INCORPORATING REAL-WORLD NOISY SPEECH IN NEURAL-NETWORK-BASED SPEECH ENHANCEMENT SYSTEMS

*Yangyang Xia**

Carnegie Mellon University
Dept. of Electrical and Computer Engineering
Pittsburgh, PA 15213, USA
raymondxia@cmu.edu

Buye Xu, Anurag Kumar

Facebook Reality Labs Research
Redmond, WA 98052, USA
{xub, anuragkr}@fb.com

ABSTRACT

Supervised speech enhancement relies on parallel databases of degraded speech signals and their clean reference signals during training. This setting prohibits the use of real-world degraded speech data that may better represent the scenarios where such systems are used. In this paper, we explore methods that enable supervised speech enhancement systems to train on real-world degraded speech data. Specifically, we propose a semi-supervised approach for speech enhancement in which we first train a modified vector-quantized variational autoencoder that solves a source separation task. We then use this trained autoencoder to further train an enhancement network using real-world noisy speech data by computing a triplet-based unsupervised loss function. Experiments show promising results for incorporating real-world data in training speech enhancement systems.

***Index Terms*—** speech enhancement, self-supervised learning, real-world data, triplet loss

1. INTRODUCTION

Supervised single-channel speech enhancement has seen considerable improvement in the last few years, primarily due to the use of deep neural networks (DNNs) [1]. Training an effective speech enhancement (SE) system requires parallel databases of simulated degraded speech signals and their reference signals as the learning objective is often a function of the clean speech signals. The performance of SE systems trained on such artificially generated noisy speech inputs depends heavily on (a) the variety and amount of noise recordings available, and (b) if the simulated degradation is realistic. While these supervised SE systems have surpassed non data-driven approaches by a large margin [2], concerns around their generalization capabilities remain. Enabling SE systems to learn from real-world noisy speech can ensure that the networks are trained on real acoustical conditions rather than synthetic ones. Moreover, these data are readily available and can be obtained with relative ease. Lastly, such methods can also enable a system trained on simulated data to adapt to a new environment.

The primary challenge in incorporating real-world noisy speech for training SE systems is the lack of corresponding clean speech signals as training targets. A few recently proposed methods seek for alternative reference signals. Mixture-invariant training (MixIT) [3] attempted unsupervised speaker separation by forcing the network to separate mixture of mixtures. However, it can suffer from over-separation problem. Following MixIT, Noisy-target Training [4] treats real-world noisy speech data as reference and mixes them with noise signals to generate "more noisy" signals for training the SE system.

Another possibility to relax supervision is through the prediction or generation of pseudo ground truth. Although it is tempting to calculate the loss through a no-reference speech quality prediction network [5], experiments have shown that DNNs might over-optimize one perceptual metric without necessarily improving others [6, 7], let alone a prediction of them. Wang *et al.* used a pair of generative adversarial networks to map speech signals from noisy to clean [8]. The trained generator is then used to generate a pseudo reference signal. A similar setup was also proposed and studied by Xiang and Bao [9] with multiple learning objectives. These studies were inspired by unpaired image-to-image translation through cycle-consistency constraints [10]. However, in [8] the cycle-consistency constraint did not enforce clean speech embeddings and degraded speech embeddings to share the same latent space by using multiple encoders.

Generation of pseudo reference signals can also be done through a latent representation. In particular, methods based on self-supervised learning (SSL) frameworks can be used. In this framework, a speech signal is typically transformed to a latent space by an autoencoder. Then, SSL tasks are assigned in the latent space to establish correlations between a measure taken in this space and a physically meaningful measure taken in the signal domain. For example, the context encoder learns to generate content of a masked region in an image based on

*Work done during internship at FRL Research

its surrounding pixels [11].

