INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Improved speech enhancement using TCN with multiple encoder-decoder layers

Vinith Kishore[1], *Nitya Tiwari*[1], *Periyasamy Paramasivam*[1]

[1]Samsung Research Institute Banglore, Bengaluru, 560037, India
<v.kishore, nitya.tiwari, periyasamy.p> @ samsung.com

Abstract

A deep learning based time domain single-channel speech enhancement technique using multilayer encoder-decoder and a temporal convolutional network is proposed for use in applications such as smart speakers and voice assistants. The technique uses encoder-decoder with convolutional layers for obtaining representation suitable for speech enhancement and a temporal convolutional network (TCN) based separator between the encoder and decoder to learn long-range dependencies. The technique derives inspiration from speech separation techniques that use TCN based separator between a single layer encoder-decoder. We propose to use a multilayer encoder-decoder to obtain a noise-independent representation useful for separating clean speech and noise. We present t-SNE –based analysis of the representation learned using different architectures for selecting the optimal number of encoder-decoder layers. We evaluate the proposed architectures using an objective measure of speech quality, scale-invariant source-to-noise ratio, and by obtaining word error rate on a speech recognition platform. The proposed two-layer encoder-decoder architecture resulted in 48% improvement in WER over unprocessed noisy data and 33% and 44% improvement in WER over two baselines.

Index Terms: multilayer encoder-decoder, speech enhancement, temporal convolutional network

1. Introduction

With the widespread adoption of automatic speech recognition (ASR) applications in mobile devices, especially in battery-operated devices such as earbuds and TV remotes, ASR systems are exposed to ever-greater background noises. Though ASR systems are robust to noisy speech to a certain level by virtue of getting trained with both real and simulated noise, it is an impossible task to cover all the kind of noise variations and levels in the training data. Hence, the performance of the ASR systems is significantly lower in real-life scenarios. Often, speech enhancement techniques are used to improve speech perception. In this paper, we focus on single-channel speech enhancement techniques that use noisy input signal from single microphone or utilize the output of a spatial beamforming filter.

Several single-channel speech enhancement techniques have been introduced in the literature that formulate the processing in the time-frequency domain [1]–[4]. The enhancement is generally carried out either using the statistical properties of speech and noise [5], [6], or by deep neural networks (DNN) based machine learning algorithms [7]–[10]. The existing deep learning methods for speech enhancement convert the noisy signal to a time-frequency (T-F) representation using a forward transformation. Enhancement is carried out by estimating a mask, and enhanced T-F representation is used to resynthesize the time domain output using inverse transformation. These methods use ideal binary mask (IBM) or ideal ratio mask (IRM) as training targets [11]. A multi-domain network, using time and frequency representation for enhancement was proposed in [12]. However, methods involving T-F representation generally use the noisy phase for the reconstruction of the output signal. This results in perceptible roughness in time domain output.

Recently, time domain techniques that process the signal directly in time-domain without converting into T-F domain and using DNNs for speech enhancement have been explored [13], [14], and [15]. Architectures with single layer encoder-decoder along with a temporal convolutional network (TCN) separator have been used successfully for speech separation that involves separating voice of target speaker from others [16], [17]. We propose to use similar architecture for speech enhancement task. As a single layer encoder-decoder may not be sufficient to obtain representation for efficient noise suppression, we propose the use of encoder-decoder with multiple layers to achieve more complex signal transformations. Use of encoder-decoder with multiple layers is inspired from the deep encoder-decoder architecture for speech source separation [18]. However, increasing the number of layers beyond a certain point may cause overfitting due to the noise-dependent encoded representation learnt by encoder. As a result, the system may have poor performance for audio containing noises, which are different from those used in training. Further, using more layers also increases computations. We conduct experiments with different number of layers for encoder and decoder while keeping the TCN architecture fixed. We use t-distributed stochastic neighbor embedding (t-SNE) [19] for analysis and visualization of overfitting. It helps us in selecting the optimum number of layers for speech enhancement. We use an objective measure of speech quality (PESQ) and scale-invariant source-to-noise ratio (SI-SNR) to evaluate the proposed technique. Further, we obtain the word error rate (WER) using an ASR system, which is one of the primary target application systems.

