Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 8, No.3, July 2026, pp. 02603033-01 ~ 02603033-011
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v8i3.2744

Optimizing Acoustic Fingerprinting for Synchronized Audio Binary Matching

Andi Bahtiar Semma[1,2*], Kusrini[2], Arief Setyanto[2], Bruno da Silva[1], An Braeken[1]

[1]ETRO-RDI, Vrije Universiteit Brussels, Pleinlaan 2, Brussel, 1050, Belgium

[2]Doctoral of Computer Science, Universitas AMIKOM Yogyakarta, Condong Catur, Yogyakarta, 55281, DIY, Indonesia

*andi.semma@vub.be

Abstract. As interconnected devices proliferate, secure and efficient pairing methods are critical. Environmental acoustic signals offer a promising solution, but their effectiveness depends on robust audio features that perform well across varying conditions. This study investigates optimal audio features for fingerprinting, focusing on synchronized audio in time-frequency domains. Six diverse datasets were collected across controlled environments to simulate real-world scenarios. Thirteen audio features were extracted and analyzed for robustness across distances, devices, scenes, and sample lengths. Cosine similarity assessed consistency, while the Youden index determined thresholds. The Mel Spectrogram, particularly with 5-second samples, achieved an AUC of 0.8758 and a J-statistic of 0.7278. Augmenting it with Tonnetz and spectral bandwidth yielded the highest performance (AUC: 0.9346, J-statistic: 0.7955, Accuracy: 0.8320, recall: 0.9648), demonstrating the potential of combining robust base features with complementary acoustic characteristics for reliable device pairing.

Keywords: Acoustic fingerprinting, device pairing, IoT security, acoustic feature extraction

(Received 2025-09-15, Revised 2026-02-16, Accepted 2026-04-02, Available Online by 2026-07-10)

1. Introduction

The proliferation of Internet of Things (IoT) devices has brought significant challenges in device pairing, necessitating a balance between connectivity, security, and user experience (UX) [1], [2]. As interconnected devices grow exponentially [3], establishing secure connections becomes increasingly critical. IoT device pairing must account for diverse communication protocols including Bluetooth Low Energy (BLE), Wi-Fi, Zigbee, and Z-Wave, each offering distinct trade-offs in range, power consumption, and data rates [4], [5]. Security remains paramount, with cryptographic key exchange and authentication mechanisms needed to safeguard against man-in-the-middle attacks and unauthorized access [6]. Advanced protocols such as Transport Layer Security (TLS) and Datagram Transport Layer

Security (DTLS) are increasingly adapted for resource-constrained IoT devices [7]. Meanwhile, "zero-touch" provisioning is gaining traction, allowing secure configuration without manual intervention [8]. The device pairing process must balance security and ease of use, with techniques such as Near Field Communication (NFC) tap-to-pair and QR code scanning explored to simplify UX without compromising security [9].

1.1 Acoustic-Based Device Pairing

Environmental acoustic signals offer a promising avenue for secure and efficient device pairing [10]. This approach leverages the ambient soundscape, a complex mixture of acoustic events, background noise, and spectral characteristics unique to each environment. These acoustic fingerprints offer rich features extractable through spectral analysis, cepstral analysis, and machine learning algorithms [11]. This paradigm capitalizes on context-based security, where the shared acoustic environment serves as a natural, difficult-to-forge authentication factor [12], [13]. Unlike conventional cryptographic key exchange, which may be susceptible to man-in-the-middle attacks, acoustic-based pairing leverages the physical co-presence of devices as an implicit security measure. However, this method also introduces new security considerations, such as acoustic eavesdropping and replay attacks [14], [15]. Beyond security, environmental acoustic signal-based IoT device pairing offers an opportunity to streamline the often cumbersome device association process, aligning with calm technology principles where complex technological processes recede into the background of user awareness [16].

1.2 Robustness Analysis of Acoustic Features

Acoustic signals are inherently complex, comprising temporal characteristics, spectral components, and amplitude variations that interact in intricate ways. The complexity is compounded by acoustic signals being often non-stationary, meaning their statistical properties change over time [17], [18]. Environmental factors such as reflections, refractions, absorptions, background noise, and interfering sources can significantly alter signal characteristics [19], [20]. These effects necessitate sophisticated signal processing techniques for accurate feature characterization. Robustness in acoustic signal analysis extends to machine learning and AI applications, where extracting meaningful and consistent features becomes critical for tasks such as voice assistants and acoustic-based diagnostics. Researchers have developed advanced techniques including adaptive filtering methods and deep learning approaches that learn complex hierarchical representations of acoustic features [21].