In this paper, we propose two unsupervised loss functions for speech enhancement enabled by self-supervised learning. These unsupervised loss functions do not require the reference clean speech and allow us to incorporate real-world noisy speech in the training process. Our semi-supervised approach consists of two stages. The first supervised stage includes a novel modification to the vector-quantized variational autoencoder (VQ-VAE) that solves a source separation task using a corpus of *paired* data. In the second semi-supervised stage, the learned VQ-VAE is used to transform any given degraded speech signal to a pseudo noise ground truth and a pseudo speech ground truth, respectively. We then construct unsupervised losses based on a triplet formulation using these estimated ground truths. These losses are used to train an enhancement system along with the supervised losses from the paired data. Note that, the framework is designed in a semi-supervised setting with the assumption that some amount of *paired* data and potentially (much) more *unpaired* (real-world) data are available during training. The *unpaired* data can be real-world noisy speech recordings for which corresponding clean references are not available.

Organization of this paper. In Section 2, we provide some necessary background on supervised speech enhancement (SE) and VQ-VAE. We then describe our method in Section 3. Experimental setups are described in Section 4 and results are discussed in Section 5. Section 6 concludes our paper.

2. BACKGROUND

2.1. Supervised DNN-based speech enhancement

We assume that the observed degraded speech contains clean speech corrupted by additive noise. This relationship can be established in the short-time Fourier transform (STFT) domain as

$$X[t,k] = S[t,k] + N[t,k] \tag{1}$$

where $X[t,k]$, $S[t,k]$, and $N[t,k]$ represent the STFT at frame t and frequency index k of the degraded speech, clean speech, and noise, respectively. One common SE method is to train a DNN to predict a magnitude gain $G[t,k]$, so that the short-time Fourier transform magnitude (STFTM) of enhanced speech signal can be obtained by

$$\left|\hat{S}[t,k]\right| = G[t,k]\,|X[t,k]|\,. \tag{2}$$

Finally, the phase of the degraded signal is combined with the enhanced STFTM to reconstruct the enhanced speech signal through inverse STFT.

Conventionally, the paired sets $(X[t,k], S[t,k])$ are required during training. The supervised training involves a reconstruction loss,

$$L_s(\widehat{\vec{S}}, \vec{S}) = d(\widehat{\vec{S}}, \vec{S}), \tag{3}$$

where $\vec{S}$ and $\widehat{\vec{S}}$ denote the clean and enhanced STFTM in vector form, and $d(\cdot)$ is a distance measure such as the mean-squared error (MSE).

2.2. Encoder-Decoder in self-supervised learning

Self-supervised learning (SSL) methods usually construct tasks in a learned representation space. These tasks can be solved without requiring any labels for a given dataset. The assumption usually is that the representation learned by solving these pretext tasks will be useful for the downstream tasks. We follow the well-known encoder-decoder framework to learn such representations from speech signals. This autoencoding process can be described by

$$\vec{e} = \text{Encoder}(\vec{f}) \tag{4}$$

$$\widehat{\vec{f}} = \text{Decoder}(\vec{e}) \tag{5}$$

$$L_{\text{rec}}(\vec{f}, \widehat{\vec{f}}) = d(\vec{f}, \widehat{\vec{f}}), \tag{6}$$

where Encoder and Decoder are realized by DNNs and L_{rec} denotes a feature reconstruction loss function such as the MSE. Within this paradigm, SSL could impose an auxiliary task to the encoded features, the decoded features, or both. A generic representation of this process can be described by

$$L_{\text{latent}}(\vec{e}) = d(\vec{e}, \text{Transform}(\vec{e})) \tag{7}$$

$$L_{\text{feature}}(\vec{f}, \widehat{\vec{f}}) = d(\text{Transform}(\widehat{\vec{f}}), \text{Transform}(\vec{f})), \tag{8}$$

where each of L_{latent} and L_{feature} denotes the loss function of an auxiliary task. "Transform" refers to manipulation that provides distinctive goals to the auxiliary task. In context encoders [11], for example, partial occlusion is applied to the input image, forcing the encoder to learn features that would extrapolate the occluded pixel values.

