University POLITEHNICA of Bucharest

Doctoral School ETTI-B

PhD THESIS

– SUMMARY –

CONTRIBUȚII LA EXPERTIZA CRIMINALISTICĂ A ÎNREGISTRĂRILOR AUDIO DIGITALE

CONTRIBUTIONS TO FORENSIC EXAMINATION OF DIGITAL AUDIO RECORDINGS

PhD student: **MSc. Ing. Gheorghe POP**

Doctoral supervisor: **Prof. Dr. Ing. Corneliu BURILEANU**

BUCHAREST 2020

Acknowledgment

This work would never have been possible without the essential contribution of professor Corneliu Burileanu, concerning both the advanced scientific studies and the research conducted during the doctoral programme.

I am particularly grateful to professor Dragoș Burileanu, whom I want to thank in this way as well, for his contributions to my scientific and personal growth, his recommendations, the patience and moral support generously offered over the past ten years and, especially, during the elaboration of this thesis. The success of this work is largely due to him.

Professors Andi Buzo and Horia Cucu also deserve deep appreciation for consolidating my specialized training in the field of speech technologies.

I am indebted to my doctoral research colleagues, Dragoș Drăghicescu, Sorin Rusu, Alexandru Caranica, and Șerban Mihalache, for their support, camaraderie and collaboration, which they offered me without hesitation.

The author's gratitude is especially directed to his fellow forensic experts from the National Institute of Forensic Expertise, especially to Grigoraș Bețiu – the director of the institute, Sorin Alămoreanu and Constantin Mirea – associate professors within the BIOSINF master's program, and, last but not least, to Cristian Dumitrescu and Florin Rușitoru, for their permanent encouragement and support in continuously raising of the scientific level of forensic research as well as of the forensic caswework for which I have been appointed.

I also thank my wife, Mona-Luiza, and my son, Radu Cristian, for the patience and moral support offered throughout the elaboration of this thesis.

Content

1. Introduction

Lawful audio recordings, produced by authorities or by private individuals, represent means of proof in the romanian law. However, it cannot be used as argument in motivating court decisions if they are not truthful. The authenticity of audio recording filed as evidence must be checked by (*judicial*) *expertise*. The Romania's Code of Civil Procedure describes the expertise as *opinion of a specialist (or of a panel of specialists) on the questionable evidence filed, in view of establishing elements of fact.*

1.1 Presentation of the field of the thesis

The domain of the thesis consists in the problems of forensic expertise concerning evidence in the form of *audio recordings*, term which we expand so as to also accomodate audio components în audio-video or multimedia files, with the most important components being:

- audio recording authentication;
- speech recognition; and
- enhancement of recorded speech.

1.2 Aim of the thesis

Presented doctoral thesis aims at contributing solutions to current scientific and practical problems in the field of the audio recording forensic expertise, still avoiding to become, by delivering ecessary details, a full-blown, self-paced initial training for those who wish to become forensic experts in the field.

1.3 Contents of the thesis

The research works presented in this thesis concentrate on four types of forensic expertize problems concerning recordings with audio content (Chapters 4 – 7).

In chapters 2 and 3, theoretical backgrounds and concepts are presented, needed as foundation of the research works presented in chapters 4 – 7.

Chapter 4 was dedicated to original contributions of te author regarding audio recording authentication by analyzing residual traces of electric network frequency (ENF) variations, so called *ENF analysis*, including all three of its components:

- efficient construction of ENF reference databases;
- accelerated search of match between traces extracted and reference databases; and
- the adaptation of ENF criterion for use in an evaluative framework.

In chapter 5, contributions are presented to audio recording authentication of audio recordings ny analysing the traces of compression, în particular the recognition of AMR (-NB) codec.

Chapter 6 introduces and evaluates an automated system for speaker recognition for forensic expertise, taking into account the intrinsic quality of recorded speech.

In chapter 7, a method for forensic speech enhancement is presented, which uses a deep neural network (DNN) and considers the quality of speech in the form of three quality measures, defined and used in the previous chapter.

Chapter 8 presents the main contributions of the author to forensic expertise of audio recordings, the results obtained, and the papers published by the author in field of the thesis, as well as the further development perspectives considered.

2. The field of forensic expertise of audio recordings

2.1 The judicial expertise in the context of forensics

From all reactions of forensic community to missing or low scientific level of forensic reporting in various speciality areas, the reaction of the Scientific Comittee for Digital and Multimedia Evidence, as part of Organization of Scientific Area Committees for Forensic Science (OSAC) stands out. It has set up a working group of academics, practitioners and international forensic organization representatives, with a mission to clearly define the place of digital/multimedia evidence in forensics. The framework defined in the report of the workgroup has made obvious the redefining of some fundamental concepts, such as *science* (*the sytematic and*

coherent study of phenomena of a certain kind) and *trace* (*any modification, subsequently observable, created in the environment as a result of an event*).

Under these circumstances, the forensic science is defined as *the science of the systematic and coherent study of traces, in order to authenticate, identify, classify, reconstruct, and evaluate it for a legal context* [Jac13].

An efficient administration of traces, as material evidence attached to the files of Romanian judicial entities, is generally performed in a Court of Law or in view of such future use, by producing *means of proof*.

The Civil Procedure Code defines the *expertise*, and lets the special laws deal with the details:

- Government Ordinance (GO) no. 2/2000 on judicial and extra-judicial technical expertise;
- GO no. 1/2000 regarding the activity and funcționing of legal medicine institutions; and
- GO no. 75/2000 regarding the organization of forensic expertise.

Forensic expertise is a kind of judicial expertise, which is distinct from the technical expertise, which deals with examination of material traces in order to authenticate it, to identify the creator object of the traces (be it persons, objects or phenomena) or the modifications thereof, as well as to evaluate the their probative strength.

Forensic expertise may include technical expertise activities, in domains including audio recordings, as long as they are necessary in order to reach its own objectives. Forensic expertises are performed in institutes and laboratories of specialty, public or private, established by law, from which the most representative are:

– The National Institute of Forensic Science (INC) as part of the General Inspectorate of Romanian Police (IGPR), under the Ministry of Internal Affairs; and

– The National Institute of Forensic Expertise (INEC), under the Ministry of Justice.

2.2 Audio recording authentication

Traces in the form of recorded acoustic events, be it audio only or audio-video recordings, are the object of *speech and voice expertise*, with its standard objectives being *audio recording authentication*, *speaker recognition* and *speech enhancement* [Pop14b].

Under the risk of being declared void, decisions of Courts of Law cannot be based on inauthentic evidence, and this is why evidence that magistrates cannot interpret on their own call for *verification* (that is, *authentication*) through expertise.

Authentication methods are generally “blind”, and aim to establish several characteristics of the creator object with no a priori knowledge about how the evidential recordings were made.

Speaker recognition expertise was the first to be used in authentication recorded conversations, and later evolved as a standalone expertise domain.

2.3 Speaker recognition

Technically speaking, the *identification* of the speaker in a *closed set* (in which all potential sources of the evidential speech are known) or *open set* (one more, unknown, potential source is considered), the *detection* of the speaker (establishing his presence by his speech) and speaker *verification* (there is only one suspected speaker) are all scenarios of the same approach in the field, called *speaker recognition*.

Scientific basis of speaker recognition consists in the collection of anatomic and behavioral properties of speech production systems which are reflected in speech.

Speech can be defined as a succession of acoustic events carrying a linguistic load. At the cerebral cortex, both the speech heard is understood (Heschl gyrus, Vernike zone etc.), and the message to be communicated by speech is elaborated (Broca zone etc.), before the acoustic signal of speech is produced, which reaches to the ears of the listener. În the context of perceived noise, people tend to speak louder, as a reflex reaction (Lombard effect).

When producing speech, the vocal tract processes the air flow expulsed by respiratory sistem, filters it and let it out through both the mouth and nose. Supraglotic air ways have anatomic structures through which *voicing* and *articulation* are performed. While voicing is produced in the larinx, articulation is performed simultaneously and in coordination by all mobile elements of supraglotic airways, thus called *articulators*.

Speaking în mother tongue is learned at the imitation age. The children repeat what they hear, then use the speech in a reflex way. Automatisms aquired during learning the mother tongue, as well as other languages, reflect in speech thus contributing to speaker recognition.

The knowledge of filter parameters realized by the vocal tract of a suspected person, estimated from reference speech samples, allows verifying if the questionable speech signals, in the form of specific spoken words, may have been produced by the speech production system of the respective person.

2.4 Speech enhancement

Low quality of recorded speech is generally caused by dificult recording or transmission conditions, while speech enhancement aims at ensuring the understanding of what was spoken.

Speech has informational redundancy, which can be reduced through *compression*, based on perception properties of human hearing. Anatomic parts of human aural analyzer (the ear) form the *external ear*, *middle ear*, and *internal ear*, while the hearing sensation is realized in the cerebral cortex, stimulated by neural action potentials collected by the auditory nerve.

