

Article

A Comparative Investigation of Cepstral Feature Extraction Methods for Deepfake Speech Detection

Nida Akıncı and Erdal Özbay *

Department of Computer Engineering, Firat University, 23119 Elazığ, Türkiye; nidakinci@gmail.com

* Correspondence: erdalozbay@firat.edu.tr

Abstract

The widespread adoption of voice-based authentication systems has been accompanied by an escalating threat from deep learning-based synthetic speech generation techniques. This study presents a comparative and experimental investigation of cepstral feature extraction methods for deepfake speech detection. Specifically, Mel-Frequency Cepstral Coefficients (MFCC), Linear-Frequency Cepstral Coefficients (LFCC), and Constant-Q Cepstral Coefficients (CQCC) are systematically evaluated with respect to their frequency scaling characteristics, spectral resolution properties, and capacity to capture artifacts specific to synthetic speech production. Experiments were conducted on 5571 audio samples drawn from the ASVspoof 2021 Logical Access evaluation partition, with all methods assessed under identical classification conditions using a linear Support Vector Machine. Results indicate that CQCC attains the highest numerical performance, achieving 83.59% accuracy, 89.15% ROC-AUC, and 15.83% Equal Error Rate (EER); however, the performance difference between MFCC and CQCC does not reach statistical significance ($p = 0.202$). Five-fold cross-validation corroborates this finding (CQCC: $87.89\% \pm 0.81\%$). McNemar's test confirms that the performance difference between LFCC and CQCC is statistically significant ($p = 0.036$). A fine-grained attack-wise analysis across 13 spoofing systems reveals that no single feature representation consistently outperforms the others across all attack types; CQCC achieves the highest accuracy on 6 out of 13 systems, while MFCC remains competitive on several attack categories. The overall findings indicate that deepfake detection performance is highly sensitive not only to the classifier architecture but also to the choice of frequency scale, cepstral transformation design, and data conditions. Empirical motivation is provided that multi-feature strategies integrating complementary frequency representations may offer more robust and generalizable detection solutions.

Keywords: deepfake speech detection; anti-spoofing countermeasure; constant-Q transform; cepstral analysis

Academic Editor: Douglas
O'Shaughnessy

Received: 1 June 2026

Revised: 27 June 2026

Accepted: 2 July 2026

Published: 4 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Voice-based technologies have become indispensable in a wide range of critical domains, including biometric authentication, intelligent voice assistant systems, call center automation, and human–computer interaction [1,2]. The inherently personal and distinctive acoustic structure of speech signals positions them as a highly valuable biometric modality for security-sensitive applications. However, the rapid advancement

of generative artificial intelligence models has introduced a new and formidable threat landscape that fundamentally challenges the reliability of voice-based systems. Deepfake speech synthesized using deep learning-based generation pipelines is now capable of replicating the acoustic characteristics of a target speaker with remarkable fidelity, thereby undermining the core assumptions upon which voice-based verification systems are built [3,4].

Deepfake speech generation predominantly relies on two technical paradigms: Text-to-Speech (TTS) synthesis and Voice Conversion (VC). Early statistical approaches to speech synthesis laid the groundwork for contemporary deepfake generation techniques [5]; the subsequent maturation of deep generative architectures such as WaveNet [6] and Tacotron 2 [7] substantially elevated the perceptual quality of synthetic speech, progressively blurring the acoustic boundary between genuine and fabricated utterances. These systems model the target speaker's frequency structure, temporal dynamics, and prosodic characteristics with high precision, producing synthetic outputs that human listeners find increasingly difficult to distinguish. Consequently, deepfake speech has evolved beyond a narrowly technical forgery problem into a multifaceted security threat with significant societal and ethical implications, encompassing identity fraud, unauthorized access, and social engineering attacks [4,8].

1.1. Related Works

Research into deepfake speech detection has undergone a pronounced methodological evolution. In its early stages, the field was dominated by handcrafted acoustic features combined with supervised machine learning algorithms. Davis and Mermelstein's seminal contribution [9], Mel-Frequency Cepstral Coefficients (MFCC), became a canonical feature set in this period owing to its perceptually motivated frequency representation. Ameer et al. [10] demonstrated that MFCC-based features, when combined with various machine learning classifiers, yield meaningful baseline performance for deepfake audio detection.

As the field matured, the diversification of cepstral representations gained momentum. Sahidullah and Kinnunen [11] conducted a systematic comparison of MFCC, LFCC (Linear-Frequency Cepstral Coefficients), and related spectral features, revealing that different frequency scales exhibit distinct sensitivity profiles across attack types. Todisco et al. [12] introduced Constant-Q Cepstral Coefficients (CQCC) specifically as a spoofing countermeasure for automatic speaker verification and reported substantial EER reductions over MFCC on ASVspoof benchmarks. Sahidullah et al. [13] further examined a broad spectrum of cepstral representations for synthetic speech detection, underscoring the importance of combining complementary feature sets.

The complementary role of long-term spectral features alongside short-term cepstral representations was highlighted by Muckenhirn et al. [14], while Tak et al. [15] provided explainability analysis of CQCC-based countermeasures, identifying which frequency regions carry the most discriminative information for spoofing detection. Lavrentyeva et al. [16] demonstrated the competitive performance of CQCC-based systems in the ASVspoof 2019 challenge across synthesis and VC attack categories.

The ASVspoof challenge series has served as the primary benchmark driving progress in this field. Kinnunen et al. [17] focused on replay attack detection in ASVspoof 2017; Wang et al. [18] introduced ASVspoof 2019 as a large-scale benchmark encompassing TTS, VC, and replay attacks; and Liu et al. [19] extended this framework through ASVspoof 2021 to better reflect real-world deployment conditions. Comprehensive surveys by Almutairi and Elgibreen [4] and Zhang et al. [8] systematically identified open research challenges, both emphasizing the limitations of single-feature approaches and advocating for multi-scale frequency representations.

1.2. Motivation and Contributions

A review of the existing literature reveals that while MFCC, LFCC, and CQCC have each been individually studied, a comprehensive and controlled comparison of all three methods under a unified experimental protocol incorporating statistical significance testing, cross-validation, and fine-grained attack-wise decomposition on the ASVspoof 2021 Logical Access partition remains absent from the literature. The present study does not propose a new feature representation; rather, it contributes a systematic experimental investigation designed to isolate and quantify the effect of frequency scale choice on deepfake speech detection performance. Motivated by this gap, the present study makes the following original contributions:

- MFCC, LFCC, and CQCC are systematically analyzed and compared with respect to their frequency scaling characteristics, spectral resolution properties, and discriminative capacity for deepfake speech detection.
- All three feature sets are experimentally evaluated on the ASVspoof 2021 Logical Access evaluation partition under identical classification conditions, ensuring that observed performance differences are attributable solely to the choice of feature representation.
- The statistical significance of performance differences is assessed using McNemar's test, and result stability is verified through 5-fold stratified cross-validation.
- A fine-grained attack-wise analysis across 13 spoofing systems quantitatively reveals which feature representation is most effective against which category of attack.
- Empirical motivation is provided that multi-feature cepstral strategies, leveraging complementary frequency representations, hold significant potential for more robust deepfake speech detection.

1.3. Paper Organization

The remainder of this paper is organized as follows. Section 2 presents deepfake speech generation techniques and their associated detection challenges. Sections 3–5 examine the MFCC, LFCC, and CQCC feature extraction methods, respectively. Section 6 provides a comparative analysis of the three methods. Section 7 reports experimental results, and Section 8 presents conclusions and future research directions.

2. Deepfake Speech Generation Techniques

The rapid advancement of deep learning-based generative models has substantially expanded the capabilities of synthetic speech production [6,7]. Modern generative architectures can learn the temporal, spectral, and prosodic structure of human speech with remarkable precision, producing synthetic utterances that closely replicate the acoustic characteristics of real speakers. The maturation of deep generative models enabling direct raw waveform synthesis has considerably elevated the perceptual quality of synthetic speech [6], progressively narrowing the acoustic gap between genuine and fabricated utterances. This section examines the two predominant deepfake speech generation paradigms—TTS synthesis and VC—before systematically analyzing the detection challenges they pose.

2.1. Text-to-Speech (TTS) Based Approaches

TTS systems constitute one of the primary technical pathways for deepfake speech generation. These systems model the complex relationship between written language and speech signals, producing natural-sounding audio from textual input [5,7]. Contemporary TTS architectures learn the interactions among linguistic content, phonetic structure, speech rhythm, and prosodic features through deep learning-based acoustic modeling

pipelines. While traditional TTS systems decomposed the generation process into separate components, language model, acoustic model, and vocoder, modern end-to-end architectures integrate these stages within a unified learning framework [6,7], substantially improving the fluency and naturalness of synthesized speech.

The primary reason TTS systems pose a particular threat in the context of deepfake speech generation lies in their ability to disentangle linguistic content from speaker identity. Once the acoustic characteristics of a target speaker have been learned, the system can reproduce that speaker's voice for arbitrary textual input, enabling the generation of content that the target speaker never actually uttered. This capability renders TTS-based deepfakes especially hazardous for voice-based biometric authentication systems. The use of mel spectrogram-based intermediate representations has been identified as a key factor in improving the naturalness and fluency of synthesized utterances [7]. The consistent prosody, balanced spectral structure, and smooth temporal continuity of TTS-generated speech substantially reduce the discriminability of synthetic from genuine utterances for both human perception and automated detection systems.

