Speech Watermarking for Tampering Detection Using SSA with a Psychoacoustic Model

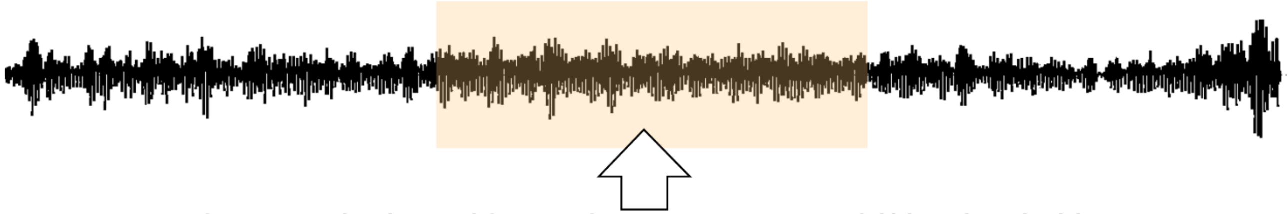
4 December 2023

26th O-COCOSDA, IGDTUW, Delhi, India

Phondanai Khanti, Pannathorn Sathirasattayanon, Patthranit Kaewcharuay, Nanthayod Termkoh, Ekachai Phaisangittisagul, Kasorn Galajit, Jessada Karnjana

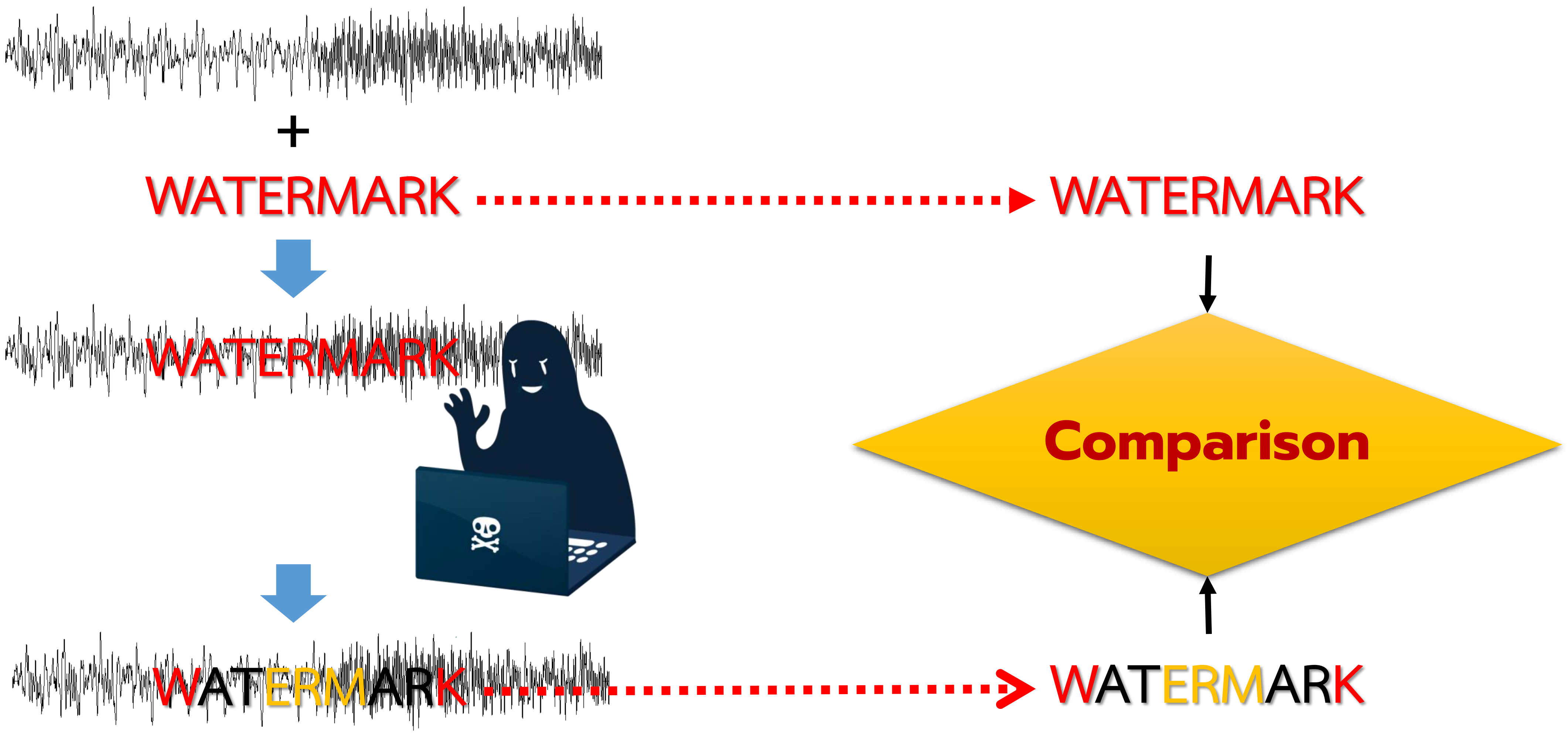
Issue: Tampering

- ❑ Unauthorized modification of speech signals can lead to misinformation, invade privacy, and reduce the reliability of individuals and agencies.
- ❑ How can we detect the tampering?

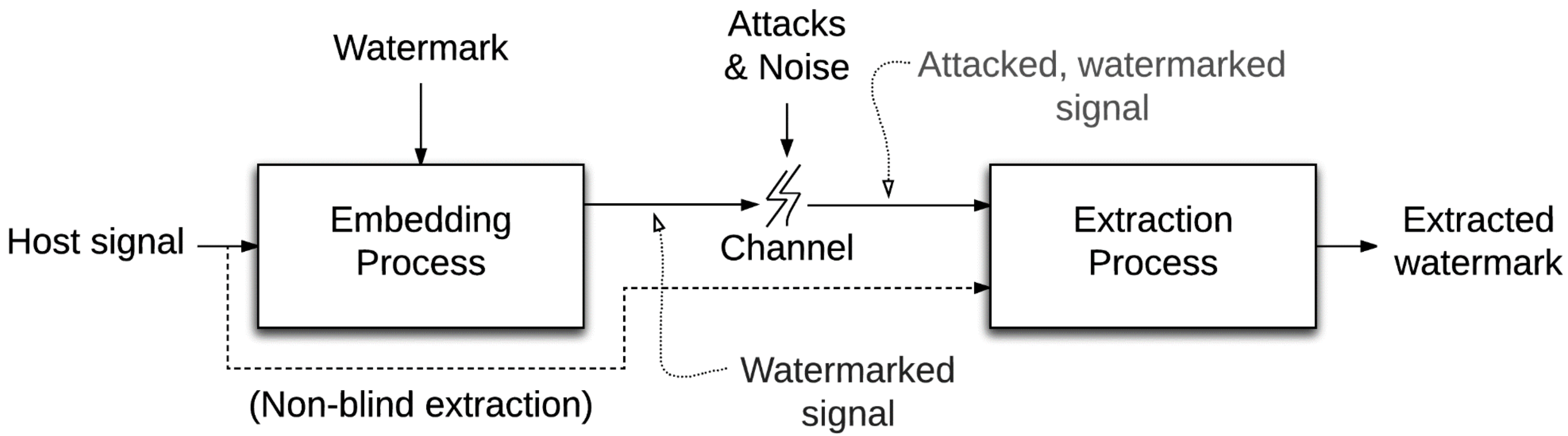


(e.g., raplacing with another segment or shifting in pitch)

Solution: Watermarking



Speech/Audio Watermarking



- ❑ Inaudibility or transparency
- ❑ Fragile to malicious attacks
- ❑ Robust against non-malicious signal processing
- ❑ Blindness
- ❑ Secrecy and security
- ❑ Capacity

