Mining Hidden Partitions of Voice Utterances using Fuzzy Clustering for Generalized Voice Spoofing Countermeasures

Sarah Mohammed Altuwayjiri[1], Ouiem Bchir[2], Mohamed Maher Ben Ismail[3]
Department of Computer Science,
College of Computer and Information Sciences,
King Saud University,
Riyadh 11543, Saudi Arabia[1,2,3]
Department of Computer Science,
College of Computing and Informatics,
Saudi Electronic University,
Riyadh 11673, Saudi Arabia[1]

***Abstract*—The high level of usability achieved by voice biometrics compared to other biometric authentication modalities has promoted the widespread use of automatic speaker verification (ASV) systems as authentication tools for several services in various domains. Despite their satisfactory performance, ASV systems are vulnerable to malicious voice spoofing attacks. Hence, voice spoofing countermeasures have emerged as essential solutions to stop such harmful attacks and protect ASV systems as well as users' confidentiality. Typically, these countermeasures classify utterances into genuine and spoofing categories. In this research, we propose two voice spoofing countermeasures that mainly aim to improve the generalization of supervised learning models. This goal is achieved through the adaptive handling of the high variance of both utterance classes, i.e., genuine and spoofing classes. The proposed spoofing countermeasure addresses the poor generalization problem by identifying the hidden structure of each utterance category prior to the classification task. Specifically, fuzzy clustering algorithms were deployed to mine the hidden partitions of utterance classes. The conducted experiments showed that the proposed approach outperforms the state-of-the-art approaches in the ASVspoof 2017 dataset, with a testing EER equal to 1.07%.**

***Keywords*—Voice spoofing; spoofing countermeasure; classification; clustering**

I. Introduction

At present, biometric authentication along with other identification features is widely deployed to manage, administrate and control systems' accessibility in order to secure the applications and stored data [43]. In particular, the widespread use of biometric recognition systems has prompted research efforts to consider various modalities such as retinal, facial and speech data. Speaker verification (SV) has been introduced as a biometric recognition paradigm that uses human voiceprints to identify individuals every time they access a given service or system. SV-based identification is typically meant to compare the speaker's voice with the voiceprints previously recorded, then grant access to the identified persons only. The advent of voice assistant and smart home devices boosted the interest in automatic speaker verification (ASV) systems as promising alternatives to ensure the security of various smart home applications, smart devices, online payment processes and phone banking [30]. However, these ASV systems have proven to be vulnerable to voice spoofing attacks [41]. In fact, such attacks have been defined as presentation attacks (PA) according to the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) [15]. In fact, spoofing attacks occur when a fraudster falsifies another identity to access some personal or secured resources [26]. For instance, they can be achieved through replay attacks which consist of collecting voice samples of a particular person, manipulating them to produce a spoofing voice, and replaying the resulting spoofing voice to mislead an automatic speaker verification (ASV) system. This kind of voice manipulation can be performed using data voice conversion or speech synthesis algorithms [41, 48]. Obviously, in order to prevent these spoofing attacks, audio anti-spoofing countermeasures are required. A voice spoofing countermeasure is a classification system that can automatically categorize voice records into two predefined categories: genuine and spoofing. Typically, it comprises two main components: (i) an audio feature extractor, and (ii) a supervised learning model. In this context, various audio features have been investigated for designing highly discriminative descriptors. Namely, the constant Q cepstral coefficients (CQCC) [45] and the linear frequency cepstral coefficient (LFCC) [42] were proposed to better discriminate between spoofing and genuine voice records. Similarly, diverse classifiers, such as the Gaussian mixture models (GMMs) [39] and deep neural networks (DNNs) [36] have been extensively used to build models that can accurately map unseen voice records into the two predefined classes. Although different deep learning architectures, such as the residual neural networks (ResNets) [22] and the recurrent neural networks (RNNs) [9] have been adapted and used for anti-spoofing, GMM-based solutions overtake the state-of-the-art anti-spoofing recognition systems [41]. Nevertheless, one of the major unsolved issues affecting the reliability and accuracy of anti-spoofing recognition systems is the poor generalization of the learned models [41]. In other words, the learned model failed to predict unseen data instances. Generalization characterizes the model's ability to predict unseen data instances.

This limitation can be attributed to the high variance in both the genuine and the spoofed utterances. Specifically, genuine speech instances exhibit high interspeaker variance, owing to the discrepancies between speaker voices, as well as an intra-speaker variance because of the inconsistency in the human voice that can be affected by aspects such as the emotional state [1, 29]. These two types of variations also apply to spoofed speeches. Moreover, spoofing utterances witnessed other types of variations caused by the recording devices used to collect the original voice records and the algorithms used to manipulate them [1, 2, 9, 15, 22, 26, 29, 36, 39, 41, 42, 45, 48]. Hence, better handling of the high variance of the genuine and spoofed classes would improve the generalization of the spoofing countermeasures and, consequently, enhance the detection performance. In this paper, we propose to improve the recognition of voice spoofing utterances by tackling the problem of intraclass variation for two categories: genuine and spoofed. More specifically, we propose learning the underlying structure of each category (genuine/spoof) by clustering them into homogeneous sub-categories. The rest of the paper is organized as follows: In Section II, existing voice-spoofing countermeasures are surveyed. In Section III, background knowledge on clustering techniques is provided. The proposed approach is presented in Section IV. The experiments conducted are described in Section V along with the reported results and their analysis.