However, research into which audio features are optimal for acoustic fingerprinting remains necessary, as no single feature provides universal performance. Each type of sound has distinct characteristics, and features effective for one sound type may underperform for another. For synchronized audio involving the time-frequency domain, features must be compact yet capture the signal's essential characteristics [22], [23]. Comprehensive robustness analysis helps overcome challenges posed by non-stationary signals and environmental factors, leading to more accurate and consistent systems. This study addresses this gap by systematically evaluating 13 audio features across varied environmental conditions to identify the most robust features for synchronized audio binary matching.

2. **Methods**

2.1 Data Acquisition

A diverse audio dataset was compiled by capturing audio samples across controlled environmental conditions, encompassing varied noise levels and ambient sounds including jungle [24], café [25], machine/factory[26], street [27], and multi-audio source settings. Recordings were captured from multiple device positions and distances (1 m, 2 m, 3 m, 5 m) within each environment. Standardized recording equipment was used throughout data collection (22050 Hz, 16-bit) with multiple sound

sources. Figure 1 illustrates the recording setup. The dataset comprises 91,756 samples organized as 50 chunks per scene, distance, sample length, and feature configuration.

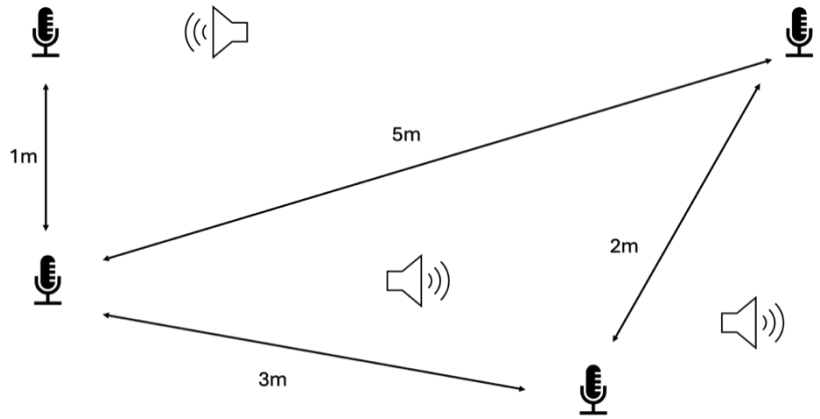

Figure 1. Data Acquisition setup

2.2 Pre-processing and Feature Extraction

Amplitude standardization was applied to normalize audio data. Thirteen features were extracted: Mel-frequency cepstral coefficients (MFCCs), chroma_cens, chroma_cqt, chroma_stft, mel_spectrogram, onset_strength, spectral_bandwidth, spectral_centroid, spectral_contrast, spectral_flatness, spectral_rolloff, tempo, tonnetz, and zero_crossing_rate.

Spectral features form the backbone of audio fingerprinting. MFCCs capture timbral characteristics crucial for identifying unique audio signatures [28]. Spectral centroid, bandwidth, contrast, flatness, and rolloff collectively describe frequency distribution and quality, helping distinguish audio samples with high precision [29], [30]. Chroma variants (CENS, CQT, STFT) represent harmonic and melodic content by mapping the spectrum into 12 pitch classes, making them invariant to octave changes while preserving musical character [31]. Onset strength and tempo capture rhythmic elements and temporal evolution. Zero crossing rate provides information about frequency content and noisiness. Mel spectrograms provide a comprehensive frequency representation aligned with human auditory perception. Tonnetz features capture harmonic relationships in a geometric space, making fingerprints more unique for audio matching.

2.3 Robustness Score

Robustness is quantified through a combination of high cosine similarity and low standard deviation. Cosine similarity measures the angle between vectors, with values closer to 1 indicating greater similarity. A high cosine similarity indicates that fundamental acoustic characteristics remain consistent despite changing recording conditions. Concurrently, a low standard deviation suggests minimal variability in feature measurements, reinforcing stability across diverse scenarios. The robustness score is expressed by Equation 1:

$$Robustness\ Score = \frac{\sqrt{(\ CS^2 + (1 - ASDCS)^2)}}{\sqrt{2}} \quad (1)$$

Where CS = Average Cosine Similarity and ASDCS = Average Standard Deviation of Cosine Similarity.

