Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 11 (2016) pp. 311-321
© MIR Labs, www.mirlabs.net/jias/index.html

Speech/music discrimination-based audio characterization using blind watermarking scheme

Eya Mezghani[1], **Maha Charfeddine**[1], **Henri Nicolas**[2] **and Chokri Ben Amar** [1]

[1]REGIM: REsearch Groups on Intelligent Machines, University of Sfax,
Route de Soukra, B.P. 1173, 3038 Sfax, TUNISIA
{eya.mezghani, Maha.Charfeddine.TN,chokri.benamar}@ieee.org

[2]LaBRI : Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux,
351 Cours de la Libration, 33405 Talence, FRANCE
henri.nicolas@labri.fr

Abstract: **The continual development of technologies of creation and diffusion has resulted in an enormous amount of multimedia document. Thus, efficient retrieval and management of these audiovisual document is needed. Therefore and in order to operate efficiently, audio and video browsing deploy automatic processes like speech/music detector systems.**

In this paper, a new concept for audio characterization based on speech/music discriminator was presented using a blind audio watermarking scheme. In fact, the class of each audio segment is embedded by watermarking. Consequently, browsing or monitoring systems can recognize audio class, i.e. speech or music, at each moment just by extracting the corresponding watermark.

Keywords: audio analysis, audio watermarking, audio classification, speech/music discrimination, multimedia content retrieval

I. Introduction

In the recent years, the increasing availability of multimedia content was strengthen by the numerous distribution channels for nonprofit or commercial goals. The resulting data wealth require powerful analysis system in order to get useful knowledge destined to users or that will be subsequently utilized by other automatic systems. Like for the other media, audio analysis has been gaining researchers interest during the past decade.

And, various audio analysis tasks were proposed namely speech recognition, speaker identification, sound emotion recognition and music information retrieval. Numerous applications can gather more than one task. Speech/music discrimination is one of the commonly used tasks.

In fact, since the sound stream generally carries both speech and music, many applications like automatic monitoring of radio content, speaker recognition and video browsing exploit speech/music discriminator systems to operate efficiently.

The novelty of our work is that it presents a new characterization strategy based on speech/music discrimination and performed using a blind watermarking scheme. In fact, audio signal is divided into fixed-length segments. An analysis strategy is applied in order to classify this segment into music or speech. The retrieved information on the audio class is then inserted by watermarking. Therefore, browsing systems can use the beforehand analyzed content just by extracting the corresponding watermark.

Digital watermarking was primarily conceived to hide a signature in a host document for security purposes as copyright protection. Later, watermarking use was extended to other applications that will be discussed in the next section.

In our work, with the proposed watermarking scheme targeting content characterization, browsing and monitoring systems will avoid repeating the audio analysis task or the use of annexe files which hold content description. Thus, using the hided information, speech tracks can be easily extracted just by detecting the watermark. Likewise, music excerpts within the host audio can be quickly located using the watermark.

Hence, browsing systems including analysis unit will gain considerably when using this system on computing and time. Similarly, monitoring systems can profit from this scheme to locate speech segment and music one within a very long audio files.

This paper is organized as follows: First, an overview about recent works in speech/music discrimination domain and in audio watermarking is given. Then, the proposed speech/music discrimination-based watermarking scheme is described in detail. After that, some experimentations are presented. And finally, we finish by a conclusion and some perspectives.

II. Literature review

In this section, a literature was given in the audio classification field and more precisely speech/music discrimination recent works was presented. Audio watermarking previous works was also presented in this next subsection.

A. Recent studies on speech/music discrimination

In recent years, more and more researchers invest in audio and video analysis and classification since they are highly demanded in information retrieval. Speech/music discrimination task is one of the highly relevant processing done on both speech and music tracks.

For some applications, audio classification and annotation is the only goal like the case of monitoring radio broadcasts for content type [1, 2]. For other applications, music/speech discriminator is used as a front-end for a downstream application namely music genre classification, automatic speech recognition, etc. [3]

Speech/music classification task was proposed in many previous work with various techniques in the last decade. Feature selection and extraction is the crucial step since performance of the resulting system is closely related to discrimination abilities of the considered features [4].

