Towards Modeling Auditory Restoration in Noisy Environments

Yating Huang*†, Yunzhe Hao*†, Jiaming Xu*‡¶ and Bo Xu*†‡§¶
*Institute of Automation, Chinese Academy of Sciences, Beijing, China
†School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
‡School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
§Center for Excellence in Brain Science and Intelligence Technology, CAS, China
Email: {huangyating2016, haoyunzhe2017, jiaming.xu, xubo}@ia.ac.cn

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534379

***Abstract*—Real-world sounds are often interrupted by various kinds of noise. The target signal of the mixture sounds is often degraded or lost. While the human auditory system can extract the target signal from the mixture and restore the degraded or lost parts simultaneously, current computational models often simplify the complex scenarios, which leads to two individual tasks, audio inpainting and speech enhancement. In this work, we take a pioneering step towards modeling auditory restoration, that is to restore the target speech signal, in which there are missing parts in the target signal and the target signal is interfered by background noise. Different from the speech enhancement task, we attempt to fill in the missing gaps with the existence of background noise. Different from the auditory inpainting task, there is some noise in our input signal and the positions of the missing gaps are unknown. In other words, we attempt to reduce interference and restore missing gaps simultaneously. We propose Hourglass-shaped Convolutional Recurrent Network (HCRN) trained with Spectro-Temporal loss to restore the target signal from the incomplete noisy mixture. Moreover, instead of restoring non-human sounds, we focus on speech restoration, which poses more challenges on reconstruction. Both the quantitative and qualitative performance show that our proposed method can suppress the background noise, identify and restore the missing gaps of the salient signal with the unreliable context information. Our code is available in https://github.com/aispeech-lab/HCRN.**

***Index Terms*—auditory restoration, audio inpainting, speech enhancement**

I. Introduction

Everyday communication often occurs in noisy environments. The human auditory system is robust enough to filter the target sound from the mixture. In some situations, besides extracting the target signal from the mixture, we also have to fill in the missing parts or greatly distorted parts of the target signal. The latter phenomenon is known as auditory restoration or induction [1]. However, there are few attempts to model auditory restoration formally. Early years, research was concerned with modeling phonemic restoration, an example of auditory restoration [2]–[4]. Recent research that is most related to auditory restoration is auditory inpainting [5] and speech enhancement [6]. Audio inpainting is a concept that borrows from image inpainting [7]–[9] in computer vision, whose aim is to recover the missing parts of a given image. Traditional methods that model audio inpainting often work in real-time settings and predict the current lost frame depending on the preceding frames, like Linear Predictive Coding (LPC) [10]. These methods work well when recovering short gaps (10-20 ms), but can not achieve satisfying performance when it comes to long audio inpainting. With the rapid developments of deep learning, deep neural networks are applied to audio inpainting tasks. [11], [12] applied encoder-decoder structure and transformed the audio into Time-Frequency (T-F) spectrogram using Short-Time Fourier Transform (STFT) to recover the missing gaps using the context information. [13] used a similar approach and introduced both the spectrogram inpainting model and waveform inpainting model to tackle the long audio inpainting problem. However, it is worth noting that they performed the quantitative and qualitative tasks on sound classification task, which is not very straightforward to evaluate the performance of audio inpainting results. Similarly, [14] used a U-Net [15] structure to recover the corrupted spectrograms and trained the network using deep feature losses by employing a VGG feature extractor network [16]. The research mentioned above often assumes that the input signal is not contaminated by noise, and the positions of the missing gaps are provided to the algorithm as a prior, which is non-blind [12], [17]. On the other hand, the modern speech enhancement task doesn't consider the cases in which missing parts exist. Most of the existing speech enhancement approaches that aim to estimate T-F representations of the target speech are divided into two groups: masking-based methods and mapping-based methods. The masking-based methods, which are the more popular choice, predict a mask and use the mask to filter the target signal from the mixture [6], which can not recover the missing gaps of the mixture. The mapping-based methods predict the clean spectral features from the noisy features directly [18], [19]. Tan and Wang [18], [19] proposed to use Convolutional Recurrent Network (CRN) to do spectral mapping. In our practice of doing multi-speaker speech separation in a noisy environment, which is also known as the "Cocktail Party Problem" [20], there are some occasions that some phonemes are greatly corrupted and the target signal is degraded by noise to some extent. In this situation, the target signal tends to be salient while the background noise is unsalient, and the context information of the lost or degraded parts are unreliable.