II. Literature Review

Recently, several spoofing countermeasures have been proposed. The development of these systems was boosted by contests in 2015 and 2017 [30, 43], which provided challenging data for anti-spoofing systems. More specifically, the training set contains five types of spoofing attack algorithms, referred to as known attacks, whereas the evaluation set, used for testing, contains the known attacks and five more types of attacks called unknown attacks. The proposed system amounts to classification systems for genuine and spoofed utterances. They aimed to discriminate spoofing utterances from genuine utterances. One of the main aspects that has been exploited is the presence of noise in the spoofing records [23]. During the playback and re-recording phases used by the replay attack, different types of noise are generated. These types of noise are mainly from the recording environment and recording device. They can potentially allow for differentiation between spoofed and genuine signals. In this context, both conventional and deep learning approaches have been reported.

A. Conventional Countermeasures

Typically, conventional spoofing/genuine recording classification comprises a feature extraction component followed by a classification component. The system proposed in [23] extracts the cepstral coefficient (CQCCs) [7] feature. A GMM [5] classifier was employed for the classification task. This system has been considered a baseline approach for recently proposed research for evaluating anti-spoofing systems [30, 43]. Alternatively, the system reported in [39] combines cochlear filter cepstral coefficients (CFCC) [33] and the instantaneous frequency (IF) [40]. The combined feature aims to capture the speech synthesis and voice conversion, thus characterizing spoofing utterances. It was then fed to the GMM classifier.

TABLE I. Summary of Conventional Approaches for Spoofing / Genuine Classification

Reference	Feature	Classifier	Dataset	Training EERor rate (%)	Testing EERor rate (%)
[45]	CQCCs	GMM	ASVspoof 2015 [49]	0.048	0.462
[27]	CQCCs	GMM	ASVspoof 2017 [27]	10.35	24.77
[39]	MFCC CFC-CIF	GMM	ASVspoof 2015 [49]	0.408	2.013
[42]	LFCC	GMM, SVM	ASVspoof 2015 [49]	0.11	1.67
[37]	MFCC, MFPC, CosPhasePCs	SVM with i-vectors	ASVspoof 2015 [49]	0.008	3.922

Similarly, the work in [42] focused on segregating spoofing records generated by voice conversion or speech synthesis algorithms. For this purpose, the authors in [42] conducted an empirical comparison of 19 different features to determine the most appropriate one for classifying spoofing versus genuine records. These features are then conveyed to both GMM [5] and SVM [6] classifiers. The experimental results reported in [42] showed that the system comprising the LFCC [53] feature extraction component and GMM classifier component outperformed all other considered systems. On the other hand, the work in [37] used SVM as a classifier [6] for different extracted features. In addition, the i-vector was used for each feature and then integrated into a one-centered i-vector with a normalized length. The experimental results in [37] showed that the MFCC [11], Mel-frequency principal coefficients (MFPC) [14], and CosPhase principal coefficients (CosPhasePC) [47] fed into the SVM classifier had better classification performance. Table I presents a brief summary of conventional approaches for spoofing/genuine classification. All of these systems have experimented on the ASVspoof 2015 dataset [49].

B. Deep Learning based Countermeasures

The successful achievement of deep neural networks (DNN) in classification tasks has motivated the application of such approaches for anti-spoofing. Recently, deep learning approaches have been proposed for voice spoofing classification. In particular, the residual network model (ResNet) found recent success in the works of [8, 30, 36]. Moreover, as voice spoofing data can be considered as a sequence classification task, recurrent neural networks (RNNs) [35] were also investigated in the works [9, 19, 31, 51, 52]. Furthermore, while raw audio data were considered as inputs to the DNN model in some studies [9, 19, 30], engineered features were considered in others [8, 31, 36, 52].

1) Residual Network based Approaches: The proposed system in [30] employs a dilated residual network (DRN) deep-learning architecture [21]. The latter is based on the ResNet model, and an attention-filtering mechanism. More precisely, the DRN uses convolution layers instead of fully connected layers, and alters the residual units by adding a dilation factor. The attention component aims to select important parts while ignoring unrelated ones, such as the background noise segments [50]. Similarly, the system proposed in [8] employed the ResNet [21] deep-learning model. However, the proposed approach applies a deep-learning architecture in conjunction with two low-level cepstral features. In fact, the input conveyed

TABLE II. SUMMARY OF DEEP LEARNING APPROACHES FOR SPOOFING / GENUINE CLASSIFICATION

Reference	Feature	Classifier	Dataset	Training EERor rate (%)	Testing EERor rate (%)
[30]	Signal Logspec via FFT	ResNet	ASVspoof 2017[27]	6.09	8.54
[8]	CQCC and MFCC	GMM, ResNet	ASVspoof 2017[27]	2.58	13.30
[36]	Fusion of HFCC and CQCC	DNN, SVM	ASVspoof 2017[27]	7.6	11.5
[9]	MFCC, Fbank	LSTM and GRU RNN	ASVspoof 2017[27]	6.32	9.81
[31]	CQT and FFT	LCNN, SVM, CNN + RNN	ASVspoof 2017[27]	3.95	6.73
[19]	Fbank	CNN + RNN (GRU)	ASVspoof 2015 [49]	0.03	1.97
[52]	Spectrogram features	CNN + RNN	ASVspoof 2015 [49]	0.40	3.33

to the network is not raw audio data but features extracted by MFCCs [11] and CQCCs [45]. Moreover, a GMM classifier is used at the back end of the network. The work in [36] uses a similar model, but it exploits high-frequency cepstral coefficients (HFCCs) instead of MFCCs [11] at the input of the network.

