Investigating Feature Reduction Strategies for Replay Antispoofing in Voice Biometrics

Sapan H. Mankad(✉), Sanjay Garg, Megh Patel, and Harshil Adalja

CSE Department, Institute of Technology, Nirma University, Ahmedabad, India
{sapanmankad,sgarg,16BIT020,16BIT063}@nirmauni.ac.in

Abstract. One of the biggest challenges for voice based biometric solutions is in handling replay spoofing attacks. These attacks pose enormous threat on speaker verification system wherein the recorded voice of a genuine user is played in front of the authentication system to attempt unauthorized access. The problem with such system is to distinguish between origin of the input signal whether it comes from a human (live signal) or a device (spoofed signal). In this work, we compare filterbank based features and attempt to choose prominent features by employing some dimensionality reduction strategies. Low level, short-term spectral features have been used to represent audio files. Three methods for feature selection and feature construction are implemented and tested on these features. Results obtained on ASVspoof 2017 version 2 corpus indicate that entropy based feature selection approach gains 9.98% relative improvement over other approaches for feature reduction studied in this work, and an overall performance gain of 13.2% in terms of equal error rate reduction.

Keywords: Replay attacks · ASVspoof 2017 · Feature selection · Entropy · Dimensionality reduction

1 Introduction

Feature reduction is an approach wherein a set of features is replaced with an alternative, lower dimensional representation without significant loss of information. This is done either by removing irrelevant features from original set (known as, feature selection), or by transforming the original input feature space to a new, reduced feature space (feature transformation) by deriving new features (feature construction). Apart from improving performance, feature reduction helps in decreasing training time for model, and computational cost.

Voice, as a biometric, is the easiest mode of communication, and suits the best for person recognition. It provides hands-free access, too. Automatic Speaker Verification (ASV) aims at identifying authenticity of a given voice signal by analysing its characteristics, determining the source of the audio signal, and comparing it with stored patterns of the claimant user. The claim is rejected

B. Deka et al. (Eds.): PReMI 2019, LNCS 11942, pp. 400–408, 2019.
https://doi.org/10.1007/978-3-030-34872-4_44

if the calculated score between the stored model and claimed model does not exceed above a predetermined threshold. An ASV system operates in two phases: (i) enrollment phase - during this phase, the system works by preparing models of all registered speakers, (ii) testing phase - in which the model of a test sample is compared with that of a claimant speaker from stored models, and decision is finally given based on some scoring mechanism. Replay attacks are carried out by presenting pre-recorded audio samples of a genuine speaker to the ASV system. Hence this kind of attacks are often referred to as presentation attacks.

The main contribution of this paper is two fold: (1) We attempt to explore the best set of features which can address the problem of replay spoofing, and (2) investigating the most suitable feature reduction strategy for this task.

The rest of the paper is organized as follows. Section 2 discusses about the existing literature related to this work. An audio representation using filterbank features is given in Sect. 3. Section 4 discusses our proposed work, experimental setup and results discussion. Finally, the paper is concluded in Sect. 5.

2 Related Work

The use of feature selection strategies for speaker identification task dates back in 1975 when [12] proposed a new probability-of-error criterion to rank 92 features derived from audio signals from a synthetic dataset. Only top 5 features were used in their experiment. Noisy and corrupted speech frame removal based feature selection was proposed in [14] on 2006 NIST SRE database. An Ant Colony Optimization (ACO) based feature selection for ASV system was proposed in [8] to optimize dimensionality of feature space on TIMIT dataset[1]. A review of ensemble for feature selection is provided in [1].

A combined approach of feature-feature and feature-class mutual information was proposed in [4] to find optimal features from a high dimensional feature set. Another study in [9] shows that minimal-redundancy-maximal-relevance criterion (mRMR) based feature selection yields promising scope on feature selection and classification accuracy over three classifiers and four datasets. A feature selection based approach using mutual information on incomplete data has been presented in [10], and found effective in terms of computational time and classification accuracy on incomplete data.

Some studies claim that high frequency components in a signal are better able to distinguish live speech from recorded speech. Inverted Mel-filtered cepstral coefficients (IMFCC) based features, which strongly emphasize high frequency regions of the signal have been found very successful in replay spoofing detection [3,6,13,15]. With this motivation, we investigate the best subset of filterbank features, and an alternative reduced representation for replay spoofing detection in this work. We explore entropy measures through decision tree feature selection. Further, we also evaluate the effectiveness of a variant of long-term spectral statistics (LTSS) based features, proposed in [7], and compare their performance with principal component analysis (PCA) based feature construction.

[1] https://catalog.ldc.upenn.edu/LDC93S1.