It should be noted that the labels used in the pretext tasks are readily available in the original dataset and therefore the training targets in Eq. (7) and Eq. (8) shall not incur additional labeling effort. More specifically, we shall design the task in such a way that it does not require the clean reference speech for real-world degraded noisy speech. This task shall ultimately enable an unsupervised loss function,

$$L_u(\widehat{\vec{S}}) = d(\widehat{\vec{S}}, \text{Transform}(\widehat{\vec{S}})). \tag{9}$$

As opposed to Eq. (3), this loss function can be used to train an SE system on real-world degraded speech data. In the next section, we will describe a procedure that enables this process.

3. METHOD

Our approach consists of two training stages. The first stage consists of training a modified VQ-VAE that is constrained

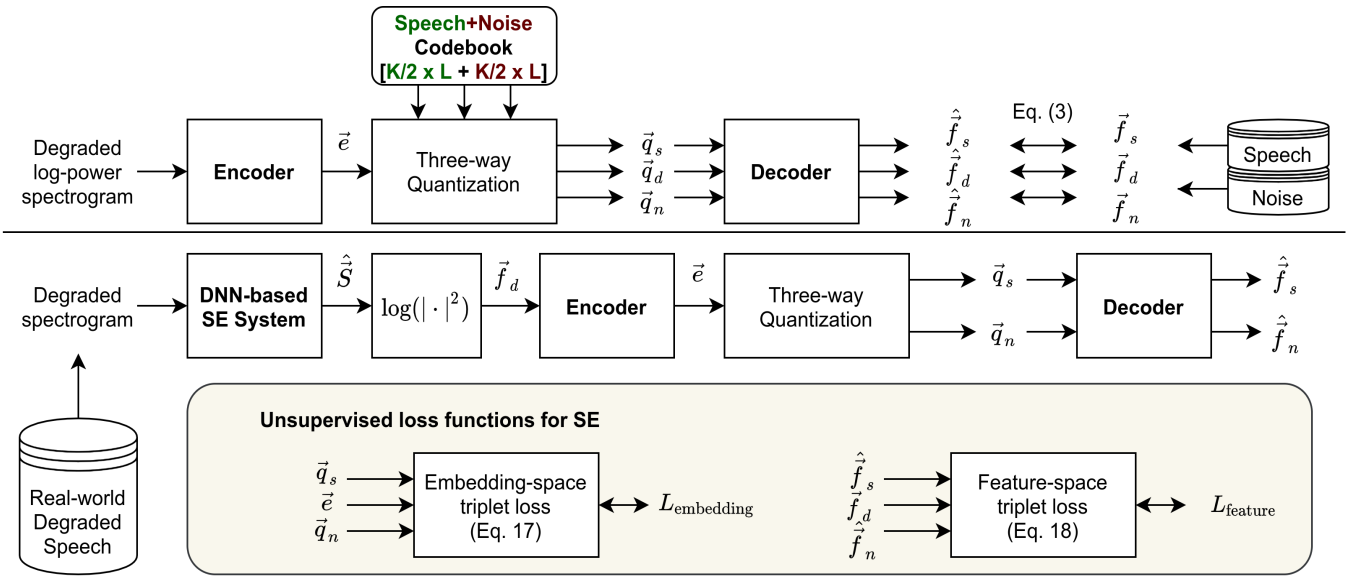

Fig. 1. Supervised training procedure of the modified VQ-VAE using *paired* data (*top*) and unsupervised training procedure of a speech enhancement system using *unpaired* data (*bottom*).

to separate speech and noise from the degraded speech signal in both latent and feature domains. Then, we describe two loss functions derived from this model that can be used to train any DNN-based SE systems in a semi-supervised manner. The unsupervised loss functions enable incorporation of real-world noisy speech during training.