¶ Corresponding author.

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

In this work, we take a pioneering step towards modeling auditory restoration, in which the target signal is contaminated by noise and the positions of the missing gaps are unknown. In other words, local audio information near the missing gaps is unreliable and we need to jointly reduce the interference signal and restore the target signal. Moreover, instead of restoring non-human sounds, we focus on speech restoration, which poses more challenges on reconstruction. We study the ability of a U-Net-like neural network to model auditory restoration, attempting to enhance the target signal and fill in the missing gaps of the target signal. We propose Hourglass-shaped Convolutional Recurrent Network (HCRN) to model auditory restoration, which is based on Hourglass structure [21] and combined with a Bidirectional Long Short-Term Memory (BLSTM) network [22] for temporal modeling. We propose to add the loss in the time domain between the target signal and the predicted signal as a normalization term while training our model, leading to better performance. We conduct quantitative and qualitative evaluations to evaluate the effectiveness of our proposed method. To the best of our knowledge, we are the first to model auditory restoration with a large gap size and the intrusion of background noise. Following the current training procedure in our work, our proposed model can perform auditory restoration at different Signal-to-Noise Ratio (SNR) levels and with different gap sizes.

II. Modelling Auditory Restoration

In this section, we first give a simple definition of the auditory restoration task. Then we describe our proposed model for solving the task and our proposed loss function for training.

A. Auditory Restoration

In this paper, we attempt to tackle the problem of auditory restoration with large gap size and the intrusion of background noise. Specifically, the input signal is the mixture of the target speech signal and interference signal at some SNR level, where some parts of the mixture are lost or degraded. Different from non-blind auditory inpainting [5], [12], [14], [17], [23], we assume that the positions of the missing gaps are unknown. Moreover, the target signal is interfered by noise, but the target signal is the salient signal in the mixture. In other words, we treat the salient sound as our target signal, and thus our goal is to enhance the target signal from the interference of noise and fill in the missing gaps of the salient sound.

B. Hourglass-shaped Convolutional Recurrent Network

We propose Hourglass-shaped Convolutional Recurrent Network (HCRN) to recover the spectrogram of the target signal. We only consider the spectrogram-based methods instead of waveform-based methods, because, in our preliminary experiments, we failed to use waveform-based methods, such as Demucs [24], to recover the lost parts of the target waveform. Demucs is only able to reduce the interference signal and enhance the target signal, but is not able to fill in the gaps. Thus, in this work, we focus on the enhancement and reconstruction of the corrupted spectrograms.