2. Proposed framework

Single-channel speech enhancement is an estimation problem that uses the noisy signal to obtain a clean speech estimate. Fig. 1 shows the proposed architecture that uses multilayer encoder-decoder to learn more generalized patterns resulting in complex signal transformation suitable for speech enhancement. TCN forms the core of the enhancement framework and is used between encoder and decoder to learn the long-range dependencies from the encoded output and to obtain the enhanced speech mask. The mask multiplied with noisy encoded representation results in enhanced encoded representation, which is finally transformed by the decoder to an enhanced time-domain output.

The noisy input is divided into M overlapping frames with L samples and a shift of S samples. The encoder-decoder layers use 1-D convolutional operations. The first encoder layer applies a linear transformation and transforms M frames with L

http://dx.doi.org/10.21437/Interspeech.2020-3122

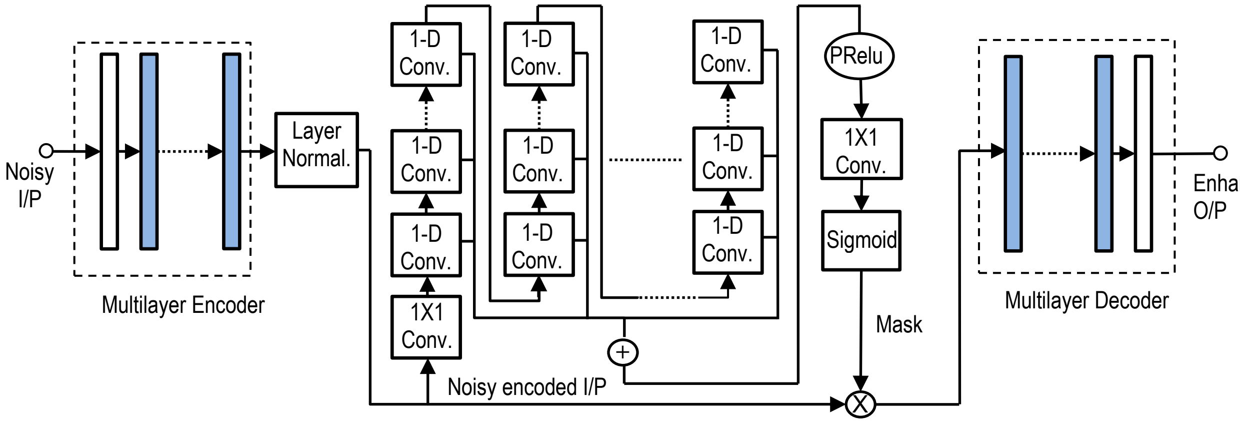

Fig 1. *Proposed architecture with encoder-decoder with multiple layers and TCN for mask estimation (adapted from [16]).*

samples to a representation with N samples. The subsequent encoder layers apply non-linear transformation using N kernels of size 3 and parametric rectified linear unit (PRelu) as the activation function. The TCN separator forms the core of the enhancement framework as it was reported to outperform long short-term memory (LSTM) recurrent neural networks in sequence modelling tasks [20]. It exhibits longer memory with relatively simple and easy to train architecture. We use TCN separator, similar to that used in speech separation network "ConvTasNet" [16], consisting of stacked dilated 1-D convolutional blocks with exponentially increasing dilation factors to capture long-range dependencies without a significant increase in model size. In our architecture, TCN uses a stack of eight convolutional blocks with dilation factors of 1, 2, 4, …, and 128, and three such stacks are concatenated. The noisy encoded input multiplied with the mask estimated using TCN is given as input to the decoder. The decoder consists of the same number of non-linear convolutional layers as used in the encoder and a final linear transformation layer that coverts N-sample representation to L-sample time-domain output. The enhanced output is resynthesized using overlap-add procedure.

