Self-Supervised Learning and Multi-Task Pre-Training Based Single-Channel Acoustic Denoising

Yi Li, Yang Sun, and Syed Mohsen Naqvi

***Abstract*— In self-supervised learning-based single-channel speech denoising problem, it is challenging to reduce the gap between the denoising performance on the estimated and target speech signals with existed pre-tasks. In this paper, we propose a multi-task pre-training method to improve the speech denoising performance within self-supervised learning. In the proposed pre-training autoencoder (PAE), only a very limited set of unpaired and unseen clean speech signals are required to learn speech latent representations. Meanwhile, to solve the limitation of existing single pre-task, the proposed masking module exploits the dereverberated mask and estimated ratio mask to denoise the mixture as the new pre-task. The downstream task autoencoder (DAE) utilizes unlabeled and unseen reverberant mixtures to generate the estimated mixtures. The DAE is trained to share a latent representation with the clean examples from the learned representation in the PAE. Experimental results on a benchmark dataset demonstrate that the proposed method outperforms the state-of-the-art approaches.**

I. Introduction

Deep learning techniques have been extensively utilized in speech denoising for teleconferencing, automatic speech recognition (ASR), and hearing aids [1] [2]. However, the neural networks are predominantly trained in a supervised mechanism. A vast training set of clean speech signals is required to be well-labelled in the training stage and suffers from drawbacks such as the strong possibility of a mismatch between the training and inference conditions [3] [4]. To relax the constraints of supervised learning approaches, self-supervised learning (SSL) based speech denoising aims to train the model without large labelled datasets to reconstruct the target speech signal from noisy speech. Therefore it becomes highly practical and attractive.

Recently, the SSL techniques have been applied in speech denoising problem. Wang et al. use an autoencoder to learn a latent representation of clean speech signals [3]. However, the pre-training stage only consists of one pre-task which is the mapping of the clean speech spectrogram. Then, Kataria et al. propose a framework called Perceptual Ensemble Regularization Loss (PERL) which shows effectiveness on SSL PASE+ models [5] [6]. However, the PERL is limited with the requirement of massive training data.

Followed by our previous work [7], to further improve the speech denoising performance, we introduce both the dereverberation mask (DM) and the estimated ratio mask (ERM) to provide the time-frequency relationships between the clean speech signal and the reverberant mixture. Hence, inspired by [8], we propose a multi pre-tasks SSL method which only needs a limited set of randomly selected clean speech signals and the corresponding mixture recordings in the pre-training.

The contributions of this paper are summarized as follows:

• Multi pre-tasks with self-training are proposed to solve the speech denoising problem.

• To address the speech denoising problem in reverberant room environments, the DM and the ERM are firstly proposed for SSL-based speech enhancement.

II. Proposed Method

A. Multi pre-tasks based autoencoders

The block diagram of the proposed method is shown in Fig. 1 (a). In this paper, we adopt the variational autoencoder (VAE) [9] as the primary framework. Because within the SSL cases where a limited training set of the labelled data is applied, as proved in [10], the VAE performs better at learning the low-dimensional latent space. In the training stage, we exploit two variational autoencoders, pre-training autoencoder (PAE) and downstream task autoencoder (DAE). The encoder and decoder of the PAE are denoted as E_1 and D_1, respectively. Similarly, we use E_2 and D_2 to present the encoder and decoder of the DAE respectively.

In the training stage, the input of the PAE consists of a very limited set of unpaired and unseen clean speech signals, background noise, and room impulse responses for both speech and noise signals. The mel-frequency cepstral coefficients (MFCC) feature [11] is first extracted. The encoder E_1 obtains the features as the input and produce the latent representation of the clean speech signal. In the proposed method, we consider two pre-tasks for pre-training: latent representation and mask estimation. The first task aims to learn the latent representation of only clean speech signals. However, the second task trains DM and ERM to describe the time-frequency (T-F) relationships from the target speech signal to the mixture. Both the latent representation and masks are trained by minimizing the discrepancy between the ground truth and the corresponding reconstruction. The decoder is trained by the losses from two tasks and produces the estimated speech signal.