3.1. Modified VQ-VAE for source separation

Compared to traditional autoencoders, the VQ-VAE has an additional quantization step in the latent space that avoids issues like high variance [12]. The whole process can be described by

$$\vec{e} = \text{Encoder}(\vec{f}) \tag{10}$$

$$\vec{q} = \text{VQ}(\vec{e}; \{\vec{c}_i\}) = \arg\min_{\vec{c}_i} d(\vec{e}, \vec{c}_i) \tag{11}$$

$$\widehat{\vec{f}} = \text{Decoder}(\vec{q}), \tag{12}$$

where $\{c_i\}, 1 \leq i \leq K$ is a set of K learnable vectors, and $d(\cdot)$ is a distance function. We define $\vec{f}$ to be the log-power spectra of degraded speech in vector form.

We design the VQ-VAE to do an acoustical source separation task with a codebook-lookup constraint. To achieve this goal, we partition the codebook $\{c_i\}$ in Eq. (11) into two equal halves,

$$C_s = \{\vec{c}_i\}, 1 \leq i \leq \frac{K}{2} \tag{13}$$

$$C_n = \{\vec{c}_i\}, \frac{K}{2} < i \leq K \tag{14}$$

$$C_d = C_s \cup C_n = \{\vec{c}_i\}, 1 \leq i \leq K. \tag{15}$$

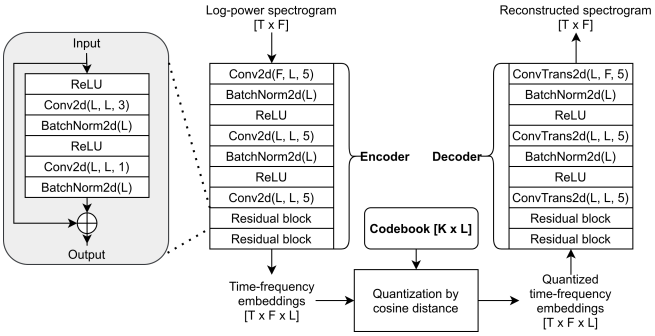

Fig. 2. Flow diagram of our VQ-VAE system. The order of parameters follows the PyTorch convention.

The quantization and decoding processes in Eq. (11) and Eq. (12) are then modified to produce three outputs,

$$\vec{q}_k = \text{VQ}(\vec{e}; C_k) = \arg\min_{C_k(i)} d(\vec{e}, C_k(i)) \tag{16}$$

$$\widehat{\vec{f}}_k = \text{Decoder}(\vec{q}_k) \tag{17}$$

where $k \in \{s, n, d\}$ denotes one of speech, noise, and degraded speech.

3.2. Encoder and decoder architectures

Both the encoder and decoder of our model are implemented using convolutional neural network (CNN). Specifically, the encoder consists of three two-dimensional convolutional layers, each followed by two-dimensional batch normalization and rectified linear unit (ReLU). The final convolutional layer

is followed by two residual blocks; a residual block is defined as two convolutional layers with an additive connection between the input and the output of the final layer [13]. The decoder's architecture is a mirror image of that of the encoder, with each convolutional layer replaced by a transposed convolutional layer.

The overall architecture is illustrated in Figure 2. For a T-by-F log-power spectrogram, each energy bin is transformed to a L-dimensional embedding by the encoder. The quantizer described in the previous step then transforms each embedding to its closest cluster center in terms of cosine distance. Finally, the decoder transforms quantized embeddings back to the log-power spectrogram.

3.3. Training procedure for modified VQ-VAE

The training procedure of the modified VQ-VAE is depicted in the top half of Figure 1. Log-power spectrograms of degraded speech signals pass through the VQ-VAE in order to obtain the embedding and reconstructed feature for each source. We train the VQ-VAE using the reconstruction loss in Eq. (3), the VQ loss, and the commitment loss as described in [12]. Note that each loss now consists of three components based on the degraded speech, clean speech, and noise, respectively. Although the mean-squared error (MSE) was originally used in [12] as the distance function $d(\cdot)$ in Eq. (11), we found that cosine distance made training more stable for this particular task. The VQ loss and the commitment loss are modified accordingly.