The TTS-based deepfake generation process can be formalized as a mapping from textual input to a spectral speech representation through a deep acoustic model, followed by waveform reconstruction via a vocoder, as implemented in systems such as Tacotron 2 [7] and WaveNet [6]. The end-to-end learning paradigm preserves micro-level waveform variations and temporal continuity with greater fidelity, rendering TTS-based deepfakes increasingly challenging to detect [4,8].

The general workflows of TTS-based and VC-based deepfake speech generation are illustrated in Figure 1.

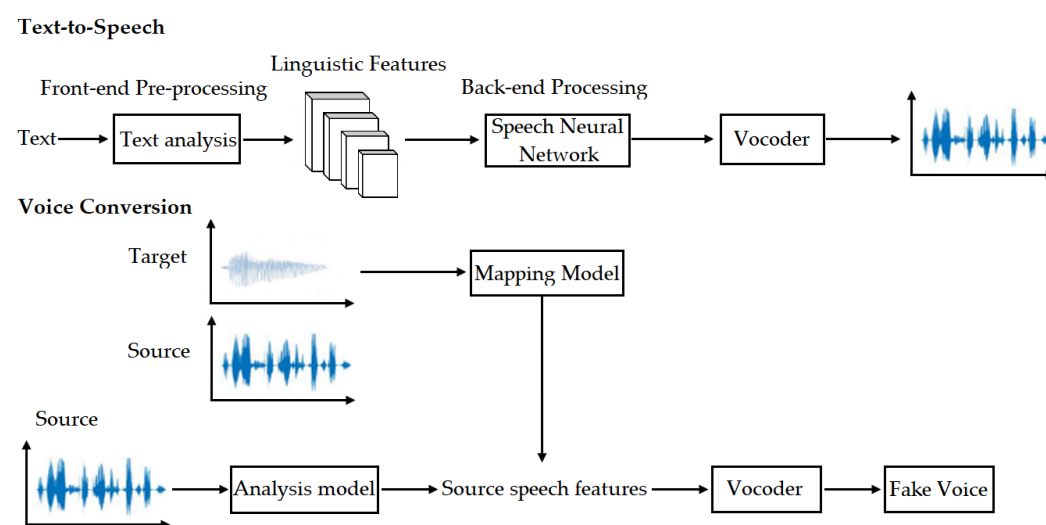


Figure 1. General workflows of TTS-based and VC-based deepfake speech generation.

2.2. Voice Conversion (VC) Techniques

VC techniques represent an alternative deepfake generation pathway that transforms the speaker-specific acoustic characteristics of an existing speech signal to match those of a target speaker, while preserving the linguistic content and temporal structure of the original utterance [18,19]. Unlike TTS systems, VC methods operate directly on source speech signals without requiring textual input, enabling the preservation of natural speech dynamics in the synthesized output.

Contemporary VC methods employ deep neural networks and generative models to learn the acoustic mapping between source and target speakers. The source speech signal is projected into the acoustic space of the target speaker to produce a transformed output. This transformation process maps the source speech signal onto the acoustic space of the

target speaker by conditioning a neural network on a target speaker embedding, thereby producing a converted output that preserves the linguistic content of the original utterance [18,19].

The primary factor making VC-based deepfakes particularly challenging to detect is that the generated speech is directly derived from a real speech signal [18]. The micro-level variations, pauses, and speech rhythm inherent in natural utterances are largely preserved in the converted output, substantially weakening the discriminative cues available to detection systems. For this reason, VC techniques are widely regarded as among the most difficult deepfake generation methods to detect [18,19].

2.3. Detection Challenges

Deepfake speech, whether generated through TTS, VC, or speaker cloning pipelines, produces synthetic or semi-synthetic outputs that replicate the identity characteristics of a target speaker with high fidelity. As state-of-the-art generative models have matured, the perceptual gap between genuine and synthetic speech has narrowed to the point where discriminative cues are no longer constituted by salient artifacts but rather by low-amplitude, context-sensitive, and transmission-condition-dependent subtle signatures. This transformation has shifted deepfake speech detection from a classical closed-set classification problem toward an open-set, domain-shifted security challenge [18,19].

2.3.1. Suppression of Artifacts and Attenuation of Discriminative Traces

Perceptually salient artifacts that were characteristic of early synthetic speech, such as metallic timbre, static prosody, and spectral discontinuities, have been substantially suppressed in modern deepfake generation pipelines. This development forces detection systems to rely on weaker indicators, including micro-level statistical irregularities, subtle deviations in harmonic structure, and phase coherence anomalies. However, such cues are highly sensitive to recording conditions and dataset characteristics, rendering generalization increasingly fragile as within-dataset performance improves [4,8].

2.3.2. Rapid Evolution of Generation Methods and Unseen Attacks

Deepfake speech generation encompasses a broad and heterogeneous family of methods, spanning diverse acoustic models, vocoder architectures, representation learning strategies, and training objectives. This diversity significantly elevates the risk of overfitting detection systems to specific attack types. When generation methods unseen during training are encountered at inference time, decision boundaries may shift unpredictably, resulting in substantial increases in error rates [19]. Consequently, deepfake speech detection must be evaluated not only in terms of classification accuracy but also with respect to generalization capability against previously unseen attacks [4,8].

2.3.3. Robustness Under Channel and Codec Effects

In real-world deployment scenarios, speech signals are rarely transmitted as pristine recordings. Instead, they typically traverse telephone networks, VoIP applications, or social media platforms, undergoing various compression and resampling operations prior to presentation to detection systems. These transmission conditions can mask spoofing-specific cues while simultaneously introducing analogous distortions in genuine speech, narrowing the inter-class decision boundary [18]. Maintaining stable detection performance under high channel and codec variability represents a persistent and critical challenge for practical deployment.

2.3.4. Dataset Bias, Partial Manipulations, and Adversarial Attacks

Detection models tend to rapidly exploit the statistical regularities of their training data. When a dataset is concentrated around specific microphone types, languages, or generation tools, models may learn dataset-specific shortcuts rather than genuine spoofing signatures. Additionally, in certain attack scenarios, only a brief and critical segment of a recording is manipulated, complicating segment-level decision-making. Furthermore, the growing deployment of detection systems has brought adversarial perturbations into scope as adaptive attacks, necessitating that detection models be evaluated not only from an accuracy standpoint but also through the lens of threat modeling and adversarial robustness [8,20].

The challenges discussed in this section provide the operational context for the cepstral feature extraction methods examined in subsequent sections, motivating the investigation of multiple frequency scale representations rather than reliance on any single feature family.

3. Mel-Frequency Cepstral Coefficients (MFCC)

Feature extraction constitutes a fundamental design decision that directly determines system performance in deepfake speech detection. Among the most widely employed methods for this purpose, MFCC represent the speech spectrum in a compact and discriminative manner by modeling the perceptual characteristics of the human auditory system [9,21]. Formally introduced by Davis and Mermelstein [9], MFCC has served as a canonical feature set in automatic speech recognition and speaker verification for decades and continues to function as the indispensable baseline representation in spoofed speech detection literature [10,22].

The core objective of MFCC is to capture the short-term spectral envelope of speech reflecting vocal tract shape while suppressing unnecessary waveform-level detail [9]. To this end, the method maps the speech signal onto a perceptually motivated frequency scale to produce a compact representation. In this study, MFCC is employed as the primary baseline feature for deepfake speech detection, computed using 20 coefficients with a 1024-point FFT; the mean and standard deviation of frame-level coefficients along the time axis are concatenated to yield a fixed 40-dimensional feature vector. MFCC-based features have been reported to provide meaningful baseline performance for deepfake audio detection when combined with classical machine learning classifiers [10].

3.1. MFCC Feature Extraction

The MFCC extraction pipeline transforms raw speech signals into structured and meaningful feature vectors through a sequence of successive transformations. The general workflow is illustrated in Figure 2.

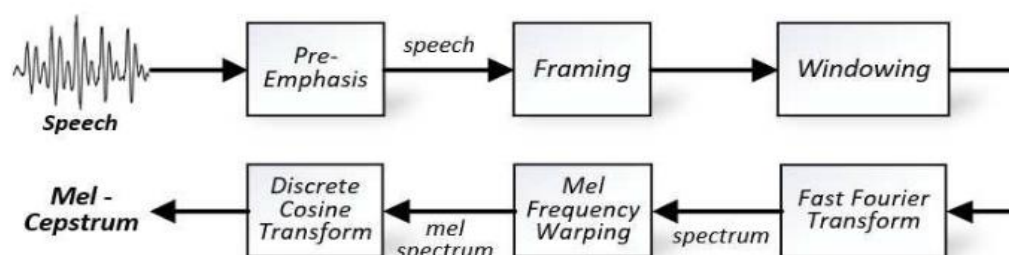


Figure 2. General workflow of the MFCC feature extraction pipeline.

