# FORENSIC APPLICATION OF SPEAKER IDENTIFICATION

Dragoş DRĂGHICESCU[1]

*This paper describes a forensic application scenario in the speaker recognition domain and explains the techniques required to develop a complete forensic analysis system based on speaker identification. It is presented a standard manner of how Gaussian Mixture Models (GMMs) could be used in order to compute and express identification results in a judicial system. A review of possible techniques in forensic expertise is made. Experiments on a filtered POTS telephone database are exemplified from the forensics examiner perspective.*

**Keywords:** audio forensics, speaker identification, GMM-UBM

## 1. Introduction

Digital audio forensic field has a lot in common with digital signal processing (DSP), in tasks such as speech recognition, speaker identification and signal quality enhancement. There are many subtle differences in the usual problems faced by judicial audio examiners, from the general practice, mostly caused by the legal reasoning. The late evolutions of the entire forensic discipline are connected to the conclusions set forth by the scientific and forensic international bodies [1], which require adaptation of the DSP techniques to the legal system paradigms.

Audio forensics refers to recovery, acquisition, analysis and evaluation of audio intended for use in legal proceedings of the court of law or some other official venue [2]. The interested legal entities usually request the audio forensic examiners to:

- Verify the authenticity of audio evidence;
- Enhance the submitted recordings for speech intelligibility or audibility of other events in the environment;
- Interpretations of acoustic evidence, such as recognizing the speakers or the spoken words, reconstruction of the timeline of a crime, and diarization of the audio (also known as „who spoke when");

[1] PhD student, researcher at Speech and Dialogue (*SpeeD*) Laboratory, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: dragos.draghicescu@etti.pub.ro

The usual practices of forensics examiners are not so far behind those seen in the DSP laboratories. For a “simple” noise removal task, argued to enhance the intelligibility of an audio source, the defense side might object that this practice can result in modifications of the meaning of the recordings content. For example, *“I didn’t do it!”* has a slight difference from the phrase with the completely opposite meaning of *“I did do it!”*. The only difference is made by the nasal alveolar consonant “n”, which may be removed by some noise reduction algorithms. So here, the methodology chosen has to scientifically guarantee that the above perspective is not possible.

The practices involved with authentication of the audio were more accessible in the past, when the audio recordings were entirely analog, eliminating the possibility to mistake the copy as being original. One of the defining moments in the history of analog authenticity assessment was the well-known Watergate case, where, in a recorded conversation in the White House office of the former American president Richard Nixon, a gap was found that lasts for 18.5 minutes. The court appointed a committee for the job, which discovered that the missing part of the conversation was erased, due to the changes in magnetic development of the tape and to the background noise inconsistencies. In digital signal processing the provenance assessment becomes harder, as digital recordings are information, so they can be perfectly replicated at any time, as opposed to the analog audio, where losses are inevitable.

## 2. Speaker Identification Procedures

Speaker identification is a task that uses individual speaker traits found in recorded speech to infer the speaker identity. Most methods in speaker identification today use low-level biometric traits, such as specific vocal tract physical dimensions and articulatory traits. The selected features are analyzed in both their static and dynamic behavior, and the probabilistic model of the extracted features is attached to known persons who uttered the speech.

Two approaches to speaker modeling are taken mostly: generative and discriminative models. A statistic model is *generative* if it has the ability to generate probability distributions for data that was never seen by the modeling process, which is in fact a *model training* procedure. Generative models have a greater ability to perform speaker recognition for unseen data, which is often the case in the forensic field. From this class of models, Gaussian Mixture Models (GMMs) are common, and they are created through an adaptation technique from an Universal Background Model (UBM), which could be regarded as the random speaker model.

When used with adapted GMMs, the technique is named accordingly, such as GMM-UBM, for UBM-adapted GMMs. The speech signal submitted to speaker identification is uttered by the speaker, captured by a telecommunication device, coded and formatted for data compression, coded again, for error protection, and sent through the transmission channel. At the other end, the signal is received, recovered from transmission errors, decoded from the compressed format and sent to a loudspeaker to reach the ear of the correspondent. There are many stages the signal passes through to communicate an idea from mouth to ear, and each of them has potential to inherently apply their *fingerprint* on the carried speech signal. Data compression stages are the most effective in doing so, and that's why they capture more interest from engineers and forensic examiners.

