A Robust Audio DNA-Based Method for OTT Content Recognition in Noisy Web Streaming Environments

Byeongchan Park[1], Sun-Jib Kim[2], Seok-Yoon Kim[1], Youngmo Kim[1] and Yoontaek Sung[3,*]

[1]*Department of Computer Science & Engineering, Soongsil University, Korea*
[2]*Department of Convergence Security, Hansei University, Korea*
[3]*Media & Advertising Research Institute, Korea Broadcasting Advertising Corporation (KOBACO), Korea*
E-mail: pbc866@ssu.ac.kr; kimsj@hansei.ac.kr; ksy@ssu.ac.kr; ymkim828@ssu.ac.kr; takarajima@kobaco.co.kr
[*]*Corresponding Author*

Received 07 March 2026; Accepted 08 May 2026

Abstract

With the proliferation of Over-The-Top (OTT) platforms via web browsers and mobile web apps, the demand for real-time copyright protection in web streaming environments has surged. However, unstructured noise generated when users consume web content in public environments (e.g., subways, cafes) increases the false positive rate of existing clean-audio-based identification systems and heavily burdens web server computations. This paper proposes a robust, audio DNA-based content recognition method capable of fast and accurate retrieval from large-scale media databases even in noisy web streaming environments. The proposed method extracts dual-stage (Coarse-Fine) features based on Mel-spectrograms at the client side and performs a highly efficient three-stage matching pipeline (Coarse Matching, Fine Matching, and Post-verification) at the web server side. Experimental results on 649

Journal of Web Engineering, Vol. 25_6, 1067–1084.
doi: 10.13052/jwe1540-9589.2562

noisy audio samples demonstrated that applying web-optimized parameters (FFT length of 4096, Hop length of 1470) achieved a precision of 0.9965, a recall of 0.8814, and an F1-score of 0.9354. The proposed method significantly accelerates retrieval speed for large-scale web streaming data through binary hash matching in the Coarse stage while maintaining high accuracy, proving its effectiveness for real-time OTT copyright protection and web media monitoring systems.

Keywords: Web streaming, audio fingerprinting, OTT content recognition, noise robustness, computational efficiency.

1 Introduction

Recent advancements in ICT and cloud infrastructure have established Over-The-Top (OTT) platforms as core web-based media streaming services [1]. Users consume content anywhere using smartphones, tablets, and laptops. While this enhances convenience, it also exacerbates copyright infringement issues such as illegal copying and distribution over the web.

For real-time copyright protection in web streaming environments, technology that can instantly identify the audio content being consumed by the user is essential. Traditional audio fingerprinting techniques, such as Shazam, rely on spectral analysis or hash values of specific frequency points to compare against a database [2, 3]. In addition, audio fingerprinting has been widely studied as a compact and robust representation for content identification, retrieval, and copyright-related applications [6, 7, 10]. However, these traditional methods are generally designed based on relatively clean audio and can be vulnerable to unstructured noise, such as conversational noise, white noise, and streaming compression loss, which are frequently encountered in mobile web environments [4]. External noise introduced at the web client may damage key feature points, causing matching failures or false positives at the server end. This can also increase unnecessary comparison operations in large-scale web-based retrieval environments [5, 8].

Therefore, to apply audio recognition technology to web-based OTT services, a distributed method is required: robust features must be extracted lightly in the restricted client environment, and the backend server – managing massive audio databases – must be able to search and verify these features rapidly without latency [5, 8].

