A Summary of :

Speech Watermarking with AudioSeal: A Comprehensive Analysis (An ICML 2024 paper)

Localized Zero-Bit/Attribution Watermarking

Presented by :

Anuj Rajan Lalla - B22AI061
Siddhesh Ayyathan - B22CS016

GitHub :
https://github.com/anuj-l22/Speech_Understanding_PA1

What Is Speech Watermarking?

Speech Watermarking

- The process of embedding an inaudible signal (the “watermark”) into an audio waveform.
- Allows identification or detection of AI-generated speech later on.

Proactive Approach

- Watermark is inserted at generation time (e.g., in a TTS pipeline).
- Detector can confirm if audio has a watermark, even after edits.

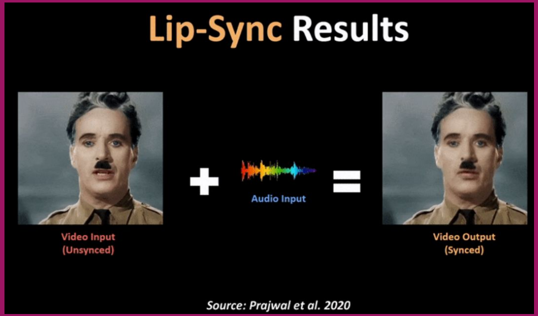
Common Goals

- Imperceptibility: No noticeable change in quality.
- Robustness: Survives noise, compression, etc.
- Localization: Identifies the exact part of the audio that is AI-generated (if partial).

Task Definition and Importance

Why Speech Watermarking?

Deepfake Era



- Rapid advances in TTS/voice cloning (Voicebox, VALL-E).
- High risk of misinformation, impersonation, voice fraud.

Regulatory Need

- Proposed AI laws (EU AI Act) mandate labeling AI-generated content.
- Watermarking can proactively tag synthetic speech..

Real-World Use Cases

- Quick authenticity checks on social media.
- Attribution of content to specific users or APIs.

Comparing SOTA Approaches

Passive Classifiers

- Examples: Voicebox classifier, or any classifier trained on real vs. AI samples.
- Strengths:
  - Straightforward, no modification to TTS pipeline
  - Quick to deploy if you have labeled data
- Limitations
  - Often fails when audio is re-synthesized (fewer artifacts)
  - Model-specific: can become outdated as TTS quality improves

.

Traditional Watermarking (Time/Frequency Domain)

- Approach: Embed bits by slightly modifying amplitude/phase in time domain or DCT/FFT domain
- Strengths:
  - Straightforward, no modification to TTS pipeline
  - Quick to deploy if you have labeled data
- Limitations
  - Often fails when audio is re-synthesized (fewer artifacts)
  - Model-specific: can become outdated as TTS quality improves

.

Comparing SOTA Approaches

Deep-Learning Data Hiding (e.g., WavMark)

- Approach: Uses invertible networks or autoencoders to hide multi-bit messages within short audio frames (1s)
- Strengths:
  - Potentially large capacity (many bits per second)
  - Can adapt to diverse audio domains
- Limitations
  - Slow detection (requires synchronization checks at each step)
  - Chunk-based embedding → coarser detection resolution

AudioSeal

- Approach: Generator–Detector architecture with sample-level detection
  Psychoacoustic losses for imperceptibility
- Strengths:
  - Localized detection (pinpoint short segments)
  - Single-pass, fast detection
  - High imperceptibility and robust to edits
- Limitations
  - Detector must remain private to avoid adversarial removal
  - Specialized training needed (less "plug-and-play")

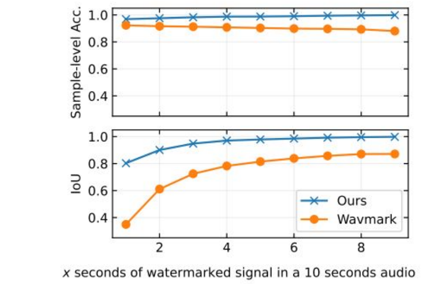


Figure 5. **Localization results** across different durations of watermarked audio signals in terms of Sample-Level Accuracy and Intersection Over Union (IoU) metrics (↑ is better).

Our's means AudioSeal method

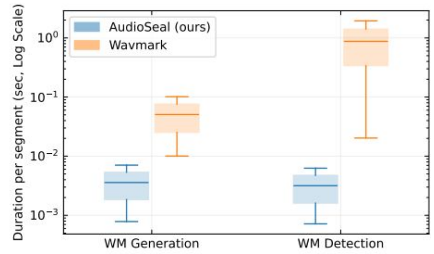


Figure 6. **Mean runtime** (↓ is better). AudioSeal is one order of magnitude faster for watermark generation and two orders of magnitude faster for watermark detection for the same audio input. See Appendix C.1 for full comparison.