For a classical landline telephone conversation, speech captured by microphones at both ends is sent in a full duplex manner on the same pair of copper wires, and effects of modules such as discontinue transmission (DTX) and comfort noise generators (CNG) are incurred.

If the signal is recorded from the telephone line, both upstream and downstream speech is caught. The mixed signal allows the codec to be identified, no matter the degree of superposition between the data streams. The channel effect might be thought as introduced by a transfer function associated to the codec used. The content received from the channel, without contribution of the local correspondent, is described as

$$Y_{channel}\left(t\right)=E_{clean}\left(t\right)\otimes H_{G712}\left(t\right)+N\left(t\right), \tag{1}$$

where $H$ is the transfer function of the channel, $E$ is the speech function and $N$ represents the noise function. For such a recorded speech signal to be accepted in speaker recognition, the known channel effects owed to the codecs have to be compensated. For this purpose, a codec recognition system is useful in which the signal $E$ may be modeled. Such a system would blindly tell the forensic examiner which codec was applied to the signal, for consistency checks needed in authentication, as well as for the right correction to be done.

In the field of speaker identification, when adapting the individual models, the specific features observed in the data are enhanced while typical features are attenuated so the adapted model mostly captures what is characteristic to the individual speaker. Patterns of other factors, like the influence of the codec function, the microphone effect function, and the background noise, should be excluded by this mechanism. That requires, if the UBM was trained with landline telephone recordings, the same to be true for individual known speaker data.

Speaker identification is a scenario in which several known persons are suspected to have uttered the questioned speech. Generally, there is no condition that the source of the questioned utterance is one of the suspected speakers, and identification is done in *open set*. If the questioned utterance is known to have been uttered by one of the suspects, a *close set identification* is performed. In a scenario with only one suspected speaker, identification task becomes *speaker verification*.

Manual and semi-automatic identification methods are prone to errors from subjectivity, lack of personal examiner experience, the choice of useful identity hints or accepted material, and so on. Automatic systems for speaker identification generally use low-level features of the speech that are emotion-insensitive, text-independent, and work in a more generalized manner than speaker verification. The first identification step is to create an UBM, or check an existing one to be a statistically correct model of the typical speaker. In this respect, about the same amount of speech should be included from persons aged between 18 (whose speech apparatus is considered to be mature) and 60 years, of all genres, for use in the training of the universal model. Acoustic features, usually Mel-Frequency Cepstral Coefficients (MFCCs), are extracted from the training files and their statistical information is stored in a GMM of the typical speaker, using the Expectation Maximization (EM) algorithm, which thus becomes UBM.

One similar-structured mixture model is necessary as a representation of each speaker known to the system, in a distinct, *enrollment* phase. The individual models of known speakers are obtained by adapting the UBM to the speaker data, usually through Maximum a Posteriori (MAP) adaptation, as described in [3].

In legacy speaker verification systems, decent performance was reached through the Dynamic-Time Warping (DTW) [4], as a time-domain comparison method between two data vectors. The DTW algorithm assigns a cost for each feature in both the known and questioned speaker feature vectors. Minimum cost, computed by dynamic programming, is compared with a preset threshold to reach a decision of whether the same speaker is the source of both compared utterances or not. The application of this method in biometric systems is good enough. However, constrains about saying exactly the same phrase are hardly applicable in digital audio forensics.

A typical speaker identification system consists of four major functions: *data acquisition and preprocessing*, *feature extraction*, and *speaker modeling* – for both enrollment and test phases, and *pattern matching and decision-making* [5]. In the following, we briefly describe all of them.

### *1. Data acquisition and preprocessing*

Usually a microphone is used to acquire speech from the analog acoustic medium. After the DC component is removed from the acquired signal, it is amplified and then sampled and quantized by an analog-to-digital converter.

Acquisition devices can show up in several aspects of speech quality. For example, a conversation recorded from a telephone line may contain lot of noise besides the actual speech. The noise may come from the quantization process, but the acoustic medium and the codec technology should also be considered. The frequency response of a microphone can be influential for the spectrum of the acquired digital original signal in a distinguishable manner, as described in [6].