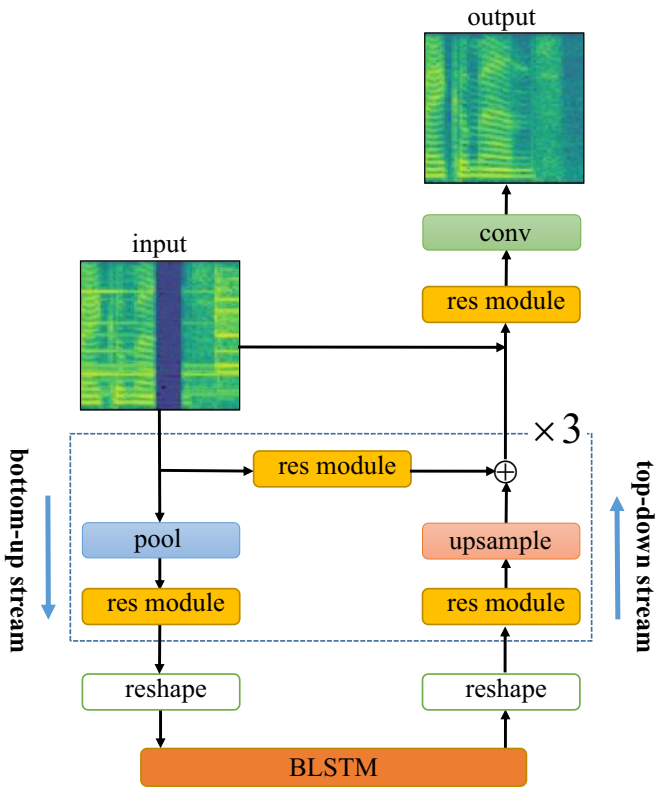


Fig. 1: Our proposed HCRN for auditory restoration. The normalized noisy incomplete spectrograms are served as input. The convolutional block in the dashed blue box is stacked 3 times in our proposed network. A 2-layer BLSTM is inserted at the bottom of the network to leverage longer-term context both from the past and the future. The output of the first convolutional block along the top-down stream is concatenated with the input magnitude spectrogram to feed into another residual module, followed by another convolutional layer and a sigmoid function to generate the final output.

HCRN is based on a U-Net-like Hourglass structure [21] to estimate and restore the target magnitude spectrogram. Figure 1 shows an overview of its structure. The U-Net-like Hourglass encoder-decoder structure can learn and integrate features at different scales. Our proposed network comprises 3 convolutional blocks, as depicted in the dashed blue block in Figure 1. The convolutional block is composed of 3 residual modules, a max pooling layer, and an upsampling layer. The residual module is a 3-layer convolutional network, as shown in Figure 2. To leverage longer context, in the residual module, two 7×7 convolutional layers with a stride of 1 and a filter size of 32 are followed by a 1×1 convolutional layer with a filter size of 64, leading to feature maps of the same size. Batch Normalization [25] and Rectified Linear Unit (ReLU) activation are applied before each convolutional layer. Residual learning [26] is enforced in the network. To be specific, a

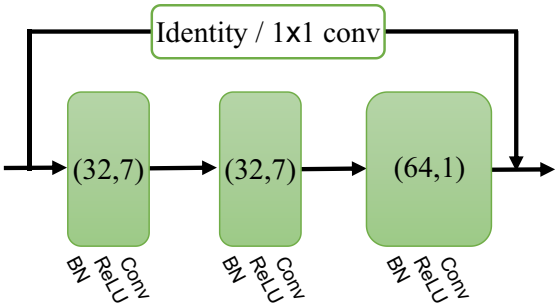


Fig. 2: The residual module used in our proposed HCRN. (Channel, kernel) of the convolutional layer is denoted. Batch Normalization (BN) and ReLU are applied before each convolutional (Conv) layer. Identity connection or 1×1 convolution is applied to enforce residual learning.

1×1 convolutional layer is added to change the channel number of input before doing element-wise addition if the channel number of input and output differ, otherwise we directly do element-wise addition of the two feature sets. The convolutional block performs the following computation:

$$\mathbf{X}_i^{bottom} = Res_i^{bottom}(Pool(\mathbf{X}_{i-1}^{bottom})) \quad (1)$$

$$\mathbf{X}_{i-1}^{top} = Res_i^{inter}(\mathbf{X}_{i-1}^{bottom}) + Upsample(Res_i^{top}(\mathbf{X}_i^{top})) \quad (2)$$

Here, Res_i^{bottom}, Res_i^{top} and Res_i^{inter} represent the residual module at the bottom-up stream, the residual module at the top-down stream and the residual module that interacts the bottom-up stream and the top-down stream in the i_{th} convolutional block, respectively. $\mathbf{X}_{i-1}^{bottom}$ and $\mathbf{X}_i^{bottom}$ denote the input and output at the bottom-up stream of the i_{th} convolutional block, respectively. $\mathbf{X}_i^{top}$ and $\mathbf{X}_{i-1}^{top}$ stand for the input and output at the top-down stream of the i_{th} convolutional block, respectively. Note that the Hourglass structure differs from U-Net [27] in interacting the bottom-up stream and top-down stream. The U-Net structure directly concatenates the feature maps from the bottom-up stream with the feature maps at the corresponding layer from the top-down stream to integrate information across scales. The Hourglass structure applies more convolutions to the feature maps from the bottom-up stream and does element-wise addition of the two feature sets. These manipulations provide more non-linearity compared to concatenation.

