Bat.ai

Yash Singh Pathania 24204265
Abhik Sarkar 24214927

Purpose

In the era of AI, there's a lot of AI-generated voice content out there. The purpose of BAT and MOTH is to detect and watermark that content in a way that remains hidden.

Literature Review

GENERATING COHERENT DRUM ACCOMPANIMENT WITH FILLS AND IMPROVISATIONS

- **Transformer-based embedding:** Inspired by transformer models generating drum patterns, we embed audio watermarks into carrier tracks to ensure seamless integration.
- **Novelty function for subtlety:** Using a novelty function concept, we minimize watermark detectability by aligning it with the carrier audio’s natural patterns.
- **In-filling for coherence:** A BERT-inspired in-filling approach hides watermarks in complex audio segments, preserving perceptual quality.

Literature Review

NEURAL DRUM ACCOMPANIMENT GENERATION

Innovative Transformer-Based Model:
Developed a transformer encoder model to generate symbolic drum patterns conditioned on melodic tracks (Piano, Guitar, Strings, Bass), addressing the underexplored challenge of drum accompaniment generation with a focus on coherence and diversity.

Scalable Data Representation:
Proposed a novel data representation scheme that incorporates silences and scales to any number of instruments, using 64-dimensional vectors for melody and 17-dimensional vectors for drums, enabling efficient processing of the Lakh Pianoroll dataset.

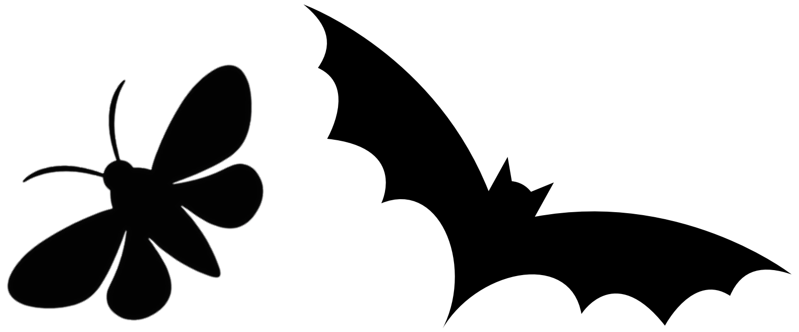
Musically Relevant Evaluation:
Achieved superior performance over benchmarks using two metrics: Polyphony Correlation (PC) and Bar Rhythm Density Correlation (BRDC), demonstrating the model's ability to partially learn drum patterns, fills, and improvisations from melodic inputs.

Literature Review

HEAR ME IF YOU CAN! - Project Back Bone [C]

Inspirations

- Their imperceptible data hiding motivated our focus on audio quality.
- Use of ML for encoding/decoding shaped our ML-driven watermark models.
- Their secure, robust approach guided our encryption and durability goals.
- We adopted their signal processing techniques for watermark embedding.
- Their PESQ-based evaluation inspired our SNR and listening tests.
- Extended their ideas to AI audio classification and conditional removal.
- Aimed for broader applications like copyright and platform integration.



Moth

The Moth encoder, inspired by Shah et al.'s "Hear Me If You Can" [C], embeds perturbations in audio clips, akin to a moth sending out a voice detected by a bat to reveal a watermark's presence, signaling AI-generated audio. This classic encoder ensures imperceptible, robust watermarking through three salient features, supported by preprocessing via time-frequency analysis for seamless integration:

- **Loss Functions:** Minimize distortion during watermark embedding using SNR-based optimization.
- **Encoder-Decoder Training:** Leverage ML models for reliable watermark insertion and detection.
- **Perturbation and Amplitude:** Create adaptive, durable watermarks detectable in 2-second segments.

Bat

The Moth encoder, inspired by Shah et al.'s "Hear Me If You Can" [C], embeds perturbations in audio clips, akin to a moth sending out a voice detected by a bat to reveal a watermark's presence, signaling AI-generated audio. This classic encoder ensures imperceptible, robust watermarking through three salient features, supported by preprocessing via time-frequency analysis for seamless integration:

- **Loss Functions:** Minimize distortion during watermark embedding using SNR-based optimization.
- **Encoder-Decoder Training:** Leverage ML models for reliable watermark insertion and detection.
- **Perturbation and Amplitude:** Create adaptive, durable watermarks detectable in 2-second segments.

Loss Functions

- **Audio steganography conceals watermarks within audio, preserving pristine sound quality.**
- Advanced loss functions drive imperceptible watermarking, targeting superior PESQ scores.
- These functions harmonize human auditory perception with robust watermark detection.
- Explored methods—MSE, Spectrogram, Log-Mel, and Psychoacoustic—offer tailored solutions.
- We have implemented this in MothEncoder with flexible loss selection for testing.

Let’s explore these loss functions in detail.

Loss Functions

- **Audio steganography conceals watermarks within audio, preserving pristine sound quality.**
- Advanced loss functions drive imperceptible watermarking, targeting superior PESQ scores.
- These functions harmonize human auditory perception with robust watermark detection.
- Explored methods—MSE, Spectrogram, Log-Mel, and Psychoacoustic—offer tailored solutions.
- We have implemented this in MothEncoder with flexible loss selection for testing.

Let’s explore these loss functions in detail.

Mean Squared Error (MSE) Loss

- Calculates precise sample-wise differences between original and watermarked audio.
- Offers computational simplicity, ideal for baseline watermarking performance.