Problem Statement

- Trade-off between the robustness and fragility (i.e., semi-fragility)
  - e.g., too fragile to some non-malicious attacks
- Trade-off between the sound quality and semi-fragility
  - e.g., sound quality is reduced in the blind scheme

Objective

NECTEC a member of NSTDA

- To develop a speech watermarking scheme based on the singular spectrum analysis (SSA) and a psychoacoustic model (PAM) for tampering detection that improves the sound quality of the watermarked speech signal

Motivation

NECTEC
a member of NSTDA

Invariance of singular values

+

Speech perception

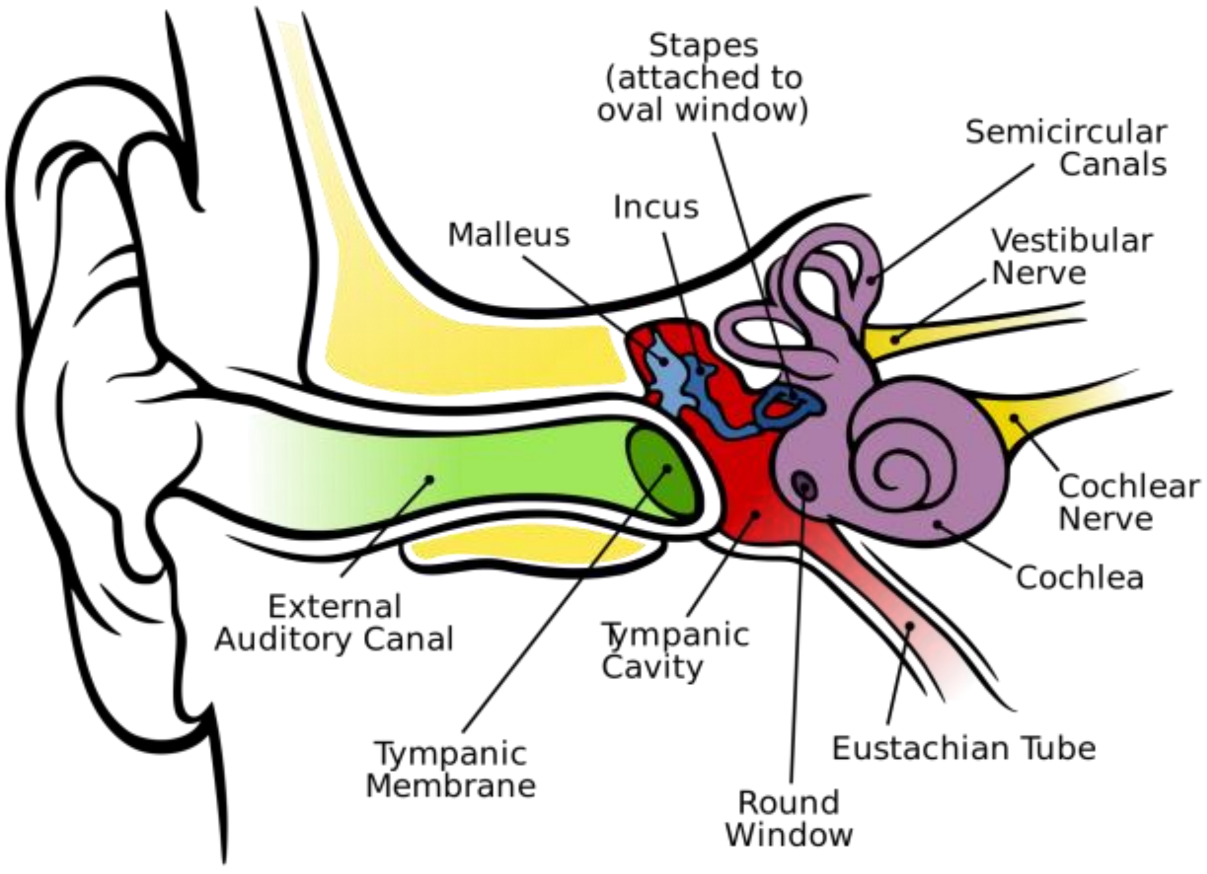
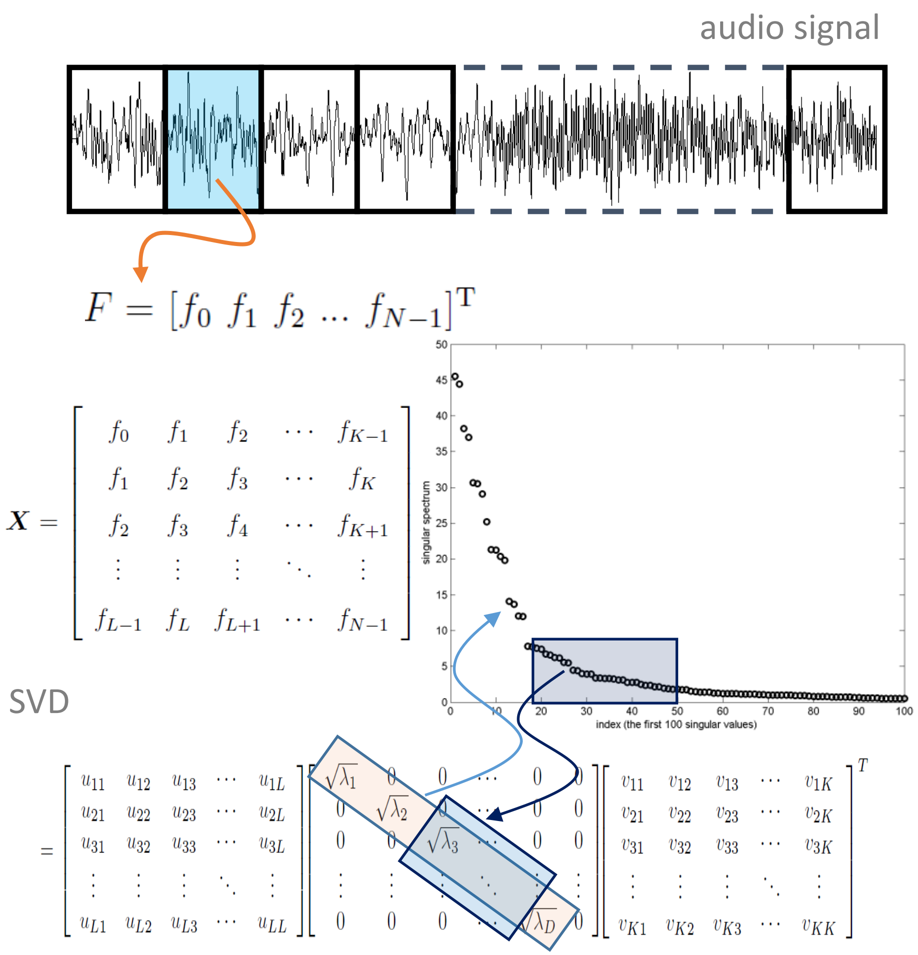
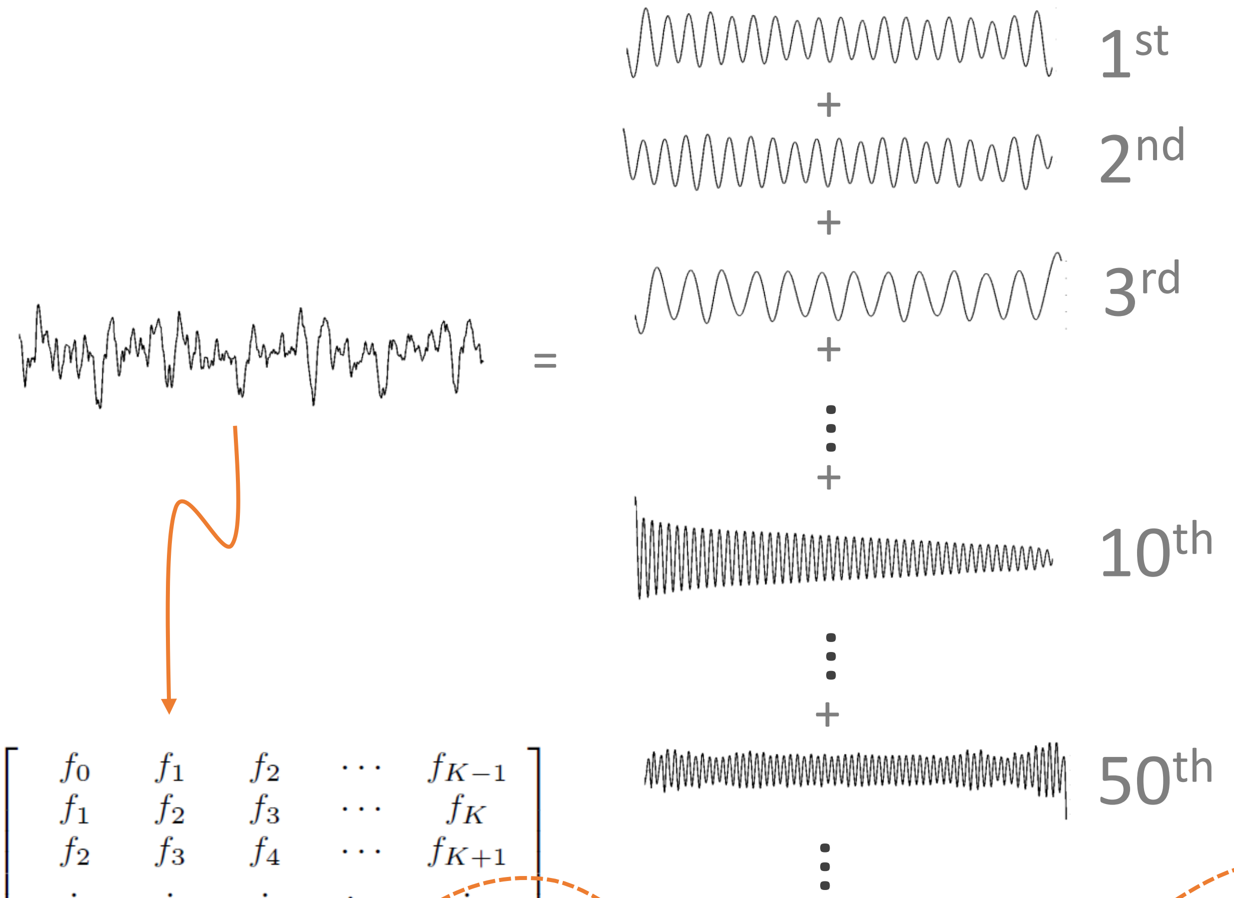