The transmission of information through neural paths, which assumes the release in the sinapses of auditory nerve of neurotransmitter substances, introduce a *relaxation time* of the ear, needed for recovery of released substances. Acoustic events closer in time than the relaxation time, thich lies in the range of 50 ms – 100 ms, cannot be distinguished.

Natural speech has a frequency band which may go from about 70 Hz to 7 kHz, although about 95% of its energy is found in a frequency band limited between 100 Hz and 5 kHz. Speech recorded through phone landlines is band-limited between 300 Hz and 3.4 kHz. These numbers emphasize a great difference between the frequency bands of recorded speech and that which can be perceived by humans, which spans from 20 Hz to 20 kHz.

3. Computational methods in audio forensics

This chapter aims to concentrate the main mathematical computational methods used during the doctoral research program – specific to the field of audio recordings and to the forensic application of speech processing – for the prezentation of conducted research to look more simple and clear.

3.1 Mathematic computations in forensics

The field of forensics often needs reasoning under uncertainty conditions, in order to ensure correct etiologic connections between traces examined and the conclusions of the expertise. Machine learning is used in order to solve problems which cannot be analitically dealt with.

In the framework of forensic speech expertise, models of speech and speaker are learned, which use the speaker recognition tasks various scenarios (identification, verification, detection, screening etc.).

A few reasonng modalities stand out:

- *The reasoning by analogy*, centered on performing known/unknown comparisons.
- *Deductive reasoning*.
- *Inductive reasoning*.
- *Probabilistic reasoning*.

International Organization for Standardization (ISO) has published a guideline to assist approaching the measurement uncertainty, to be applied to all numeric result determinations.

Uncertainty is communicated both numerically and by verbal scale expressions. *The European Network of Forensic Science Institutes* (ENFSI) recommends the use of the *likelihood ratio* (LR), as well as its equivalence to conventionally established verbal scales.

When the forensic expert has to reach an answer to the question set forth by the judiciary, finds himself of herself in a position to enquire facts he or she has not perceived directly, at the time they happened. Their complete investigation is not feasible, as not all hypotheses can be tested. In such contexts, *the method of negative tests* is used, in which only tests which may lead to the *exclusion* of evidence authenticity are relevant. Any *study* (or *method* or *experiment*) designed to confirm a hypothesis or theory is, by this definition, *not scientific* [Pop63].

The analysis of recorded speech in order to change its short time representation assumes processing by short duration signal fragments, 10-35 ms, with or without overlap. The initial representation can be restored by synthesis.

In the every day forensic expertise practice, digital signals are seen as sequences of numbers, to which several representatins may be applied through analysis, such as: the discrete power spectrum, the spectrogram, filterbank decompositions (especially for cepstrum calculation), sets of coefficients (LPC, MFCC etc.) or speech signal quality measures.

The sets of short term representation features are described by their probability distributions.

3.2 Machine learning. Methods and algorithms

The main objective of machine learning is to assist computational systems in performing tasks without a need for human intervention.

Several learning modes are used by inteligent systems to aquire data about hidden states of the environment:

- *unsupervised*, for clustering or data dimensionality reduction;
- *supervised*, for function approximations, generative or discriminative modelling;
- *semi-supervised*, for partially labelled data;
- *reinforced*, to establish, in cooperation with an expert in the field, what actions are to be performed in the application domain.

3.3 Algorithms used in speaker recognition

Comparable speakers are examined according to the recognition scenario selected and to speaker and population models, which are obtained through machine learning. Using the models learned, the hypothesys supported by the findings is identified, as well as the degree of support lent by forensic findings to the respective hypothesis.

Individual speakers can be modelled for example by a probability density of *Mel-Frequency Cepstral Coefficients* (MFCC), also called *Mel-cepstral coefficients* or *MFCC features.* Seing that voice parameters vary with large numbers of factors, the MFCC feature probability density model resembles more a *Gaussian Mixture Model* (GMM).

Gaussian mixture models are learned through Expectation Maximization algorithm (EM), based on a large amount of data available. This is the case of Universal Background Models (UBM) which are usually built for populations. In usual cases, available data for each speaker are not enough to use EM for individual model learning, so that the Maximum A Posteriori algorithm (MAP) is preferable. In automated systems, similarity scores or likelihood scores are computed. Similarity scores raise problems generated by the random similarity between speakers.

A discrimination instrument is used in such cases, between tipical features and specific ones, readily offered by the very MAP algorithm. The usual likelihood score is the *likelihood ratio*,

$$\mathrm{LR}_k = \frac{p(\mathrm{X} \mid \mathrm{H}_{0k})}{p(\mathrm{X} \mid \mathrm{H}_1)}, \qquad (3.67)$$

where H_{0k} and H_1 are the forensic hyotheses used in the evaluative reasoning:

H_{0k} – unknown speaker is the same as speaker *k*; and

H_1 – unknown speaker is a ramdon member of relevant population, respectively.

The LR value conveys to the output both the decision of recognition and the strength of the evidence. Professionals in the field of Law cannot operate with numerically expressed strength of the evidence, so that the paralel use of verbal scales is preferred.

3.4 Deep learning

Forgotten at the level of interesting idea, because of the lack of computation power at the time, artificial neural networks were reactivated in the second half ofthe ’80s. Nowadays, machine learning us used in automatically adjusting the neural network parameters even for deep architecture networks, which is now called Deep Learning (DL).

Paul Werbos proposed in 1990 one of the best algorithms for training neural networks, called *backropagation* [Wer90].

As a reaction against the so-called *superior order efects* has been the *batch normalization*, proposed in [Iof15]. Measures taken in order to normalize data by batch make the gradients more statistically relevant for training the network, but do not eliminate the superior order efects which, stimulated by the input data, may divert the learning. A more thorough solution was the *layer normalization*. The main difference between the two normalization means is the action radius of normalization statistics.

In batch normalization, statistics are common to all neurons on the same batch, while in layer normalization they are the same across each feature in data vectors and are independent of other examples.

One method to prevent neuron activation co-dependence is the *probabilistic omission* (or *dropout*).

4. Audio recording authentication through ENF analysis

The electricity distribution network with national coverage, called the electrical network, works in interconnection and sincronicity with most of the european electric network.

4.1 Audio recording integrity check through ENF analysis

The load of electric networks varies for obvious reasons, and temporary unballance fire up control mechanisms which are designed to maintain the network standard parameters. Generated power is controlled to follow the power consumed, leading to variations of electric network frequency (ENF) in the form of rapid, unpredictibile frequency jumps, followed by asympthotic recovery of the balance. First research works in the field, conducted by Cătălin Grigoraș [Gri03], have defined the *ENF criterion*, which assumes three stages:

- reference database collection;
- detection and recovery ot ENF traces in questionable files; and
- matching the traces against a reference.

Traces detected in recorded material are recovered for analysis as a series frequency values, which is compared to the reference at the time alleged. Reference databases are managed by Transmission System Operators (TSOs), but forensic laboratories may benefit from building one.

4.2 Proposed method for building ENF reference databases

In Fig. 4.1, the block diagram id shown of the collection module implemented by the author [Pop17]. Two opposed phase signals are connected to analog DC circuits, which completely translate the waveform in the positive voltage range. At each periodmeter input, the signal il shaped regtangular and activates a gate for a local clock impulses, which are numbered.

Variations of electric network frequency are small,

$$\varepsilon = \frac{|\Delta f|}{f_G} < \frac{0.15}{50} = 0.3\%. \qquad (4.2)$$

Considering $\varepsilon^2 < 0.003^2 = 0.000009$, terms with powers larger than 1 in the MacLaurin infinite series can be omitted,

$$1 + \varepsilon \approx \frac{1}{1 - \varepsilon}. \qquad (4.4)$$

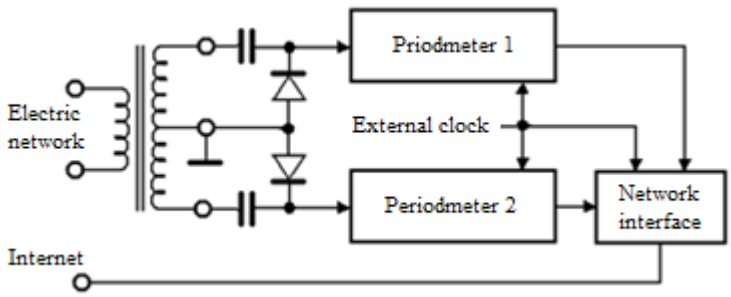

Fig. 4.1 *The block diagram of the proposed collector module* [Pop17].

Equation (4.4) has double impact. It makes possible a very fast conversion of relative periods in relative frequencies. On the other hand, if 1+ε is Gaussian, 1–ε is Gaussian too. Expected values of both series 1+ε and 1–ε, are 1, while their standard deviations are identical.