3.1.1. Signal Acquisition and Pre-Emphasis

A pre-emphasis filter is applied to the raw input signal to amplify high-frequency components and equalize the spectral energy distribution prior to further processing [23]. The effect of pre-emphasis filtering on a 25 ms signal frame is illustrated in Figure 3.

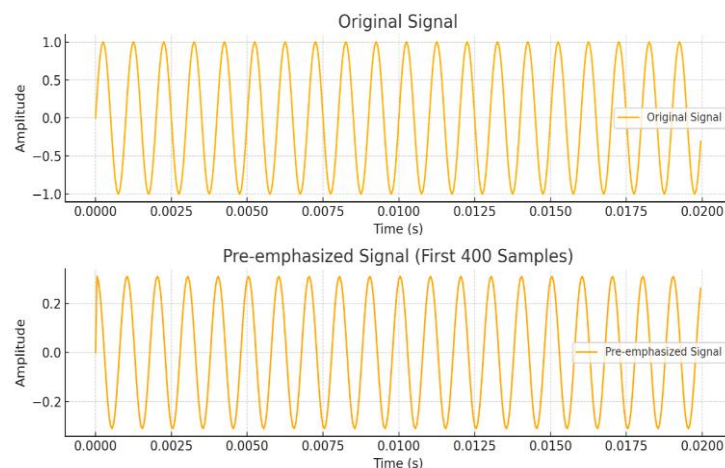


Figure 3. Effect of pre-emphasis filtering on a 25 ms signal frame in the time domain.

3.1.2. Framing

Since speech signals can be considered approximately stationary within short temporal windows [21], the signal is segmented into overlapping frames of typically 20–40 ms duration with 30–75% overlap between consecutive frames [21,23].

3.1.3. Windowing

A windowing function is applied to each frame to reduce spectral leakage at frame boundaries [23–27]. In this study, the Hamming window is employed, as defined in Equation (1):

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right) \quad (1)$$

This function is applied element-wise to each sample within the frame, smoothly attenuating amplitudes at the frame boundaries while preserving the central signal structure. The effect of Hamming windowing on a 25 ms signal frame is illustrated in Figure 4.

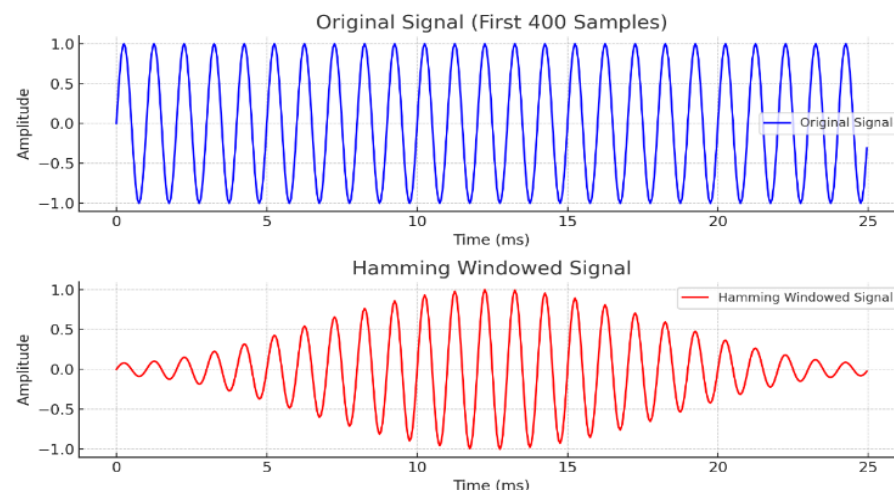


Figure 4. Effect of Hamming windowing on a 25 ms signal frame in the time domain.

3.1.4. Fast Fourier Transform (FFT)

Each windowed frame is transformed to the frequency domain using a 1024-point Fast Fourier Transform (FFT) [24,25], and the magnitude spectrum is computed as expressed in Equation (2):

$$|X(k)| = \sqrt{[Re(X(k))]^2 + [Im(X(k))]^2} \quad (2)$$

This magnitude spectrum serves as input to the subsequent Mel filterbank stage.

3.1.5. Mel-Scale Filterbank

The human auditory system perceives frequencies on a non-linear scale, exhibiting higher resolution at low frequencies and progressively coarser resolution at higher frequencies [9,21]. To model this psychoacoustic property, the FFT magnitude spectrum is passed through a bank of triangular filters distributed along the Mel frequency scale. In this study, a filterbank consisting of 20 Mel-scaled filters is employed. The Mel-scale transformation is expressed in Equation (3):

$$Mel = 2595 \times \log_{10} \left(1 + \frac{f}{700} \right) \quad (3)$$

This transformation maps the linear frequency axis onto a perceptually motivated non-linear representation. The Mel filterbank provides finer resolution at low frequencies and coarser resolution at high frequencies, effectively emphasizing the perceptually salient components of speech [9].

3.1.6. Discrete Cosine Transform (DCT)

Following log-compression of the Mel filterbank outputs, the Discrete Cosine Transform (DCT) is applied to yield cepstral coefficients [9,25]. The DCT decorrelates the filterbank outputs and compacts the spectral information into a lower-dimensional representation. In this study, 20 cepstral coefficients are retained. The MFCCs are computed as expressed in Equation (4):

$$c_n = \sum_{m=1}^M \log(E_m) \cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right] \quad (4)$$

where E_m denotes the energy in the m -th Mel filter band, M is the total number of filters, and n is the cepstral coefficient index. Lower-order coefficients capture the broad spectral envelope of the signal, while higher-order coefficients are more susceptible to noise and are therefore typically excluded from analysis [9]. In this study, the mean and standard deviation of the 20 frame-level coefficients along the time axis are concatenated, yielding a fixed 40-dimensional feature vector.

3.2. Advantages, Limitations, and Role in Deepfake Speech Detection

3.2.1. Advantages

The primary advantage of MFCC lies in its perceptually motivated frequency resolution, achieved through the Mel scale [9]. Since the most discriminative components of speech are predominantly concentrated in the low-to-mid frequency range, this scaling yields a practically beneficial trade-off. The compact representation, typically 12–20 coefficients, efficiently encodes the essential structure of the speech signal while reducing both storage and computational requirements. MFCC is readily compatible with both classical classifiers such as GMM and SVM, and deep learning architectures including CNN and RNN [22]. Moreover, temporal dynamics can be incorporated by appending first- and second-order delta coefficients.

3.2.2. Limitations

The compression of high-frequency content by the Mel scale renders fine-grained spectral structures in this region less visible [9,12]. In certain deepfake speech signals, however, synthetic artifacts manifest precisely at high frequencies or as low-energy broadband irregularities. Furthermore, MFCC operates under the assumption of approximate stationarity within each frame; rapidly varying signals, including manipulated speech, may violate this assumption, causing distortions to be masked through averaging [11,21]. Environmental factors such as microphone type, room reverberation, and codec compression alter the spectral envelope and may adversely affect the stability of MFCC representations [18]. Most critically, since modern synthetic speech generation replicates the fundamental spectral structure of natural speech with high fidelity, summary representations such as MFCC may fail to capture discriminative spoofing traces [4,8].

3.2.3. Role in Deepfake Speech Detection

MFCC has long served as the default feature set for speech recognition and speaker verification [21,22]. In the spoofed speech detection literature, it continues to function as the primary baseline against which novel approaches are benchmarked. In deep learning-based systems, MFCC representations are frequently converted to spectrogram-like input formats for use in CNN architectures [13,14]. In this study, MFCC serves as the primary baseline feature representation. The achieved accuracy of 82.33% and EER of 17.09% provide a quantitative reference point for comparison with LFCC and CQCC. While MFCC remains an indispensable reference in the literature, its limitations become increasingly apparent as deepfake generation techniques diversify [4,8,19].

These limitations motivate the investigation of LFCC in the following section, which aims to address the high-frequency suppression tendency of the Mel scale through linear frequency resolution.

4. Linear-Frequency Cepstral Coefficients (LFCC)

The high-frequency suppression tendency of MFCC, attributable to Mel-scale compression, risks overlooking synthetic artifacts that may manifest in higher frequency regions of deepfake speech signals. As an alternative aimed at addressing this limitation, LFCC partition the frequency axis into uniform intervals, representing the entire speech spectrum with equal resolution across all bands [11,12]. While sharing the same general processing pipeline as MFCC, LFCC is distinguished by the adoption of a linear filterbank in place of the perceptually motivated Mel-scale design, enabling balanced spectral representation across the full frequency range [11,21].

While MFCC prioritizes the low-frequency region through Mel-scale weighting, yielding strong performance in speech recognition tasks, LFCC preserves spectral detail in the high-frequency bands, providing a potential advantage for detecting synthetic artifacts introduced during deepfake speech generation [11,12]. This property may enhance the discriminative power of LFCC in scenarios involving replay attacks, codec-induced distortions, and certain synthetic speech production techniques [18]. In this study, LFCC is computed under the same parametric configuration as MFCC—20 filters, 1024-point FFT, and a 40-dimensional feature vector formed by concatenating frame-level mean and standard deviation, ensuring that any observed performance differences are attributable solely to the frequency scale choice.

