



SelfRemaster: Self-Supervised Speech Restoration with Analysis-by-Synthesis Approach Using Channel Modeling

Takaaki Saeki, Shinnosuke Takamichi, Tomohiko Nakamura, Naoko Tanji, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo, Japan.

{takaaki.saeki, shinnosuke.takamichi}@ipc.i.u-tokyo.ac.jp

Abstract

We present a self-supervised speech restoration method without paired speech corpora. Because the previous general speech restoration method uses artificial paired data created by applying various distortions to high-quality speech corpora, it cannot sufficiently represent acoustic distortions of real data, limiting the applicability. Our model consists of analysis, synthesis, and channel modules that simulate the recording process of degraded speech and is trained with real degraded speech data in a self-supervised manner. The analysis module extracts distortionless speech features and distortion features from degraded speech, while the synthesis module synthesizes the restored speech waveform, and the channel module adds distortions to the speech waveform. Our model also enables audio effect transfer, in which only acoustic distortions are extracted from degraded speech and added to arbitrary high-quality audio. Experimental evaluations with both simulated and real data show that our method achieves significantly higher-quality speech restoration than the previous supervised method, suggesting its applicability to real degraded speech materials.

Index Terms: Speech restoration, self-supervised learning, neural vocoder, historical speech

1. Introduction

Speech data used in modern speech processing research is usually recorded with sophisticated digital recording equipment and in a well-designed environment. However, most existing audio data contains acoustic distortions and have different data properties from high-quality audio data. To expand the range of applications of speech processing, it is necessary to make use of speech data in varying qualities, e.g., speech material recorded with poor or outdated audio devices [1, 2] and endangered languages' voices [3–5]. However, restoring or analyzing such degraded audio is challenging because paired high-quality and degraded speech data is not generally available. The previous general speech restoration method [6] relies on artificial paired corpora for supervised learning. It cannot sufficiently represent acoustic distortions of real data, which limits the applicability.

We propose a self-supervised speech restoration method without paired speech data. Our proposed model consists of *analysis*, *synthesis*, and *channel* modules based on the recording process of degraded speech. The analysis module extracts *time-variant* distortionless speech features and *time-invariant* acoustic distortion features from degraded speech (hereafter, “channel feature”). The other modules carry out the generation process of distorted speech; the synthesis module synthesizes distortionless speech (i.e., restored speech), and the channel module adds acoustic distortion. The model is trained in a self-supervised manner by minimizing the reconstruction loss between degraded input and reconstructed speech. We also propose a dual-learning method for stable self-supervised learning,

where a backward process defines distortionless speech feature loss with arbitrary high-quality speech corpora while the forward process calculates the reconstruction loss. Experimental evaluations show that our method achieves significantly higher-quality speech restoration than the previous supervised method. Our implementation¹ and audio samples² are publicly available. The contributions of this paper are as follows:

- We propose self-supervised speech restoration that can learn various acoustic distortions without a paired speech corpora and is applicable to real degraded speech data.
- Our method achieved significantly higher-quality speech restoration than the previous supervised method on both simulated data and real historical materials.
- Our model enables audio effect transfer, in which only channel features are extracted from degraded speech and applied to other high-quality speech.

2. Related work

Speech restoration. Many previous studies have focused on specific speech restoration tasks such as bandwidth extension [7, 8], dereverberation [9, 10], denoising [11, 12], and de-clipping [13, 14]. Similar to our work, some previous studies [15, 16] jointly perform multiple speech restoration tasks, and some methods leverage waveform synthesis [6, 17] to perform speech restoration. Unlike these studies, our work does not focus on removing additive noises and enables self-training. A previous study on general speech restoration with a neural vocoder [6] is the closest work to ours. The most significant difference is that our method does not require paired data, so it can be trained with real audio materials in a self-supervised manner instead of artificially creating degraded speech data. Furthermore, as our proposed method also extracts channel features at the same time, it can transfer audio effects, which adds the desired channel distortion to arbitrary audio.

Self-supervised learning for speech. Self-supervised speech representation learning [18–20] and methods using those models for various downstream tasks [21, 22] have widely been studied. There are several studies related to our work which involve self-supervised learning on speech enhancement [23–25], though they only focus on denoising while our method restores speech from various kinds of distorted speech. Similar to our method, the DDSP autoencoder [26] learns disentangled features from acoustic signal in a self-supervised manner and manipulates each feature. It uses a simple sinusoidal vocoder and assumes only reverberation as acoustic distortion, while our method leverages a more expressive waveform synthesis model to achieve high-quality speech restoration and channel modules to capture various acoustic distortions.