A similar architecture, "TCNN" [15], for speech enhancement that uses seven 2-D convolutional encoder-decoder layers with skip connection and a TCN with dilation rates of 1, 2, 4, …, 32 has been reported earlier. However, TCNN uses the TCN to directly obtain the enhanced representation. In our architecture, TCN is used to obtain the enhanced mask. Enforcing TCN to learn mask rather than enhanced representation helps it in learning a filter that must select only the target speech from the noisy encoded representation. We use TCN with dilation rates of up to 128 for obtaining long-range dependencies with smaller frame size and give a t-SNE based analysis to show that a two-layer encoder-decoder gives a representation appropriate for speech enhancement application. We present results for both causal and non-causal implementations and compare the results with ConvTasNet [16] and TCNN [15]. The ConvTasNet model, which was originally proposed for speaker separation, is trained for speech enhancement task for comparison.

3. Experiments

The speech enhancement experiments were carried out using "train-clean-360" subset from Librispeech [21] dataset and single-channel speech from Chime-3 [22] dataset. The noisy mixtures, at 0, 5, and 10 dB SNR, were generated by adding babble, cafeteria, pedestrian and street noises from Chime-3 [22] dataset to clean speech. The enhancement networks were trained and validated using generated Librispeech noisy mixtures of 544 and 68 hours, respectively. Testing was done with noisy Librispeech and Chime-3 mixtures of 68 and 6 hours, respectively. To examine the effect of encoder-decoder depth, number of convolutional layers in encoder-decoder were selected as 0, 2, and 4 resulting in three architectures "enc-dec0", "enc-dec2", and "enc-dec4". ConvTasNet [16] and TCNN [15] were used as a baseline models for comparison. Non-causal implementation used TCN with symmetric non-causal convolutions where output at time t is convolved with elements earlier as well as after time t in the previous layer. Causal convolutions, with no information leakage from future to past, were used for causal implementation. All the implementations used a sampling rate of 16 kHz. The parameters for ConvTasNet were chosen as $\{L = 20, S = 10$, and N = 256$\}$ as preliminary experiments showed it to be better for speech enhancement than those reported in [16]. For TCNN the parameters were chosen as $\{L = 320, S = 160$, and $N = 256\}$ as reported in [15]. For proposed architectures, the parameters were empirically chosen as $\{L = 16, S = 8$, and $N = 512\}$ that resulted in the best performance. All the implementations used scale-invariant signal-to-noise ratio (SI-SNR) as objective function for training using Adam [23] optimizer. The learning rate, initially set as 10^{-3}, was halved every 3 epochs for no change in validation SI-SNR. The training was carried out till the validation SI-SNR reached 13 dB, or for 50 epochs, whichever was achieved earlier.

The t-SNE analysis of the representation obtained at the output of the last encoder layer was carried out to analyze the effect of increase in the number of non-linear convolutional layers in encoder and decoder. Unlike T-F transformation, where visualization of the transformed representation is possible through spectrograms, the visualization of high dimensional representation obtained at the encoder output is not feasible. Thus, the t-SNE method that embeds high dimensional vectors in a two-dimensional space was used for analysis. Scatter plots were used for visualization of the two-dimensional representation. Initial analysis was carried out by giving clean speech and noise as inputs to the proposed architectures separately. The experiment was to see if the encoders learn a separate representation for speech and noise resulting in separate speech and noise clusters when plotted on the same scatter plot. A set of clean sentences from a male and a female speaker and four noises were used in this experiment. In the next step, the analysis was carried by giving noisy speech as input to the proposed architectures to see if the encoders are able to learn a noise-independent representation for noisy input. For a noise-independent representation, the scatter plots for

