Journal of Engineering and Applied Sciences 13 (12): 4512-4517, 2018
ISSN: 1816-949X

Cross Validation of Machine Learning Classifiers and Features for Audio Forensics Verification

Jhon Kevin Segura, Diego Renza and L. Dora M. Ballesteros
Universidad Militar Nueva Granada, Carrera 11 No. 101-80, Bogota, Colombia

Abstract: In literature, there are several manuscripts related to finding the best feature or the best classifier for audio verification systems. However, cross validation with both criteria has not been widely discussed. In this research, 15 classifiers and six features have been selected to obtain ninety options for audio forensics verification. The aim is to provide suggested combinations for forensics researches. The evaluated classifiers are based on decision trees, discriminant analysis, support vector machines, nearest neighbour and hybrid classifiers. The feature extraction is based on Mel-Frequency Cepstral Coefficients (MFCC) and cochleagrams, using principal component analysis optionally. The tests are performed on a database of 50 speakers and 10 utterances per speaker and the assessment of classifiers is made by means of accuracy. According to the results, the best combination is MFCC with linear discrimination, followed very close by MFCC+PCA with linear discriminant.

Key words: Speaker recognition, Mel-Frequencies Cepstral Coefficients (MFCC), cochleagram, feature selection, classiffier evaluation, assessment

INTRODUCTION

The main goal in audio forensics consists in applying digital signal processing procedures to acquire, preserve, analyse and evaluate audio data that could be presented as potential digital evidence in a legal proceeding (Maher, 2009, 2010). According to ISO/IEC 27037:2012 the digital evidence corresponds to "information or data, stored or transmitted in binary form that may be relied on as evidence". Moreover, the sources of the potential digital evidence can be different types of digital devices, networks, databases, etc. (Anonymous, 2012).

Among the tasks to perform in audio forensics, three approaches can be highlighted: authenticity, enhancement and interpreting (Maher, 2009). The authenticity is important because of the fragility of digital evidence, for instance, audio recordings can be easily edited by means of applications and tools widely available. Therefore, authenticity practices are oriented to ensure that the potential evidence is complete and unaltered from the time of acquisition until its final disposition (Maher, 2009; Anonymous, 2012). Audio enhancement is used to get a better audibility and to improve the intelligibility of the contents, here, non-destructive signal processing techniques are required to avoid degradations and loss of relevant content (Maher, 2010). Finally, interpreting relates to discovering the significance of the potential evidence in the official investigation, it includes aspects such as speaker recognition, speech transcription, etc. (Maher, 2010).

Speaker recognition can be divided into 2 main areas: speaker verification and speaker identification. The fundamental task in verification is to compare two speech utterances and determine if they correspond to the same person in a general sense, the system verifies if a person is the one he/she claims to be (e.g., biometric authentication). In speaker identification there is a set of known speakers (suspects) and an unknown speaker (from the digital evidence), the task is to find out which of the suspects sounds closest to the unknown speaker (Hansen and Hasan, 2015).

Regarding speaker recognition in audio forensics, it can be accomplished in a human-based approach or in a computer-based approach. Human-based approach can be accomplished by qualified listeners that compare systematically the speech samples of the unknown speaker and the known speakers giving a report about the similarities between them. On the other hand, the analysis and decision in a computer-based approach are completely done by means of computation analysis (Hansen and Hasan, 2015). In any case, the forensic experts act as witnesses showing the strength of their findings by means of for instance, the likelihood ratio.

To perform the comparison between the speech samples, speaker idiosyncratic characteristics should be extracted. Ideally, these feature parameters should show

Corresponding Author: Jhon Kevin Segura, Universidad Militar Nueva Granada, Carrera 11 No. 101-80, Bogota, Colombia

high inter-speaker variability and low intra-speaker variability, must be resilient to imitation have a high frequency of occurrence and their extraction must be easy. In a broad sense, the feature parameters can be classified into auditory versus acoustic, linguistic versus non-linguistic and short-term versus long-term features (Hansen and Hasan, 2015). In computer based-approaches, acoustic, non-linguistic and short-term features are generally used.