In PAE, both E_1 and D_1 consist of 4 1-D convolutional layers. In E_1, the size of the hidden dimension sequentially decreases from 512 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 64. Consequently, the dimension of the latent space is set to 64, and a stride of 1 sample with a kernel size of 7 for the convolutions. Different

Y. Li and S. M. Naqvi are with the Intelligent Sensing and Communications Group, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (e-mails: y.li140, mohsen.naqvi@newcastle.ac.uk)

Y. Sun is working with the Big Data Institute, University of Oxford, Oxford OX3 7LF, U.K. (e-mail: Yang.sun@bdi.ox.ac.uk)

E-mail for correspondence: y.li140@newcastle.ac.uk

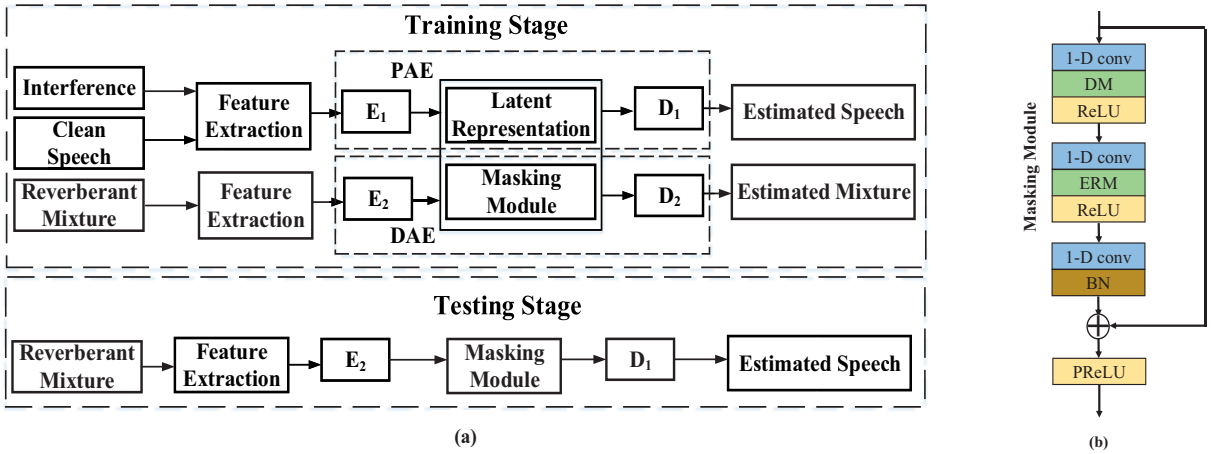

Fig. 1. The block diagram of the proposed method is shown in (a). The masking module is shown in (b). The very limited unseen clean speech and interference signals generate the mixtures for the pre-task autoencoder (PAE). The interference consists of the background noise and room impulse responses of both the clean speech and background noise. The MFCC features are extracted for each stage. The latent representation from the input clean speech signals is learnt in the PAE, meanwhile, the target speech signal in the reverberant mixture is estimated in the masking module. The reverberant mixtures used in the downstream task autoencoder (DAE) are unseen and unpaired to the signals in PAE and mixtures in the testing stage. Then, the estimated mixture is produced and shares the learned representation. The enhanced signal is obtained from the output of the decoder in the testing stage.

from E_1, the D_1 increases the size of the latent dimensions inversely.

Different from the PAE, the DAE only needs access to the reverberant mixtures which are unseen from the PAE. The features are extracted from the reverberant mixture and input to the E_2. Consequently, the latent representation of the mixtures is obtained as the output of E_2. The learnt representation and masks from the PAE are exploited to modify the loss functions and learn a shared latent space between the clean speech and mixture representations. Benefited from the pre-tasks, a mapping from the mixture domain to the target speech domain is learnt with the latent representation of the clean speech signal. Furthermore, D_2 is trained to produce the estimated mixture as the downstream task.

The DAE network follows a similar architecture to PAE. E_2 consists of 6 1-D convolutional layers where the hidden layer sizes decrease from 512 $\rightarrow$ 400 $\rightarrow$ 300 $\rightarrow$ 200 $\rightarrow$ 100 $\rightarrow$ 64, and D_2 increases the sizes inversely.