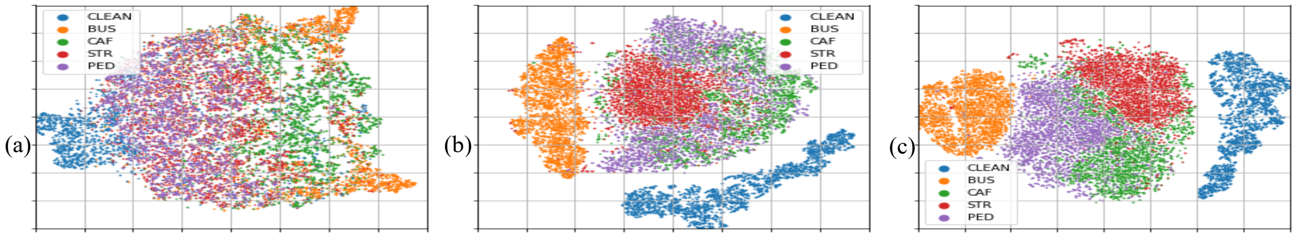

Fig 2. *T- SNE scatter plots for clean speech and noises using non- causal implementations of (a) enc- dec0, (b) enc- dec2, (c) enc- dec4.*

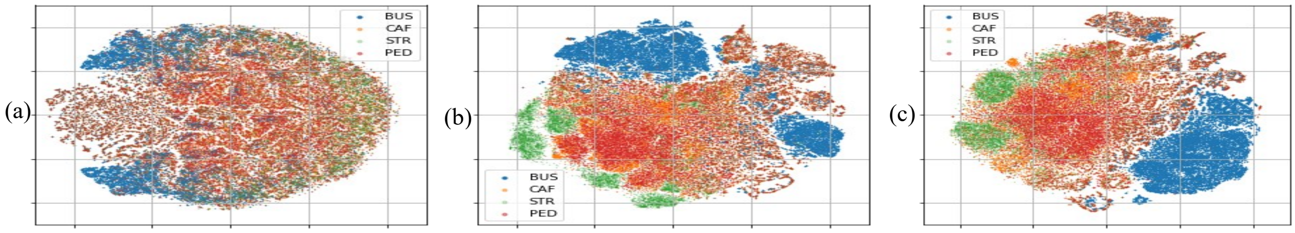

Fig 3. *T-SNE scatter plots for noisy speech using non-causal implementations of (a) enc-dec0, (b) enc-dec2, (c) enc-dec4.*

noisy input should not exhibit noise-dependent clusters. In this experiment, two noisy sentences (one from male and one from female) corrupted with different noises at 5 dB SNR were used.

The performance evaluation was carried out using PESQ [24] and SI-SNR objective measures. Further, word error rate (WER) was obtained on a hybrid ASR platform with deep bidirectional LSTM based acoustic model (AM) that has look-ahead convolutional layers [25], [26]. The ASR used an N-gram language model (LM) with two pass decoding with first pass using smaller LM (up to 3-grams) followed by on-the-fly rescoring with larger LM (up to 6-grams) with a vocabulary of 1 million words.

4. Results

The results obtained from t-SNE analysis for non-causal implementations are shown as scatter plots in Fig. 2 and Fig. 3. The results using clean speech from a male speaker and four noises as separate inputs to three proposed architectures are shown in Fig. 2. It can be observed from scatter plot corresponding to enc-dec0 (in Fig. 2(a)) that it results in overlapped points indicating that it is not able to learn a separate representation for clean speech and noise. However, in the scatter plots for enc-dec2 and enc-dec4 (in Fig. 2(b) and Fig. 2(c), respectively), there is a distinct cluster for clean speech. In addition, there is a different cluster for bus noise and a trend for segregation based on different noise types. This indicates that enc-dec2 and enc-dec4 are able to learn a separate representation for clean speech and noise.