According to a recent interpol survey, the used approaches in speaker identification by law enforcement agencies are as follows: auditory by forensic practitioners, spectrographic or auditory-spectrographic by forensic practitioners, auditory-acoustic-phonetic by forensic practitioners, acoustic-phonetic by forensic practitioners, human-supervised automatic approaches by forensic practitioners and fully automatic approaches by non-forensic practitioners. From the above list, the two most used methods are the auditory-acoustic-phonetic and auditory-spectrographic ones (Morrison *et al*., 2016). The latter approaches mostly use short-term features, mainly extracted from the speech spectrum (Hansen and Hasan, 2015). It should be noted that most of the schemes require human intervention which may have a subjective component and depend strongly on the experience of the forensic practitioner. Moreover, considering multiple feature parameters, it is critical, since, 2 different speakers may have the same $feature_a$ but a different $feature_b$.

Regarding spectrographic analysis, there are several options to analyse and compare voice signals (Tirumala *et al*., 2017): Mel-Frequency Cepstral Coefficients (MFCC), Mean Hilbert Envelope Coefficients (MHEC), Frequency Domain Linear Prediction (FDLP), Power-Normalized Cepstral Coefficients (PNCCs), spectrograms and cochleagrams (gammatone filters). Among the short-term acoustic features, the MFCC approach is the most popular (Hansen and Hasan, 2015); in MFCC, the speech signal is represented as a sequence of cepstral vectors, given by the application of Short-Fourier-Transform (STFT), passing its magnitude through Mel-filters and decorrelating with Discrete Cosine Transform (DCT) (Morrison *et al*., 2016). On the other hand, spectrograms and cochleagrams are time-frequency representations of voice signals as a sequence of spectral vectors. Their difference lies in the frequency scale representation, for spectrograms, the scale is linear while it is logarithmic for cochleagrams (Larrota *et al*., 2017; Camacho and Renza, 2017). As the human ear concentrates on only certain frequency components, for speech analysis it is more interesting to have more components in the low frequency than in the high frequency regions, whereby a logarithmic scale is more suitable.

Alternatively, in the human-supervised automatic approaches, the core of the system (classifier) can include several methods of machine learning such as linear prediction, neural networks and support vector machines, among others.

In literature, there are many researches focused on finding the "best feature" or the "best classifier" but it is not common to find a research that combines both types of decisions. According to the above, the objective of this study is to make a c omparison of several classifiers (i.e., fifteen choices) and six different features. The selected classifiers are based on decision trees, discriminant analysis, Support Vector Machines (SVM), nearest neighbour and hybrid classifiers. The feature extraction is based on Mel-Frequency Cepstral Coefficients (MFCC) and cochleagrams or a combination of them in any case, it is feasible to reduce the dimensions of the data through Principal Component Analysis (PCA). With the results, a researcher in the field of audio forensics can use the obtained data to choose an adequate combination of feature/classifier at the time of developing a text-dependent audio forensic verification system.

MATERIALS AND METHODS

MFCC: The MFCC (Mel-Frequency Cepstral Coefficients) is a method to obtain a parametric representation of a signal which has been widely used in speech recognition tasks. To obtain this representation, a 1st order FIR (Finite Impulse Response) filter is used as pre-emphasis filter to flatten the spectrum of the signal. Then, the Short-Time Fourier Transform (STFT) is applied to the signal keeping an overlapping time between the frames. Its magnitude spectrum is passed through a filter-bank comprising M triangular filters where their linear frequency is mapped to mel-frequency through (Eq. 1) (Han *et al*., 2006):

$$\text{Mel}(f) = 2595\log_{10}(1+f/700) \quad (1)$$

In the mel domain, filters are equally spaced in a frequency band (Han *et al*., 2006). After signal filtering, the Discrete Cosine Transform (DCT) is applied to the logarithm of the previous data (Morrison *et al*., 2016) in order to de-correlate them. Finally, a lifter is applied to obtain the MFCC coefficients.

