A CLOSER LOOK AT WAV2VEC2 EMBEDDINGS FOR ON-DEVICE SINGLE-CHANNEL SPEECH ENHANCEMENT

Ravi Shankar[1], *Ke Tan*[2], *Buye Xu*[2], *Anurag Kumar*[2]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, USA
[2]Meta Reality Labs, Redmond, Washington, USA

ABSTRACT

Self-supervised learned models have been found to be very effective for certain speech tasks such as automatic speech recognition, speaker identification, keyword spotting and others. While the features are undeniably useful in speech recognition and associated tasks, their utility in speech enhancement systems is yet to be firmly established, and perhaps not properly understood. In this paper, we investigate the uses of SSL representations for single-channel speech enhancement in challenging conditions and find that they add very little value for the enhancement task. Our constraints are designed around on-device real-time speech enhancement – model is causal, the compute footprint is small. Additionally, we focus on low SNR conditions where such models struggle to provide good enhancement. In order to systematically examine how SSL representations impact performance of such enhancement models, we propose a variety of techniques to utilize these embeddings which include different forms of knowledge-distillation and pre-training.

Index Terms: Speech Enhancement, Wav2Vec2, GCRN, Pre-training, Knowledge Distillation, Conditioning

1. INTRODUCTION

Speech enhancement (SE) is a fundamental problems in the domain of speech signal processing. Broadly speaking, its goal is to enhance the quality (naturalness) and intelligibility of any given speech signal with or without making apriori assumptions about the noise or other distortions in the signal. SE systems have multiple applications such as noise suppression in phone calls, better communication in noisy environment, and in designing more robust hearing aids [1].

Speech enhancement is a very challenging problem to solve, mainly due to the blind nature of the noise (in the statistical sense), non-stationarity of the signal and, the duality in type of signal corruption, i.e., additive "vs" convolutional. Having said that, significant improvements have been made in recent times in separating noise-like component from speech using supervised machine learning methods. Common techniques for speech enhancement formulates it as a discriminative task where the goal might be to learn a mask or directly predict the clean speech [2]. This can be done in time-domain as well as time-frequency (TF) domain.

To this end, multiple novel neural network architectures have been proposed such as [3, 4, 5, 6, 7, 8, 9, 10, 11]. Generative modeling via either score matching or denoising diffusion methods have also been proposed to synthesize clean speech from noisy inputs [12, 13, 14, 15]. Beyond supervised training of deep neural networks, some recent works have explored semi and self-supervised approaches [16, 17, 18, 19, 20, 21].

Among the model architectures proposed in the above works, gated convolutional recurrent network (GCRN) proposed in [7], is a special type of hybrid convolutional-recurrent model for enhancement in time-frequency domain. This model consists of a gated convolutions [22] followed by grouped long short-term memory (LSTM) layers to reduce parameters in a U-net [23] style architecture. The architecture can be conveniently used to formalize speech enhancement using various targets (mask, magnitude spectrogram, complex spectrogram) and loss functions to obtain state-of-the-art speech enhancement system. Moreover, the encoder-decoder design of GCRN makes it suitable for our study. Hence, we use a GCRN architecture based model in this study.

Recent years have also seen development of a variety of works on speech representation learning and it's application to different tasks. The goal of representation learning is to extract meaningful features from a signal (audio/video/text) in a self-supervised/unsupervised manner. These learned representations are helpful for downstream tasks. Some of the most popular models for speech representation learning are Wav2Vec2 [24], HuBERT [25] and WavLM [26]. With similar underlying neural architectures, they are trained in slightly different ways. The key objective of all these models is to capture the broad phonetic-linguistic structure in the speech.

While SSL models have been found useful in ASR tasks, only a few works, [27, 28, 29] have focused on explicitly using self-supervised learned features for speech enhancement. One rationale behind using SSL representations for enhancement can be to inject phonetic information which have been found to be useful for enhancement [30]. [27, 28, 29] use these SSL embeddings as inputs to train the enhancement model, either alone or in concatenation with short-time Fourier transform (STFT) representations. The authors