3.4. Unsupervised loss functions for enhancement

After training using *paired* data and supervised loss functions, the VQ-VAE is frozen while training a speech enhancement system. Specifically, a supervised SE system takes in a degraded speech signal and outputs $\widehat{\vec{S}}$, the STFTM of enhanced speech. It is then transformed to a log-power spectrogram and passed through the frozen VQ-VAE that outputs the continuous embedding, the quantized embeddings, and the decoded features in the process. We define the unsupervised embedding-space loss as

$$L_{\text{embedding}}(\vec{e};\Theta) = max(d(\vec{e},\vec{q}_s) - d(\vec{e},\vec{q}_n) + m, 0), \quad (18)$$

where m is a constant and $d(\cdot)$ is the cosine distance. Note that the continuous embedding $\vec{e}$ is used instead of the quantized embedding $\vec{q}_d$ as the latter is not differentiable. Similarly, the unsupervised feature-space loss can be derived from the decoded features,

$$L_{\text{feature}}(\vec{f}_d;\Theta) = max(d(\vec{f}_d,\widehat{\vec{f}}_s) - d(\vec{f}_d,\widehat{\vec{f}}_n) + m, 0), \quad (19)$$

where $\widehat{\vec{f}}_s$ and $\widehat{\vec{f}}_n$ are decoded from their corresponding quantized embeddings using Eq. (17). Note that both losses are calculated per time-frequency bin. We believe that the source separation task imposed on the VQ-VAE makes $\vec{q_s}$ and $\widehat{\vec{f}}_s$ pseudo-positive targets, and $\vec{q_n}$ and $\widehat{\vec{f}}_n$ pseudo-negative targets. We used the triplet margin [14] because neither target is ideal.

The bottom half of Figure 1 shows how to train a DNN-based SE system using real-world data. After obtaining the enhanced log-power spectrogram from the system, the frozen VQ-VAE is used to calculate the continuous embedding, the quantized embeddings, and the reconstructed features. The unsupervised embedding loss can be calculated by Eq. (18); the unsupervised feature loss can be calculated by Eq. (19). This loss is backpropagated to adapt the parameters of the SE system. If this system is also trained on paired data, the entire procedure is a semi-supervised training process.

In the next section, we will describe the experimental setup used to evaluate the effectiveness of these unsupervised losses for speech enhancement.

4. EXPERIMENTAL SETUP

4.1. Dataset

We used the clean speech of the Interspeech 2020 Deep Noise Suppression (DNS) Challenge dataset [15] and the ESC-50 dataset [16] for simulating *paired data* in all our experiments. The DNS training set contains a total of 500 hours of clean speech. The ESC-50 dataset contains 50 different types of environmental sounds (noises). In our experiments, we used fractions of these datasets to synthesize the *paired* data for training both the VQ-VAE and the supervised part of the SE system. The *real-world noisy speech* or the *unpaired data* was obtained from the Audioset dataset [17]. Audio recordings in Audioset tagged with "speech" class were further filtered by a sound event detector [18] to ensure that a large part of the recording contains speech along with other sounds. All audio recordings are sampled at 16k Hz. The average SNR of the filtered Audioset data estimated by the WADA algorithm [19] is around 10 dB.

4.2. Training procedure for VQ-VAE

To train our VQ-VAE, we randomly sampled 1-second speech segment from the DNS dataset and 1-second noise segment from the ESC-50 dataset, respectively. We then mixed the two signals at a SNR randomly sampled from the range $[-10, 30)$ dB. The mixed signals are scaled to provide a dynamic range of 40 dB. The resulting degraded signal was the input to the VQ-VAE.

4.3. Training and evaluation procedure for enhancement

We used stacked Gated Recurrent Units described in [20] as the baseline system for real-time speech enhancement in

our experiments. Similar to the training procedure for the VQ-VAE, we simulated degraded speech from speech signals in the DNS dataset and noise from the ESC-50 dataset. The degraded-clean *pairs* were used to train the SE system with the supervised loss function in Eq. (3). We consider three different conditions for training the enhancement system: (1) *Baseline*: the enhancement model is trained using only the paired data with supervised losses, (2) *Paired-Unsupervised*: the unsupervised loss functions (either Eq. (18) or Eq. (19)) are calculated from the *paired* data, and (3) *Unpaired-Unsupervised*: the unsupervised losses calculated from the real-world *unpaired* data in addition to the *paired* data. We summarize the setup of these systems in Table 1.

