

Supplementary Material

Anonymous Author(s)
Submission Id: 1178

A MORE DETAILS OF EXPERIMENTAL SETTINGS

Model settings. The three diffusion-model-based vocoders described in Section 5.1.3 take Gaussian noise as input and mel-spectrogram as the condition. The conversion of mel-spectrogram is performed using the transformation corresponding to each diffusion model. The parameters for converting mel-spectrograms are detailed as follows. Regarding to DiffWave and PriorGrad, the size of Fast Fourier Transform (FFT) was set to 1024. The length of hop between Short Time Fourier Transform (STFT) was 256. We set the number of mel filterbanks and the window size to 80 and 1024, respectively. Furthermore, the length of hop between STFT was modified to 300, while all other parameters remained unchanged for WaveGrad.

Training settings. In the training strategy, our initial emphasis is on refining the accuracy of watermark extraction, followed by the imposition of constraints on the quality of the generated audio. Therefore, we initialized the hyper-parameters τ of audio quality loss to 0. Upon surpassing a predefined recovery accuracy threshold, we reset τ to 1.

B MORE DETAILS OF WATERMARK DECODER

For completeness of Section 4.1, this section offers the detailed configuration of the watermark decoder. The detailed architecture of the watermark decoder is depicted in Fig. 8. It consists of seven MGCNNs followed by two fully connected layers. The output of the first fully connected layer undergoes transformation via a ReLU activation function, resulting in the final output. This final output shares the same length as the binary watermark \mathbf{w} . In particular, each MGCNN is composed of two parallel one-dimensional convolutional layers with kernel size 3×3 , strides of 2, and padding set to 1. The watermark decoder accepts the watermarked audio $\mathbf{x}_T \in \mathbb{R}^\omega$ that generated by DMs, where $\omega = b \times c_h \times l_s$, b is the batch size, c_h represents the channels of the watermarked audio and l_s denotes the input length of DMs.

C FIDELITY AND CAPACITY OF GROOT ON CROSS-LINGUISTIC DATASET

To further evaluate the effectiveness of the proposed Groot, we conducted cross-lingual experiments utilizing the Aishell-3 Chinese audio dataset [34]. This is a multi-speaker dataset sampled at 44.1 kHz. It comprises 88,035 utterances totaling 85 hours, recorded by 218 speakers from different accent regions across China. The experimental configuration is identical to that described in Section 5.3. Similarly, the audio of all datasets is segmented into 1-second utterances and downsampled to a sampling rate of 22.05 kHz. As shown in Table 6, we present experimental results across capacities of 100, 500, 1000, and 2000 bps, comparing the watermarked audio with the generated audio.

Table 6: Fidelity of Groot on Aishell-3.

	Capacity (bps)			
	100	500	1000	2000
STOI \uparrow	0.9132	0.9131	0.9150	0.9131
MOSL \uparrow	3.0386	3.0378	3.0636	3.0379
SSIM \uparrow	0.7880	0.7880	0.7885	0.7880
MCD \downarrow	1.1772	1.1791	1.1747	1.1779
ACC \uparrow	0.9840	0.9865	0.9856	0.9848

The experimental results clearly demonstrate that the proposed method maintains relatively strong fidelity performance on the Chinese dataset. Despite exhibiting suboptimal performance in terms of the SSIM metric, the STOI and MOSL consistently hold steady at 0.91 and 3.03 across various capacities. Furthermore, the watermark recovery accuracy remains high at 98.5% for all tested capacities. Regarding capacity, as the watermark length increases, our method experiences only minor degradation. At the highest tested capacity, the audio quality remains nearly indistinguishable from that at the lowest capacity, achieving an accuracy rate of 98.48%.

D MORE DETAILS OF ROBUSTNESS

D.1 Individual Attacks

As outlined in Section 5.4.1, this section will delve into the details of the six speech post-processing operations used in the robustness experiments.

Gaussian Noise. Gaussian noise with intensities of 5, 10, and 20 dB is added to the watermark audio.

Low-pass Filtering. The watermarked audio is subjected to low-pass filtering with a cutoff frequency of 3 kHz.

Band-pass Filtering. The watermarked audio undergoes band-pass filtering with a passband ranging from 0.3 to 8 kHz.

Stretching. The watermarked audio is time-stretched to double its original duration and then compressed back to its initial length through interpolation.

Cropping. The watermarked audio is truncated to half its original length, with the first and second halves separately clipped.

