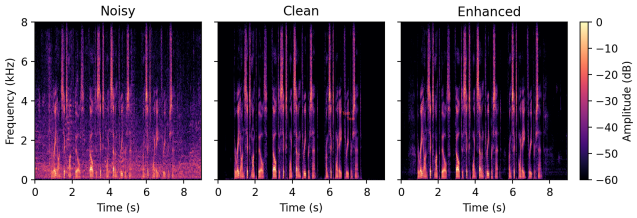

Figure 1. Spectrograms in Speech Enhancement

Owing to the performance of several image-based deep learning techniques for supervised speech enhancement, we look at transformers for the same. Transformers have al-

ready been analyzed for speech enhancement in the time domain [14]. Given the performance of models like the Vision Transformer (ViT) [15], we explore the application transformer models on the spectrogram to perform supervised speech enhancement.

1. Prior Work

1.1. Traditional Methods

Spectral Sampling involves the estimation of noise followed by subtraction from the power spectrum of the original speech signal. It follows the assumption that noise is additive in nature and attempts to identify regions where noise is present, but speech is absent [17]. The estimated noise is then subtracted from segments where speech is active.

The Wiener filter [4] works under the assumption that the power spectral densities (PSDs) of the speech and noise signals can be estimated or modeled. It aims to minimize the Mean Square Error (MSE) between clean and noisy speech by determining how much to amplify or attenuate different frequency components of the signal. Like spectral subtraction, this approach operates in the frequency domain, adjusting the signal to reduce noise while preserving speech quality.

The Kalman filter [5] is a recursive algorithm that continuously updates its estimate of clean speech as it processes new data from the noisy signal. It models noisy speech as a combination of clean speech and noise. With each new input, the filter refines its prediction by incorporating the latest data alongside its previous estimate.

While simple to implement, these techniques struggle when noise is strong compared to the speech signal, i.e, when the signal-to-noise ratio (SNR) is low. Unlike classical methods that depend on assumptions about noise and speech, deep neural networks (DNNs) learn complex mappings from noisy to clean speech directly. This approach enabled DNNs to adapt to diverse noise types and speech characteristics, leading to superior performance in both objective and subjective metrics [18].

1.2. Time Domain DNN Methods

Some of the well-known time domain deep neural network models are TCNN, Wave U-Net, and ConvTasNet.

TCNN [8] architecture features a CNN-based encoder-decoder framework, with an additional temporal convolutional module (TCM) added in between the encoder and decoder. The encoder is used to compress noisy input frames into a low-dimensional representation. The TCM leverages causal and dilated convolutions to incorporate information from the current and previous frames. Finally, the decoder utilizes the TCM output to reconstruct the enhanced frame.

Wave-U-Net [9] architecture is inspired by the U-Net model, commonly used in image processing. It uses an encoder-decoder structure with downsampling to capture features and upsampling to reconstruct the enhanced waveform. Skip connections help the model retain feature information. Several works have also extended Wave-U-Net by incorporating, for example, attention gates to identify more suitable features from the downblocks in the U-Net architecture paths. [19]

SEGAN architecture [20] typically consists of two main components: a fully convolutional generator and a discriminator, structured in a Generative Adversarial Network framework. The generator takes in noisy speech signals and produces cleaner versions, employing convolutional layers to capture temporal features and spatial patterns. Meanwhile, the discriminator assesses the realism of the generated speech by distinguishing between enhanced and real clean signals, which encourages the generator to produce high-fidelity outputs.

While the models shown above operate directly on the raw waveform, others operate in the time-frequency domain, i.e., employing techniques like Short-Time Fourier Transform (STFT) to process spectrograms and use them as input.

1.3. Time-Frequency Domain DNN Methods

The RNNoise [21] architecture combines traditional digital signal processing techniques with a recurrent neural network to effectively model the temporal dependencies in speech signals. The audio signal is preprocessed using techniques such as Spectral Subtraction to provide an initial noise reduction. It is then transformed into the time-frequency domain using the STFT operation. The magnitude of the complex output is then passed to a gated recurrent unit (GRU) based RNN, which learns the temporal patterns in both speech and noise. Finally, the enhanced magnitude spectrum is combined with the original phase information (from the noisy signal), and an inverse STFT is performed to convert the signal back to the time domain.