The results of t-SNE analysis using speech corrupted with different noises at 5 dB SNR are shown in Fig. 3. It can be observed that for all three architectures, there are no separate clusters formed depending on the noise type. For enc-dec0, the points corresponding to all four noise types are uniformly spread. When plots for enc-dec2 and enc-dec4 are compared, enc-dec4 shows a trend for noise-dependent segregation of the points indicating that it tends to learn a noise-dependent representation of noisy input. Analyzing results in Fig.2 and Fig. 3 together, we can conclude that enc-dec2 based architecture is able to learn a separate representation of clean speech and noise, and at the same time has a noise-independent representation for noisy input. Thus, out of the three proposed architectures, enc-dec2 is more suitable for speech enhancement.

Table 1. *PESQ and SI-SNR scores for unprocessed (Unpr.) noisy input and enhanced output obtained using non-causal implementations of different architectures.*

Score	Data	Enhancement approach					
		Unpr.	Conv TasNet	TCNN	enc-dec0	enc-dec2	enc-dec4
PESQ	Libri	1.1	2.0	2.2	2.5	2.6	2.6
	Chime	1.6	2.6	2.1	2.3	2.3	2.2
SI-SNR	Libri	1.6	12.7	2.1	11.4	10.8	10.6
	Chime	−2.6	11.7	6.4	7.7	7.9	7.9

The average PESQ, SI-SNR, and WERs for Librispeech (Libri) and Chime-3 (Chime) were obtained for three proposed architectures and for ConvTasNet [16] and TCNN [15]. The PESQ and SI-SNR scores for unprocessed noisy input (Unpr.) and for enhanced output obtained using five architectures with non-causal implementations are given in Table 1. It is observed that PESQ scores for enc-dec2 are the highest and those for TCNN are the lowest. However, the SI-SNR scores are the highest for ConvTasNet and are not in agreement with results from PESQ. This may be attributed to the fact that PESQ measure incorporates a perceptual quality criterion, which is not the case with SI-SNR that uses ratio of the power in error between target and estimate and the power of target. Thus, further evaluation was carried out using WER on the speech recognition platform. The WER for unprocessed input and for the enhanced output from five architectures with non-causal implementations are given in Table 2. It is observed that enc-dec2 results in the lowest WER and ConvTasNet results in the highest WER. This indicates that although ConvTasNet removes more noise resulting in higher SI-SNR improvements, it causes more speech distortions resulting in higher WER. The PESQ and SI-SNR scores and WER for enc-dec4 are slightly poorer than enc-dec2. This may be attributed to the noise-dependent learning in enc-dec4, as observed in the t-SNE analysis.

The algorithmic delays (obtained using the size of the receptive fields of the separator and the encoder-decoder depths) for non-causal implementations of ConvTasNet,

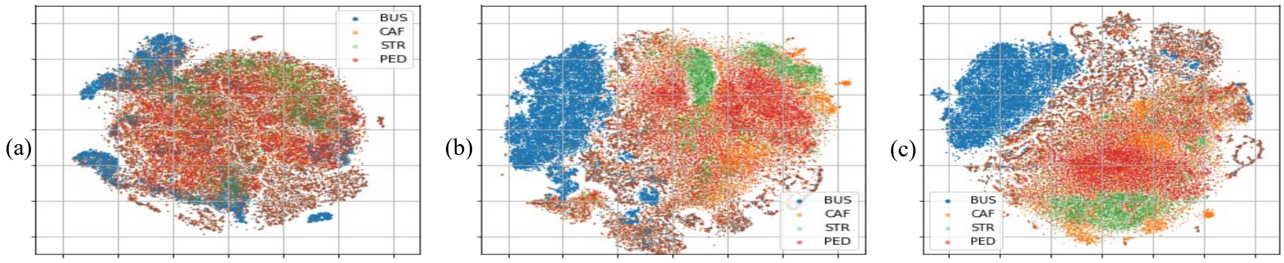

Fig 4**.** *T-SNE scatter plots for noisy speech with 4 noises at 5 dB SNR using causal implementations of (a) enc-dec0, (b) enc-dec2, and (c) enc-dec4.*

