Lightweight Clustering of Cepstral Features for Deepfake Audio Detection

Jean Alex Firmino Pinhata
Dept. of Data Science
Ourinhos College of Technology
Ourinhos, Brazil
jeanpinhata@gmail.com

Douglas Rodrigues
Department of Computing
Sao Paulo State University
Bauru, Brazil
d.rodrigues@unesp.br

Alex M. G. de Almeida
Department of Computing
Ourinhos College of Technology
Ourinhos, Brazil
alex.marino@fatecourinhos.edu.br

Kelton Costa
Department of Computing
Sao Paulo State University
Bauru, Brazil
kelton.costa@unesp.br

João Paulo Papa
Department of Computing
Sao Paulo State University
Bauru, Brazil
papa@fc.unesp.br

***Abstract*—This paper presents an exploratory, fully unsupervised analysis of how short-term cepstral representations cluster bona fide (legitimate) and spoofed (illegitimate) speech samples from the ASVspoof 2021 DF corpus. The analysis employs MFCC, MFCC Δ, MFCC $\Delta\Delta$, LFCC, and CQCC features, constructing utterance-level vectors by aggregating frame-wise coefficients using means and standard deviations. Subsequently, the K-means clustering algorithm is applied, with the number of clusters (k) predetermined by the Elbow method.**

Cluster quality is evaluated using both intrinsic and extrinsic metrics, along with per-cluster analysis of homogeneity and Shannon entropy. Two-dimensional visualizations are generated via UMAP dimensionality reduction. The results reveal that CQCC achieves the best overall separation between bona fide and spoofed speech, whereas MFCC produces several clusters that are almost exclusively composed of spoofed samples. In contrast, the dynamic coefficients ($\Delta\Delta$ and Δ) degrade the cluster structure.

These findings demonstrate that lightweight, unsupervised cepstral front-ends can uncover meaningful spoofing attack patterns and emerge as promising candidates for pre-filtering or routing stages in anti-spoofing pipelines. Nevertheless, their actual impact on supervised performance metrics, such as Equal Error Rate, still warrants robust investigation in future work.

***Index Terms*—audio deepfake, unsupervised learning, handcratfted features, UMAP, K-Means, ASVspoof**

I. Introduction

Recently, with the rapid evolution and popularization of Artificial Intelligence—exemplified by Large Language Models (LLMs) and underpinned at their core by advances in generative models such as *Generative Adversarial Networks* (GANs) and *WaveNet*—it has become possible to create deepfake audio derived from synthetic recordings or manipulations capable of reproducing the human voice with impressive realism [10].

These artifacts faithfully replicate timbre, prosody, and affective nuances, thereby making it increasingly challenging to distinguish between bona fide and spoofed speech. In this scenario, serious risks have emerged, including financial fraud through voice cloning in banking systems, disinformation in political campaigns, and social manipulation.

The pervasive integration of artificial intelligence into daily life has made advanced speech synthesis tools widely accessible—such as *Tacotron* [15] and commercial voice-cloning platforms—exacerbating the problem of deepfake audio and underscoring the need for robust countermeasures. Although supervised methods still dominate the literature and achieve high accuracy with remarkably low Equal Error Rate (EER), they depend on large labeled datasets and often generalize poorly to unseen spoofing attacks [14]. In response, unsupervised approaches—especially clustering techniques—have emerged as promising alternatives, as they can uncover latent patterns in the speech signal without relying on labels and help identify distinctive spectral anomalies characteristic of synthesized and deepfake audio; motivated by this perspective, this work presents a fully unsupervised exploratory analysis of how short-term cepstral features cluster bona fide and spoofed utterances in the ASVspoof 2021 Deepfake (DF) subset.

This work proposes a fully unsupervised approach for the exploratory analysis of deepfake audio samples using the ASVspoof 2021 dataset [16]. Short-term spectral features—namely MFCC, MFCC Δ, MFCC $\Delta\Delta$, LFCC, and CQCC—are extracted and normalized, then submitted to the K-means algorithm to generate partitional clusters, with the number of clusters k automatically determined via the Elbow method.

The separability between bona fide and spoofed samples was evaluated using two-dimensional UMAP projections, combined with intrinsic clustering metrics (Silhouette and Calinski–Harabasz) and extrinsic metrics (Rand Index (RI), Adjusted Rand Index (ARI), Homogeneity (HOM), Completeness (COM), and V-Measure (VME)). In addition, we performed a per-cluster analysis to quantify entropy and homogeneity for each partition individually.

The main objective of this work is to assess the discriminative power of short-term cepstral representations and to identify latent patterns that allow us to distinguish between bona fide and spoofed audio even in the absence of labels.

We hypothesize that it is possible to obtain homogeneous groupings in clustered partitions and leverage them as preliminary countermeasures for more computationally expensive supervised methods.

This study reinforces the viability of lightweight and interpretable approaches for audio deepfake detection, offering a promising avenue for building hybrid countermeasure systems in voice biometrics and digital security.

The remainder of this paper is organized as follows. Section II presents the theoretical background on deepfakes, acoustic features, and clustering algorithms; Section III describes the materials and methods used; Section IV discusses the analysis of the results; and Section V provides the final remarks and suggestions for future work.

II. Background

This section describes the main components of this work, encompassing the formal definition of deepfake audio, the definitions of the acoustic features explored, the K-means clustering algorithm, the UMAP dimensionality reduction technique, and the evaluation metrics. In addition to describing these components, the section concludes with a review of related work, situating this study within the current state of the art in the field.