PercepNet [22] is another efficient speech enhancement model that relies heavily on human perception of speech signals. It builds on RNNoise by further leveraging human auditory perception principles. It incorporates the concept of critical bands i.e., frequency ranges where the human ear is particularly sensitive, to more effectively separate speech from background noise. The architecture integrates both convolutional layers and GRU (Gated Recurrent Unit) layers to capture local spectral features as well as model the long-term temporal dependencies.

The noisy phase information was combined with the enhanced magnitude spectrum in the models referred to above.

Phase has a significant effect on the perceived speech quality. Works by Paliwal et al.[23] show that both objective and subjective criterias evaluated the speech processed in

	NoisySNR	Spectral Subtraction	Wiener Filter	Kalman Filter
White Noise (dB)	0.02227	8.5699	12.253	17.8932
Random Noise (dB)	-5.0952	2.2479	3.067	5.45473
Color Noise (dB)	-16.4817	1.123	9.2333	12.2185

Table 1. SNR Comparison of traditional methods[16]

Metric	Noisy	Wiener	SEGAN	Wave-U-Net
PESQ	1.97	2.22	2.16	2.40
SSNR	1.68	5.07	7.73	9.97

Table 2. PESQ and SSNR of various models

both magnitude and phase domain to be better than when only the magnitude is processed. Accurate reconstruction of the phase of the clean sources is a nontrivial problem, and the erroneous estimation of the phase introduces an upper bound on the accuracy of the reconstructed audio. Multiple efforts were made to incorporate the phase, and researchers started to look into other domains for motivation.

In image processing applications, the use of 3 separate channels as input to convolutional neural networks is common. One example is image classification using the Deep Convolutional Neural Networks [24] which achieved excellent results. Using these results as motivation, attempts were made to use various CNN-based architectures while dealing with the real and imaginary parts as two channels. [25], [26], [27] use similar techniques achieving improved results with DCU-Net [28] being one of the most prominent. Improvements can be seen when both the phase and magnitude information is used as opposed to only the magnitude.

Recently, there has been a surge in the use of the Vision Transformer [15] for several image processing tasks, including classification, and segmentation, among many others. ViT uses an architecture consisting of layers of self-attention mechanisms and feedforward neural networks. The final model is composed of a stack of these layers.

However, despite significant improvements in the field of image processing using transformer techniques, the use of the same in the field of speech processing has been rather limited. Following in the footsteps of using certain image processing techniques on speech processing (like CNN and DCNN), there is a significant scope for research in this field.

2. Model Architecture

We motivate the enhancement model through a fully convolutional autoencoder designed based on the brief idea described in Section 1.

2.1. STFT-based Input Processing

The input audio is transformed into its corresponding spectrogram form (i.e., the time-frequency domain) using the STFT operation with an FFT size of 512, a hop length of 160, and a window length of 400 samples. The real and imaginary components are treated as separate channels for the CNN. We improve upon the architecture proposed in the MPR by adding additional batch normalization blocks.

2.2. Description

2.2.1 An improved CNN-based model

The network consists of seven convolutional layers. The first five layers (`conv1` to `conv5`) perform downsampling through convolution, gradually increasing the number of channels while reducing the spatial resolution of the features. The $PReLU$ activation function is used post each of these layers. The convolutions have a kernel size of $(3, 4)$ and strides varying from $(1, 2)$ to $(2, 2)$, and a padding of 1, which progressively extracts features from the input spectrogram.

After the downsampling, two additional convolutional layers (`conv6` and `conv7`) maintain the spatial resolution while enhancing the features. These layers are followed by five deconvolutional layers (`deconv1` to `deconv5`), which perform upsampling, effectively reversing the downsampling process. The deconvolutions use transposed convolutions with matching kernel sizes and strides to the convolutional layers.