¹https://github.com/Takaaki-Saeki/ssl_speech_restoration

²https://takaaki-saeki.github.io/ssl_remaster_demo/

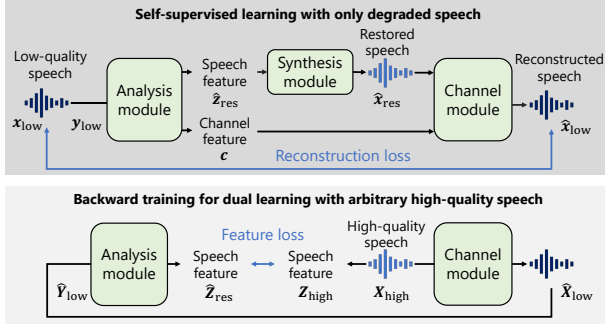


Figure 1: *Proposed self-supervised speech restoration method, consisting of analysis, synthesis, and channel modules that simulate recording process of degraded speech. Basic training process (top) minimizes reconstruction loss between degraded and reconstructed speech. We also propose dual learning (top & bottom) with arbitrary high-quality speech corpora.*

3. Proposed method

3.1. Basic framework

We begin with an overview of the recording process of degraded speech underlying our method. High-quality speech (i.e., distortionless speech) is emitted from the mouth through a human speech production process, which can be parameterized by *time-variant* speech features such as traditional source-filter vocoder features or mel spectrograms. Then acoustic distortions (e.g., non-linear response of recording equipment and lossy audio coding) are added to the high-quality speech, resulting in a final recorded audio. We assume the distortions (i.e., channel features) to be *time-invariant*, where the recording equipment or audio coding does not change in each audio sample.

The proposed speech restoration model consists of analysis, synthesis, and channel modules based on the abovementioned process, as shown in Fig. 1 (top). All of these modules are composed of neural networks, so they can be trained in an end-to-end manner. We denote the degraded-speech waveform and the degraded-speech features as x_{low} and y_{low} , respectively. The analysis module estimates the speech features of the restored speech, \hat{z}_{res} , from y_{low} . At the same time, the intermediate features of the analysis module are input to other layers in order to finally estimate the time-invariant channel features c . Note that we use mel spectrogram for y_{low} and the two features described in Section 4.1 for \hat{z}_{res} . The synthesis module simulates human speech production; it takes \hat{z}_{res} as input and estimates the speech waveform of the restored speech \hat{x}_{res} . We use HiFi-GAN [27] for the synthesis module, where it is trained on an arbitrary high-quality speech corpus and the model parameters are frozen. The channel module simulates the acoustic distortion; it is conditioned by the channel feature c and distorts the restored speech \hat{x}_{res} to estimate the degraded speech \hat{x}_{low} . The model is trained to minimize the frame-level reconstruction loss, which can be defined as a multi-scale spectral loss [26] as follows:

$$\mathcal{L}_{recons} = \sum_i \{ \|s_i - \hat{s}_i\|_1 + \alpha |\log s_i - \log \hat{s}_i| \}, \quad (1)$$

where s_i and \hat{s}_i are the amplitude spectrograms of x_{low} and \hat{x}_{low} , respectively. The subscript i is the window length of the Fourier transform and α is a weight of the log term, where we used $i = (2048, 1024, 512, 256, 128, 64)$ and $\alpha = 1.0$ in our experiment. Speech restoration is performed by driving the trained analysis and synthesis modules. The analysis module

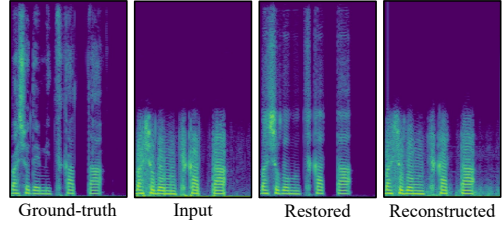


Figure 2: *Spectrograms obtained with proposed method*

receives the acoustic feature of degraded speech, y_{low} , from the analysis module, and the synthesis module synthesizes the restored-speech waveform \hat{x}_{res} .

