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 tasks such as automatic speech recognition, speaker identification, and others. However, their utility in speech enhancement systems is yet to be firmly established, and perhaps slightly misunderstood. In this paper, we investigate the uses of SSL representations for single-channel speech enhancement in challenging conditions and establish the impact they can have on the enhancement task. Our constraints are designed around on-device real-time speech enhancement – model being causal, and the compute footprint being small. Additionally, we focus on low SNR conditions where such models struggle to provide good performance.

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 processing. Its goal is to enhance the quality (naturalness) and intelligibility of any given speech signal with or without making apriori assumptions about the noise. SE systems have multiple applications in real time communications, such as noise suppression in phone calls, in designing more robust hearing aids [1], to mention a couple.

While it is a challenging problem, significant improvements have been made recently in separating noisy component from speech using supervised machine learning. These techniques formulate it as a discriminative task where the goal is to learn a mask to be applied to the noisy speech [2] or directly estimate the clean speech [3]. To this end, multiple novel neural network architectures have been proposed such as [4, 5, 6, 7, 3]. Generative modeling via diffusion methods have been suggested by authors in [8, 9, 10] to synthesize clean speech from noisy inputs by conditioning the process on noisy speech. Beyond supervised training, some recent works have also explored semi and self-supervised approaches [11, 12, 13, 14, 15, 16] for speech enhancement.

Moreover, there has been a surge of research in representation learning and its application in speech processing. Self-supervised learning is aimed at learning such representations without any human labels. Prominent models for speech representation learning include Wav2Vec2 [17], HuBERT [18], and WavLM [19]. The primary objective here is to capture the phonetic and linguistic structure embedded within input speech signals. Previous works [20, 21, 22] have proposed leveraging these features for speech enhancement. Notably, incorporating additional phonetic information can be beneficial for enhancing speech quality [23]. Furthermore, in [22], SSL embeddings are employed to supervise and regularize the enhancement network. However, these enhancements have shown some limitations, and the experiments conducted have yet to offer comprehensive insights into their effectiveness.

In this paper, we set out to systematically investigate different ways of using SSL embeddings in order to improve an SE system. We focus on on-device and real-time processing which constrains how SSL embeddings can be used. Such SE systems are expected to *(a)* be causal - no future look ahead, and, *(b)* have low compute footprint. However, they may provide satisfactory performances in high-SNR conditions [24]. Therefore, the key question we study is - **Can SSL embeddings improve on-device SE systems in low-SNR conditions?** In particular, we focus on using pre-trained Wav2Vec2 network to improve GCRN based SE model.

Our proposed approaches are based on using SSL networks as teachers for knowledge distillation, as well as, for pre-training of the enhancement model. Along with comprehensive quantitative analysis we also bring an nuanced understanding of the SSL embeddings. We show that it is non-trivial to transfer the structure and information captured by SSL models (such as Wav2Vec2) to small student models.

2. METHOD

In this section, we describe different approaches for using SSL model for enhancement. The input to these models is the spectrogram representation of speech signal extracted using a window of length 25ms and a stride of 20ms to achieve the downsampling factor of 320. We make this choice to be consistent with the Wav2Vec2 model. Note: *Our experiments use the Wav2Vec2 model as SSL model, hence SSL and Wav2Vec2 embeddings are used interchangeably throughout this paper*.

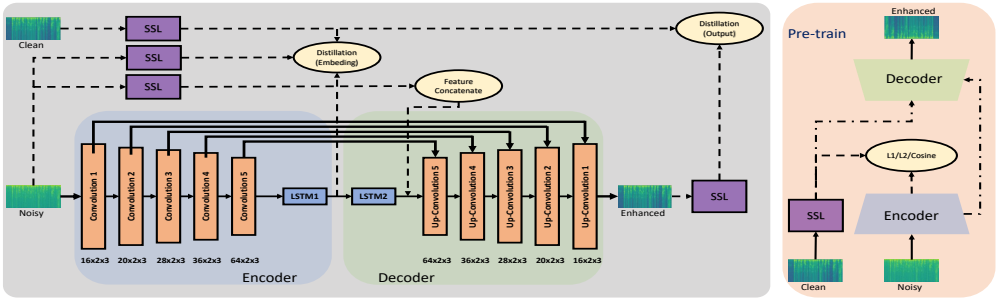

Fig. 1. GCRN model and different modes of using SSL embeddings to guide enhancement model. The right panel (red) shows our pre-training modes. The left panel (gray) shows knowledge distillation modes and uses of SSL embeddings as inputs.