This approach uses the Euclidean norm (L2 norm) of the two metrics. The square root of the sum of squares combines the metrics while preventing any single metric from dominating. Using (1 - ASDCS) accounts for its inverse relationship with robustness. Since the maximum possible value for each metric is 1 (all values scaled from 0 to 1), dividing by √2 ensures the final score is between 0 and 1. This approach naturally penalizes inconsistency: if one metric is significantly lower, it has a pronounced effect on lowering the overall score.

2.4 Matching Criteria

Effectiveness of audio features is assessed based on their ability to correctly identify matching and non-matching audio samples. A correct match (True Positive) occurs when two audio samples share the same scene and chunk (time segment), leveraging the principle that audio signals are characterized by unique patterns persisting across different recordings of the same acoustic event [32]. Any other combination (same scene but different time, same time but different scene, or both different) is considered a non-match (True Negative). Features that consistently yield high similarity scores for true matches and low scores for non-matches are considered more effective.

2.5 Optimal Threshold Determination

The optimal threshold is determined using Receiver Operating Characteristic (ROC) curve analysis, computing the Area Under the Curve (AUC) and Youden's J statistic, formulated in Equation 2:

$$J = Sensitifity - Speciticity - 1 = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \quad (2)$$

Youden's J identifies the specific optimal threshold that maximizes both sensitivity and specificity simultaneously [33]. Geometrically, this corresponds to the point where the ROC curve extends furthest above the diagonal line, as depicted in Figure 2. Youden's J spans from -1 to 1, where 1 indicates perfect classification, 0 represents random chance, and negative values suggest worse-than-random performance. Additionally, classical metrics including accuracy, precision, recall, and F1 score are calculated for comprehensive evaluation.

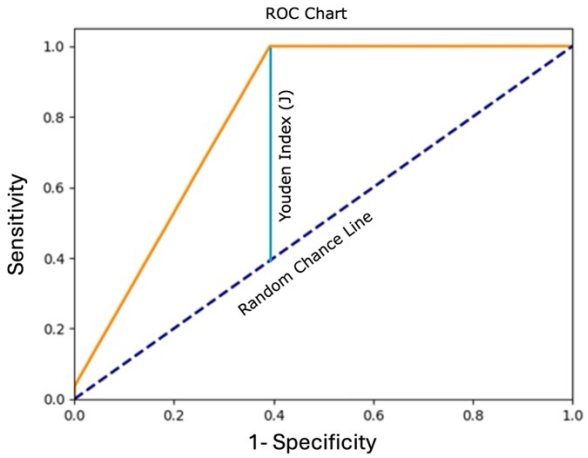

Figure 2. Youden Index (J)

3. Results and Discussion

3.1 Robustness Across Conditions

Spectral features demonstrate exceptional robustness, with spectral centroid (0.9982), spectral bandwidth (0.9975), and spectral rolloff (0.9961) leading the results. These high scores indicate that overall frequency distribution and energy characteristics remain remarkably stable despite changes in physical distance, scene, and sample length. Zero crossing rate (0.9931) also shows excellent robustness, indicating well-preserved temporal characteristics. Chroma features exhibit strong robustness, particularly Chroma CENS (0.9828) and Chroma CQT (0.9693), suggesting pitch class information remains largely consistent. MFCCs (0.9797) and onset strength (0.9650) also demonstrate high robustness. While most features show high robustness (above 0.9), the mel spectrogram (0.7850) displays moderate robustness, suggesting some susceptibility to environmental changes. The tonnetz feature (0.6477) shows the lowest robustness, indicating that harmonic relationships are more sensitive to recording conditions.

Distance-based analysis reveals that spectral features maintain consistently high robustness across all distances. At 1 m vs 2 m, spectral bandwidth achieves 0.9990, spectral centroid 0.9984, and spectral rolloff 0.9975. MFCCs show strong robustness (0.9895) at close distances. At 5 m, a noticeable decrease occurs for some features: MFCCs drop to 0.9714 and Chroma CQT to 0.9649, though spectral centroid (0.9979) and spectral bandwidth (0.9959) remain excellent. The mel spectrogram shows moderate robustness across all distances (0.7783-0.7889), while tonnetz consistently scores lowest (0.6344-0.6638), confirming sensitivity to recording distance variation.