2) Recurrent Neural Network based Approaches: The authors of [9] used recurrent neural networks (RNN) [35] for spoofing/genuine record classification. More specifically, the proposed system employs long short-term memory (LSTM) [24] and a gated recurrent unit (GRU) [10]. LSTM has also been used in [44], where the proposed architecture consists of multiple dense layers followed by one or more LSTM layers. Similarly, the works in [19, 52] exploited the RNN [35] deep-learning model. However, an RNN is used with a convolutional neural network (CNN) [20]. More precisely, CNN is used as a feature extractor and RNN is used for processing long dependencies. The work in [19] crops the input records and trains the two models separately and uses a linear discriminant analysis (LDA) [25] as a back-end classifier. However, the work in [52] uses an end-to-end model. It uses a context window for input, and trains both models simultaneously by conjointly optimizing them through backpropagation. The combination of CNNs and RNNs was also exploited in [31]. Specifically, it fuses three approaches: the i-vector [12] approach, light convolutional neural network (LCNN) [46] approach, and CNN+RNN approach. Three inputs were considered separately in the first convolution layer yielding three variants of the LCNN-based model. More precisely, the first input consists of truncated normalized fast Fourier transform (FFT) spectrograms [34], the second is constant Q transform (CQT) [7], and the third is FFT with a sliding window. Table II provides a brief summary of deep learning approaches for spoofing/genuine classification.

C. Discussion

Previous studies tackled the generalization problem by mainly investigating various feature selection and fusion approaches [31, 36, 37], studying feature representations and diverse classification approaches [30, 39, 42, 45], and applying diverse deep learning architectures such as ResNet [8, 30, 36] and RNN [9, 19, 31, 51, 52]. Furthermore, for deep learning approaches both raw audio data [9, 19, 30], and engineered features [8, 31, 36, 52] are considered. Nevertheless, neither the engineered features nor those learned automatically by deep learning succeeded in alleviating the generalization problem. In fact, there was a discrepancy between the training and testing performances; they improve the prediction for seen utterances, but they are not able to generalize to unseen ones. Nevertheless, the baseline approach [45], which is based on extracting the CQCC feature and GMM-based classification approach, outperforms the other state-of-the-art approaches in terms of generalization on the ASVspoof 2015 dataset but failed on the ASVspoof 2017 dataset. On the latter dataset, the approach proposed in [31] is the best reported approach with a testing EERor rate of 6.73%.

III. CLUSTERING

Clustering is an unsupervised learning approach that groups unlabeled instances into homogeneous clusters based on certain criteria or similar functions. This allowed the exploration and analysis of the data. There are three main approaches to clustering: (i) hierarchical clustering, (ii) partitioning, and (iii) density-based clustering approaches [5]. Hierarchical clustering [17] creates a hierarchy of clusters following either a top–down (divisive) or a bottom–up (agglomerative) strategy. Alternatively, partitioning or centroid-based clustering [32], is characterized by learning a representative of each cluster such as the cluster centers. Accordingly, data instances were assigned to the cluster corresponding to the closest representative. For this purpose, the distance to the representatives is calculated using distance metrics such as the Euclidean distance or Manhattan distance. On the other hand, density-based clustering approaches [28] consider the density rather than distance to assign an instance to a cluster. Clustering is performed in such a way that dense instances form clusters, whereas sparse instances are considered noise and outliers. Density-based approaches are characterized by the arbitrary shapes of the clusters. Clustering approaches can also be categorized as crisp or fuzzy. Although crisp clustering approaches assign an instance exclusively to one cluster, fuzzy clustering can assign an instance to more than one cluster using a membership degree. The latter can be perceived as an instance's probability of belonging to a given cluster. In this way, fuzzy approaches can deal with real-world applications in which clusters exhibit overlapping boundaries [5]. In the following section, we describe three fuzzy clustering approaches that will be investigated to uncover the underlying structure of voice-spoofing data. Specifically, we consider fuzzy c-means (FCM) clustering [4], simultaneous clustering and attribute discrimination (SCAD) [18], and competitive agglomeration CA [17] algorithms.

A. Fuzzy C-Means

Fuzzy c-means (FCM) [4] clustering performs a fuzzy partitioning of the unlabeled data by minimizing the intra-cluster distances. More precisely, for a set of instances, x_j, it simultaneously learns the cluster representatives (centers), c_i,

and the fuzzy memberships, (u_{ij}), by minimizing the objective function:

$$J(\mathbf{B}, \mathbf{U}; \mathcal{X}) = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m \parallel x_j - c_i \parallel^2 \quad (1)$$

subject to

$$u_{ij} \in [0,1] \& \ \forall i, j \ and \sum_{i=1}^{C} u_{ij} = 1 \& \ \forall j \quad (2)$$

In Equation (1), x_j and $c_i \in R^d$ where d is the dimension of the vectors, m is a fuzzier that controls the membership fuzziness, C is the number of clusters and N is the number of instances.