Residual connections are incorporated between corresponding convolutional and deconvolutional layers to retain information from earlier layers.

Output of the final deconvolution block is passed through a *tanh* activation function. This is done to get a mask (each element in the output is a real number, hence multiplying by an activation function that takes values from -1 to 1), which is then multiplied by the input spectrogram to get the enhanced spectrogram. An inverse STFT (ISTFT) is performed to convert the enhanced spectrogram into a time domain output signal.

	Noisy	RNN Noise	PercepNet	**DCUnet** (20 layers)	**DCUnet** (16 layers)	Wiener	SEGAN
PESQ	1.97	2.29	2.54	3.07	2.91	2.22	2.16
SSNR	3.40	3.70	4.05	15.54	14.25	5.07	-

Table 3. PESQ and SSNR of DCU compared to other models

2.2.2 The updated Transformer+CNN-based model

Attention: Attention helps take into account the global relationship schemes present in the data being analyzed.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

Q is the query matrix,
K is the key matrix,
V is the value matrix,
d_k is the dimension of the key vectors.

We leverage the global context provided by self-attention to improve the enhancement performance of the model. The transformer is added post the fourth convolutional layer and three extra convolutional layers surplus to requirements are removed. This counteracts the increased number of parameters introduced by the transformer. The convolutional layers learn the fine details of the model while the transformer helps capture a global context. Given a pixel in the STFT, the transformer helps utilize the relations caused as a result of previous timestamps.

3. Pipeline

3.1. Dataset

We used the *Valentini* dataset which contains speech samples from 28 speakers (including 14 male and 14 female voices), captured under various noise conditions. The dataset includes approximately 10,000 clean speech samples mixed with various real-world and synthetic noise types, such as white noise, street sounds, and babble. These samples are sampled at 16 kHz and cover a wide range of signal-to-noise ratios (SNRs).

For training: The input to the model is fixed to be a certain dimension to ensure the output of the deconvolutional blocks matches the corresponding convolutional outputs. We chose this length to be 32160 frames, which is slightly longer than 2 seconds (sr = 16000 Hz). For signals longer than 2 seconds, we randomly selected a starting point (more than 2 seconds from the end) and used the subsequent 2-second segment. For signals shorter than 2 seconds, we padded them by repeating the signal until it reaches the required length.

For testing: To ensure, the autoencoder doesn't throw any errors while passing the signal through it, we padded the input with the required number of frames before passing it through the model and removed those frames from the enhanced output, once obtained.

One issue that we experienced with testing is that the model parameter `n_dim` must be initialized in the model, but this presents a challenge. The value of `n_dim` is determined by the product of the number of channels and the number of time frames after the convolutional blocks. While the number of channels remains fixed, the number of time frames post-convolution is variable for sequences of different lengths. This results in a mismatch in the dimensions. Thus, inference cannot be performed unless the duration of test and training samples is same.

Now, to resolve this issue arising in the transformer setup, we introduced a new variable called *dimension*, which represents the number of time frames, computed after the STFT. This allows us to compute the input dimension dynamically, based on the actual sequence length at inference time, rather than relying on a fixed number. This solution ensures that the transformer model can handle variable-length sequences without requiring the test data to have the same duration as the training data.

3.2. Training setup

We make use of the AdamW optimiser with learning rate of 10^{-3}, with a schedule of "reduce on plateau with patience 50" and weight decay of 10^{-6}. We run training for 200 epochs with a batch size of 32.

Following are the loss functions that our model was trained on:

Mask Loss: Mask Loss is a metric used to evaluate the performance of speech enhancement models. An oracle mask is derived from the clean and noisy spectra, which represents the ideal mask that would perfectly separate the speech and noise components. By minimizing the difference between the estimated and oracle masks, the model can learn to produce more accurate mask predictions, leading to improved speech enhancement quality.

PESQ Loss: Perceptual Evaluation of Speech Quality (PESQ) [29] is a widely used objective metric for assessing the quality of speech signals, particularly in the context of speech enhancement. PESQ compares a processed speech signal to a reference clean signal, focusing on the human auditory perception, to provide a score that is able to properly reflect the perceived quality of the enhanced speech.