We adopt a 2-layer stacked Bidirectional Long Short-Term Memory (BLSTM) network [22] with a hidden size of 512 at the bottom of our proposed network to capture temporal dynamics of speech in both directions. By using a BLSTM, long-term context from the past and the future are leveraged to convolutional neural networks.

As shown in Figure 1, for feature map $\mathbf{X}_3^{bottom} \in \mathbb{R}^{C \times T \times F}$ from the third convolutional block, C, T, F denote channel dimension, time dimension and frequency dimension, respectively. We transpose time dimension and channel dimension of $\mathbf{X}_3^{bottom}$, then flatten the channel dimension and frequency dimension, leading to $\hat{\mathbf{X}}_3^{bottom} \in \mathbb{R}^{T \times (C \cdot F)}$, which can be treated as a sequence $\hat{\mathbf{X}}_3^{bottom} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T], \boldsymbol{x}_t \in \mathbb{R}^{C \cdot F}$ and fed to the BLSTM network. The hidden states from the last layer of BLSTM of both directions are concatenated as $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]$. The output sequences are then reshaped back to fit the convolutional network and serve as input to the third convolutional block along the top-down stream.

The normalized linear magnitude spectrograms $\mathbf{S}^{mix}(t, f)$ with missing gaps are extracted as the input and fed to the network. $\mathbf{X}_0^{top}$ is concatenated with the input magnitude spectrogram, followed by another residual module, a 3×3 convolutional layer, and a sigmoid function to generate the final output. In this work, we leave the task of phase reconstruction for future research and the original complete phase information of the mixture is used in inverse Short-Time Fourier Transform (iSTFT) to reconstruct the time-domain signal from the estimated magnitude. Though previous research often uses Griffin-Lim algorithm [28] to reconstruct the phase [13], [14], there is a big difference between our experimental settings and theirs. In our work, the non-loss parts of our corrupted incomplete spectrogram are interfered by noise almost everywhere, while the non-loss parts of their input are clean and reliable. We find that Griffin-Lim algorithm doesn't perform well when it comes to the enhanced spectrogram. In practice, most of the time phase information of the audio clips are available and the reconstruction of the spectrogram is not very sensitive to noisy phase. Therefore, we used the mixture phase for iSTFT here.

C. Considering Loss in Time Domain

We use L_1 loss to compute the loss in the frequency domain. Since spectrograms are lossy compared to time-domain signals, besides computing L_1 loss between the estimated magnitude and the target magnitude, we also consider L_1 loss of the estimated waveform and the target waveform. The motivation is that by considering the loss in the time-domain waveform, the phase information of the mixture is combined to recover the time-domain signals from spectrograms, which exposes more temporal information when training the model. The loss of auditory restoration task is therefore formulated as Eq. 3 in our work,

$$L = L_1(\mathbf{S}^{est}(t, f), \mathbf{S}^{gt}(t, f)) + \alpha L_1(s^{est}(t), s^{gt}(t)) \quad (3)$$

where $\mathbf{S}^{est}(t, f)$, $\mathbf{S}^{gt}(t, f)$, $s^{est}(t)$ and $s^{gt}(t)$ denote estimated spectrogram, target spectrogram, estimated waveform and target waveform, respectively. α is a hyperparameter that balances the two terms. To be specific, we use the phase of the mixture signal to reconstruct the predicted waveform. The second term of Eq. 3 can be considered as a normalization term. For simplicity, we refer to Eq. 3 as Spectro-Temporal loss and refer to the loss in the frequency domain as Spectrum loss, which is the first term in Eq. 3.

TABLE I: Improvements of SDR, PESQ and STOI results on test set. The evaluation scores listed here are all the higher, the better. The best score for each metric is denoted in bold.