In the testing stage, after the features are extracted from the unpaired and unseen reverberant mixtures, they are fed into the trained $E2$ as the input. As aforementioned, the loss function in $E2$ is trained with the mapping of the latent space from the mixture domain to the target speech domain. Thus, the trained $E2$ produces an estimated latent representation of the reverberant mixture. Then, the estimated masks are used to dereverberate and denoise the mixture representation. Finally, the trained $D1$ obtains the reconstructed representation and maps to the target speech signal.

B. Masking Module

As aforementioned, the masking module is exploited to train the DM and ERM to describe the T-F relationships between the target speech signal to the mixture. The architecture of the masking module is depicted in Fig. 1 (b).

The masking module has three sub-layers and obtains the mixture spectrogram $\mathbf{Y}$ which consists of the clean speech, background noise, and reverberations of both. The aim of the masking module is to estimate the target speech spectrogram $\hat{\mathbf{S}}$. The first two sub-layers consist of two time-frequency (TF) masks, DM and ERM. According to [12], the DM is presented:

$$\mathbf{DM} = (\mathbf{S} + \mathbf{N}) \cdot \mathbf{Y}^{-1} \tag{1}$$

where '$\cdot$' is the dot product, and $\mathbf{S}$ and $\mathbf{N}$ are the spectra of the clean speech and the background noise, respectively. It is highlighted that $\mathbf{S}$ and $\mathbf{N}$ are both non-reverberant. The dereverberated mixture is obtained as:

$$\hat{\mathbf{Y}}_d = \mathbf{Y} \cdot \widehat{\mathbf{DM}} \tag{2}$$

where $\widehat{\mathbf{DM}}$ is the estimated DM. However, in practice, obtaining the dereverberated mixtures is very challenging [13]. Although most of the reverberations are removed by DM, the remaining reverberations in $\hat{\mathbf{Y}}_d$ still limit the performance [7]. Thus, we exploit ERM in the second sub-layer to further improve the speech denoising in reverberant room environments, which can be defined as:

$$\widehat{\mathbf{ERM}} = \frac{|\mathbf{S}|}{|\hat{\mathbf{Y}}_d|} \tag{3}$$

Then, the background noise and the remaining reverberations are removed by ERM. Moreover, the ReLU activation is added to each mask and produces the output for the next sub-layer. Additionally, a residual connection [14] is applied in the masking module to ease the training of the module. Finally, the target speech spectrogram is obtained with a PReLU activation [15] as:

$$\hat{\mathbf{S}} = \widehat{\mathbf{ERM}} \cdot \widehat{\mathbf{DM}} \cdot \mathbf{Y} \tag{4}$$

The overall loss to train the masking module is a combination of three loss terms as:

$$\mathcal{L}_{\text{masking}} = \lambda_1 \cdot \mathcal{L}_{\text{KL-masking}} + \mathcal{L}_{\mathbf{S}} + \mathcal{L}_{\text{cyc}} \tag{5}$$

where $\mathcal{L}_{\text{KL-masking}}$ denotes the Kullback-Leibler (KL) loss and is applied to train the latent representation closed to a normal distribution [3]. Then, λ_1 is the coefficient of $\mathcal{L}_{\text{KL-masking}}$ and empirically set to 0.001. Besides, $\mathcal{L}_{\mathbf{S}}$ denotes the loss between the target speech signal and the corresponding reconstruction. The $L2$ norm of the error $\mathcal{L}_{\mathbf{S}} = \|\mathbf{S} - \hat{\mathbf{S}}\|_2^2$ is exploited as the loss function. Similarly, the cycle loss $\mathcal{L}_{\text{cyc}}$ consists of $\mathcal{L}_{\text{S}}$ and the loss between the latent representation and the corresponding reconstruction:

$$\mathcal{L}_{\text{cyc}} = \|\mathbf{S} - \hat{\mathbf{S}}\|_2^2 + \lambda_2 \cdot \left\| \mathbf{X}_S - \hat{\mathbf{X}}_{\mathbf{S}} \right\|_2^2 \tag{6}$$

where $\hat{\mathbf{X}}_{\mathbf{S}}$ is the estimated representation of the target speech signal. The $L2$ norm of the error $\mathcal{L}_{\mathbf{S}}$ is calculated twice because the calculation of the cycle loss is based on the $\mathcal{L}_{\mathbf{S}}$. Moreover, λ_2 is the coefficient of representation loss and empirically set to 0.001. Finally, the combination of losses $\mathcal{L}_{\text{masking}}$ are utilized in PAE to improve the speech denoising performance.