4.1. LFCC Feature Extraction

The LFCC extraction pipeline follows the classical cepstral approach of analyzing speech under the short-term stationarity assumption, but adopts a linear frequency

scaling at the filterbank stage to achieve balanced spectral representation across all bands [21,22]. The general workflow is illustrated in Figure 5.

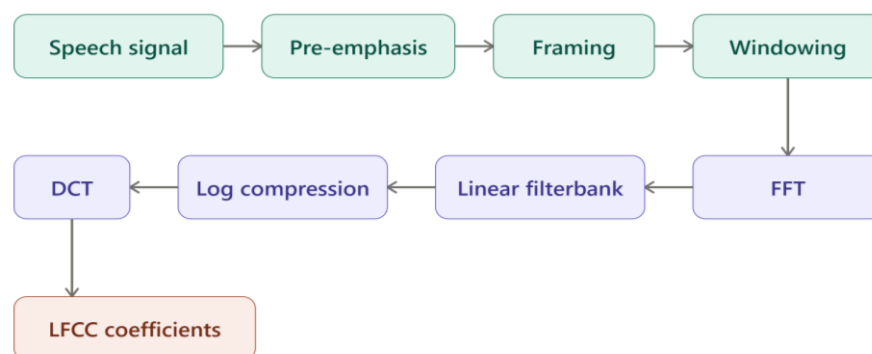


Figure 5. General workflow of the LFCC feature extraction pipeline.

The initial stages, framing and windowing, are identical in principle to MFCC, aiming to mitigate the adverse spectral effects of frame boundary discontinuities. Following windowing, the FFT is computed for each frame to obtain the frequency-domain representation of the signal.

The fundamental distinction of LFCC emerges at this point. Rather than a Mel-scaled filterbank, a linearly spaced filterbank is applied to the FFT output. This filterbank partitions the frequency axis into uniformly spaced bands and computes the energy within each band. In this study, 20 linear filters are employed. The uniform band placement preserves spectral detail in the high-frequency region without attenuation, contributing to a more balanced representation of the physical characteristics of the speech signal [11]. The filterbank outputs are subsequently subjected to logarithmic compression to normalize amplitude dynamics.

In the final stage, the DCT is applied to the log-compressed filterbank outputs to yield cepstral coefficients. The LFCCs are computed as expressed in Equation (5):

$$c_k = \sum_{m=1}^M \log(E_m) \cdot \cos\left[\frac{\pi k(m - 0.5)}{M}\right] \quad (5)$$

where E_m denotes the energy in the m -th linear filter band, M is the total number of filter bands, k is the cepstral coefficient index, and the summation runs from $m = 1$ to M . Consistent with the MFCC implementation in this study, 20 cepstral coefficients are extracted per frame; the mean and standard deviation along the time axis are concatenated to form a fixed 40-dimensional feature vector.

4.2. Application Context

LFCC has been investigated as a cepstral-based feature set across various speech classification tasks, particularly in scenarios where preserving spectral detail beyond perceptual modeling is of primary concern [11,12]. Its potential to retain high-frequency spectral information has been explored in contexts involving replay attacks, channel distortions, and recording environment variability [11,18]. Dewi et al. [3] demonstrated that linear frequency resolution can yield meaningful results in certain audio classification tasks such as infant cry analysis, illustrating the cross-domain applicability of LFCC. In anti-spoofing research, LFCC has frequently been examined as a complementary representation alongside MFCC and CQCC [11,19].

In this study, LFCC achieves 81.26% accuracy and 18.53% EER on the ASVspoof 2021 LA evaluation partition, the lowest general performance among the three methods. However, a notable finding from the attack-wise analysis is that LFCC achieves the highest accuracy on the A09 attack system (83.67%), outperforming both MFCC and

CQCC on this specific category. This result suggests that linear frequency resolution may retain complementary discriminative capacity for certain attack types, supporting its consideration within multi-feature detection frameworks.

4.3. Advantages, Limitations, and Position in the Literature

4.3.1. Advantages

The most distinguishing advantage of LFCC is its uniform frequency resolution across the entire spectrum, which preserves spectral detail in the high-frequency region without the attenuation characteristic of perceptually scaled methods [11,21]. Contemporary synthetic speech generation techniques typically succeed in replicating low-to-mid frequency components with high fidelity, while inadvertently leaving subtle statistical irregularities and spectral anomalies in specific frequency bands. By preserving such regions without compression, LFCC may make these synthetic traces more visible to the detection system, offering a richer feature set for discriminative modeling [11,19]. Its cepstral-based formulation also ensures straightforward integration with both classical and machine learning-based classifiers.

4.3.2. Limitations

The linear frequency representation of LFCC introduces certain limitations in specific application scenarios. Its departure from perceptual modeling results in reduced advantage relative to perceptually scaled methods in tasks where auditory similarity is the primary criterion, such as speech recognition and speaker verification [21,22]. Under challenging real-world conditions—including environmental noise, channel distortions, and codec compression—the high-frequency detail preserved by LFCC does not always translate into meaningful discriminative information [18]. In low-quality recordings or short-duration utterances, high-frequency band representations may become unstable, limiting the effectiveness of LFCC. Consequently, LFCC is most commonly positioned in the literature as a complementary feature rather than a standalone solution [11,12].

4.3.3. Position in the Literature

Although LFCC has not achieved the broad adoption of MFCC, it occupies an important position as a comparative reference in anti-spoofing research [11]. Sahidullah and Kinnunen [11] established a foundational comparison demonstrating that LFCC and MFCC exhibit distinct sensitivity profiles across different attack types, supporting the case for their joint use. Sahidullah et al. [13] further confirmed that LFCC offers competitive performance on specific spoofing categories in their comprehensive feature comparison study. Dewi et al. [3] reported meaningful results for LFCC in audio classification beyond the speech domain, attesting to its broader applicability. In multi-feature deepfake detection frameworks, LFCC is consistently positioned as a complementary representation that contributes incremental discriminative information not captured by MFCC or CQCC alone [11,12,19].

The following section examines CQCC, which aims to transcend both the perceptual scaling limitation of MFCC and the fixed bandwidth constraint of LFCC through a logarithmic, variable-resolution frequency representation.

5. Constant-Q Cepstral Coefficients (CQCC)

The feature extraction methods examined in the preceding sections, MFCC and LFCC, differ in their frequency axis scaling but share a common structural constraint: both employ a fixed bandwidth across all frequency bands. The harmonic structure of speech signals, however, inherently demands higher spectral resolution at low frequencies and

benefits from improved temporal resolution at higher frequencies. To address this requirement, Todisco et al. [12] proposed CQCC specifically as a spoofing countermeasure for automatic speaker verification. CQCC is grounded in the Constant-Q Transform (CQT), which adjusts frequency resolution logarithmically as a function of center frequency. This design enables CQCC to transcend both the high-frequency suppression tendency of MFCC and the fixed-bandwidth constraint of LFCC.

CQCC is derived by first applying the CQT to obtain energy distributions over logarithmically spaced frequency bands, followed by log-compression and cepstral transformation via the DCT [12]. In this study, CQCC is computed using 84 CQT bins with 12 bins per octave and 20 cepstral coefficients; the same statistical summarization strategy employed for MFCC and LFCC concatenation of frame-level mean and standard deviation is adopted, yielding a consistent 40-dimensional feature vector. CQCC-based features have been shown to provide high discriminability across diverse spoofing scenarios when combined with appropriate classification backends [12,16,18].

5.1. Constant-Q Transform (CQT)

The Constant-Q Transform (CQT) constitutes the analytical core of the CQCC feature extraction framework. It analyzes speech signals in the time-frequency plane using a logarithmically spaced frequency scale [12,28]. In contrast to the Short-Time Fourier Transform (STFT), which employs uniform frequency band spacing, the CQT distributes frequency bands logarithmically, providing higher spectral resolution at low frequencies and improved temporal resolution at high frequencies, a characteristic that closely mirrors the perceptual properties of the human auditory system [12,28].

The Q factor defined for each frequency band in the CQT represents the ratio of center frequency to bandwidth, as expressed in Equation (6):

$$Q = \frac{f_k}{\delta f_k} \quad (6)$$

where f_k denotes the center frequency of the k -th band and δf_k denotes its bandwidth. Maintaining the Q factor constant across all frequency bands imposes a logarithmic structure on frequency resolution, constituting the fundamental property that distinguishes the CQT from fixed-bandwidth transforms [28]. In this study, the Q factor is determined by the standard librosa parametrization with 84 bins at 12 bins per octave.

The CQT is mathematically expressed in Equation (7):

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} x[n] \cdot W^*[n, k] \cdot e^{-j2\pi Q \frac{n}{N[k]}} \quad (7)$$

where $x[n]$ denotes the input signal, $W^*[n, k]$ is the conjugate of the analysis window associated with the k -th frequency band, $N[k]$ is the frequency-dependent window length, and Q is the constant Q factor. The frequency-dependent variation in window length enables the CQT to provide finer spectral resolution at low frequencies, facilitating more precise representation of harmonic structures in speech signals. The absence of an analytically defined inverse CQT renders this transform particularly suited for feature extraction rather than signal reconstruction.