Figure source: wikipedia.org

Singular Spectrum Analysis

- Born in 1986, it has become a standard tool in the analysis of climate, meteorological, and geophysical **time series**.





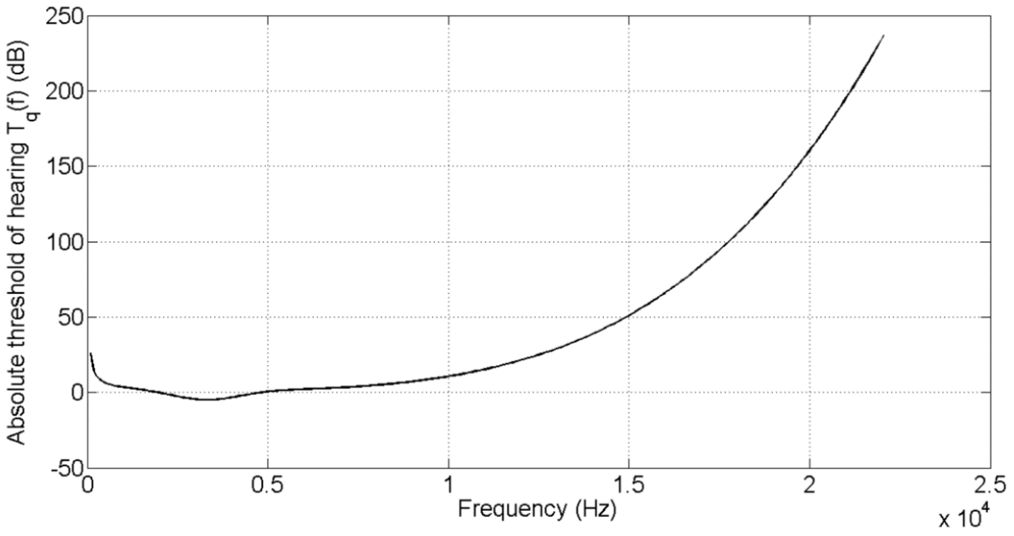
$$
\begin{bmatrix}
f_0 & f_1 & f_2 & \cdots & f_{K-1} \\
f_1 & f_2 & f_3 & \cdots & f_K \\
f_2 & f_3 & f_4 & \cdots & f_{K+1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1}
\end{bmatrix}
$$

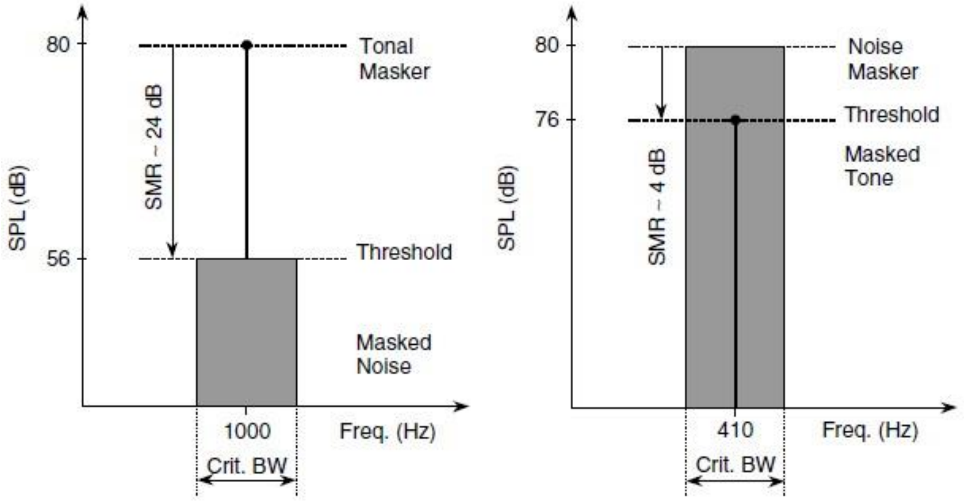
$$
= \begin{bmatrix}
u_{11} & u_{12} & u_{13} & \cdots & u_{1L} \\
u_{21} & u_{22} & u_{23} & \cdots & u_{2L} \\
u_{31} & u_{32} & u_{33} & \cdots & u_{3L} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
u_{L1} & u_{L2} & u_{L3} & \cdots & u_{LL}
\end{bmatrix}
\begin{bmatrix}
\sqrt{\lambda_1} & 0 & 0 & \cdots & 0 & 0 \\
0 & \sqrt{\lambda_2} & 0 & \cdots & 0 & 0 \\
0 & 0 & \sqrt{\lambda_3} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \sqrt{\lambda_D} & 0
\end{bmatrix}
\begin{bmatrix}
v_{11} & v_{12} & v_{13} & \cdots & v_{1K} \\
v_{21} & v_{22} & v_{23} & \cdots & v_{2K} \\
v_{31} & v_{32} & v_{33} & \cdots & v_{3K} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_{K1} & v_{K2} & v_{K3} & \cdots & v_{KK}
\end{bmatrix}^T
$$