SNR/gap size	5 dB/ 0 ms			10 dB/ 0 ms			15 dB/ 0 ms		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Unprocessed	3.12	1.94	0.85	7.75	2.25	0.91	12.63	2.62	0.95
Model	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi
U-Net	4.28	0.19	0.02	2.91	0.21	0.02	0.36	0.13	0.01
HCRN (Spectrum loss)	6.95	0.64	0.05	6.16	0.79	0.04	4.20	0.76	**0.03**
HCRN (Spectro-Temporal loss)	**8.07**	**0.68**	**0.06**	**7.49**	**0.81**	**0.05**	**5.87**	**0.77**	**0.03**
SNR/gap size	5 dB/ 96 ms			10 dB/ 96 ms			15 dB/ 96 ms		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Unprocessed	2.35	1.54	0.81	6.28	1.70	0.87	9.80	1.89	0.91
Model	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi
U-Net	3.94	0.42	0.03	2.84	0.53	0.03	1.14	0.57	0.02
HCRN (Spectrum loss)	6.75	0.90	**0.08**	6.02	1.12	**0.07**	4.66	1.22	**0.05**
HCRN (Spectro-Temporal loss)	**7.77**	**0.91**	**0.08**	**7.09**	**1.14**	**0.07**	**5.79**	**1.23**	**0.05**
SNR/gap size	5 dB/ 176 ms			10 dB/ 176 ms			15 dB/ 176 ms		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Unprocessed	1.81	1.46	0.78	5.36	1.58	0.83	8.32	1.73	0.87
Model	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi	SDRi	PESQi	STOIi
U-Net	3.94	0.41	0.04	2.91	0.51	0.04	1.47	0.55	0.03
HCRN (Spectrum loss)	6.28	**0.81**	**0.09**	5.42	1.00	**0.09**	4.05	1.09	**0.07**
HCRN (Spectro-Temporal loss)	**7.12**	**0.81**	**0.09**	**6.22**	**1.01**	**0.09**	**4.83**	**1.10**	**0.07**

III. Experiments

A. Datasets

In our experiments, the Wall Street Journal (WSJ) speech corpus[1], Environmental Sound Classification (ESC-50) dataset [29], and AudioSet [30] are used to generate the mixture. We randomly choose a sample from WSJ dataset to act as the target signal, and randomly choose another sample from ESC-50 dataset, AudioSet, and WSJ dataset with equal opportunity as the interference signal, to generate the mixture input. All audio clips are sampled at 8 kHz.

WSJ corpus is a well-known English corpus of reading sentences from the Wall Street Journal, recorded by different speakers under clean conditions. The split of training set si284 (37416 utterances, 81 hours), validation set dev93 (503 utterances) and test set eval92 (333 utterances) of WSJ is according to the official split. We use WSJ corpus to mimic the salient speech signal as target and the unsalient speech as background noise.

ESC-50 and AudioSet are used to mimic the unsalient non-speech background noise. ESC-50 has 2000 5-second-long environmental audio recordings of 50 classes. In our work, 90% of ESC-50 dataset is used for training and validation, while 10% is used for the test set. AudioSet is a large-scale collection of human-labeled sound clips drawn from YouTube videos. In our experiments, we used the inside-small-room subset of AudioSet, which are sounds that appear to have been recorded within a small room. The unbalanced train set of the subset is used as the training set, the balanced train set is used as the validation set and the balanced evaluation set is used as the test set.

[1]Obtained from LDC under the catalog numbers LDC93S6B and LDC94S13B.

B. Baseline

As a baseline, we train a U-Net model based on [27], but both the encoder and decoder in our U-Net have one less layer overall, as our input is of a lower resolution. Since there are missing gaps in our input, we use the baseline U-Net model to predict the magnitude instead of a mask.

C. Experimental Setups

In the training process, we generate the training samples on the fly. To make the noisy training set, we randomly select a sample from the WSJ corpus as the target signal. With a probability of $1/3$, we randomly select another sample from WSJ, ESC-50, or AudioSet as the interference signal, to generate the mixture. We mix the target speech signal and the interference signal at a random SNR level ranging from 0 to 15 dB so that the target speech is the salient signal. STFT is used to generate the input magnitude spectrogram. The window size is 32 ms and the hop size is 16 ms, resulting in 128 frequency bins. The magnitude spectrograms are normalized to the range [0,1]. We segment a sample to several 128-frame patches and randomly delete consecutive 10 frames (176 ms) from every patch to mimic the lost parts of the input signal. The resulted segmented spectrograms are fed to the networks. To investigate the influence of the loss term in the time domain, we also train our proposed HCRN with Spectrum loss. All the models are trained with Adam optimizer [31] and an initial learning rate of 0.0002. We train the models for 30 epochs. α in Eq. 3 is 0.025. Finally, we use the complete mixture phase to recover the time-domain signal from estimated spectrograms using iSTFT.

