UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Blind speaker identification for audio forensic purposes

Dora María Ballesteros-Larrota, Diego Renza-Torres & Steven Andrés Camacho-Vargas

Programa Ingeniería en Telecomunicaciones, Universidad Militar Nueva Granada, Bogotá, Colombia. dora.ballesteros@unimilitar.edu.co, diego.renza@unimilitar.edu.co, u1400943@unimilitar.edu.co

Received: October 3rd, 2016. Received in revised form: May 23th, 2017. Accepted: May 31th, 2017.

Abstract
This paper presents a blind method for speaker identification for audio forensics purposes. It is based on a decision system with fuzzy rules and works with the correlation between the cochleagrams of the audio proof and of the audios of the suspects. Our proposed system can give a null output, a unique selected suspect or a group of identified suspects. According to several tests, our Overall Accuracy (OA) is 0.97 with agreement (κappa index, κ) of 0.75. Additionally, unlike typical systems in which a low false acceptance (FP) implies high false rejection (FN), our system can work simultaneously with FN and FP equal to zero (i.e. OA=1; κ=1). Finally, our system works with blind identification, it means, without preliminary knowledge of the audio recordings or a training step; an imperative characteristic for audio forensics.

Keywords: Speaker identification; cochleagram; fuzzy logic; true acceptance; false acceptance.

Identificación del hablante de forma ciega para fines de audio forense

Resumen
Este artículo presenta un método ciego para identificación del hablante, con fines de audio forense. Se basa en un sistema de decisión que trabajo con reglas difusas y la correlación entre los cocleagramas del audio de prueba y de los audios de los sospechosos. Nuestro sistema proporciona salida nula, con único sospechoso o con un grupo de sospechosos. De acuerdo a las pruebas realizadas, el desempeño global del sistema (OA) es 0.97 con un valor de concordancia (índice kappa) de 0.75. Adicionalmente, a diferencia de sistemas clásicos en los que un bajo valor de selección incorrecta (FP) implica un alto valor de rechazo incorrecto (FN), nuestro sistema puede trabajar con valores de FP y FN igual a cero, de forma simultánea. Finalmente, nuestro sistema trabaja con identificación ciega, es decir, no es necesaria una fase de entrenamiento o conocimiento previo de los audios; característica importante para audio forense.

Palabras clave: Identificación del hablante; cocleagrama; lógica difusa; aceptación correcta; aceptación incorrecta.

1. Introduction

In audio forensics there are two main scenarios for treating a suspect: identification and verification of the speaker [1]. In the first scenario, there is a specific group of suspects, in which the perpetrator of the crime is; the proof given as evidence is a voice recording. To identify the perpetrator of the crime, the voices of the members of the group are compared against the audio proof, and the voice that matches the evidence is the villain's voice. In the second scenario, there is only one suspect and the purpose is to determine whether s/he participated in an audio recording given as evidence.

Speaker identification algorithms can be used for authentication purposes or for audio forensics. In the first case, the system is trained with some utterances of every speaker and then the system should identify the speaker with a new set of words. However, in the second case, the identification process is a blind task, and there is not preliminary knowledge about the owner of the audio recordings.

Generally, one of the main blocks in any speaker identification system is a decision algorithm, which makes a choice from the information obtained from the voices (feature extraction). Among decision techniques, the

How to cite: Ballesteros-Larrota, D. M., Renza-Torres, D., and Camacho-Vargas, E. A., Blind speaker identification for audio forensic purposes DYNA 84(201), pp. 259-266, 2017.

artificial intelligence schemes such as neural networks [1], fuzzy logic [2,3] or genetic algorithms [4] have been widely used. Since these kind of solutions need a training stage, they are very useful for authentication but not for audio forensics.

In the context of audio forensics, one of the main decision algorithms for speaker identification used by law enforcement agencies is the analysis based on voice spectrogram, according to official information reported by INTERPOL [5]. This technique is the second most used approach around the world, and the first in Asia, Africa, the Middle East, South and Central America. In the case of North America, the most frequent approaches are the semi-supervised algorithms which use signal processing methods and statistical models; here, the algorithms can use voice spectrogram but unlike visual comparison of formant shapes, human inspection is replaced by automatic analysis.

In the literature, another approach suggests carrying out voice analysis by means of cochleagrams instead of spectrograms. Although spectrograms give useful information to analyze a voice signal [6] and to compare two or more recordings [7,8], and it has been verified that impostors cannot mimic the behavior of target spectrograms [9], they work with lineal frequency resolution and are not the best way to analyze low frequencies of the signal. On the other hand, cochleagrams work with a finer frequency resolution at low frequencies, which may be useful for differentiating voice characteristics, since its energy is high at low frequencies; therefore, some recent studies of audio forensics have used cochleagrams as a feature input in speaker identification systems [10-12]

According to the above, this paper proposes a blind speaker identification algorithm that uses cochleagrams for feature analysis, and a fuzzy system for classification. With our proposal, a training step is not required before the identification task, and therefore, it is intended for audio forensics purposes.