The success of the method is ensured by expressing on 15 bits only the rapidly changing part of the measured duration, laying between 19.9 – 20.1 ms for 99.99% of time. For an internal clock frequency of 30 MHz, the method ensures a temporal resolution of 33 ns, and a relative resolution of $33 \cdot 10^{-9}/20 \cdot 10^{-3} = 1.65 \cdot 10^{-6}$.

Compared to the opriginal approach in [Gri03], in which 120 audio samples are stored per second, 16 bits each, the method proposed in this thesis uses the storage 120 more efficient, while the annual data fit in **121 MB**, that is, less than 20% of the storage capacity of a optical compact disc (CD) per year.

The relative measuring precision is $\mathbf{1.65 \cdot 10^{-6}}$, that is, **0.00008 Hz** in absolute terms. The method's precision compares in Table 4.1 to the precision of other methods.

Table 4.1 *Precision and duration of analysis frame for compared methods*

Extraction method	Precision [Hz]	Frame duration [ms]
FDR [Zha10]	0.0005	20
ESPRIT [Haj12]	0.0001	2000
MUSIC [Haj12]	0.0002	2000
Spectrographic (QIFFT) [Coo09]	0.0006	4000
Proposed method	**0.00008**	1000

4.3 Proposed method for accelerated match search

Short term behavior of a reference ENF sequence may be evaluated by comparison to its *moving average* (MA) as shown by the example in Fig. 4.4. The search domain is pruned by eliminating positions where the binary correlation factor,

$$\mathrm{C}(k) = 1 - \frac{1}{N}\sum_{n=1}^{N} \mathrm{XOR}(bq_n, br_{n+k}) \,, \qquad (4.5)$$

is lower than an empiric threshold.

A final decision regarding the match of trace sequence to the reference comes in a second step, based on the usual sequence similarity indicators. The search speedup obtained by pruning is at least twice for trace sequences of 240-sample lonh sequences, and **at least 7 times** for sequences of 600 samples.

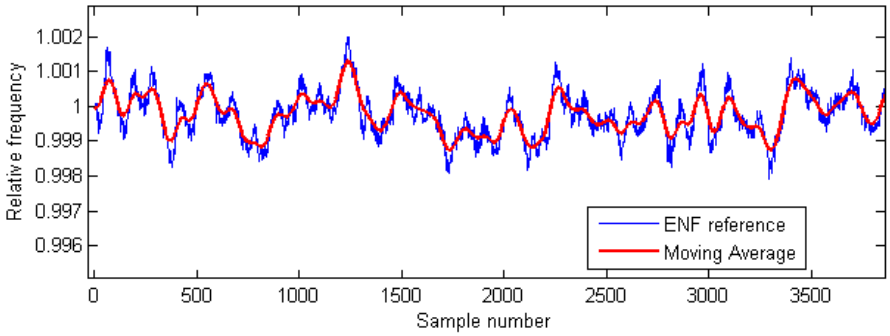

Fig. 4.4 *Exemple of reference sequence and the relation to its moving average.*

4.4 Evaluative framework for ENF-based authentication of digital audio

Evaluative frameworks consider all possible interpretations of the evidence. The background of the problem comes out from the description in [Gri03] of the reference framework as well as from the necessity in forensics to express conclusions in an evaluative manner. Real life ENF trace carrier file databases are still scarce, so that authentication frameworks discussed in this chapter are compared on corpora prepared by mixing electric network signal with steady and variable noise.

Reference framework uses both the correlation coefficient (CC) and the mean square error (MSE) as similarity indicators between the recovered trace and the reference. A pair of same length and same time-step sequences is labelled either as “match” or “mismatch”, with no evaluation of the strength of evidence.

Proposed framework is **an evaluative implementation of ENF criterion**, in which the strength of evidence is established, together with the strength of evidence. Before the use of proposed framework, the existence and unicity of the trace is established. Starting from the clues available in the file, te search domain is pruned, by using the method proposed for that, described in section 4.3. Each reference sequence potentially matching the trace is checked using the **similarity measure proposed by the author**, *the relevant similarity* (RS), as *the number of trace ENF sequence samples which differ from the corresponding reference sequence samples by less than a given threshold*.

In order to test the proposed evaluative framework, described in detail in subsections 4.4.2 and 4.4.3, I have used a database of 4464 electric network signal files, consecutively recorded as audio for 31 days recorded as audio, each lasting for 10 minutes, in a fixed arrangement. From this database two corpora were prepared afterwards: one with a steady level of white noise, at four signal-to-noise ratios (SNR), and another with a variable SNR, with rock music as noise.

The first corpus was used for an experimental evaluation of the proposed authentication framework on ENF traces over a stationary background. For the evaluation of proposed framework with variable quality traces, both prepared corpora were used. A total of four experiments were conducted.

- First experiment provided the reference framework with evaluative capability,

$$\mathrm{LR} = \frac{\mathrm{pdf}(\mathrm{CC}|\mathrm{H}_0, D)}{\mathrm{pdf}(\mathrm{CC}|\mathrm{H}_1, D)}, \qquad (4.11)$$

where pdf(.) is the probabilkity density function, CC is the correlation coefficient, and *D* represents context data, such as the SNR value, trace duration, etc.

Probability distributions of CC are shown in Fig 4.9 (a), (b), (c), and (d). From the figure it comes out that a decision threshold depends on the knowledge of SNR and other unknown factors, so that this way of transforming the reference framework in an evaluative one is not feasible.

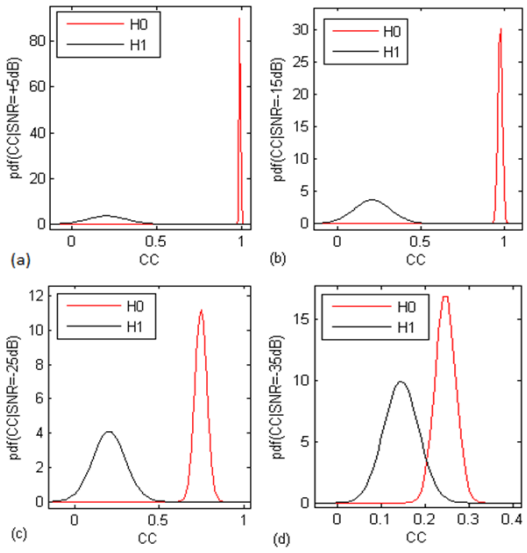

Fig 4.9 *CC probability distributions for trace sequences with SNR of* (a) *+5 dB*, (b) *–15 dB*, (c) *–25 dB, and* (d) *–35 dB.*

- During the second experiment, the distributions of RS probability were built. As shown in Fig. 4.10 (a), (b), (c) și (d). It becomes clear from the figure that directly replacing CC with RS does not produce an appearent advantage, for the same values of the SNR.

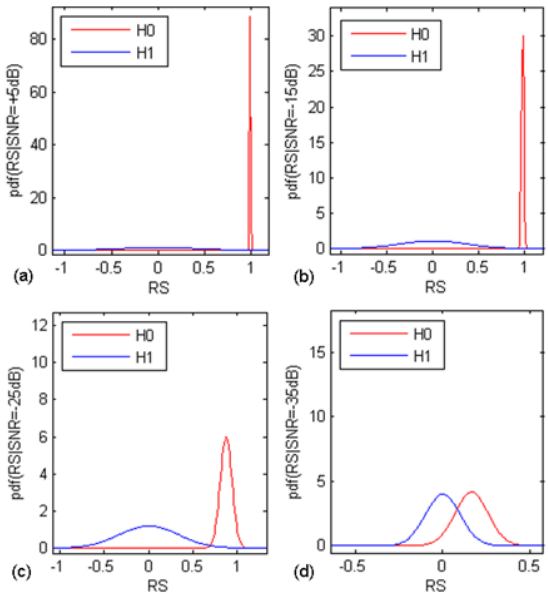

Fig. 4.10 *RS probability distributions, for SNR values of* (*a*) *+5 dB,* (*b*) *–15 dB,* (*c*) *–25 dB, and* (*d*) *–35 dB.*

- For the third experiment, the probability distributions for STRS were built on the first corpus, as shown in Fig. 4.11 (a), (b), (c) și (d).

The consequence of proposed matching method, shown in Fig. 4.11, is that for SNR values of at least –25 dB, the sequence match probability distributions in the two forensic hypotheses used are *fixed*. Un such circumstances, the strength of the sequence match evidence could be computed as a likelihood ratio,

$$\mathrm{LR} = \frac{\mathrm{pdf}(E|\mathrm{H}_0, D)}{\mathrm{pdf}(E|\mathrm{H}_1, D)}, \tag{4.12}$$

where H_0 și H_1 are the forensic hypotheses, and D is the set of context data.