AudioSeal: Generator–Detector Architecture

Generator (G)

- Encodes original audio
- Adds an imperceptible watermark δ
- Final speech: $sw = s + \delta$

Detector (D)

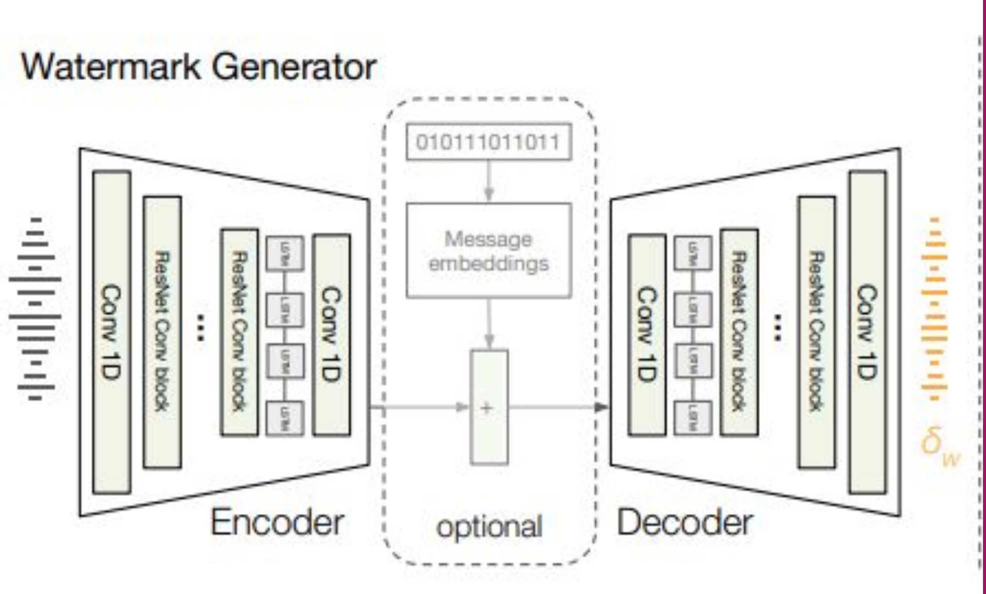
- Outputs a sample-level probability of watermark presence
- (Optional) Decodes bits for attribution.

Key Innovation

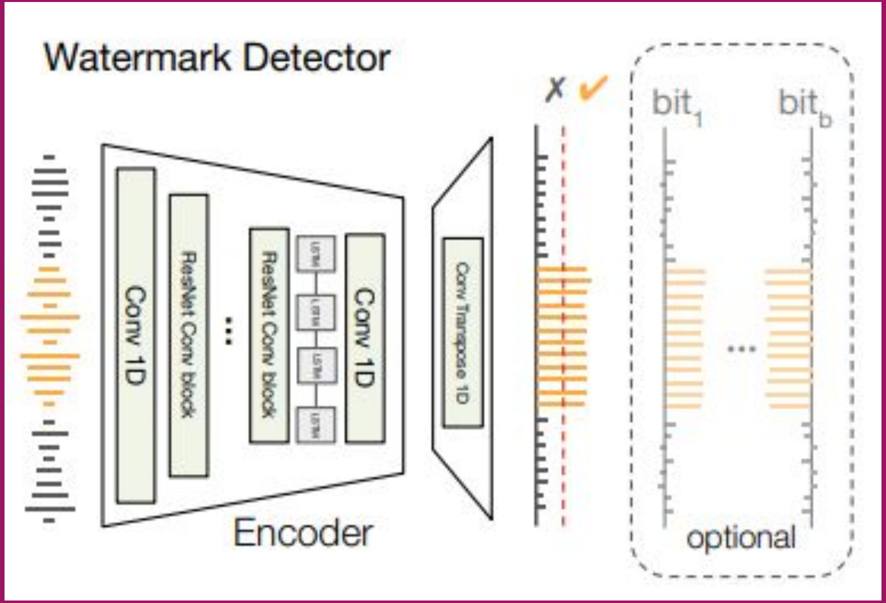
- Localized detection, high imperceptibility using psychoacoustic cues

AudioSeal: Generator–Detector Architecture

Generator (G)



Detector (D)



Psychoacoustic Band-Splitting and Perceptual Losses

Band-Splitting

- Splits audio into frequency subbands to exploit auditory masking.
- Penalizes watermark more in low-energy (highly audible) regions..