To generate the test set, we use the same strategy as in the training process to randomly select a noise list of 333 samples. And then the noise is added to the WSJ eval92 set under 3 SNR conditions: 5 dB, 10 dB, and 15 dB. Afterward, we delete n ($n = 0, 5, 10$) consecutive frames every 128-frame

patch at the same randomly selected starting position in each condition, which are no missing gap, 96-ms missing gap, and 176-ms missing gap, respectively. The resulted test set has 333 utterances on each occasion. Note that there will be multiple missing gaps in a sample according to the length of the target signal.

IV. Results

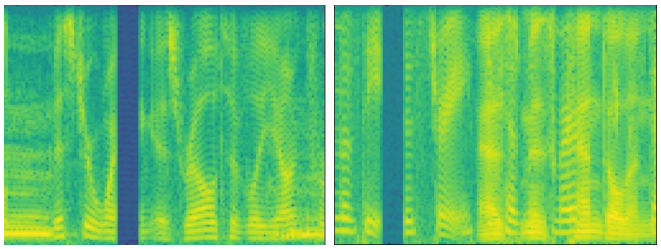
A. Quantitative Performance

In this study, we use Signal-to-Distortion Ratio (SDR) [32], Perceptual Evaluation of Speech Quality score (PESQ, ranging from 1 to 5) [33] and Short-Time Objective Intelligibility (STOI, ranging from 0 to 1) [34] to evaluate the improvements of applying auditory restoration. We define SDR improvement (SDRi), PESQ improvement (PESQi), and STOI improvement (STOIi) as the difference of the corresponding metric between the processed audio and the unprocessed audio, respectively. All the evaluation metrics used here are the higher, the better. Note that all the models are only trained once as described in Section III-C, but the models are supposed to be able to process missing gaps shorter than the gap size used in the training process. We can see from Table I that HCRN (Spectro-Temporal loss) achieves the best performance compared to HCRN (Spectrum loss) and the baseline U-Net in all metrics and has the ability to restore the missing gaps. It is worth noting that with the increment of the length of missing gaps, the intelligibility of speech decreases, and it's harder to restore the missing gaps at a lower SNR level. At the SNR level of 5 dB, HCRN (Spectro-Temporal loss) achieves a 35.05% PESQ improvement and a 7.06% STOI improvement over the noisy input when there is no missing gap, and a 55.48% PESQ improvement and an 11.54% improvement over the noisy incomplete input when the missing gaps are 176 ms; while at the SNR level of 15 dB, HCRN (Spectro-Temporal loss) achieves a 29.39% PESQ improvement and a 3.16% STOI improvement over the noisy input when there is no missing gap, and a 63.58% PESQ improvement and an 8.05% improvement over the noisy incomplete input when the missing gaps are 176 ms. These results demonstrate that the improvements are not only due to suppressing the background noise, but also restoring the missing gaps. In other words, our proposed method has the capability of restoring the gaps with the unreliable context information.

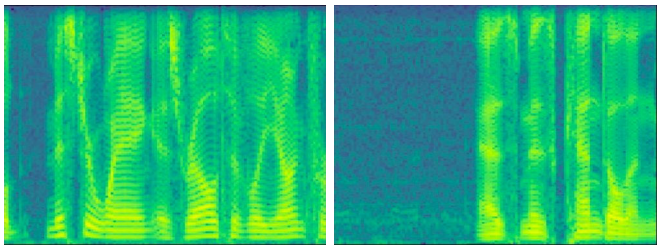
Comparing HCRN (Spectro-Temporal loss) and HCRN (Spectrum loss), the former outperforms the latter in SDRi to a margin. By introducing L_1 loss in the time domain, HCRN (Spectro-Temporal loss) provides 0.78 to 1.67 SDR improvement over HCRN (Spectrum loss), which demonstrates the effectiveness of Spectro-Temporal loss to suppress background noise and improve the SDR metric. PESQ and STOI metrics of HCRN (Spectrum loss) and HCRN (Spectro-Temporal loss) don't make a big difference in most cases.