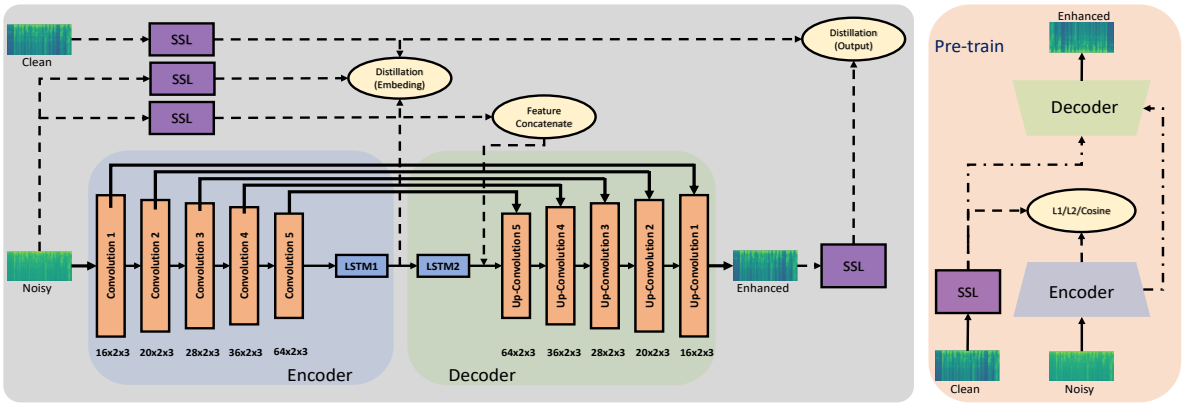

Fig. 1. GCRN model and different modes of using SSL embeddings to guide enhancement model. The right panel (red) shows our pre-training modes. The left panel (green) shows knowledge distillation modes and uses of SSL embeddings as inputs.

in [29] use SSL embeddings to supervise and regularize their enhancement network. In above works, the improvements from SSL models are fairly limited, and they fail to provide clear explanation for their results.

Our goal in this paper is to systematically investigate different ways of using SSL embeddings to improve an SE system. More specifically, we focus on on-device and real-time processing which constrains how SSL embeddings can be used. Such SE systems are expected to be *(a)* causal - no future look ahead, and, *(b)* of low compute footprint. The GCRN neural architecture can be used to design and develop an SE models satisfying these characteristics. However, they may suffer from unsatisfactory performances in low-SNR conditions [31]. Hence, the key question we study is - **Can SSL embeddings improve on-device SE systems in low-SNR conditions?**. In particular, we study the popular wav2vec2.0 SSL model and attempt to utilize it to improve a GCRN based on-device SE model.

The conditions outlined above constrain how any SSL model can be used for enhancement. SSL models are usually very large, non-causal and hence fine-tuning them [32] is not a possible path for using them in our case. More generally, feeding SSL embeddings as inputs (or any form of conditioning) to the enhancement model will break causality as well as compute constraints as the SSL model will be needed during inference. Hence, the core idea which motivates the solution space here is that *any approach to use SSL models for enhancement should not change the behavior of the underlying SE model during inference*, that is if the original SE model is causal and small, uses of SSL models to improve them should not change these characteristics.

Keeping the above motivation in mind, we propose and analyze different ways in which an SSL model can be used to improve an enhancement model. Our approaches include those based on using SSL models as teachers for knowledge distillation as well as for pre-training of enhancement models. Along with comprehensive quantative analysis we also try to bring an understanding of Wav2vec2 structure. We show that it is difficult to transfer the structure and information captured by Wav2Vec2 model to small enhancement models.

2. METHOD

In this section, we describe different approaches for using SSL model for enhancement. The input to each of these models is the spectrogram representation of speech signal extracted using a window of length 25ms with a 20ms stride to achieve the downsampling factor of 320. Note that, this choice is made for feature extraction to be consistent with the Wav2Vec2 model. Note: *In our experiments, we use the Wav2vec2 model as SSL model. Hence, SSL and Wav2Vec2 embeddings are used interchangeably in this paper.*

2.1. Overview

An overview of our framework is shown in Fig. 1. Our base enhancement model consists of a GCRN (details in Sec 2.2) model. We propose 3 approaches to employ the SSL model to improve our enhancement model. (a) *Feature Concatenate:* (Sec. 2.5) In this case the SSL embeddings are used to condition the decoder. Clearly, providing SSL embeddings as input to the GCRN model would make the overall inference non-causal and of large compute, breaking the constraint outlined earlier. Since this method does not satisfy the requirements it is a baseline for comparison. (b) *Knowledge Distillation:* (Sec. 2.5/2.6) Employing a teacher-student framework we propose a variety of ways to distill knowledge from the SSL model to the enhancement model. (c) *Pre-training:* Lastly, we use the SSL model to pre-train the enhancement model.

