Real-time Speech Watermark for Defending Hidden Phone Call Recording

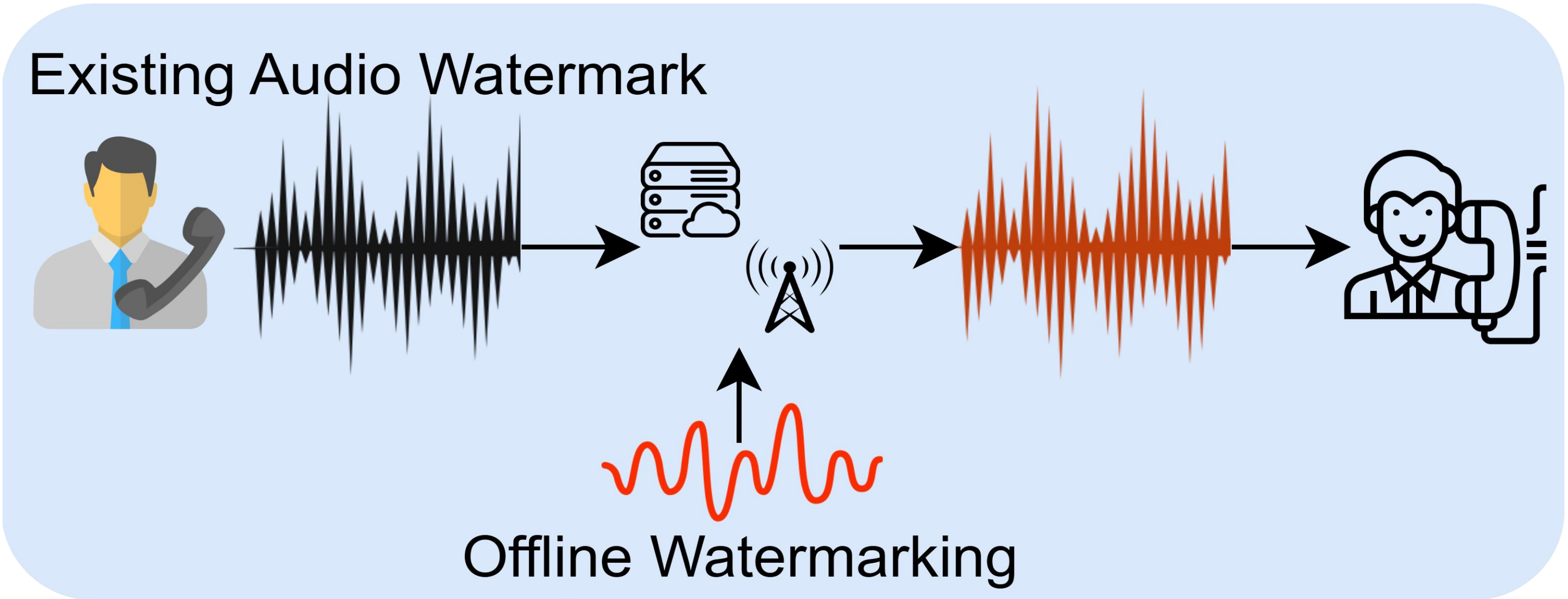
Yizhu Wen[1], Rui Duan[2], Yisheng Zhong[3], Zhuangdi Zhu[3], Hanqing Guo[1]
University of Hawaii at Manoa[1], University of Missouri-Kansas City[2], George Mason University[3]