Furthermore, the trained model can also perform an “audio effect transfer” in addition to speech restoration. The audio effect is transferred without the synthesis module. We first extract the channel feature c from the degraded speech x_{low} using the trained analysis module. The trained channel module conditioned by c receives arbitrary high-quality audio and distorts it so that the resulting audio sounds distorted like the referred degraded speech.

Fig. 2 shows the spectrograms obtained by the proposed method. The input low-quality speech is simulated on the basis of the quantization and resampling described in Section 4, and is missing high-frequency bands that are present in the ground-truth high-quality speech and has distorted low-frequency bands. In the restored speech, the missing and distorted bands are restored. Furthermore, the reconstructed speech output by the channel module faithfully reproduces the input speech.

Unfortunately, the above basic self-supervised training process is unstable because each module has a high expressive power, and the modules are only trained with the reconstruction loss between the degraded and reconstructed waveform. Specifically, the analysis module can represent the effect of the channel module, so there is no guarantee that the analysis module outputs the speech features of high-quality speech. Therefore, we propose the dual-learning method described in Section 3.2. In Section 3.3, we also introduce supervised pre-training to improve the performance when only a very limited amount of data is available for the model size, such as when we use rare historical audio.

3.2. Dual learning for stable self-supervised learning

Fig. 1 shows the proposed dual-learning method. In addition to the basic training framework described in Section 3.1, we introduce a training task that propagates information in the backward direction. We denote the two training tasks in the forward direction and backward direction as $\mathcal{T}_{forward}$ and $\mathcal{T}_{backward}$, respectively. $\mathcal{T}_{forward}$ is the basic learning framework described in Section 3.1; degraded speech is sent to the analysis module to reconstruct the degraded waveform. Meanwhile, in $\mathcal{T}_{backward}$, high-quality speech is input to the channel module to estimate the features of high-quality speech. Our method is analogous to the forward and backward learning between two domains in machine translation [28].

Here, we use arbitrary high-quality speech, which does not align with the degraded speech and only consists of other speaker’s utterances. The high-quality speech waveform X_{high} is first passed through the channel module to generate the degraded speech waveform \hat{X}_{low} with channel features. Next, the speech features of the degraded speech, \hat{Y}_{low} , are input to the analysis module to estimate the speech feature of the re-

stored speech, \hat{Z}_{res} . Then a feature loss is defined as the mean squared error between \hat{Z}_{res} and the speech feature Z_{high} is obtained from the high-quality speech as:

$$\mathcal{L}_{\text{feature}} = \|Z_{\text{high}} - \hat{Z}_{\text{res}}\|_2. \quad (2)$$

The final loss function is given as the weighted sum of the loss function between the reconstruction loss and the feature loss as:

$$\mathcal{L} = (1 - \beta) * \mathcal{L}_{\text{recons}} + \beta * \mathcal{L}_{\text{feature}}, \quad (3)$$

where β is a weight of the feature loss. Here, the gradient due to feature loss is not propagated to the channel module, and the channel module is only trained on the reconstruction loss. Intuitively, $\mathcal{T}_{\text{forward}}$ and $\mathcal{T}_{\text{backward}}$ can be interpreted as:

- $\mathcal{T}_{\text{forward}}$ obtains the analysis module that adapts to degraded speech and the channel module that expresses the acoustic distortion.
- $\mathcal{T}_{\text{backward}}$ obtains the analysis module to estimate high-quality speech features.

This dual learning method enables self-supervised learning of the analysis module and channel module without paired data.

3.3. Supervised pretraining for low-resource settings

Data scarcity is an obstacle in restoring historical audio materials; only a small amount of audio is available due to the limited amount of recording storage size available in the past. Thus, we introduce a supervised pre-training method to obtain the initial weights for such low data resource conditions as:

1. We supervisedly train the analysis and channel modules using paired data artificially created from arbitrary high-quality speech data.
2. With the initialization from supervised pre-training, we then carry out the self-supervised learning described in Section 3.1 and Section 3.2 with real degraded speech data.

In the supervised pretraining, we generate pseudo-degraded speech waveform $X_{\text{pseudo-low}}$ by randomly applying distortion to the original high-quality speech. When we denote the restored speech feature estimated from the pseudo-degraded speech feature $Y_{\text{pseudo-low}}$ as \hat{Z}_{res} , we define the feature loss as $\mathcal{L}_{\text{feature}} = \|Z_{\text{high}} - \hat{Z}_{\text{res}}\|_2$. Note that we input the high-quality speech waveform to the channel module instead of using the output of the synthesis module. Then we train the model to minimize the weighted sum of the reconstruction loss and feature loss in the same manner as Eq. (3).