Echo. An echo effect is applied to the watermarked audio, introducing a delayed and attenuated replica of the original waveform.

The *Band-pass filtering* and *Echo* were obtained through the open-source code of *AudioSeal* [33], while the remaining post-processing operations were sourced from torchaudio 2.0.1.

D.2 Compound Attacks

For the composite attacks mentioned in section 5.4.2, the specific details are as follows:

Low-pass filtering + Gaussian noise. We first pass the watermark audio through a low-pass filter with a threshold of 3kHz, then add Gaussian noise with an intensity of 10dB.

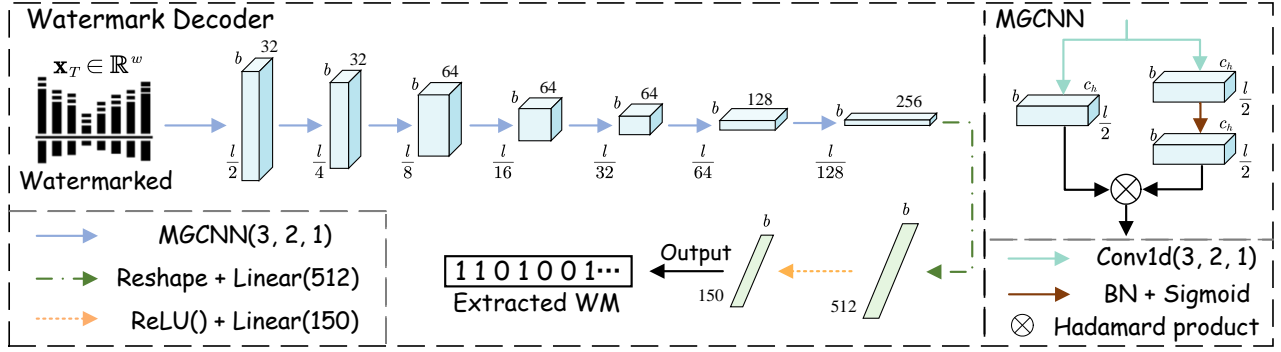


Figure 8: The detailed architecture of watermark decoder.

Band-pass filtering + Echo. We first pass the watermark audio through a band-pass filter with a threshold of 0.3 to 8kHz, then apply an echo effect.

Cropping + Stretch. We first crop the first half of the watermark audio signal, then stretch it to twice its original length.

Gaussian noise + Echo. Gaussian noise of 10dB is added to the watermark audio, followed by applying an echo effect.

Gaussian noise + Band-pass filtering. Similarly, 10dB Gaussian noise is added to the watermark audio and then passed through a band-pass filter with a threshold of 0.3 to 8kHz.

Gaussian noise + Band-pass filtering + Echo. Gaussian noise of 10dB intensity is applied to the watermarked audio, followed by band-pass filtering and echo addition.

E MORE EXPERIMENTS OF ROBUSTNESS

Robustness experiments evaluating the effectiveness of Groot at 100 bps capacity were also conducted on multi-speaker *LibriTTS*, *LibriSpeech* and *Aishell-3* datasets. We also validated the robustness of the multi-speaker datasets against individual attacks and compound attacks. Experiments and corresponding analysis are detailed as follows.

E.1 Robustness against Individual Attacks

The audio post-processing operations described in section 5.4.1 were similarly employed to conduct experiments on the multi-speaker datasets. Table 7 presents experimental results, with the results derived from comparisons between the *attacked audio* and the unattacked watermarked audio.

The experimental results clearly indicate that Groot exhibits strong robustness across various multi-speaker datasets, especially in maintaining high watermark recovery accuracy in the presence of Gaussian noise. Even at a noise level of 5 dB, the three datasets achieve accuracies of 98.17%, 97.43%, and 96.82%, respectively. In the case of *LibriTTS*, even after undergoing low-pass and band-pass filtering, the accuracy remains virtually unchanged at 99.22% and 99.50%. In addition, Groot exhibits exceptional balance performance on this dataset, boasting an average accuracy of 98.86%. After applying band-pass filtering and stretching on *LibriSpeech*, the accuracy reaches 99.36% and 99.39%, respectively. Furthermore, the average accuracy of 98.08% showcases a balanced level of robustness. Although the robustness on *Aishell-3* is relatively lower compared

to the other datasets, potentially due to its Chinese language nature, Groot maintains a respectable average accuracy of 95.51%, illustrating robustness on this dataset as well.