2.2. Baseline Enhancement Model

The baseline enhancement framework used in this paper is a causal GCRN model which learns a complex spectral mapping from noisy speech to clean speech [7]. The GCRN (Fig.1) consists of a stack of down-sampling convolutional layers followed by 2 layers of uni-directional LSTMs. The output of these LSTMs are then up-sampled

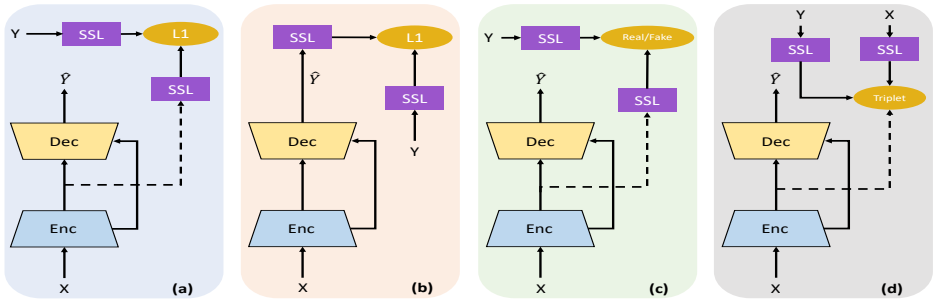

Fig. 2. Different ways of knowledge distillation from Wav2Vec2 embeddings used in this paper. (a) Sample-wise distillation via encoder output, (b) distillation via enhanced signal (c) adversarial distillation, and (d) triplet loss based distillation.

by a set of transposed convolutions and added to the residue from down-sampling layers for generating final output. The causal structure of GCRN facilitates streaming capability for continuous on-the-fly operation. Furthermore, the recurrent operation in GCRN is performed group-wise (along feature dimension) to reduce the number of trainable parameters to < 4M resulting in a memory footprint of only 16 megabytes.

Mathematically, we denote the noisy speech as $\mathbf{X}$, clean speech as $\mathbf{Y}$ and the enhancement model as f_e. The training objective is to maximize $\mathcal{L} = SISDR(f_e(\mathbf{X}), \mathbf{Y})$ with respect to parameters of f_e. Scale-invariant signal-to-distortion ratio (SISDR) [33] is defined via the following equation:

$$SISDR(f_e(\mathbf{X}), \mathbf{Y}) = 10 \log_{10} \frac{\|\alpha \mathbf{Y}\|^2}{\|\alpha \mathbf{Y} - f_e(\mathbf{X})\|^2}, \quad (1)$$

where $\alpha = \frac{f_e(\mathbf{X})^T \mathbf{Y}}{\|\mathbf{Y}\|^2}$.

Left panel in Fig. 1 shows the three approaches by which SSL embeddings are used to guide the SE model. The details are described in Sec. 2.4 and Sec. 2.5.

2.3. Wav2Vec-2 Embeddings

Wav2Vec2 [24] is one of the most popular self-supervised learned speech model. The model consists of a stack of convolutional layers to exploit short-time stationarity of speech followed by a stack of multi-head attention layers. A vector quantization layer discretize the convolutional feature space. By masking the output representation at random time-frames, the network is trained via a contrastive loss to maximize its similarity with the corresponding code-book vector from quantization layer. A maximum entropy criterion is added to the loss function to encourage equal uses of each code-book.

In this paper, we use Wav2Vec2 embeddings in two different ways: *(a)* by using the last transformer layer output, and *(b)* by convex combination of multi-layered outputs where weights are estimated ad-hoc. The rationale behind using multiple layer outputs hinges on the observations made by [34, 35] that, intermediate features can store important para-linguistic information for speech reconstruction.