3 Audio Representation

An audio signal can be represented by several features which characterize the inherent details inside the speech signal for a particular task. Often, in case of this high dimensionality, it becomes difficult to manage or visualize the feature space. Moreover, separability among classes is also not clear, feature selection and dimensionality reduction plays key role in such cases.

3.1 Filterbank Based Features

We extracted four filterbank based features using Bob toolkit[2] as suggested in [11], and computed their derivatives to obtain the dynamic information. Figure 1 depicts these filterbank structures. A typical process of extracting mel frequency cepstral coefficients (MFCC) features is shown in Fig. 2. Rest of the features are computed using the same process by replacing triangular mel-filter bank with respective filterbank.

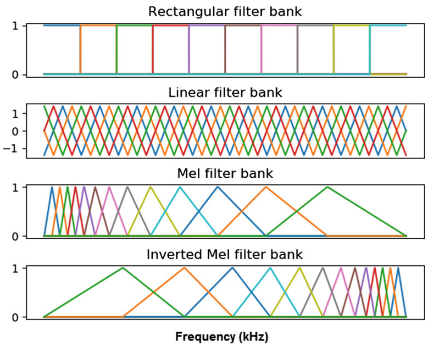

Fig. 1. Filterbanks used for computation [11]

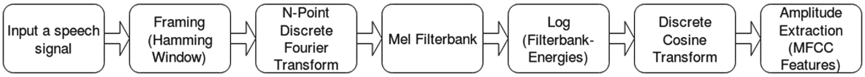
Input a speech signal
Framing (Hamming Window)
N-Point Discrete Fourier Transform
Mel Filterbank
Log (Filterbank-Energies)
Discrete Cosine Transform
Amplitude Extraction (MFCC Features)

Fig. 2. MFCC feature extraction

A feature vector is 156-dimensional vector consisting of all four features (each with 39 coefficients). A feature set is 39-dimensional vector consisting of all features of same kind. In nutshell, feature set is a subset of original feature vector. For example, an MFCC feature set comprises 39 coefficients, whereas the entire feature vector is comprised of 156 coefficients. This feature vector is made up of concatenation of coefficients from all four feature sets.

[2] https://www.idiap.ch/software/bob/.

3.2 Entropy Based Feature Selection

Decision trees were primarily meant for classification and regression purpose, but they have also got popularity for feature selection task. A decision tree is typically built by arranging the most representative features near the root of the tree, and discarding the least important features from the dataset. This helps in feature elimination. Decision trees use measures like entropy (information gain) or gini index to decide the feature containing the maximum information with respect to class label information. Information Gain (Entropy) represents the discriminative power of a specific attribute. The foundation of decision trees is based on choosing the most important and meaningful attributes or features in a supervised manner. In this work, we exploit this phenomenun to reduce the dimensionality of our original feature vector.

3.3 Feature Construction

Feature construction is carried out by transforming the existing feature space into new feature space which has less dimensions, yet most of the information is preserved.

PCA transforms data points from existing feature space to a lower dimensional feature space. These lower dimensional features, termed as principal components, are linear combination of original features. These components are obtained in such a way that the first principal component (PC) captures the direction of the maximum variance, and subsequent components get the next possible highest variance. Every principal component is orthogonal to all other PCs.

Long-term spectral statistics (LTSS) features are obtained by computing the mean and variance of frame level short-term features. These features are not calculated exactly using the methodology used in [7], but motivated from there.

4 Proposed Approach

In this work, we compare two main approaches for reducing features: (i) feature construction through dimensionality reduction i.e. feature construction using (a) principal component analysis, and (b) LTSS based features, and (ii) feature subset selection using entropy.

4.1 Approach

Figure 3 describes the scenario of experiments. In one scenario, a PCA based reduction was carried out and impact of principal components over accuracy was computed, whereas in other scenario, entropy based feature selection was exploited. In this scenario, entropy values of all feature sets were computed. All the coefficients with entropy above 0.1 were selected, and rest were discarded. Entropy values for different coefficients of IMFCC feature set is shown in Fig. 4.

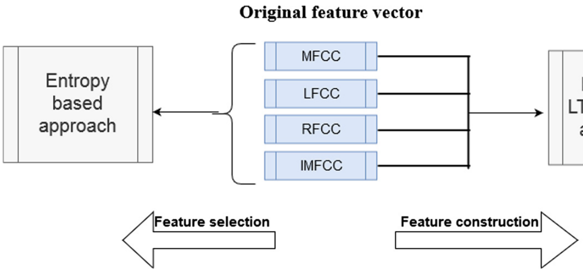

Fig. 3. Two scenarios for feature reduction used in this work.