5.2. Logarithmic Scaling

The spectral magnitudes produced by the CQT may span a wide dynamic range on a linear scale, potentially complicating the learning process for statistical and deep learning-based classifiers. Logarithmic scaling is therefore applied to the spectral magnitude values to obtain a more numerically stable and learning-friendly representation. The logarithmic spectrum is defined in Equation (8):

$$\log|X[k]|^2 = 2\log|X[k]| \quad (8)$$

where $X[k]$ denotes the CQT spectral coefficients and the logarithm is applied to the power spectrum magnitude. This operation is equivalent to scaling the log magnitude by a factor of two. Log-compression reduces the dynamic range of spectral magnitudes, yielding a more numerically stable representation that better approximates the amplitude perception characteristics of the human auditory system [9,21]. This operation is applied prior to the cepstral transformation stage, facilitating the extraction of more discriminative and compact cepstral coefficients by the subsequent DCT [12].

5.3. Cepstral Transformation (DCT)

To derive compact, time-stationary features from the log-compressed spectrum, the Discrete Cosine Transform (DCT) is applied [9]. The DCT decorrelates the log-spectral components and concentrates the information into lower-order coefficients [9,25], enabling a more effective representation of the spectral envelope and making the resulting coefficients better suited for classification algorithms. The cepstral transformation for CQCC is expressed in Equation (9):

$$c_m = \sum_{k=1}^K \log|X[k]|^2 \cdot \cos\left[\frac{\pi m(k-0.5)}{K}\right] \quad (9)$$

where $X[k]$ denotes the CQT spectral coefficients, K is the total number of frequency bands, m is the cepstral coefficient index, and the summation runs from $k = 1$ to K . Lower-order cepstral coefficients represent the broad spectral envelope of the signal, while higher-order coefficients reflect finer spectral detail. In this study, 20 cepstral coefficients are retained; the mean and standard deviation along the time axis are concatenated to yield a 40-dimensional feature vector consistent with the MFCC and LFCC implementations.

5.4. Technical Characteristics and Application Domains

CQCC features have been widely adopted in automatic speaker verification and spoofed speech detection systems. Todisco et al. [12] reported substantial EER reductions over MFCC when CQCC was combined with a GMM backend on the ASVspoof 2015 evaluation set. Lavrentyeva et al. [16] demonstrated competitive CQCC-based system performance across synthesis and VC attack categories in the ASVspoof 2019 challenge. Tak et al. [15] provided explainability analysis of CQCC-based countermeasures, identifying the 4–8 kHz frequency region as particularly informative for spoofing detection.

Beyond security-oriented applications, CQCC has been applied to pathological speech analysis, the auditory examination of neurological disorders, and emotion recognition [29]. In the assessment of motor speech disorders, CQCC features have been reported to effectively capture subtle spectral changes in the signal. Mittal and Dua [30] demonstrated that combining CQCC features with LSTM architectures yields strong results in speaker verification applications, attesting to the compatibility of CQCC with deep learning-based classification frameworks.

5.5. Advantages, Limitations, and General Assessment

5.5.1. Advantages

The primary advantage of CQCC lies in its logarithmic frequency resolution, which enables detailed analysis of speech signals across a wide spectral range [12]. The constant-Q principle adjusts frequency resolution as a function of center frequency, providing high spectral resolution at low frequencies and improved temporal resolution at high frequencies, closely mirroring the sensitivity characteristics of the human auditory system. This design enables fine-grained spectral differences critical for detecting

synthetic content to be rendered more visible [12,18]. The direct compatibility of CQCC with deep learning architectures such as LSTM and CNN facilitates integration into modern end-to-end detection frameworks [14,30]. The capacity of constant-Q representations to capture both narrow- and broad-band spectral variations has been validated across diverse application domains [29].

5.5.2. Limitations

The logarithmic frequency representation of CQCC introduces certain practical limitations. The computational cost of CQT is comparatively higher than FFT-based methods, which may constitute a constraint in real-time processing applications. Furthermore, CQCC performance varies with the dataset characteristics and attack type distribution. The attack-wise analysis conducted in this study reveals that CQCC underperforms MFCC on certain traditional TTS-based attack systems, specifically A07 and A08. This finding indicates that CQCC does not constitute a universally superior solution across all attack categories, and that its generalization capacity against unseen attacks is sensitive to the diversity of the training data [18,19].

5.5.3. General Assessment

CQCC integrates the constant-Q transformation mechanism with cepstral representation techniques, yielding a powerful analytical tool for both theoretical and practical speech signal analysis. In this study, CQCC attains the highest numerical performance across all three evaluation metrics on the ASVspoof 2021 LA partition: 83.59% accuracy, 89.15% ROC-AUC, and 15.83% EER; however, the difference relative to MFCC does not reach statistical significance ($p = 0.202$). These findings are corroborated by 5-fold cross-validation results ($87.89\% \pm 0.81\%$) and are consistent with the earlier findings of Todisco et al. [12]. The attack-wise analysis indicates that CQCC achieves the highest accuracy on 6 of 13 spoofing systems, with a particularly notable result of 95.83% on the A16 attack. These results collectively suggest CQCC as a competitive feature representation for deepfake speech detection, while acknowledging that its advantage over MFCC is not statistically confirmed under the current evaluation conditions.

The following section provides a comparative analysis of MFCC, LFCC, and CQCC in terms of frequency scaling characteristics, spectral resolution properties, and discriminative power across attack categories.

6. Comparative Analysis of Cepstral Representations

The preceding sections examined MFCC, LFCC, and CQCC individually. This section presents a systematic comparative analysis of these three methods in the context of deepfake speech detection, integrating both qualitative characterization and quantitative evidence drawn from the literature. The discriminative capacity of each cepstral representation, its frequency scaling rationale, and its potential to capture synthetic speech artifacts are evaluated alongside their respective strengths and limitations.

6.1. Comparison of Frequency Scales and Representation Approaches

The fundamental distinction among the three methods lies in how the frequency axis is scaled. MFCC employs the perceptually motivated Mel frequency scale to emphasize the perceptually salient components of speech [9,21]. The Mel scale provides higher resolution at low frequencies while progressively widening band spacing at higher frequencies. This design has achieved high success in perception-centric tasks such as speech recognition and speaker verification, but may suppress a portion of the synthetic artifacts introduced during deepfake generation due to high-frequency compression

[11,12]. Ameer et al. [10] confirmed that MFCC achieves competitive detection accuracy on benchmark datasets when paired with conventional classifiers.

LFCC employs a linear frequency scale, representing all frequency bands of the speech spectrum with uniform resolution [11,21]. This approach preserves spectral detail in the high-frequency region and facilitates the analysis of statistically irregular signatures specific to synthetic generation. CQCC, grounded in the constant-Q transform, adapts frequency resolution logarithmically—providing high spectral resolution at low frequencies and broader bandwidth at higher frequencies [12,28]. These distinct scaling approaches result in each cepstral representation extracting complementary types of information relevant to deepfake speech detection.

The discriminative value of long-term spectral features alongside short-term cepstral representations has been emphasized in the literature [14], suggesting that cepstral representations should be evaluated not only with respect to frequency scaling but also in terms of temporal resolution and window length parameters.

6.2. Literature-Supported Quantitative Comparison

Experimental findings reported in the literature confirm that the three cepstral representations exhibit distinct performance profiles across different spoofing scenarios. Todisco et al. [12] reported substantial EER reductions for CQCC over MFCC using a GMM backend on ASVspoof 2015. Sahidullah and Kinnunen [11] systematically demonstrated that MFCC and LFCC exhibit distinct sensitivity profiles across attack types, supporting the case for their joint use. Sahidullah et al. [13] confirmed that LFCC offers competitive performance on specific spoofing categories. Lavrentyeva et al. [16] demonstrated competitive CQCC-based performance across TTS and VC attack categories in ASVspoof 2019.

Representative EER values drawn from the literature are summarized in Table 1.

Table 1. Representative EER values reported in the literature (ASVspoof benchmarks, GMM backend).

Method	ASVspoof 2015 EER (%)	ASVspoof 2019 LA EER (%)	Source
MFCC + GMM	7.42	14.33	[12,13]
LFCC + GMM	5.91	11.96	[13,18]
CQCC + GMM	3.69	9.57	[12,18]

EER values are drawn from results reported under comparable experimental conditions in the cited studies. ASVspoof 2015 values are sourced from Todisco et al. [12] and Sahidullah et al. [13]; ASVspoof 2019 LA values are sourced from Wang et al. [18] and Sahidullah et al. [13]. Exact values may vary depending on evaluation subset, GMM configuration, and protocol version.

These results demonstrate that CQCC-based representations provide stronger discriminability against TTS and VC attacks, while LFCC consistently reduces error rates relative to MFCC. It should be noted, however, that these values are sensitive to the classification backend, dataset, and evaluation protocol [18,19].

6.3. Discriminative Power Across Attack Categories

Since different deepfake speech generation methods leave distinct acoustic traces, the discriminative power of each cepstral representation varies with attack type. In TTS-based attacks, the synthesized speech typically exhibits consistent spectral structure and smooth prosody; the logarithmic frequency resolution of CQCC renders such regularities more detectable by amplifying subtle spectral deviations [12]. Tak et al.'s explainability analysis

[15] revealed that CQCC is particularly informative in the 4–8 kHz frequency region relative to MFCC.