B. Qualitative Performance

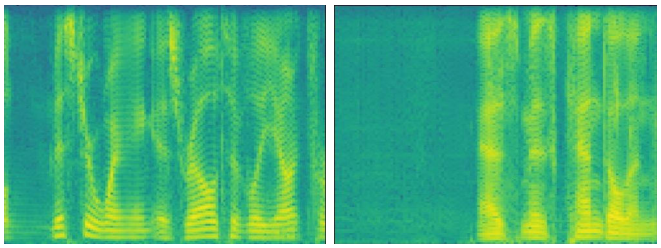
For qualitative performance, spectrograms from two audio clips of two test samples are visualized in Figure 3. Here, the



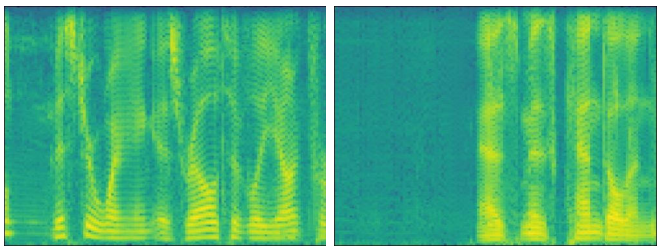
(a) Input



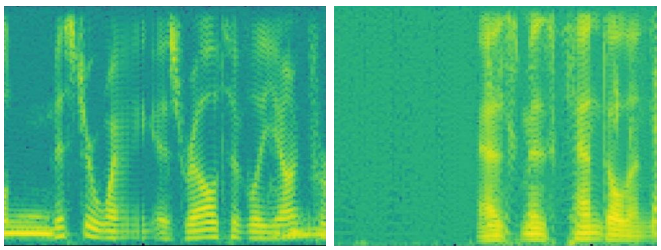
(b) Target



(c) HCRN (Spectro-Temporal loss)



(d) HCRN (Spectrum loss)



(e) U-Net

Fig. 3: Spectrograms of audio clips from test samples mixed at a SNR level of 5dB. The gap size is 176 ms. Each column represents an audio clip from a test sample.

test set is mixed at an SNR level of 5 dB and the consecutive missing gap size is 176 ms. We can see from the first column that HCRN (Spectro-Temporal loss) and HCRN (Spectrum loss) don't show a lot of differences in restoring the missing gap. They both fill in the missing gap naturally and close to the target, at the same time, suppress the noise in the mixture to some extent. However, the spectrogram of HCRN (Spectrum loss) is noisier than that of HCRN (Spectro-Temporal loss), which implies that Spectro-Temporal loss helps the model suppress the background noise. Whereas, the baseline U-Net model does a worse job in both aspects. Interestingly, we can see from the second sample that, under the current training procedure, all the models are capable of only restoring the salient signal and ignoring the lost parts of the unsalient signal.

V. Discussion

A. Differences between Our Work and Previous Research

To the best of our knowledge, we are the first to investigate the task of auditory restoration with large gap size and the intrusion of background noise. The scenarios of our experimental settings are more practical and closer to real-world scenarios compared to previous research in the following four aspects: 1) The missing gaps and the noise exist simultaneously in our mixture signal. And our goal is to jointly reduce the background noise and restore the missing parts of the input depending on the unreliable contents, which is more challenging compared to performing speech enhancement or audio inpainting individually. 2) Besides, the size of the missing gaps is quite large in our experiments, which further poses challenges for restoring the target signal from the noisy mixture input. 3) The positions of the missing gaps are unknown in our experimental settings. 4) The target signal we used to generate the mixture is a speech from multiple speakers, which is more difficult than the sounds of musical instruments or simple notes.

B. Inspiration from Image Inpainting

The task of image inpainting is to recover the lost part of a given image and there is an abundance of research on the task [7]–[9]. Audio inpainting often views T-F spectrograms as images, and therefore it's very natural to get inspiration from the literature on image inpainting. In image inpainting literature, generative adversarial networks [35] are often applied in general image inpainting configurations. However, in our preliminary experiments, the application of GAN doesn't make a difference. It is worth further investigating the similarity and the difference between the image inpainting task and the auditory restoration task and find if other techniques in image inpainting can play a role in auditory restoration. It's hopeful that inspiration from image inspiration may shed light on the research of auditory restoration.