2.1. Overview

Fig. 1 shows the overview of our framework. The base enhancement model is a causal GCRN network which learns a complex spectral mapping from noisy to clean speech [4]. GCRN (Fig.1) consists of a stack of down-sampling convolutional layers followed by uni-directional LSTMs. The output of LSTMs are up-sampled by a set of transposed convolutions to generate the final output. The causal structure of GCRN facilitates streaming capability. Further, the recurrent operation in GCRN is performed group-wise (along feature dimension) to reduce the total number of trainable parameters to < 4M. We analyze 3 approaches to employ SSL for improving the base model. (a) *Concatenation:* (Sec. 2.2) In this case the SSL embeddings are used to condition the decoder. Clearly, this makes the overall inference non-causal and extremely computational, breaking the required constraint. (b) *Knowledge Distillation:* (Sec. 2.3/2.4) Using a teacher-student framework we study a variety of ways to distill knowledge from the SSL model. (c) *Pre-training:* Lastly, we use the SSL model to pre-train the GCRN encoder and decoder.

Mathematically, denoting the noisy speech by $\mathbf{X}$, clean speech by $\mathbf{Y}$ and the GCRN model as f_e. The training objective of enhancement is to maximize $\mathcal{L} = SISDR(f_e(\mathbf{X}), \mathbf{Y})$ with respect to the parameters of f_e. Scale-invariant signal-to-distortion ratio (SI-SDR) [25] is defined as follows:

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

where $\alpha = \frac{f_e(\mathbf{X})^T \mathbf{Y}}{\|\mathbf{Y}\|^2}$ is the scaling factor of clean speech Y.

Finally, [26] showed that, intermediate features can store important para-linguistic information for speech reconstruction. Therefore, 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 (weighted sum).

2.2. Wav2Vec2 Embeddings as Input

In this regime, we provide SSL embedding as an extra input to the GCRN network. Prior works [20] have shown that concatenation of SSL after the bottleneck layers work better in providing phonetic guidance. The concatenated features (bottleneck + SSL) are passed via a projection layer to maintain the dimensionality for up-sampling. 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.

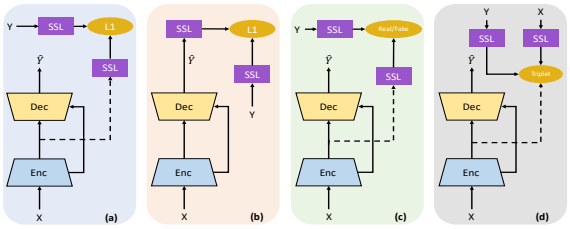

Fig. 2. Different ways of knowledge distillation from Wav2Vec2 used in this paper. (a) Sample-wise distillation via encoder output, (b) distillation via enhanced signal (c) adversarial distillation, and (d) triplet loss based distillation.

2.3. Distillation to SE Embeddings

2.3.1. Distillation via L1 Loss

In this approach, we force the output of the SE encoder to have the same semantic information as the SSL embedding of clean speech (Fig. 2(a)). We hypothesize that the output of LSTM layers in GCRN should retain language-specific knowledge for reconstruction. Note that, this technique uses the Wav2Vec2 embeddings extracted from the clean signal. The loss function in this mode can be expressed as:

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

2.3.2. Distillation via Adversarial Loss

Another way of knowledge distillation is via distribution matching enforced through objectives such as KL-divergence. We experiment with the distribution level matching of the Wav2Vec2 embeddings with GCRN encoder via adversarial loss (Fig. 2(c)) using a block-wise convolutional discriminator from [27]. Denoting the discriminator network by $\mathcal{D}$, the enhancement and discriminator objectives are:

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

2.3.3. Distillation via Triplet Loss

Triplet loss enforces a similarity between the GCRN embeddings and Wav2Vec2 representations in the latent manifold. Authors in [11] proposed triplet loss for unsupervised training of speech enhancement. The goal is to maximize the margin between the GCRN embeddings and SSL embeddings from clean and noisy speech (Fig. 2(d)). The triplet objective is:

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

Here, $\mathbf{a} = f_e^{enc}(\mathbf{X})$ is the GCRN encoder representations, $\mathbf{p} = g_s(\mathbf{Y})$ is the Wav2Vec2 embeddings from clean speech and $\mathbf{n} = g_s(\mathbf{X})$ is the same from noisy speech. We set the margin m to 100, accounting for embedding dimensions.

2.4. Distillation to SE Outputs

Finally, we enforce similarity in the latent space of enhanced and clean speech by adding a loss term between Wav2Vec2 representations of the enhanced signal and the ground-truth speech (Fig. 2(b)). This method does not require additional linear layer, but it backpropagates through the SSL model during training. The overall loss function is given by:

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

2.5. Pre-training via Wav2Vec2 Embeddings

Pre-training is a popular technique to initialize model parameters from a related task followed by fine-tuning it on a target task to overcome data scarcity. The hypothesis is that pre-training might provide better initialization of model parameters than random (illustration in Fig. 1's right panel).

Encoder Pre-training: 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 GCRN network.

Decoder Pre-training: We pre-train the decoder to 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.