Table 2. *WER in % for unprocessed (Unpr.) noisy input and enhanced output obtained using non-causal implementations of different architectures.*

Data	Processing					
	Unprc.	Conv TasNet	TCNN	enc-dec0	enc-dec2	enc-dec4
Libri	37.1	34.8	29.0	20.7	19.4	19.6
Chime	47.8	46.1	37.2	30.0	28.8	29.0

Table 3. *PESQ SI-SNR, and WER for enhanced output using causal implementation of three proposed architectures.*

Score	Data	Processing		
		enc-dec0	enc-dec2	enc-dec4
PESQ	Libri	2.2	2.3	2.3
	Chime	1.9	2.0	2.0
SI-SNR	Libri	9.8	9.6	9.8
	Chime	6.3	6.6	6.8
WER	Libri	26.4	23.9	23.4
	Chime	37.9	34.7	33.3

TCNN, enc-dec0, enc-dec2, enc-dec4 were 956 ms, 2000 ms, 765 ms, 767 ms, and 769 ms, respectively. It can be seen that algorithmic delay for the proposed architectures is lower than the corresponding implementations of the baselines. However, the algorithmic delays obtained from non-causal implementations may not be suitable for real-time applications. Thus, preliminary investigations were carried out using causal implementations of the three proposed architectures.

The t-SNE scatter plots for clean speech and noise as separate inputs to the causal implementations were similar to those in Fig. 2 and thus are not shown. The scatter plots for noisy speech as input to the causal implementations are shown in Fig. 4. It can be observed that for enc-dec0, the points corresponding to all noises are uniformly spread. Unlike Fig. 3, the differences in scatter plots for causal enc-dec2 and enc-dec4 is not clearly visible. The PESQ, SI-SNR, and WER for the causal implementations of the three proposed architectures are given in Table 3. The scores are poorer than those corresponding to non-causal implementations, as the non-causal implementations make better predictions using future samples. The scores from enc-dec0 are the poorest among the three, indicating that it is not suitable for enhancement. The scores for enc-dec4 are better than those for enc-dec2. The results of the t-SNE analysis and objective evaluation indicate that enc-dec4 is better than enc-dec2. However, evaluation using a higher number of layers is needed to obtain the optimal number of layers for causal implementations. For causal implementations, the algorithmic delay (frame duration + frame shift) of the three proposed architectures was 1.5 ms, and those for ConvTasNet and TCNN were 1.85 ms and 30 ms, respectively. Algorithmic delays for causal implementations are significantly lower than non-causal implementations. For real-time applications that can afford a slightly higher delay, a tradeoff exists between performance degradation resulting from causal implementation and delay resulting from non-causal implementation. An implementation that buffers future frames and uses asymmetric non-causal convolutions for enhancement may be more suitable for such applications.

The number of parameters used in both causal and non-causal implementations of ConvTasNet, TCNN, enc-dec0, enc-dec2, enc-dec4 were 5.1 M, 5.1 M, 6 M, 9 M, and 12 M, respectively. This shows that the improved performance of the proposed architectures comes at the cost of an increase in the number of trainable parameters, and model compression techniques should be used to reduce the number of parameters.