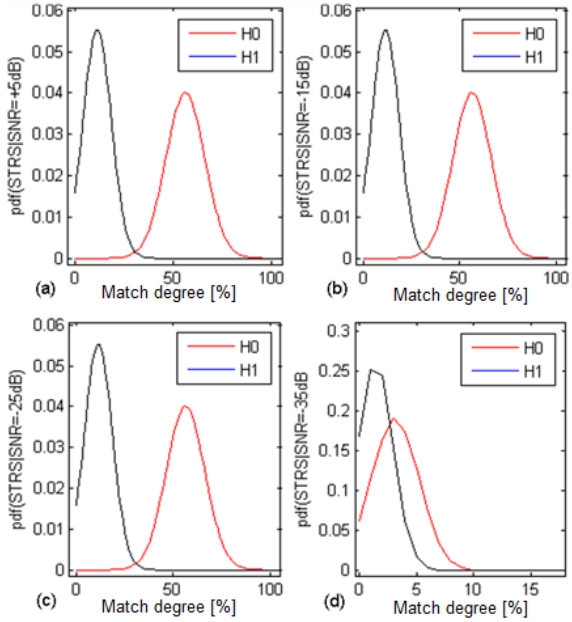

Fig. 4.11 *STRS probability distributions in both forensic hypotheses, for SNR values of* (a) *5 dB,* (b) *–15 dB,* (c) *–25 dB, and* (d) *–35 dB.*

- In experiment 4, both the reference and proposed frameworks were used on the variable noise corpus, and a comparison was made based on the ROC curves in Fig. 4.13. It is clear that for variable ENF trace quality, proposed framework has higher accuracy than the reference one, allowing for the match of such traces with a minimum false positive rate.

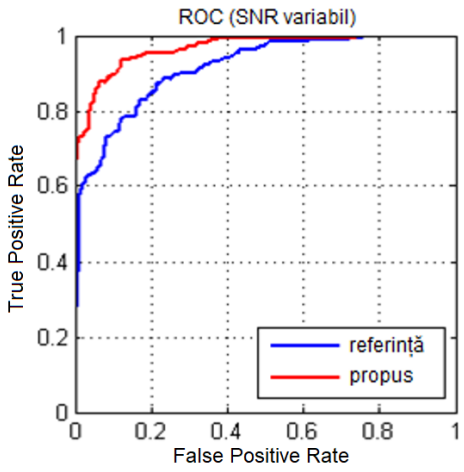

Fig. 4.13 *ROC curves for SNR values variable between –45 dB and –25 dB.*

4.5 Conclusions

In this chapter, original contributions of the author to ENF analysis were presented, regarding all the three stages of the criterion:

• a method for ENF reference database build;
• a method for accelerated match search; and
• an evaluative framework for applying the criterion.

On the path to these contributions, the author of the thesis has also proposed:

- a similarity measure for ENF sequence shapes – *relevant similarity* (RS) – either globally or on a short term basis; and

- a method to determine the strength of sequence match evidence, based on applying the RS measure on short time intervals (STRS).

5. Audio recording authentication through compression analysis

Lossy compression in recording equipments is usually applied by *coders* on an analysis frame basis. The restauration of the signal is done with *decodors*. Various coders, including the *Adaptive Multi-Rate* (AMR) coder, are able to adapt with transmission channel width variations or with the limited storage available.

5.1 Literature overview

Research papers presented in literature on simple or double AMR compression have approached both the waveform analysis and the compressed signal as transmitted data packs.

In [Luo16] a compression detection system is described which examines the Sample Repetition Rates (SRR) in order to decide if the signal was previously AMR-compressed or not.

The author proposed in [Dra14a] and [Dra15] a holistic algorithm for codec recognition based on specific features, by analysing signals in decompressed format. The GMM-UBM paradigm was used in order to detect the traces of codecs like AMR, G.729, MP3 or WMA in uncompressed fiormat signals.

A DNN-based method for AMR-NB compression detection and bitrate recognition in recorded speech was described in [Shi18]. For each signal frame, 184 featrures were extracted, consisting in speech distortion-sensible features, such as the LPC coefficients of order 10, the first 12 mel-cepstral coefficients and the Zero-Crossing Rate (ZCR).

5.2 AMR codec recognition

The general aim in forensic expertise of AMR compression analysis is the recording authentication, by:

• detection and identification of compression artifacts;
• bitrate verification; and
• data flow integrity verification after decompression.

Speech is recovered at the receiver using the *ACELP synthesis model*, shown in Fig. 5.1. Source and filter approximation errors rise with the decrease of available bits, that is, with the increase of compression degree.

For every signal *frame* (20 ms), the compression *mode* (expressed as an integer value between 0 and 7) is decoded from data packets received, followed by *the index of curent quantized Line Spectrum Pairs (LSP)*, which are then interpolated for all four subframes.

For every *subframe* (5 ms) remaining indexes are decoded and a subframe of signal is synthesized. Before delivery at the output, synthesized speech is *postfiltered*, in order to put an emphasis on formant structure, compensate for spectral tilt, and remove components lower than 60 Hz. The recognition of AMR codec in this chapter is based on using neural networks to classify feature vectors extracted from 20 ms long signal fragments, in steps of 5 ms. In order to train the neural networks and test the propised method, high quality speech files were used, both unmodified and AMR compressed.

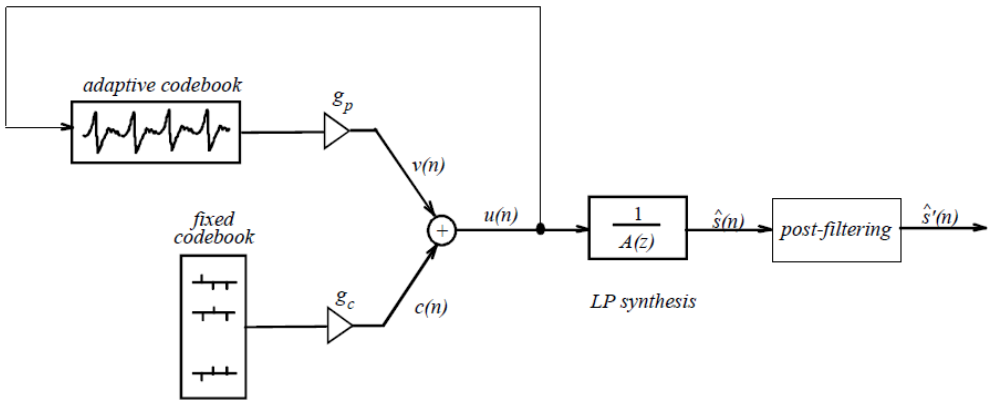

Fig. 5.1 *Block diagram of ACELP speech synthesis în AMR decoder* [3GP02].

In order to concentrate the relevance of input data in a reduced number of features, I have elaborated enhanced feature vectors, with 176 componenta per frame.

During the research presented in this chapter, two AMR codec recognition experiments were conducted, based on using four layer neural networks, as ilustrated in Fig. 5.4, with one classification decision per signal frame.

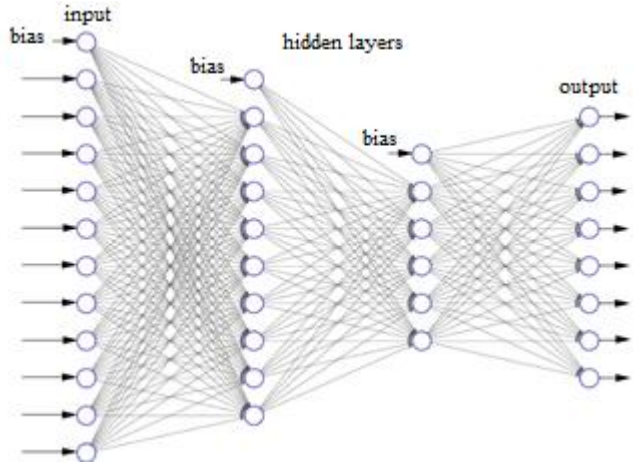

Fig. 5.4 *Schematic of neural networks used in AMR codec recognition.*

- The first experiment used a 2-output neural network (“compressed”/“uncompressed”). By analyzing subsequences of 20 to 500 labels (0.4 s – 10 s) an accuracy of 100% was reached, starting from sequence lengths of 80 frames (1.6 s).

Proposed method outperformed the methods described in [Luo14] (based on MDCT features) and in [Shi18] (DNN-based), **as the shorter label series needed for detection allows for action on shorter intervals than reference methods**.

- For the second experiment, an 8-output (“0”, “1”, ..., “7”) neural network was trained, as the signal was already known to be AMR compressed.

By modifying the lebgth of analyzed intervals, **the accuracy of bitrate recognition** varies, as shown in Table 5.1.