This study proposes a multi-stage DNA-based audio content identification method that overcomes the limitations of existing recognition technologies by

considering both noise robustness and server computational efficiency. The main contributions of this paper are as follows:

- **Dual-stage Coarse-Fine Audio DNA Extraction Optimized for Web Transmission:** By utilizing the Short-Time Fourier Transform (STFT) and Mel filter banks, we generate spectrograms that reflect human auditory characteristics. We designed a structure that extracts a 64-bit binary Coarse DNA for fast server transmission and retrieval, alongside a 256-bit ternary Fine DNA for precise matching.
- **Computationally Efficient 3-Stage Matching Pipeline:** To prevent web server overload, we propose a three-stage pipeline consisting of Coarse Matching, Fine Matching, and Post-verification. The Coarse stage conducts a high-speed search based on Hamming space, while the post-processing stage eliminates false positives using Hough transforms and peak ratio filtering, ensuring real-time performance in large-scale web services.
- **Parameter Optimization and Performance Verification:** Through various noise-infused experiments, we derived the optimal feature extraction parameters, including FFT length, Hop length, and feature dimension, that achieve the best balance between discriminability and bitrate, proving the practical applicability of the proposed method in real-world environments.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed audio DNA extraction and multi-stage recognition method customized for web streaming environments. Section 4 analyses the parameter optimization experiments and performance evaluation results, and Section 5 concludes the paper.

2 Related Work

2.1 Existing Audio Identification Technologies and Limitations in Web Environments

To identify audio content, acoustic fingerprinting technologies that extract unique feature values from audio signals have been widely developed [2–4, 10]. Existing technologies primarily extract specific frequency points of an audio signal to generate unique hash values and compare them against a database [2, 3, 6]. Commercial technologies such as Shazam typically utilize peak-pairing-based local frequency features, providing fast and accurate results in static environments [2].

However, these methods suffer from significantly reduced reliability in mobile web and streaming environments where users are easily exposed to external noise [4]. When key feature points are damaged by white noise or conversational noise, incorrect matching results or identification failures occur. Beyond simple recognition errors, this increases the false positive rate of the web server and generates unnecessary computational overhead, posing clear limitations for practical application in large-scale real-time monitoring systems [4, 5, 8].

2.2 Time-Frequency Transformation and Auditory Model-Based Feature Extraction

As a fundamental approach to extracting features robust to external noise, the STFT is utilized [9]. STFT is suitable for audio signal analysis because it can simultaneously represent information in the time and frequency domains. It divides the input signal into frame units and extracts features by applying methods such as the Hamming window [9].

Subsequently, to process bandwidth efficiently, a Mel-Scale Filter Bank reflecting human auditory characteristics is applied. Because humans are more sensitive to low frequencies than high frequencies, applying the Mel scale effectively compresses and retains the key information of the signal in a way that is advantageous for web transmission, while removing unnecessary high-frequency components. The Mel-Spectrogram generated through this process is utilized as the core foundational data for audio fingerprint extraction [3, 9].

2.3 Hamming Space Matching and Coarse-Fine Structure

To enhance the retrieval speed and computational efficiency of web servers operating large-scale media databases, sequence matching techniques based on Hamming distance have been attempted [5, 8, 10]. This method extracts features, or DNA, of a fixed bit length from the audio signal and has the advantage of extremely high processing speed by utilizing simple bit-level operations such as XOR operations or bit masking [5].

Recently, research has been conducted to reduce server load by separating these fingerprints into a dual Coarse-Fine structure [6–8]. The Coarse fingerprint is utilized as a low-dimensional binary vector to perform rapid candidate searches on the web server, while the Fine fingerprint is used as a multi-class vector for precise matching. However, existing Hamming-based methods primarily remain at the level of Coarse comparison, lacking precise

matching and post-processing stages to effectively eliminate false positives. Therefore, this study proposes an identification structure optimized for web architectures by combining rapid retrieval, Coarse, precise matching, Fine, and Post-verification [5, 8, 10].

3 Web-Optimized Audio DNA Recognition Method

3.1 Overview

The proposed method consists of a highly accurate DNA extraction process and a matching module to effectively compare audio infused with unstructured noise against original web content, as shown in Figure 1.