Preprocessing is also employed at this stage. Audio data is processed in 10 to 30 ms frames with 30 to 50% overlap. Hamming windowing and first order pre-emphasis filter are usually applied to each frame.

### *2. Feature extraction*

This phase should be carefully considered in every speaker recognition system design, as the final recognition performance heavily relies on it. The best features need to have characteristics that forensic science also claims:

- individuality;
- relative stability; and
- reflectivity.

In other words, selected features should correspond to individual traits of the speaker, such as hidden dimensions of the speech production system, articulatory stereotypes, or other behavioral traits. They should be met frequently in natural speech, and should not change with time or under normal emotional states.

The age of subjects must also be considered, as young and very old speakers might show particular variability of voice. From the forensic perspective, voice samples collection could span across multiple weeks or months.

Speech was shown to be restlessly variable in time, but in short-time intervals of 10 ms to 30 ms a stationarity assumption is acceptable in most applications. The vast majority of systems use speech framing in this range and this is one important pillar of the good results shown by Gaussian mixture models. As their probability distribution remains unchanged in normal voice signals, parameters like mean value and variance of features are important.

The most used acoustic features are the Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding Coefficients (LPCCs). The MFCCs are taken on a scale closer to the human frequency perception, and perform better in noisy conditions, typical in forensic speaker identification systems.

Mel-frequency cepstrum coefficients are sets of coefficients obtained for each short-time analysis frame through a series of transformations applied to the signal. The following sequence of processing steps is necessary for the MFCC vectors to be obtained [7]: windowing, Discrete Fourier Transform (DFT), Mel frequency warping (the linear frequency scale of the Fourier domain is warped to fit the perceptual frequency scale of Mel), logarithm, and inverse DFT. Sometimes the first- and second-order derivatives of MFCCs are also taken as features.

### *3. Speaker modeling*

In both enrollment and test (recognition) phases, speech features are used to create a model of the speaker. For enrollment, enough speech utterances are needed, taken at various moments, for the model to be representative. In the GMM-UBM approach, individual speaker models are obtained by adapting the UBM to the speaker features extracted from their utterances. In the enrollment phase, speakers' models are stored in a reference database.

### *4. Pattern matching and decision-making*

The pattern matching domain is very complex, stretching from simple bit patterns to face recognition or matching of other complex models. Many algorithms qualify for use in speaker identification, such as Vector Quantization (VQ), Support Vector Machines (SVMs), and even Artificial Neural Networks (ANNs). In this paper we use the GMM-UBM as a statistical pattern recognition solution which assumes that features of the speech uttered by a certain person can be modeled as a weighted sum of Gaussian distributions.

The Gaussian mixture is completely described by the weights of the mixture components, the mean vectors and the covariance matrix of the individual features around the mean vector. The GMM only becomes statistically representative for feature vectors of a speaker when enough data variation was seen by the model in the adaptation phase.

The UBM accounts for what is typical for the average speaker. It is not a model of an existing person, but a general model in which all typical traits are collected, each with its own likelihood. When trained with representative data for all traits that could possibly appear in the speech of a person, it represents the background every speaker is expected to have. There are objective measures of the model optimization, such as Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC), but there is no translation of these into how much data should be involved in training the UBM.

Decision-making varies greatly depending on the algorithm type. In discriminative methods such as SVMs, the decision is made by finding the separation hyperplanes between two classes that maximizes a distance between the two classes. In generative methods like GMM-UBM, the decision-making approach follows the recommendation of using the Bayes theorem of hypotheses testing, as the only scientifically sound approach to forensic speaker recognition. Speaker identification is then based on the speaker recognition results, and other data available from the case file.

A general scheme that encompasses the previous discussed steps required by a conventional speaker recognition systems is depicted in Fig. 1. The suspected speaker's model is created in the enrollment phase, thereafter the speaker becomes known to the system. The newly created GMM is used in the consequent verification or identification procedures. In the recognition phase of either type, the GMM of the suspected speaker or the database, respectively, is used to create the corresponding likelihood distributions. The ratio of likelihood densities are taken from the generated distributions, for a certain speaker's model and the UBM, and a Bayesian decision can be made at either short-time analysis frame level or globally, at the questioned utterance level. The global decision is generally made by fusing the local decisions. If the time between two analysis frames is long enough, and we look at the MFCC vectors as random variables, a statistical independence assumption is reasonable between the consecutive frames, therefore the locally computed likelihoods are fused by multiplication.