A. Conceptual Definitions

*1) **Deepfake Audio**:* The term deepfake audio refers to the synthetic generation of human speech with high fidelity, using artificial intelligence models capable of imitating the timbre, prosody, and vocal style of real individuals [6]. This technology has evolved rapidly with the advent of architectures such as *Generative Adversarial Networks* (GANs), *autoencoders*, and neural vocoders (such as *WaveNet* and *HiFi-GAN*), making falsified audios virtually indistinguishable from authentic recordings.

Two main approaches dominate the creation of voice deepfakes:

- **TTS (Text-to-Speech)**: converts text into synthetic speech without requiring source audio.
- **VC (Voice Conversion)**: transforms real audio from one speaker to imitate another, preserving linguistic content.

*2) **Short-Term Acoustic Features**:* Speech signal analysis heavily relies on representations that capture spectro-temporal structure within short windows. The selected *features*—MFCC, Δ, $\Delta\Delta$, LFCC, and CQCC—are widely established in the literature for their discriminative power in speech recognition and *spoofing* detection tasks [18].

*a) **MFCC, Δ, and $\Delta\Delta$.**: Mel-Frequency Cepstral Coefficients* (MFCC) model human auditory perception through the mel scale [3]. The process involves frame segmentation, FFT application, mel-scale filtering, log-energy computation, and DCT:

$$\text{MFCC}_n = \sum_{m=1}^{M} E_m \cos\left[\frac{\pi n}{M}(m - 0.5)\right]. \tag{1}$$

First- and second-order derivatives (Δ and $\Delta\Delta$) incorporate the signal's temporal dynamics, enriching the representation and making it more sensitive to voice manipulations.

*b) **CQCC.**: Constant-Q Cepstral Coefficients* (CQCC) use the *Constant-Q Transform* (CQT), which provides logarithmic resolution, ideal for detecting phase artifacts in synthetic audio [12]. The CQT is followed by linear resampling and DCT:

$$\text{CQCC}_n = \sum_{k=1}^{K} \log|X(k)| \cos\left[\frac{\pi n}{K}(k - 0.5)\right]. \tag{2}$$

CQCC outperforms MFCC in *ASVspoof* challenges due to its greater sensitivity at low frequencies and ability to capture spectral irregularities associated with artificial synthesis [11].

*c) **LFCC.**: Linear Frequency Cepstral Coefficients* (LFCC) are linear variants of MFCC, where the mel scale is replaced by uniformly spaced filters. They serve as a *baseline* to evaluate the impact of perceptual compression on the discriminability of cepstral representations [1].

These *features*, taken together, form the basis of the unsupervised *clustering* process, enabling the exploration of latent patterns without manual labels.

B. Clustering and Analysis Techniques

*1) **K-Means and Elbow Method**:* The *K-Means* algorithm partitions data into groups by minimizing the sum of intra-*cluster* distances [4]. The *Elbow* method is used to determine the optimal number of clusters (k), by analyzing the distortion curve (*Within-Cluster Sum of Squares* – WCSS) and identifying the inflection point where marginal compactness gains become negligible [8]. In this work, the method was applied independently to each feature set, resulting in k values between 5 and 8.

*2) **Dimensionality Reduction with UMAP**:* For visualization of clusters, the *Uniform Manifold Approximation and Projection* (UMAP) algorithm was used [5], which preserves both local and global data structures through weighted graph construction and cross-entropy-based optimization. Configured with *n_neighbors* = 30 and *min_dist* = 0.3, the method generated two-dimensional projections that facilitated visual analysis of separability between *bona fide* and *spoof* samples.

*3) **Evaluation Metrics**:* The quality of the clusters was assessed using a set of **intrinsic** and **extrinsic** metrics to provide a comprehensive analysis of the coherence and structure of the generated *clusters*.

Intrinsic metrics, such as the *Silhouette* coefficient [7] and the *Calinski–Harabasz* index [2], quantify group separability and compactness using only internal data distances. Extrinsic metrics, such as the Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and the Homogeneity (HOM), Completeness (COM), and V-Measure (VME) indices, compare the obtained clusters with reference labels when available. Labels are never used during clustering and serve only as ground truth for extrinsic evaluation metrics (RI, ARI, AMI, HOM, COM, VME).

Regarding the assessment of cluster homogeneity, we employed the Shannon entropy [9], which enabled us to quantify the degree of internal heterogeneity, or the informational purity, of each cluster. The combination of these metrics enabled a rigorous quantitative comparison of the different spectral representations, clearly revealing the relative performance of each feature-extraction approach.

C. *Related Work*

Several recent studies address deepfake audio detection and the application of *clustering* techniques for this purpose. The literature covers both the use of classical acoustic features (MFCC, CQCC, LFCC) and unsupervised approaches evaluated via entropy.

The exponential growth of voice synthesis techniques driven by deep neural networks has enabled the generation of highly realistic synthetic audios capable of imitating real individuals and compromising automatic authentication systems [6]. These systems, widely used in banking, virtual assistants, and voice biometrics, become vulnerable to sophisticated *spoofing* attacks.

According to [17], voice synthesis and conversion technologies have evolved to the point of generating signals indistinguishable from real ones, even capable of deceiving automatic verifiers. *Text-to-Speech* models produce entirely synthetic speech, while *Voice Conversion* models operate on existing natural speech. Both, when trained with adequate data, pose significant threats to biometric system security.