Cochleagram: Cochleagrams and spectrograms are time versus frequency representations of speech signals. In the case of cochleagram, it generates a 2D representation of the spectral characteristics of the audio signal but using a logarithmic frequency scale. To obtain the cochleagram, the audio signal is passed through n-order

band-pass filters, known as gammatone filters. Also, the response of each filter is analysed in small time windows, offering a time-frequency representation (Han *et al*., 2015). With a j-order gammatone filter centred at frequency fc an amplitude A, a bandwidth B and a phase φ, the impulse response is given by Eq. 2 (Sharan and Moir, 2015):

$$g(t) = At^{j-1}e^{-2\pi Bt}\cos(2\pi ft+\varphi) \quad (2)$$

To measure the bandwidth of each cochlea-filter a psycho-acoustic measure of the auditory filter width at each point along the cochlea can be used Eq. 3 (Sharan and Moir, 2015):

$$f_{c,ERB} = \left[\left(F_{c,HZ}/Q_{ear}\right)^p + \left(B_{min}\right)^p\right]^{1/p} \quad (3)$$

From $f_{c,ERB}$ an approximation of the filter bandwidth can be obtained through Eq. 4:

$$B = 1.019 \times f_{c,ERB} \quad (4)$$

It should be noted that a cochleagram extracts the voice signals characteristics giving a graphic representation of the signal, i.e., a 2D representation (time versus frequency). In this graphic, the intensity of each pixel represents the energy of each band in each time and its frequency axis is not linear (Hz), showing a higher frequency resolution in low frequencies than in high frequencies (Shao and Wang, 2008; Shao *et al*., 2009).

Proposed method: The purpose of this phase is to validate the scheme in terms of accuracy for the forensic audio verification process. To evaluate the performance of the proposed system, a database of 50 speakers and 10 utterances per speaker was used. To compare the results of the proposed system, three feature extraction methods (plus PCA) and fifteen classifiers were used as shown in Fig. 1. The PCA parameters were adjusted in such a way that only the components that explain 95% of the variance were kept.

Feature selection: To characterise the input data, two main feature extraction methods were used: MFCC and Cochleagram. A third method consists in joining the MFCC features and the cochleagram features. For each case, the reduction of dimensions using PCA was also evaluated. The 6 feature extraction methods are listed below:

Feature type 1: MFCC of the input utterance. For 0.5 sec signals, this gives a 13×48 matrix per each utterance, this matrix is converted to a row vector being an observation for training or testing.

Feature type 2: MFCC with PCA. This method is similar to feature type 1 with the difference that PCA is applied to a row vector to reduce its dimensions.

Feature type 3: Cochleagram of the input utterance. For 0.5 sec signals, this gives a 128×4000 matrix per each utterance, this matrix is converted to a row vector, being an observation for training or testing.

Feature type 4: Cochleagram with PCA. This method is similar to feature type 3, with the difference that PCA is applied to the row vector to reduce its dimensions.