These features can be classified into two groups: time domain and frequency domain feature according to the computing domain of each metric. Zero-crossing rates, amplitudes and pitches are some of popular time domain audio feature. While, spectrograms, cepstral coefficients and Mel-frequency cepstral coefficients (MFCC) are some of most known frequency domain features. [4]

In some applications, spectral descriptors have achieved better discrimination performance than temporal ones. Classification abilities of descriptor can't be confirmed in absolute but according to the targeted goal. In recent works, various features was selected for the speech/music discriminator.

In [4], audio classification was based on MFCC, detla M-FCC, improvement of MFCC, RASTA-PLP cepstra and 12th order PLP spectra.

A frame-level narrowband speech/music classification was proposed in [5] using combination between line spectral frequencies (LSFs) and zero-crossing-based features. The presented techniques has shown powerful discriminating abilities.

In [6], a low complexity speech/music classification techniques was proposed using only one descriptor only Warped LPC-based Spectral Centroid (WLPC-SC). Gaussian Mixture Model (GMM) classifier was employed which have shown higher performance.

Besides the feature set, the choice of the classifier is very crucial.

In the literature, different supervised classifiers were employed for speech and music discrimination task such as Gaussian mixtures models (GMM) [**?**, 22, 23], k Nearest Neighbor (kNN) [5], Support Vector Machine (SVM) [2,22], etc. Thus, in this paper, we propose to use one of the more powerful classifier which is SVM combined with a set a features.

B. Audio watermarking previous works

In the twentieth century, digital watermarking has appear as a mean to identify musical pieces and to prevent document piracy. Later many approaches for embedding and detecting have been presented.

Over the past ten years, digital watermarking has gained considerable interest, and began to take its place in a some fields of applications. Watermarking is often confused with two other technical terms which are data hiding and steganography.

These three terms have a lot of overlap and share many concepts. However, these terms have few points of difference, both on the design and also on the application constraints.

In fact, Data hiding is a generic term that involves a wide range of issues beyond information burying in a document. It can refer either to the imperceptible information like watermarking or maintain the secrecy of the existence of information as in steganography.

In the other hand, steganography is the art of hiding the existence of a message transmitted via a support which can be a text, an image or an audio stream between a transmitter and a receiver. It concentrates on the mechanisms for rendering the presence of the message secret and undetectable.

However, watermarking is a technique to embed a signature or a mark with special information, generally related to the carrier signal.

The watermarking can be used in several types of applications that target two different contexts; the first to prevent from documents piracy and the second for the data transmission.

In the security context, watermarking aims to adapt the inserted mark according to the action on the document held by the hacker. In this case, integrated information within the original document must be robust toward different intentional piracy attacks. Among the security applications, we find :

- The owner identification : the embedded watermark in the original document hold information on copyright. The watermarking had to be very robust and secure, allowing the owner to justify the presence of this watermark in the case of property dispute.
- Proof of ownership : it is also possible to use the watermarking not only to identify the ownership of copyright, but as a real proof of ownership.

The problem arises when an opponent attempts to replace the reference to the original copyright by another, and then claims to own the copyright. In this case, instead of having direct proof of ownership by inserting, for example, another signature in the original document, the algorithm will instead try to prove that the opponent document is derived from that watermarked.

- Integrity : the signature is embedded in the original document, and is used more-later to check if the content has been altered or not. If we take as an example the recording of the speech, it would be easy to remove part of a recorded phrase.

This can completely change the understanding of a whole speech. Thus, to avoid this illegal treatment, inserts a mark in the document so that if we remove a part of the sentence, part of the watermark will also be removed and this will prevent the correct detection. If the mark is not detected, we can conclude that the document was modified.