Recent studies show that spectral features, such as MFCC coefficients, remain relevant in detecting anomalies in synthesized audios, especially when combined with robust machine learning techniques. Furthermore, representations like LFCC and CQCC have demonstrated superior performance in differentiating *bona fide* from *spoof* samples [13], exploring distinct aspects of the signal's spectral structure.

In the field of unsupervised learning, algorithms such as *K-Means*, *Fuzzy C-Means*, and *DBSCAN* have been used to identify latent patterns in voice databases and group samples with similar acoustic behavior. Evaluating the purity of formed groups through entropy measures has proven to be a promising strategy for estimating the discriminative capacity of representations.

The *ASVspoof 2021* dataset stands out as a benchmark for validating *spoofing* detection methods, offering a wide variety of *bona fide* and *spoof* samples generated by different synthesis techniques [16]. Its diversity and careful curation make it a gold standard for experiments in the field.

III. Materials and Methods

This study adopts a quantitative-descriptive experimental approach to qualitatively evaluate deepfake audio samples by extracting acoustic features and applying unsupervised clustering techniques. The materials used and the methodological procedures adopted are described below.

A. *Materials*

The dataset used and the computational tools employed for processing, analysis, and visualization are presented.

*1) **Dataset**:* Samples from the *ASVspoof 2021* dataset were used, widely recognized in the specialized literature on automatic speaker verification and voice forgery detection. The *dataset* contains samples labeled as *bona fide* (authentic) and *spoof* (falsified)—the distribution can be observed in Table I—generated by different voice synthesis and conversion techniques. The samples were converted to a uniform audio format and normalized in amplitude before processing.

TABLE I: Distribution of bona fide and spoofed utterances in the ASVspoof 2021 Deepfake partition [16].

Partition	**Bonafide**	**Spoof**	**Total**	**% Spoof**
Train	18000	37000	55000	67.3
Development	2500	2500	5000	50.0
Evaluation	3500	3500	7000	50.0
TOTAL	**24000**	**43000**	**67000**	**64.2**

For the experiments reported in this work, the training and development subsets of ASVSpoof 2021 were used.

*2) **Implementation and Experimental Considerations**:* The entire *pipeline* was implemented in Python 3.12 and executed on standard Google Colab CPU runtime, each full clustering run (feature extraction, standardization, Elbow search and K-Means fitting) completed in the order of a few minutes, indicating that the proposed pipeline can be executed with modest computational resources. The code is modular, with separate *scripts* for each stage (feature extraction, k determination, *clustering*, and visualization).

The `umap-learn` library was used for dimensionality reduction with *UMAP*, and hyperparameters were selected empirically on small development splits. We observed that small variations around $n_{\text{neighbors}} = 30$ and min_dist $= 0.3$ did not qualitatively change the main clustering patterns discussed in this paper. Preliminary tests were conducted on subsets of *ASVspoof 2021* (e.g., 10% of the samples) for hyperparameter tuning, runtime estimation, and bottleneck identification, ensuring scalability to the full dataset.

Specialized libraries in audio signal processing, machine learning, statistical analysis, and data visualization were used:

- **Librosa**: audio signal loading and manipulation; extraction of MFCCs, LFCCs, and Δ.
- **SoundFile (sf)**: efficient reading of FLAC format files.
- **NumPy** and **Pandas**: matrix operations, *array* manipulation, and *DataFrame* structuring.
- **Scikit-learn**: normalization (*StandardScaler*), *K-Means clustering*, and intrinsic metric computation (*Silhouette Score, Calinski–Harabasz*.
- **Spafe**: extraction of CQCC coefficients.
- **SciPy**: statistical functions and entropy calculation when needed.
- **Matplotlib** and **Seaborn**: generation of *Elbow* curves, *scatterplots*, and other visualizations.

- **umap-learn**: non-linear data projection for 2D/3D visualization.

This setup enabled a robust unsupervised analysis suitable for identifying hidden patterns without manual annotation—a relevant scenario for voice deepfake detection.

B. *Methods*

The methodology is based on unsupervised machine learning techniques, using *clustering* to identify patterns in acoustic data without prior labels. The workflow is illustrated in Figure 1, which presents the complete *pipeline*: from raw data input to quantitative evaluation.

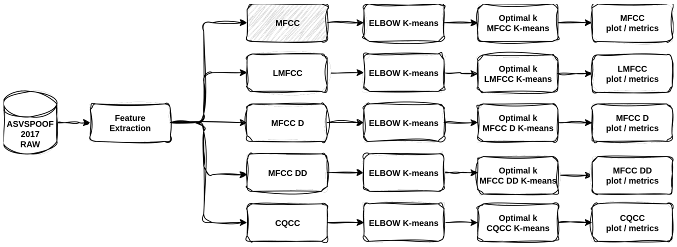


Fig. 1: Overview of the experimental *pipeline*: feature extraction, dimensionality reduction, *clustering*, and evaluation.

The dataset used is *ASVspoof 2021*, widely employed in *spoofing* detection research. It contains *raw* (unprocessed) recordings of genuine and falsified speech, simulating attacks via voice synthesis and conversion. The choice of *raw* data reduces biases introduced by preprocessing.