Speaker identity is inferred after global likelihoods ratio is compared to a threshold established on a case-by-case basis. If from testing the questioned speaker utterance the prosecutor hypothesis receives more support than the threshold, then an identification decision is made. While this decision does not reflect a certainty, there are costs from possible errors of false identification or false rejection, for which only a magistrate is enabled to set a tradeoff.

Generally, the prosecutor brings suspected people in court because in the investigations he couldn't exclude the suspect from the possible authors of a criminal offense. This hypotheses is called thus *the prosecutor hypothesis*.

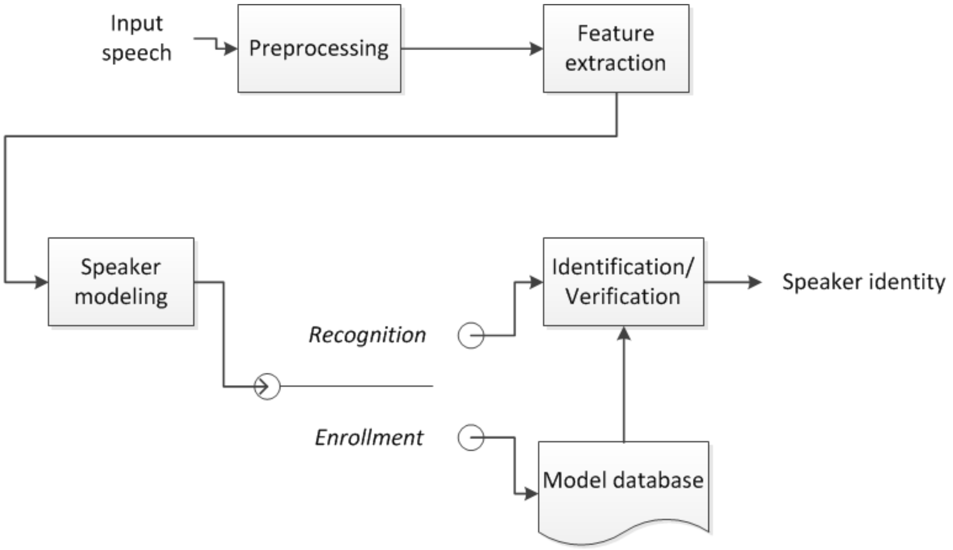

Fig. 1. The principle of a speaker recognition system

Defense counselors might choose from different ways to hypothesize the innocence of their clients, so theirs is *the alternative hypothesis*, which ever they choose. Sometimes defense makes no hypothesis. Because the Bayesian reasoning needs two hypotheses, sometimes the alternative hypothesis is the most general one. Either of the defense options is called a *defense hypothesis*.

The hypotheses being considered in speaker recognition are that of the prosecutor, for which the likelihood of the questioned speech being generated by the specific GMM of the enrolled suspect is taken at the numerator, and the defense hypothesis, for which the likelihood of the questioned speech is generated by the background model is taken at the denominator of the likelihood ratio (LR).

The LR is the final evidence for the forensic examiner, which essentially tells how many times the questioned speech is more likely to occur in the speech of the suspect (represented by a specific GMM) than to be observed by chance (represented by the UBM).

No decision is to be made by a forensic examiner about whether the suspect is guilty or innocent of the charges against him, as only a judge may make such decisions in the Romanian judicial system.

Final likelihood of the questioned speech being generated by a specific model may be regarded as a combination of decisions made for each short-time frame, or feature vector. Given an utterance of *T* frames from a speaker *j*, and considering a *D*-dimensional feature vector, the utterance can be mathematically expressed as $x_t \in \mathbb{R}^D : 1 \le t \le T$, where $x_t$ is the feature vector.