Table 5.1 *Accuracy variations by fragment length of bitrate classification.*

***F* [frames] (Length [s])**	**20 (0.4)**	**200 (4)**	**500 (10)**
Accuracy [%]	23.2	72.1	**78.0**

Peak accuracy obtained at AMR compression bitrate recognition was **78%**, for subseries of 500 enhanced feature vectors (or 10 s). Given that reference methods do not explore the dependence of accuracy on the length of the signal fragment analyzed, the performance of proposed method is compared to others in Table 5.2, for 4 s fragments only.

Table 5.2 *Bitrate recognizer methods comparison, on 4 s fragments.*

Compared methods	**Accuracy [%]**
Method described in [Luo14] (based on MDCT)	32.4
Method described in [Shi18] (based on DNN)	69.7
Proposed method	**72.1**

5.3 Conclusions

Proposed method has detected AMR compression with maximum accuracy – 100% corect. Also, by an accuracy of **72.1%**, proposed method outscores the reference methods for the task of bitrate recognition.

6. Quality-aware speaker recognition

6.1 The problem of speaker recognition

From the very beginning of first technical speaker reconition methods, the possibility of error was aknowledged, caused by random similarity between voices. Large scale experiments conducted by academic institutions have made clear the extension of such methods and provoqued ,measures to be taken by Courts of Law in relation evidence admissibility.

Na important problem of speecj signals used in *Forensic Automated Speaker Recognition* (FASR) is their conformance with both qualitative and quantitative criteria, set forth in order to avoid obtaining uninterpretable results.

6.2 Typical FASR system

An important FASR system category is the one base on GMM-UBM paradigm. These include a *preprocessing* stage, where MFCC features are extracted from signal frames of 15-30 ms with 10-15 ms overlap. The Expectation Maximization Algorithm (EM), detalied in subsection 3.3.2, examines multi-dimensional data points and identifies clusters they form.

The Maximum A Posteriori Algorithm (MAP), also detailed in subsection 3.3.2, makes use of statistic relevance of the background model, and adapts it to a dataset specific to one speaker.

In Fig. 6.2 the bloc diagram of a tiypical FASR system is ilustrated, which uses the GMM-UBM paradigm. After initialization (setting up the UBM), the FASR system is ready for use by running the stages of *enrollment* for each speaker. For the Bayesian inference, the LR value at the test stage is considered as *evidence*.

6.3 Proposed quality-aware FASR system

In this section, the proposed forensic speaker recognition system, based on GMMs, is described. A speaker recognition system becomes *quality-aware* **if it gets a schema for recognition decisions assistence, based on input signal quality data**. While in telecommunication speech quality is considered in relation to the average listener, Forensic experts define it according to the task at hand, and the biometrics define a *quality measure* as „information which contributes to establishing the probability that a biometric decision is correct.”

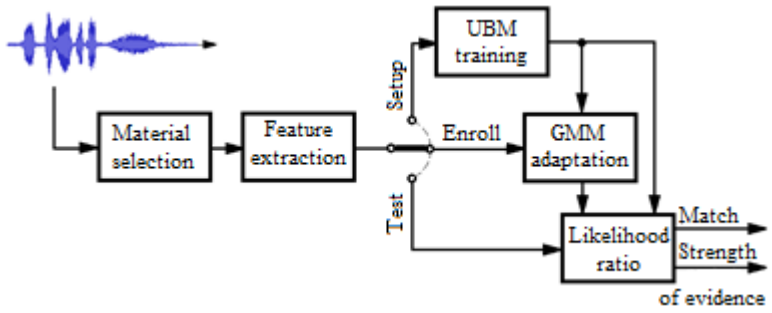

Fig 6.2 *Block diagram of a GMM-UBM based FASR system.*

In a FASR system having a UBM and a speaker model with *M d*-dimensional Gaussian densities, conditions are created for determining the contributions of each signal frame to the likelihood ratio, either liniar or logaritmic.

6.4 Evaluation of proposed FASR system

Proposed FASR system, made quality-aware with thw quality measures prezented in the thesis, has a block diagram like the one in Fig. 6.4. As related to the typical system, the proposed system, described in section 6.2, has three quality information awareness channels:

1. *The selection of input material*;
2. *The computation of matchind score*; and
3. *The structure of speech feature vectors*.

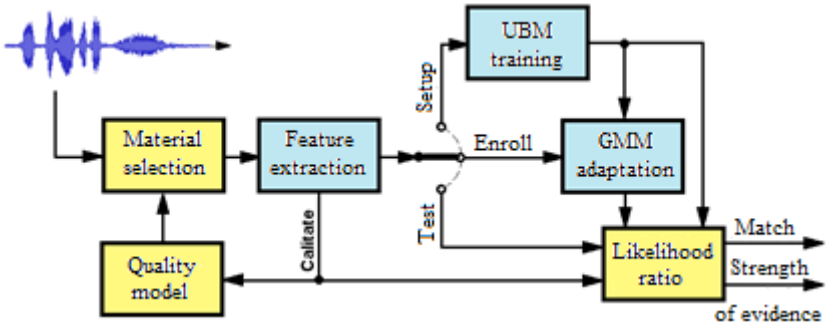

Fig. 6.4 *Bloc diagram of proposed, quality-aware, FASR system.*

The three FASR configurations were comparate by *accuracy* and the *equal error rate* (EER). For a synthetic evaluation, the *confusion matrix and* the DET curves were used.

Audio speech files, used in theining the background model, individual models and for evaluation are from the *test* section of NIST2008 SRE [NIS11] dataset. From the 942 hours of speech, offered by 1460 individuals, both sexes, 368 hours are landline phone recordings, while the other 574 hours are studio microphone recordings.

Speech from 420 persons was used to build a UBM, and two files from 39 of the remaining speakers were used for enrollment and testing.

- The first experiment has evaluated the effect of best material selection on speaker recognition. The selection criterion was set up by imposing a set of thresholds to mormalized quality measures, as described in the thesis. The confusion matrix, in a non-standard form, is shown in Fig. 6.5. The accuracy of the proposed system was 93.1%, with an equal error rate of 4.7%.

The improvement of the EER, as compared to the case of not using the quality data was **0.9%**, still outperforming the methods described in [Gar06] and [God93].

- A second experiment was dedicated to weighting contributions of feature vectors in forming matchin score with quality measure data. Such configuration of the proposed system has also been evaluated by the author in [Dra14b], for the case of co-channel speech.

The confusion matrix for configuration 2 is shown in Fig. 6.6.

Attained accuracy for configuration 2 was **99.08%**, with an equal error rate of **2.0%**.

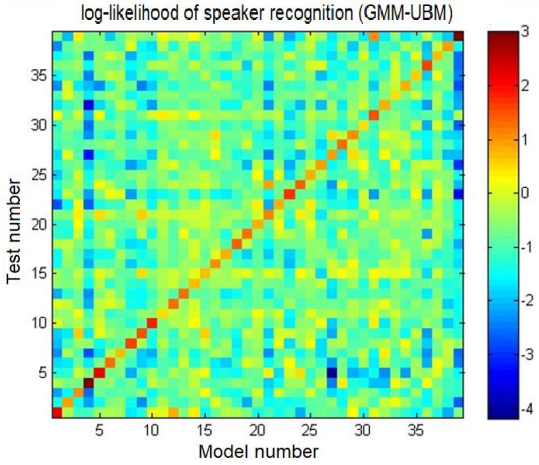

Fig. 6.5 *Confusion matrix of proposed system in configuration 1.*

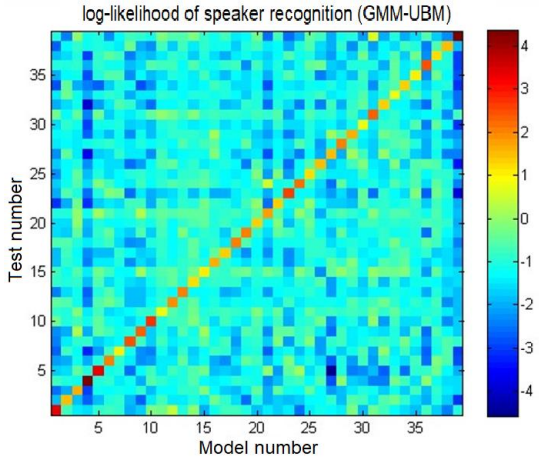

Fig. 6.6 *Confusion matrix of proposed system in configuration 2 (score weighting).*

- The extension of feature vectors has imposed a new initialization, with training a new, quality-aware UBM, from the speech files in the same corpus, as well as preparing new individual GMMs for known speakers, and the confusion matrix of the system in the third configuration of proposed FASR, is shown in Fig. 6.7.