Table 1. System configurations

Method	Training Data	Unsupervised Loss Function
Baseline	*paired*	-
Paired-Embedding	*paired*	Eq. (18)
Paired-Feature	*paired*	Eq. (19)
Unpaired-Embedding	*paired* & *unpaired*	Eq. (18)
Unpaired-Feature	*paired* & *unpaired*	Eq. (19)

To evaluate the quality of enhanced speech signals, we used the perceptual evaluation of speech quality (PESQ) [21] and scale-invariant signal-to-distortion ratio (SI-SDR) [22] metrics.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1. Effect of the amount of *paired* data

We present the absolute improvement of all SE systems under the *seen* noise condition as a function of the amount of *paired* training data in Figure 3. With the minimum *paired* data (10%), the unsupervised losses based training were not able to improve over the supervised baseline; in fact, many performed noticeably worse than the baseline. As the amount of *paired* data increased to 20% of DNS speech and ESC-50 noise, all unsupervised loss functions were able to largely improve and surpassed the baseline performance. This indicates that a decent amount of *paired* data is necessary for making the VQ-VAE learn a reliable representation of speech and noise. Finally, as more amount of *paired* data was presented in training, the significance of unsupervised losses goes down. This suggests that the supervised loss function eventually outweighs the unsupervised losses.

5.2. Generalization to unseen noise types

We present the absolute improvement of all SE systems under the *unseen* noise condition as a function of the amount of *paired* training data in Figure 4. We note the similar trend

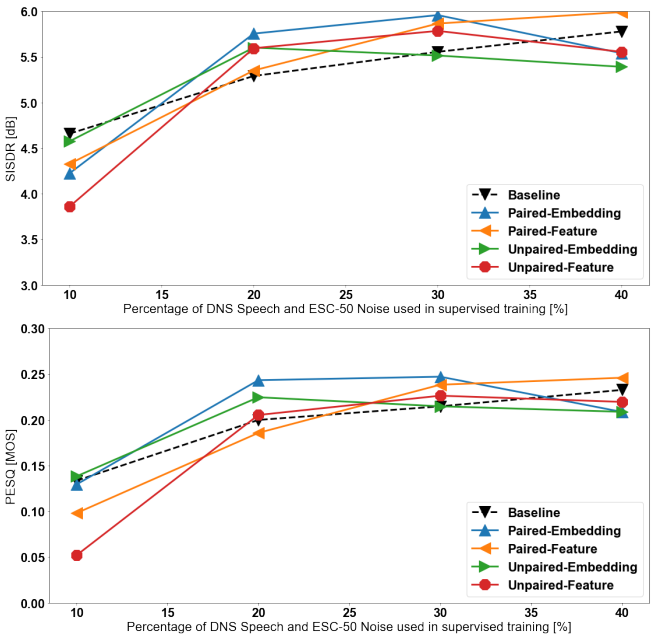

Fig. 3. Absolute SI-SDR improvement (*top*) and PESQ improvement (*bottom*) averaged across all evaluation conditions for *seen* noise types during training. The averaged SNR and PESQ of unprocessed speech are 0 dB and 1.39 MOS, respectively.