Disadvantages

- Lacks alignment with human auditory perception, limiting PESQ optimization.
- Overly sensitive to phase shifts, penalizing inaudible waveform deviations.
- Serves as a foundational approach, effective yet perceptually constrained.

Mean Squared Error (MSE) Loss

- Calculates precise sample-wise differences between original and watermarked audio.
- Offers computational simplicity, ideal for baseline watermarking performance.

Disadvantages

- Lacks alignment with human auditory perception, limiting PESQ optimization.
- Overly sensitive to phase shifts, penalizing inaudible waveform deviations.
- Serves as a foundational approach, effective yet perceptually constrained.

Spectrogram Based Loss

- Harnesses Short-Time Fourier Transform to analyze time-frequency disparities.
- Captures audible spectral shifts, surpassing MSE in perceptual relevance.
- Enhances PESQ scores by prioritizing frequency content over raw samples.
- Incurs moderate computational cost due to STFT processing demands.
- Provides a robust bridge to more advanced perceptual loss strategies.

Disadvantages

- Humans perceive frequency logarithmically.
- Vanilla Spectrograms use linear frequency bins, which are misaligned with human auditory system.
- It also lacks loudness scaling, limiting sensitivity to perceptual nuances.

Spectrogram Based Loss

- Harnesses Short-Time Fourier Transform to analyze time-frequency disparities.
- Captures audible spectral shifts, surpassing MSE in perceptual relevance.
- Enhances PESQ scores by prioritizing frequency content over raw samples.
- Incurs moderate computational cost due to STFT processing demands.
- Provides a robust bridge to more advanced perceptual loss strategies.

Disadvantages
- Humans perceive frequency logarithmically.
- Vanilla Spectrograms use linear frequency bins, which are misaligned with human auditory system.
- It also lacks loudness scaling, limiting sensitivity to perceptual nuances.

Log-Mel-Spectrogram Loss

- Employs mel-scale spectrograms, mirroring human auditory frequency perception.
- Logarithmic scaling aligns with natural loudness sensitivity, boosting PESQ accuracy.
- Utilizes L1 loss for resilient optimization, minimizing distortion impacts.
- Streamlines computation by reducing frequency bins via mel transformation.
- Optimal choice for efficient, high-quality watermark concealment.
- We have used 128 mel bands for our use case.

Disadvantages

- It doesn't model auditory masking, where loud sounds hide quieter ones in nearby frequencies, missing opportunities to strategically place the watermark.
- Reducing to 128 mel bands (from 1025 STFT bins) loses some frequency detail, which might affect precision in complex audio

Log-Mel-Spectrogram Loss

- Employs mel-scale spectrograms, mirroring human auditory frequency perception.
- Logarithmic scaling aligns with natural loudness sensitivity, boosting PESQ accuracy.
- Utilizes L1 loss for resilient optimization, minimizing distortion impacts.
- Streamlines computation by reducing frequency bins via mel transformation.
- Optimal choice for efficient, high-quality watermark concealment.
- We have used 128 mel bands for our use case.

Disadvantages

- It doesn’t model auditory masking, where loud sounds hide quieter ones in nearby frequencies, missing opportunities to strategically place the watermark.
- Reducing to 128 mel bands (from 1025 STFT bins) loses some frequency detail, which might affect precision in complex audio

Psychoacoustic Model Based Loss

- Integrates auditory masking and Bark scale for unparalleled perceptual fidelity.
- Conceals watermarks in frequency regions obscured by dominant audio components.
- Leverages Power Spectral Density and spreading functions to set masking thresholds.
- Maximizes PESQ by ensuring watermark audibility remains negligible.
- Complex but transformative, ideal for cutting-edge steganography applications.

Psychoacoustic Model Based Loss

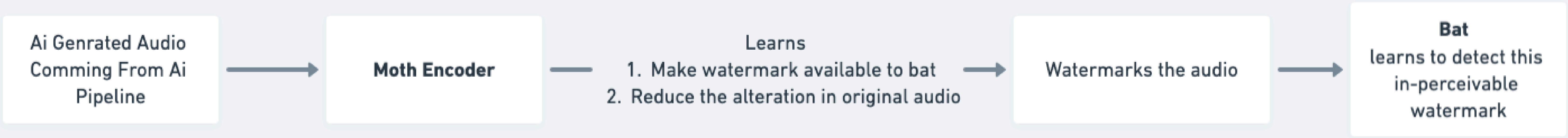
- Integrates auditory masking and Bark scale for unparalleled perceptual fidelity.
- Conceals watermarks in frequency regions obscured by dominant audio components.
- Leverages Power Spectral Density and spreading functions to set masking thresholds.
- Maximizes PESQ by ensuring watermark audibility remains negligible.
- Complex but transformative, ideal for cutting-edge steganography applications.

Results

- Through our experiments, we found that using Log-Mel-Spectrogram Loss results in audio that carries an embedded watermark which is nearly imperceptible to the human ear.
- The PESQ scores for various loss functions are as follows:

LOSS FUNCTION	AUDIO	PESQ
N/A	ORIGINAL x ORIGINAL	4.64
MSE	ORIGINAL x WATERMARKED	2.28
SPECTROGRAM	ORIGINAL x WATERMARKED	3.69
MEL-SPECTROGRAM BASED LOSS	ORIGINAL x WATERMARKED	4.25
PSYCHOACOUSTIC LOSS	ORIGINAL x WATERMARKED	4.37

Demo



Made with Whimsical