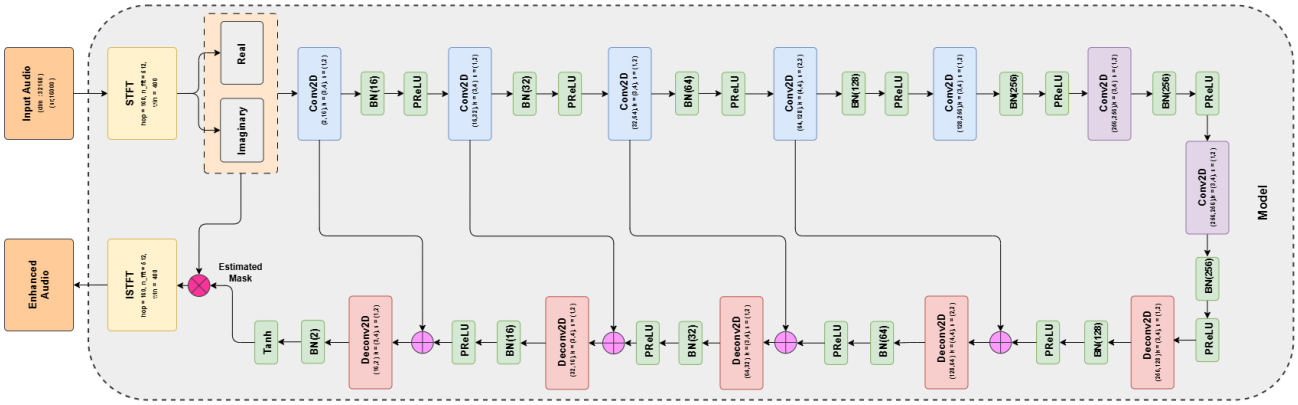

Figure 2. CNN-based Architecture

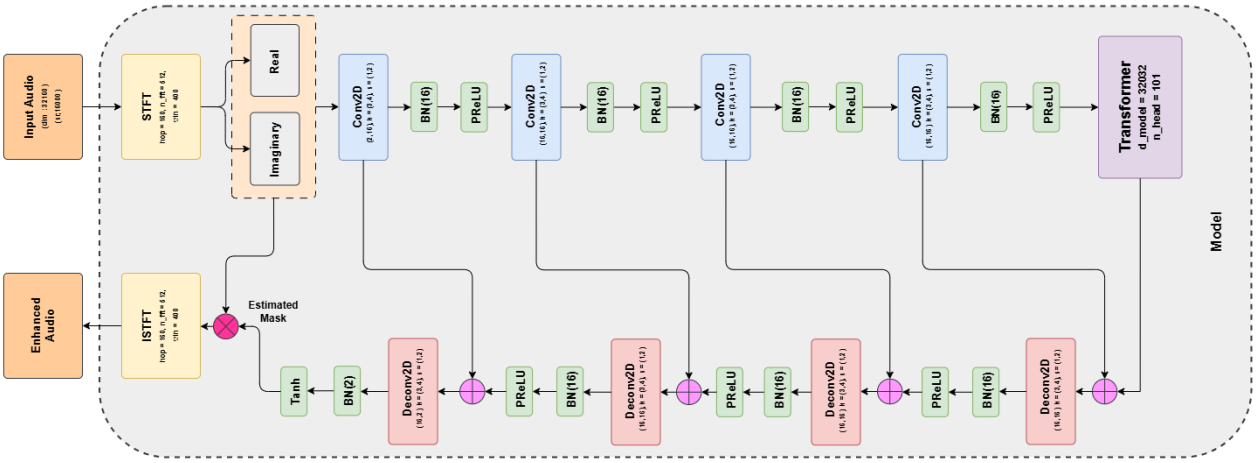

Figure 3. CNN+Transformer-based Architecture

SI-SSNR Loss: Scale-Invariant Segmental Signal-to-Noise Ratio (SI-SSNR) measures the quality of reconstructed signals by evaluating the enhanced signal, while operating on short, overlapping frames of audio. It works by calculating the signal-to-noise ratio for each frame and then calculating the average over the entire signal.