E.2 Robustness against Compound Attacks

In accordance with the description provided in Section 5.4.2 regarding compound attacks, we conducted further robustness validation on multi-speaker datasets using these composite attacks. Table 8 compares the results of attacked audio with those of the unattacked audio. The specific attack combinations are detailed as follows. 1) low-pass filtering succeeded by Gaussian noise, 2) band-pass filtering followed by an echo attack, 3) cropping and subsequently stretching, 4) Gaussian noise followed by echo, 5) Gaussian noise coupled with band-pass filtering and, 6) Gaussian noise succeeded by band-pass filtering coupled with echo.

Especially concerning *LibriTTS*, when facing compound attack 1, the watermark extraction accuracy achieves 98.77%. Even when encountering compound attack 6, comprising three attacks, the accuracy remains stable at 97.42%. Although the performance on *LibriSpeech* slightly lags behind *LibriTTS*, the accuracy reaches 94.64% when resisting compound attack 6. Moreover, after undergoing compound attacks 3 and 5, the accuracy maintains levels of 98.68% and 98.57%, respectively. *Aishell-3*, on the other hand, shows a marginal decline in performance against individual attacks, with accuracy falling to 89.78% after compound attack 6. Nevertheless, it sustains accuracy levels of 96.94% and 96.12% against compound attacks 1 and 5, respectively. Furthermore, the average recovery accuracy across the three datasets is 98.10%, 96.68%, and 93.45%, respectively, effectively demonstrating balanced robustness of Groot against compound attacks.

F VISUALIZATION

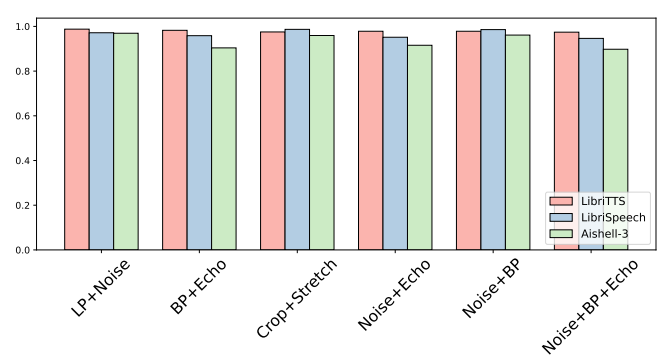
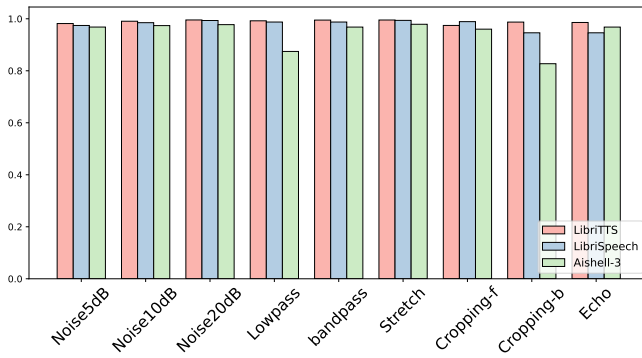
To enhance understanding of the fidelity of watermark audio synthesized by Groot, we present visualizations of the generated audio, watermark audio, and their corresponding residual, illustrated from Fig. 11 to Fig. 14. These displayed audio samples are randomly chosen from the *LJSpeech*, *LibriTTS*, *LibriSpeech*, and *Aishell-3* datasets, operating at a capacity of 100 bps. Darker colors in the residual indicate larger deviations from the generated audio. The visualization demonstrates that Groot effectively disperses watermark features with minimal impact on audio quality, thereby ensuring its integrity.

Table 7: Comparison of Robustness Against Individual Attacks on Multi-speaker Datasets.

Dataset		Noise			LP-F	BP-F	Stretch	Cropping		Echo
		5 dB	10 dB	20 dB	3k	0.3-8k	2	front	behind	default
LibriTTS	STOI↑	0.7927	0.8742	0.9620	0.9931	0.8601	0.9937	0.8899	0.9113	0.9224
	MOSL↑	1.0619	1.1432	1.7058	4.6173	3.3331	3.9927	1.1051	1.2243	1.2607
	ACC↑	0.9817	0.9906	0.9955	0.9922	0.9950	0.9953	0.9744	0.9873	0.9858
LibriSpeech	STOI↑	0.8357	0.9997	0.9244	0.9905	0.8575	0.9915	0.9511	0.3428	0.8334
	MOSL↑	1.1105	1.2427	1.9152	4.6164	3.3653	4.1850	1.4262	1.1110	1.1911
	ACC↑	0.9743	0.9850	0.9934	0.9875	0.9936	0.9939	0.9890	0.9461	0.9642
Aishell-3	STOI↑	0.9132	0.6790	0.7647	0.8869	0.9913	0.8245	0.9824	0.2302	0.8245
	MOSL↑	1.1012	1.2276	2.0128	4.6220	3.4356	4.1528	1.4909	1.0945	3.4356
	ACC↑	0.9682	0.9738	0.9775	0.9744	0.9681	0.9789	0.9601	0.8272	0.9681