The likelihood ratio for an observed feature vector, namely a vector of *D* Mel-frequency cepstral coefficients, can be computed as the ratio of probability densities. For a *D*-dimensional Gaussian random variable, the *D*-variate probability distribution function (PDF) will be

$$g\left(\mathbf{x}|\mu_i,\Sigma_i\right)=\frac{1}{\left(2\pi\right)^{D/2}\Sigma_i^{1/2}}e^{-\frac{1}{2}\left(\mathbf{x}-\mu_i\right)^{'}\Sigma_i^{-1}\left(\mathbf{x}-\mu_i\right)}, \tag{2}$$

where $\mathbf{x}$ is the feature vector, $\mu_i$ is the mean vector, and $\Sigma_i$ is the covariance matrix, often considered just diagonal.

As the *D*-dimensional feature vectors from the questioned speaker are supposedly modeled as GMMs, a mixture for an *M*-component Gaussian probability density is

$$p\left(\mathbf{x}|\lambda\right)=\sum_{i=1}^{M}w_i g(\mathbf{x}\mid\mu_i,\Sigma_i), \tag{3}$$

where $w_i$ are the mixture weights and satisfy the stochastic constraint that $\sum_{i=1}^{M}w_i=1$.

A GMM is completely described by the mixture weights, mean vectors, and covariance matrices and can be denoted in a condensed form by

$$\lambda = \{w_i, \mu_i, \Sigma_i\}; \; i = 1, \ldots, M \, . \tag{4}$$

It is now obvious that the likelihoods of each feature vector could be extracted from the GMM probability distributions of the known speaker, in a close vicinity of the point determined by the feature vector in the *D*-dimensional space.

### 3. Example of a Forensic Proceeding

This paper does not intend to present real case data or give instruction to the readers about how the criminal procedure works. We will only point out background elements of use for the presentation, and avoid debatable parts.

Consider a case where a multimedia camera captures images and sound at a crime scene. Firstly, the reported possible crime is examined by the crime scene investigation (CSI) teams of the fact finder agencies (which is the Police, in Romania). Starting from the way the possible criminal fact was discovered (victim or witness reporting, Police officer discovered it by chance, automatic surveillance systems triggered alarms, or something else), the crime scene is identified and protection is set up against contamination. Situational data are then taken by marking positions of each evidentiary entity, or by completely recording it in 3D, by using the Spheron technology, for example. Reported facts are then checked with the situation at scene, where at least one criminal law article is held as applicable.

Assuming an armed burglary was reported, possible eye witnesses or ear witnesses are looked for, and evidence such as fingerprints, shoeprints, fibers, gun, powder, projectile, shells, and gunshot residues, DNA traces, as well as digital information storage and multimedia evidence are collected from the scene by the forensic team. A case file is started and assigned to a prosecutor which takes the lead of the investigation.

All of the above evidentiary items are seized, and preserved for subsequent analysis. Every possible clue is considered, not only the ones which point at some person. Findings of the investigators might trigger and guide different lines of investigation, which are tested for consistency with existing evidence in the case. An extended list of suspects is built based on the opportunities and the possible motives to commit the crime.

From a law enforcement agency point of view, when an explanation was clearly established about who are the victims, what is the prejudice, and what is the way the crime was committed, the case is solved. The allegations against remaining suspects are sent by the prosecutor to the court of law.

In all phases of the investigation, the forensic examiner services might be asked for, but especially in evidence interpretation. Multimedia evidence seized from the camera at the crime scene could be important in reducing the list of suspects. This is a case of speaker identification in which the questioned audio evidence is the relevant fragment of the seized multimedia camera recording.

Assuming the armed burglary was committed by two persons, from which one has possibly uttered the questioned speech, the preliminary recognition result takes the form of a list with known speakers that are more likely to have uttered the questioned speech evidence. The speakers that are very unlikely to have uttered the relevant speech are excluded from the list of possible sources, while the most likely ones are set for a new line of investigation. These speakers are then added to the suspect list and enrolled in the speaker identification system.

In order to conduct the identification, an UBM is prepared by the forensic examiner, mainly from speech recordings that seem similar to the questioned speech. If the two burglars are as likely to be female or male, two separate UBMs are necessary, one for each gender. At the identification phase, positive log-likelihood ratios (LLRs) show that the prosecutor hypothesis is more supported by the speech evidence than alternative, while negative LLRs give more support to alternative hypotheses.