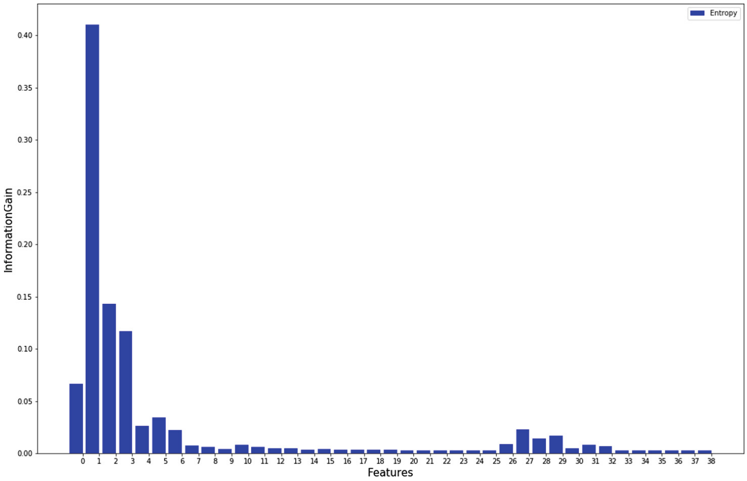

Fig. 4. Entropy values for IMFCC features

4.2 Implementation Scenario

In this section, we discuss scenario for implementation for our proposed work. Since the lower order coefficients contain more information pertaining to speaker-specific characteristics, we used the first 13 coefficients as the static features. After appending dynamic features, each audio file was represented by a feature of 39 coefficients.

A neural network classifier was trained with single hidden layer. Python Scikit-learn library was used for implementation. A Gaussian Mixture Model (GMM) with default parameters was also used for classification. A typical performance metric for speaker verification systems is equal error rate (EER). It is calculated based on threshold at which false acceptance rate (FAR) and false rejection rate (FRR) are equal. We have presented our results in terms of EER. We tested the proposed approach on ASVspoof 2017 version 2.0 corpus [5]. More details of the dataset can be found from [2].

Table 1. Results of four systems in terms of Equal Error Rate (EER) over features selected using entropy measure on two classifiers

Original feature set	No. of coefficients	Index	ANN	GMM
ALL156	156	—	0.4578	NA
BEST10	10	—	0.4567	NA
MFCC39	3	3,5,7	0.4391	0.4908
LFCC39	1	2	0.5156	0.5094
RFCC39	3	3,5,7	0.4445	0.51
IMFCC39	3	1,2,3	**0.397**	0.4926

4.3 Results and Discussion

Initially, a 156-dimensional feature vector was used to classify the dataset, and it achieved 0.4578 equal error rate (EER) on a multilayer perceptron classifier.

Table 2. EER performance of PCA based dimensionality reduction on Eval set

Feature	# of components	EER
MFCC	2	**0.4704**
LFCC	2	0.5502
RFCC	2	0.5180
IMFCC	2	0.5708
MFCC	3	0.5393
LFCC	3	0.5535
RFCC	3	0.5150
IMFCC	3	0.5660

Table 3. LTSS features performance (EER)

System	Using mean and variance	Mean only
IMFCC	0.5008	0.5433
MFCC	0.4986	0.4410
RFCC	0.4995	0.5
LFCC	0.5	0.4648

Table 1 lists the number of coefficients chosen from each feature set using entropy based feature selection. It is evident that all are static coefficients. A combination of ten best coefficients was used as a feature representation for audio which produced an EER value of 0.4567, stating that there is no significant improvement on performance if we combine the best coefficients from all four feature sets. As seen from Table 1, it is observed that using only three best coefficients of IMFCC feature yield considerable improvement of 13.2% on the performance (0.4578 to 0.397). GMM classifier also provided similar performance, hence we carried out further experiments with neural network. Using maximum mutual information (MMI) feature selection method, we obtained the same set of coefficients for each feature.

Table 2 summarizes the results on reduced data with two and three principal components. It can be seen that for 2-dimensional case, MFCCs show the best performance, whereas IMFCCs show the worst performance. This indicates that PCA is not able to capture high frequency characteristics of features while performing dimensionality reduction. Moreover, results with three PCs show either degradation or slight improvement in performance for all features. MFCC performance is highly affected due to addition of one more principal component. In no way, PCA based model is able to outperform the entropy based model.