The other sections of the paper are structured as follows: Section 2 explains the auditory features used in our proposal. Section 3 presents the proposed method. Section 4 shows the implementation and validation of the method. Section 5 presents the conclusions.

2. Auditory features

As discussed previously, in traditional approaches of audio forensics the spectrogram has been widely used as a tool for auditory feature extraction. A spectrogram is a graphic representation of a one-dimensional signal into a two-dimensional time-frequency display. In the case of the spectrogram of a voice signal, a linear frequency scale from 0 Hz to 8 KHz (or 4 KHz depending of the frequency sampling of the voice signal) is used. This linear characteristic is not adequate for speaker identification purposes, since there are little differences in the voice features for people with similar timbre and pitch. Therefore, these differences could not be detected through methodologies using spectrograms. On the other hand, the cochleagram is a two-dimensional representation of sound signals, but unlike the spectrogram, it uses a bank of gammatone filters. The result has a finer frequency resolution at low frequencies respect to high frequencies and its main consequence is a better contrast around the features [13].

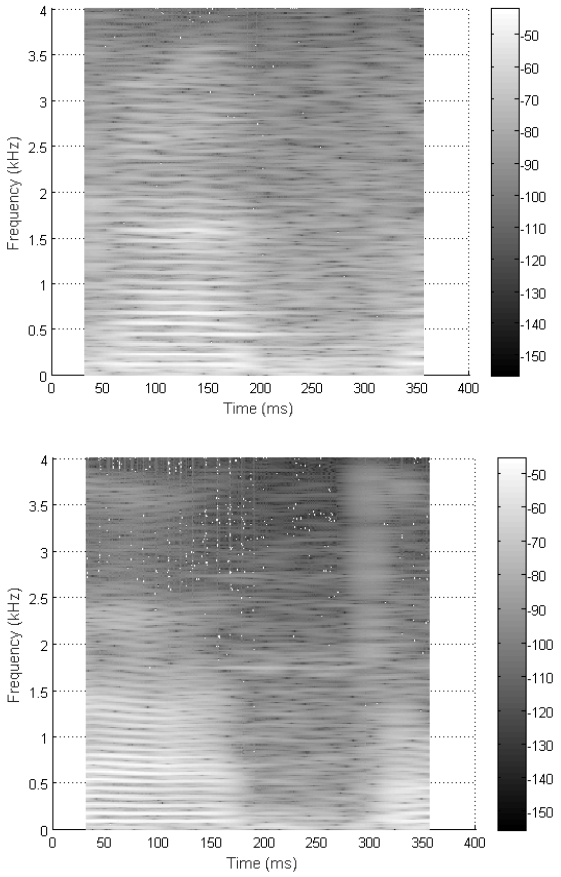

Figure 1. Spectrograms of the word "gato". Speaker$_1$ and speaker$_2$. Source: The authors

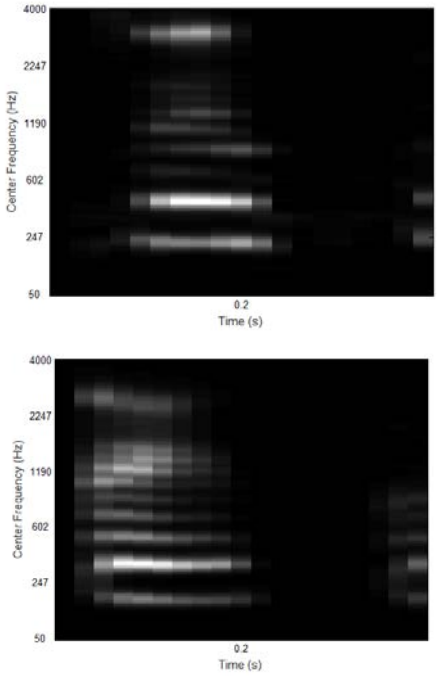

Figure 2. Cochleagrams of the word "gato". Speaker$_1$ and speaker$_2$. Source: The authors

In order to illustrate the representation of a voice signal using spectrograms and cochleagrams, Figs. 1, 2 show an example of a voice signal with the pronunciation of the word "gato" (Spanish for "cat") for two different speakers. Here, white represents high values of energy and black represents low values (or mute).

According to the above graphics, dissimilarity between the plots of the same figure is more noticeable in the case of Fig. 2. In that instance, there are (visually) remarkable differences in the region between 602 Hz and 1190 Hz.