Figure 1 illustrates the overall flow where the query audio is input through a microphone and undergoes a preprocessing stage to be converted into audio DNA, enabling identification even when noise is included. The DNA extracted in the client environment is transmitted to the web server. To minimize computational load, the server backend processes the data through a three-stage pipeline (Coarse Matching, Fine Matching, Post-verification) to finally output the content ID and position information.

3.2 Audio Feature Extraction Based on Time-Frequency Transformation

In the proposed method, the input audio is divided into frames of a specific length in the time domain, with overlaps between adjacent frames to secure temporal resolution. A Hamming window is applied to each frame to mitigate spectral distortion that may occur at the boundaries, as illustrated in Figure 2 and defined in Equation (1).

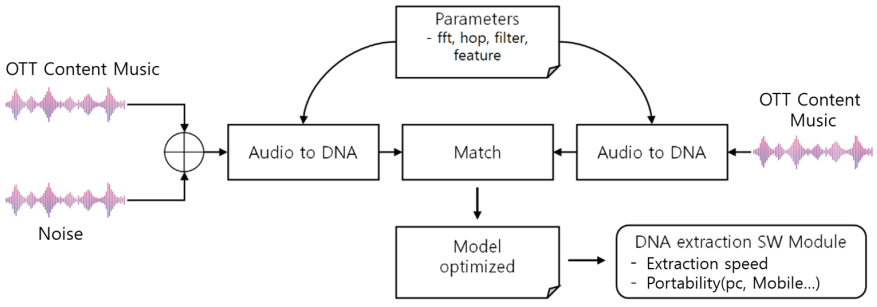

Figure 1 Configuration for robust audio matching in noisy environments.

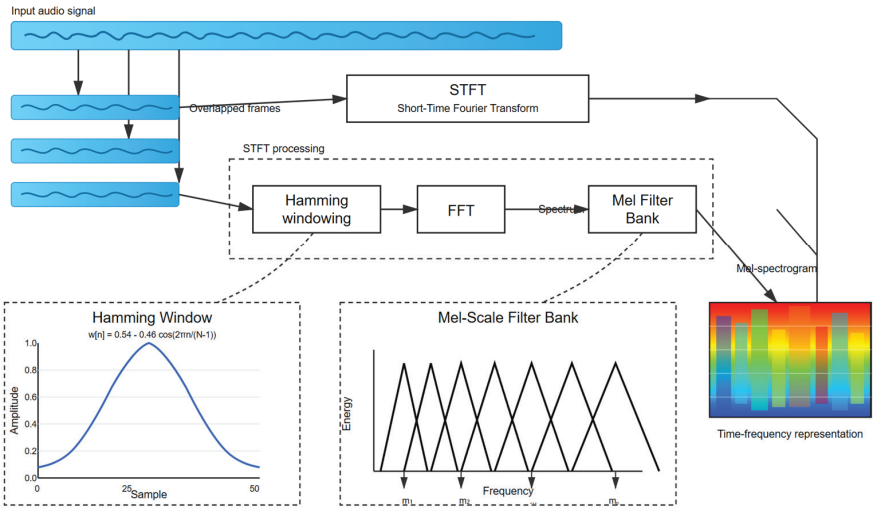

Figure 2 STFT and Mel filter-based audio feature extraction process.

The Hamming window $\omega[n]$ is defined as:

$$\omega[n] = 0.54 - 0.45cos\left(\frac{2\pi N}{N-1}\right) \tag{1}$$

Mathematically, this formula smoothly tapers the ends of the frame to zero, effectively suppressing the discontinuous noise (spectral leakage) that occurs when truncating the audio signal. Subsequently, a STFT is performed over short time intervals to obtain the time-frequency domain distribution. This is executed based on Equation (2):

$$STFTx(m,\omega) = \sum_{n=-\infty}^{\infty} x[n] \cdot \omega\,[n-m] \cdot e^{-j\omega n} \tag{2}$$

where $x[n]$ represents the input audio signal. By multiplying the windowed signal by a complex exponential function, the system calculates the variations of frequency components over time. A Mel filter bank, which reflects human auditory characteristics, is applied to the linear spectrum obtained from the STFT to generate a Mel-spectrogram. The process of converting frequency to Mel frequency m is performed by Equation (3):

$$m = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

This transformation formula reflects the structure of the human ear, which is more sensitive to low-frequency bands. By compressing unnecessary high-frequency data, it efficiently reduces the volume of feature data the server must process, making it highly advantageous for web transmission.