.

Multi-Scale Spectral Loss

- Compares Mel-spectra of sw vs. s at multiple scales.
- Preserves speech formants.

.

$\ell 1$ Loss on δ

- Minimizes watermark amplitude
- Minimizes signal distortion

.

Adversarial Loss

- A small discriminator tries to detect the watermark artifacts
- Generator “fools” it, improving realism

.

AudioSeal Training Process

Differentiable Augmentations

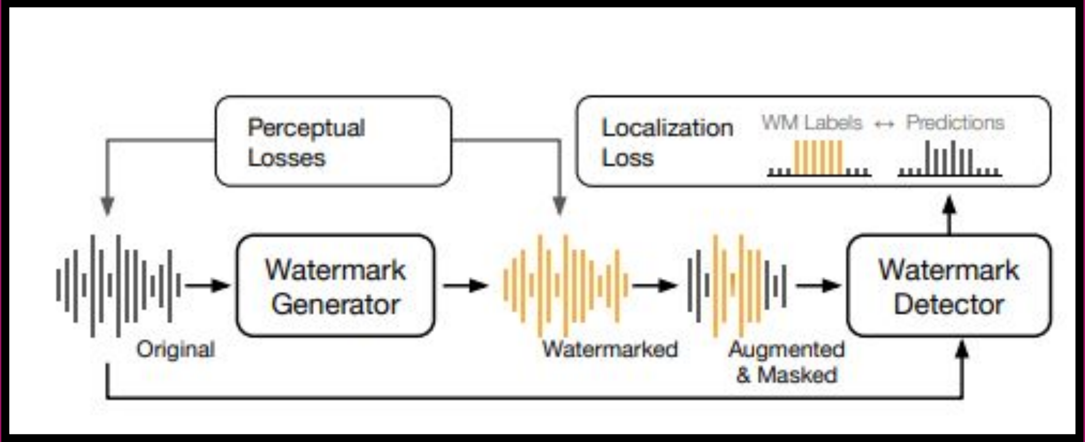
- Noise, time-stretch, partial masking, compression
- Teaches the generator to embed a robust watermark

Localization Loss

- Detector uses BCE at the sample-level to identify watermark presence

Overall Optimization

- $L = L_{perceptual} + \lambda L_{localization}$



Datasets and Code Structure

Datasets

- NPTEL (Indian English): Lectures with varied Indian accents
- Hindi_test (OpenSLR): Cross-lingual test for Hindi

Code Organization

- Main Evaluation:
  1. Extract WAV,
  2. Embed watermark,
  3. Random mask,
  4. Compute metrics (PESQ, STOI, SI-SNR, Detector Score)

- Visualization and Playback : Waveform plotting, interactive audio

Key Findings (Paper Tables)

High Perceptual Scores

- PESQ ~4.47, STOI ~0.997, MUSHRA ~77
- Beats WavMark's ~4.30 PESQ, MUSHRA ~71

Table 1. **Audio quality metrics**. Compared to traditional watermarking methods that minimize the SNR like WavMark, AudioSeal achieves same or better perceptual quality.

Methods	SI-SNR	PESQ	STOI	ViSQOL	MUSHRA
WavMark	**38.25**	4.302	0.997	4.730	71.52 ± 7.18
AudioSeal	26.00	**4.470**	0.997	**4.829**	$\mathbf{77.07} \pm 6.35$

Robustness

- TPR/FPR near 1.0 / 0.0 across edits (AUC ~0.97)
- Passive classifier fails on re-synth but AudioSeal remains perfect

Attribution

- Up to 1,000 versions with better accuracy than WavMark
- Slightly higher false attribution rate at large scale

Key Findings (Paper Tables)

Table 2. **Comparison with Voicebox binary classifier.** Percentage refers to the fraction of masked input frames.

% Mask	AudioSeal (Ours) Acc.	TPR	FPR	Voicebox Classif. Acc.	TPR	FPR
Original audio vs AI-generated audio						
30%	1.0	1.0	0.0	1.0	1.0	0.0
50%	1.0	1.0	0.0	1.0	1.0	0.0
90%	1.0	1.0	0.0	1.0	1.0	0.0
Re-synthesized audio vs AI-generated audio						
30%	**1.0**	**1.0**	**0.0**	0.704	0.680	0.194
50%	**1.0**	**1.0**	**0.0**	0.809	0.831	0.170
90%	**1.0**	**1.0**	**0.0**	0.907	0.942	0.112

Table 3. **Detection results** for different edits applied before detection. Acc. (TPR/FPR) is the accuracy (and TPR/FPR) obtained for the threshold that gives best accuracy on a balanced set of augmented samples. AUC is the area under the ROC curve.