From the video surveillance, police could also obtain an estimate of the constitution of the two aggressors, which can be used in exclusion of some suspects. The gunshot too might be an object of the forensic examination. It can give effective clues to reduce even more the circle of suspects. Is takes the role of verifying an eye witness or an ear witness testimony, for example. The muzzle blast, the sound of the shot itself, may be more informative of the bullet speed, than the gun identification itself, due to possible reverberations.

After all important details of the case were established by investigators, and the case went to the court, new challenges might appear for the forensic examiner. Authenticity of the multimedia recording of the camera could be verified by checking the following requirements [8]:

1. Was the recording device capable of making the questioned recording?
2. Was the surveillance recorder set up to use the device accordingly?
3. Is the recording device the one presented by the Police?
4. Was any alteration done to the recording (such as deletions, insertions, or any other kind)?
5. Was the evidence correctly preserved at all times?
6. Are the persons those identified by the Police using voice or image?
7. Were the participants to the conversation speaking and acting freely, or were they roleplaying some script?

One of the most important issues in legal practice is the evidence authenticity assessment. The evidence should be in continuous custody, and the general rule is that nothing has to be changed with evidence, from its discovery. For example, the contents of digital audio is not likely to change, but all the metadata such as time stamps could easily change.

The file structure should also be analyzed. Usually, the speech evidence is presented to court in a container file. This format may be analyzed for information about the structure of data, sampling rate, bit depth, number of channels and other clues. Any mismatch with the recording device could question or destroy evidence authenticity. For example, in wave files, RIFF format header may show information about the recording device. Also, the storage medium is important.

When the case gets to the point where the suspect is charged with burglary, ear or eye witnesses can have a great importance in the process, as they are indirect authentication means. They actually heard the acoustic events and seen images captured in the recording at the time of the crime.

One of the first cases known in history of a person being identified by his voice was the trial of William Hulet, in 1660, who was accused of executing King Charles I. A witness identified Hulet "by his speech" while his face was obscured [3]. This is also one of the first misidentification cases in history. The witness associated the voice he heard to the one of the persons he knew and asserted the voice has that of the defendant.

A verification of the ear witness statements with the opinion of a forensic expert is useful in many cases. A voiceprint that resides in someone's memory must be considered as decaying, resulting in only 57% recognition rate after a month and only 13% after five months [5].

The prosecution may postulate in court that "the recorded voice on the camera recorder belongs to the suspect". This will be considered hypothesis $H_0$. The defense may state the alternative that "the voice belongs to some unknown speaker", which will be considered $H_1$. In a forensic approach, the hypotheses of parts, $H_0$ from the prosecutor and $H_1$ from defense, are not being tested against each other, but the evidence would need to be tested against both of them. From a statistical point of view, we learn how to adjust our current odds about the matter by applying the Bayes rule of hypothesis testing, that is by multiplying the a priori odds by the likelihood ratio.

The numerator of the ratio is the likelihood that the questioned speech, $E$, was uttered by the suspect, while the denominator is the likelihood of evidence to randomly appear from an average speaker. The LR value thus includes all new knowledge inferred from speaker examination, mathematically expressed as:

$$\mathrm{LR} = \frac{p(E \mid H_0)}{p(E \mid H_1)} \tag{5}$$

The likelihood ratio presented to the judge in this case will be the ratio between the two likelihoods of the evidence. Beware of the "probabilities, given the evidence" which is an interpretation error known as the *prosecutors fallacy*.

This ratio can be better perceived by the judges when expressed as a log-likelihood ratio (LLR), because this way a negative value would show support for the defense hypothesis, while a positive value would give support the prosecutor hypothesis. There is a school of thought according to which the LLR should also be translated according to established *verbal scales*. In this case, the degrees of support the resulting LLR gives to which hypothesis is expressed verbally.

## 4. GMM-UBM system evaluation

The likelihood ratio represents a scientifically result for the expertise, but remains to be interpreted by the judge. Further it will be discussed how the LR can be obtained in the context of a GMM-UBM speaker identification procedure. Here, the LR could also be interpreted as a score. The identification is achieved by choosing the maximum score when the probability distribution of the verified utterance is compared with each model distribution.