5. Conclusion and future direction

A single-channel speech enhancement technique that uses a multilayer encoder-decoder with a TCN has been proposed. The use of multilayer encoder-decoder helps in achieving complex signal transformations suitable for speech enhancement. Visual analysis using t-SNE scatter plots was carried out for the three proposed architectures (enc-dec0, enc-dec2, and enc-dec4). The analysis for non-causal implementations showed that enc-dec2 results in a noise-independent representation appropriate for speech enhancement. The objective evaluation using PESQ, SI-SNR, and WER was carried out for non-causal implementations, and the results were compared with ConvTasNet and TCNN. The PESQ scores and WER for proposed architectures were better than the baselines, and best scores were obtained for enc-dec2. The algorithmic delay for the proposed architectures was lower than the baselines. To examine the suitability of the proposed architectures for real-time applications, investigation was carried out using causal implementations. The causal implementations resulted in lower algorithmic delays than non-causal implementations at the cost of degradation in performance. The number of parameters in the proposed architectures were higher than the baselines. Future work involves model optimization and use of model compression technique to reduce the number of parameters to make the proposed architecture suitable for applications demanding low memory and computational requirements. Further tests need to be carried out to examine the generalization capability and suitability of the proposed technique for use in unseen noisy environments.

6. Acknowledgements

This work is supported by the AI R&D Group and Samsung Research in Korea. We especially thank Vice President Kyoung-Gu Woo, Vice President Chanwoo Kim, Dr. Chang Woo Han, and Dr. Sky Jung for their support and guidance.

7. References

[1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 32, pp. 1109–1121, 1984.

[2] M. N. Schmidt and J. Larsen, “Reduction of non-stationary noise using a non-negative latent variable decomposition,” in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Cancun, Mexico, 2008, pp. 486–491.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, “DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art,” in *Synthesis Lectures on Speech and Audio Processing*, B. H. Juang and San Rafael, Eds. San Rafael, CA, USA: Morgan and Claypool, 2013.

[4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. Interspeech*, Lyon, France, 2013, pp. 436–440.

[5] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, 2003.

[6] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, vol. 48, pp. 220–231, 2006.

[7] S. R. Park and J. W. Lee, “A fully convolutional neural network for speech enhancement,” in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 1993–1997.

[8] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 23, pp. 7–19, 2015.

[9] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, “Speech enhancement using Bayesian wavenet,” in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2013–2017.

[10] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, “Convolutional-recurrent neural networks for speech enhancement,” in *Proc. ICASSP*, Calgary, AB, Canada, 2018, pp. 2401–2405.

[11] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, pp. 1849 – 1858, 2014.

[12] J.-H. Kim, J. Yoo, S. Chun, A. Kim and J.-W. Ha, "Multi-domain processing via hybrid denoising networks for speech enhancement", arXiv preprint arXiv:1812.08914, 2018.

[13] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” arXiv preprint arXiv:1703.02205, 2017.

[14] A. Pandey and D. Wang, “A new framework for supervised speech enhancement in the time domain,” *in Proc. Interspeech*, Hyderabad, India, 2018, pp. 1136–1140.

[15] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” *in Proc. ICASSP*, Brighton, UK, 2019, pp. 6875–6879.

[16] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, pp. 1256 – 1266, 2019.

[17] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. Liu, “FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing,” *in Proc. ASRU*, Sentosa, Singapore, 2019.

[18] B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, “An empirical study of Conv-Tasnet,” *in Proc. ICASSP*, Barcelona, Spain, 2020.

[19] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] S. Bai, J. Z. Kolter, V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” arXiv preprint arXiv:1803.01271, 2018.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” *in Proc. ICASSP*, Brisbane, Queensland, Australia, 2015.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines,” *in Proc. ASRU*, Scottsdale, AZ, USA, 2015.

[23] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[24] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Rec. ITU-T P.862, International Telecommunications Union, Geneva, Switzerland, 2001.

[25] A. Graves, N. Jaitly and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding*, Olomouc, 2013, pp. 273-278.

[26] C. Wang, D. Yogatama, A. Coates, T. Han, A. Hannun, and B. Xiao, “Lookahead convolution layer for unidirectional recurrent neural networks,” in *Proc. ICLR 2016 workshop,* https://openreview.net/forum?id=91EowxONgIkRlNvXUVog

.