In most cases, these differences can be enough to reject a suspect as the source of an audio proof. However, to enhance the performance of the system, the comparison can be done quantitatively through mathematical operations. Here, the aim is to obtain a value of similarity/dissimilarity that can be used by a classification block. One way to make the comparison is through Normalized Correlation (*NC*) between two cochleagrams. If the *NC* value is close to 1, it implies that the behavior (in time-frequency) is highly correlated between them, and therefore it is highly probable that the suspect is the source of the audio proof. Otherwise, if the *NC* value is close to 0, it means that the suspect is not. The challenge of deciding a positive coincidence relies on cases of middle values of *NC*. According to several tests, in some cases of the same speaker and in some cases of different speakers, middle values of *NC* are achieved. Therefore, datasets of high similarity and low similarity are not exclusive, meaning that some region of *NC* values simultaneously belong to high similarity and low similarity. This is the reason for selecting fuzzy logic within the analysis of positive or negative coincidence between the suspect and the audio proof.

3. Proposed speaker identification method

Our proposal can be discriminated in three main parts: feature analysis of speech recordings based on cochleagrams, fuzzy system, and selection. Fig. 3 shows a general outline of the proposed speaker identification method. Each of these three blocks is described in detail below.

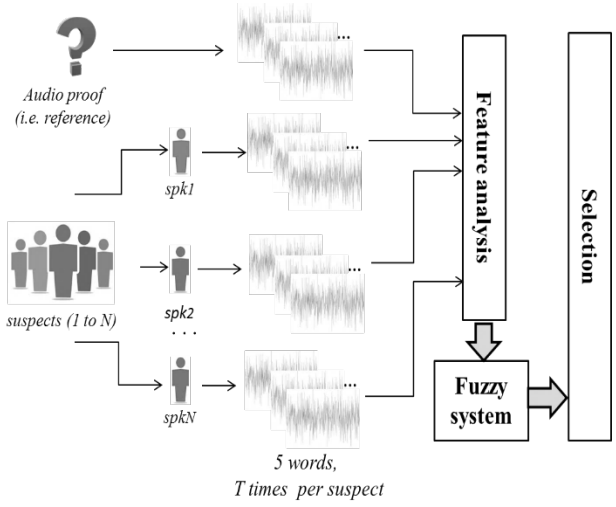

Figure 3. Block diagram of the proposed method.
Source: The authors.

3.1. Feature Analysis of speech recordings based on cochleagrams

For audio forensics purposes, short utterances are selected for speaker identification in our proposal. Forensics extract five words of the audio proof (*ref*) and ask the specific group of suspects (i.e. spk_1 to spk_N) to pronounce the selected words (i.e. w_1 to w_M). These signals are the input of the block Feature Analysis (see Fig. 4).

Being *N* the number of suspects, it will be (*N+1*)*5 voice recordings corresponding to the five words pronounced by each suspect and the recordings of the five words extracted from the audio proof. Then, the cochleagram of each recording is calculated, and stored as *cref_wi* for the case of the recordings extracted from the audio proof, or $cspk_j\_w_i$, for the case of voice recordings pronounced by each suspect (where *i* is the number of the word, *j* is the number of the suspect).

Then, the NC between the cochleagrams of the recordings extracted from the audio proof and each voice recording pronounced by each suspect (NCj,i) is computed. According to the above, there will be five values of NC per suspect (i.e.NCj,1 to NCj,5). These values are the inputs of the fuzzy system.

With the same example of Fig. 2, the visual similarity between the cochleagrams is very low, which is confirmed mathematically by an NC equal to 0.0376.

However, full certainty is not guaranteed regarding about the individual pointed by the evidence with a unique comparison. For this reason, the comparison of five different words with subsequent selection using a fuzzy system is proposed.

3.2. Fuzzy system

The aim of this system is to determine the degree of match between the recordings of each suspect and the recordings extracted from the audio proof. The selection of fuzzy system

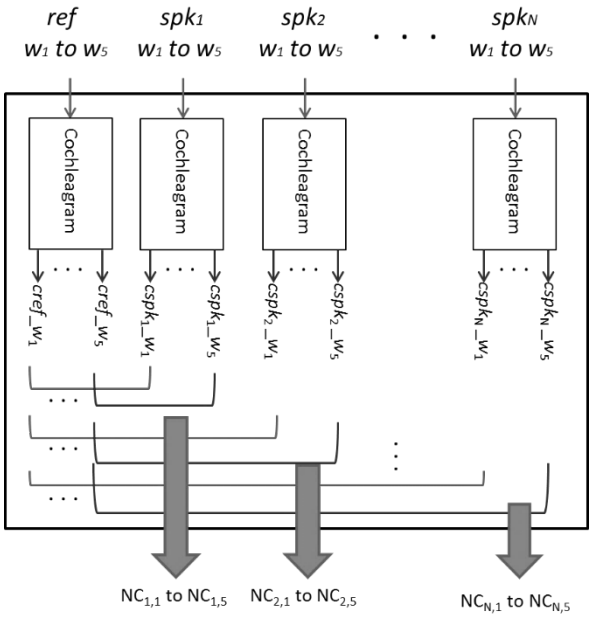

Figure 4. Internal process of the "Feature Analysis" block.
Source: The authors.