Scene-based analysis shows spectral features consistently demonstrate high robustness across different acoustic environments, with spectral bandwidth and centroid typically scoring above 0.99. Spectral flatness shows more variability (0.85-0.94) but still outperforms many other feature types. Chroma CENS demonstrates the highest robustness within chroma features, consistently above 0.97. MFCC scores range from 0.9757 to 0.9864, confirming their effectiveness at capturing the spectral envelope resistant to environmental changes. Zero crossing rate stands out with scores consistently above 0.98, attributable to its fundamental nature. Onset strength shows good robustness (0.9525-0.9755), indicating reliable capture of musical events across varying acoustic conditions. Tonnetz consistently shows lowest robustness (0.6173-0.7164) as its complex harmonic representations are easily disrupted by environmental noise.

Sample length-based analysis reveals spectral centroid leads at 0.9980 for 1-second samples, increasing slightly to 0.9983 for 3 and 5-second samples. Spectral bandwidth improves from 0.9969 to 0.9981 across sample lengths. Chroma CENS robustness increases from 0.9822 (1 s) to 0.9835 (5 s), while Chroma STFT shows more noticeable improvement from 0.9154 to 0.9215. MFCCs maintain high robustness across all sample lengths (0.9792-0.9801). Mel spectrogram exhibits a slight decrease for longer samples (0.7917 to 0.7812). Tonnetz, while lowest overall, shows the most significant improvement with increased sample length (0.6288 to 0.6663), indicating that harmonic representations benefit considerably from longer durations.

3.2 Optimal Threshold and Matching Performance

Figure 3 presents the ROC curves and Youden Index for the best-performing features.

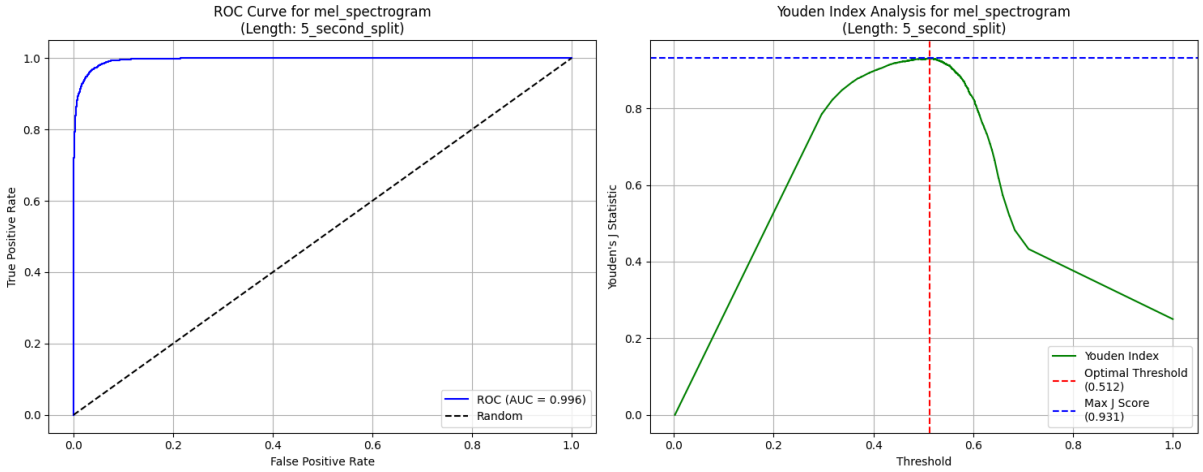

Figure 3. ROC Curve and Youden Index of Best Feature

Single-Feature Performance

The Mel Spectrogram with 5-second samples demonstrated the strongest single-feature performance, achieving AUC 0.8758 and J-statistic 0.7278. The optimal threshold was 0.5368, yielding accuracy of 0.7859, precision of 0.24, recall of 0.9545, and F1 score of 0.3836. Zero crossing rate (5 s) achieved AUC 0.8442 and accuracy 0.7354 with threshold 0.9832, showing high recall (0.9596) but low precision (0.2036). Spectral centroid (5 s) achieved AUC 0.8686 at threshold 0.9961. Notably, spectral bandwidth achieved the highest accuracy in the dataset at 0.8946 with the highest precision among all features (0.3613) and F1 score of 0.468. Chroma CENS (5 s) achieved the highest AUC at 0.8889 with threshold 0.9715. Longer samples (5 s) consistently produced better results across most features, though this benefit must be balanced against increased computational requirements. Detailed data is presented in Table 2.