B. Simultaneous Clustering and Attribute Discrimination

Simultaneous clustering and attribute discrimination (SCAD) [18] is an extension of FCM that addresses the problems of feature selection and aggregation. It learns the relevance feature weights, $\nu = [\nu_{ik}]_{i=1...c, j=1...d}$ with respect to each cluster. In addition to the centers, $C = [c_{ik}]_{i=1...c, j=1...d}$, and fuzzy membership, $U = [u_{ik}]_{i=1...c, j=1...N}$. This is achieved by minimizing the objective function, as follows:

$$J(\mathbf{C}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m \sum_{K=1}^{d} v_{ik} (x_{jk} - c_{ik})^2 \quad (3)$$

subject to

$$u_{ij} \in [0,1] \& \ \forall i, j \ and \sum_{i=1}^{C} u_{ij} = 1 \& \ \forall j \quad (4)$$

and

$$v_{ik} \in [0,1] \& \ \forall i, k \ and \sum_{i=1}^{d} v_{ik} = 1 \& \ \forall i \quad (5)$$

where C is the number of clusters, N is the number of instances and d is the feature size, and vik, cik and $uij \in R^d$ where d is the dimension of the vectors.

C. Competitive Agglomeration

Competitive agglomeration (CA) [17] is another extension of FCM that addresses the problem of estimating the number of clusters in an unsupervised manner. In fact, it learns the number of clusters while learning the cluster representatives and the fuzzy memberships. It combines hierarchical and partitioning clustering approaches, and thus benefits from their advantages. Specifically, CA applies the competitive agglomeration in order to select the best number of clusters. It begins by dividing the instances into small clusters. During the optimization process, the clusters compete over instances, and the empty clusters disappear gradually. The CA optimizes the following objective function:

$$J(\mathbf{B}, \mathbf{U}; \mathcal{X}) = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^2 . d_{ij}^2 (x_j, \beta_i) - \alpha \sum_{i=1}^{C} [\sum_{j=1}^{N} u_{ij}]^2 \quad (6)$$

where $B = (1, \ldots, c)$ are the cluster representatives, $dij^2(xj, \beta i)$ is the distance between feature vectors xj and prototype βi, and uij is the fuzzy membership of instance j with respect to cluster i. As can be seen, the objective function in (6) incorporates two terms: the first one is inherited from the FCM objective function (1). On the other hand, the second term in (6) is the competitive term that allows cluster competition to enclose data instances.

IV. Proposed Approach

Owing to the high intra-class variance of the spoofing and genuine categories, the sub-groups of these two categories are scattered. Moreover, spoofing subgroups overlap with genuine subgroups, and vice versa. This renders the classification problem even more challenging. In fact, the learned classification model is too complex and may result in the overfitting of the training dataset. This reflects the low generalization of the supervised learning model. Therefore, we propose splitting each category into homogeneous groups, and then classifying unknown instances by considering the closest sub-group. The proposed spoofing countermeasure based on homogeneous subcategories is illustrated in Fig. 1. As one can see, it starts by extracting audio features from the genuine and spoofing utterances. Then, the genuine instances are clustered separately to determine the representatives of the genuine sub-categories in an unsupervised manner. Similarly, spoofing instances were clustered in order to obtain the spoofing representatives. The learned sub-category representatives are then used to classify unknown instances. Specifically, for the clustering task, we propose employing prototype-based fuzzy clustering approaches. This choice is motivated by the need to learn cluster representatives, and the fact that fuzzy memberships are better at handling the overlapping boundaries of clusters. In other words, we intend to investigate several prototype-based fuzzy clustering approaches such as FCM [4] and SCAD[17, 18]. In fact, FCM-based clustering approaches learn the cluster centers which is not the case for other types of clustering approaches such as density or hierarchical-based clustering algorithms. Moreover, SCAD learns the relevance feature weights while clustering the data. This allows for the automatic selection and aggregation of the features. Similarly, CA automatically estimated the number of clusters while clustering the data. The three clustering algorithms under consideration are optimized iteratively by alternating the update of the centers, the fuzzy memberships, and eventually the relevance feature weights and the number of clusters through the use of closed-form update equations. Furthermore, we plan to explore the number of clusters that generate the optimal subcategories for the spoofing and the genuine classes. In fact, estimating the number of clusters allows the correct structure of the data to be uncovered. Therefore, it yields a better local classification which helps to lessen the generalization issue for unseen instances. An illustrative example of the proposed spoofing countermeasure based on homogeneous subcategories is shown in Fig. 2. As it can be seen, the spoofing utterances are clustered into six clusters ($S1, S2, S3, S4, S5$, and $S6$), while genuine utterances are clustered into four clusters ($G1, G2, G3$, and $G4$). Then, the blue unseen instance was compared to the ten learned representatives before assigning it to one of the clusters. Because $G1$, one of the representatives of the genuine category, is the closest to the blue unseen instance, the latter