III. EXPERIMENTAL RESULTS

A. Experimental Setup

The proposed method is trained by using the Adam optimizer with a learning rate of 0.001 and the batch size is 20. The number of epochs for PAE and DAE are 700 and 1500, respectively. All the experiments are run on a work station with 4 Nvidia GTX 1080 GPUs and 16 GB of RAM. The magnitude spectrograms have 513 frequency bins for each frame as a Hanning window and a discrete Fourier transform (DFT) size of 1024 samples are applied.

To evaluate the proposed model, we use composite metrics that approximate the Mean Opinion Score (MOS) including COVL: MOS predictor of overall signal quality, CBAK: MOS predictor of background-noise intrusiveness, CSIG: MOS predictor of signal distortion [16] and Perceptual Evaluation of Speech Quality (PESQ). Higher values of the measurements imply that the desired speech signal is better estimated.

B. Comparisons and Datasets

We compare the proposed method with two recent SSL speech denoising approaches [3] [8]. The first SSE in [3] exploits two autoencoders to process pre-task and downstream task, respectively. The architecture is similar to the proposed method. The second one is pre-training fine-tune (PT-FT) [8], which uses three models and three SSL approaches for pre-training: speech denoising, masked acoustic model with alteration (MAMA) used in TERA [17] and continuous contrastive task (CC) used in wav2vec 2.0 [18]. We reproduce the PT-FT method with DPTNet model [19] and speech denoising as the pre-task because it shows the best denoising performance in [8].

There are two factors in the self-supervised learning (SSL). According to [3], the first factor is the limited training data and in the proposed method, we only use 12 clean utterances as [3] in the pre-training and 188 mixtures in the downstream task, the clean utterances and mixtures are unpaired. Hence, we think we meet the claim of using limited training data in the SSL.

Besides, the second factor in SSL is using the pre-training stage and downstream task to map the mixture into a latent representation feature space and then obtain the enhanced speech by decoding the feature [8]. We consider two pre-tasks to improve the speech denoising performance and has the same workflow as shown in [8]. Therefore, we again think the proposed work is an SSL-based speech denoising algorithm.

Meanwhile, in some SSL studies, a limited amount of paired data is allowed. For example, in [8], the input of the pre-training stage also consists of both clean speech signals and the background noise which are similar to the proposed work. To further confirm that the proposed method does not require extra information as in [3], we directly apply unpaired and unseen speech mixtures as the interference (a more challenging scenario) in the proposed method and we still outperforms the state-of-the-art methods. The settings of the proposed method and the baselines are described in TABLE I. The cross mark ✗ means the method does not use the setting such as no reverberations in [8] but does not mean it cannot be handled in the method. Besides, 3 pre-tasks are trained in the PT-FT method and we train 2 pre-tasks in the proposed method.

TABLE I

COMPARISON OF SSL SPEECH DENOISING APPROACHES WITH THE PROPOSED METHOD. MORE SPECIFICALLY, THE PT-FT METHOD USE 50,800 PAIRED UTTERANCES IN THE TRAINING STAGE. HOWEVER, ONLY 200 UTTERANCES ARE REQUIRED IN THE PROPOSED METHOD.

	SSE [3]	PT-FT [8]	*Proposed*
Noise	✗	✓	✓
Paired Data	✗	✓	✓
Multiple Models	✓	✗	✓
Single Pre-Task	✓	✗	✗
Reverberation	✓	✗	✓

In the PAE training, 12 clean utterances from 4 different speakers with three reverberant room environments (ipad_livingroom1, ipad_bedroom1, and ipad_confroom1) are randomly selected from the DAPS dataset [20]. The training data consists of 2 male and 2 female speakers each reading out 5 utterances and recorded in different indoor environments with different real room impulse responses (RIRs) [20]. In the DAE training, the unseen and unpaired 300 noisy mixtures from 20 different speakers with three reverberant room environments are randomly selected from the DAPS dataset. The training data consists of 10 male and 10 female speakers each reading out 5 utterances and recorded in different indoor environments with different real room impulse responses (RIRs) [20]. In order to improve the ability of the proposed method in adapting to unseen speakers, the speakers in the DAE training are different from the speakers in the PAE training. Moreover, three background noises ($factory$, $babble$, and $cafe$) from the NOISEX dataset [21] and four SNR levels (-10, -5, 0, and 5 dB) are used to generate the mixtures in both the PAE

and DAE. The validation data contains 50 noisy mixtures generated by the randomly selected reverberant speech from the DAPS dataset and the background noise. In the testing stage, 200 reverberant utterances of 10 speakers are randomly selected and used to generate the mixtures with the same background noises and SNR levels for the configuration in the training stage.