Table 1. Chunk matching performance (single feature)

Sample Length	Feature Name	Optimal Threshold	AUC	J Statistic	Accuracy	Precision	Recall	F1
5 seconds	mel_spectrogram	0.5368	0.8758	0.7278	0.7859	0.2400	0.9545	0.3836
5 seconds	zero_crossing_rate	0.9832	0.8442	0.6782	0.7354	0.2036	0.9596	0.336
3 seconds	mel_spectrogram	0.4714	0.8633	0.6778	0.7068	0.1913	0.9924	0.3208
5 seconds	spectral_centroid	0.9961	0.8686	0.6504	0.7204	0.1932	0.947	0.3209
5 seconds	chroma_cens	0.9715	0.8889	0.6232	0.7777	0.2188	0.8510	0.3481
3 seconds	tonnetz	0.2494	0.8797	0.609	0.7819	0.2193	0.8308	0.3470
1 second	mel_spectrogram	0.5919	0.8487	0.5947	0.7403	0.1941	0.8636	0.3170
3 seconds	chroma_cens	0.9702	0.875	0.5866	0.7567	0.201	0.8359	0.3240
5 seconds	tonnetz	0.2411	0.8728	0.5825	0.7311	0.1882	0.8611	0.3089
5 seconds	spectral_bandwidth	0.9987	0.8632	0.5761	0.8946	0.3613	0.6641	0.4680

Combined-Feature Performance

The combination of mel_spectrogram-tonnetz-spectral_bandwidth consistently demonstrated superior performance, particularly at 5-second sample length. This configuration achieved AUC 0.9346, accuracy 0.8320, and F1 score 0.1018. Precision and recall were 0.0538 and 0.9648, indicating strong positive case identification despite numerous false positives. A clear performance degradation emerges as sample lengths decrease: AUC drops from 0.9346 (5 s) to 0.9082 (3 s) to 0.8620 (1 s) for the top-

performing combination. Optimal thresholds show interesting variations: mel_spectrogram-tonnetz-spectral_bandwidth consistently requires lower thresholds (0.59-0.60), while other combinations demand significantly higher thresholds (0.75-0.88). A noteworthy observation is the stark contrast between precision and recall: recall values remain consistently high (0.87-0.96), while precision values are notably low (0.02-0.05), suggesting the models excel at identifying positive cases but struggle with false positives. The second-best combination was mel_spectrogram-mfcc-spectral_bandwidth-chroma_cens, consistently trailing the top performer across all metrics. Detailed data is presented in Table 3.

Table 2. Chunk matching performance (combined features)

Sample Length	Feature Name	Optimal Threshold	AUC	J Statistic	Accuracy	Precision	Recall	F1
5 seconds	mel_spectrogram-tonnetz-spectral_bandwidth	**0.5958**	0.9346	0.7955	0.8320	0.0538	0.9648	0.1018
5 seconds	mel_spectrogram-mfcc-spectral_bandwidth-chroma_cens	**0.8712**	0.9106	0.7571	0.8299	0.0513	0.9281	0.0973
5 seconds	mel_spectrogram-mfcc-chroma_cens	**0.8288**	0.9096	0.7520	0.8237	0.0497	0.9293	0.0943
5 seconds	mel_spectrogram-mfcc	**0.7533**	0.9059	0.7492	0.8169	0.0481	0.9335	0.0914
3 seconds	mel_spectrogram-tonnetz-spectral_bandwidth	**0.5915**	0.9082	0.7473	0.7949	0.0341	0.9537	0.0658
3 seconds	mel_spectrogram-mfcc-spectral_bandwidth-chroma_cens	**0.8664**	0.8692	0.6828	0.7378	0.0267	0.9466	0.0519
3 seconds	mel_spectrogram-mfcc-chroma_cens	**0.8229**	0.8687	0.6799	0.7348	0.0264	0.9466	0.0513
3 seconds	mel_spectrogram-mfcc	**0.7472**	0.8621	0.6762	0.7318	0.0261	0.9461	0.0508
1 second	mel_spectrogram-tonnetz-spectral_bandwidth	**0.5970**	0.8620	0.6104	0.7318	0.0242	0.8797	0.0470
1 second	mel_spectrogram-mfcc-spectral_bandwidth-chroma_cens	**0.8823**	0.8263	0.5578	0.6878	0.0206	0.8714	0.0403