We extracted a variant of long-term spectral statistics (LTSS) based features in following way. Mean and variance of all coefficients across all frames was computed to represent these long-term parameters. Thus, every audio file was represented by two dimensions for each feature. Table 3 shows the results of LTSS features on the evaluation set of the dataset. Each feature is represented by two coefficients i.e. mean and variance. It can be seen that in this case also, MFCCs give good results and they achieve the minimum EER. However, it is interesting to note that its performance is not better than the one shown in entropy based feature selection scenario. This shows that feature construction is related to human auditory perception and its representation (in lower dimension) goes alongwith auditory nerve behaviour. The best performance is given by a subset of static IMFCC features which is a relative improvement of roughly 10% compared to LTSS mean based classification.

Table 4. Performance of systems with combination of mean and variance coefficients.

# of coefficients	Features	EER	# of coefficients	Features	EER
4	MFCC+LFCC	0.4815	6	MFCC+LFCC+RFCC	0.4712
4	MFCC+RFCC	**0.4645**	6	MFCC+LFCC+IMFCC	0.4779
4	MFCC+IMFCC	0.4666	6	MFCC+RFCC+IMFCC	0.4728
4	LFCC+RFCC	0.476	6	LFCC+RFCC+IMFCC	0.4958
4	LFCC+IMFCC	0.5436	8	MFCC+LFCC+RFCC+IMFCC	0.5
4	RFCC+IMFCC	0.5047			

Results also show that RFCC features are not sensitive to standard deviation or variance coefficients as there is negligible change in their performance with or without variance. However, other three features show some sensitivity towards these coefficients. The equal error rate drops by almost 7.82% (0.5433 to 0.5008) when variance information is appended for IMFCCs. In contrast, MFCC and LFCC show good results in absence of variance information.

We tried different combination of these features to examine its effect on overall performance. From the results shown in Table 4, we can observe that minimum EER is achieved with combination of mean and variance of MFCC and RFCC features, yet they cannot outperform the standalone MFCC feature performance on the same case.

5 Conclusion and Future Work

We presented a comparative analysis of filterbank based short-term spectral features using feature selection and feature construction approach. A comparison was given among entropy based feature selection, and PCA and LTSS based feature construction. In this paper, we attempted to explore the possibility of finding best representative feature reduction strategies for the task of voice playback spoofing detection. Results indicate that entropy based feature selection gives promising results. Further, it can be seen that IMFCC features capture the replay attack information more significantly than any other feature. This shows that human perception (MFCCs) may not help much for detection of playback spoofing attacks, and this may also fail while performing subjective evaluation of such systems.

In future, recent techniques for feature visualization can be adopted to see class distribution among input data points. This information may be helpful for designing robust feature selection strategies.

References

1. Boln-Canedo, V., Alonso-Betanzos, A.: Ensembles for feature selection: a review and future trends. Inf. Fusion **52**, 1–12 (2019)
2. Delgado, H., et al.: ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. In: Odyssey (2018)
3. Hanilci, C.: Features and classifiers for replay spoofing attack detection. In: 10th International Conference on Electrical and Electronics Engineering (ELECO), pp. 1187–1191, November 2017
4. Hoque, N., Bhattacharyya, D., Kalita, J.: MIFS-ND: a mutual information-based feature selection method. Expert Syst. Appl. **41**(14), 6371–6385 (2014)
5. Kinnunen, T., et al.: The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection (2017)
6. Mankad, S.H., Shah, V., Garg, S.: Towards development of smart and reliable voice based personal assistants. In: TENCON 2018, pp. 2473–2478 (2018)
7. 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. **25**(11), 2098–2111 (2017)
8. Nemati, S., Basiri, M.E.: Text-independent speaker verification using ant colony optimization-based selected features. Expert Syst. Appl. **38**(1), 620–630 (2011)
9. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
10. Qian, W., Shu, W.: Mutual information criterion for feature selection from incomplete data. Neurocomputing **168**, 210–220 (2015)
11. Sahidullah, M., Kinnunen, T., Hanilçi, C.: A comparison of features for synthetic speech detection. In: Interspeech (2015)
12. Sambur, M.R.: Selection of acoustic features for speaker identification. IEEE Trans. Acoust. Speech Sig. Process. **23**(2), 176–182 (1975)
13. Sriskandaraja, K., Suthokumar, G., Sethu, V., Ambikairajah, E.: Investigating the use of scattering coefficients for replay attack detection. In: APSIPA, pp. 1195–1198, December 2017

14. Sun, H., Ma, B., Li, H.: An efficient feature selection method for speaker recognition. In: 2008 6th International Symposium on Chinese Spoken Language Processing, pp. 1–4, December 2008
15. Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., Galka, J.: Audio replay attack detection using high-frequency features. In: Interspeech (2017)