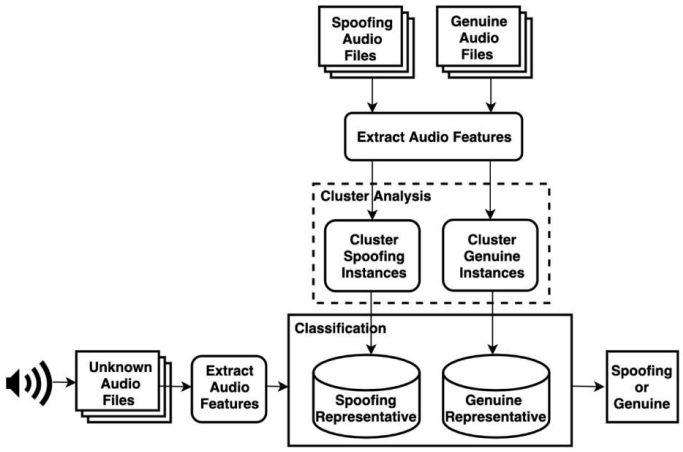

Fig. 1. General Overview of the Proposed Approach.

is classified as genuine.

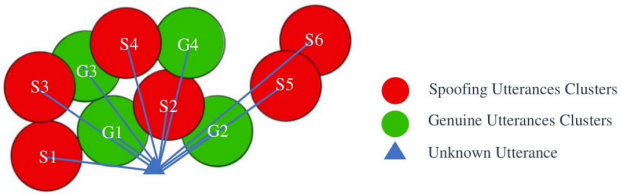

Fig. 2. Illustrative Example the Proposed Spoofing Countermeasure based on Homogenous Sub-Categories.

V. Experiments

The dataset used for the experiments was ASVspoof 2017 version 2.0 [13]. The dataset contained audio files with a sampling rate of 16 kHz and a 16-bit resolution, which were divided into three subsets: a training set containing 3016 files, a development set containing 1710 files, and an evaluation set with 13,306 files. The training set had 1507 genuine and 1507 replay files, the development set had 760 genuine and 950 replay files, and the evaluation set had 1298 genuine and 12,008 replay files [13]. The Mel-frequency cepstral coefficients' (MFCCs) [11] and the constant Q cepstral coefficients' (CQCCs) [45] features are extracted from the audio files. MFCC is computed by applying the discrete cosine transform (DCT) type 2 on a 20 ms audio frame. This generates an audio spectrum that reflects energy in different frequency bands. After spectrum computation, a bank of triangular filters was employed to warp the spectrum into the Mel-scale. Finally, the results of a Mel-scale filter bank are logarithmized and decorrelated by applying the DCT. Alternatively, CQCC is based on a constant Q transform (CQT). The latter is a time-frequency analysis tool for short-time Fourier transform (STFT) [7]. The number of bins per octave was set to 12, and the sampling frequency was set to 44,000 Hz. The CQCCs' spectrum was derived by first performing CQT transform on the audio frame. Next, the logarithm non-linearity and linearization of the CQT's geometric scale were applied. Then, the final 167 CQCC cepstral coefficients were obtained by applying the DCT [45]. Similar to the approaches described in Section III, the performance of the proposed approach is evaluated using an equal error rate (EER) [38]. It was calculated using a receiver operating characteristic (ROC) curve. More specifically, the EER is defined as the operating point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal [16, 38].

A. Experiment 1: Discovering the Underlining Structure using Fuzzy C-Means

In this experiment, we clustered the genuine and the spoof classes from the training subset separately, using fuzzy c-means [4]. The same number of clusters was used for both classes, and it was tuned from 2 to 16 with a step of 2. The learned sub-category representatives were then used for the classification of unknown instances from the testing subset, using the K-nearest neighbor classifier KNN [3] with K = 1. The experiment was conducted on the CQCC, MFCC, and a concatenation of the CQCC and MFCC independently. The EER with respect to the number of clusters on the considered features is shown in Fig. 3.

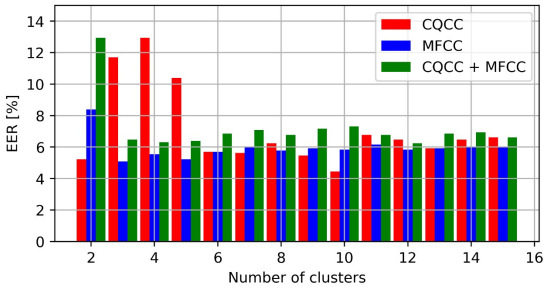

Fig. 3. EER with respect to the Number of Clusters when using Fuzzy C-Means [20] on CQCC, MFCC and the Concatenation of CQCC and MFCC.

As shown in Fig. 3, the EER varies with respect to the number of clusters and features considered. The best result is obtained when using a number of clusters equal to 2 for both classes on the CQCC feature. It reached an EER of 4.46%.

B. Experiment 2: Discovering the Underlining Structure using the Competitive Agglomeration

In this experiment, the underlining structure was learned using the competitive agglomeration (CA) [17] clustering algorithm in order to simultaneously cluster the training data and learn the optimal number of clusters. The same number of clusters was initially set to 100 for both classes (genuine and spoofed). Similar to Experiment 1, the learned cluster representatives are used for the classification of unknown instances from the testing subset, using the K-nearest neighbor classifier KNN [3] with K = 1. Moreover, an experiment was conducted on the CQCC, MFCC, and concatenation of the CQCC and MFCC independently. Table III reports the obtained EER, and the learned number of clusters with respect to the considered features.