An application was run on the TSP speech database [9]. The speaker identification experiment was conducted on the database after all files were filtered by passing through a G.712 codec. The database we selected consists of over 1,400 utterances spoken by 24 speakers, 11 women, 12 men, and a child. Children have rapidly changing vocal tracts, so models of their voices never last long. This is why in the evaluation tests we only used the 23 speakers that were mature.

The database speech files were recorded with a Digital Audio Tape recorder in an anechoic room, at a sampling frequency of 48 kHz, each lasting for about 3 seconds. After the filtering was applied, the new sampling frequency was 8 KHz. A number of 18 utterances of each speaker were selected for use in individual speaker model adaptation, while 150 of the files remaining from all speakers were used in UBM training. In the mixture models we considered 42 MFCCs, 64 Gaussians, and 10 Expectation Maximization (EM) steps until the convergence of the UBM.

We performed 828 tests (36 for each of the 23 speakers), and the recognition rate achieved was about 99.88%. The evaluation recordings was assured to be different than the training set. A well-known system performance indicator is the Equal Error Rate (EER), which was 0.53%. The evaluation was done in a closed set, as the system responses with the best match from the database, after calculating the likelihood ratio and comparing it to each database model.

Comparison results of each speaker test data to each speaker model are presented in Fig. 2 through a matrix of confusion, where files with data from the same speaker are vertically grouped together so as they may look as a small square. The vertical axis shows the number of the test, while on the horizontal axis the index of considered speaker model is shown.

The log likelihood ratio (LLR) is visually indicated for each test, in colors taken from a rainbow-like color map, named “Jet”, to be precise. High scores are shown by colors close to dark red, while lower scores are close to dark blue.

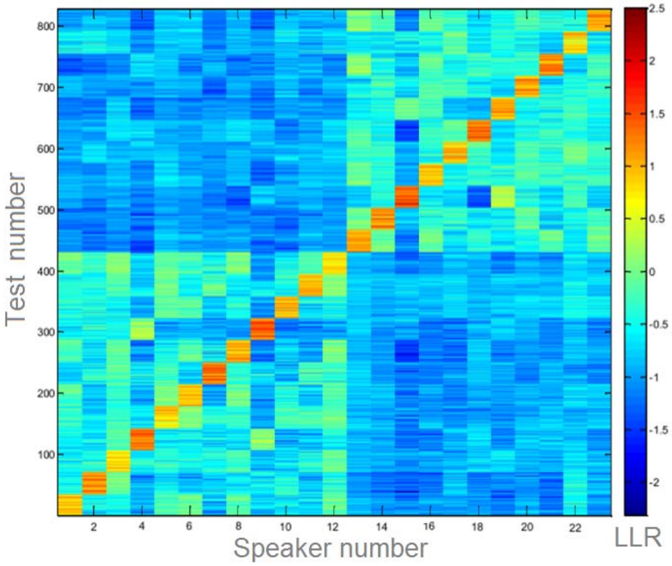

Fig. 2. Visual representation of the confusion matrix

The visual form of the confusion matrix shown in Fig. 2 reveals a lot of information useful in a forensic case. First, the higher scores are grouped on the reddish diagonal line, on which evaluation speech files are compared to the speaker models of the real source, *the target speakers*.

It is easy to see that the scores in the matrix are grouped in four rectangular parts. Rectangular zones which include parts of the high scores have consistent confusion degrees, demonstrated by their background in greener hues. Described rectangles correspond to the group of 12 male speakers and 11 females, respectively. Discriminative power of the system is indicated by the background of the other rectangles, which shows a low likelihood of women being mistakenly identified as men and vice versa.

The performance of the system under evaluation is depicted in Fig. 3 by a detection error tradeoff (DET) curve, along which the false negative error rate

(FNR) is shown as a function of the false positive error rate (FPR); the same error rates are sometimes called false rejection rate (FRR), and false acceptance rate (FAR), respectively. The figure shows that the achieved equal error rate was 0.53%, as it was previously mentioned.

The numerical value of the EER may be understood as a balanced tradeoff between FNR and FPR, which have very different effects in forensics. The criminal law in Romania is based on the principle of no punishment for the innocent, while nobody escapes law. At this point, a tradeoff is considered by magistrates, which are the only entitled persons to weigh social costs of the crime and of the possible identification errors.