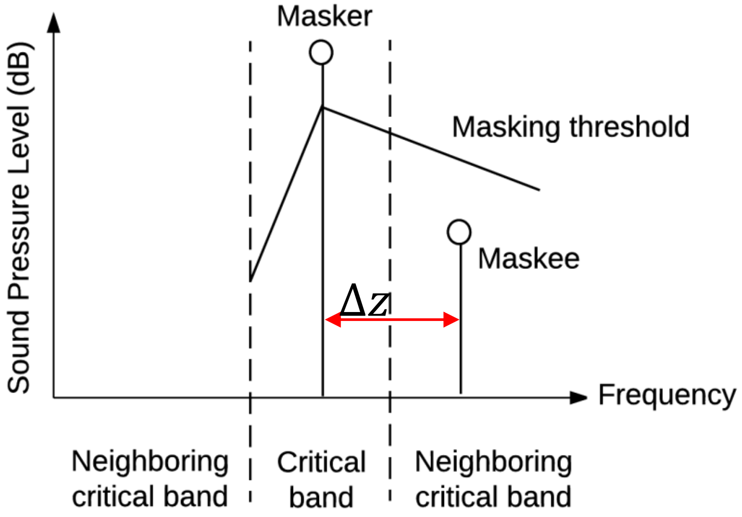
Psychoacoustic Principles

- ❑ Absolute threshold of hearing
- ❑ Masking

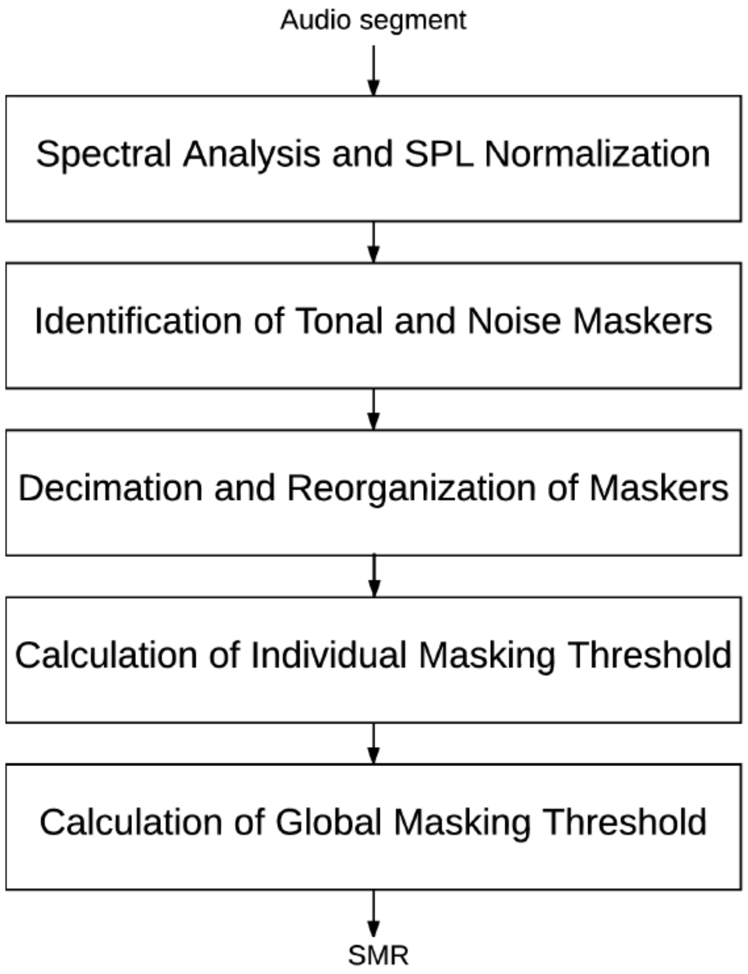
$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 0.001\left(\frac{f}{1000}\right)^4$$