obeys the following reason: the *NC* between two recordings of the same word and the source is expected to be close to 1, and for the case of different individuals is expected to be close to 0; however, in some cases, people with similar voice patterns (e.g. timbre, pitch) can have high values of *NC*, and in other cases, recordings of the same word from the same speaker cannot have a high value of *NC* (e.g. if the quality of the recordings is not good). Hence, it is necessary to classify the degree of matching in two fuzzy groups (high-similarity and low-similarity), if they are overlapped in a region of *NC*. This means that a given value of *NC* will have a certain degree of membership to the low-similarity set and a certain degree of membership to the high-similarity set. For example, an *NC* value of 0.4 has a degree of membership to the low-similarity set of 0.8 and a degree of membership to the high-similarity of 0.2. Therefore, these coincidence values suggest the probability of belonging to the high- similarity set.

The fuzzy system block is divided in three parts: fuzzy input, fuzzy operator and implication operator as showed in Fig. 5. The Inputs of the fuzzy system are the scalar values (*NC*) of the output of the previous block (i.e. $NC_{j,1}$ to $NC_{j,5}$ for every j[th] suspect). The output of this block is 1 or 0. Below, the three parts of the fuzzy system are explained.

3.2.1. Fuzzy input

The fuzzy input works with two fuzzy sets: high- similarity and low-similarity. The level of "truth value" of a particular membership function is named μ. The trapezoidal function was selected as the membership function because it has a flat top with belonging equal to 1 and a break-point which decreases linearly to 0. This condition allows that the sum of μ in both sets of the same *NC* value to be always 1, and therefore, the challenge is to determine the break-point of each membership function. After several tests (with words of different phonetic characteristics), we find the *NC* values shown in Fig. 6a, where each box contains the 95% of the results. These graphs show that the *NC* value for the cochleagram of two voice recordings of the same person (saved under different conditions) ranges around 0.7, while the corresponding *NC* value for two different people ranges around 0.3. From these results, the membership functions of Fig. 6b were proposed.

Summarizing, in the fuzzification process, the $NC_{j,i}$ values are mapped by the membership functions showed in Fig. 6b. The output of the process correspond to the truth values $\mu_{j,i}$ (where *i* is the number of the word, *j* is the number of the suspect). Since there are five words for analysis, there are five values of *NC* and therefore five values of μ per suspect.

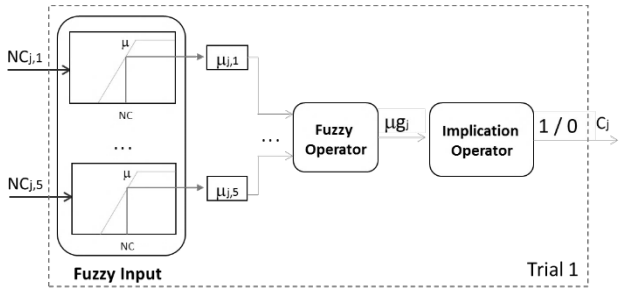

Figure 5. Proposed fuzzy system.
Source: The authors

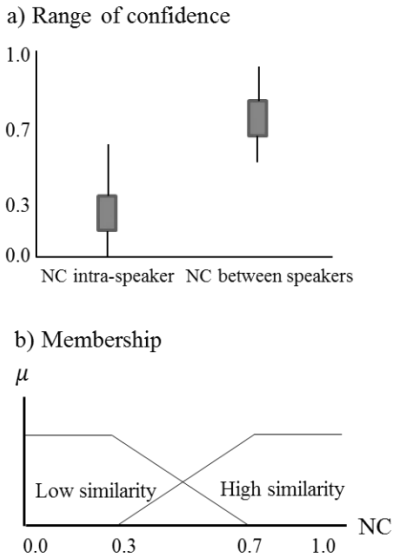

Figure 6. NC of inter and intra-speakers: a) range of confidence of several tests, b) proposed membership.
Source: The authors