- Traceability or transactions tracking: also known as fingerprint. watermarking is used here to trace the sender of the multimedia document copy. The idea is to use a specific mark for each copy. In this case, if there is an illegal copy in the market, we can easily identify the person who has distribute this copy. [8]

In the non security applications, watermarking consists in transferring additional information in the digital document.

Although the robustness against intentional attacks is not required, a certain amount of robustness against licit disturbances as the compression is necessary.

In such applications, the watermark should generally contain a high capacity informations and must be extracted using a blind detection algorithm. Among the applications for data transmission, we find :

- Broadcast control: this application is regarded as a non security context since the watermark does not serve as proof of ownership but aim to compile statistics on the use of the document.

In radio broadcasts, advertisers generally want to ensure that their announcements and their advertisements were properly distributed according to the number of times specified in the contract. Thus, a watermark is inserted in each advertisement.

The information identified the advertising record when transferred in a broadcast network (radio, television or Internet). It Allows, for example, to know in which radio the audio signal was broad-casted, how often and at what time.

- Separation of musical extracts : a set of information, with certain characteristics, can be extracted from source signals. This information are embedded by watermarking inaudibly in the mixture of source audio signals. After the extraction of this inserted watermark, the retrieved information allows us to the separation of original music signals. [24]

- Increase the intelligibility of television programs: it is an application that works in real time which aims to replace the teletext display by inserting a cloned into the television programs.

This will enable deaf and hard hearing people to improve their understanding thanks to the movement of a face and hands that reproduce the Cued Speech. [25]

- Sound documents Annotation: This application can be used to transfer a label to help signals indexing. The hidden information can involve meta-data describing the signal content or information concerning a target application. For example, the hidden message can indicate the name of the artist, the place of registration or any other data relating to the signal. [9, 10]

According to the targeted applications, the hided information varies from one technique to another. In state of the art, some watermarking schemes consider a bit stream as a signature.

In fact, these techniques are proposed without focusing on the applicative goal but presented to solve some weakness of previous works: the problem of synchronization loss, [27] low pass, resampling and MP3 attacks [28] and also targeting higher perceptual quality [26, 27].

Others schemes consider textual signature to be hided in the host signal [11]. These works are generally presented for ownership proof or identification by embedding the author name.

Binary image [14] or gray scale images [13] are also used in some approaches as a watermark. These works hided mostly the owner logo [13] or qr-codes [8].

The novelty of our paper is the watermark type which contain information about the host audio signal. And more precisely at each moment the watermark informs about the nature of the host audio segment, music or speech.

before the embedding stage, most of watermarking techniques propose a watermark preprocessing step generally used to enhance the robustness and the efficiency of the proposed scheme. Depending on the targeted application, the author may suggest one or more operation on this phase. The most popular preprocessing methods are the use scrambling algorithms and the error correcting codes.

Scrambling technique are used to blend the original content and render it meaningless.

It was primarily applied in encryption and digital right management context. And more precisely, the scrambling techniques were performed on digital images. After that, the notion of scrambler device was adopted in digital communications issues in order to ensure the security of the transmitted signals. Then, this concept was applied in watermarking techniques since it affords a good robustness against various attacks and ensure the security.

In fact, even if a malicious attack contribute to extract the embedded watermark, by scrambling it becomes incomprehensible. The scrambling was carried out by different methods. The Arnold transform is one of the widely used techniques in watermarking thanks to its periodicity and its good decentralization.

However, it was noticed that the inverse transform (anti-Arnold) is time-consuming operation. [15, 16]

On the other hand, error correcting codes have shown their efficiency. It was shown that cyclic codes like Hamming code, BCH and Reed Solomon codes are very powerful in detecting and correcting erroneous bits frequently occurring after malicious attacks. [17]

In this paper, error correcting code is applied to the constructed watermark and more precisely a Hamming encoder.

III. Music/Speech content-based watermarking

In the next subsection, we will describe in detail our watermarking scheme beginning by the watermark construction basing on music/speech discriminator system and ending by the different steps of watermark embedding.

A. Audio feature extraction

Feature extraction is a crucial step in audio analysis. In fact, the main purpose is to select a set of features and calculate their values for the training dataset.