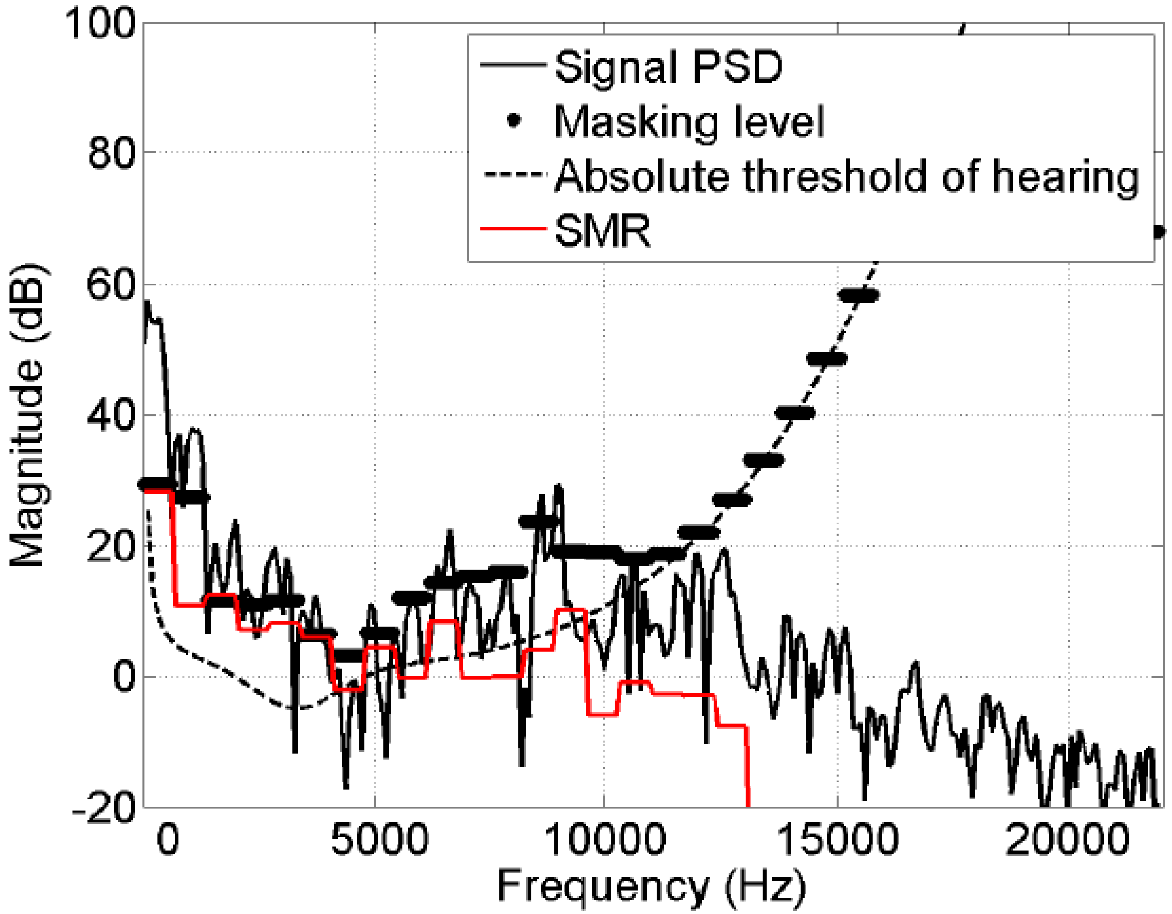






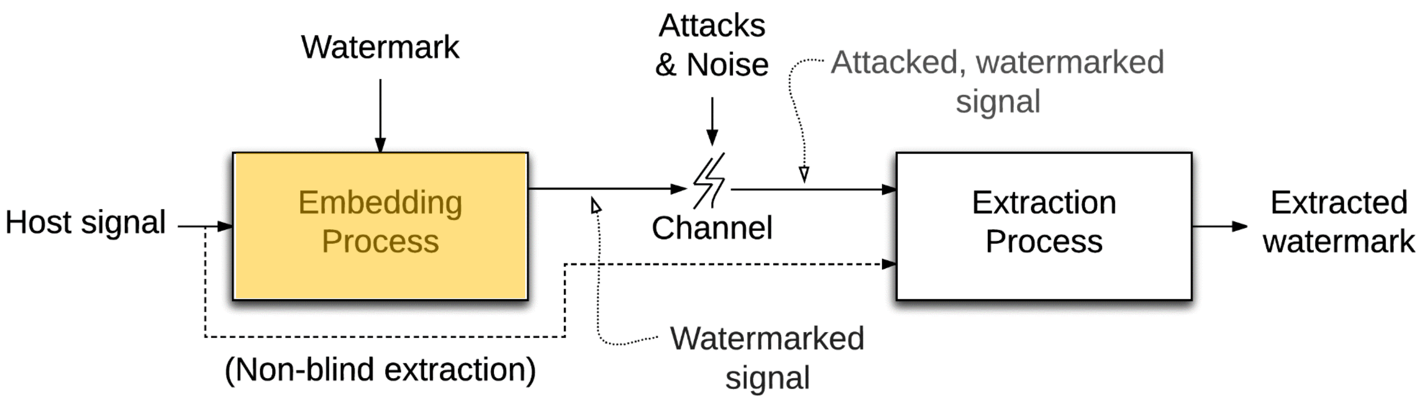
PA Model 1 (ISO/IEC 11172-3)



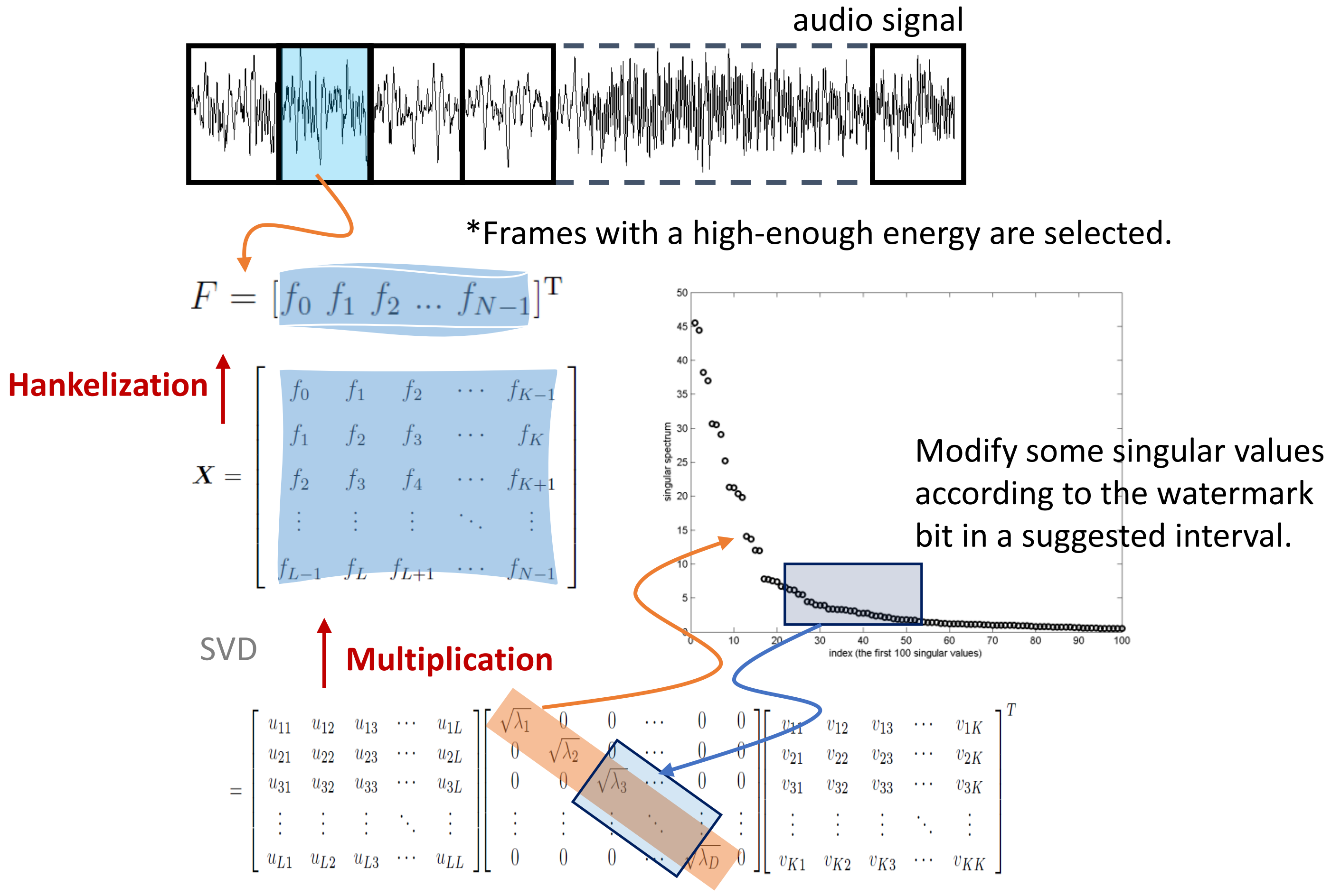


Proposed Method

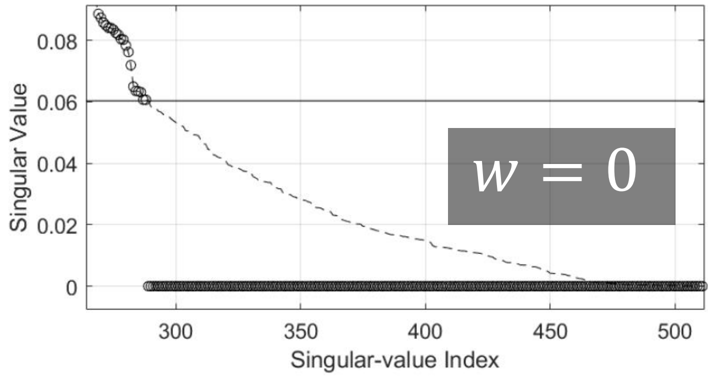
- ❑ Embedding process
- ❑ Extraction process