*1) **Feature Extraction**:* The initial stage transforms *raw* audio signals into compact and discriminative vector representations suitable for *clustering* algorithms. The following short-term *features* were extracted: **MFCC**, **MFCC** Δ, **MFCC** $\Delta\Delta$, **LFCC**, and **CQCC**.

Each recording was segmented into 20–30 ms *frames* with 50% overlap, using Hamming windows to mitigate spectral leakage. The resulting *features* were normalized (standardized) to ensure numerical stability and comparability in subsequent *clustering*.

*2) **Utterance-level feature representation**:* For each recording, short-term feature extraction produces a matrix $\mathbf{X} \in \mathbb{R}^{T\times D}$, where T is the number of frames and D is the number of coefficients per frame (e.g., $D = 20$ for MFCCs or LFCCs and $D = 30$ for CQCCs in our experiments). To obtain a fixed-dimensional vector per utterance, we aggregate simple statistics over time. For each coefficient dimension d, we compute the sample mean μ_d and standard deviation σ_d across all frames of that utterance:

$$\mu_d = \frac{1}{T}\sum_{t=1}^{T} X_{t,d}, \qquad \sigma_d = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\big(X_{t,d} - \mu_d\big)^2}.$$

The final utterance-level representation is the concatenation of these statistics,

$$\mathbf{f} = [\mu_1, \ldots, \mu_D, \sigma_1, \ldots, \sigma_D] \in \mathbb{R}^{2D}.$$

This procedure yields, for example, a 40-dimensional vector for MFCCs and LFCCs and a 60-dimensional vector for CQCCs. All utterance-level vectors are subsequently standardized using a global `StandardScaler` fitted on the ASVspoof 2021 DF split before clustering. For the Δ and $\Delta\Delta$ variants, the same aggregation scheme is applied to the corresponding frame-level sequences, producing utterance-level vectors of identical dimensionality.

*3) **Determination of the Optimal Number of Clusters**:* For each feature set, the *Elbow* method was applied alongside the *K-Means* algorithm to determine the optimal number of *clusters* (k). The *Elbow* method was executed by varying k from 2 to 20 *clusters*, using the *Within-Cluster Sum of Squares* (WCSS) metric to evaluate clustering quality at each configuration. *K-Means* partitions the data into k groups by minimizing the sum of intra-*cluster* distances (inertia), with the inflection point in the WCSS vs. k curve considered as the indicator of the optimal number.

Once the optimal k was defined via *Elbow* analysis, the final *K-Means* was executed for each representation. The algorithm was initialized with *K-Means++* using a fixed seed (*random_state*=42) to ensure full reproducibility of the experiment, and iterated until convergence, assigning each feature vector to its nearest centroid and updating the centroids at each iteration. Hyperparameters:

- Maximum number of iterations: 300;
- Convergence tolerance: 10^{-4};
- Initialization: *K-Means++*;
- Random seed: *random_state*=42;
- Tested k range in *Elbow*: 2 to 20 *clusters*;
- *Elbow* evaluation metric: WCSS.

After defining k, *UMAP* was used to reduce dimensionality and generate *scatterplots* for visualization, facilitating qualitative inspection of separability between *bona fide* and *spoof* samples.

*4) **Generation of Evaluation Artifacts**:* The metrics were computed *per feature* – MFCC, LFCC, MFCC Δ, MFCC $\Delta\Delta$, and CQCC – enabling a systematic comparison of the performance of each representation. The results were consolidated into comparative tables, highlighting the representation that best separates *bona fide* from *spoof* samples in a strictly unsupervised scenario.

To support qualitative analysis, high-resolution *Elbow* curves and *UMAP* projections were generated. These artifacts support subsequent discussions and informed selection of *features* for *anti-spoofing* countermeasures.

IV. Results and Discussion

In this section, we analyse the results of our unsupervised clustering pipeline applied to different acoustic representations from the ASVspoof 2021 corpus. Using Elbow curves, UMAP projections, and intrinsic/extrinsic metrics, we compare MFCC, LFCC, MFCC Δ, MFCC $\Delta\Delta$, and CQCC in terms of how well they separate *bona fide* and *spoof* samples without supervision. The analysis contrasts traditional cepstral features with Constant-Q-based representations and discusses practical implications for designing robust, hybrid, or semi-supervised anti-spoofing systems.

A. Elbow Curve Analysis

Figure 2a shows a clear elbow for CQCC at $k = 5$ (distortion $\approx 8.2 \times 10^5$), after which inertia reductions become marginal, indicating a good balance between compactness and model complexity. For MFCC and LFCC (Figures 2b and 2e), the elbow is located at $k = 7$ with similar distortion values ($\approx 3.9 \times 10^7$), characterised by a sharp decrease up to $k \approx 6$ followed by stabilisation.

Dynamic derivatives behave differently: MFCC Δ favours $k = 7$, whereas MFCC $\Delta\Delta$ suggests $k = 6$ and exhibits stronger post-elbow oscillations, indicating higher sensitivity to temporal artefacts. Overall, the curves consistently indicate k between 5 and 7, compatible with a binary structure (*bona fide* vs. *spoof*) plus attack subgroups. CQCC stands out by requiring fewer clusters and presenting the sharpest elbow, supporting its use for automatic k selection in efficient unsupervised partitioning.