As shown in Table III, the CQCC exhibited the lowest EER of 2.46%. Moreover, the optimal cluster learned by CA

TABLE III. EER AND THE LEARNED NUMBER OF CLUSTERS WHEN USING COMPETITIVE AGGLOMERATION (CA) [17] ON CQCC, MFCC, AND THE CONCATENATION OF CQCC AND MFCC

	CQCC	MFCC	CQCC+MFCC
EER	2.46%	14.86%	9.24%
No of genuine clusters	2	3	3
No of spoof clusters	2	2	3

TABLE IV. LEARNED FEATURE WEIGHTS WITH RESPECT TO EACH CLUSTER

	CQCC	MFCC
Cluster 1 (genuine)	0.92	0.08
Cluster 2 (genuine)	0.92	0.08
Cluster 1 (spoof)	0.91	0.09
Cluster 2 (spoof)	0.91	0.09

when using CQCC is two clusters for the genuine class and two clusters for the spoof class. This is similar to the result obtained in the first experiment using fuzzy c-means by tuning the number of clusters. We can then conclude that the CA clustering approach can learn the underlying structure of both the genuine and spoof classes while learning the optimal number of clusters.

C. Experiment 3: Discovering the Underlining Structure using Simultaneous Clustering and Attribute Discrimination

In this experiment, partitions of the genuine and spoof classes are mined using simultaneous clustering and attribute discrimination (SCAD) [18]. It aims to discover the underlying partitions while learning the optimal relevance feature weights of CQCC and MFCC audio features. In fact, the weights of each of these two considered features are learned with respect to each class. The number of clusters was set to two for the genuine class and two for the spoof class according to the results obtained in Experiment 2. The obtained partitions were then used for the classification task using the K-nearest neighbor classifier KNN [3] with K = 1. Table IV presents the obtained EER, and the learned number of clusters with respect to the considered features.

As reported in Table IV, the feature weights with the largest relevance were learned for the CQCC. This is in concordance with the results obtained in the previous experiments, which showed that the CQCC is more relevant for the classification of genuine/spoof utterances. To further investigate the CQCC feature, we applied SCAD to its 167 CQCC cepstral coefficients. In other words, each dimension of the CQCC is considered as a single feature. A feature relevance weight was then learned for each dimension. For Experiment 1, the number of clusters was tuned from 2 to 16 in steps of 2.

As shown in Fig. 4, when using the same number of clusters, the lowest EER, equal to 1.07, was obtained for the number of clusters equal to two. We can conclude then that performing SCAD on the CQCC yields better results. This is because it deals with the high dimension of CQCC by performing an optimal weighted sum of the coefficients. Fig. 5 shows the relevance feature weights for each cluster.

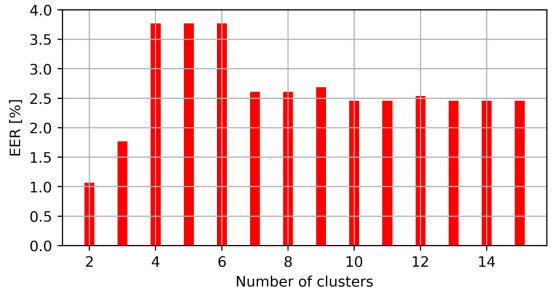

Fig. 4. EER with respect to the Number of Clusters when using SCAD [18] on CQCC Dimensions.

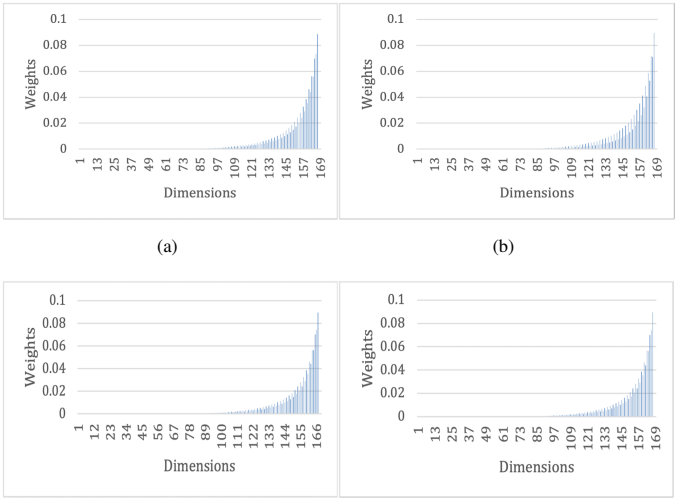

Fig. 5. Relevance Feature Weights with respect to (a) Cluster 1 (Genuine), (b) Cluster 2 (Genuine), (c) Cluster 1(Spoof), (d) Cluster 2 (Spoof).

D. Experiment 4: Performance Comparison with the State-of-the-Art Approaches

In this experiment, the performance of the proposed approach was compared to that of the state-of-the-art approach in the ASVspoof 2017. More specifically, the best results obtained using the considered clustering algorithms are compared to the conventional KNN [3] approach and to the countermeasures reported in the literature [8, 9, 27, 30, 31, 36]. To compare the proposed approach to KNN, we classified ASVspoof 2017 using KNN while tuning the neighboring parameter from 3 to 9. The experiment was conducted on MFCCs, CQCCs, features, and their corresponding concatenation. As shown in Fig. 6, the lowest EER (EER=3.62%) was obtained when using CQCC with K equal to 9.