4. Experimental evaluations

4.1. Experimental condition

Database. We used both simulated and real degraded speech datasets with the sampling rate set to 22.05 kHz. We created simulated datasets by applying various distortions to the JSUT corpus [29], which consists of around six hours of high-quality speech utterances of a Japanese female speaker. We randomly selected 25 and 25 sentences for the validation and test sets, respectively. For the real data, we used a Japanese historical speech dataset [30]. This dataset was recorded in the 1960s–1970s and consists of nine speakers telling folktales, and the whole duration was about 22-minutes. The analog audio clips recorded on the cassette tape were digitized using a radio cassette player. Because this data contains a lot of additive noise, iZotope RX9 was used for noise reduction as preprocessing.

We randomly selected two-minute audio clips for the validation and test sets. We used the JVS corpus [29] as a high-quality speech corpus for pretraining the synthesis module, supervised pretraining in Section 3.3, and $\mathcal{T}_{\text{backward}}$ of dual learning.

Speech feature analysis. We used two types of speech features \hat{z}_{res} : mel spectrogram and source-filter vocoder features. The 80-dimensional mel spectrogram was extracted with a frame size of 1024 and frame shift of 256. The vocoder features consists of F0 and 41-dimensional mel cepstrum coefficients with a 5-ms frame shift, which were extracted by the WORLD vocoder [31]. Hereafter, we denote these types as “Melspec” and “SourceFilter”, respectively.

Model details. We used the U-Net architecture [32] for the analysis and channel modules. In each down sampling of the U-Net, the temporal resolution was reduced by half with four layers of residual convolution blocks and average pooling. Each residual convolution block consists of convolutional layers and batch normalization layers [33] with a skip connection [34]. During up-sampling, the temporal resolution is doubled by applying deconvolution, resulting in time-variant high-quality speech features with the same temporal resolution as that of the input features. For more details, refer to the published implementation. For the synthesis module, we used the pretrained multi-speaker model³ under the “MelSpec” condition and we trained the model with the JVS corpus under the “SourceFilter” condition.

Training settings. The batch size was four and the maximum number of epochs was set to 50, and the model used for evaluation was selected based on the validation results. Adam [35] was used as the optimizer with the initial learning rate set to 0.001. We applied a learning rate scheduling that multiplies the learning rate by 0.5 if the validation loss did not decrease over three epochs. In dual learning and supervised pretraining, we set β of Eq. (3) to 0.1 and 0.001, respectively. For the evaluation with real data, we used the supervised pretraining described in Section 3.3, in which quantization and resampling were randomly applied to each speech waveform. After μ -law quantization [36] with random quantization bits ranging from 6 to 10 bits, the data was resampled with random sampling frequencies of {8, 11.25, 12, 16} kHz.

4.2. Evaluation on speech restoration with simulated data

We evaluated the proposed method by adding various artificial distortions to the high-quality speech corpus described in Section 4.1. The purpose of this evaluation is to investigate the performance in an experimental setting similar to that of the previous supervised speech restoration model [6]. We compared our proposed method described in Section 3.2 (“SSL-dual”) with the previous supervised model (“Liu, et al.”) for generalized speech restoration tasks [6], where we used a pretrained model provided by the author⁴. This supervised model is trained with artificial paired data generated by randomly applying noise, reverberation, clipping, and low-pass filtering to multi-speaker and multi-lingual clean speech corpora.

We created datasets simulating two in-domain and two out-of-domain distortions. “In-domain” and “out-of-domain” indicate whether the training data of the supervised model includes the distortion or not. The in-domain distortions are (a) **Band-limited**: Applying biquad-lowpass filtering to the high-quality speech with a cut-off frequency of 4 kHz; and (b) **Clipped**: Clipping the high-quality speech waveform with an absolute-