3.3 Dual Fingerprint (Coarse & Fine) Generation

The Mel-spectrogram is used as input for the feature extraction stage, forming an audio DNA with a dual structure consisting of a Coarse fingerprint and a Fine fingerprint, as depicted in Figure 3.

Figure 3 visualizes the process where features extracted from the spectrogram are separated into a low-dimensional binary vector for rapid search in large-scale databases and a multi-class vector for precise matching.

- **Coarse DNA:** This is a 64-bit binary vector used for rapid candidate selection based on Hamming distance. It is generated through binary processing based on a threshold θ_c applied to the spectrogram's intensity value s_i, as shown in Equation (4):

$$f_c[i] = \begin{cases} if\ s_i \geq \theta_c \\ otherwise \end{cases} \tag{4}$$

- **Fine DNA:** This is a 256-bit fingerprint composed of 128-dimensional ternary values. It is generated through differential normalization using a scaling parameter Δ for precise sequence matching, as expressed in Equation (5):

$$f_f[i] = \left[\frac{s_{i+1} - s_i}{\Delta} \right] \tag{5}$$

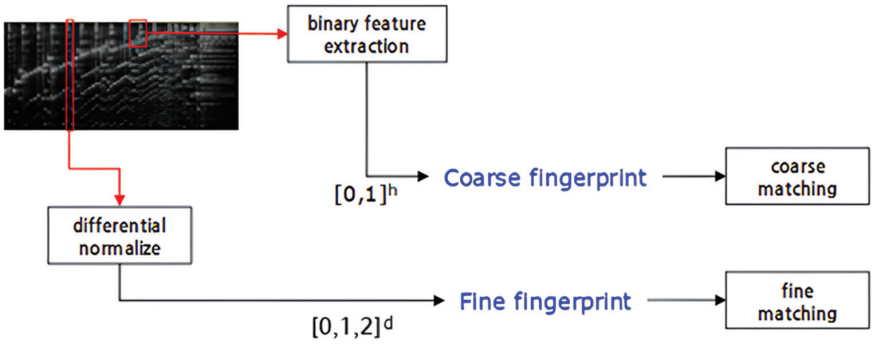

Figure 3 Audio recognition architecture based on dual fingerprint structure.

3.4 Computationally Efficient 3-Stage Matching Pipeline

3.4.1 Stage 1: Coarse matching

In this stage, candidates are rapidly selected from the large-scale server database by calculating the Hamming distance between the 64-bit binary Coarse fingerprints. To enhance the computational efficiency of the proposed method, a 64K lookup table-based optimization is applied, as shown in Figure 4. The Hamming distance H is calculated using Equation (6):

$$H = (D_c^q, D_c^r) = \sum_{i=0}^{64} 1\{D_c^q[i] \neq D_c^r[i]\} \tag{6}$$

This formula performs a logical XOR operation between two binary DNA sequences (D_c^q and D_c^r), counting the number of differing bits to determine similarity rapidly and lightly, thus minimizing server load.

3.4.2 Stage 2: Fine matching

For the candidates filtered from the first stage, a precise similarity comparison is performed using the 256-bit ternary DNA D_f. Because the number of comparison targets is significantly reduced, real-time processing is achievable even with CPU-based computations, effectively filtering out items with a high probability of being false positives.