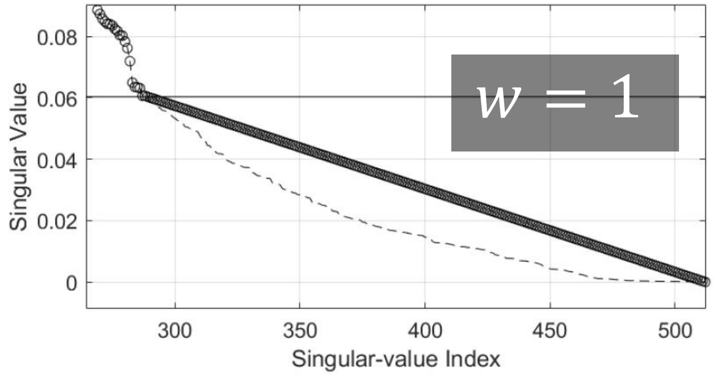


Embedding Process



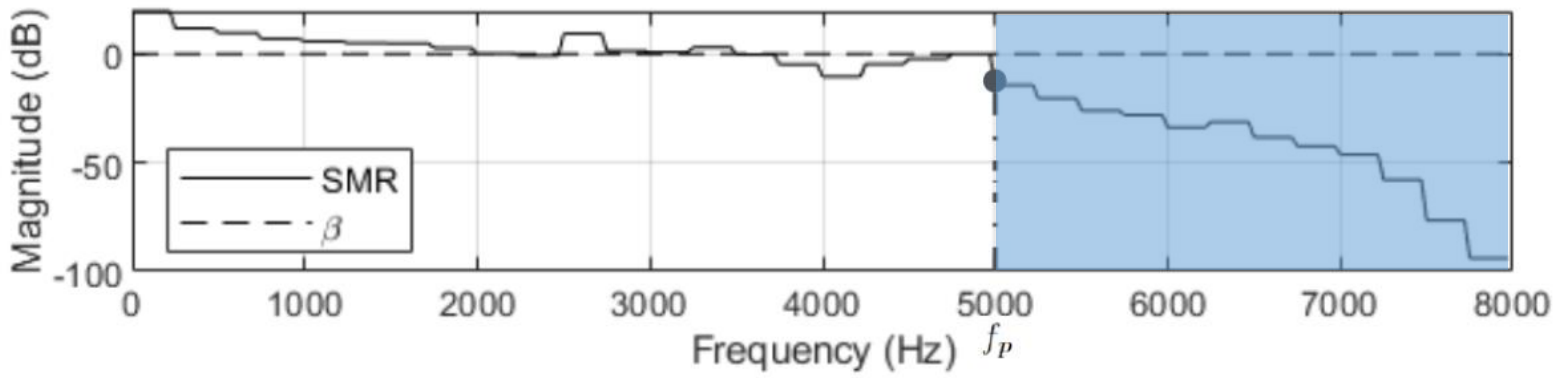
$$\sqrt{\lambda_i} = \begin{cases} \left(\frac{\sqrt{\lambda_q}-\sqrt{\lambda_p}}{q-p}\right)\cdot(i-p)+\sqrt{\lambda_p}, & \text{if } w=1 \\ 0, & \text{if } w=0 \end{cases}$$





Suggested Interval

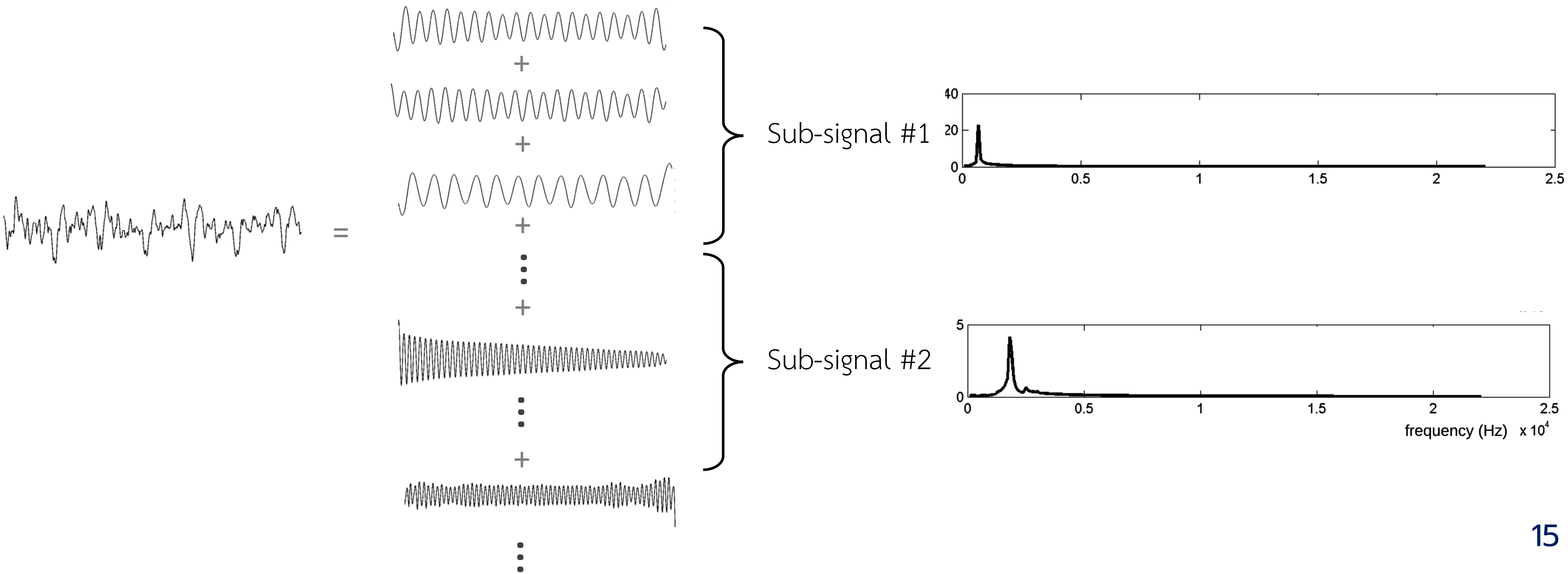
- We set a predefined SMR threshold (β) such that the frequency components in which its SMR is lower than the threshold are considered suitable for hiding the watermark bit.



- Convert f_p to a singular-value index p.

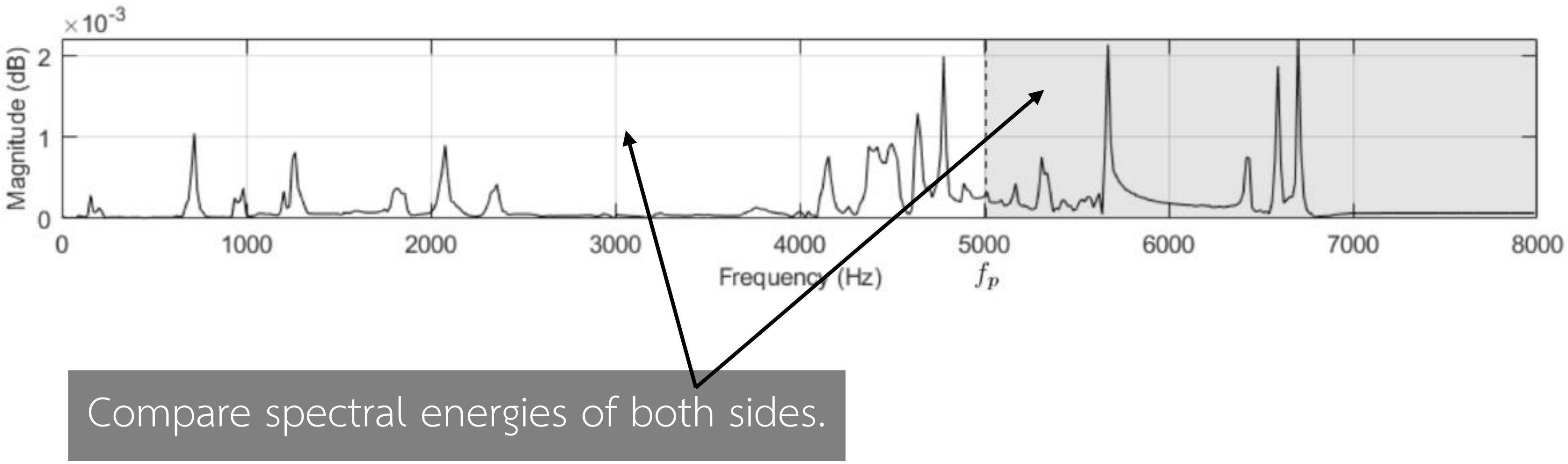
Frequency-to-Index Conversion

❑ STEP 1: Find a spectrum of each sub-signal.