2.4. Wav2Vec2 Embeddings as Input

The simplest approach for using SSL embeddings is providing them as an extra input to the enhancement model. This conditioning is done via concatenating Wav2Vec2 features with noisy speech spectrogram which are then fed to the SE model. Prior works [27] have shown that concatenation of SSL after the bottleneck LSTM layers work better as it provides useful phonetic guidance right before the generative upsampling operation begins. Hence, we adopt the same strategy in our experiments. The concatenated features (bottleneck features + SSL) are passed via a linear projection layer to maintain the dimensionality for residual skip connections. We further investigate two different strategies for concatenating the embeddings: (a) using the SSL embedding from the last layer of Wav2Vec2 model and (b) using weighted combination of embeddings obtained from each self-attention layer calculated adaptively. Let g_s be the SSL model then, we minimize $\mathcal{L}_E = -SISDR(f_e[\mathbf{X}, g_s(\mathbf{X})], \mathbf{Y})$, where $g_s(\mathbf{X})$ is the feature extracted from Wav2Vec2 model.

We refer to this approach as *feature concatenation* since the SSL embeddings are concatenated with LSTM outputs. Note that, in this case, *Wav2Vec2 embeddings would be required during inference and making the overall system large and non-causal*.

2.5. Distillation to SE Embeddings

2.5.1. Distillation via L1 Loss

In this approach, the Wav2Vec2 representations of the target clean speech are directly used to inject knowledge in the SE model. It forces the outputs of the SE encoders to have the same semantic information as SSL embeddings of the clean speech (Fig. 2(a)). Note that, this technique uses the Wav2Vec2 embeddings extracted from the ground-truth signal itself. Therefore, it can act as a stronger prior than the naive concatenation because self-supervised models use clean speech to learn their parameters. The outputs of the LSTM layers which are decoded to produce the clean speech are expected to be enriched through the SSL representations.

We employ a linear projection layer to match dimensionality of bottleneck features with the SSL embeddings which are learned via backpropagation. Similar to the concatenate mode, we use two forms of SSL embeddings, i.e., output of the last layer, and weighted combination of each layer of

Wav2Vec2. The overall loss function for training in this case is:

$$\mathcal{L}_E = -SISDR(f_e(\mathbf{X}), \mathbf{Y}) \; + \; \lambda \times \|g_s(\mathbf{Y}) - f_e^{enc}(\mathbf{X})\| \quad (2)$$

2.5.2. Distillation via Adversarial Loss

The previous distillation approach use sample-wise similarity constraint that puts a strong prior on the enhancement network. We propose another way of doing knowledge distillation via distribution matching enforced through objectives such as KL-divergence penalty. We experiment with the distribution level matching of the Wav2Vec2 embeddings with GCRN encoder via an adversarial loss term (Fig. 2(c)). We use a block-wise convolutional discriminator proposed in [36] with a period of 1 to consider each frame of the embeddings. Mathematically, denoting the discriminator network by $\mathcal{D}$, the enhancement and discriminator objective are written as:

$$\begin{aligned} \mathcal{L}_E &= -SISDR(f_e(\mathbf{X}), \mathbf{Y}) \; + \; \lambda \times \|\mathcal{D}(f_e^{enc}(\mathbf{X})) - 1\|_2^2 \\ \text{and, } \mathcal{L}_D &= \frac{1}{2}\|\mathcal{D}(f_e^{enc}(\mathbf{X})) - 0\|_2^2 \; + \; \frac{1}{2}\|\mathcal{D}(g_s(\mathbf{Y}) - 1\|_2^2 \end{aligned} \quad (3)$$

This is the least-square formulation of GAN [37] which has been shown to minimize the Pearson χ^2 divergence between the Wav2Vec2 embeddings and GCRN encoder output.

2.5.3. Distillation via Triplet Loss

Triplet loss is yet another way to enforce a similarity between the GCRN encoder embeddings and Wav2Vec2 representations in the manifold space. Authors in [16] proposed triplet loss in the context of unsupervised training of speech enhancement. The underlying idea in triplet loss is to maximize the margin (up to a certain threshold) between the set of GCRN embeddings and embeddings obtained for clean and noisy speech from Wav2Vec2 (Fig. 2(d)). Note that, this is a stronger penalty than contrastive loss which only forces dissimilarity among different representations. The objective is:

$$\begin{aligned} \mathcal{L}_E &= -SISDR(f_e(\mathbf{X}), \mathbf{Y}) \; + \lambda \times \mathcal{L}_T \\ \text{where, } \mathcal{L}_T &= \max(\|\mathbf{a} - \mathbf{p}\| - \|\mathbf{a} - \mathbf{n}\| + m, 0) \end{aligned} \quad (4)$$

Here, $\mathbf{a} = f_e^{enc}(\mathbf{X})$ represents the GCRN encoder representations (also called anchor), $\mathbf{p} = g_s(\mathbf{Y})$ is the Wav2Vec2 embeddings from clean speech and $\mathbf{n} = g_s(\mathbf{X})$ is the same from noisy speech. The margin m is set to 100 in this task to account for high dimensionality of the embeddings.