3.4.3 Post-verification

To prevent final false positives caused by repeating patterns or noise, the following three procedures are executed:

1. **Hough Transform (Line Detection):** Detects diagonal line shapes in the similarity image to determine sequence alignment. This is highly

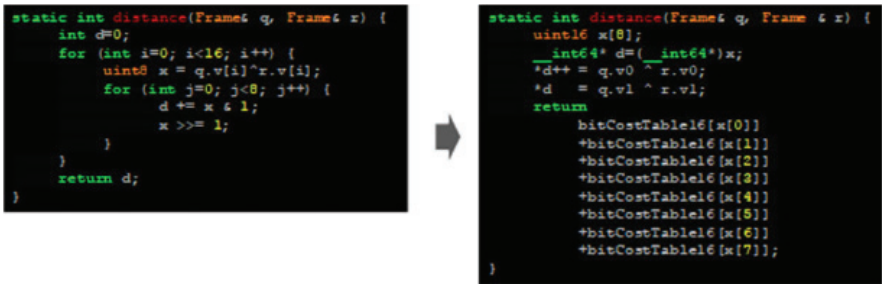

Figure 4 Optimized architecture for sequence matching based on hamming distance.

effective in verifying the continuity of the content segment even in environments with temporal drift.

2. **Peak Ratio Filtering:** If the ratio between the 1st peak and the 2nd peak in the similarity image falls below a standard threshold, it is considered a false positive caused by repeating patterns and is filtered out, as defined in Equation (7):

$$\frac{Peak2}{Peak1} < 0.7 \geq Reject \tag{7}$$

This formula mathematically screens out instances where high similarity (peaks) redundantly appear across multiple segments due to repetitive rhythms (e.g., strong musical beats).

3. **Sigmoid Similarity Scaling:** A sigmoid function is applied to the distance-based similarity to maintain discriminative power near the classification boundaries, as expressed in Equation (8):

$$S(d) = \frac{1}{1 + e^{-k(d-\delta)}} \tag{8}$$

Here, d represents the distance value, k is the function's slope, and δ is the center threshold. This scaling equation makes the score variation based on distance non-linear, mathematically preventing the overestimation of ambiguous matching results.

4 Experiments and Results

4.1 Experimental Environment and Dataset

To verify the performance of the proposed audio DNA extraction and recognition method in a noisy web streaming environment, an experimental server environment was configured as shown in Table 1.

The dataset used for the experiment consisted of a total of 649 audio samples, simulating various web content. Each sample was formatted as a 20-second, 22 kHz mono WAV file. The experimental data was structured as reference-query pairs. The query audio was generated by recording the reference audio output through a speaker using a microphone in various noisy

Table 1 Experimental web server environment

Component	Specification
CPU	Intel Core i9 12900K
GPU	NVIDIA GeForce RTX 4090

environments, reflecting the actual conditions under which users consume OTT content via mobile web browsers.

4.2 Parameter Optimization for Web Streaming

The parameters used for audio DNA extraction significantly impact both the discriminability of the content and the computational load on the web server. To find the optimal balance, we measured the sequence similarity of the fingerprints.

Discriminability is defined by the difference between the intra-similarity (same audio) and inter-similarity (different audio), as expressed in Equation (9):

$$Discriminability = mean(sim_{intra}) - mean(sim_{inter}) \tag{9}$$

Higher discriminability (lower intra value and higher inter value) indicates better performance for accurate server retrieval.

4.2.1 Discriminability according to FFT length

We compared the discriminative power of Coarse and Fine fingerprints by setting the FFT length to 2048 and 4096. The results are presented in Table 2 and visualized in Figure 5.

As shown in Table 2 and Figure 5, the 4096-setting demonstrated superior discriminability, with the inter/intra ratio increasing by approximately 6% for both Coarse and Fine fingerprints. This indicates that higher frequency resolution enables more precise feature extraction, which is highly effective for accurate content recognition in OTT environments.

4.2.2 Discriminability according to Hop length

With the FFT length fixed at 4096, we tested Hop lengths of 882 and 1470 to evaluate the impact on temporal resolution and bitrate, as shown in Table 3.

