Incorporating Symbolic Sequential Modeling for Speech Enhancement

Chien-Feng Liao[1], Yu Tsao[1], Xugang Lu[2], Hisashi Kawai[2]

[1]Research Center for Information Technology Innovation, Academic Sinica, Taiwan
[2]National Institute of Information and Communications Technology, Japan
`r06946002@ntu.edu.tw`, `yu.tsao@citi.sinica.edu.tw`, {`xugang.lu,hisashi.kawai`}`@nict.go.jp`

Abstract

In a noisy environment, a lossy speech signal can be automatically restored by a listener if he/she knows the language well. That is, with the built-in knowledge of a "language model", a listener may effectively suppress noise interference and retrieve the target speech signals. Accordingly, we argue that familiarity with the underlying linguistic content of spoken utterances benefits speech enhancement (SE) in noisy environments. In this study, in addition to the conventional modeling for learning the acoustic noisy-clean speech mapping, an abstract symbolic sequential modeling is incorporated into the SE framework. This symbolic sequential modeling can be regarded as a "linguistic constraint" in learning the acoustic noisy-clean speech mapping function. In this study, the symbolic sequences for acoustic signals are obtained as discrete representations with a Vector Quantized Variational Autoencoder algorithm. The obtained symbols are able to capture high-level phoneme-like content from speech signals. The experimental results demonstrate that the proposed framework can obtain notable performance improvement in terms of perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) on the TIMIT dataset.

Index Terms: Speech enhancement, deep learning, symbolic representation, multi-head attention

1. Introduction

Speech enhancement (SE) has been commonly used as a front-end module in speech-related applications, such as robust automatic speech recognition (ASR) [1–3], automatic speaker recognition, and assistive listening devices [4–6]. Recently, deep learning (DL)-based SE models have also been proposed and extensively investigated [7–11]. The main idea in these DL-based SE models is to learn the complex mapping functions between noisy speech and clean speech. In most studies, the mapping functions are learned based on a large quantity of well-prepared noisy-clean speech pairs in the acoustic domain without considering the underlying linguistic structure.

In a noisy environment, audiences can automatically restore a noise-masked speech based on their knowledge of a "language model", and the restoring ability depends on the effectiveness of this internal "language model". For example, in noisy environments, great effort is required for non-native listeners [12]. These studies indicate that the linguistic-related information is helpful to retrieve target speech signals from the noisy ones. Accordingly, it is argued in this study that it is beneficial to incorporate text information (phonemes or words) into an SE system for improved performance.

In [13], oracle transcription is used to extract time-aligned text features as auxiliary input to the DNN model. Even though this can be formulated as a text-to-speech application, it is not practical under SE scenarios to assume to have ground-truth transcription. Several studies incorporate recognition results or outputs from acoustic models. In [14], a phone-class feature is augmented to standard acoustic features as input for de-reverberation. In [9], an ASR and an SE system are trained iteratively, where each system's input depend on the other's output. In [15, 16], a set of DNNs were trained as enhancement models, one for each specific phoneme. During inference time, an ASR or a phoneme classifier was used to determine which DNN to use. Even though promising results have been obtained, these approaches have major drawbacks. First, the recognition model is not jointly trained and thus optimization cannot be achieved for both systems. If the ASR system is incorrect, errors will be propagated to the downstream SE system. Secondly, heavily equip SE with an ASR system may be undesirable because SE is commonly used as a preprocessor. To overcome these obstacles, [17] proposed learning a Deep Mixture of Experts (DMoE) network where the experts are DNNs, whose outputs are combined by a gating DNN. The gating DNN is trained to assign a combination weight to each expert. This results in splitting the acoustic space into sub-areas in an unsupervised manner, which is similar to our proposed method.