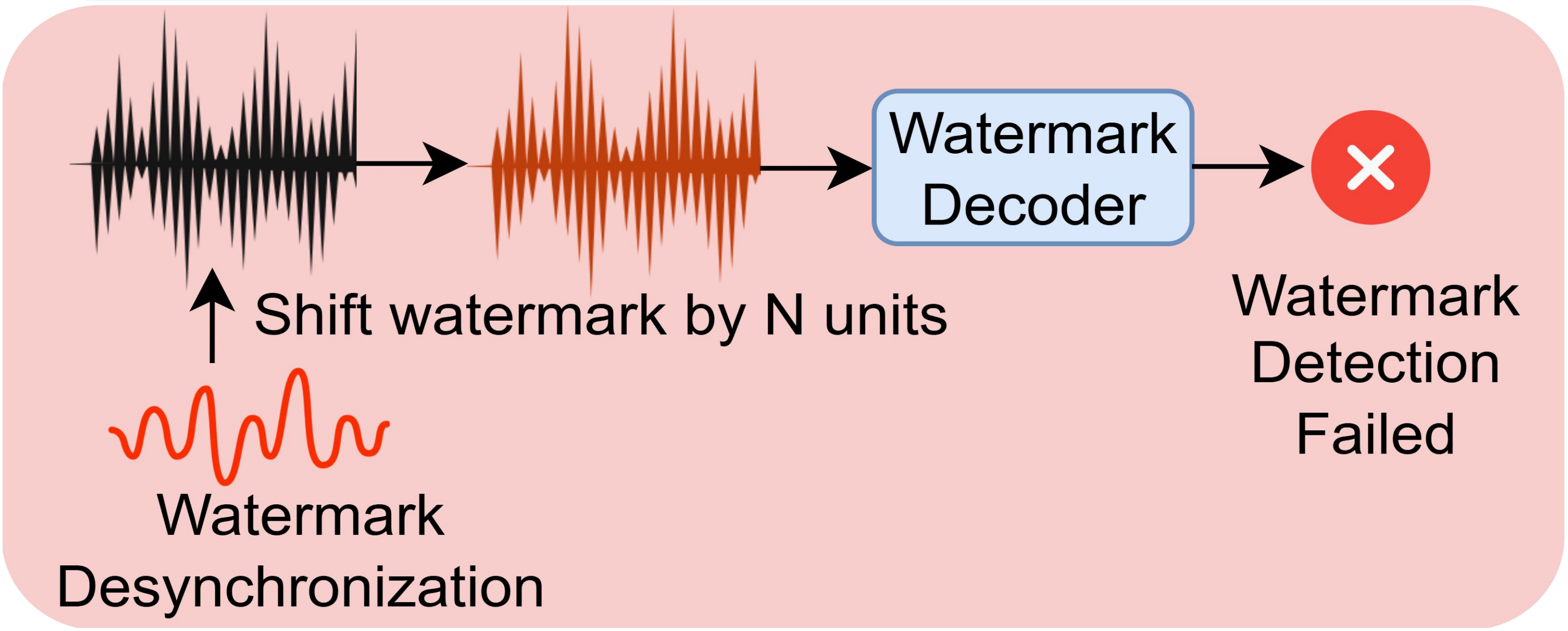
Contact Email: {yizhuw, guohanqi}@hawaii.edu[1],
ruiduan@umkc.edu[2], {yzhong7, zzhu24}@gmu.edu[3]

Motivation:

- **How to prove the recording has my consent?**
- Idea:
  - - Add watermark as consent during the call.
  - - Check the presence of watermark to prove.
- However, all existing **watermarking** fail to embed watermarks **in real time**.

Introduction:

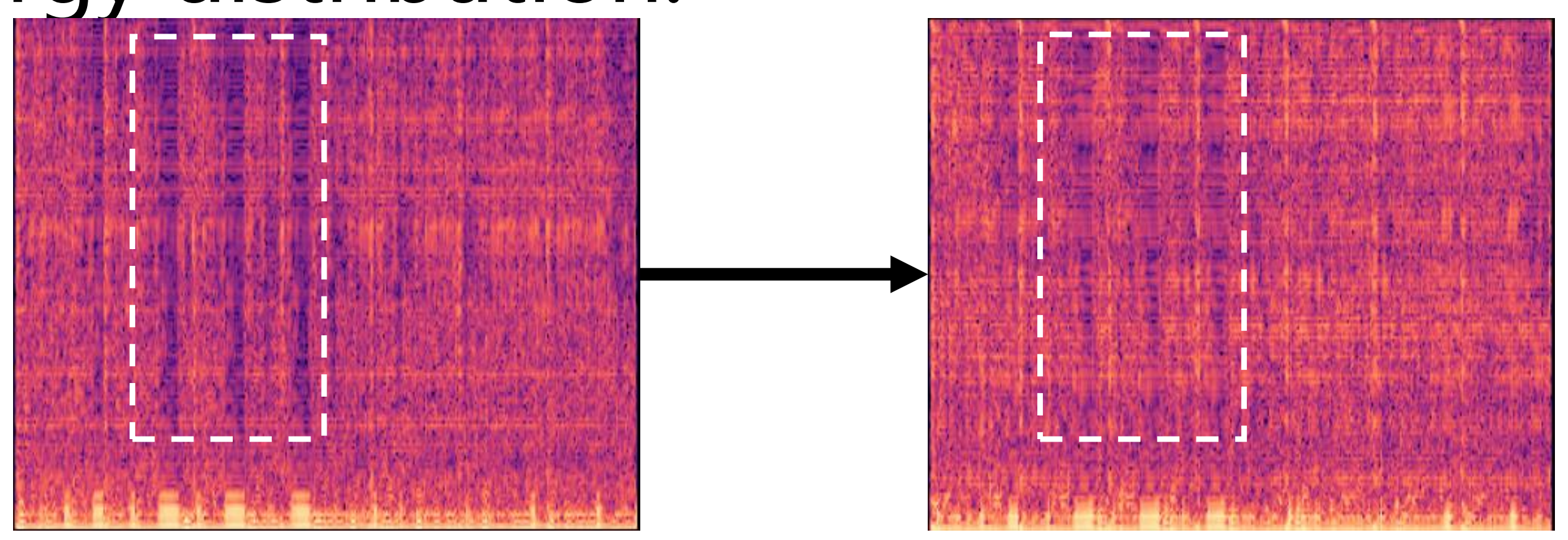



Watermark Decoder
Shift watermark by N units
Watermark Detection Failed
Watermark Desynchronization

Frequency Repetition:

To reduce spectrogram watermark visibility, we project the bitstring onto half the frequency bins and repeat it, avoiding time-axis repetition for more even energy distribution.

No Frequency Repetition

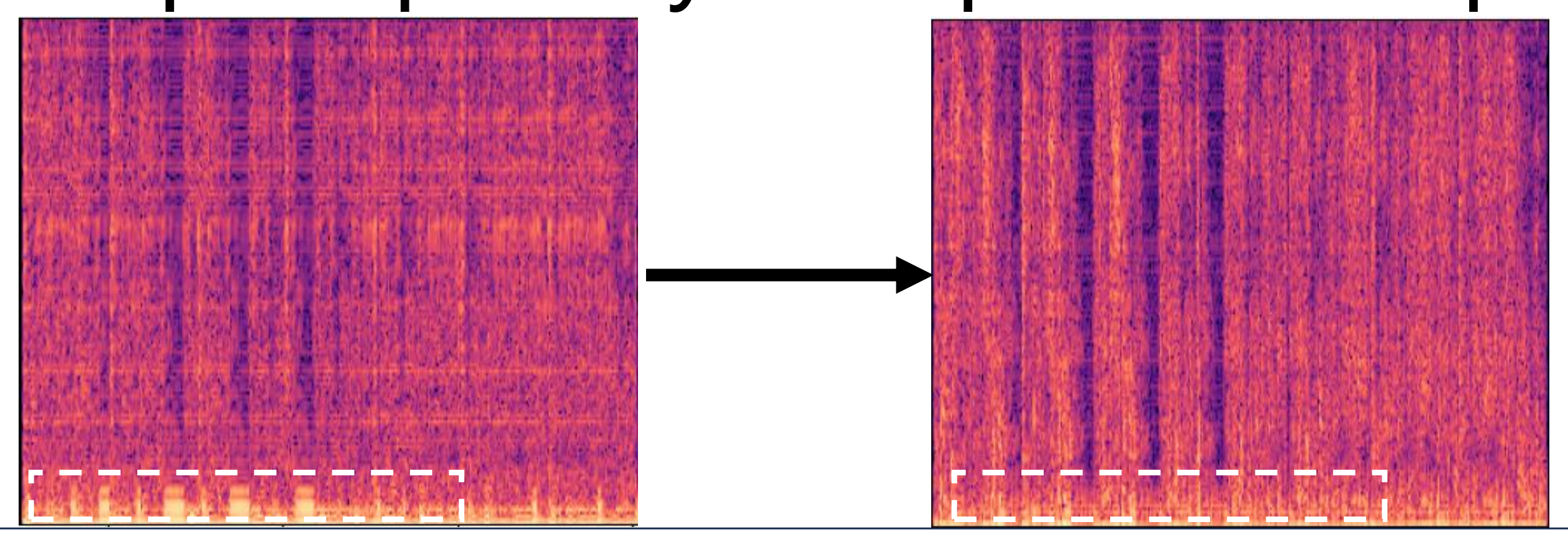


Frequency Repetition

VAD Filtering:

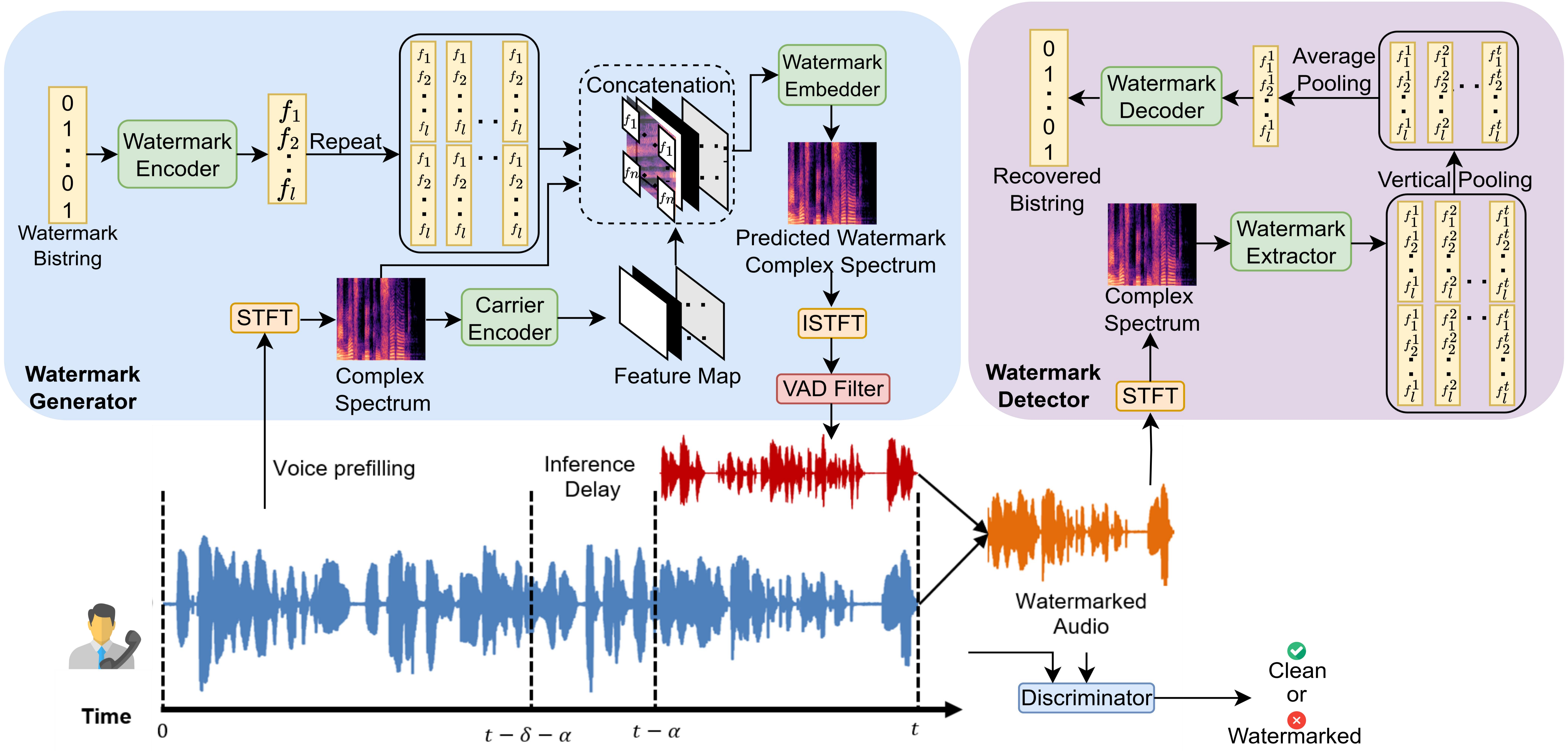
Voice Activity Detection (VAD) suppresses watermarks during silence to prevent perceptibility in unpredictable pauses in speech.

No VAD



VAD Filtering

Framework Overview:



- **Real-time Requirement**: To operate on live phone calls, the technique must predict the audio watermark **ahead of time**.
- **Predictive Approach:** The method uses **forward prediction** to forecast the watermark vector for future time steps **by conditioning on past audio input**.
- **Timing Constraint:** To embed a watermark starting at time $t-\delta$, computation must begin by time $t-\delta-\sigma$, where δ represents the maximum time required to record, compute, and play the watermark.

Loss Functions:

Watermark Embedding Loss:

MSE = $\frac{1}{N}\sum_{n=0}^{N-1}(x_n - y_n)^2$,where x is the watermarked audio and y is the clean audio.

Time-frequency Loudness Loss:

Loudness difference: $l_b^w = Loudness(a_b^w) - Loudness(x_b^w)$

Loudness Loss: $L_{loud} = \sum_{b,w}(softmax(l)_b^w * l_b^w)$, where a is watermark audio, b is frequency bands, and w is window size. The detailed code is from Meta’s Audiocraft library.

Conclusion:

We introduce a real-time speech watermarking system against hidden call recording by predicting and embedding stealthy watermarks into future live audio. Using frequency repetition and VAD filtering, it preserves speech quality while enabling post-call verification. It achieves **96.71%** bit recovery accuracy on LibriSpeech and LJSpeech datasets demonstrates effectiveness, highlighting predictive watermarking’ s potential as a practical real-world privacy tool.