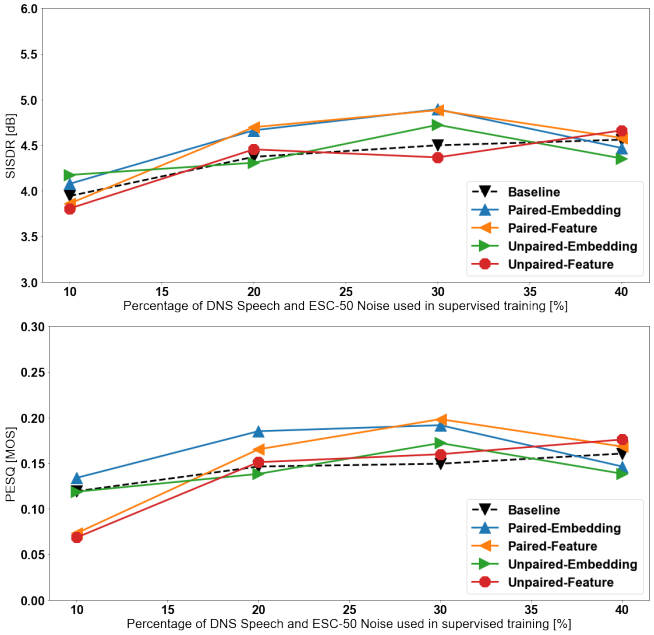

Fig. 4. Absolute SI-SDR improvement (*top*) and PESQ improvement (*bottom*) averaged across all evaluation conditions for *unseen* noise types during training. The averaged SNR and PESQ of unprocessed speech are 0 dB and 1.41 MOS, respectively.

Table 2. Evaluation of speech enhancement systems trained on 20% supervised data: *seen* (*unseen*) noise conditions

Method	SNR									
	-10 dB		-5 dB		0 dB		5 dB		10 dB	
	PESQ	SI-SDR	PESQ	SI-SDR	PESQ	SI-SDR	PESQ	SI-SDR	PESQ	SI-SDR
Degraded	1.16 (1.09)	-9.98 (-10.0)	1.17 (1.15)	-4.99 (-5.02)	1.31 (1.30)	0.01 (-0.01)	1.53 (1.56)	5.00 (4.99)	1.80 (1.95)	10.0 (10.0)
Baseline	1.29 (1.22)	-3.13 (-2.85)	1.41 (1.27)	1.46 (0.576)	1.55 (1.44)	5.77 (4.25)	1.73 (1.74)	9.61 (8.15)	1.98 (2.11)	12.8 (11.6)
Paired-Embedding	**1.31** **(1.24)**	**-2.44** (-2.44)	**1.45** **(1.32)**	**2.16** **(1.12)**	**1.60** **(1.48)**	**6.32** (4.61)	**1.78** **(1.78)**	**9.93** (8.28)	**2.05** **(2.16)**	12.9 (11.7)
Paired-Feature	1.28 **(1.24)**	-2.69 **(-2.33)**	1.41 (1.29)	1.66 (1.08)	1.53 (1.47)	5.75 **(4.65)**	1.71 (1.76)	9.51 **(8.32)**	1.96 (2.11)	12.6 (11.7)
Unpaired-Embedding	1.30 (1.20)	-2.64 (-2.84)	1.43 (1.27)	1.90 (0.530)	1.57 (1.43)	6.05 (4.19)	1.76 (1.72)	9.82 (7.98)	2.02 (2.12)	12.9 (11.6)
Unpaired-Feature	**1.31** (1.21)	-2.56 (-2.88)	1.43 (1.27)	1.83 (0.543)	1.56 (1.45)	6.05 (4.37)	1.73 (1.75)	9.73 (8.25)	1.97 (2.13)	**13.0** **(11.9)**

as observed in the results for the *seen* noise conditions: unsupervised loss functions require a certain amount of supervised training to benefit the system. The overall performance compared to the *seen* noise condition is generally worse and improves slower as the amount of *paired* data increased. This phenomenon is generally true for supervised SE systems. At 30% of supervised data, however, we observe similar improvement to the *seen* noise condition by including the unsupervised losses calculated from the *paired* data. This indicates that the unsupervised losses calculated from the *paired* data is generalizable to unseen noise conditions.

5.3. Effect of unsupervised loss functions

As Figure 3 and Figure 4 revealed that 20% of supervised data is the minimum from our setting that the SE systems start benefitting from unsupervised loss functions, we present the detailed evaluation across noise conditions in Table 5. Results show that the paired-embedding loss is the best across most SNR conditions. The paired-feature loss is slightly more superior under some unseen noise types. The unpaired loss functions had more impact at higher SNRs. We believe that this could be because the filtered Audioset has relatively high SNR.