Discussion

The performance analysis reveals a significant difference in difficulty between scene matching and chunk matching. Scene matching achieves outstanding results: mel_spectrogram reaches AUC 0.9957 and F1 0.8931 for 5-second samples, MFCC achieves AUC 0.9847 and F1 0.8077, and spectral_centroid and chroma_cens show strong performance with AUC values of 0.9769 and 0.9622. These high metrics reflect that scene-level audio characteristics are relatively distinct. In contrast, chunk matching is substantially more challenging. The mel_spectrogram, while still the best-performing single feature, achieves AUC 0.8758 and F1 of only 0.3836 for 5-second samples. Low precision values (0.2400 for mel_spectrogram, 0.2036 for zero_crossing_rate) highlight the difficulty in avoiding false positives in chunk matching. Certain features, including mel_spectrogram, MFCC, chroma_stft, and chroma_cens, demonstrate consistent performance across both scenarios, making them the most reliable choices for dual-purpose applications. Mel_spectrogram stands out as the most robust choice, particularly at 5-second sample length.

The combination of features significantly improves AUC, J Statistic, and accuracy compared to single-feature models. The 5-second mel_spectrogram alone achieves AUC 0.8758, while the combined mel_spectrogram-tonnetz-spectral_bandwidth achieves AUC 0.9346, an improvement of 0.0588.

Similarly, the J Statistic improves from 0.7278 to 0.7955, and accuracy increases from 0.7859 to 0.8320. These improvements confirm the benefit of leveraging multiple features to capture more nuanced patterns. However, precision and F1 scores decline with feature combinations due to the imbalanced nature of the dataset, where only 1.5% of labels are true. Precision drops from 0.2400 (single mel_spectrogram) to 0.0538 (combined mel_spectrogram-tonnetz-spectral_bandwidth). Despite this, recall remains consistently high (best: 0.9648 for the top combination at 5 s). These findings align with prior research [31], [34], confirming that feature effectiveness is application-dependent and no single feature is universally optimal. This study's primary limitation is class imbalance (1.5% positive labels), which drives low precision and F1 scores particularly in combined-feature configurations. Future work should address this through resampling, data augmentation, or weighted cosine similarity that assigns higher weights to discriminative features such as mel_spectrogram. Introducing audio landmarks (e.g., clap, buzz) could also provide more discriminative data for matching.

4. Conclusion

This study systematically analyzed 13 audio features for robustness in synchronized audio binary matching. Spectral features, particularly spectral centroid, spectral bandwidth, and spectral rolloff, demonstrated the highest robustness across all conditions (scores above 0.996). Mel spectrogram with 5-second samples showed the strongest single-feature matching performance (AUC 0.8758, J-statistic 0.7278). Combining mel_spectrogram, tonnetz, and spectral_bandwidth achieved the highest overall performance (AUC 0.9346, recall 0.9648). Longer sample lengths consistently improved matching reliability. These findings reinforce the importance of tailoring feature selection to specific use cases and leveraging feature combinations to capture nuanced audio characteristics. While class imbalance limits precision, this study provides a foundational approach for audio binary matching. Future work should address imbalance and explore weighted feature fusion.

Declaration of AI and AI assisted technologies in the writing process

During the preparation of this work, the author(s) used Grammarly in order to do grammar check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Ding, M. Nemati, C. Ranaweera, and J. Choi, 'IoT connectivity technologies and applications: A survey', *IEEE Access*, vol. 8, pp. 67646–67673, 2020, Accessed: Apr. 24, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9057670/

[2] A. Karale, 'The challenges of IoT addressing security, ethics, privacy, and laws', *Internet Things*, vol. 15, p. 100420, 2021, Accessed: Aug. 15, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660521000640

[3] S. Bansal and D. Kumar, 'IoT Ecosystem: A Survey on Devices, Gateways, Operating Systems, Middleware and Communication', *Int. J. Wirel. Inf. Netw.*, vol. 27, no. 3, pp. 340–364, Sep. 2020, doi: https://doi.org/10.1007/s10776-020-00483-7.