B. Global Metrics Analysis

After determining the optimal k values using the Elbow method (ranging from 5 to 7 clusters depending on the feature), the final clustering was quantitatively evaluated using the intrinsic and extrinsic metrics presented in Table II. The results reveal significant differences in discriminative performance across acoustic representations, with CQCC ($k = 5$) emerging as the most promising due to its superior global balance.

Detailed analysis shows that CQCC achieved the best overall performance, with higher ARI (0.46569) and AMI (0.089344), indicating stronger agreement with ground-truth labels. It also attained superior homogeneity (0.155906), V-Measure (0.089441), and the highest Calinski–Harabasz index (14,283.95), evidencing compact, well-separated clusters.

MFCC and LFCC (both with $k = 7$) obtained identical but intermediate results, with ARI of 0.33466 and Silhouette score of 0.21144. Although their Calinski–Harabasz index (13,028.29) suggests reasonable internal separation, the higher class mixing indicates lower discriminative ability than CQCC.

Dynamic features clearly underperformed. MFCC Δ ($k = 7$) yielded the lowest ARI (0.19738) among static representations, while MFCC $\Delta\Delta$ ($k = 6$) showed the weakest metrics overall (ARI 0.049268, Calinski–Harabasz 9,262.38), indicating that temporal derivatives introduce noise that hinders class separation.

In summary, the combined evaluation of adjusted (ARI, AMI) and intrinsic (Calinski–Harabasz) metrics consistently indicates that CQCC is the most robust representation for unsupervised clustering in spoofing detection. MFCC and LFCC remain viable but less effective alternatives, whereas dynamic derivatives would require additional preprocessing to become competitive, reinforcing the advantage of features explicitly designed for anti-spoofing countermeasures.

C. ***Qualitative Analysis via UMAP Projections***

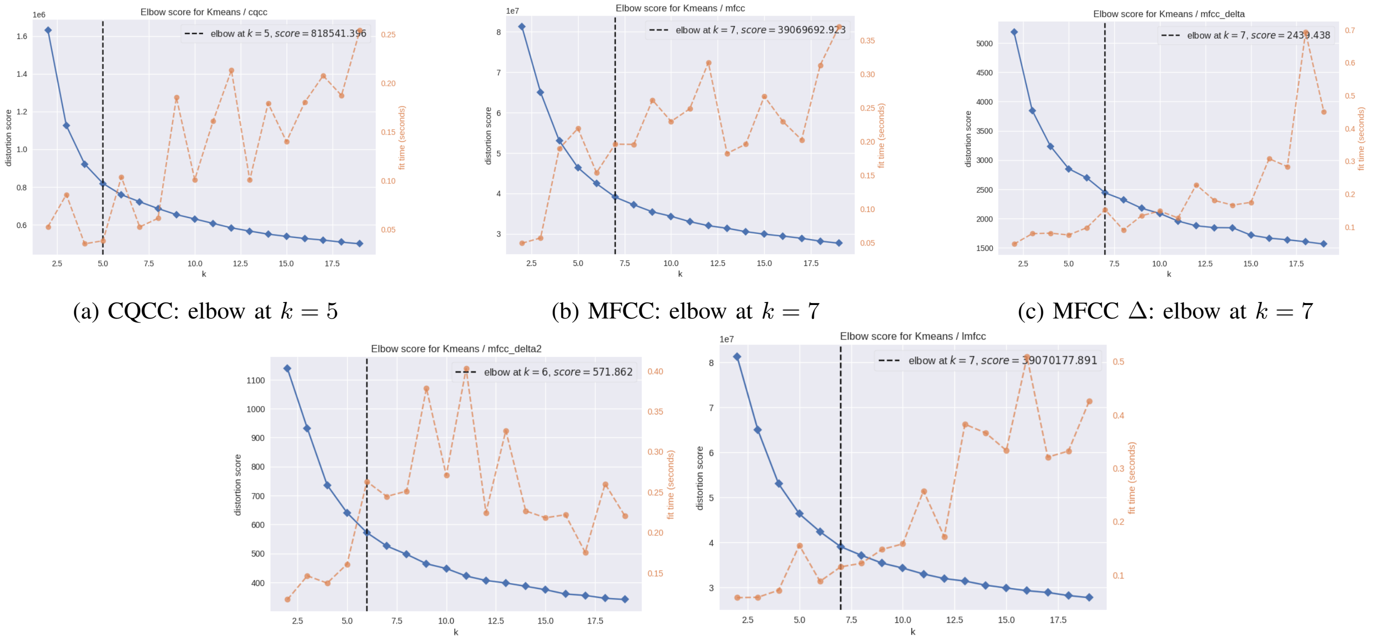
For a rigorous qualitative assessment of separability between *bona fide* (black squares) and *spoof* (inverted triangles) samples in reduced feature spaces, we examined 2D UMAP scatterplots (n_neighbors=30, min_dist=0.3) colored by optimized K-means clusters. Figure 3 organizes these visualizations in a 2×3 grid for direct comparison across acoustic representations. The analysis is anchored in the metrics from Table II, focusing on homogeneity (class purity per cluster), completeness (class coverage), Adjusted Rand Index (adjusted agreement with ground truth), Silhouette Score (intrinsic cohesion/separation), and Calinski-Harabasz (inter/intra-cluster variance ratio). Separability is quantified by minimal visual overlap between *bona fide*/*spoof* symbols and cluster alignment with true classes, avoiding excessive fragmentation or mixing.