C. Results and Discussions

In the evaluations, we first conduct the experiments in three cases as different interferences in the PAE to further confirm that the proposed method does not require extra information as other SSL works e.g., [3]. :

• Case 1: The interference only consists of the background noises ($factory$, $babble$, and $cafe$).

• Case 2: In the SSE method shown in [3], only limited amount of clean speech signals and unlabeled mixtures are available in the training stage. Therefore, to further evaluate the proposed masking module, we randomly generate a Gaussian noise to produce the reverberant mixture as the interference. Hence, compared with [3], no extra information is introduced. The mixtures used in the PAE and DAE are unseen.

• Case 3: To evaluate the performance with various interferences, we use both the background noise (Case 1) and the unlabelled mixture (Case 2) to generate the interference. In both Cases 2 & 3, the mixtures used in the PAE and the DAE are unseen to the each other.

TABLE II

AVERAGED SPEECH DENOISING PERFORMANCE (CASE 1) IN TERMS OF THREE ROOM ENVIRONMENTS, THREE NOISE INTERFERENCES AND THREE SNR LEVELS.

Method	PESQ	CSIG	CBAK	COVL
SSE [3]	1.48	2.28	1.90	1.84
PT-FT [8]	1.58	2.34	2.04	1.91
Proposed	**1.71**	**2.45**	**2.16**	**1.97**

1) Case 1: From Table 2, it is clearly observed that the proposed method outperforms the state-of-the-art methods in terms of all three performance measurements. In [8], the original PT-FT method is trained with Libri1Mix train-360 set [22] which contains 50,800 utterances. However, in the comparison experiments, we use the limited amount of training utterances (200). Therefore, the speech denoising performance of the PT-FT suffers a significant degradation compared with the original paper. The latent representation and the masking module have limitations, however, the proposed method takes advantage of both approaches and mitigates the speech denoising problem. Thus, the speech denoising performance is improved compared with only learning the clean speech representation in the SSE method.

2) Case 2: It can be seen from Table 3 that the proposed method always achieves highest denoising performance compared with SSE and PT-FT. However, the denoising performance at Case 2 suffers a degradation compared with Case 1. Because in Case 2, the interference consists of the undesired speech signal, the background noise, and reverberation of both speech signals and noises. It is highlighted that, due to different distributions between speech and noise interference domains, the task of personalized speech denoising from the mixture with undesired speech signals is more challenging than from noise interferences [23].

TABLE III

AVERAGED SPEECH DENOISING PERFORMANCE (CASE 2) IN TERMS OF THREE ROOM ENVIRONMENTS, THREE NOISE INTERFERENCES AND THREE SNR LEVELS.

Method	PESQ	CSIG	CBAK	COVL
SSE [3]	1.39	2.31	1.82	1.75
PT-FT [8]	1.44	2.34	1.89	1.90
Proposed	**1.64**	**2.38**	**2.11**	**1.92**

3) Case 3: In this case, we use two background noises ($restaurant$ and $f16$) from the NOISEX dataset [21] and two SNR levels (-5 and 5 dB). The experimental results are shown in Table 4.

TABLE IV

AVERAGED SPEECH DENOISING PERFORMANCE (CASE 3) IN TERMS OF THREE ROOM ENVIRONMENTS, TWO NOISE INTERFERENCES AND TWO SNR LEVELS.