While the shorter Hop length (882) showed slightly better discriminability (about 1% higher), a shorter Hop length directly leads to an increase in

Table 2 Quantitative comparison of Coarse/Fine discriminative power according to FFT length

	Coarse			Fine		
FFT Size	Intra	Inter	Intra/Inter	Intra	Inter	Intra/Inter
2048	23.41	29.10	1.257	21.25	26.50	1.261
4096	22.26	29.13	1.332	19.46	25.64	1.338

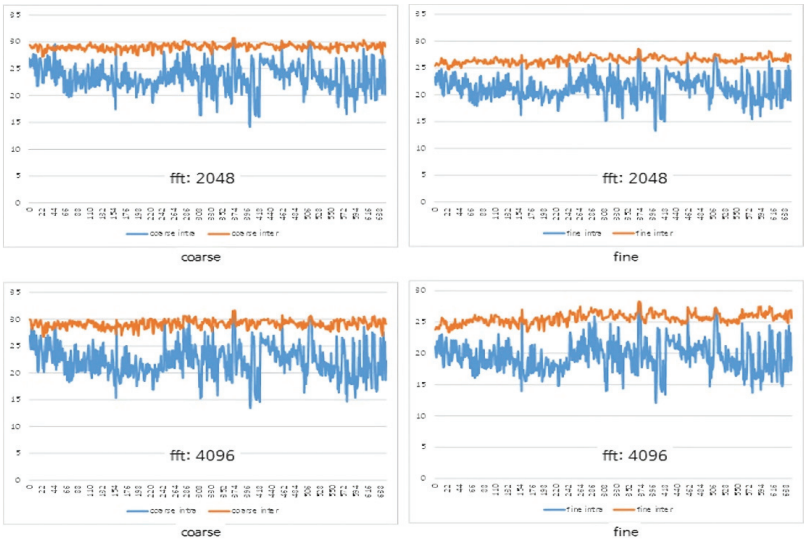

Figure 5 Comparison of discriminative power according to FFT length.

Table 3 Quantitative comparison of Coarse/Fine discriminative power according to Hop length

	Coarse			Fine		
Hop Length	Intra	Inter	Intra/Inter	Intra	Inter	Intra/Inter
882	22.25801	29.13482	1.331695	19.4586	25.63156	1.337812
1470	22.42108	29.13616	1.320801	19.5983	25.6275	1.326934

Table 4 Comparison of discriminative power according to Coarse feature dimension

Coarse Dim	Intra	Inter	Intra/Inter
64	21.72236	29.7282	1.395886
128	22.25801	29.13482	1.331695

bitrate. Considering the balance between system resources and processing speed required for a web server managing large-scale traffic, a Hop length of 1470 was determined to be optimal.

4.2.3 Discriminability according to FFT length

We evaluated the discriminative power based on the dimensionality of the Coarse and Fine features, which directly affects the data payload transmitted over the web, as presented in Tables 4 and 5.

For the Coarse dimension, the 64-bit setting improved discriminability by approximately 5% compared to the 128-bit setting. This reduction

Table 5 Comparison of discriminative power according to Fine feature dimension

Fine Dim	Intra	Inter	Intra/Inter
64	19.45858	25.63156	1.337812
128	18.66766	25.77689	1.40604

Table 6 Optimal parameter settings based on experimental results

FFT Length	FFT Hop	Coarse Dimension	Fine Dimension
4096	1470	64	128

in dimensionality decreases inter-pair similarity (interference), making it a highly suitable standard for fast candidate retrieval in the initial web matching stage. Conversely, for the Fine dimension, the 128-dimensional setting showed about a 5% improvement over the 64-dimensional setting, as it represents the detailed information of the audio more precisely for the final matching stage. Based on these comprehensive analyses, the most effective parameter combination for balancing recognition accuracy and server efficiency is summarized in Table 6.