3.3. Evaluation Metric

The model is evaluated based on PESQ scores.

PESQ: As discussed above, Perceptual Evaluation of Speech Quality (PESQ) [29] is a widely used objective metric for assessing the quality of speech signals, particularly in the context of speech enhancement. By prioritizing perceptual quality, PESQ loss helps ensure that generated outputs are perceptually better for human listeners.

Using PESQ as both a loss function and an evaluation metric may initially seem counterintuitive due to concerns about circular reasoning, where a model optimizes for the same measure it's evaluated against. However, using the SISSNR loss creates an essential trade-off that addresses these issues. PESQ focuses on the perceptual quality of audio, aligning model outputs with human auditory preferences, while SISSNR loss emphasizes the relative power of the desired signal compared to noise. This dual approach allows the model to achieve a balanced performance - effectively capturing audio details and also minimizing noise.

4. Performance

Audio signal	**PESQ**
Clean	4.64
Noisy	1.46
Enhanced (updated CNN) output	2.84
Transformer+CNN output	2.60

Table 4. PESQ

As compared to the noisy PESQ of the dataset, which is at 1.46, our CNN-based model achieves a score of 2.90 and

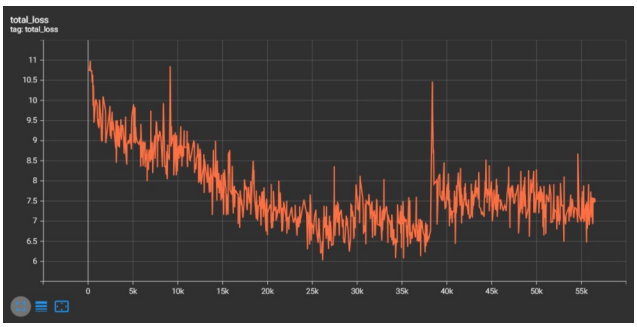

Figure 4. Loss (Mask+PESQ+SISNR) during training (updated CNN-based model)

a best performance of 2.84, which is significantly higher than the 2.02 achieved without batch normalization and the use of PReLU. The new transformer+CNN-based model achieves an error of 2.60 and a best-case performance of 2.70. The loss during training and final results are shown in Fig. 4, 5, 6, and 7.

Figure 5. PESQ obtained during testing updated CNN-based model

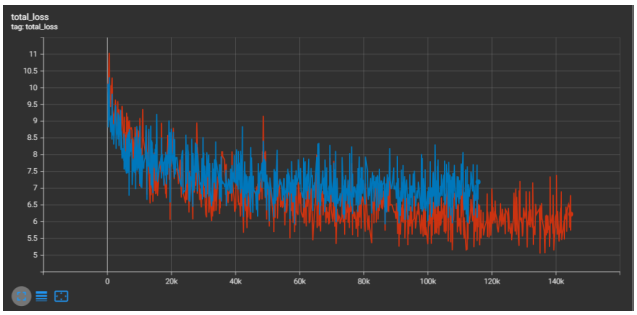

Figure 6. Loss during training (Transformer+CNN-based model)

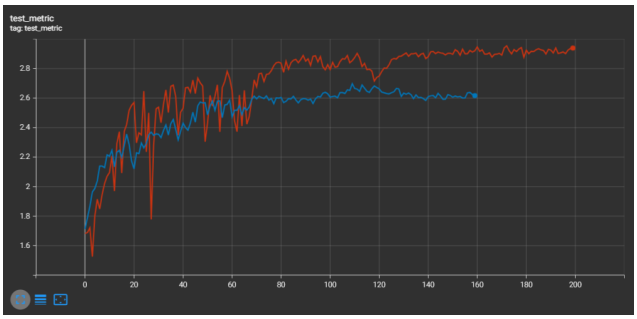

Figure 7. PESQ obtained during testing (Transformer+CNN-based model)