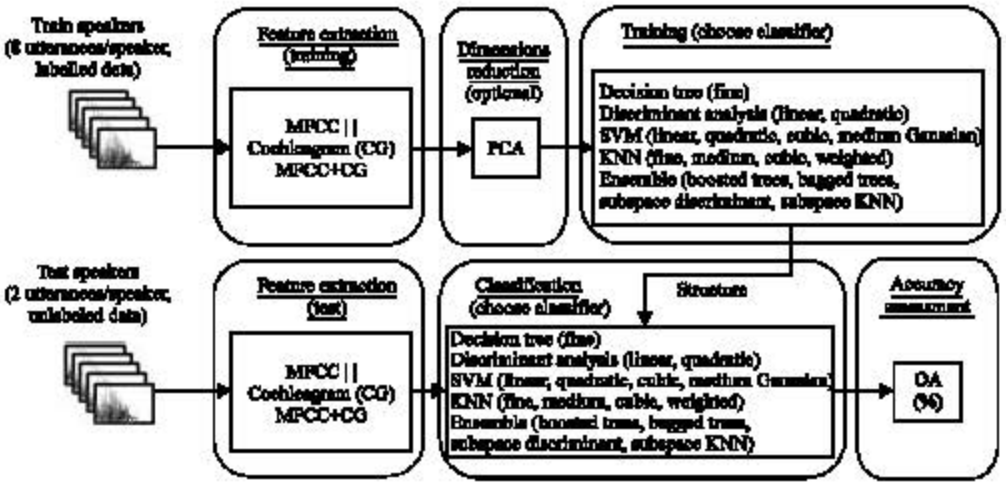

Fig. 1: Study methodology; Full outline

Table 1: Selected classifiers

Classifier	Description
Decision tree	Its response is based on successive decisions in a tree structure from the root node down to the respective leaf node. The selected type is fine tree (a tree where many leaves are used to make many fine distinctions between classes)
Discriminant analysis	The separation is based on the parameters of a Gaussian distribution for each class. The types tested in this research are linear discriminant and quadratic discriminant
SVM	The separation of classes is based on finding the best hyperplane that separates data in one class from data in other class. The types tested in this research are linear, quadratic, cubic and medium Gaussian
KNN	The separation of classes is based on finding the closest points to a particular class in the training dataset. The types tested in this research are fine KNN (k = 1), medium KNN (k = 10), cubic KNN (k = 10) and weighted KNN (k = 10)
Hybrid classifiers	These classifiers combine different methods in order to enhance the results. The types tested in this work are boosted trees (AdaBoost +decision tree), bagged trees (Random forest+Decision tree), subspace discriminant (Subspace+discriminant learners) and subspace KNN (Subspace+Nearest neighbors)

Feature type 5: MFCC and cochleagram. This method concatenates the vectors of features type 1 and 3.

Feature type 6: MFCC and cochleagram with PCA. This method concatenates the vectors of features type 2 and 4, it differs from the previous methods since PCA is applied to reduce the dimension of data.

Classifier selection: After feature extraction in each of the methods, every speaker gives 10 feature vectors. The 80% of these vectors are used to train a classifier whereas the other 20% are used to test the scheme. Fifteen classifiers were evaluated as shown in Table 1.

Evaluation: For the training process, 80% of the database (8 utterances/speaker, 400 utterances in total) is used for training and cross-validation of the classifier. To evaluate the accuracy of the classifier, 20% of the Fig. 1 database (2 utterances/speaker, 100 utterances in total) is used. The Accuracy (ACC) is calculated through Eq. 5:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Where:
TP = True Positives
TN = Depicts true Negatives
FP = False Positives
FN = Corresponds to False Negatives

RESULTS AND DISCUSSION

As discussed so far, the precision of the results was evaluated by varying both the feature extraction method and the classifier type. According to this, the results are shown from three points of view. In the first place, the results of each type of characteristic are shown considering all the classifiers. Secondly, the results of each classifier are shown considering all types of characteristics. Finally, the performance of all the evaluated cases will be shown.

Figure 2 shows the cumulative accuracy per feature type. The total accuracy in each feature is obtained through the sum of the accuracy in each classifier. The objective of this analysis is to determine the best type of characteristic, regardless of the classifier. According to the results shown in this Fig. 2, the best method for the extraction of characteristics is MFCC followed by MFCC with PCA and MFCC plus cochleagram. In the case of MFCC, the best performance is observed with linear discriminant (92%) and subspace KNN (91%), the worst case is using a subspace discriminant classifier.

Figure 2 shows the cumulative accuracy per classifier type. The total accuracy in each classifier is obtained through the sum of the accuracy in each feature type. The objective of this analysis is to determine the best classifier, regardless of the feature type. According to the results shown in this Fig. 2, the best classifier method is subspace KNN followed by linear discriminant. For these two cases, the method that contributes the most to global accumulation is MFCC followed by MFCC with PCA.