[4] D. R. Komilov, 'Application of zigbee technology in IOT', *Int. J. Adv. Sci. Res.*, vol. 3, no. 09, pp. 343–349, 2023, Accessed: Aug. 15, 2024. [Online]. Available: https://sciencebring.com/index.php/ijasr/article/view/425

[5] J. Yang, C. Poellabauer, P. Mitra, and C. Neubecker, 'Beyond beaconing: Emerging applications and challenges of BLE', *Ad Hoc Netw.*, vol. 97, p. 102015, 2020, Accessed: Aug. 15, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870518307170

[6] P. M. Chanal and M. S. Kakkasageri, 'Security and Privacy in IoT: A Survey', *Wirel. Pers. Commun.*, vol. 115, no. 2, pp. 1667–1693, Nov. 2020, doi: https://doi.org/10.1007/s11277-020-07649-9.

[7] G. Restuccia, H. Tschofenig, and E. Baccelli, 'Low-power IoT communication security: On the performance of DTLS and TLS 1.3', in *2020 9th IFIP International Conference on Performance Evaluation and Modeling in Wireless Networks (PEMWN)*, IEEE, 2020, pp. 1–6. Accessed: Aug. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9293085/

[8] I. Al Ridhawi, M. Aloqaily, F. Karray, M. Guizani, and M. Debbah, 'Realizing the tactile internet through intelligent zero touch networks', *IEEE Netw.*, 2022, Accessed: Aug. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/iel7/65/7593428/09921199.pdf

[9] S. S. Madsen, A. Q. Santos, and B. N. Jørgensen, 'A QR code based framework for auto-configuration of IoT sensor networks in buildings', *Energy Inform.*, vol. 4, no. S2, p. 46, Sep. 2021, doi: https://doi.org/10.1186/s42162-021-00152-w.

[10] A. B. Semma, Kusrini, A. Setyanto, B. da Silva, and A. Braeken, 'Authenticating devices based on audio feature selection with scene-specific tuning and landmark augmentation', *Intell. Syst. Appl.*, vol. 28, p. 200593, Dec. 2025, doi: https://doi.org/10.1016/j.iswa.2025.200593.

[11] J. Lopez-Ballester, S. Felici-Castell, J. Segura-Garcia, and M. Cobos, 'AI-IoT platform for blind estimation of room acoustic parameters based on deep neural networks', *IEEE Internet Things J.*, vol. 10, no. 1, pp. 855–866, 2022, Accessed: Aug. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9873918/

[12] S. A. Anand, J. Liu, C. Wang, M. Shirvanian, N. Saxena, and Y. Chen, 'Echovib: Exploring voice authentication via unique non-linear vibrations of short replayed speech', in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 67–81.

[13] Z. Wang, Y. Ren, Y. Chen, and J. Yang, 'Earable authentication via acoustic toothprint', in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2390–2392.

[14] D. Han *et al.*, '(In) secure acoustic mobile authentication', *IEEE Trans. Mob. Comput.*, vol. 21, no. 9, pp. 3193–3207, 2021.

[15] Y. Ren *et al.*, 'Proximity-Echo: Secure Two Factor Authentication Using Active Sound Sensing', in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, May 2021, pp. 1–10. doi: https://doi.org/10.1109/INFOCOM42981.2021.9488866.

[16] A. B. Semma, Kusrini, A. Setyanto, B. Da Silva, and A. Braeken, 'Exploring the Landscape of Acoustic Authentication: Techniques, Performance and Security', in *2024 IEEE International Conference on Technology, Informatics, Management, Engineering and Environment (TIME-E)*, Aug. 2024, pp. 61–68. doi: https://doi.org/10.1109/TIME-E62724.2024.10919631.