Table 8: Comparison of Robustness Against Compound Attacks on Multi-speaker Datasets.

Dataset	Lowpass+Noise			Bandpass+Echo			Cropping+Stretch		
	STOI↑	MOSL↑	ACC↑	STOI↑	MOSL↑	ACC↑	STOI↑	MOSL↑	ACC↑
LibriTTS	0.8539	1.1235	0.9877	0.8017	1.2523	0.9825	0.8859	1.1054	0.9752
LibriSpeech	0.8164	1.2135	0.9716	0.7500	1.2015	0.9585	0.9467	1.4243	0.9868
Aishell-3	0.7422	1.2054	0.9694	0.7211	1.1874	0.9039	0.9810	1.4923	0.9593
Dataset	Noise+Echo			Noise+Bandpass			Noise+Bandpass+Echo		
	STOI↑	MOSL↑	ACC↑	STOI↑	MOSL↑	ACC↑	STOI↑	MOSL↑	ACC↑
LibriTTS	0.8419	1.0705	0.9783	0.7630	1.1573	0.9883	0.7310	1.0784	0.9742
LibriSpeech	0.6877	1.0885	0.9515	0.7423	1.2542	0.9857	0.6177	1.0986	0.9464
Aishell-3	0.6074	1.0801	0.9156	0.6690	1.2159	0.9612	0.5542	1.0850	0.8978

**Figure 9: Visualization of Groot Against Individual Attacks. Figure 10: Visualization of Groot Against Compound Attacks.**

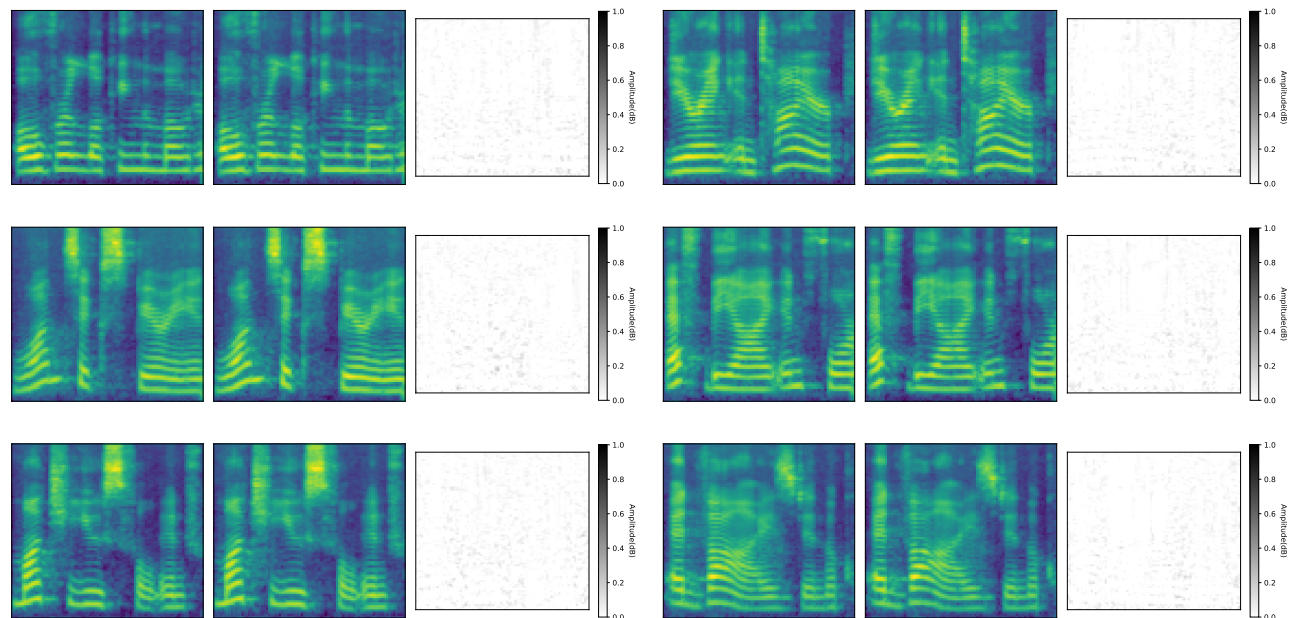


Figure 11: Fidelity Visualization of Groot on LJSpeech.