2.6. Distillation to SE Outputs

Another distillation approach is to enforce similarity in the latent space of the enhanced and the clean speech. We do this by adding an extra loss term to the objective function which forces similarity between Wav2Vec2 representations of the enhanced signal and the ground-truth speech (Fig. 2(b)). The underlying hypothesis is same as the previous method, i.e., the enhanced speech should have the same phonetic-linguistic content as the clean speech, thereby, improving its intelligibility. This method does not require additional linear layer, but it does require gradient backpropagation through the SSL model during training. The overall loss function is given by:

$$\mathcal{L}_E = -SISDR(f_e(\mathbf{X}), \mathbf{Y}) \; + \; \lambda \times \|g_s(\mathbf{Y}) - g_s(f_e(\mathbf{X}))\| \quad (5)$$

2.7. Pre-training via Wav2Vec2 Embeddings

SSL is primarily used as unsupervised pre-training strategy, models are then fine-tuned on different downstream tasks. However, our goal here is to use the Wav2Vec2 model to improve the given SE network under consideration. To achieve this, we propose to pre-train the GCRN enhancement model using SSL embeddings. The general motivation is that pre-training provide better initialization of the model parameters than random. This approach is illustrated in right panel of Fig. 1. GCRN facilitates a straightforward decomposition of the model architecture into an encoder and a decoder. The LSTM blocks in the bottleneck layers are split equally among the encoder and decoder components. We pre-train both the encoder and the decoder.

Encoder Pre-training: For pre-training of the encoder, we provide noisy spectrogram as input and predict the Wav2Vec2 embeddings as output. This corresponds to knowledge distillation from the large scale SSL model to the encoder of the SE network.

Decoder Pre-training: We pre-train the decoder by formulating the task as a speech generation problem. To achieve this, the decoder is trained by making it predict the clean spectrogram conditioned on the ground truth SSL embeddings. We also experiment with the decoder training conditioned on the encoder outputs rather than ground truth embeddings. In the former case, the residual connection based on up-sampling operation is replaced by locally duplicating the learned features by a factor of 2.

2.8. Training Details

We train the enhancement model with complex STFT features extracted from the noisy and clean speech pair. We use the Adam optimizer with a fixed learning rate of 0.001 for 4 million steps with a batch size of 200. SI-SDR loss between the ground truth and predicted signal is used for training.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We use the DNS challenge [38] corpus in our experiments. The clean speech and the noise samples are mixed at random SNRs between -5dB and 5dB for training and validation set. The testing set consists of 500 samples of clean speech from Librispeech test-clean mixed with noise (from the test sets) at a fixed -5dB SNR to simulate challenging real-world scenario.

3.2. Baseline, Feature Concatenation and Distillation

We first compare different approaches outlined in Section 2. Table. 1 summarizes the result of this experiment. We can see that weighted sum concatenation (*-ws) and distillation via output works best on all three metrics, i.e. PESQ, STOI and SI-SDR. The differences however, are very small in practice. The last column of Table 1 shows whether the model complies with the constraint set in the beginning of this paper namely, causality and memory footprint. As expected, distillation via adversarial loss performs poorly in comparison to other modes because distribution level prior is weak. Triplet loss however, performs similar to sample-wise modes since it enforces similar constraint as the SE embedding distillation with a margin component.

Model	PESQ	STOI	SI-SDR	Constr.
Base	1.59	0.84	9.1	✓
Feature Concat	1.55	0.83	8.9	✗
Feature Concat-ws	1.60	0.84	9.3	✗
Distillation Embed.	1.52	0.83	9.1	✓
Distillation Embed.-ws	1.56	0.83	9.2	✓
Distillation Output	1.60	0.84	9.3	✓
Distillation Adversarial	1.52	0.82	7.5	✓
Distillation Adversarial-ws	1.55	0.81	8.4	✓
Distillation Triplet	1.56	0.81	8.5	✓
Distillation Triplet-ws	1.57	0.83	8.9	✓