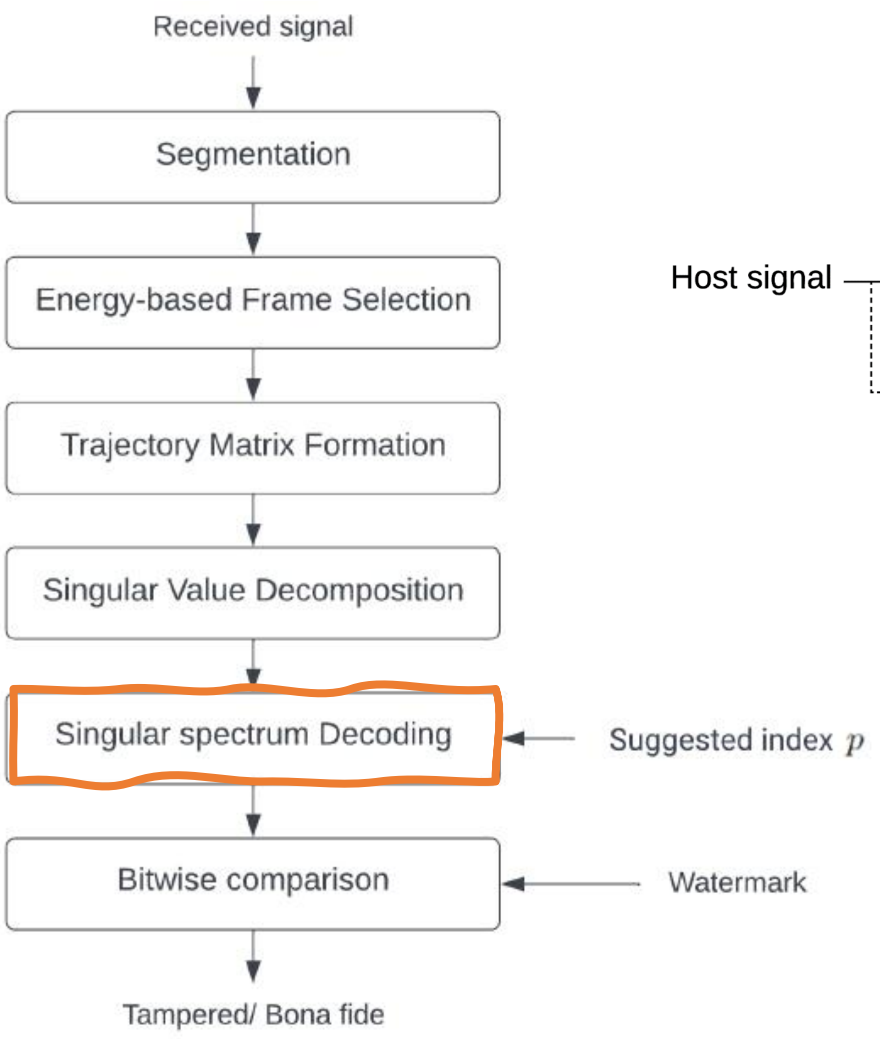
Freq-to-Ind Conversion (cont'd)

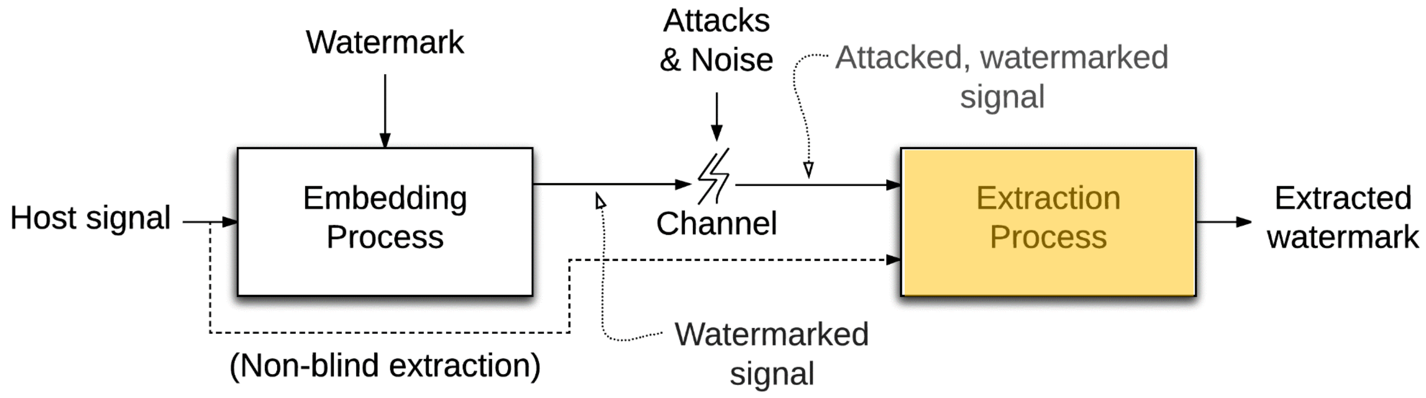
- STEP 2: Divide the sub-signal spectrum into two parts at f_p.

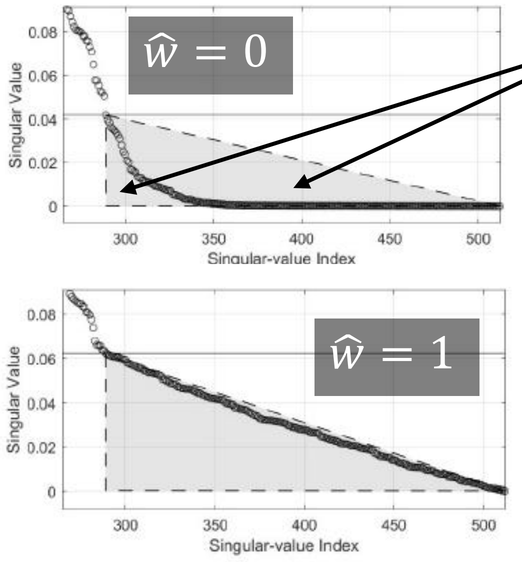


The first singular-value index of the first sub-signal that satisfies a condition that the spectral energy on the left is greater than the spectral energy on the right is chosen as the index p.

Extraction Process







Compare the area under the singular spectrum and the area of shaded triangle.

Experimental Data

- 12 Japanese speech signals from the ATR dataset (B set)
- 16 kHz sampling rate
- 16-bit quantization
- single channel signal
- frame size = 1024 samples
- 100 watermark bits per signal (i.e., duration = 6 seconds)
- 10 signal-processing operations: Gaussian-noise addition, G.711, G.726, band-pass filtering, MP3, MP4, **pitch shifting**, **single echo addition**, **replacing a segment**, and **changing the speed**

Evaluation

- Robustness and fragility: Bit Error Rate (**BER**, in %)
    - BER < 10% for untouched or non-tampered signals
    - BER > 20% for malicious attacks
    - BER between 10% and 20% for unintentionally modified or tampered with a low amount
- Sound quality
    - Perceptual Evaluation of Speech Quality (**PESQ**, in ODG)
        - ODG > 3 (Note that ODG = -0.5 means highly othersome, and ODG = 4.5 means imperceptible)
    - Log-spectral Distance (**LSD**, in dB)
        - LSD < 1 dB
    - Signal-to-Distortion Ratio (**SDR**, in dB)
        - SDR < 25 dB

Experimental Result: BER

	LSB-based method [1]	CD-based method [12]	FE-based method [13], [14]	SSA-based method [6]	SSA-based method [6] with frame selection	Proposed method
No attack	0.00	∼0.00-1.00	0.00	0.49	0.00	0.34
G.711	0.00	∼4.00	0.00	0.49	0.00	0.34
G.726	51.77	∼20.00-25.00	0.00	27.66	16.50	47.50
MP3	50.49	-	-	3.69	31.47	1.39
MP4	49.53	-	-	32.79	35.22	22.40
BPF	50.83	-	-	50.23	43.86	47.42
AWGN (15, 40 dB)	50.70, 49.53	-	∼54.00	49.69, 24.53	55.68, 0.00	56.21, 27.54
PSH (−4%, −10%, −20%)	35.64, 35.33, 4.08	-	∼31.00, -, -	10.58, 22.03, 47.83	19.24, 21.41, 43.08	17.23, 26.34, 43.08
PSH (+4%, +10%, +20%)	34.42, 34.36, 38.03	-	-	12.44, 15.33, 20.47	20.56, 25.27, 18.47	20.42, 22.79, 30.79
Echo (20, 100 ms)	50.18, 51.34	-, ∼50.00	-, ∼5.00	15.76, 20.33	30.28	30.73
Replace (1/3, 1/2)	16.51, 24.97	-	∼57.00, -	17.08, 25.78	32.84, 32.91	36.36, 36.56
SCH (-4%, +4%)	49.47, 48.72	-	∼20.00, -	47.00, 47.19	35.79, 39.23	39.41, 40.28