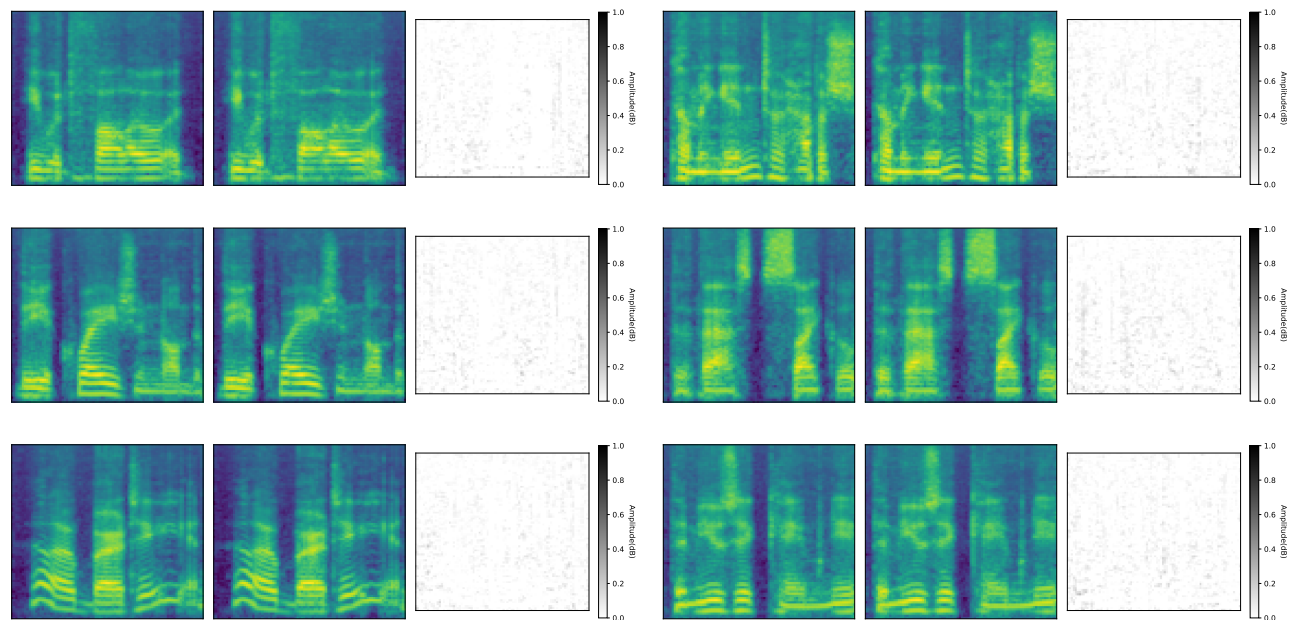


Figure 12: Fidelity Visualization of Groot on LibriTTS.