Edit	AudioSeal (Ours) Acc. TPR/FPR	AUC	WavMark Acc. TPR/FPR	AUC
None	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
Bandpass	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
Highpass	0.61 0.82/0.60	0.61	**1.00 1.00/0.00**	**1.00**
Lowpass	**0.99 0.99/0.00**	**0.99**	0.50 1.00/1.00	0.50
Boost	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
Duck	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
Echo	**1.00 1.00/0.00**	**1.00**	0.93 0.89/0.03	0.98
Pink	**1.00 1.00/0.00**	**1.00**	0.88 0.81/0.05	0.93
White	**0.91 0.86/0.04**	**0.95**	0.50 0.54/0.54	0.50
Fast (1.25x)	**0.99 0.99/0.00**	**1.00**	0.50 0.01/0.00	0.15
Smooth	**0.99 0.99/0.00**	1.00	0.94 0.93/0.04	0.98
Resample	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
AAC	1.00 1.00/0.00	1.00	1.00 1.00/0.00	1.00
MP3	**1.00 1.00/0.00**	**1.00**	1.00 0.99/0.00	0.99
EnCodec	**0.98 0.98/0.01**	**1.00**	0.51 0.52/0.50	0.50
Average	**0.96 0.98/0.04**	**0.97**	0.85 0.85/0.14	0.84

NPTEL + Hindi Experiments

NPTEL

- Watermarked: PESQ ~4.44, STOI=0.998, Detector=1.0
- Masked: Detector~0.686 → partial removal only

```
===== Average Metrics Across Samples =====
Watermarked Audio Metrics:
  Average SI-SNR: 28.32 dB
  Average PESQ: 4.45
  Average STOI: 0.998
  Average Detector Score: 1.000
Masked Audio Metrics:
  Average SI-SNR: 30.25 dB
  Average PESQ: 4.49
  Average STOI: 0.999
  Average Detector Score: 0.682
  Average Generation Time: 14192.84 ms
  Average Detection Time: 596.25 ms
  Average Attribution Hamming Distance: 0.000
```

Hindi_test

- Watermarked: PESQ=4.45, STOI=0.998, Detector=1.0
- Masked: Detector=0.682

```
===== Average Metrics Across Samples =====
Watermarked Audio Metrics:
  Average SI-SNR: 27.91 dB
  Average PESQ: 4.44
  Average STOI: 0.998
  Average Detector Score: 1.000
Masked Audio Metrics:
  Average SI-SNR: 30.03 dB
  Average PESQ: 4.49
  Average STOI: 0.999
  Average Detector Score: 0.686
  Average Attribution Hamming Distance: 0.000
```

Interpretation

- Near-invisible watermark in both datasets
- Cross-lingual performance remains strong

Evaluating Quality and Detection

Audio Quality Metrics

- PESQ, STOI → strongly correlate with human perception
- SI-SNR → purely signal-based, not always aligned with perceived quality

Detection Metrics

- Detector Score, TPR/FPR, AUC
- Hamming distance for multi-bit extraction

Limitations

- Real listening tests (like MUSHRA) remain key for subtle artifact detection

Challenges and Future directions

Adversarial Removal

- Detector exposure → attackers can craft noise to remove the watermark

Extreme Compression

- Surviving 2–3 kbps streams is underexplored

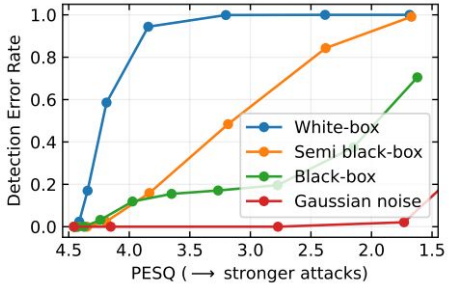


Figure 7. **Watermark-removal attacks.** PESQ is measured between attacked audios and genuine ones (PESQ < 4 strongly degrades the audio quality). The more knowledge the attacker has over the watermarking algorithm, the better the attack is.

Cross-Lingual Expansion

- More languages, code-switching

Security vs. Transparency

- **Open-source generator vs. private detector**

Real-Time Embedding/Detection

- Scaling for live streaming at large platforms

Conclusion

AudioSeal

- Localized, near-invisible watermarking for speech

High Quality

- PESQ ~4.4, STOI ~0.998

Robust

- Survives various edits, single-pass detection

Cross Lingual

- Maintains performance on Indian English + Hindi

THANK YOU !