Table V reports the training EER and the testing EER of the state-of-the-art approaches, the best KNN result, and the best results of the proposed approach with respect to the different clustering approaches under consideration.

As shown in Table V, the proposed approach outperforms the KNN classifier and the methods reported in the literature, regardless of the considered clustering algorithm. Moreover, it solves the generalization problem by reducing the performance

TABLE V. PERFORMANCE COMPARISON IN TERMS OF EER BETWEEN THE PROPOSED APPROACHES AND THE STATE-OF-THE-ART APPROACHES

Countermeasures	Training EER %	Testing EER%
Reported work in [30] (Featurs: Signal Logspec via FFT, Model: ResNet)	6.09	8.54
Reported work in [8] (Features: CQCC and MFCC, Model:GMM + ResNet)	2.58	13.30
Reported work in [36](Features: Fusion of HFCC and CQCC, Model:DNN + SVM)	7.6	11.5
Reported work in [9](Features: MFCC, Fbank, Model: LSTM + GRU RNN)	6.32	9.81
Reported work in [31](Features: CQT and FFT, Model: LCNN, SVM, CNN + RNN)	3.95	6.73
Reported work in [27](Feature: CQCC, Model: GMM)	10.35	24.77
KNN (K=9, Feature: CQCC)	1.05	3.62
Proposed approach based on Fuzzy C-Means (No of genuine clusters= 2, No of spoof clusters=2, Feature: CQCC)	3.15	4.46
Proposed approach based on CA (feature: CQCC)	2.67	2.46
Proposed approach based on SCAD No of genuine clusters= 2, No of spoof clusters=2, Features: coefficient of CQCC)	0.13	1.07

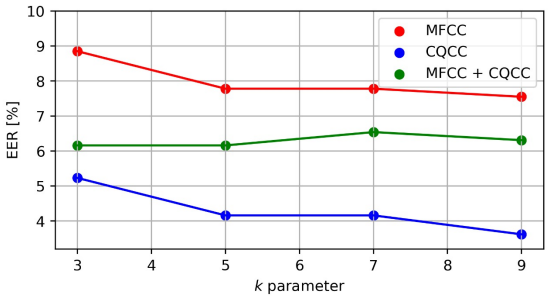

Fig. 6. EER with respect to the Number of Neighbor Parameter K when using KNN [3] on CQCC, MFCC, and the Concatenation of CQCC and MFCC.

gap between the training EER and the testing EER. This is achieved by mining the underlying structure of both genuine and spoof utterances. Furthermore, the lowest EER is obtained when using SCAD with a number of genuine clusters equal to 2, and a number of spoof clusters equal to 2 on CQCC coefficients. The achieved EER was 1.07%. This improved the result by a ratio of 5.66% compared with the best reported work [31], which achieved an EER of 6.73%.

VI. CONCLUSION

Voice spoofing is a prominent security risk that requires effective countermeasures to protect the user's information when using ASV systems. These countermeasures amount to the classification of the utterances into genuine or spoofing categories. However, this classification problem is challenging because of high class variance. In fact, genuine utterances are subject to variability due to the differences in speakers' voices and the discrepancies within the human voice due to emotions or other effects. Similarly, spoofing utterances are subjected to the same variability, in addition to the variability caused by the recording devices employed. Moreover, the diversity in the methods that produce spoofing utterances, such as voice manipulation and synthesis, contributes significantly to the variance in the spoofing class. This high variance in utterances drastically affects the performance of the spoofing classification task. Specifically, it limits the model's generalization and yields a less accurate system. Recently, considerable attempts have been made to address the low generalization of spoofing countermeasures. The reported works focused mainly on investigating various feature selection and fusion approaches, studying feature representations, and applying diverse deep learning architectures. However, neither handcrafted features nor deep learning-based descriptors have succeeded in alleviating the generalization problem. In fact, although they have shown slight improvements in the prediction performance of the models, they fail to generalize to unseen utterances. In fact, they are prone to overfitting, as indicated by the discrepancy between the training and testing performances. In this study, we devised a new countermeasure to address the low-generalization problem. Specifically, the proposed approaches mined the understructure of the genuine and spoofing utterances. This was achieved by integrating the clustering component into the classification process. The experimental results showed that mining hidden partitions of voice utterances using fuzzy clustering yielded a better generalization of the voice-spoofing countermeasure. In fact, the proposed approach outperformed the state-of-the-art approaches. Specifically, when using the CA clustering approach, the training and testing EERs were similar. Moreover, when using SCAD on the CQCC feature for a number of genuine clusters equals to 2, and a number of spoof clusters equal to 2, the performance is drastically improved with an EER of 1.07%. In future work, we suggest investigating additional audio features. Moreover, we intend to use CA as the first step in order to learn the number of clusters and the initial fuzzy memberships. Then, SCAD would be performed using the obtained results.

ACKNOWLEDGMENT

This work was supported by the Research Center of the College of Computer and Information Sciences at King Saud University. The authors are grateful for this support.

REFERENCES

[1] Moez Ajili. Reliability of voice comparison for forensic applications. Avignon, 2018.

[2] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013*, pages 1–8, October 2013.