5.4. Learned embedding margin

To verify that the source separation task imposed on the VQ-VAE was effective, we present the averaged triplet margin calculated on the validation set in Figure 5. As defined in Eq. (18), the margin should be high when global SNR is low, and the margin should be low when global SNR is high. As Figure 5 shows, using 10% and 20% supervised data were not enough to learn the correct relationship. While using 30% supervised data worked, using 40% data made a more drastic improvement. This shows that the more training data the better the VQ-VAE learns the SSL tasks, which in turn would improve the quality of unsupervised losses for SE.

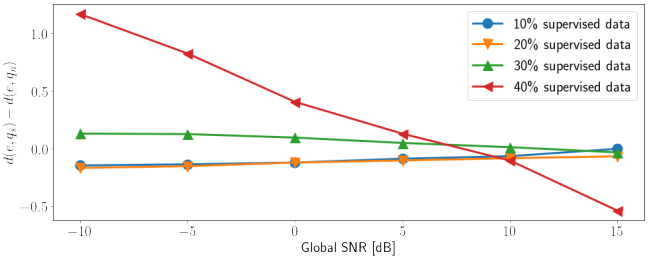

Fig. 5. Averaged triplet margin on the validation set as a function of global SNR. Lines with more negative slopes correspond to better learned representation. The margin was calculated on the validation set in three-dimensional latent space. The standard deviations from smallest amount of data to largest amount of data were 0.55, 0.28, 0.31, and 1.53, respectively.

6. CONCLUSIONS

In this paper, we introduced two novel unsupervised loss functions for speech enhancement that were enabled by a modified vector-quantized variational autoencoder and a self-supervised learning task. We showed that the loss functions calculated on supervised data were able to improve supervised speech enhancement systems when the amount of training data is small. We also showed that the loss functions calculated on real-world noisy speech data were able to improve the supervised SE systems in some noise conditions. In the future, we plan on fine-tuning the VQ-VAE on enhanced speech data. We will also explore sampling techniques of real-world data to better match the evaluation condition.

7. REFERENCES

[1] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[2] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, pp. 1256–1266, 2019.

[3] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J. Weiss, Kevin Wilson, and John R. Hershey, "Unsupervised sound separation using mixture invariant training," in *NeurIPS*, 2020.

[4] Takuya Fujimura, Yuma Koizumi, Kohei Yatabe, and Ryoichi Miyazaki, "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech," *arXiv preprint arXiv:2101.08625*, 2021.

[5] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 331–335.

[6] Juan Manuel Martin-Donas, Angel Manuel Gomez, Jose A Gonzalez, and Antonio M Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[7] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5074–5078.

[8] Yu-Che Wang, Shrikant Venkataramani, and Paris Smaragdis, "Self-supervised learning for speech enhancement," *arXiv preprint arXiv:2006.10388*, 2020.

[9] Yang Xiang and Changchun Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1826–1838, 2020.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE ICCV*, 2017, pp. 2223–2232.

[11] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[12] Aaron van den Oord, O Vinyals, and K Kavukcuoglu, "Neural discrete representation learning," in *31st International Conference on Neural Information Processing Systems*, 2017, p. 6309–6318.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] Chandan KA Reddy, E Beyrami, H Dubey, V Gopal, R Cheng, R Cutler, S Matusevych, R Aichner, A Aazami, S Braun, et al., "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," *arXiv preprint arXiv:2001.08662*, 2020.

[16] Karol J Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[17] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R Channing Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[18] Anurag Kumar and Vamsi Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5447–5457.

[19] C. Kim and R M Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[20] Yangyang Xia, Sebastian Braun, Chandan KA Reddy, Harishchandra Dubey, Ross Cutler, and Ivan Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 871–875.

[21] 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 *IEEE ICASSP*, 2001, vol. 2, pp. 749–752.

[22] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.