In criminal investigations, a tradeoff could be set by using detection error tradeoff plots. Technically, when a FPR is deemed acceptable, this corresponds to a point on the curve that becomes the selected operating point. If we consider 0.05% as the maximum acceptable FPR, we automatically select the operating point marked in Fig. 3 in red color. As a consequence of our option, a tradeoff FNR of about 4% is also accepted. For a forensic application, FPR should be as low as possible, as a false positive is not acceptable.

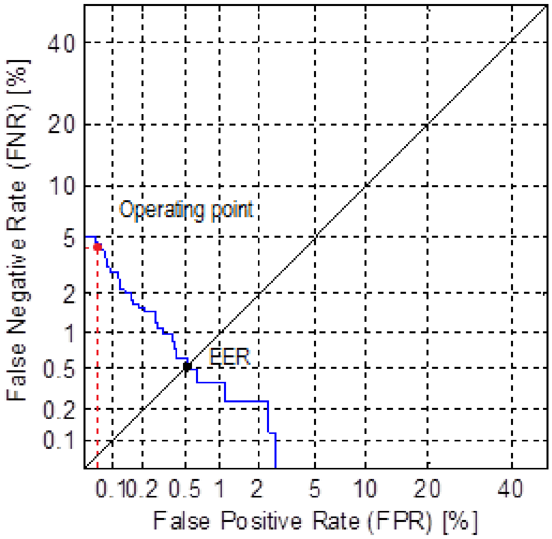

Fig 3. DET curve of the GMM-UBM speaker identification system

## 5. Conclusions

In this work we described forensic application scenarios and explained several techniques that can be used in developing speaker identification systems for use in the forensic domain. Expressions needed in quantification of the matching degree between a questioned speech signal and the model of a suspected speaker were discussed, without trying to be exhaustive.

Attention has been given to proper evaluation of the material subjected to forensic analysis, including proper authentication. General system performance issues have been discussed.

The perspective of using the general scientific method in the judicial system was briefly presented and applied, as an example, to a GMM-UBM speaker identification system. If used properly, the system may produce reliable evidence about the identity of the unknown speakers whose utterances appear in the questioned audio.

The evaluation system presented in Chapter 4 has promising accuracy and reliability. The recognition rate achieved was about 99.88% and the equal error rate was only 0.53%.

It is worth mentioning that most current commercial speaker recognition systems have their best EER values starting from about 1.5%, while requiring at the same time questioned audio files to include at least 90 seconds of pure speech.

## REFERENCES

[1] *National Academy of Science*, “Strengthening Forensic Science in the United States: A Path Forward”, 2009. [Online]. Available: http://www.nap.edu/openbook.php?record_id=12589&page=R1.

[2] *A. Eriksson*, “Tutorial on forensic speech science. Part I: Forensic phonetics”, Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Sept. 4-8, 2005.

[3] *R. Douglas and R. Rose*, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, in IEEE Transactions on Speech and Audio Processing, **vol. 3**, 1995, pp. 72-81.

[4] *S. Segărceanu and T. Zaharia*, “Speaker Verification Using Dynamic Time Warping”, in U.P.B. Sci. Bull., Series C, **vol. 75**, Iss. 1, 2013, pp. 179-194.

[5] *J. Campbell*, “Speaker Recognition: A Tutorial”, in Proceedings of IEEE, **vol. 85**, no. 9, 1997, pp. 1437-1462.

[6] *B. Plichta*, “Best Practices in the Acquisition, Processing, and Analysis of Acoustic Speech Signals”, Michigan State University, 2000, [Online]. Available: http://www.historicalvoices.org/flint/extras/Audio-technology.pdf

[7] *S. Therese and C. Lingam*, “Review of Feature Extraction Techniques in Automatic Speech Recognition”, in International Journal of Scientific Engineering and Technology, **vol. 2**, no. 6, 2013, pp. 479-484.

[8] *R. C. Maher*, “Audio Forensic Examination: Authenticity, Enhancement, and Interpretation”, in IEEE Signal Processing Magazine, **vol. 26**, no. 2, 2009, pp. 84-94.

[9] *Telecommunications & Signal Processing Laboratory*, “TSP Database”, Electrical & Computer Engineering, McGill University. [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Data/index.html.