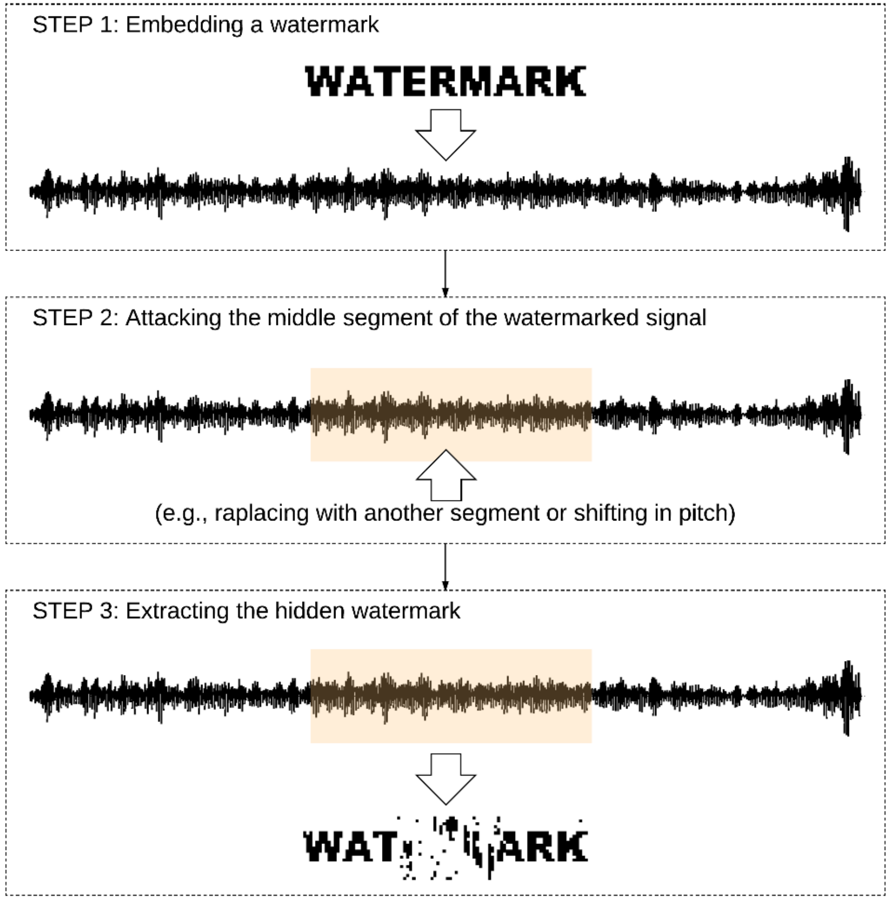
- The proposed method is better than the CD-based and FE-based methods and is comparable to the SSA-based method.
- It is fragile to G.726.

Experimental Result: Sound Quality

	ODG	LSD	SDR
LSB-based method [1]	4.49	0.19	65.35
CD-based method [12]	∼3.10-4.30	∼0.60-0.80	-
FE-based method [13], [14]	∼3.90	∼0.40	-
SSA-based method [6]	3.64	0.69	30.96
SSA-based method [6] with frame selection	3.29	0.61	27.00
Proposed method	3.92	0.33	33.10

- The sound quality of the watermarked signal from the proposed method is better than the others, except the LSB-based method.
- It should be noted that the LSB-based method is too sensitive to noise and non-malicious attacks.

Result: Tampering Detection



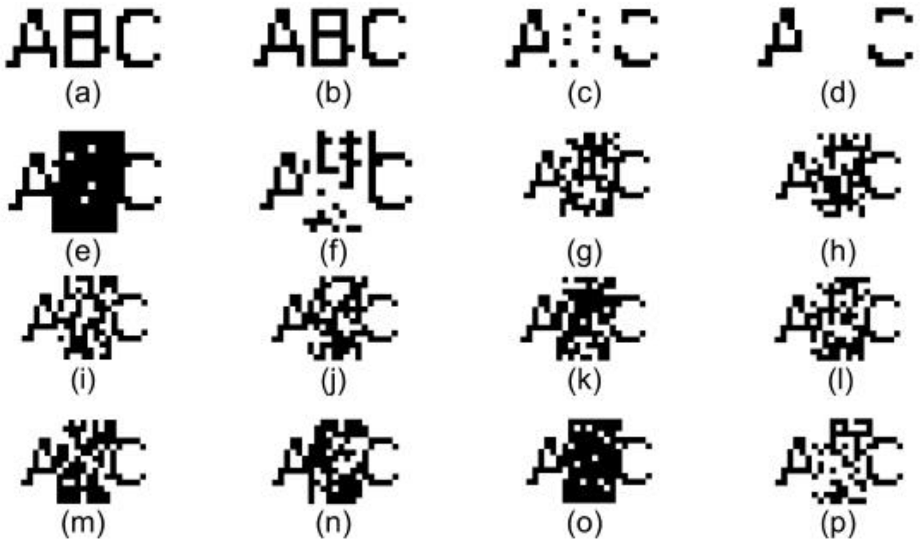
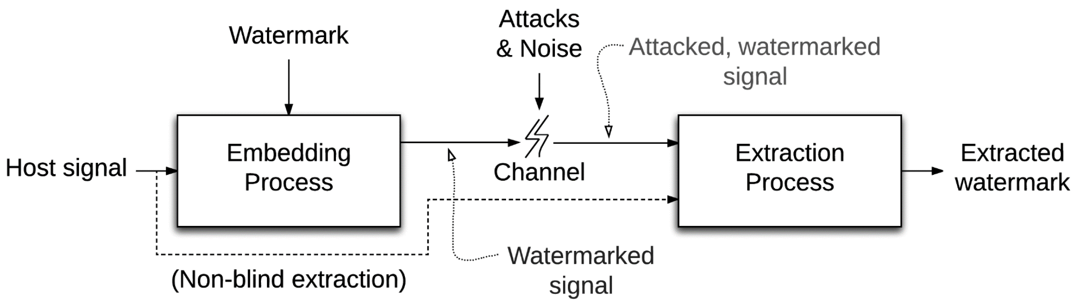


Fig. 10. Results of the tampering detection. Original image (a) and the reconstructed images after performing the following signal-processing operations: (b) G.711, (c) G.726, (d) AWGN (15 dB), (e) BPF, (f) Echo (100 ms), (g) PSH -4%, (h) PSH +4%, (i) Replace (1/3), (j) Replace (1/2), (k) PSH -10%, (l) PSH +10%, (m) SCH -4%, (n) SCH +4%, (o) PSH -20%, and (p) PSH +20%.

Discussion

- The adoption of energy-based selection trades embedding capacity for partial improvement in the watermarked sound quality.
- The tampering detection requires a sequence of suggested indices to decode singular spectra precisely. That is, the extraction process is not completely blind.
- The parameters used in the proposed method have yet to optimize.



Summary

- **Issue**: Speech tampering
- **Aim**: To improve a speech-tampering detection scheme based on the watermarking approach in terms of transparency
- **Method**: SSA + PAM
- **Result**: 7.69% ODG improvement

  6.91% SDR improvement

  52.17% LSD reduction

THANK YOU FOR LISTENING

Contact: Jessada Karnjana
jessada.karnjana@nectec.or.th
NECTEC
a member of NSTDA