van den Oord et al. [18] recently proposed the Vector Quantized Variational Autoencoder (VQ-VAE), in which the stochastic continuous latent variables from the original VAE are replaced with deterministic discrete latent variables. It maintains a set of prototype vectors, i.e., a predefined size of learnable codebook. During forward pass, feature vectors produced by the encoder are replaced with their nearest-neighbor in the codebook. Although this quantization component acts as an information bottleneck and can regularize the power of the encoder, the discrete latent variables are more interpretable and tend to learn higher level representations, which can naturally correspond to phoneme-like features for given speech signal inputs. In [19], a comprehensive study of VQ-VAE applied to speech data was carried out, and it was demonstrated that VQ-VAE achieves better interpretability and information separation (such as disentangling speaker characteristics) than VAEs and AEs. Furthermore, the extracted representation allowed for accurate mapping into phonemes and achieved competitive performance on an unsupervised acoustic unit discovery task. Overall, the characteristics of the VQ-VAE make it a suitable component to reinforce an SE system with high-level linguistic information.

In this study, an SE system with U-Net architecture [20–23] is proposed. Moreover, a "symbolic encoder" is developed, consisting of DNNs and the vector quantization mechanism in VQ-VAE. The extracted symbolic sequence is then connected to the U-Net via multi-head attention mechanism [24]. Thereby, the two components can be jointly trained without the need of any supervised transcription or explicit constraints. The results demonstrate a notable improvement in terms of objective measures including perceptual evaluation of speech quality (PESQ) [25] and short-time objective intelligibility (STOI) [26].
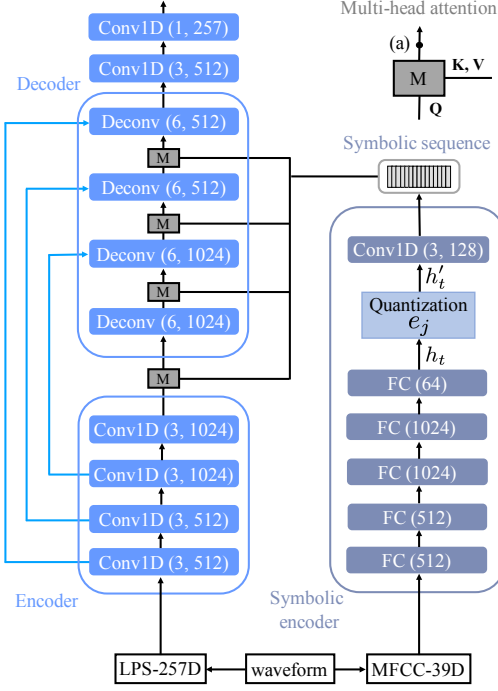
Figure 1: *Proposed system consisting of a U-Net architecture, a symbolic encoder, and an attention mechanism. Conv1Ds and Deconvs are in the format (filterWidth, outputChannels), and the down-sample\up-sample rates are both 2. FC (outputChannels) denotes the fully connected layer.*

The rest of the paper is organized as follows. In Section 2, the proposed approach is detailed, including each components of the system and the objective functions. The experiment settings and results are presented in Section 3. Finally, Section 4 concludes the paper.

2. System architecture

A paired training dataset $\{x_i, y_i\}_{i=1}^N$, where x_i is the input noisy speech and y_i is the target clean speech. The proposed system is shown in Figure 1. It consists of the following parts: an encoder network $E_{se}(x)$ consisting of convolutional layers that extracts the feature sequence; another encoder network called **symbolic encoder** $E_{symb}(x)$ consists of fully connected layers and extracts the **symbolic sequence** by vector quantization. Multi-head attention function and skip-connection are used to connect the two encoder outputs with the decoder $Dec(E_{se}(x), E_{symb}(x))$. All components are jointly trained using mean-squared-error (MSE) loss function between the clean speech and the enhanced speech:

$$\mathcal{L}_{mse} = \frac{1}{N} \sum_{i=1}^N ||Dec(E_{se}(x_i), E_{symb}(x_i)) - y_i||_2^2 \quad (1)$$

The quantization mechanism and the multi-head attention mechanism will now be briefly explained; for more detailed information readers may refer to [18] and [24], respectively.