³<https://github.com/jik876/hifi-gan>

⁴<https://github.com/haoheliu/voicefixer>

Table 1: Evaluation results with simulated data

	(a) Band-limited		(b) Clipped		(c) Quantized & Resampled		(d) Overdrive	
	MCD	MOS	MCD	MOS	MCD	MOS	MCD	MOS
Ground-truth	-	4.51 ± 0.092	-	4.58 ± 0.083	-	4.67 ± 0.098	-	4.65 ± 0.098
Input	17.85	2.38 ± 0.115	11.21	2.45 ± 0.124	22.25	1.73 ± 0.115	16.73	1.54 ± 0.105
Liu, et al. [6]	6.98	3.74 ± 0.115	9.33	3.01 ± 0.119	9.34	2.80 ± 0.109	10.23	2.00 ± 0.112
SSL-dual (MelSpec)	5.61	4.20 ± 0.090	7.20	3.49 ± 0.114	8.07	3.27 ± 0.111	8.93	2.68 ± 0.108
SSL-dual (SourceFilter)	5.88	3.46 ± 0.115	7.59	2.49 ± 0.190	9.60	2.66 ± 0.109	9.06	2.58 ± 0.125

value threshold of 0.25. The out-domain ones are more significant distortions than in-domain ones, which are (c) **Quantized & resampled**: Quantizing the high-quality speech to 8-bit μ -law and resampling it to 8 kHz; and (d) **Overdrive**: Overdriving the high-quality speech using the SoX implementation [37].

We conducted an objective evaluation using mel cepstral distortion (MCD) [38] with the dimension of the cepstrum set to 24. We also evaluated with subjective mean opinion score (MOS) test, where we recruited 40 Japanese speakers for each evaluation and calculated MOS with 95 % confidence intervals. Table 1 lists the results, with the best results shown in bold.

As shown in the table, “SSL-dual (Melspec)” significantly improves the quality of input degraded speech in all cases, indicating that it can greatly improve the speech quality. We also verified that the proposed method significantly outperforms the previous supervised method in both out-of-domain and in-domain cases. Particularly, in the simple band-limited case, the proposed method achieves a high MOS of 4.20. In the proposed methods, the quality of “Melspec” is higher than that of “SourceFilter”. The difference in MOS may be due to the lower naturalness of the neural vocoder in the SourceFilter, while MCD shows a significant improvement over the previous supervised model.

4.3. Evaluation on speech restoration with real data

We also evaluated the proposed method with real data. Note that there is no ground-truth speech in this data. Because this is a small dataset, we applied supervised pre-training as shown in Section 4.1. Table 2(a) lists the results of the MOS test. The suffix “-pre” indicates the use of the supervised pre-training.

The proposed method “SSL-dual-pre (Melspec),” achieved the highest score and significantly outperformed the previous supervised model. Here, the performance of the previous method and some of the proposed methods were lower than the input, which is most likely due to the quality degradation in the neural vocoder. We also determined that the MOS of both MelSpec and SourceFilter can be improved by supervised pre-training. “SSL-dual-pre (Melspec)” showed a higher MOS than the input but there was not significant difference. Therefore, we additionally conducted an AB test with 40 Japanese evaluators to clarify the difference in quality. Table 2(b) indicate that the proposed method achieved significantly higher preference score than the input. The above evaluation results show that our proposed method is feasible when only a small amount of real data is available.

4.4. Evaluation on audio effect transfer

To investigate the effectiveness of the audio effect transfer, we retrieved the acoustic distortion from the training data and added it to different high-quality speech data. There were two kinds of training data: “Simulated” generated with quantization and resampling as in Section 4.2 and “Real” used in Section 4.3. The high-quality speech to be modified was sampled from the

Table 2: Evaluation results with real data

(a) MOS tests of all methods	
	MOS
Input	2.98 ± 0.136
Liu, et al. [6]	2.80 ± 0.138
SSL-pre (MelSpec)	2.70 ± 0.127
SSL-dual (MelSpec)	2.96 ± 0.135
SSL-dual-pre (MelSpec)	3.06 ± 0.140
SSL-pre (SourceFilter)	2.14 ± 0.123
SSL-dual (SourceFilter)	2.41 ± 0.124
SSL-dual-pre (SourceFilter)	2.55 ± 0.135
(b) Preference AB test	
SSL-dual-pre (MelSpec) vs. Input	<i>p</i> -value
0.565 - 0.435	0.047

Table 3: Evaluation results of audio effect transfer

	Simulated	Real
Reference	3.98 ± 0.152	2.99 ± 0.146
High-quality	1.16 ± 0.069	1.30 ± 0.093
Mean spec. diff.	1.68 ± 0.113	-
Proposed	3.44 ± 0.159	2.12 ± 0.155

JVS corpus. We conducted a five-level similarity MOS (SMOS) test with 40 listeners to evaluate the degree to which the output speech sounded distorted like the training data. “Reference” is a sample that has the same channel features as the training data. For comparison, we prepared a method (“Mean spec. diff”) that multiplies the amplitude spectrum of the high-quality speech by the time-averaged amplitude spectrum difference of training data and high-quality data.