Table 1. Baseline GCRN model and different techniques considered towards using Wav2Vec2 embeddings for enhancement. **Noisy PESQ: 1.11, STOI: 0.69, SI-SDR: -4.99dB**

Overall, we can conclude that the Wav2Vec2 embeddings do not provide any significant improvement over the vanilla model under on-device requirement scenario. We hypothesize the main reasons why it fails to do so as follows: first, concatenation adds an overhead of extra parameters in the model which may change the loss landscape of the parameterized function making it harder to find any better local optima. Second, distillation via embedding act by adding a penalty on the latent space representation of noisy speech. This might be a weak signal because the Wav2Vec2 embeddings retain the qualitative aspects of speech in trace amounts [34].

3.3. Pre-training with SSL

We use the Wav2Vec2 embedding extracted from clean speech to train the GCRN encoder. The goal is to learn the latent representation of clean speech. We also pre-train the decoder from (a) encoder's output and (b) SSL embeddings directly to generate clean speech. In the latter case, the upsampling operation is performed locally instead of using skip connections from encoder. We also experiment with different types of encoder losses namely, L1, L2 and Cosine metric and pick only the best one for final fine-tuning. The selection is based on which model performs best for enhancement.

Encoder Loss/Decoder Input	PESQ	STOI	WER %
L1/Frozen	1.24	0.74	71.2
L1/Wav2Vec2	1.25	0.83	7.4
L2/Frozen	1.22	0.73	76.7
L2/Wav2Vec2	1.24	0.81	16.5
Cosine/Frozen	1.21	0.72	79.4
Cosine/Wav2Vec2	1.29	0.82	12.5

Table 2. Pre-training of speech enhancement model using SSL. **Noisy PESQ: 1.11, STOI: 0.69, SI-SDR: -4.99dB**

Table. 2 summarizes the result of enhancement task performed directly using the pre-trained models. Note that, we did not train this GCRN on any enhancement task - just encoder and decoder training as outlined before. Even though we do not train the model for noise removal, *the model manages to do some form of enhancement*. This shows that a pre-training based on prediction of SSL representations can on it's own lead to some denoising capabilities.

Further, we can see that when we train the decoder directly from SSL embeddings, the intelligibility of generated speech is higher. In fact, we obtain a word error rate of $< 20\%$ upon decoding the reconstructed speech using a pre-trained CRDNN model from Speechbrain library [39]. Note that, the CRDNN model is trained on Librispeech 960-h corpus itself. Nevertheless, this shows that SSL captures the phonetic linguistic information but completely ignores other aspects of speech such as tonality, loudness and voice quality.

Finally, we pick the model trained with L1 loss on encoder for fine-tuning on the enhancement task. As we can see

Encoder Loss / Decoder Input	PESQ	STOI	SI-SDR
L1/Frozen	1.53	0.83	8.60
L1/Wav2Vec2	1.54	0.84	8.90

Table 3. Speech enhancement assessment from fine-tuned models. **Noisy PESQ: 1.11, STOI: 0.69, SI-SDR: -4.99dB**

from (Table. 3), pre-training provides no significant advantage over the base model for the speech enhancement task. In fact, the model performance worsens slightly compared to the GCRN model (see Table. 1) trained from scratch. We conjecture that this happens due to two main reasons: (a), the Wav2Vec2 embeddings only capture the information required for reconstruction of smoothed quantized features and need further fine-tuning for generative tasks such as enhancement. The evidence for this hypothesis is provided by Table. 2. We can observe that the STOI scores are relatively high, meaning we get intelligible speech of poor quality when generated directly from Wav2Vec2 embeddings. This is in contrast to

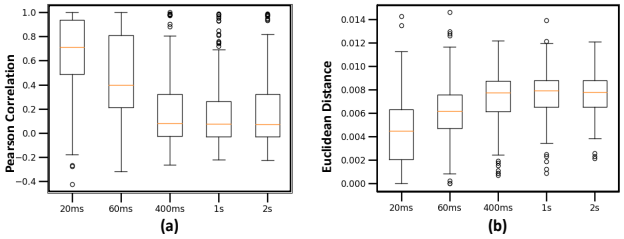

Fig. 3. **Wav2Vec2**: Box plot of (a) correlations and (b) Euclidean distances obtained from frames of Wav2Vec2 embeddings separated by 20ms, 60ms, 400ms, 1sec and 2sec.