4.3 Final Recognition Performance

To verify the overall performance of the proposed architecture, we applied the optimal parameters and the 3-stage identification pipeline to the 649 query audio samples. The identification performance was evaluated based on Precision, Recall, and F1-score, defined in Equations (10), (11), and (12).

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

The final performance evaluation results are presented in Table 7 and visualized in Figure 6.

As shown in Table 7 and Figure 6, the precision was extremely high at 99.65%, which clearly demonstrates the effectiveness of the Post-verification stage in eliminating false positives generated in web environments. While a few false negatives occurred in environments with extremely low SNR or very short similarity durations, the overall F1-score of 93.54% confirms

Table 7 Final identification performance results

Metric	Value
Total Tests	649
True Positives (TP)	572
False Negatives (FN)	75
False Positives (FP)	2
Precision	0.9965
Recall	0.8814
F1-score	0.9354

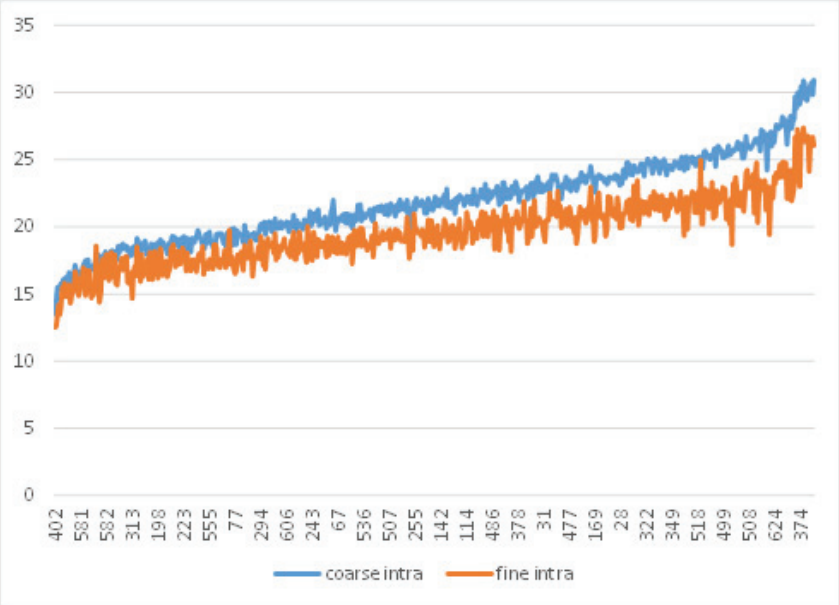

Figure 6 Final recognition performance metrics of the system with optimal parameters.

the excellent balance between accuracy and detection rate. Furthermore, the application of GPU parallel computing in the Coarse matching stage confirmed that the processing speed of the entire web architecture is fully capable of real-time application.

5 Conclusion

This study proposed a multi-stage, audio DNA-based content recognition architecture that robustly addresses unstructured noise occurring

in web-based OTT streaming environments (e.g., via smartphones and browsers) while maximizing the computational efficiency of the web server. The proposed method utilizes STFT and Mel filters to extract lightweight and robust dual-structure (Coarse-Fine) audio features in the client environment. To prevent bottlenecks in large-scale media database environments, the receiving web server executes a three-stage pipeline consisting of Coarse Matching, Fine Matching, and Post-verification. Through various noise-infused experiments, we derived parameters optimized for web transmission and server processing, such as an FFT length of 4096 and a Hop length of 1470. Applying these to 649 real noise test datasets achieved a high performance with a precision of 0.9965, a recall of 0.8814, and an F1-score of 0.9354.

The significance of this study lies in effectively solving the problems of increased false positive rates and server overloads that traditional clean-audio-based fingerprinting technologies face in web environments. The proposed architecture is expected to be practically applicable to future cloud-based real-time streaming monitoring systems or large-scale web content Digital Rights Management (DRM) solutions. Future research will focus on integrating real-time streaming protocols, such as Web Sockets, to evaluate and optimize in-browser audio recognition performance in live broadcasting environments.