3.2.2. Fuzzy operator

Once the five values of μ per suspect have been calculated in the previous block, the following step consists in determining a unique value per suspect related to the global similarity between him(her) and the audio proof. According to preliminary tests, it was found that a good operator among values of $\mu_{j,i}$ (for the same suspect) is the min function. However, it is neccesary to take into account that audio recordings with bad quailty can give low values of *NC* and then low values of μ. A solution is to calculate the second-lowest value (min_2) of $\mu_{j,i}$, for *i* [1 to 5] and *j* as the number of suspect. Therefore, output of this block is the global similarity (μ_g), calculated as follows:

$$\mu g = min_2(\mu_{j,1}, \mu_{j,2}, \dots \mu_{j,5}) \quad (1)$$

For example, suppose that the probabilities of the first suspect are $\mu_{1,1}$=0.79, $\mu_{1,2}$=0.72, $\mu_{1,3}$=0.75, $\mu_{1,4}$=0.5, and $\mu_{1,5}$=0.81, then the result of this blocks is $\mu g = \min_2(0.79, 0.72, 0.75, 0.5, 0.81) = 0.72$. The above result is the global degree of similarity of the suspect with respect to the audio proof.

3.2.2. Implication operator

Finally, a value of coincidence (Cj) is given according to the value of μg and a fixed threshold, calculated through eq. (2). This result is calculated by each suspect (*j*).

$$C_j = \begin{cases} 1 & \mu g \geq 0.7 \\ 0 & \mu g < 0.7 \end{cases} \quad (2)$$

3.3. Selection

According to preliminary results, a unique set of five words is not enough to determine if a suspect is the author of

the audio proof. Therefore, it is necessary to work with many trials (*T*). In each trial (*t*), five new words are selected, and the above steps (feature analysis and fuzzy system) are run again. For every trial, the suspect has a value of coincidence (C_{tj}). Then, the total score of coincidence (S_j) of the T trials is calculated, as follows:

$$S_j = \sum_T C_{tj} \quad (3)$$

If a speaker has a score of coincidence equal to or higher than 8/10, the system selects the speaker; otherwise, the system gives a null identification. The aim of this condition is to prevent false acceptance, which means selecting a wrong suspect.

4. Implementation and evaluation of the method

The purpose of this phase is to validate the proposed scheme in terms of the accuracy of the identification process. To evaluate the performance of the proposed system, we work with databases of 28 suspects. Also, there is an audio proof, from which five words have been extracted. Each suspect has pronounced the five selected words ten times (i.e. T=10) and their recordings are compared to the recordings extracted from the audio proof. If the total number of coincidences is at least 8/10 (of the total trials), the suspect is identified as positive (whole evaluation). This procedure is repeated for all suspects.

4.1. Evaluation measures

To measure the accuracy of the identification process, the following metrics are selected: overall accuracy (OA) and Kappa (**κ**) index. OA ranges from 0 to 1, being the latter the ideal value (i.e. all sources and not sources of the audio proof are correctly identified). On the other hand, **κ** ranges from -1 to 1, where -1 means perfect disagreement, 1 means perfect agreement, and 0 means a random level of agreement/disagreement.

Knowing the recordings that correspond to the same speaker and those that are not the source of the audio proof, it is possible to compute the true positives (TP), the true negatives (TN), the false positives (FP), and the false negatives (FN). TP is the number of sources of audio proofs identified by the system as positive coincidence (correct identification); TN is the number of suspects that are not the source of audio proofs and were identified by the system as negative coincidence (correct rejection); FN is the number of sources of audio proofs identified by the system as negative coincidence (incorrect rejection); FP is the number of suspects that are not the source of audio proofs but were identified by the system as positive coincidence (incorrect identification).

With the above metrics OA is calculated, as follows:

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

And the kappa index is obtained according to:

$$\kappa = \frac{OA-P_e}{1-P_e} \quad (5)$$

Where Pe is defined as:

$$P_e = \{P_1 * P_2\} + \{(1-P_1) * (1-P_2)\} \quad (6)$$

P_1 is the number of suspects identified as positive coincidence divided by the total number of suspects (i.e. P_1=(TP+FP)/(TP+TN+FP+FN)); P_2 is the real number of sources of the audio proofs divided by the total number of suspects (i.e. P_2=(TP+FN)/ TP+TN+FP+FN)).

4.2. Results and discussion

When evaluating the accuracy of the proposed method, 10 different cases were detected. Table 1 shows the meaning of each case in terms of **κ** and OA. According to the results, OA and **κ** are equal to 1 if and only if the performance of the system is perfect (1st case); it means the system identifies only the correct source of the audio proof. If the system identifies two suspects positively, i.e. the correct source and one incorrect source of the audio proof (i.e. the 2nd case), values of OA and **κ** are 0.96 and 0.65, respectively.