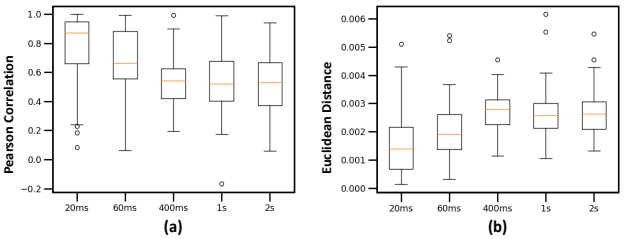

Fig. 4. **Distilled model**: Box plot of (a) correlations and (b) Euclidean distances obtained from frames of Wav2Vec2 embeddings separated by 20ms, 60ms, 400ms, 1sec and 2sec.

the enhancement task, where the representation learnt by encoder should encode both, the style and speaker information in conjunction with the phonetic/linguistic component of the speech utterance. Second, it is challenging to distill knowledge from large SSL models due to the inherent structure of embeddings themselves. We discuss this phenomenon in the next subsection.

3.4. Structure of Wav2Vec2 embeddings

We analyze the features from Wav2Vec2 for a variety of utterances and show the interesting correlation patterns in these embeddings. Fig. 3 shows the box plot of correlations and L2 norm between features separated by 20ms, 60ms, 400ms, 1sec and 2sec, respectively. We can see that the features are highly correlated up until 60ms (typical phoneme length), and are similar in magnitude (Euclidean distance) throughout an utterance leading to a conclusion that the phonetic/linguistic content is stored in the small magnitude variations between frames. The main implication of this pattern is that capturing such small differences is extremely difficult. Therefore, it forces the GCRN encoder to learn an average noise-like representation for each utterance in pre-training.

3.5. Knowledge Distillation from SSL

In this experiment, we probe into the question of finding out whether knowledge distillation from models such as Wav2Vec2 is possible or not. The GCRN encoder learns an average representation, but it could be due to the low complexity of the model itself. Therefore, we train a convolution-transformer stack identical to the Wav2Vec2 architecture

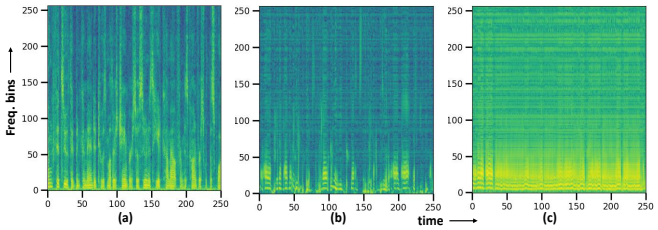

Fig. 5. Illustration of spectrograms: (a) ground truth speech (b) speech decoded using Wav2Vec2 embeddings and (c) speech decoded using embeddings extracted from the trained knowledge distillation model (same encoder as Wav2Vec2).

itself, to predict the SSL embeddings from speech. We use a mix of L1 and cosine loss in this regard. Fig. 4 shows the correlation and Euclidean distance pattern of embeddings obtained from this new distilled model. Note that, it does not exhibit the same characteristics as the original embeddings from Fig. 3. The change in correlation coefficient and normalized L2 distance is completely different from the pre-trained Wav2Vec2 model. Furthermore, Fig. 5 shows an example generation from the GCRN decoder when prompted with original Wav2Vec2 embeddings (Fig. 5(b)) and the embeddings extracted from the distilled encoder (Fig. 5(c)). We can see that the original Wav2Vec2 embeddings allow the reconstruction of energy in the higher frequency bands whereas, the distilled model completely loses that information. The main reason for this behavior is the inability of even an expressive model to capture the intricate details present in the original embeddings. Therefore, we conclude that it is fairly challenging to distill knowledge from Wav2Vec2 model even if the student model follows the exact same architecture.