[3] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*, 46(3):175–185, 1992.

[4] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, USA, 1981.

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[6] Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. A Training Algorithm for Optimal Margin Classifier. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 5, Aug 1996.

[7] Judith C. Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[8] Zhuxin Chen, Zhifeng Xie, Weibin Zhang, and Xiangmin Xu. ResNet and model fusion for automatic spoofing detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 102–106, 2017.

[9] Zhuxin Chen, Weibin Zhang, Zhifeng Xie, Xiangmin Xu, and Dongpeng Chen. Recurrent neural networks for automatic replay spoofing attack detection. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2052–2056, 2018.

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Y Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Jun 2014.

[11] Steven B. Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[12] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009.

[13] Héctor Delgado, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. 2018.

[14] P Ding and L Zhang. Speaker Recognition Using Principal Component Analysis. *Proc. ICONIP2001*, Jan 2001.

[15] Project Editor. DRAFT INTERNATIONAL STANDARD ISO / IEC DIS 30107-3 Information technology — Biometric presentation attack detection. 2017.

[16] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 2006.

[17] Hichem Frigui and Raghu Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):450–465, 1999.

[18] Hichem Frigui and Olfa Nasraoui. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 158–163. IEEE, 2000.

[19] Alejandro Gomez-Alanis, Antonio M. Peinado, Jose A. Gonzalez, and Angel M. Gomez. A deep identity representation for noise robust spoofing detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 676–680, 2018.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[22] Guy A. Hembury, Victor V. Borovkov, Juha M. Lintuluoto, and Yoshihisa Inoue. Deep Residual Learning for Image Recognition Kaiming. *Chemistry Letters*, 32(5):428–429, 2003.

[23] H.-G Hirsch and D Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, 4:29–32, jan 2000.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9(8):1735–1780, Dec 1997.

[25] Alan Izenman. *Linear Discriminant Analysis*. Springer, New York, NY, Jan 2008.

[26] Madhu R. Kamble, Hardik B. Sailor, Hemant A. Patil, and Haizhou Li. Advances in anti-spoofing: From the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*, 9, 2020.

[27] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[28] Hans-Peter Kriegel, Peer Kröger, Joerg Sander, and Arthur Zimek. Density-based Clustering. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 1(3):231–240, May 2011.

[29] Yoohwan Kwon, Soo Whan Chung, and Hong Goo Kang. Intra-class variation reduction of speaker representation in disentanglement framework. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 14–18, Shanghai, China, September 2020.

[30] Cheng-I Lai, Alberto Abad, Korin Richmond, Junichi Yamagishi, Najim Dehak, and Simon King. Attentive Filtering Networks for Audio Replay Attack Detection. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 6316–6320. IEEE, 2019.

[31] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 82–86, 2017.

[32] Y Leung, J Zhang, and Z Xu. Clustering by space-space filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1396–1410, Jan 2000.

[33] Qi Li. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 181–184. IEEE, 2009.

[34] Charles Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM, Jan 1992.

[35] L R Medsker and L C Jain. *Recurrent Neural Networks: Design and Applications*. International Series on Computational Intelligence. CRC Press, 1999.

[36] Parav Nagarsheth, Elie Khoury, Kailash Patil, and Matt Garland. Replay attack detection using DNN for channel discrimination. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 97–101, 2017.

[37] Sergey Novoselov, Alexander Kozlov, Galina Lavrentyeva, Konstantin Simonchik, and Vadim Shchemelinin. STC anti-spoofing systems for the ASVspoof 2015 challenge. pages 5475–5479. Mar 2016.

[38] John Oglesby. What's in a number? Moving beyond the equal error rate. *Speech Communication*, 17(1-2), 1995.

[39] Tanvina B. Patel and Hemant A. Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Germany, September 2015.

[40] T.F. Quatieri. *Discrete-time Speech Signal Processing: Principles and Practice*. Pearson Education Taiwan, 2005.

[41] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. Introduction to voice presentation attack detection and recent advances. In *Handbook of biometric anti-spoofing*, pages 321–361. Springer, 2019.

[42] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[43] Zia Saquib, Nirmala Salam, Rekha Nair, and Nipun Pandey. Voiceprint Recognition Systems for Remote Authentication-A Survey. *International Journal of Hybrid Information Technology*, 4(2), April 2011.

[44] Simone Scardapane, Lucas Stoffl, Florian Rohrbein, and Aurelio Uncini. On the use of deep recurrent neural networks for detecting audio spoofing attacks. In *Proceedings of the International Joint Conference on Neural Networks*, pages 3483–3490. IEEE, 2017.

[45] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In *Odyssey 2016*, pages 283–290, 2016.

[46] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[47] Zhizheng Wu, Eng Chng, and Haizhou Li. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2, Jan 2012.

[48] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.

[49] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[50] Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[51] Sung-Hyun Yoon, Hee-Soo Heo, Hye-Jin Shim, Ha-Jin Yu, and Jee-Weon Jung. Replay Spoofing Detection System for Automatic Speaker Verification Using Multi-Task Learning of Noise Classes. pages 172–176, 2018.

[52] Chunlei Zhang, Chengzhu Yu, and John H.L. Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE*

Journal on Selected Topics in Signal Processing, 11(4):684–694, 2017.

[53] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 559–564. IEEE, Dec 2011.