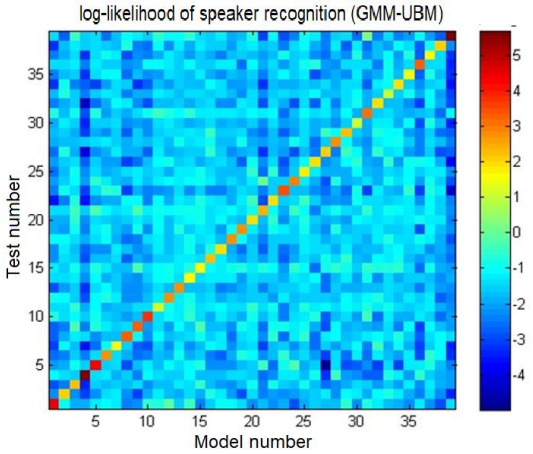

Fig. 6.7 *Confusion matrix of proposed system in configuration 3 (extended vectors).*

Accuracy of the new system configuration of proposed system was **99.8%**, with an equal error rate of **0.6%**. This is better than the best result obtained by the reference system on TIMIT corpus (EER of 1.01%, after [Sad13]). In Fig. 6.8 the Detection Error Tradeoff (DET) curves are presented comparatively, for the three configurations of the proposed system.

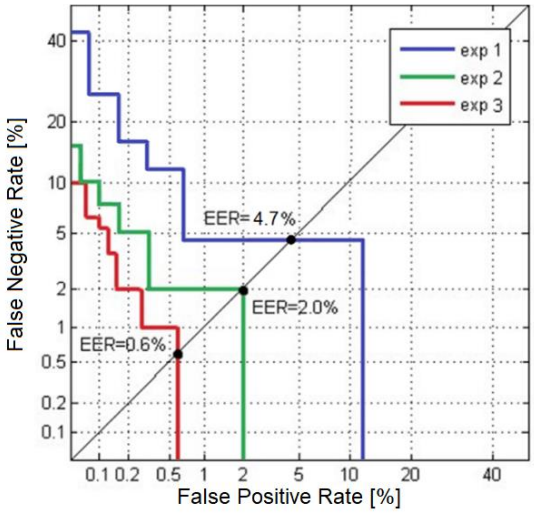

Fig. 6.8 *DET curves of the three experimental configurations.*

6.5 Conclusions

Quality awareness at score computation stage has allowed the FASR system to considerably improve speaker recognition accuracy, from 93.1% to **99.08%**, which is already acceptable for use in forensic expertise. In the configuration with quality measures used at the same time as speaker features, in input material selection and in score computation, the speaker recognition system reached a state-of-the art performance, with an accuracy of **99.8%** and an EER value of **0.6%**.

7. Speech enhancement

7.1 Introduction

In the field of forensic expertise, speec henhancement is defined as "the process of restoring a speech signal in order to make it understandable by humans." This process aims at producing a means of proof, by subsequent transcription of recorded speech, where the main criterion is the level of inteligibility of speech.

Speech enhancement has undergone a long process from the analog graphic equalizers to the use of Deep Neural Networks (DNN).

7.2 Quality-related problems of recorded speech

From the speaker to being stored in a file, recorded acoustic events are processed in three stages: an acoustic one, an analog electric one and a digital one.

The quality of the speech depends on:

- *intrinsic quality of the speech*;
- *source – microphone acoustic channel*;
- *the superposition of target signal with other acoustic signals*;
- *non-linearity of sensitivity curve of the microphone*;
- *analog-digital conversion noise*; and
- *lossy compression noise*.

Speech enhancement methods aim at one or more of the problems listed, and their results are evaluated according to the objective assumed. In case of forensic expertise, the evaluation criterion is the *inteligibility* of speech.

7.3 "Classical" methods for speech enhancement

One of he first sound enhancement techniques was based on the use of an analog device called *graphic equualizer* ("graphic" for the operator to be able to apreciate by just looking at the knobs on its command panel, which is the momentary effect introduced).

Some of the "classical" speech enhancement methods are:

- spectral subtraction (and its various versions);
- optimal filtering (Wiener);

- wavelet decomposition techniques;
- Independent Component Analysis (ICA); and
- using an Ideal Binary Mask (IBM) in time-frequency domain.

7.4 Speech reconstruction using DNNs

Neural networks have demonstrated they can learn complex correspondances between input and output data, and even their own data representations. With a proper training, *a deep feed-forward neural network*, whose structure and learning were discussed in section 3.4, is able to recognize or synthetize speech, or perform various tasks, including the enhancement of recorded speech.

7.5 Proposed quality-aware speech enhancement

The method proposed in this chapter for use in forensic expertise relies on enhancing the speech as a sequence of short time feature vectors.

In Fig. 7.1 the block diagram used is illustrated.

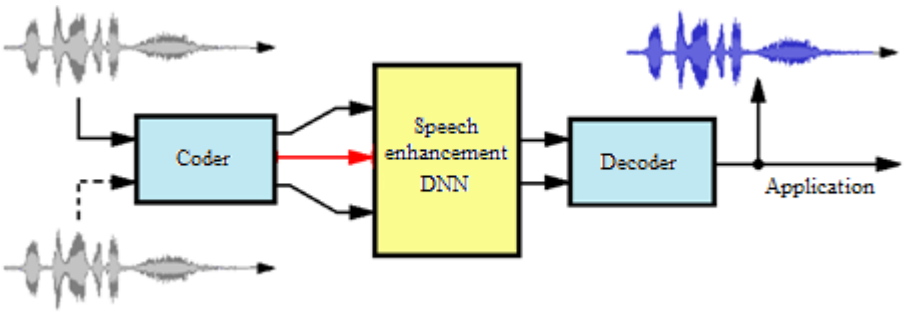

Fig. 7.1 *Speech enhancement block diagram based on proposed method.*

- The Coder module applies Fast Fourier Transform in 512 points each channel, with 32 ms frames and 16 ms overlap. By further processing the amplitude spectrum, 256 *log-spectral features*, 93 *cepstral features* (31 MFCC and 31 orders one and two variations of them), and *quality features*.
- *Proposed* DNN for speech enhancement has a feed-forward structure, ilustrated in Fig. 7.2, and was trained using the *transfer learning* technique, based on importing parameters from an already trained network, described in [Xu15].

The results of the tests performed are presented in Table 7.1. The values of word error rate (WER) obtained on NSDTSEA and SSC-eval corpora, described in the thesis, were determined using the speech *automatic speech recognition* (ASR) system, described in [Geo18], based on the Kaldi toolkit. In the forensic expertise, the most desirable effects of using a speech enhancement method are the increase in short time objective inteligibility (STOI) and the Itakura-Saito distance (DIS) between enhanced speech and the recorded one.

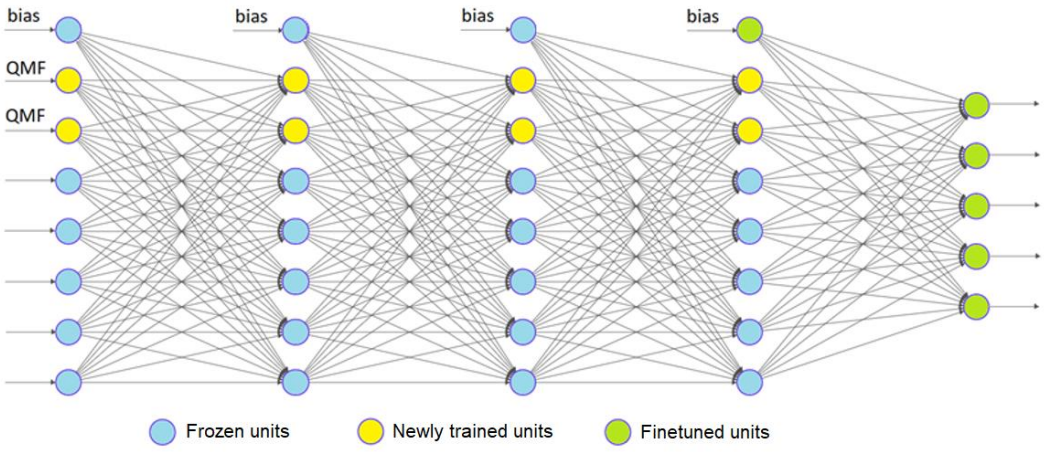

Fig. 7.2 *The schematics of the DNN used for speech enhancement.*

Tabble 7.1 *Comparison of proposed method with others*

Method	WER [%] on SSC-eval	WER [%] on NSDTSEA	PESQ	STOI	DIS
Blind noise estimation, with Bayesian filtering [Doi16]	21.73	9.42	2.47	0.88	53.1
Dilated CNN (WaveNet) [Qia17]	47.32	10.23	1.15	0.06	57.0
Feed-forward DNN in spectral domain [Xu15]	20.68	8.47	2.26	0.87	19.4
Clean signal	20.02	8.60	–	–	–
Proposed method	**20.34**	**8.21**	**2.49**	**0.90**	**16.6**

7.6 Conclusions

From comparisons presented in Table 7.1 it comes that the neural network training with quality measures as supplemental information, and with input signals corupted with rando noise types, contributes to a state-of-the-art generalization capacity to unknown acoustic environments, to unknown speakers and noises.