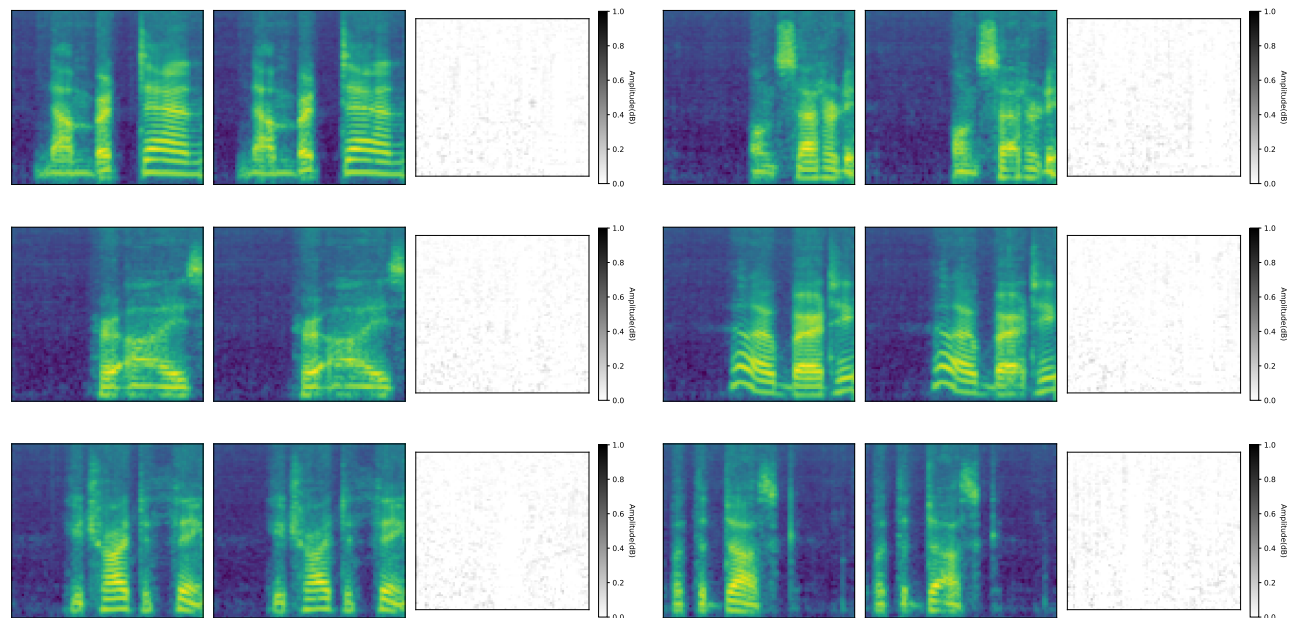


Figure 13: Fidelity Visualization of Groot on LibriSpeech.

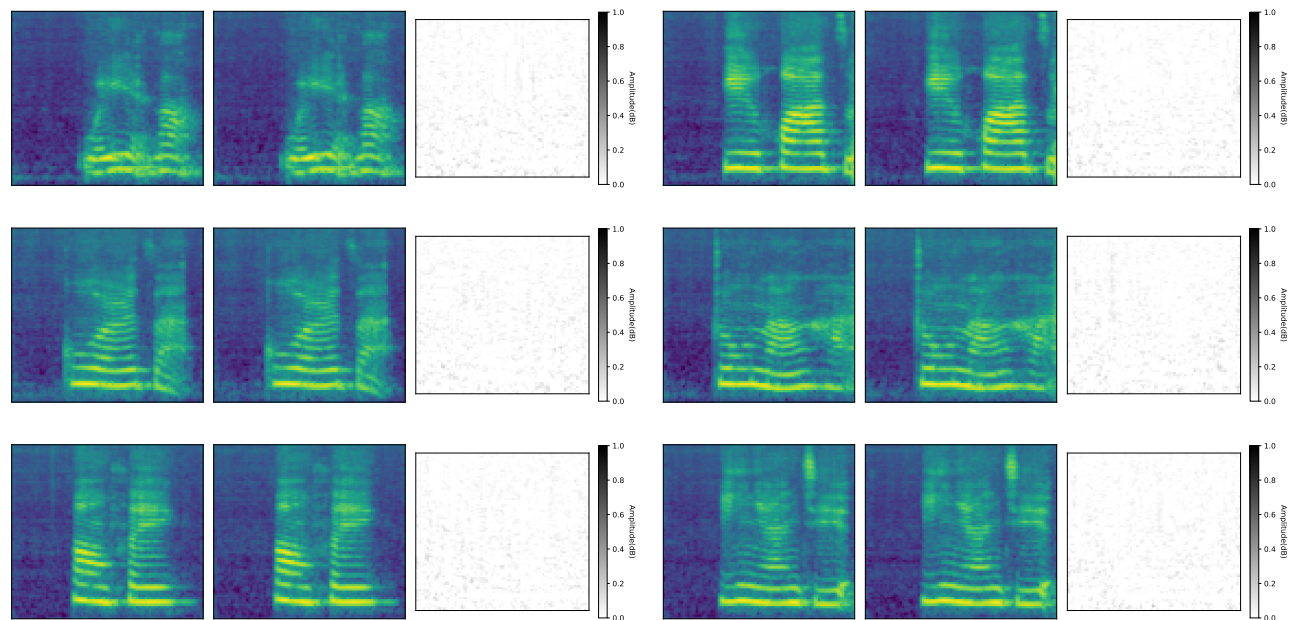


Figure 14: Fidelity Visualization of Groot on Aishell-3.