Table 1.
Meaning of parameters κ and OA

Case	κ	OA	Meaning
1st	1.00	1.00	TP=1; TN=27; FP=0; FN=0 Only one suspect is positively identified and s/he is the source of the audio proof
2nd	0.65	0.96	TP=1; TN=26; FP=1; FN=0 Two suspects are positively identified and one of them is the source of the audio proof
3rd	0.47	0.92	TP=1; TN=25; FP=2 FN=0 Three suspects are positively identified and one of them is the source of the audio proof
4th	0.36	0.89	TP=1; TN=24; FP=3; FN=0 Four suspects are positively identified and one of them is the source of the audio proof
5th	0.29	0.85	TP=1; TN=23; FP=4; FN=0 Five suspects are positively identified and one of them is the source of the audio proof
6th	0	0.96	TP=0; TN=27; FP=0; FN=1 No suspect is positively identified
7th	-0.037	0.92	TP=0; TN=26; FP=1; FN=1 Only one suspect is positively identified and s/he is not the source of the audio proof
8th	-0.05	0.89	TP=0; TN=25; FP=2; FN=1 Two suspects are positively identified and none of them is the source of the audio proof
9th	-0.056	0.85	TP=0; TN=24; FP=3; FN=1 Three suspects are positively identified and none of them is the source of the audio proof
10th	-0.06	0.82	TP=0; TN=24; FP=3; FN=1 Four suspects are positively identified and none of them is the source of the audio proof

Source: The authors

Table 2.
Average of the results by database

Database	OA	κ
1	0.97	0.66
2	0.98	0.78
3	0.97	0.76
4	0.97	0.70
5	0.97	0.79
6	0.97	0.67
7	0.98	0.83
8	0.97	0.76
9	0.98	0.82
10	0.97	0.69
Average	0.974	0.752

Source: The authors

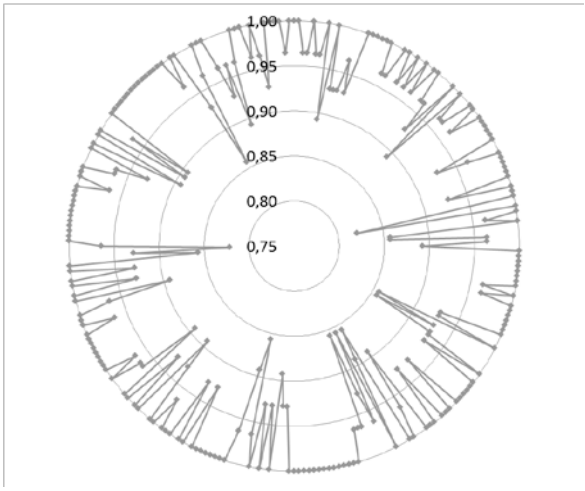

Figure 7. Radial plot of OA for 280 tests.
Source: The authors.

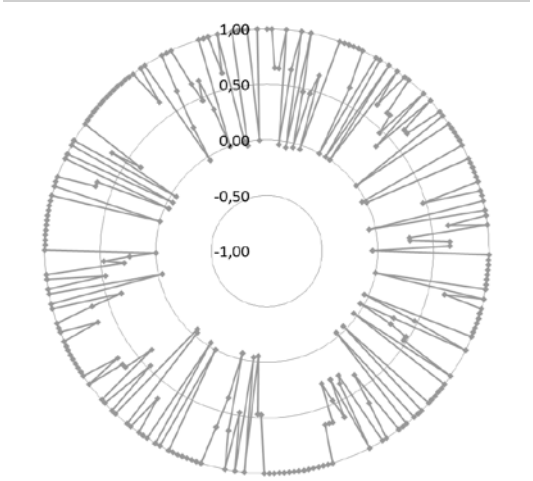

Figure 8. Radial plot of κ for 280 tests.
Source: The authors.

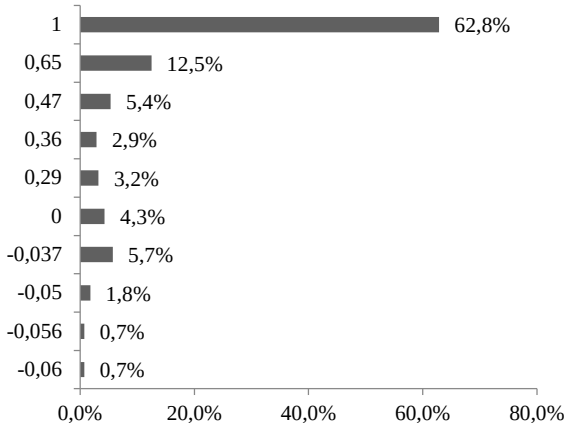

Figure 9. Normalized histogram of κ for the 280 tests.
Source: The authors.

But if none of the two identified suspects are the source of the audio proof (i.e. the 8th case), OA is 0.89 and **κ** is -0.05. It is worth noting that in the case of four positive identifications with one of them as correct (i.e. the 4th case), the value of OA is the same of the above situation (ie. equal to 0.89), but the value of **κ** is 0.36. It means, for the parameter **κ**, having a higher number of positive identifications with one of them as the correct is better than having a lower number of positively identifications without the correct source of the audio proof. In terms of OA, identifying a correct suspect is as important as rejecting a false source of the audio proof.