The comparison results also show that proposed method incurrs the least loss in WER, as compared to clean speech, outperforming the reference methods in all considered criteria, and especially the method described in [Xu15].

8. Conclusions

In this thesis, a series of contributions in the main subdomains of audio forensics are presented, that is, in forensic digital audio authentication, speaker recognition, and speech enhancement.

The introductory part of the thesis describes both the organization of judicial and extra-judicial extertise acrivity, and the scientific basis of the forensic expertise, which comprises anatomy and phisiology of speech production and perception, concepts of statistics and probability, as well as of forensic methodology and reasoning.

Chapters 4 to 7 follow the main objectives of the thesis:

a. the seting up of algorithms for increasing the performance of ENF analysis, in all the three application stages:

1. an efficient method of reference databases collection;
2. accelerated search of references to match the traces; and
3. evaluative reporting of sequence match evidence.

b. the elaboration of a AMR codec trace recognition, using enhanced feature sets;

c. the implementation of automated speaker recognition system using short term measures of speech quality; and

d. a method to enhance recorded speech, on either one or two channels, by using a neural network in the signal spectral, cepstral, and quality feature domain.

8.1 Original contributions of the author

Chapter 4, dedicated to forensic expertise applications of residual electric network frequency traces, hosts the presentation of three main contributions of the author.

1. The elaboration of a method to efficiently build reference databases for the electric network frequency variations, using a versatile collector module in a client-server architecture [Pop17].
2. The elaboration of an accelerated match search between the sequence of frequency values from as trace ENF and the sequence of the reference values for the real electric network frequency variations [Pop17].
3. The definition of a similarity measure between sequences of electric network frequency sequences, so as to allow the evaluative reporting of traces matching the reference, no

matter if the time interval during which the recording was known beforehand or it was founf by match searching in a reference database [Pop18b].

Chapter 5 was dedicated to research on audio recording authentication through compression traces analysis, especially through AMR compression trases, with emphasis on its narrow band version. The author's main original contributions, presented in this chapter, are:

4. The setup of a recognition method for the detection of traces of AMR compression and for recognition of previous compression bitrate, using a deep neural network [Pop19a].
5. The conception and use of a new data type, with primary features obtained by short term analysis of the signal, at the subframe level, and feature vectors obtained by primary feature enhancement on medium term (0.2 s – 4 s) [Pop19a].

Chapiter 6 introduces and evaluates from the viewpoint of speech signal quality an automated system of speaker recognition for the forensic expertise. The original contrinutions of the author, presented in this chapter, consist of:

6. A demonstration of the individual character of speech quality measures, by using them as supplemental components of the extended feature vectors trăsături o a GMM-UBM based FASR [Pop18a].
7. Putting an emphasis on the superiority of the audio input material selection based on speech quality measures, as opposed to the signal selection based on applying a mere voice activity detection algorithm [Pop18a].

In chapter 7 este a speech signal enhancement method is presented for forensic expertise. The original contributions of the author in this chapter are:

8. The enhancement of speech based on both expanding feature vectors at the neural network input with speech quality measures and transferring the learning aquired by another network [Pop19b].
9. The conception of a mixing-demixing system for using the single channel enhancement system with both single and double channel signals [Pop19b].

8.2 Reported results

Chapter 4:

- Relative precision of electric network frequency determination: $\mathbf{1.65 \cdot 10^{-6}}$ [Pop17].
- Absolute precision of electric network frequency: $\mathbf{8 \cdot 10^{-5}}$ **Hz** [Pop17].
- Storage of frequency and duration sequencies in relative terms, so that duration-frequency duality is concentrated **in the sign** of relative deviation [Pop17].
- Collection of each ENF sample on 15 biți, which allows the collected data to fit in **121 MB** per year [Pop17].
- Storage is immune with harddrive write errors [Pop17].
- Database search speed for ENF sequences of 240 to 600 seconds: **2 to 7 times** [Pop17].
- Evaluative reporting of ENF sequence match [Pop18b].
- Capacity to work with variable quality ENF trace recordings [Pop18b].
- Comparison of framework performance using ROC curves [Pop18b].

Chapter 5:

- Minimum fragment length for 100% AMR compression detection: 1.6 s [Pop19a].
- Minimum fragment length for bnitrate recognition: 40 ms [Pop19a].
- Bitrate recognition accuracy for 10-second long fragments: 78.0% [Pop19a].

Chapter 6:

- Quality aware forensic speaker recognition system, with speech quality measures considered as individual features in extended feature vectors, which reached 99.8% acuracy [Pop18a].
- EER on NIST 2008 SRE corpus: 0.6%, state-of-the-art at publication time [Pop18a].
- Implementation of an a posteriori VAD, based on thresholding normalized quality measures, which improves the model of unusable speech [Pop18a].

Chapter 7:

- The enhancement of speech above the intrinsic quality of clean speech [Pop19b].
- Aplicability of proposed method in forensic context, justified by increased inteligibility and other objective speech quality indicators [Pop19b].
- Aplicability for both single and double channel speech recordings [Pop19b].

8.3 List of author publications in the domain of the thesis

- **Gh. Pop**, Ș. Mihalache, and D. Burileanu, "Forensic Recognition of Narrowband AMR Signals," *Proceedings of the 10th International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, 6 p., Oct. 2019, IEEE NY, ISBN: 978-1-7281-0984-8 Indexare IEEE Xplore DOI10.1109/SPED.2019.8906279
- **Gh. Pop** and D. Burileanu, "Speech Enhancement for Forensic Purposes," *UPB Scientific Bulletin, Series C – Electrical Engineering and Computer Science*, Vol. 81, Issue 3, Ed. Politehnica Press, Bucharest, pp. 41-52, 2019, ISSN: 2286-3540. **ISI WOS:000477996400004**.
- **Gh. Pop**, D. Burileanu, and Ș. Mihalache, "An Evaluative ENF-Based Framework for Forensic Authentication of Digital Audio Recordings," *Proceedings of the Romanian Academy Series A – Mathematics, Physics, Technical Sciences, Information Science*, Ed. of Romanian Academy, Bucharest, Vol. 19, Issue 4, pp. 605-612, Dec. 2018, ISSN: 1454-9069.
  ISI WOS:000454140900012 (Q2, IF: 1.752 – Dec. 2018).
- **Gh. Pop**, Ș. Mihalache, and D. Burileanu, "Forensic Speaker Identification Using Speech Quality Data," *Proceedings of the 12th International Conference on Communications* (*COMM*), Bucharest, pp. 509-512, 2018, ISBN: 978-1-5386-2350-3. IEEE Xplore Index (DOI:10.1109/ICComm.2018.8484766). **ISI WOS:000449526000096**.
- **Gh. Pop**, D. Drăghicescu, D. Burileanu, H. Cucu, C. Burileanu, "Fast Method for ENF Database Build and Search," *Proceedings of the 9th International Conference* "*Speech Technology and Human-Computer Dialogue* (*SpeD*)," Bucharest, 6 p., 2017, IEEE NY, ISBN: 978-1-5090-6497-7. IEEE Xplore Index (DOI:10.1109/SPED.2017.7990447). **ISI WOS:000425849600022**.
- **Gh. Pop**, "Digital Forensics," Invited scientific paper (*keynote speaker*) in *The International Conference* "*Modern Forensics Techniques – Possibilities of Access via International Mutual Legal Assistance Procedures*," Poiana Brașov, Nov. 16-18, 2016.
- D. Drăghicescu, **Gh. Pop**, D. Burileanu, and C. Burileanu, "GMM-based Audio Codec Detection with Application in Forensics," *Proceedings of the 37th International Conference on Telecommunications and Signal Processing – TSP 2014*, Berlin, pp. 442-446, 2014, ISBN: 978-1-4799-8498-5 (published in 2015, together with *TSP 2015 Proceedings*). IEEE Xplore Index (DOI:10.1109/TSP.2015.7296421). **ISI WOS:000375231000047**.
- **Gh. Pop**, D. Drăghicescu, and D. Burileanu, "A Quality-Aware Forensic Speaker Recognition System," *Romanian Journal of Information Science and Technology*, vol. **17**, Issue 2, Ed. of Romanian Academy, pp. 134-149, 2014, ISSN: 1453-8245. **ISI WOS:000350281600002**.
- D. Drăghicescu, **Gh. Pop**, and C. Burileanu, "Codec Recognition from Decoded Audio," *Proceedings of the 10th International Conference on Communications* (*COMM*), Bucharest, pp. 37-40, May 29-31, 2014, ISBN: 987-1-4799-2385-4. IEEE *Xplore* Index (DOI: 10.1109/ICComm.2014.6866684). **ISI WOS:000345844600028**.
- **Gh. Pop**, D. Drăghicescu, and D. Burileanu, "On Forensic Speaker Recognition Case Pre-Assessment," *Proceedings of the 7th International Conference* "*Speech Technology and Human-Computer Dialogue* (*SpeD*)," Cluj-Napoca, pp. 169-176, 2013, ISBN: 978-1-4799-1065-6. IEEE *Xplore* Index (DOI:10.1109/SpeD.2013.6682668).
  ISI WOS:000330672700025.