Figure 3 shows the results of each of the ninety cases evaluated, grouped by classifier type. In this case, the good behaviour of the KNN subspace and linear discriminant classifiers is evident. As particular cases, the best performances (accuracy >90%) were obtained in the following order:

- MFCC with linear discriminant (92%)
- MFCC (PCA) with linear discriminant (91%)
- MFCC plus cochleagram (PCA) with linear discriminant (91%)
- MFCC with subspace KNN (91%)
- MFCC (PCA) with subspace discriminant (91%)

From the previous results, the following aspects can be inferred:

- The feature extraction method based on MFCC presents the best results
- The use of PCA in some classifiers presents similar results which can mean an advantage in terms of computational cost
- The linear discriminant classifier presented results above 90% in three of the six cases evaluated which together with its low computational cost can be key in the selection of this classifier

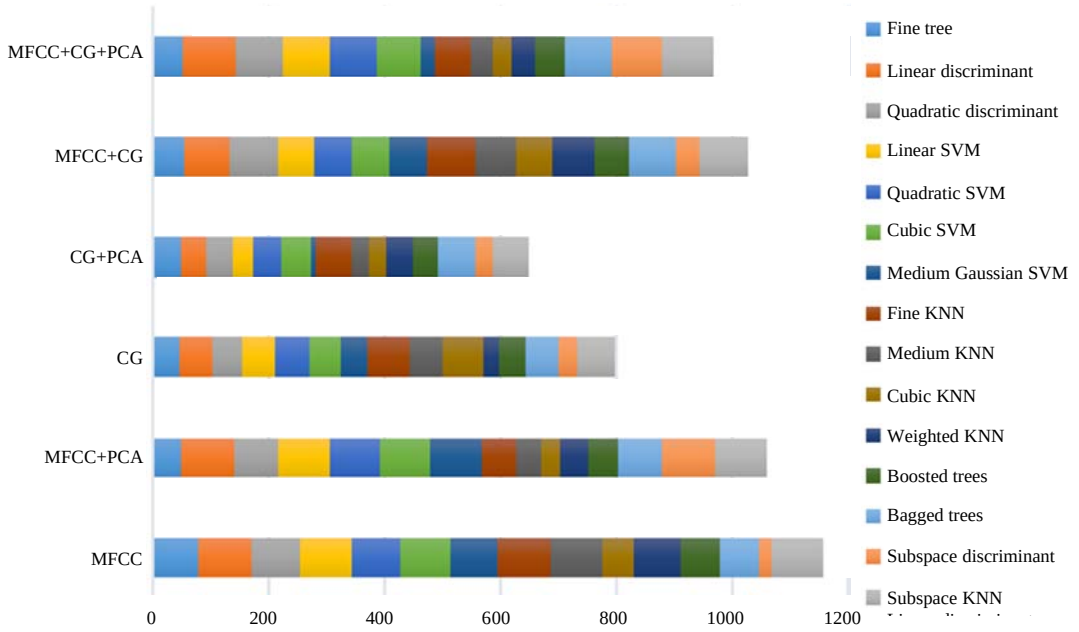

Fig. 2: Cumulative accuracy per feature type

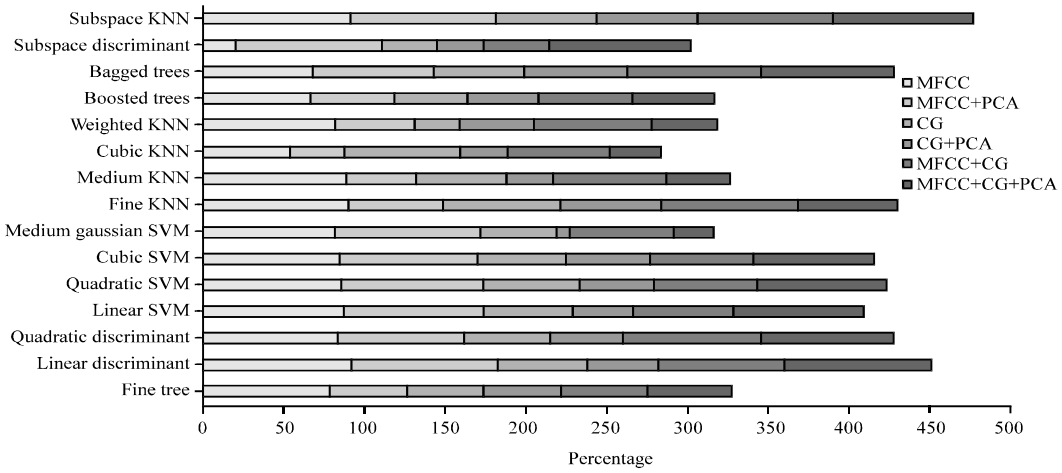

Fig. 3: Cumulative accuracy per classifier type

CONCLUSION

In this research, ninety cases of machine learning systems for audio forensic verification were evaluated (using six features and fifteen classifiers). The selected criteria to compare the performance of every choice was the overall accuracy, measured through the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The methodology was focused on text-dependent verification in which 10 utterances of the same word were selected for each speaker, 8 of them were used for training the machine learning system and the others were used for the testing phase. In total, 50 speakers were taken into account. At the end, our experimental phase evaluated 450 results, obtained from 50 speakers by 90 choices. The results were grouped by feature and by classifier. In the first case, MFCC provided a cumulative accuracy of 1180/1500 in the latter, subspace KNN gave a cumulative accuracy of 480/600. The best combination was MFCC+Linear discriminant followed very closely by MFCC+PCA+Linear discriminant.

ACKNOWLEDGEMENT

This research is supported by the\Universidad Militar Nueva Granada-Vicerrectora de Investigaciones under the grant IMP-ING-2136 of 2016.

REFERENCES

Anonymous, 2012. ISO/IEC 27037: Information technology-security techniques-guidelines for identification, collection, acquisition and preservation of digital evidence. International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/44381.html.