2.1. Symbolic Encoder

The symbolic encoder reads a sequence of acoustic features as input. Here, mel-frequency cepstral coefficients (MFCCs)

are used, as suggested in [19]. A sequence of hidden vectors $\{h_t \in R^D, t = 1, ..., T\}$ is extracted by the fully connected layers, where D is the dimensionality and T denotes the sequence length. A **symbolic book** that contains a set of prototype vectors $\{e_j \in R^D, j = 1, ..., M\}$ is maintained, where M is the size of the book. The hidden vectors h_t will be replaced by the nearest prototype vector in the symbolic book. That is, $h_t' = e_k$, where $k = argmin_j ||h_t - e_j||_2^2$. During the training phase, the prototypes in the symbolic book are updated as a function of exponential moving averages of h. This method is presented in the original paper as an alternative way to update the book, and has the advantage of faster training speed than using an auxiliary loss. To prevent the symbolic encoder diverge in h with unbounded value, [18] also uses the "commitment loss" to encourage the symbolic encoder to produce vectors lying close to the prototypes. Overall, the full system is optimized with two loss terms: the MSE between the enhanced acoustic features and the clean target features, and the commitment loss:

$$\mathcal{L}_{total} = \mathcal{L}_{mse} + \lambda ||h_t - sg(e_k)||_2^2 \quad (2)$$

where λ is a hyperparameter that controls the importance of the commitment loss and $sg(.)$ denotes the stop-gradient operation. It should be noted here that the gradient of the loss can be back-propagated to the symbolic encoder using the straight-through estimator presented in [27].

2.2. Multi-head Attention

Multi-head attention (MHA) was first proposed in the transformer architecture [24] for machine translation, and recently explored in various speech-related tasks including end-to-end ASR [28] and text-to-speech system [29]. MHA extends the conventional attention mechanism to have multiple heads, where each head generates a different attention weight vector. This allows the decoder to jointly retrieve information from different representation subspaces at different positions, which facilitates focusing on the various structures of the symbolic sequence. The input argument consists of queries Q, keys K, and values V, i.e., $Attn(Q, K, V)$. In this study, MHA is used before each layer in the decoder. Every time-step of the decoder output acts as an query to attend on the symbolic sequence. The output of MHA will be concatenated with the skip-connection and fed to the proceeding decoder layer together. Formally, we have the symbolic sequence h and the skip-connection from the encoder at each layer $\{s^{(l)}, l = 1, ..., L\}$. The output of each layer in the decoder is the following:

$$d^{(l)} = Deconv(Concat(s^{(L-l+1)}, Attn(d^{(l-1)}, h', h')))$$

where $l = 1, ..., L$ is the depth of the decoder layer and $d^{(0)}$ is the encoder output.

2.3. Model Details

The symbolic encoder consists of four fully connected layers, each followed by a ReLU activation function and a dropout layer [30] with a drop rate of 0.2. A linear projection layer then maps the hidden vectors h_t to $D = 64$ dimensions in order to perform quantization. After the quantization, an one-dimensional (1-D) convolutional layer is used to give the symbolic sequence contextual information. Four heads are used in MHA, leading to point (a) in Figure 1, with a dimensionality of 4×128. As in the original transformer, the positional encodings

Table 1: *Average PESQ, and STOI scores for evaluating baseline models and the proposed method on the test set under three unseen noise environments at five SNR levels and the average scores across all SNRs. The unprocessed test set is denoted by **Noisy**. Size of the symbolic book is shown in the parenthesis. The highest scores per metric are highlighted with bold text, excluding **Oracle**.*