Method	PESQ	CSIG	CBAK	COVL
SSE [3]	1.37	2.27	1.77	1.66
PT-FT [8]	1.49	2.25	1.84	1.87
Proposed	**1.69**	**2.29**	**2.13**	**1.90**

It can be observed from Table 4 that the speech denoising performance is significantly improved by the proposed method compared to the baselines. Although Table 3 indicates the case where the interference consists of the background noise and the reverberant mixture. The improvement in terms of PESQ, CBAK, and COVL is more obvious than the other two cases. The DM can mitigate the adverse effect of acoustic reflections to extract the target speech from the noisy mixture. Then, the ERM is estimated by using the desired speech and the estimated dereverberated mixture, which can further improve the dereverberation. Thus, the proposed ERM can better model the relationship between the clean speech and the estimated dereverberated mixture. As a result, the proposed masking module has a better ability in adapting to unseen speakers and leading to improved performance in highly reverberant scenarios.

In all comparison experiments, we can observe that: (1) The proposed method outperforms the recent SSL-based speech denoising methods. (2) When the interference contains both background noise and the undesired speech signal, the denoising performance is degraded. (3) The proposed method still improves the speech denoising performance in the hardest case (Case 3) because unseen scenario is also considered. Moreover, the improvement is more significant as the case is more challenging.

In the proposed method, we solve the mismatch of the speakers between the training and testing stages, which is most important in practical scenarios e.g., speaker indepen-

dent. Moreover, the proposed method can be used where both SNR levels and noise types are unseen, however, the speech denoising performance suffers a slight degradation, which will be handled in future work.

IV. Conclusion

In order to address the single-channel speech denoising problem in reverberant room environments, a multi pre-tasks SSL method was proposed. In the pre-training stage, the latent representation of the clean speech signal was learnt as the first pre-task. Meanwhile, in the PAE, the DM- and ERM-based masking module was applied to assist to estimate the target speech spectrogram. We evaluated the proposed method in three cases with different interferences. The experimental results showed that pre-training with multi pre-tasks provides better speech denoising performance than the state-of-the-art approaches within the benchmark dataset.

References

[1] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network'," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359 – 369, 2019.

[2] A. Pandey and D. L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[3] Y.-C. Wang, S. Venkataramani, and P. Smaragdis, "Self-supervised learning for speech enhancement," *International Conference on Machine Learning (ICML)*, 2020.

[4] Z. H. Du, M. Lei, J. Q. Han, and S. L. Zhang, "Self-supervised adversarial multi-task learning for vocoder-based monaural speech enhancement," *Interspeech*, 2020.

[5] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] M. Ravanelli, J. Y. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[7] Y. Li, Y. Sun, and S. M. Naqvi, "Single-channel dereverberation and denoising based on lower band trained SA-LSTMs," *IET Signal Processing*, vol. 14, no. 10, pp. 774 – 782, 2021.

[8] S.-F. Huang, S.-P. Chuang, D.-R. Liu, Y.-C. Chen, G.-P. Yang, and H.-Y. Lee, "Stabilizing label assignment for speech separation by self-supervised pre-training," *Interspeech*, 2021.

[9] Y. Li, Y. Sun, K. Horoshenkov, and S. M. Naqvi, "Domain adaptation and autoencoder based unsupervised speech snhancement," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 43 – 52, 2021.

[10] Z. Ding, Y. F. Xu, W. J. Xu, G. Parmar, Y. Yang, M. Welling, and Z. W. Tu;, "Guided variational autoencoder for disentanglement learning," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] M. Xu, L.-Y. Duan, J. F. Cai, L.-T. Chia, C. S. Xu, and Q. Tian, "HMM-based audio keyword generation," *Advances in Multimedia Information Processing: 5th Pacific Rim Conference on Multimedia*, pp. 566 – 574, 2004.

[12] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 125–138, 2019.

[13] Y. Zhao and D. L. Wang, "Noisy-reverberant speech enhancement using DenseUNet with time-frequency attention," *Interspeech*, 2020.

[14] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," *IEEE International Conference on Computer Vision*, 2015.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 – 238, 2008.

[17] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351 – 2366, 2021.

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," *Neural Information Processing Systems (NeurIPS)*, 2020.

[19] J. J. Chen, Q. R. Mao, and D. Liu, "Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation," *Interspeech*, 2020.

[20] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006 – 1010, 2014.

[21] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 12, no. 3, pp. 247 – 251, 1993.

[22] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: an open-source dataset for generalizable speech separation," *Interspeech*, 2020.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: deep audio-visual speech enhancement," *Interspeech*, 2018.