In VC-based attacks, the substantial preservation of natural speech characteristics in the converted signal poses greater challenges for detection systems [18]. In this scenario, the linear frequency resolution of LFCC may offer complementary detection capacity by capturing fine spectral distortions introduced during formant transformation [11]. In replay attacks, the channel signature and room reverberation introduced during recording and playback affect all cepstral representations, but the variable time-frequency resolution of CQCC enables more stable capture of such distortions [17,18].

6.4. Robustness and Generalization

Real-world deployment conditions, including environmental noise, channel effects, and codec compression, can substantially degrade the performance of cepstral representations [17,18]. MFCC may exhibit relative robustness to certain noise types due to its perceptual scaling, but risks suppressing synthetic artifacts. LFCC and CQCC, by preserving broader spectral information, may offer higher generalization potential under certain conditions [11,12]. However, the detailed spectral representations provided by LFCC and CQCC may become unstable in low-quality recordings or short-duration utterances, indicating that robustness depends not only on frequency scaling but also on dataset characteristics and experimental conditions. The literature consistently interprets this as evidence that cepstral representations should be evaluated as complementary approaches rather than standalone solutions [11,12,19]. Speech signal representation performance depends not only on feature type but on the holistic design of the system [31].

6.5. Comprehensive Summary Table

The comparative analysis presented in this section demonstrates that MFCC, LFCC, and CQCC offer distinct and mutually complementary perspectives on the deepfake speech detection problem. The comparative summary is presented in Table 2.

Table 2. Comparative summary of MFCC, LFCC, and CQCC for deepfake speech detection.

Property	MFCC	LFCC	CQCC
Frequency scale	Mel (perceptual)	Linear	Logarithmic, constant-Q
Core motivation	Model human auditory perception	Represent spectral detail with uniform resolution	Variable-resolution frequency analysis
High-frequency representation	Compressed/limited	Detailed and preserved	Variable-resolution
Perceptual alignment	High	Low	Moderate
Synthetic artifact visibility	May be suppressed in some scenarios	More visible under certain conditions	Prominent in spoofing and replay attacks
TTS attack performance	Moderate	Moderate-low	High
VC attack performance	Moderate	Moderate	High
Replay attack performance	Moderate	Moderate	High
Sensitivity to noise and channel effects	Moderate	Condition-dependent	Relatively robust

Prevalence in literature	Very common	Limited	Common in anti-spoofing
Usage role	Baseline/reference feature	Complementary feature	Security-oriented feature
Role in deepfake detection	Perceptually grounded discrimination	Structural and spectral analysis	Spoofing-focused robust representation

This comparative theoretical and quantitative analysis motivates the experimental evaluation presented in the following section.

7. Experimental Results

7.1. Dataset and Experimental Setup

The experimental evaluation in this study is conducted on the ASVspoof 2021 Logical Access (LA) evaluation partition [19], publicly available at <https://www.asvspoof.org>. ASVspoof 2021 is a widely adopted benchmark dataset for spoofing countermeasure research against automatic speaker verification systems. The evaluation partition encompasses 13 distinct spoofing systems (A07–A19) based on TTS synthesis and VC techniques, alongside genuine (bonafide) utterances. To address the pronounced class imbalance in the full evaluation set, a balanced sampling strategy is applied: equal numbers of bonafide and spoof samples are drawn, forming a balanced subset of 10,000 audio recordings, of which 5571 are successfully processed and used in the experimental analysis. The class distribution is summarized in Table 3.

Table 3. Class distribution of the experimental dataset.

Class	Sample Count	Proportion
Bonafide (Genuine)	2786	50.0%
Spoof (Synthetic)	2785	50.0%
Total	5571	100.0%

All audio recordings are processed in .flac format. Feature extraction is implemented in Python using the librosa and scipy libraries (Python 3.x, librosa 0.11.0, scikit-learn 1.8.0, numpy 2.3.5). Of the 10,000 targeted recordings, 5571 were successfully processed; the remaining 4429 files could not be loaded due to technical incompatibilities between the librosa/soundfile backend and certain FLAC encodings in the evaluation partition. These failures were not systematic with respect to attack type, speaker identity, or recording condition, and the 50/50 class balance was preserved in the final subset. A stratified train-test split of 80–20% was applied within the balanced subset. It should be noted that this custom split replaces the official ASVspoof evaluation protocol; consequently, overlap between attack systems or speakers across training and test partitions cannot be fully excluded, which should be considered when interpreting the generalization of the reported results.

7.2. Feature Extraction Parameters

The sole sources of variation across methods are the frequency scale and filterbank design. The parameters were not tuned independently for each method; instead, standard values widely adopted in the literature and verified to yield reasonable performance across all three representations were employed. No pre-emphasis filtering was applied, and all recordings were processed at their original sampling rate of 16,000 Hz without resampling. This deliberate design choice guarantees that observed performance differences are attributable solely to frequency scale selection rather than parameter

optimization, since independent tuning would confound the effect of feature representation with that of configuration choices, thereby undermining the primary research objective. The complete extraction parameters for all three methods are summarized in Table 4.

Table 4. Feature extraction parameters of the three methods.

Parameter	MFCC	LFCC	CQCC
Number of coefficients	20	20	20
FFT size (n_fft)	1024	1024	—
Hop length	512 samples	n_fft//4	—
Window type	Hann	Hanning	—
Number of filters	20 (Mel)	20 (Linear)	—
CQT bins	—	—	84
Bins per octave	—	—	12
Fmin (CQT)	—	—	32.70 Hz (C1)
Pre-emphasis	Not applied	Not applied	Not applied
Numerical stabilization (ϵ)	1×10^{-10}	1×10^{-10}	1×10^{-10}
Feature dimensionality	40 (mean+std)	40 (mean+std)	40 (mean+std)
Sample rate	16,000 Hz	16,000 Hz	16,000 Hz
Resampling	None	None	None
SVM kernel	Linear	Linear	Linear
SVM C	1.0	1.0	1.0
Calibration	Platt scaling	Platt scaling	Platt scaling
Class weighting	None	None	None
Random state	42	42	42
librosa	0.11.0	0.11.0	0.11.0
scikit-learn	1.8.0	1.8.0	1.8.0
numpy	2.3.5	2.3.5	2.3.5

Each feature vector is formed by concatenating the temporal mean and standard deviation of frame-level cepstral coefficients, yielding a fixed 40-dimensional representation for all three methods.

7.3. Classifier and Evaluation Metrics

A linear-kernel Support Vector Machine (SVM) is employed for the comparative classification experiments. Fixing the same classifier across all experiments ensures that observed performance differences are exclusively attributable to the feature representation rather than classifier choice. Prior to classification, feature normalization is applied using StandardScaler; probability estimates are obtained by calibrating the SVM with Platt scaling.

The following evaluation metrics are employed:

Accuracy: The proportion of correctly classified samples relative to the total sample count. Serves as a primary performance indicator for balanced-class datasets.

ROC-AUC: The area under the Receiver Operating Characteristic curve, measuring the classifier's overall discrimination capability across decision thresholds. A value of 1.0 represents perfect discrimination; 0.5 corresponds to random classification.

Equal Error Rate (EER): The error rate at the operating point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). EER is the standard evaluation metric in anti-spoofing literature, where lower values indicate superior detection performance [18,19].

5-fold Cross-Validation: Stratified 5-fold cross-validation is applied to assess result stability and generalization capability. Mean ROC-AUC and standard deviation are reported.

McNemar's Test: McNemar's test is applied to assess the statistical significance of pairwise performance differences ($\alpha = 0.05$).

7.4. Results

7.4.1. Overall Performance Comparison

The performance of the three feature sets under the linear SVM classifier is presented in Table 5.

Table 5. Overall performance comparison of MFCC, LFCC, and CQCC.

Feature	Accuracy (%)	95% CI (Acc)	ROC-AUC (%)	EER (%)	95% CI (EER)
MFCC	82.33	[79.96, 84.50]	87.42	17.09	[15.00, 19.37]
LFCC	81.26	[78.85, 83.48]	87.10	18.53	[16.35, 20.90]
CQCC	83.59	[81.29, 85.69]	89.15	15.83	[13.84, 18.01]

CQCC attains the highest numerical performance across all three metrics, with a numerical difference of 1.73 percentage points in ROC-AUC relative to MFCC and 2.05 percentage points relative to LFCC. In terms of EER, CQCC attains the lowest error rate at 15.83%. It should be noted, however, that these differences should be interpreted in light of the McNemar test results presented in Section 7.4.3. The ROC curves of the three methods are presented comparatively in Figure 6.