SNR	Noisy		U-Net		U-Net-MOL		Proposed (64)		Oracle	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-6	1.213	0.532	1.685	0.602	1.800	0.619	**1.828**	**0.624**	1.961	0.703
-3	1.353	0.598	1.880	0.669	1.974	0.681	**2.045**	**0.693**	2.140	0.741
0	1.517	0.669	2.071	0.725	2.140	0.736	**2.240**	**0.750**	2.306	0.776
3	1.702	0.739	2.237	0.770	2.290	0.779	**2.416**	**0.794**	2.456	0.806
6	1.902	0.823	2.387	0.805	2.424	0.813	**2.581**	**0.830**	2.592	0.831
Avg.	1.537	0.669	2.052	0.714	2.126	0.725	**2.222**	**0.738**	2.291	0.771

are also added to the inputs of the MHA, providing some information about the position of the tokens in the sequence. Queries and keys are first passed through a linear projection layer with 256 nodes before being divided into multiple heads. For the encoder, the frequency axis is treated as channel; thus, 1-D convolutional layers are used. The sequence length is down-sampled at each layer using a stride of 2 instead of pooling layers. The decoder is a mirrored version of the encoder with deconvolutional layers and larger kernel width. LeakyReLU is used as activation function in both the encoder and the decoder. Finally, the decoder output is projected back to frequency dimension using 1-D convolution with a kernel width of 1.

3. Experiments

The experiments were conducted on the TIMIT database [31]. A total of 3696 utterances from the TIMIT training set (excluding SA files) were randomly sampled and corrupted with 100 noise types from [32] at six SNR levels, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to obtain 40-hour multi-condition training set, consisting of pairs of clean and noisy speech utterances. Another 100 utterances were randomly sampled to construct the validation set. They are mixed with cafeteria babble noise at 4 SNR levels (-4 dB, 0 dB, 4 dB, and 8 dB), which is unseen from the training set. The 192 utterances from the core test set of the TIMIT database were used to construct the test set for each combination of noise types and SNR levels. To evaluate the system on unseen noise types, three other noise types, namely Buccaneer1, Destroyer engine, and HF channel from the NOISEX-92 corpus [33], were adopted. In the following experiments, the SE algorithm will be evaluated in terms of speech quality and speech intelligibility. Therefore, PESQ and STOI, respectively, will be used to evaluate the enhanced speech, respectively. Higher scores represent better performance.

3.1. Implementation

The sampling rate of the speech data was 16 kHz. For the encoder input, time-frequency (T-F) features were extracted using a 512-point short time Fourier transform (STFT) with a hamming window size of 32 ms and a hop size of 16 ms, resulting in feature vectors consisting of 257-point STFT log-power spectra (LPS). For the symbolic encoder, standard 13 MFCC features (extracted at a rate identical to that for the LPS features) were used and concatenated with their temporal first and second derivatives. MFCCs are often used in speech recognition because they are pitch invariant and slightly robust to noise. A better quantization behavior was observed using MFCC compared to LPS in the preliminary experiments. The input was a

Table 2: *Average PESQ and STOI performance on the validation set for different size of the symbolic book.*

Book size M	PESQ	STOI
39	2.061	0.711
64	**2.108**	**0.713**
128	2.027	0.712
256	2.041	0.711

segment of 64 frames (approximately 1 s), and was normalized by mean and standard-deviation before being fed to the system. Finally, the decoder outputs were synthesized back to the waveform signal via inverse Fourier transform and an overlap-add method. The phases of the noisy signals were used for the inverse Fourier transform. All models were trained on mini-batches of 32. The Adam optimizer [34] was used with learning rate $lr = 0.0001$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. The weight of the commitment loss λ was set to 0.2, which is close to the original setting in VQ-VAE, and it did not have significant impact on performance. Early stopping was performed based on the validation set to prevent overfitting.