In MFCC Delta ($k = 7$, Figure 3a), a globular distribution with well-defined clusters (vibrant colors) is observed, but with moderate mixing—many *spoof* triangles infiltrate *bona fide*-dominated clusters (e.g., dark blue). This reflects low ARI (0.197), HOM (0.047), and COM (0.019), indicating weak separability (clusters capture only 19% of classes completely). The high SIL (0.285) suggests reasonable internal cohesion (CAL=12,582), but post-elbow visual oscillation implies sensitivity to dynamic noise, reducing generalization for binary detection.

LFCC ($k = 7$, Figure 3b) shows a configuration similar to standard MFCC, with elongated clusters and significant overlap (e.g., green and red mix *bona fide* and *spoof*). Identical metrics to MFCC (ARI=0.335, HOM=0.254, COM=0.082, SIL=0.211, CAL=13,028) confirm intermediate separability: 25% pure clusters (HOM) but only 8% completeness, with moderate ARI suggesting partial label alignment. The linear structure (colored bands) indicates capture of low-frequency variations, but fragmentation dilutes the primary distinction.

For CQCC ($k = 5$, Figure 3d), the projection reveals superior separability, with compact and minimally overlapping clusters (e.g., blue for dominant *bona fide*, orange/red for isolated *spoof*). Despite slightly lower HOM (0.156 vs. 0.254 for MFCC), the highest ARI (0.466) and CAL (14,284) attest to better global agreement and intrinsic quality. Completeness (COM=0.063) indicates well-delineated attack subgroups, with a superior V-Measure (0.089), confirming that CQCC's variable-frequency resolution promotes more discriminative manifolds, reducing visual mixing and enhancing unsupervised accuracy.

Finally, MFCC Double Delta ($k = 6$, Figure 3e) exhibits the worst separability, with fragmented clusters and high overlap (e.g., dark blue extensively mixes symbols). Metrics confirm: minimum ARI (0.049), HOM (0.023), COM (0.011), and VME (0.015), with an inflated SIL (0.342) due to small clusters, but a low CAL (9,262) indicating high intra-cluster variance. Temporal oscillation introduces noise, resulting in spurious partitions that fail to capture the *bona fide*/*spoof* dichotomy.



(a) CQCC: elbow at $k = 5$

(b) MFCC: elbow at $k = 7$

(c) MFCC Δ: elbow at $k = 7$

(d) MFCC $\Delta\Delta$: elbow at $k = 6$

(e) LFCC: elbow at $k = 7$

Fig. 2: Elbow curves for determining the optimal number of clusters (k).

TABLE II: Clustering evaluation metrics for each feature with its respective optimal k.

Feature	K	RI	ARI	AMI	HOM	COM	VME	SIL	CAL
MFCC	7	0.754481	0.33466	0.124266	0.253709	0.082343	0.124333	0.211144	13028.293910
MFCC Δ	7	0.463013	0.19738	0.026552	0.046819	0.018600	0.026624	0.285214	12582.082400
MFCC $\Delta\Delta$	6	0.512873	0.049268	0.014751	0.022603	0.011079	0.014869	0.341935	9262.375438
LFCC	7	0.754481	0.33466	0.124266	0.253709	0.082343	0.124333	0.211144	13028.293910
CQCC	5	0.486192	**0.46569**	**0.089344**	**0.155906**	**0.062707**	**0.089441**	0.235166	**14283.954828**

In summary, CQCC results suggest superior separability (maximum ARI/CAL, minimal mixing), aligning clusters with spoofing subtypes without supervision, while dynamic derivatives degrade performance (low HOM/ARI). This visual analysis quantitatively validates feature selection for anti-spoofing countermeasures, with implications for downstream EER reduction.

D. Per-Cluster Metrics Analysis

Tables III and IV provide a detailed breakdown of entropy (impurity: values close to 0 indicate pure clusters) and homogeneity (purity relative to the majority class: 1.0 represents perfect homogeneity) metrics for K-means clusters on MFCC ($k = 7$) and CQCC ($k = 5$), respectively.