The results show that the proposed method is closer to the target channel features than the original high-quality speech on both simulated and real data. Regarding the results on the simulated data, the SMOS is significantly higher than that of the case corrected by the mean amplitude spectral difference. Although the score of “Proposed” is 0.5–0.8 smaller than the reference, it performed sufficiently.

5. Conclusion

We proposed a self-supervised speech restoration method consisting of analysis, synthesis, and channel modules. Our method achieved significantly higher-quality speech restoration than the previous supervised method in several simulated data and was effective even with low-resource real data. Our future work includes applying our method to a wider variety of real data, such as where multiple speakers are speaking at the same time.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 21H04900.

6. References

- [1] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.
- [2] J. H. L. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless Steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. INTERSPEECH*, 2018, pp. 2758–2762.
- [3] B. Abraham, D. Goel, D. Siddarth, K. Bali, M. Chopra, M. Choudhury, P. Joshi, P. Jyoti, S. Sitaram, and V. Seshadri, "Crowdsourcing speech data for low-resource languages from low-income workers," in *Proc. LREC*, 2020, pp. 2819–2826.
- [4] S. Takamichi and H. Saruwatari, "CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects," in *Proc. LREC*, 2018, pp. 434–437.
- [5] K. Matsuura, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Speech corpus of ainu folklore and end-to-end speech recognition for ainu language," in *Proc. LREC*, 2020, pp. 2622–2628.
- [6] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv preprint arXiv:2109.13731*, 2021.
- [7] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, 2007.
- [8] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017.
- [9] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using kalman filter and em algorithm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 394–406, 2015.
- [10] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. EUSIPCO*, 2018, pp. 390–394.
- [11] 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*, vol. 27, pp. 1256–1266, 2019.
- [12] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," *Proc. ICASSP*, pp. 181–185, 2020.
- [13] F. Bie, D. Wang, J. Wang, and T. F. Zheng, "Detection and reconstruction of clipped speech for speaker recognition," *Speech Communication*, vol. 72, pp. 218–231, 2015.
- [14] P. Závřiska, P. Rajmic, A. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, pp. 5–24, 2021.
- [15] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.
- [16] X. Shu, Y. Zhu, Y. Chen, L. Chen, H. Liu, C. Huang, and Y. Wang, "Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation," *arXiv preprint arXiv:2107.09298*, 2021.
- [17] A. Polyak, L. Wolf, Y. Adi, O. Kabeli, and Y. Taigman, "High fidelity speech regeneration with application to speech enhancement," *Proc. ICASSP*, pp. 7143–7147, 2021.
- [18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.
- [19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv preprint arXiv:2106.07447*, 2021.
- [20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.
- [21] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2.0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.
- [22] H. Nguyen, F. Bougares, N. A. Tomashenko, Y. Estève, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Proc. INTERSPEECH*, 2020, pp. 1466–1470.
- [23] Y. Wang, S. Venkataramani, and P. Smaragdis, "Self-supervised learning for speech enhancement," *arXiv preprint arXiv:2006.10388*, 2020.
- [24] R. E. Zezario, T. Hussain, X. Lu, H.-M. Wang, and Y. Tsao, "Self-supervised denoising autoencoder with linear regression decoder for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6669–6673.
- [25] Z. Du, M. Lei, J. Han, and S. Zhang, "Self-supervised adversarial multi-task learning for vocoder-based monaural speech enhancement," in *Proc. INTERSPEECH*, 2020, pp. 3271–3275.
- [26] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. ICLR*, 2020.
- [27] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.
- [28] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Proc. NIPS*, 2016, pp. 820–828.
- [29] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, pp. 761–768, 2020.
- [30] T. Sasaki, *Tono no mukashibanashi / Sasaki Tokuo hen*. Ohfusha, 1985.
- [31] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.
- [32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
- [33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc ICML*, 2015, p. 448–456.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. CVPR*, pp. 770–778, 2016.
- [35] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [36] C. Recommendation, "Pulse code modulation (pcm) of voice frequencies," in *ITU*, 1988.
- [37] C. Bagwell, et al., "SoX - Sound eXchange." [Online]. Available: <http://sox.sourceforge.net/>
- [38] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.