3.2. Baseline model

We constructed the baseline model by excluding the symbolic encoder component, i.e., the left part of Figure 1 without MHA. This model is denoted by **U-Net**. Subsequently, the multi-objective learning method proposed in [35] was adopted in the baseline model. The input of the **U-Net** was augmented by MFCC features, and an additional objective was added to $\mathcal{L}_{total}$ during training to predict clean MFCCs. This baseline is denoted by **U-Net-MOL**. Finally, the benefit of using real text information as in [13] should be demonstrated. The phoneme level transcriptions provided by the TIMIT corpus were used to obtain frame-wise phoneme labels. The input MFCCs of the symbolic encoder were then replaced by the phoneme embeddings (embeddings are jointly learned). Quantization was discarded because the real phonetic information was provided. This is considered as an oracle model, as it takes correct transcriptions as input. This system will be called **Oracle**.

3.3. Results

Table 1 presents the results of the average PESQ, SSNR, and STOI scores on the test set for different systems. "Noisy" denotes unprocessed noisy speech, and the proposed model is shown with the symbolic book size of 64 as a representative. From this table, it can be observed that **Oracle** performed the
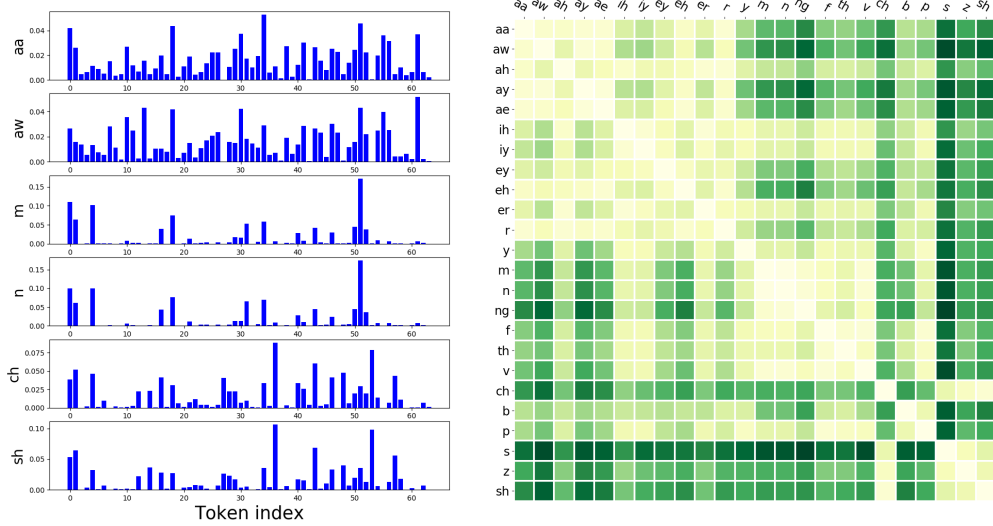
Figure 2: *Left: Histogram: each bin represents the token index, and the value shows how many times this token was chosen, given the corresponding phoneme. Right: The element on location (i,j) represents JS-divergence between the histogram from the i-th phoneme and the histogram from the j-th phoneme. Darker color implies larger divergence. Some phonemes were omitted owing to space limitations.*

best, as expected. This also confirmed the hypothesis that, given correct text information, the SE system can be more robust to noisy environments. Furthermore, the proposed model outperformed **U-Net** and **U-Net-MOL** at every SNR levels. It should be noted here that the system had fewer trainable parameters compared to the baselines, as MHA reduces the dimension to 4×128, as mentioned in Section 2.3. Thus, the improvement was not due to model complexity. Table 2 shows the de-noise ability of the proposed method with different size symbolic book. It can be seen that performance peaked for a size of 64. During the experiments, it was also observed that the symbolic book suffered from the "index collapse" problem [36] (some tokens are not activated through out training) for sizes larger than 256, implying that 256 tokens are sufficient for exploring the acoustic units, whereas adding more will be of no benefit.