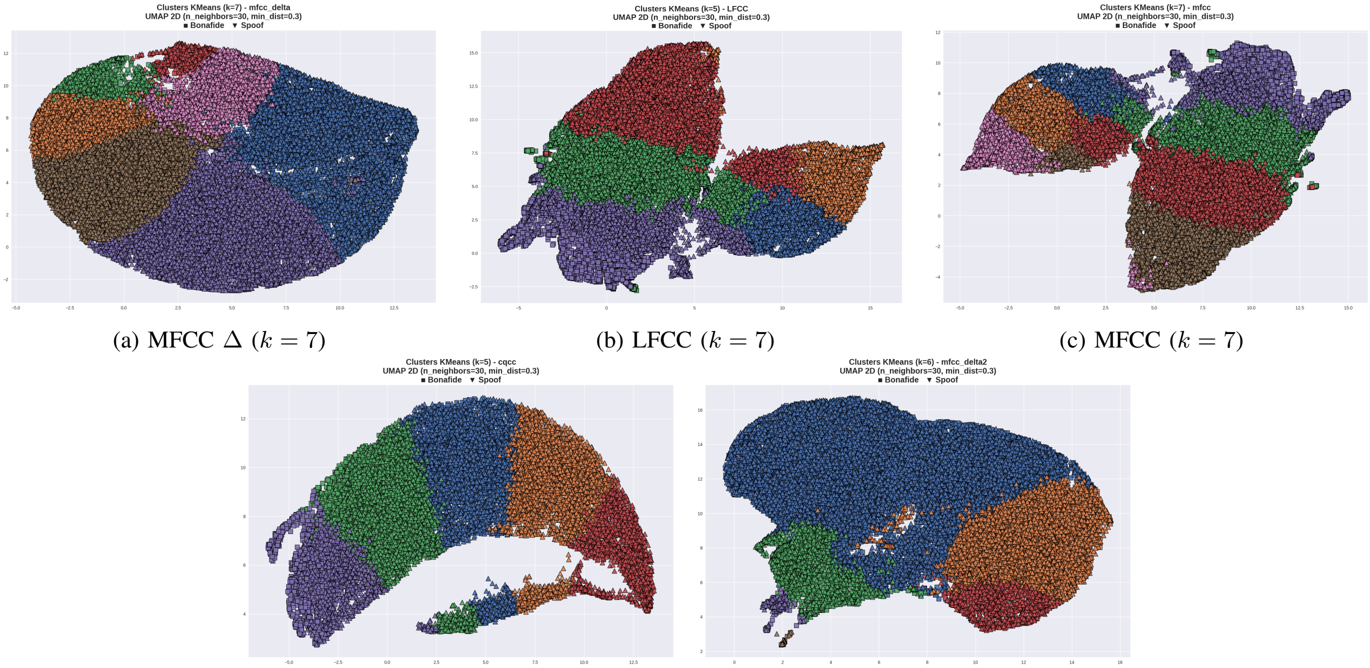
In MFCC (Table III), wide variation in entropy (0.0000 to 0.9848) and homogeneity (0.5725 to 1.0000) is observed, with three highly pure spoof-dominant clusters (0, 1, and 6: entropy $\leq$0.0303, homogeneity $\geq$0.9969) but four significantly mixed clusters (e.g., cluster 2: entropy 0.9848, homogeneity 0.5725), including the only bona fide-majority cluster (cluster 4: entropy 0.7843, homogeneity 0.7664). The global weighted entropy is $\approx$0.633 and the weighted homogeneity is $\approx$0.785, indicating large mixed clusters that capture spoof substructures but dilute binary separability.

In CQCC (Table IV), entropy is more moderate (0.1970 to 0.9077) and homogeneity consistently high (0.6769 to 0.9695), with one nearly pure cluster (cluster 3: entropy 0.1970, homogeneity 0.9695) and a balanced bona fide-majority cluster (cluster 4: entropy 0.9077, homogeneity 0.6769). Weighted entropy $\approx$0.756 and homogeneity $\approx$0.764 reflect more cohesive and discriminative clusters, with lower variation and better global alignment (ARI 0.466 vs. 0.335 for MFCC), highlighting CQCC's superior ability to isolate deepfake/spoof artifacts with reduced impurity.

TABLE III: Per-cluster metrics for MFCC ($k = 7$).

Cluster	Size	Entropy	Hom.	Bonafide	Spoof	Majority Class
0	2654	0.0048	0.9996	1	2653	Spoof
1	3523	0.0000	1.0000	0	3523	Spoof
2	7181	0.9848	0.5725	3070	4111	Spoof
3	9155	0.7813	0.7681	2123	7032	Spoof
4	3681	0.7843	0.7664	2821	860	Bonafide
5	4964	0.7208	0.8006	990	3974	Spoof
6	1615	0.0303	0.9969	5	1610	Spoof

Identifying homogeneous groups offers clear advantages:



(a) MFCC Δ ($k = 7$)

(b) LFCC ($k = 7$)

(c) MFCC ($k = 7$)

(d) CQCC ($k = 5$)

(e) MFCC $\Delta\Delta$ ($k = 6$)

Fig. 3: 2D UMAP projections of the analyzed acoustic representations, colored by optimized K-means clusters and marked with *bona fide* (black squares) and *spoof* (inverted triangles).

TABLE IV: Per-cluster metrics for CQCC ($k = 5$).

Cluster	Size	Entropy	Hom.	Bonafide	Spoof	Majority Class
0	4912	0.8221	0.7431	1262	3650	Spoof
1	3709	0.5423	0.8754	462	3247	Spoof
2	4599	0.9885	0.5632	2009	2590	Spoof
3	1736	0.1970	0.9695	53	1683	Spoof
4	2408	0.9077	0.6769	1630	778	Bonafide

it facilitates the interpretation of sub-patterns (pure clusters as prototypes of specific attacks), improves downstream tasks (pseudo-labels for semi-supervision, reducing EER), and optimizes efficiency (less refinement is needed for mixed clusters). CQCC maximizes these benefits with superior consistency, while MFCC excels in local purity but suffers from mixing in large clusters.

The results indicate superior performance of CQCC and MFCC representations in the unsupervised approach for spoofing analysis on the ASVspoof 2021 dataset, positioning them as the most robust and discriminative configurations among those evaluated. CQCC with $k = 5$ stands out due to the highest Adjusted Rand Index (ARI $\approx$ 0.466), Calinski-Harabasz index ($\approx$14,284), and global alignment with ground-truth labels, combined with compact clusters of high local homogeneity (up to 0.970 in spoof subgroups) and moderate weighted entropy (0.756), reflecting its superior ability to capture phase artifacts and variable frequency resolution typical of deepfakes. MFCC (and its identical LFCC variant) with $k = 7$ ranks second, with solid ARI ($\approx$0.335), lower weighted entropy ($\approx$0.633), and perfectly pure clusters (homogeneity 1.000 in three cases), ensuring exceptional local purity and interpretability of spoof attack subtypes—far outperforming dynamic derivatives (Δ and $\Delta\Delta$), which exhibit fragmentation, high impurity, and near-random metrics. This dual leadership validates CQCC as ideal for global accuracy and MFCC for granular purity, recommending their prioritization or combination in unsupervised anti-spoofing countermeasure systems.