8.4 Development perspectives

The complexity of forensic expertise as a task concerning audio recordings is continuously increasing, mainly because of larger and larger access of public at powerful technical means, with top performance in the field of audio recording and processing. Such societal evolution require that forensic processes minimize their necessary time from new forgery mode discovery to the scientific validation of the recognition method to counter it.

The research efforts must continue, on the examination directions presented in the thesis as well as alonf the following directions:

1. The use in speaker recognition of voice pathology features.
2. Identification of new speech quality estimators.
3. Setting up speciality-based dedicated expert systems, using machine learning.
4. Speech enhancement through synthesis from parameters extracted from contaminated audio.
5. Active authentication by inserting speaker facial images in the audio recording, as filmed while speaking.

Selected references

[Coo09] A.J. Cooper, "An automated approach to the electric network frequency (ENF) criterion – theory and practice," *International Journal of Speech Language and the Law*, vol. **16**, no. 2, pp. 193-218, 2009.

[Doi16] C.S.J. Doire, *Single-channel enhancement of speech corrupted by reverberation and noise*, PhD thesis, Imperial College of London (Great Britain), 2016.

[Dra14a] D. Drăghicescu, **G. Pop**, and C. Burileanu, "Codec Recognition From Decoded Audio," *Proc. of the 10th International Conference on Communications (COMM)*, Bucharest, May 29-31, 2014, pp. 1-4.

[Dra14b] D. Drăghicescu, *Analiza înregistrărior audio și recunoașterea vorbitorului în criminalistică* (*Audio Recording Analysis and Speaker Recognition in Forensics*), PhD thesis, Faculty of Electronics, Telecommunications and Information Technology, University "Politehnica" of Bucharest, 2014.

[Dra15] D. Drăghicescu, **G. Pop**, D. Burileanu, and C. Burileanu, "GMM-based Audio Codec Detection with Application in Forensics," *The 38th International Conference on Telecommunications and Signal Processing* (*TSP*), Berlin (Germany), Jul. 1-3, 2015, pp. 1-5.

[Gar06] D. Garcia-Romero, J. Fierrez-Aguillar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Using Quality Measures for Multi-Level Speaker Recognition," *Computer Speech and Language*, vol. **20**, no. 2-3, pp. 192-209, 2006.

[Geo18] A.L. Georgescu and H. Cucu, "Automatic Annotation of Speech Corpora Using Complementary GMM and DNN Acoustic Models," *Proceedings of the 41-st International Conference on Telecommunications and Signal Processing* (*TSP*), Athens (Greece), Jul. 4-6, 2018, pp. 1-4.

[God93] J.J. Godfrey and E. Holliman, *Switchboard-1, Release 2, Telephone Speech Corpus, LDC97S62*, Philadelphia, PA (SUA), Linguistic Data Consortium, 1993.

[Gri03] C. Grigoraș, "Digital audio recording analysis – The electric network frequency criterion," *Application note AN-4*, Diamond Cut Productions Inc., Oct. 2003.

[Haj12] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," *Proceedings of The Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA*), Hollywood, CA (SUA), Dec. 3-6, 2012, pp. 1-10.

[Iof15] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning* (*ICML*), Lille (Franța), Jul. 7-9, 2015, vol. **37**, pp. 448-456.

[Jac13] D.-O. Jaquet-Chiffelle, *Traces: New Definition*, Introductory Course in Digital Forensic Evidence, School of Criminal Justice, University of Lausanne (Switzerland), https://serval.unil.ch/resource/serval: BIB_DFE9D125DECA.P001/REF, 2019.

[Luo14] D. Luo, W. Luo, R. Yang, and J. Huang, "Identifying compression history of wave audio and its applications," *ACM Trans. Multimedia Computing, Communications and Applications*, vol. **10**, no. 3, pp. 1-19, 2014.

[NIS11] NIST Multimodal Information Group, *2008 NIST Speaker Recognition Evaluation Test Set LDC2011S08*, https://catalog.ldc.upenn.edu/LDC2011S08, Philadelphia, PA (SUA), Linguistic Data Consortium, 2011.

[Pop13a] **G. Pop**, D. Drăghicescu, and D. Burileanu, "On Forensic Speaker Recognition Case Pre-Assessment," *Proceedings of the 7th Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Cluj-Napoca, pp. 169-176, 2013.

[Pop13b] **G. Pop**, *Recunoașterea vorbitorului în criminalistică* (*Forensic Speaker Recognition*), Master's dissertation, masteral program BIOSINF ("Multimedia Technologies in Biometry and Information Security Applications"), University "Politechnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Jun. 2013.

[Pop14a] **G. Pop**, D. Drăghicescu, and D. Burileanu, "A Quality-Aware Forensic Speaker Recognition System," *Romanian Journal of Information Science and Technology (ROMJIST)*, Ed. of Romanian Academy, Bucharest, vol. **17**, nr. 2, pp. 134–149, 2014.

[Pop14b] **G. Pop**, "Posibilități și limite ale expertizei criminalistice a vocii și a vorbirii" ("Possibilities and Limits of Forensic Speaker Recognition"), *Review of Criminology, Criminalistics and Penology – Studies*, vol. **2**, no. 1-2, pp. 178-183, 2014.

[Pop16] **G. Pop**, *Digital Forensics*, Invited scientific paper (keynote speaker), The International Conference "Modern Forensics Techniques – Possibilities of Access Via International Mutual Legal Assistance Procedures," Poiana Brașov, Nov. 16-18, 2016.

[Pop17] **G. Pop**, D. Drăghicescu, D. Burileanu, H. Cucu, and C. Burileanu, "Fast Method for ENF Database Build and Search," *Proceedings of the 9-th International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Bucharest, 6-9 Jul. 2017, pp. 1-6.

[Pop18a] **G. Pop**, Ș. Mihalache, and D. Burileanu, „Forensic Speaker Identification Using Speech Quality Data", *Proceedings of the 12th International Conference on Communications* (*COMM*), Bucharest, Jun. 14-16, 2018, pp. 509-512.

[Pop18b] **G. Pop**, D. Burileanu, and Ș. Mihalache, "An Evaluative Framework for ENF-based Authentication of Audio Recordings," *Proceedings of the Romanian Academy, Series A – Mathematics, Physics, Technical Sciences, Information Science*, Ed. of Romanian Academy, Bucharest, vol. **19**, no. 4, pp. 605-612, Dec. 2018.

[Pop19a] **G. Pop**, Ș. Mihalache, and D. Burileanu, "Forensic Recognition of Narrowband AMR Signals," *Proceedings of the 11-th International. Conference On Human-Computer Dialogue* (*SpeD*), Timișoara, Nov. 18-20, 2019, pp. 1-6.

[Pop19b] **G. Pop** and D. Burileanu, "Speech Enhancement for Forensic Purposes," *UPB Scientific Bulletin, Series C*, Ed. Politehnica Press, Bucharest, vol. **81**, no. 3, pp. 41-52, 2019.

[Pop63] K. Popper, *Conjectures and Refutations: The growth of scientific knowledge*, Routledge and Keagan Paul, London (Great Britain), pp. 33–39, 1963.

[Qia17] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Speech Enhancement Using Bayesian WavNets," *Proceedings of the 18th Annual Conference of the International Speech Communication Association* (*INTERSPEECH*), Stokholm (Sweden), Aug. 20-24, 2017, pp. 2013-2017.

[Sad13] S.O. Sadjadi, M. Slaney, L. Heck, *MSR Identity Toolbox: A MATLAB Toolbox for Speaker Recognition Research, v1.0*, în Microsoft Research Technical Report, Oct. 17, 2013.

[Shi18] S.-H. Shin, W.-J. Jang, H.-W. Yun, and H. Park, "Encoding Detection and Bit Rate Classification of AMR-Coded Speech Based on Deep Neural Network," *IEICE Transaction on Information and Systems*, vol. **E101 D**, no. 1, Jan. 2018, pp. 269-272.

[Xu15] Y. Xu, J. Du, L.-R. Dai, and C.-H.Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. **23**, no. 1, pp. 7-19, 2015.

[Zha10] Y. Zhang, P. Markham, T. Xia, L. Chen, Y. Ye, Z. Wu, Z. Yuan, L. Wang, J. Bank, J. Burgett, R.W. Conners, and Y. Liu, "Wide-area frequency monitoring network (FNET) architecture and applications," *IEEE Transactions on Smart Grid*, vol. **1**, no. 2, pp. 159-167, 2010.