References

[1] Li-Ping Yang and Qian-Jie Fu. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The journal of the Acoustical Society of America*, 117(3):1001–1004, 2005. 1

[2] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech communication*, 52(5):450–475, 2010. 1

[3] Huijun Ding, Yann Soon, Soo Nee Koh, and Chai Kiat Yeo. A spectral filtering method based on hybrid wiener filters for speech enhancement. *Speech Communication*, 51(3):259–267, 2009. 1

[4] Marwa A Abd El-Fattah, Moawad I Dessouky, Alaa M Abbas, Salaheldin M Diab, El-Sayed M El-Rabaie, Waleed Al-Nuaimy, Saleh A Alshebeili, and Fathi E Abd El-samie. Speech enhancement with an adaptive wiener filter. *International Journal of Speech Technology*, 17:53–64, 2014. 1, 2

[5] Stephen So and Kuldip K Paliwal. Modulation-domain kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829, 2011. 1, 2

[6] K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 177–180. IEEE, 1987. 1

[7] Sharon Gannot, David Burshtein, and Ehud Weinstein. Iterative and sequential kalman filter-based speech enhancement algorithms. *IEEE Transactions on speech and audio processing*, 6(4):373–385, 1998. 1

[8] Ashutosh Pandey and DeLiang 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)*, pages 6875–6879. IEEE, 2019. 1, 2

[9] Craig Macartney and Tillman Weyde. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018. 1, 2

[10] Dongheon Lee, Seongrae Kim, and Jung-Woo Choi. Inter-channel conv-tasnet for multichannel speech enhancement. *arXiv preprint arXiv:2111.04312*, 2021. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[12] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*, 2016. 1

[13] Mojtaba Hasannezhad, Zhiheng Ouyang, Wei-Ping Zhu, and Benoit Champagne. An integrated cnn-gru framework for complex ratio mask estimation in speech enhancement. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 764–768. IEEE, 2020. 1

[14] Bryce Irvin, Marko Stamenovic, Mikolaj Kegler, and Li-Chia Yang. Self-supervised learning for speech enhancement through synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[16] Mariyadasu Mathe, Siva Prasad Nandyala, and T Kishore Kumar. Speech enhancement using kalman filter for white, random and color noise. In *2012 International conference on devices, circuits and systems (ICDCS)*, pages 195–198. IEEE, 2012. 3

[17] Hwai-Tsu Hu, Fang-Jang Kuo, and Hsin-Jen Wang. Supplementary schemes to spectral subtraction for speech enhancement. *Speech Communication*, 36(3-4):205–218, 2002. 2

[18] Mousa Al-Akhras, Khaled Daqrouq, and Abdul Rahman Al-Qawasmi. Perceptual evaluation of speech enhancement. In *2010 7th International Multi-Conference on Systems, Signals and Devices*, pages 1–6. IEEE, 2010. 2

[19] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. Attention wave-u-net for speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253, 2019. 2

[20] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: speech enhancement generative adversarial network. *CoRR*, abs/1703.09452, 2017. 2

[21] Jean-Marc Valin. A hybrid dsp/deep learning approach to real-time full-band speech enhancement. *CoRR*, abs/1709.08243, 2017. 2

[22] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy. A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech, 2020. 2

[23] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *Speech Communication*, 53(4):465–494, 2011. ISSN 0167-6393. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 3

[25] Donald S. Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for joint enhancement of magnitude and phase. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224, 2016. 3

[26] Aidan E.W. George, Christine Pickersgill, Belinda Schwerin, and Stephen So. A study on subspace-based estimation of stft real and imaginary modulation signals for speech enhancement. In *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–6, 2016. 3

[27] Belinda Schwerin and Kuldip Paliwal. Speech enhancement using stft of real and imaginary parts of modulation signals. 12 2012. 3

[28] Phase-aware speech enhancement with deep complex u-net. *CoRR*, abs/1903.03107, 2019. Withdrawn. 3

[29] 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)*, volume 2, pages 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023. 4, 5