EchoMark: Perceptual Acoustic Environment Transfer with Watermark-Embedded Room Impulse Response

Core Implementation

In EchoMark, models comprise the RIR Encoder, RIR Generator, and Watermark Detector, where the first two jointly take a target reverberant and message information as input, and the last one is later used in the potential watermarked environment-transferred audio detection. In the following tables, we list all model hyperparameters and configurations for the three modules for reproduction reference. Note that these parameters may be further optimized for better accuracy or efficiency.

Table 1: Configuration of the RIR Encoder Module

Submodule	Hyperparameter	Value
Spectrogram Transform	FFT size	1024
	Window length	1024
	Hop length	256
Conformer Encoder	Input feature dimension	513
	Model hidden size	256
	Feedforward layer size	1024
	Number of attention heads	4
	Number of layers	12
	Convolution kernel size	31
Attentive Pooling	Projection output dimension	128
	Pooling mechanism	Attn

Table 2: Configuration of the RIR Generator Module

Submodule	Hyperparameter	Value
Watermark Embedder	Message length	5 bits
	Dimension	128
	Network	MLP(Swish+LN)
Early RIR Generator	Pre-conv	Conv1d (128, 100, stride=1)
	HiFi-GAN	3 ResBlocks, 4 upsample stages
Late RIR Generator (Decor style)	Noise-shaping filter	7 octaves in 32.5–8000 Hz
	Backbone network	3-layer MLP(Swish+GN(8))
	Amplitude head	MLP(512→200→7×20)
	Gate head	MLP(512→200→7×20, Sigmoid)
	Number of decays	20

Optional Dereveberation Network

As mentioned in the main submission, a byproduct of **RIR Encoder** is we can train a time-frequency mask for denoising and dereverberation, which is also useful if the source speech for AEM is not clean or dry. Although it is not the main contribution of this paper, this module completes the overall EchoMark design. Please refer to our code for implementation details. The overall result with dereveberation is shown in spectrogram, see Fig. 1.

Table 3: Configuration of the Watermark Detector Module

Submodule	Hyperparameter	Value
RIR Encoder	Conformer+Attn	Same as (Table 1)
Detector	Layer 1	MLP(Swish+LN+Dropout)
	Layer 2	Linear(hidden→6)
	Hidden dimension	128
	Output dimension	6 (5-bit message + 1 presence)

Training Process

In our code, we include the dereverberation network in the joint training, and it can be removed if clean source input can be readily captured. Although the overall training can be long and the loss gradually decreases, early stop when overall loss is around 6.1 yields acceptable perceptual quality and sufficient watermark detection and decoding accuracy.

Dataset Information

To train and evaluate EchoMark, we utilize a diverse set of real-world RIR datasets, shown in Tab 4. These datasets collectively ensure a broad coverage of acoustic variability for robust model development.

Table 4: Summary of RIR Datasets Used

Dataset	RIR Count	Environment
BUT Reverb Database	1300+	8 rooms
REVERB Challenge	24	Small to large rooms
Aachen Impulse Response	344	5 sites incl. church
RWCP Sound Scene	143	14 rooms

Code Reproduction

The supplementary material includes training and inference code. Due to file size limits and conference policy, pretrained models will be shared after the review process if permitted. For reproduction, use `train.py` or `train_ddp.py` for GPU training. We also provide watermarked audio samples with ground-truth messages from a subset of the listening test. Since model behavior may vary, please use `inference.ipynb` to generate and decode audio with your own trained model.