4. DISCUSSION AND CONCLUSIONS

In this paper, we have explored different mechanisms to leverage Wav2Vec2 representation for the task of speech enhancement. We showed that under on-device constraints and low-SNR conditions, SSL models add little to no value in improving the base enhancement model. We hypothesized that the SSL embeddings retain only the phonetic/linguistic component of speech and ignores the qualitative aspects of the signal. Our experiments with the pre-training of GCRN model using SSL embeddings further confirmed this hypothesis. We showed that GCRN decoder is capable of generating intelligible speech from Wav2Vec2 embeddings, but it lacks in quality as demonstrated by the poor PESQ and SI-SDR scores (Table 2). In addition, we showed that the structure of these embeddings makes it difficult to pre-train the GCRN encoder. These features are difficult to reproduce even with a more expressive model, due to the phonetic details encoded in tiny variations across time. These subtle variations make it difficult for speech enhancement models to extract any meaningful information. Therefore, the Wav2Vec2 features require more refined understanding to use them for speech enhancement in challenging scenarios.

5. REFERENCES

[1] Arata Kawamura, Weerawut Thanhikam, and Youji Iiguni, "Single channel speech enhancement techniques in spectral domain," *ISRN Mechanical Engineering*, vol. 2012, 07 2012.

[2] 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)*, 2016, pp. 5220–5224.

[3] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[4] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.

[5] Xiaohuai Le, Hongsheng Chen, Kai Chen, and Jing Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," *arXiv preprint arXiv:2107.05429*, 2021.

[6] Chiheb Trabelsi, Olexa Bilaniuk, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher Joseph Pal, "Deep complex networks," *ArXiv*, vol. abs/1705.09792, 2017.

[7] Ke Tan and DeLiang Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[8] Hyeong-Seok Choi, Janghyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2019.

[9] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[10] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis R. Bach, "Music source separation in the waveform domain," *CoRR*, vol. abs/1911.13254, 2019.

[11] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.

[12] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," 2023.

[13] Joan Serrà, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini, "Universal speech enhancement with score-based diffusion," 2022.

[14] Bryce Irvin, Marko Stamenovic, Mikolaj Kegler, and Li-Chia Yang, "Self-supervised learning for speech enhancement through synthesis," 2022.

[15] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7402–7406.

[16] Yangyang Xia, Buye Xu, and Anurag Kumar, "Incorporating real-world noisy speech in neural-network-based speech enhancement systems," *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 564–570, 2021.

[17] Efthymios Tzinis, Yossi Adi, Vamsi K Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[18] 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.

[19] Takuya Fujimura, Yuma Koizumi, Kohei Yatabe, and Ryoichi Miyazaki, "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 436–440.

[20] Ying Cheng, Mengyu He, Jiashuo Yu, and Rui Feng, "Improving multimodal speech enhancement by incorporating self-supervised and curriculum learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4285–4289.

[21] Ryandhimas E Zezario, Tassadaq Hussain, Xugang Lu, Hsin-Min Wang, and Yu Tsao, "Self-supervised denoising autoencoder with linear regression decoder for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6669–6673.

[22] Yann Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*, 2016.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015, pp. 234–241, Springer International Publishing.

[24] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[25] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[26] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[27] Kuo-Hsuan Hung, Szu wei Fu, Huan-Hsin Tseng, Hsin-Tien Chiang, Yu Tsao, and Chii-Wann Lin, "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 186–190.

[28] Zili Huang, Shinji Watanabe, Shu wen Yang, Leibny Paola García-Perera, and Sanjeev Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6837–6841, 2022.

[29] Ori Tal, Moshe Mandel, Felix Kreuk, and Yossi Adi, "A systematic comparison of phonetic aware techniques for speech enhancement," in *Interspeech*, 2022.

[30] Yen-Ju Lu, Chien-Feng Liao, Xugang Lu, Jeih-weih Hung, and Yu Tsao, "Incorporating broad phonetic information for speech enhancement," 2020.

[31] Yaser Yurtcan and Banu Günel Kılıç, "Speech recognition on mobile devices in noisy environments," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 1–4.

[32] William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe, "Reducing barriers to self-supervised learning: Hubert pre-training with academic compute," *arXiv preprint arXiv:2306.06672*, 2023.

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

[34] Yohan Lim, Namhyeong Kim, Seung Yun, Sanghun Kim, and Seung-Ik Lee, "A preliminary study on wav2vec 2.0 embeddings for text-to-speech," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 2021, pp. 343–347.

[35] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Hwan Lee, Hoon Heo, and Kyogu Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, Eds., 2021.

[36] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020, NIPS'20, Curran Associates Inc.

[37] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang, "Multi-class generative adversarial networks with the L2 loss function," *CoRR*, vol. abs/1611.04076, 2016.

[38] Chandan K A Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," 2021.

[39] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, and et.al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.