3.4. Interpretation of symbolic sequence

An advantage of the discrete representation learned by the VQ-VAE is the interpretability of individual tokens in the symbolic book. Here, a visualization method was developed to connect input acoustic features to the activated token. Figure 2 (left) shows histograms corresponding to phoneme classes (39-way). More specifically, noisy speech from the test set were passed through the symbolic encoder to obtain the symbolic sequences. Given the frame-wise phoneme labels, a histogram for individual phoneme class can then be plotted. Each bin represents the token index, and the value shows how many times this token was chosen, given the frame that belongs to the corresponding phoneme. The histograms were normalized to become probability distribution functions (PDFs), i.e., the summation equals 1. Here, it can be seen that phonemes with similar pronunciation also have similar distribution in the histograms. For example, the phonemes in each of the pairs (*aa*, *aw*), (*m*, *n*), and (*ch*, *sh*) have similar distributions, whereas phonemes in different pairs have different distributions.

For a complete understanding of the relations within the phoneme set, the Jensen-Shannon divergence between the

phonemes was measured. Figure 2 (right) shows a heat map. Each element represents the distance between two PDFs, and darker color corresponds to larger distance. As JS-divergence is symmetric, the heat map is also a symmetric matrix. Some squares in light color are located on the diagonal, which implies that phonemes with similar pronunciation are clustered together, e.g., vowels have lighter colors with each other, and are completely separated from fricatives. The heat map greatly facilitates the visualization of the relationship between phonemes. For instance, it shows that *ch* is very close to *s*, *z*, and *sh*. In conclusion, the symbolic encoder was demonstrated to be reactive to phonetic content. It was observed that some of the phonemes that are pronounced differently lie near each other. The obvious explanation is that the noise affected the input MFCCs, thus confusing the symbolic encoder. One possible solution is to constrain explicitly the symbolic encoder so that it may become noise-invariant by adding a discriminator and using adversarial training as in [37]. This is left as future work.

4. Conclusion and future work

A novel approach for incorporating phonetic content into a SE system was proposed, without the need for a recognition system or any transcriptions during training. The symbolic encoder used the vector quantization method proposed in VQ-VAE to extract discrete representations. Consequently, the symbolic encoder learned to divide the input MFCCs into acoustic units automatically, and achieved notable performance improvement compared to the baseline systems. The representations were further interpreted by visualizing the symbolic encoder behavior, and it was confirmed that it was phoneme-sensitive. In future studies, the effect of different noise types on the symbolic encoder will be investigated, and noise-invariant training will be performed to extract purer symbolic sequence. Furthermore, an explicit language model constraint based on the learned symbolics may be even more useful to the SE system.

5. References

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications.* Academic Press, 2015.

[3] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[4] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2007.

[5] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2017.

[6] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[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 Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[9] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] M. Kolbk, Z.-H. Tan, J. Jensen, M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[11] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[12] G. Borghini and V. Hazan, "Listening effort during sentence processing is increased for non-native listeners: a pupillometry study," *Frontiers in neuroscience*, vol. 12, p. 152, 2018.

[13] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] M. Mimura, S. Sakai, and T. Kawahara, "Deep autoencoders augmented with phone-class feature for reverberant speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4365–4369.

[15] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 146–150.

[16] S. E. Chazan, S. Gannot, and J. Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[17] S. E. Chazan, J. Goldberger, and S. Gannot, "Speech enhancement using a deep mixture of experts," *arXiv preprint arXiv:1703.09302*, 2017.

[18] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[19] J. Chorowski, R. J. Weiss, S. Bengio, and A. v. d. Oord, "Unsupervised speech representation learning using wavenet autoencoders," *arXiv preprint arXiv:1901.08810*, 2019.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[21] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[22] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[23] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[25] I.-T. Recommendation, "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*, 2001.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[28] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[29] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[32] G. Hu, "100 nonspeech environmental sounds, 2004."

[33] 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.

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

[35] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.

[36] Ł. Kaiser, A. Roy, A. Vaswani, N. Pamar, S. Bengio, J. Uszkoreit, and N. Shazeer, "Fast decoding in sequence models using discrete latent variables," *arXiv preprint arXiv:1803.03382*, 2018.

[37] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.