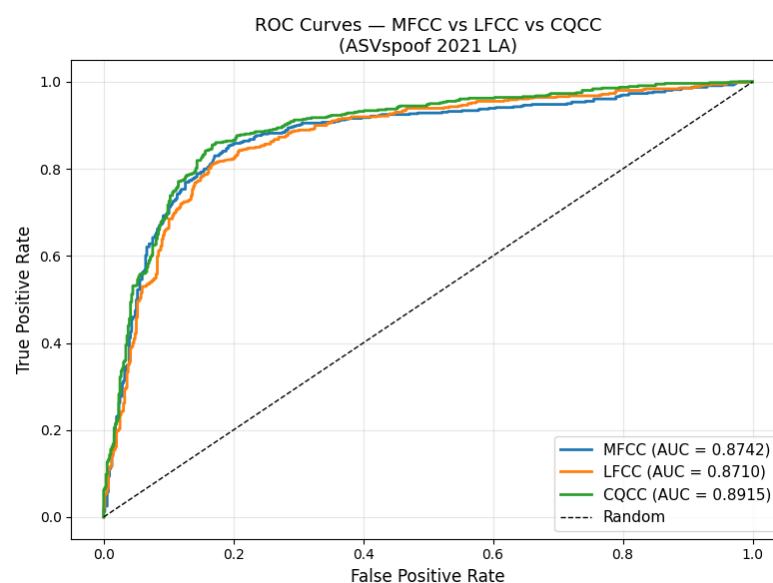


Figure 6. ROC curves for MFCC (AUC = 0.8742), LFCC (AUC=0.8710), and CQCC (AUC = 0.8915) on the ASVspoof 2021 LA evaluation partition.

7.4.2. Cross-Validation Results

The results of 5-fold stratified cross-validation, applied to verify result stability, are presented in Table 6.

Table 6. 5-Fold cross-validation ROC-AUC results.

Feature	Mean ROC-AUC (%)	Std (%)
MFCC	86.86	±0.85
LFCC	85.90	±0.92
CQCC	87.89	±0.81

Cross-validation results are consistent with the single-split findings, confirming that CQCC achieves both higher mean ROC-AUC and lower standard deviation than its counterparts. The lowest standard deviation observed for CQCC (0.81%) indicates that this representation produces more stable performance across different data partitions. The highest standard deviation observed for LFCC (0.92%) reflects its relatively greater sensitivity to the data distribution.

7.4.3. Statistical Significance Analysis

The statistical significance of pairwise performance differences is assessed using McNemar’s test. Results are presented in Table 7.

Table 7. McNemar’s test results ($\alpha = 0.05$).

Comparison	χ^2	<i>p</i> -Value	Significance
MFCC vs CQCC	1.6250	0.2024	Not significant
LFCC vs CQCC	4.4014	0.0359	Significant
MFCC vs LFCC	1.0254	0.3112	Not significant

McNemar’s test reveals that the performance difference between LFCC and CQCC is statistically significant ($\chi^2 = 4.40$, $p = 0.036 < 0.05$). The difference between MFCC and CQCC does not reach the significance threshold ($p = 0.202$), nor does the difference between MFCC and LFCC ($p = 0.311$). These findings are consistent with the limited size of the test set; evaluations on larger datasets are expected to yield increased statistical power.

7.4.4. Attack-Wise Performance Analysis

The attack-wise analysis across 13 spoofing systems in the ASVspoof 2021 LA evaluation partition reveals distinct performance profiles for each feature representation. Results are presented in Table 8.

Table 8. Attack-wise accuracy analysis.

Attack	N	MFCC (%)	LFCC (%)	CQCC (%)	Best
A07	43	88.37	88.37	81.40	MFCC/LFCC
A08	42	88.10	80.95	80.95	MFCC
A09	49	79.59	83.67	79.59	LFCC
A10	39	76.92	82.05	84.62	CQCC
A11	57	80.70	75.44	78.95	MFCC
A12	43	86.05	79.07	81.40	MFCC
A13	38	81.58	81.58	86.84	CQCC
A14	44	77.27	77.27	81.82	CQCC
A15	48	81.25	77.08	79.17	MFCC
A16	48	91.67	87.50	95.83	CQCC
A17	38	78.95	76.32	86.84	CQCC
A18	30	90.00	83.33	90.00	MFCC/CQCC
A19	50	82.00	82.00	86.00	CQCC

Bonafide	546	81.50	81.50	83.33	CQCC
----------	-----	-------	-------	-------	------

The attack-wise analysis yields several important findings. CQCC achieves the highest accuracy on 6 of the 13 spoofing systems (A10, A13, A14, A16, A17, A19), with a particularly notable result of 95.83% on the A16 system. MFCC attains the best performance on A08, A11, A12, and A15. Two ties were observed: A07 (MFCC and LFCC, both 88.37%) and A18 (MFCC and CQCC, both 90.00%). LFCC achieves the highest accuracy solely on the A09 system (83.67%). It should be noted that the bonafide row in Table 8 represents genuine speech classification performance and is not counted among the 13 spoofing systems. Collectively, these findings confirm that no single cepstral representation consistently dominates across all attack categories, providing strong empirical support for multi-feature detection strategies.

7.5. Discussion

The experimental results provide quantitative support for the central hypothesis of this study: the discriminative power of cepstral representations varies with attack type, and single-feature approaches may prove insufficient for robust deepfake speech detection [11,12]. The achievement of above 80% accuracy and above 87% ROC-AUC by all three methods demonstrates that cepstral features combined with a linear SVM provide a meaningful detection baseline. However, EER values in the 15–18% range indicate that more advanced feature representations or classifiers are warranted for high-security biometric authentication deployments [18,19].

To contextualize the performance of the hand-crafted cepstral features relative to end-to-end deep learning approaches, a RawNet2-based model was trained on the same balanced subset of 5571 samples under identical data conditions. Training was conducted in a CPU-only environment (PyTorch 2.12.0, no GPU acceleration) for 10 epochs with a batch size of 16 and a learning rate of 0.0001. The training loss curve across all 10 epochs is presented in Figure 7.

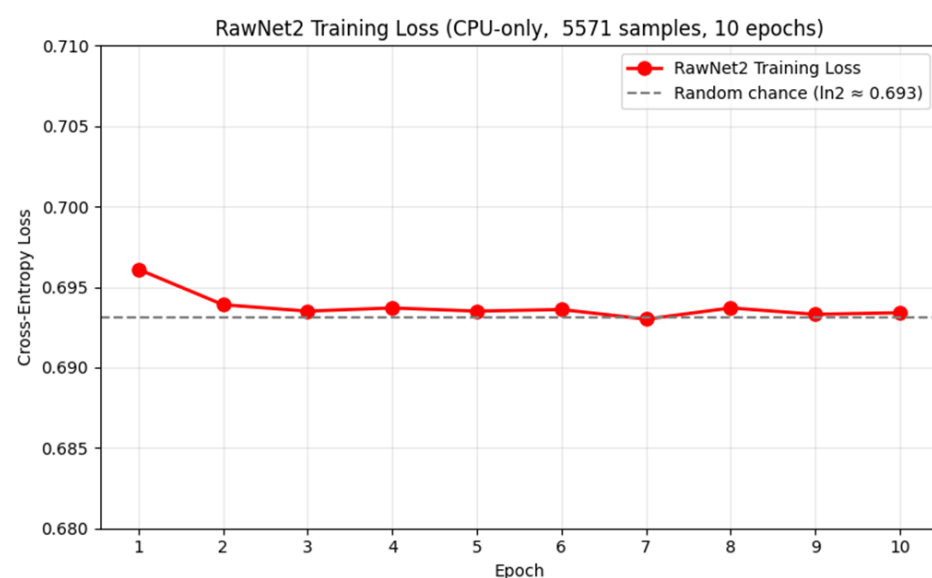


Figure 7. RawNet2 training loss curve over 10 epochs (CPU-only, 5571 samples). The dashed line represents the theoretical random chance level ($\ln 2 \approx 0.693$).

As illustrated, the loss remained at approximately 0.693 throughout training, the theoretical value for random binary classification confirming that no meaningful learning occurred.

The corresponding classification performance is summarized in Table 9. RawNet2 achieved 50.04% accuracy (95% CI: [47.11, 52.97]), 50.00% ROC-AUC, and 50.00% EER (95% CI: [47.07, 52.93]) values that are statistically indistinguishable from random binary classification. This outcome reflects a well-documented practical limitation of end-to-end deep models operating on raw waveforms: such architectures require GPU-accelerated hardware and substantially larger training corpora to converge reliably under conditions that were not available in the present computational environment. In contrast, the cepstral feature representations examined in this study, combined with a linear SVM classifier, consistently achieved above 80% accuracy without requiring specialized hardware, demonstrating their practical advantage in computationally constrained settings.

Table 9. Performance of RawNet2 deep learning baseline under CPU-only training conditions. * CPU-only training (PyTorch 2.12.0, no GPU acceleration), 10 epochs, batch size 16, learning rate 0.0001. Training loss remained at approximately 0.693 throughout, indicating no convergence.