C. Future Work

Since only time gaps are considered in this work, frequency gaps and irregular gaps should be considered in future experiments. What's more, we only focus on the reconstruction of the spectrograms and don't consider phase reconstruction in this work. Though in previous audio inpainting methods, the recovered spectrograms are transformed to waveform using Griffin-Lim algorithm or a WaveNet decoder [36], we don't follow them for the following two reasons: 1) We find in our preliminary experiments that Griffin-Lim algorithm causes great performance degradation for enhanced signals. 2) It is difficult to get a publicly available pre-trained WaveNet decoder for multiple speakers. Meanwhile, it is cumbersome to train such a satisfying WaveNet decoder by ourselves. Therefore, phase reconstruction in spectrogram-based auditory restoration is another problem we plan to investigate in the future.

VI. Conclusion

This work explored the pioneering task of auditory restoration with large gap size and the intrusion of background noise. We propose Hourglass-shaped Convolutional Recurrent Network (HCRN) and Spectro-Temporal loss function to train the model, which helps the model improve the SDR metric. Our experiments give some promising results. Both the quantitative and qualitative performance show that our proposed method can suppress the background noise to some extent, at the same time, identify and restore the missing gaps of the salient signal with the unreliable context information. The model has the ability of processing missing gaps shorter than the gaps used during training. Moreover, with the current training procedure, the proposed model tends to ignore missing gaps of the unsalient signal and doesn't restore them. We hope the application of the auditory restoration task can serve as an extension of modern source separation frameworks, leading to a general-purpose source separation model solving the "Cocktail Party Problem".

VII. Acknowledgments

This work was funded by a grant from the National Key Research and Development Program of China (2018AAA0100400), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB32070000).

References

[1] G. M. Bidelman and C. Patro, "Auditory perceptual restoration and illusory continuity correlates in the human brainstem," *Brain Research*, vol. 1646, pp. 84 – 90, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006899316304127

[2] M. P. Cooke and G. J. Brown, "Computational auditory scene analysis: Exploiting principles of perceived continuity," *Speech Communication*, vol. 13, no. 3-4, pp. 391–399, 1993.

[3] I. Masudakatsuse and H. Kawahara, "Dynamic sound stream formation based on continuity of spectral change," *Speech Communication*, vol. 27, no. 3, pp. 235–259, 1999.

[4] S. Srinivasan and D. Wang, "A schema-based model for phonemic restoration," *Speech Communication*, vol. 45, no. 1, pp. 63–87, 2005.

[5] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[6] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] C. Guillemot and O. Le Meur, "Image inpainting: Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, 2013.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[10] P. Vary, R. Hofmann, K. Hellwig, and R. J. Sluyter, "A regular-pulse excited linear predictive codec," *Speech Communication*, vol. 7, no. 2, pp. 209–215, 1988.

[11] A. Marafioti, N. Holighaus, P. Majdak, N. Perraudin *et al.*, "Audio inpainting of music by means of neural networks," in *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019.

[12] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[13] Y.-L. Chang, K.-Y. Lee, P.-Y. Wu, H.-y. Lee, and W. Hsu, "Deep long audio inpainting," *arXiv preprint arXiv:1911.06476*, 2019.

[14] M. Kegler, P. Beckmann, and M. Cernak, "Deep speech inpainting of time-frequency masks," in *Interspeech 2020*, 2020, pp. 3276–3280. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1532

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[16] P. Beckmann, M. Kegler, H. Saltini, and M. Cernak, "Speech-vgg: A deep feature extractor for speech processing," *arXiv preprint arXiv:1910.09909*, 2019.

[17] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 283–292.

[18] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[19] ——, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2019, pp. 6865–6869.

[20] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[21] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[22] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 436–440.

[23] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1083–1094, 2018.

[24] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 745–751. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/171_Paper.pdf

[28] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast griffin-lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[29] K. J. Piczak, "Esc:dataset for environmental sound classification," *IEEE Transactions on Wireless Communications*, vol. 9, no. 2, pp. 1015–1018, 2015.

[30] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*. IEEE, 2017, pp. 776–780.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *ArXiv*, 06 2014.

[36] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.