[17] S. Cortesi, C. Vogt, E. Reinschmidt, and M. Magno, 'Latency and Power Consumption in 2.4GHz IoT Wireless Mesh Nodes: An Experimental Evaluation of Bluetooth Mesh and

Wirepas Mesh', *2023 19th Int. Conf. Wirel. Mob. Comput. Netw. Commun. WiMob*, pp. 200–205, 2023, doi: https://doi.org/10.1109/WiMob58348.2023.10187799.
[18] D. Uchida, Y. Yonezawa, and K. Akita, 'Measurement-Based Latency Evaluation and the Theoretical Analysis for Massive IoT Applications Using Bluetooth Low Energy', *2023 IEEE 97th Veh. Technol. Conf. VTC2023-Spring*, pp. 1–5, 2023, doi: https://doi.org/10.1109/VTC2023-Spring57618.2023.10200332.
[19] H. Nam, S.-H. Kim, and Y.-H. Park, 'Filteraugment: An acoustic environmental data augmentation method', in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4308–4312. Accessed: Sep. 25, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9747680/
[20] S. S. Sethi *et al.*, 'Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set', *Proc. Natl. Acad. Sci.*, vol. 117, no. 29, pp. 17049–17055, Jul. 2020, doi: https://doi.org/10.1073/pnas.2004702117.
[21] J. Abeßer, 'A review of deep learning based methods for acoustic scene classification', *Appl. Sci.*, vol. 10, no. 6, p. 2020, 2020, Accessed: Aug. 15, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/10/6/2020
[22] C. Bisogni, V. Loia, M. Nappi, and C. Pero, 'Acoustic features analysis for explainable machine learning-based audio spoofing detection', *Comput. Vis. Image Underst.*, vol. 249, p. 104145, Dec. 2024, doi: https://doi.org/10.1016/j.cviu.2024.104145.
[23] G. Sharma, K. Umapathy, and S. Krishnan, 'Trends in audio signal feature extraction methods', *Appl. Acoust.*, vol. 158, p. 107020, Jan. 2020, doi: https://doi.org/10.1016/j.apacoust.2019.107020.
[24] Trundlefly, 'Amazon Jungle Morning | Royalty-free Music'. Accessed: Sep. 10, 2024. [Online]. Available: https://pixabay.com/sound-effects/amazon-jungle-morning-24939/
[25] Hitrison, 'Restaurant Sounds (Sunny Point Cafe)'. Accessed: Sep. 10, 2024. [Online]. Available: https://pixabay.com/sound-effects/restaurant-sounds-sunny-point-cafe-25092/
[26] CSNmedia, 'industrial sounds | Royalty-free Music'. Accessed: Sep. 10, 2024. [Online]. Available: https://pixabay.com/sound-effects/industrial-sounds-25817/
[27] Aatreya_v, 'Busy Street Ambience | Royalty-free Music'. Accessed: Sep. 10, 2024. [Online]. Available: https://pixabay.com/sound-effects/busy-street-ambience-195884/
[28] D. Prabakaran and S. Sriuppili, 'Speech processing: MFCC based feature extraction techniques-an investigation', in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012009. Accessed: Aug. 14, 2024. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1717/1/012009/meta
[29] I. A. Thukroo, R. Bashir, and K. J. Giri, 'A comparison of cepstral and spectral features using recurrent neural network for spoken language identification', *Comput. Artif. Intell.*, vol. 2, no. 1, 2024, Accessed: Mar. 19, 2025. [Online]. Available: http://2oldtoojsacad.acad-pub.com/index.php/CAI/article/view/440
[30] A. B. Semma, M. Saerozi, K. Kusrini, A. Syukur, and A. Maimun, 'An Extreme Programming Approach to Streamlining Thesis Writing', *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 13, no. 6, Art. no. 6, Dec. 2023, doi: https://doi.org/10.18517/ijaseit.13.6.18701.
[31] G. Sharma, K. Umapathy, and S. Krishnan, 'Trends in audio signal feature extraction methods', *Appl. Acoust.*, vol. 158, p. 107020, Jan. 2020, doi: https://doi.org/10.1016/j.apacoust.2019.107020.

[32] M. Crocco, M. Cristani, A. Trucco, and V. Murino, 'Audio Surveillance: A Systematic Review', *ACM Comput Surv*, vol. 48, no. 4, p. 52:1-52:46, Feb. 2016, doi: https://doi.org/10.1145/2871183.

[33] E. F. Schisterman, N. J. Perkins, A. Liu, and H. Bondell, 'Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples':, *Epidemiology*, vol. 16, no. 1, pp. 73–81, Jan. 2005, doi: https://doi.org/10.1097/01.ede.0000147512.81966.ba.

[34] C. Bisogni, V. Loia, M. Nappi, and C. Pero, 'Acoustic features analysis for explainable machine learning-based audio spoofing detection', *Comput. Vis. Image Underst.*, vol. 249, p. 104145, 2024, Accessed: Mar. 13, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314224002261