Acknowledgment

This work was supported by the Software Copyright Research and Development Program funded by the Ministry of Culture, Sports and Tourism and managed by the Korea Institute of Culture Technology Evaluation and Planning (KCTEP) (Project Name: Development of Copyright Technology for OTT Contents Copyright Protection Technology Development and Application, Project Number: RS-2023-00225267, Contribution Rate: 100%).

References

[1] Korea Information Society Development Institute, *Digital Content Industry Trends Report*, KISDI, 2021.

[2] A. Wang, "An Industrial-Strength Audio Search Algorithm," in *Proceedings of the 4th International Conference on Music Information Retrieval*, pp. 7–13, 2003.

[3] J.-S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo, "Audio Fingerprinting Based on Normalized Spectral Subband Moments," *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 209–212, 2006, doi:10.1109/LSP.2005.863678.

[4] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-by-Example Applications," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 801–806, 2011.

[5] Q. Xiao, M. Suzuki, and K. Kita, "Fast Hamming Space Search for Audio Fingerprinting Systems," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 133–138, 2011.

[6] A. Baez-Suarez, N. Shah, J. A. Nolazco-Flores, S.-H. S. Huang, O. Gnawali, and W. Shi, "SAMAF: Sequence-to-Sequence Autoencoder Model for Audio Fingerprinting," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–23, 2020, doi:10.1145/3380828.

[7] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural Audio Fingerprint for High-Specific Audio Retrieval Based on Contrastive Learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3025–3029, 2021, doi:10.1109/ICASSP39728.2021.9414337.

[8] J. Six, "Olaf: A Lightweight, Portable Audio Search System," *Journal of Open Source Software*, vol. 8, no. 87, article 5459, 2023, doi:10.21105/joss.05459.

[9] A. Marafioti, N. Holighaus, and P. Majdak, "Time-Frequency Phase Retrieval for Audio – The Effect of Transform Parameters," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3585–3596, 2021, doi:10.1109/TSP.2021.3088581.

[10] Y. Zhou, X. Li, C. Xiong, H. Yao, and C. Qin, "A Survey of Perceptual Hashing for Multimedia," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025, doi:10.1145/3727880.

Biographies

Byeongchan Park received the bachelor's degree in computer engineering through the Academic Credit Bank System in 2015, and the master's and Doctor of Philosophy degrees in computer science and engineering from Soongsil University, Korea, in 2018 and 2023, respectively. He is currently a Visiting Professor at the Department of Computer Science, Soongsil University. His research areas include copyright technology and the promotion of its utilization.

Sun-Jib Kim is currently a Professor in the School of IT, Hansei University, Korea. His research areas include information security, the Internet of Things (IoT), cloud computing, and AI system authentication.

Seok-Yoon Kim received the B.S. degree in electrical and electronic engineering from Seoul National University, Korea, in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin, USA, in 1990 and 1993, respectively. From 1982 to 1987, he was a Researcher with the Electronics and Telecommunications Research Institute (ETRI). From 1993 to 1995, he worked as a Senior Researcher at Motorola. Since 1995, he has been a Professor at Soongsil University. His main research interests include copyright protection and the promotion of its utilization.

Youngmo Kim received the B.S., M.S., and Ph.D. degrees in computer engineering from Daejeon University, Korea, in 2003, 2005, and 2011, respectively. Since 2012, he has been a Professor at Soongsil University. His main research interests include copyright protection and the promotion of its utilization.

Yoontaek Sung received his Ph.D. in Communication from Sungkyunkwan University, Korea. He is currently a Principal Research Fellow at the Media Advertising Research Institute of the Korea Broadcasting Advertising Corporation (KOBACO). His research interests include the application of technology and data-driven approaches to advancing the media and advertising industries, on which he continues to conduct related projects and studies.