A closer inspection of the resulting clusters reveals an even more promising practical application: several clusters (e.g., clusters 0, 1, and 6 for MFCC) exhibit near-perfect spoof homogeneity (**Shannon entropy $<$ 0.2 and homogeneity $>$ 0.97**) as shown in Table III. These highly pure spoof groups can be directly exploited to bootstrap lightweight one-class classifiers or semi-supervised systems without any manual labeling. Future work will explore training One-Class SVMs using only the most compact bona-fide cluster as inliers and the identified spoof-dominant clusters to set decision thresholds, potentially yielding an EER that competes with existing methods, with the additional computational overhead expected to remain small compared to full neural-network-based countermeasures.

V. Conclusion

Overall, this work is an exploratory, fully unsupervised study on the ASVspoof 2021 DF corpus. Its findings have clear limits of generalisation, since they have not yet been validated on other datasets, partitions or attack conditions.

We examined how short-term cepstral representations (MFCC, MFCC Δ, MFCC $\Delta\Delta$, LFCC and CQCC) cluster bona fide and spoofed speech when utterance-level vectors are built from frame-wise means and standard deviations and K-Means is applied with k chosen by the Elbow method. Using intrinsic and extrinsic cluster metrics, plus Shannon entropy and UMAP visualisations, CQCC with $k = 5$ emerged as the best option for global separation, with the highest Adjusted Rand Index and Calinski–Harabasz indices. MFCC and LFCC with $k = 7$ do not separate classes as well globally, but still yield several clusters with very high spoof purity. Dynamic coefficients (Δ and $\Delta\Delta$) tend to reduce compactness and class separation and appear less useful in this setting.

The per-cluster analysis shows that, for MFCC in particular, some clusters are almost entirely spoof, with homogeneity above 97% and low entropy. These groups can be viewed as high-confidence "spoof-dominant" regions that might support early rejection of likely spoofed traffic or help initialise or regularise supervised models. This remains a hypothesis grounded solely on cluster compositions, not on supervised performance metrics.

As future work, we will investigate how these unsupervised clusters can be integrated into end-to-end anti-spoofing pipelines. In particular, we plan to train compact one-class SVMs on the most bona-fide-like cluster, use spoof-dominated clusters to define conservative thresholds, and evaluate the resulting system with metrics such as EER on ASVspoof 2021 DF and ASVspoof 2025 DF. This will clarify the practical impact of lightweight cepstral clustering front-ends on automatic speaker verification robustness.

References

[1] Moustafa Alzantot, Ziqi Wang, and Mani B. Srivastava. "Deep Residual Neural Networks for Audio Spoofing Detection". In: *arXiv preprint arXiv:1907.00501* (2019).

[2] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1 (1974), pp. 1–27. DOI: 10.1080/03610927408827101.

[3] Steven Davis and Philip Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366.

[4] Stuart P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137.

[5] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018). Available: https://arxiv.org/abs/1802.03426.

[6] Yisroel Mirsky and Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". In: *ACM Computing Surveys* 54.1 (2021), 7:1–7:41. DOI: 10.1145/3425780.

[7] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[8] Ville Satopää et al. "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior". In: *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2011, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20. URL: https://doi.org/10.1109/ICDCSW.2011.20 (visited on 06/08/2025).

[9] Claude E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[10] Jonathan Shen et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4779–4783. DOI: 10.1109/ICASSP.2018.8461368.

[11] Hemlata Tak et al. "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification". In: *arXiv preprint arXiv:2004.06422* (2020).

[12] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification". In: *Computer Speech & Language* 45 (2017), pp. 516–535.

[13] Massimiliano Todisco et al. "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection". In: *Interspeech*. 2019, pp. 1008–1012. DOI: 10.21437/Interspeech.2019-2249.

[14] Luisa Verdoliva. "Media Forensics and DeepFakes: An Overview". In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932. DOI: 10.1109/JSTSP.2020.3002101. URL: https://doi.org/10.1109/JSTSP.2020.3002101 (visited on 06/08/2025).

[15] Yuxuan Wang et al. "Tacotron: Towards End-to-End Speech Synthesis". In: *Interspeech*. 2017, pp. 4006–4010. DOI: 10.21437/Interspeech.2017-1452. URL: https://arxiv.org/abs/1703.10135 (visited on 06/08/2025).

[16] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, et al. "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection". In: *Proc. ASVspoof 2021 Workshop*. 2021, pp. 1–6. DOI: 10.21437/ASVSPOOF.2021-1.

[17] Bowen Zhang et al. "Audio deepfake detection: What has been achieved and what lies ahead". In: *Sensors (Basel, Switzerland)* 25.7 (2025), p. 1989.

[18] Tao Zhang. "Deepfake generation and detection, a survey". In: *Multimedia Tools and Applications* 81.5 (2022), pp. 6259–6276.