Feature	Accuracy (%)	95% CI (Acc)	ROC-AUC (%)	EER (%)	95% CI (EER)
RawNet2 *	50.04	[47.11, 52.97]	50.00	50.00	[47.07, 52.93]

The attack-wise decomposition reveals that MFCC, despite its limitations in capturing high-frequency artifacts, retains a competitive advantage on specific TTS-based attack systems, a finding that underscores the complementary nature of the three representations. Future work should investigate the integration of these cepstral representations through early or late fusion strategies, as the complementarity demonstrated by the attack-wise analysis provides strong empirical motivation for this direction. Combining cepstral features with CNN, LSTM, or Transformer-based classifiers, or conducting comparative evaluations against self-supervised learning models such as wav2vec 2.0 and HuBERT under adequate GPU infrastructure, represents a promising avenue for substantially improving generalization performance. Additionally, the robustness of these representations under real-world transmission conditions telephone channel distortions, background noise, and codec compression warrants systematic investigation in future studies. The scope of detection may also be extended beyond speech signals to general audio content, as demonstrated by recent systems such as FakeSound [32]. Speech signal representation performance depends not only on feature type but on the holistic design of the detection system [31].

8. Conclusions

This study presented a systematic comparative and experimental investigation of three cepstral feature extraction methods, MFCC, LFCC, and CQCC, for deepfake speech detection on the ASVspoof 2021 Logical Access evaluation partition. Evaluated under identical experimental conditions using a linear SVM classifier, CQCC achieved the highest numerical performance across accuracy, ROC-AUC, and EER metrics; however, the difference between MFCC and CQCC does not reach statistical significance, indicating that the observed numerical advantage of CQCC over MFCC is not statistically supported under the current evaluation conditions, while MFCC demonstrated competitive results on specific TTS-based attack categories and LFCC exhibited complementary discriminative capacity on the A09 attack system. The attack-wise decomposition across 13 spoofing systems confirmed that no single cepstral representation universally dominates, providing compelling empirical evidence for the value of multi-feature detection strategies. These findings establish that deepfake speech detection performance is substantially dependent on the choice of frequency scale and cepstral transformation design, and that complementary cepstral representations hold significant potential for

building more robust and generalizable anti-spoofing systems. It is anticipated that the findings of this study will serve as a solid reference point for the development of more comprehensive deepfake speech detection frameworks in future research.

Author Contributions: Conceptualization, E.Ö. and N.A.; data curation, E.Ö. and N.A.; investigation, N.A.; methodology, E.Ö. and N.A.; supervision, E.Ö. All authors declare equal and joint responsibility for the study. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ASVspooF 2021 dataset used in this study is publicly available at <https://www.asvspooF.org>. The evaluation partition and associated metadata files can be accessed through the official ASVspooF challenge repository [19].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, Z.; Yamagishi, J.; Kinnunen, T.; Hanilçi, C.; Sahidullah, M.; Sizov, A.; Evans, N.; Todisco, M.; Delgado, H. ASVspooF: The automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 588–604. <https://doi.org/10.1109/JSTSP.2017.2671435>.
2. Bonastre, J.-F.; Delgado, H.; Evans, N.; Kinnunen, T.; Lee, K.A.; Liu, X.; Nautsch, A.; Noé, P.-G.; Patino, J.; Sahidullah, M.; et al. Benchmarking and challenges in security and privacy for voice biometrics. In Proceedings of the 2021 ISCA Symposium on Security and Privacy in Speech Communication (SPSC 2021), Magdeburg, Germany, 10–12 November 2021; pp. 52–56. <https://doi.org/10.21437/SPSC.2021-11>.
3. Dewi, S.P.; Prasasti, A.L.; Irawan, B. The study of baby crying analysis using MFCC and LFCC in different classification methods. In Proceedings of the 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 16–18 July 2019; IEEE: New York, NY, USA, 2019; pp. 18–23. <https://doi.org/10.1109/ICSIGSYS.2019.8811070>.
4. Almutairi, Z.; Elgibreen, H. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms* **2022**, *15*, 155. <https://doi.org/10.3390/a15050155>.
5. Yamagishi, J.; Nose, T.; Zen, H.; Ling, Z.-H.; Toda, T.; Tokuda, K.; King, S.; Renals, S. Robust speaker-adaptive HMM-Based Text-to-Speech Synthesis. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1208–1230. <https://doi.org/10.1109/TASL.2009.2016394>.
6. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. In Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW 9), Sunnyvale, CA, USA, 13–15 September 2016; p. 125.
7. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 4779–4783. <https://doi.org/10.1109/ICASSP.2018.8461368>.
8. Zhang, B.; Cui, H.; Nguyen, V.; Whitty, M. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. *Sensors* **2025**, *25*, 1989. <https://doi.org/10.3390/s25071989>.
9. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>.
10. Ameer, H.; Javed, A.R.; Iqbal, F.; Kryvinska, N.; Almadhor, A.S.; Jalil, Z.; Borghol, R. Deepfake audio detection via MFCC features using machine learning. *IEEE Access* **2022**, *10*, 134018–134028. <https://doi.org/10.1109/ACCESS.2022.3231480>.
11. Sahidullah, M.; Kinnunen, T. Local spectral variability features for speaker verification anti-spoofing. *Digit. Signal Process.* **2016**, *50*, 1–11. <https://doi.org/10.1016/j.dsp.2015.10.011>.
12. Todisco, M.; Delgado, H.; Evans, N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.* **2017**, *45*, 516–535. <https://doi.org/10.1016/j.csl.2017.01.001>.

13. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A comparison of features for synthetic speech detection. In Proceedings of Interspeech 2015, Dresden, Germany, 6–10 September 2015; pp. 2087–2091. <https://doi.org/10.21437/Interspeech.2015-472>.
14. Muckenhirn, H.; Korshunov, P.; Magimai-Doss, M.; Marcel, S. Long-term spectral statistics for voice presentation attack detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2098–2111. <https://doi.org/10.1109/TASLP.2017.2743340>.
15. Tak, H.; Patino, J.; Nautsch, A.; Evans, N.; Todisco, M. An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification. In Proceedings of Odyssey 2020—The Speaker and Language Recognition Workshop, Tokyo, Japan, 1–5 November 2020; pp. 214–221. <https://doi.org/10.21437/Odyssey.2020-47>.
16. Lavrentyeva, G.; Novoselov, S.; Tseren, A.; Volkova, M.; Gorlanov, A.; Kozlov, A. STC antispooofing systems for the ASVspoof2019 challenge. In Proceedings of Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1033–1037. <https://doi.org/10.21437/Interspeech.2019-1768>.
17. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Proceedings of Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2–6. <https://doi.org/10.21437/Interspeech.2017-1111>.
18. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K.A.; et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101114. <https://doi.org/10.1016/j.csl.2020.101114>.
19. Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H.; Kinnunen, T.; Todisco, M.; Yamagishi, J.; Evans, N.; Nautsch, A.; et al. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2507–2522. <https://doi.org/10.1109/TASLP.2023.3285283>.
20. Zhang, X.; Ge, Z.; Tivig, M. Adversarial attacks on audio deepfake detection models. *Appl. Sci.* **2023**, *13*, 5933. <https://doi.org/10.3390/app13105933>.
21. Rabiner, L.; Schafer, R.W. *Digital Processing of Speech Signals*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1978; ISBN 9780132143608.
22. Rabiner, L.R.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1993; ISBN 9780130151575.
23. Smith, S.W. *The Scientist and Engineer's Guide to Digital Signal Processing*; California Technical Publishing: San Diego, CA, USA, 1997; ISBN 9780966017632.
24. Cooley, J.W.; Tukey, J.W. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **1965**, *19*, 297–301. <https://doi.org/10.1090/S0025-5718-1965-0178586-1>.
25. Oppenheim, A.V.; Schafer, R.W. *Discrete-Time Signal Processing*, 3rd ed.; Pearson Education: Upper Saddle River, NJ, USA, 2010; ISBN 9780131988422.
26. Harris, F.J. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE* **1978**, *66*, 51–83. <https://doi.org/10.1109/PROC.1978.10837>.
27. Bracewell, R.N. *The Fourier Transform and Its Applications*; McGraw-Hill: New York, NY, USA, 1986; ISBN 9780070070158.
28. Brown, J.C. Calculation of a constant Q spectral transform. *J. Acoust. Soc. Am.* **1991**, *89*, 425–434. <https://doi.org/10.1121/1.400476>.
29. Patil, H.A.; Patil, A.T.; Kachhi, A. Constant Q cepstral coefficients for classification of normal vs. pathological infant cry. In Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 7392–7396. <https://doi.org/10.1109/ICASSP43922.2022.9746946>.
30. Mittal, A.; Dua, M. Constant Q cepstral coefficients and long short-term memory model-based automatic speaker verification system. In Proceedings of International Conference on Intelligent Computing, Information and Control Systems (ICICCS 2020); Pandian, A.P., Palanisamy, R., Ntalianis, K., Eds.; Springer: Singapore, 2021; pp. 895–904. https://doi.org/10.1007/978-981-15-8443-5_76.
31. Haykin, S.; Van Veen, B. *Signals and Systems*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2005; ISBN 9780471164746.
32. Xie, Z.; Li, B.; Xu, X.; Liang, Z.; Yu, K.; Wu, M. FakeSound: Deepfake general audio detection. In Proceedings of Interspeech 2024, Kos Island, Greece, 1–5 September 2024; pp. 112–116. <https://doi.org/10.21437/Interspeech.2024-1703>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.