Camacho, S. and D. Renza, 2017. A semi-supervised speaker identification method for audio forensics using cochleagrams. Proceedings of the 4th Workshop on Engineering Applications (WEA'17), September 27-29, 2017, Springer, Cartagena, Colombia, ISBN:978-3-319-66962-5, pp: 55-64.

Han, K., Y. Wang, D. Wang, W.S. Woods and I. Merks *et al*., 2015. Learning spectral mapping for speech dereverberation and denoising. IEEE. Trans. Audio Speech Lang. Proc., 23: 982-992.

Han, W., C.F. Chan, C.S. Choy and K.P. Pun, 2006. An efficient MFCC extraction method in speech recognition. Proceedings of the 2006 IEEE International Symposium on Circuits and Systems (ISCAS'06), May 21-24, 2006, IEEE, Island of Kos, Greece, pp: 1-4.

Hansen, J.H. and T. Hasan, 2015. Speaker recognition by machines and humans: A tutorial review. IEEE. Sig. Process. Mag., 32: 74-99.

Larrota, D.M.B., D.R. Torres and S.A.C. Vargas, 2017. Blind speaker identification for audio forensic purposes. Dyna., 84: 259-266.

Maher, R.C., 2009. Audio forensic examination. IEEE. Sig. Process. Mag., 26: 84-94.

Maher, R.C., 2010. Overview of Audio Forensics. In: Intelligent Multimedia Analysis for Security Applications, Sencar, H.T., S. Velastin, N. Nikolaidis and S. Lian (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-11754-1, pp: 127-144.

Morrison, G.S., F.H. Sahito, G. Jardine, D. Djokic and S. Clavet *et al*., 2016. Interpol survey of the use of speaker identification by law enforcement agencies. Forensic Sci. Intl., 263: 92-100.

Shao, Y. and D. Wang, 2008. Robust speaker identification using auditory features and computational auditory scene analysis. Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), March 31-April 4, 2008, IEEE, Las Vegas, Nevada, ISBN:978-1-4244-1483-3, pp: 1589-1592.

Shao, Y., Z. Jin, D. Wang and S. Srinivasan, 2009. An auditory-based feature for robust speech recognition. Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09), April 19-24, 2009, IEEE, Taipei, Taiwan, pp: 4625-4628.

Sharan, R.V. and T.J. Moir, 2015. Cochleagram image feature for improved robustness in sound recognition. Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP'15), July 21-24, 2015, IEEE, Singapore, ISBN:978-1-4799-8057-4, pp: 441-444.

Tirumala, S.S., S.R. Shahamiri, A.S. Garhwal and R. Wang, 2017. Speaker identification features extraction methods: A systematic review. Expert Syst. Appl., 90: 250-271.