The enhancement model is trained with complex STFT features extracted from the noisy and clean speech pair. We use the Adam optimizer having a fixed learning rate of 1e-3 for 4 million steps and a batch size of 200.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We use DNS challenge [28] 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 operation 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. The last column of Table 1 shows the model compliance with the constraint set. We can see that distillation via output works best on all three metrics, i.e. PESQ, STOI and SI-SDR **while meeting the constraints**. The differences however, are very small in practice. The base model performs relatively well and is on par with the best model. As expected, distillation via other modes perform poorly as the clean speech and enhanced speech manifolds may not have overlap.

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 do not observe any huge improvements over the baseline. The main reason is due to distillation providing a weak feedback signal, since Wav2Vec2 embeddings contain qualitative aspects of speech in trace amounts [29].

3.3. Pre-training with SSL

We use the Wav2Vec2 embeddings of 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 to generate ground truth. We experiment with different encoder losses namely, L1, L2 and Cosine, and pick the best one for fine-tuning.

Table. 2 summarizes the result of enhancement task performed directly using the pre-trained models. Note that, we did not fine-tune this GCRN, yet *the model manages to do some form of enhancement*. 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 Speechbrain [30]. Therefore, SSL captures the

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**

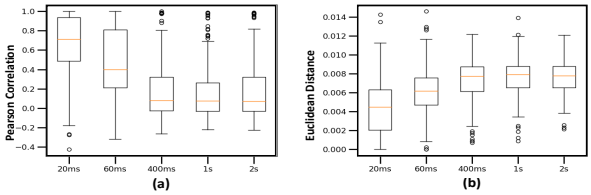

Fig. 3. **Wav2Vec2**: Box plot of (a) correlations and (b) Euclidean distances obtained from frames of Wav2Vec2 embeddings separated by 20ms, 60ms, 400ms, 1sec and 2sec.

phonetic 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. Table. 3 shows

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**

that pre-training does not help in speech enhancement task. In fact, the model performance worsens slightly compared to the baseline (see Table. 1). We conjecture that this happens due to two main reasons: first, the Wav2Vec2 embeddings only capture the information required for reconstruction of smoothed quantized features. Second, it is challenging to distill knowledge from large SSL models due to structure of embeddings themselves. We discuss this phenomenon in 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 only until 60ms (typical phoneme length), but are similar in magnitude (Euclidean distance) throughout an utterance. It means that phonetic content is stored in the small magnitude variations between frames. Capturing such

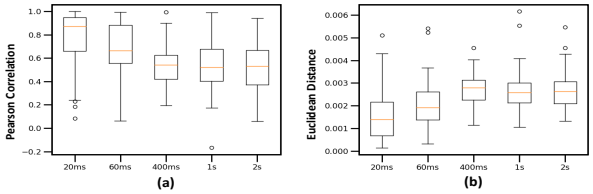

Fig. 4. **Distilled model**: Box plot of (a) correlations and (b) Euclidean distances obtained from frames of Wav2Vec2 embeddings separated by 20ms, 60ms, 400ms, 1sec and 2sec.

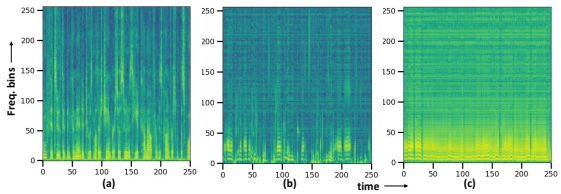

Fig. 5. Illustration of spectrograms: (a) ground truth speech (b) speech decoded using Wav2Vec2 embeddings and (c) speech decoded using embeddings extracted from the trained knowledge distillation model (same encoder as Wav2Vec2).

small differences is extremely difficult. Therefore, the GCRN encoder learns an average representation in pre-training.

3.5. Knowledge Distillation from SSL

We finally probe into the question of whether knowledge distillation from Wav2Vec2 is possible or not. We train a convolution-transformer stack identical to the Wav2Vec2 model to predict the embeddings using a mix of L1 and cosine loss. Fig. 4 shows the correlation and Euclidean distance pattern of embeddings obtained from the new model. Note that, it exhibits similar characteristics as the original embeddings from Fig. 3. However, the dip in correlation and increase in L2 distance is lower than the pre-trained. Fig. 5 shows speech 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.

4. CONCLUSION

We have explored different mechanisms to leverage Wav2Vec2 representation for speech enhancement. We showed that under causal, on-device and low-SNR constraints, SSL model adds very little value in improving the baseline model. We hypothesized that the SSL embeddings retain only the phonetic/linguistic component of speech and ignore the qualitative aspects which was confirmed by the experiments. In addition, we showed that the structure of SSL embeddings makes it difficult to pre-train a small encoder. These features are difficult to reproduce even with an expressive model, due to the phonetic details encoded in tiny variations across time.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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