For an extensive validation, we work with 10 mini-databases each one with 28 suspects and 28 audio proofs. Every audio proof is compared with the 28 suspects of the database. At the end, there are 28 results by database, with a total of 280 tests. Table 2 shows the average of the validation parameters by database.

According to the results of Table 2 (i.e. $\kappa_{average}$=0.752; $OA_{average}$=0.974), the system is expected to mostly work between the 1st and the 2nd cases of Table 1.

Figs. 7 and 8 show the radar charts for the results of the 280 tests (OA and **κ**, respectively).

It is noteworthy that all results of Fig. 7 are higher than 0.8 and most of them are on the unit circle (ideal value). This means that all results are classified in some of the ten cases of Table 1.

Throughout the entire circumference of Fig. 8, there are a lot of values of κ equal to 1, which means that in some tests of each database, the performance is perfect (only one identified suspect as the source of the audio proof). However, it is necessary to count the cases in which the result is not the ideal. Then, the number of results by each value of κ is divided by the number of tests (280), obtaining a normalized histogram (sum of ocurrences equal to 100%).

According to Fig. 9, most of the results of κ (62.8%) correspond to the ideal performance, and in the second place (12.5%) two suspects are identified with one of them as the source of the audio proof. The percentage of results in which one of the identified suspects is the correct one (1st case to 5th case of Table 1) is 86.7%.

4.3. Other parameter of validation

One of measure in speaker recognition/identification systems is the DET or ROC curve [14], which consists in the plot of the false acceptance (i.e. FP) vs false rejection (i.e.

false negative). The AUC (area under the curve) is very useful to compare the behavior of two or more models, in which the best model is the one with the highest AUC [15]. Typically, a zero value of FP needs a high value of FP and vice versa. Then, it is very common that the plot starts in a high value of FN and it decreases as the value of FN increases.

However, with our proposal, the above behavior is not the rule. For example, our system works simultaneously with FN=0 and FP=0 in the 62.8% of the cases. It corresponds to the cases with κ equal to 1. Therefore, this parameter is not used in our project as a measure of the performance of the proposed system.

5. Comparison to related works

In this section, some related works are analyzed in terms of the selected features, the method of identification and the findings. In Avci's proposal, the speaker identification system is based on a genetic algorithm and a fuzzy inference system. Inputs of the system are 25 Turkish words by suspect. Its correct classification rate ranges from 87.7% to 91.04%. In Almaadeed's work, neural networks and wavelet analysis are used to identify the speaker. Several sentences are used to extract the features and results depend on the number and the kind of selected features; its performance accuracy ranges from 84% to 99%. Finally, in Daqroup's work, the features are five formants and seven Shannon entropies extracted from vowels which are used as the inputs of a feed-forward neural network. In this proposal, recognition rate is 90.09%. In all three methods above, it is mandatory to train the system prior to the identification task; this means the system needs to know the correct answer in advance. Therefore, these systems are useful in applications of security and authentication in which a training phase is feasible.

In the context of audio forensics, the audio proof is not known by the system in advance, and then, the identification process is a 'blind' task. For this reason, a good way of identifying the suspect is through the similarity between the suspect's voice and the voice in the audio proof. Unlike speaker identification for authentication, in which the output is a unique identified suspect, our output is one of three cases: null, a unique identified suspect, or with multiple identifications. Due to the above reasons, a quantitative comparison among our proposal and other methods for authentication is not feasible, and only a qualitative analysis can be performed.

In the specific area of speaker identification for audio forensics, it is remarkable that in Central and South America, identification is based on voice spectrogram (visual comparison). With our proposal, we select a fine time-frequency representation of the speech/voice signal at low frequencies (i.e. the cochleagram), as well as performing the comparison through a mathematical parameter (i.e Normalized Correlation).

6. Conclusions

In this paper a method for speaker identification is presented. The method is based on the normalized correlation between the Cochleagram of the suspect's voice and the Cochleagram of the voice in the audio proof; then, the NC value enters a fuzzy system. Cochleagrams are selected because they can represent the time-frequency behavior of the sound in the low frequencies in a better manner, and the value of similarity/dissimilarity is closest to the perpectual assessment.

The significance of this proposal is that the system works without a training phase, which means it is not necessary to have a knowledge of the suspects in advance. Furthermore, the five words extracted from the recordings can be selected by forensics every time.

The proposed method was validated in terms of overall acuracy (OA) and kappa (k) index. According to 280 tests (every one with 28 suspects and one audio proof), averages are 97.4% and 75.2%, respectively. For the fisrt parameter (OA), selecting the source of the audio proof is as important as rejecting the other suspects. For the second parameter (k), it is better to have a higher number of positive identifications with one of them as the source of the audio proof, than having a lower number of positive identifications all of them incorrect. As a result, our proposal has a good trade-off between correct identification, correct rejection and number of identified suspects.

References

[1] Almaadeed, N., Aggoun, A. and Amira, A., Speaker identification using multimodal neural networks and wavelet analysis. IET Biometrics, 4, pp. 18-28, 2015. DOI: 10.1049/iet-bmt.2014.0011

[2] Avci, E. and Avci, D., The speaker identification by using genetic wavelet adaptive network based fuzzy inference system. Expert Systems with Applications, 36, pp. 9928-9940, 2009. DOI: 10.1016/j.eswa.2009.01.081

[3] Daqroug, K. and Tutunji, T.A., Speaker identification using vowels features through a combined method of formants, wavelet, and neural network classifiers. Applied Soft Computing, 27(2), pp. 231-239, 2015. DOI: 10.1016/j.asoc.2014.11.016

[4] Pham, T., Genetic learning of multi-attribute interactions in speaker verification, Proceedings of the 2000 Congress on Evolutionary Computation, 2000, pp. 379-383. DOI: 10.1109/CEC.2000.870320

[5] Morrison, G., Sahito, F., Jardine, G., Djokic, D., Clavet, S., Berghs, S., et al., INTERPOL survey of the use of speaker identification by law enforcement agencies. Forensic Science International, 263(6), pp. 92-100, 2016. DOI: 10.1016/j.forsciint.2016.03.044

[6] Kober, V., Diaz-Ramirez, V.H. and Sandoval-Ibarra, Y., Speech enhancement with local adaptive rank-order filtering. Computación y Sistemas, 18(1), pp. 123-136, 2014. DOI: 10.13053/CyS-18-1-2014-023

[7] Ajmera, P.K., Jadhav, D.V. and Holambe, R.S., Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition, 44(10), pp. 2749-2759, 2011. DOI: 10.1016/j.patcog.2011.04.009

[8] Maher, R.C., Audio forensic examination. IEEE Signal Processing Magazine, 26, pp. 84-94, 2009. DOI: 10.1109/MSP.2008.931080

[9] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F. and Li, H., Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66(2), pp. 130-153, 2015. DOI: 10.1016/j.specom.2014.10.005

[10] Gao, B., Woo, W. and Khor, L., Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation. The Journal of the Acoustical Society of America, 135, pp. 1171-1185, 2014. DOI: 10.1121/1.4864294

[11] Zhao, X., Shao, Y. and Wang, D., CASA-based robust speaker identification. IEEE Transactions on Audio, Speech, and Language Processing, 20, pp. 1608-1616, 2012. DOI:

10.1109/TASL.2012.2186803
[12] Shao, Y. and Wang, D., Robust speaker identification using auditory features and computational auditory scene analysis, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 1589-1592. DOI: 10.1109/ICASSP.2008.4517928
[13] Patterson, R.D. Holdsworth, J. and Allerhand, M., Auditory models as preprocessors for speech recognition. In: Schouten, M.E., Ed., The Auditory Processing of Speech, 1992, pp. 67-84. DOI: 10.1515/9783110879018.67
[14] Beigi, H., Fundamentals of speaker recognition. Springer Science and Business Media, 2011. DOI: 10.1007/978-1-4419-5906-5_747
[15] Mazaira-Fernandez, L.M., Álvarez-Marquina, A., and Gómez-Vilda, P., Improving speaker recognition by biometric voice deconstruction. Frontiers in bioengineering and biotechnology, 2015, vol. 3. DOI: 10.3389/fbioe.2015.00126

D.M. Ballesteros-Larrota, received the MSc. degree in Electronic Engineering from University of Los Andes, Colombia in 2004, and the PhD. in Electronic Engineering from the Technical University of Cataluña, Spain in 2014. She is currently working at Universidad Militar Nueva Granada, Colombia. His research interests include signal processing and data hiding.
ORCID: 0000-0003-3864-818X

D. Renza-Torres, received the MSc. degree in Telecommunications Eng. from the Universidad Nacional de Colombia in 2010, and the PhD. in Advanced Computing from the Technical University of Madrid, Spain in 2015. He is currently working at Universidad Militar Nueva Granada, Colombia. His research interests include signal processing and remote sensing.
ORCID: 0000-0001-8073-3594

S. Camacho-Vargas, received the BSc.. Eng in Telecommunication Eng. in 2015, from the Universidad Militar Nueva Granada (UMNG). Bogotá, Colombia. He is auxiliary researcher at the UMNG. His research interests include software and signal processing.
ORCID: 0000-0002-9290-7798