The chosen features should be the most informative and having high discrimination abilities according to the targeted audio classification task.

Since it's hard to work directly on the original data, which is in our case the audio signal, it's necessary to reduce the data volume. So, audio features are extracted for the reason that we need to get more compact data representation using the properties of the considered audio signal.

As shown in the figure 1 and before features extraction, the audio signal is divided into non-overlapping short-term frames.

After the blocking step, a Hamming window function is applied on each frame in order to avoid discontinuities at block boundaries.

The output signal Y(n) after hamming window is given by the following equation :

$$Y(n) = w(n)x(n) \tag{1}$$

Where x(n) is the original signal and w(n) is the Hamming window given by the formula :

$$w(n) = 0.54 - 0.46cos(2\pi\frac{n}{N}), 0 \leq n \leq N \tag{2}$$

Where N represents the frame samples number.

After windowing, the selected features will be computed per frame. Extracted features can be categorized according to the computation way into time-domain and frequency-domain features. In the next subsections, we will detail each one of these features.

1) Time-Domain Audio Features

Time-domain audio features are computed directly from the audio samples. The most used time-domain features (Short-term energy, zero crossing rate and energy entropy) will be defined and described in more detail in the next subsections. These features will be utilized in the feature extraction stage of our scheme since they afford a good and simple mean for audio signals analysis.

III-A.1.a Short Time Energy

The short-term energy is a time domain audio feature which was calculated according to the following formula:

$$STE = \frac{1}{W_l}\sum_{n=1}^{W_l} \mid (x_i(n)) \mid^2, \tag{3}$$

Where: i is the frame index within the audio signal,W_l is the frame length and $x_i(n)$ are the audio samples.

In general, short-term energy present high fluctuation over speech frames unlike music clips and changes quickly between high and low energy values since speech signals hold weak phonemes and short silence periods between spoken words.

III-A.1.b Zero-Crossing Rate

The Zero-Crossing Rate (ZCR) compute the rate of samples sign changes during the frame and is normalized by dividing it with window length in order to remove the dependency on the frame length. The ZCR is calculated according to the formula below :

$$ZCR = \frac{1}{2W_l}\sum_{n=1}^{W_l} \mid (sgn[x_i(n)] - sgn[x_i(n-1)]) \mid, \tag{4}$$

where the sign function, sgn(.), is given by the following equation :

$$sgn[x_i(n)] = \begin{cases} 1 & \text{if } x_i(n) \geq 0, \\ -1 & \text{if } x_i(n) < 0. \end{cases}$$

The ZCR show the level of noisiness of a given signal. In fact, noisy recorded audio signals correspond to high values of ZCR.

III-A.1.c Entropy of Energy

In order to calculate the energy entropy, short-term frames are first divided in K fixed duration sub-frames. After that, we calculate e_j, the probability of each sub-frame, as the quotient between the sub-frame energy and the total energy of sub-frames of the considered short-term frame i. Finally, the energy entropy is given by the equation below :

$$Entropy = -\sum_{j=1}^{K} e_j log_2(e_j) \tag{5}$$

where :

$$e_j = \frac{E_{subframe(j)}}{\sum\limits_{k=1}^{K} E_{subframe(k)}} \tag{6}$$

the energy entropy feature measure sharp changes in the energy envelop of the corresponding signal.

2) Frequency-Domain Audio Features

In order to ensure accurate audio analysis, it is necessary to combine time-domain features and frequency-domain one, called also spectral features.

These metrics are computed using Discrete Fourier Transform (DFT) coefficients of the considered audio frame.

III-A.2.a Spectral Centroid

The spectral centroid (SC) measures the center of gravity of the spectrum. The spectral centroid characterize the sound brightness and is computed as following:

$$SC = \frac{\sum\limits_{k=1}^{Wf_L} kX_i(k)}{\sum\limits_{k=1}^{Wf_L} X_i(k)} \tag{7}$$

III-A.2.b Spectral Flux

The spectral flux (SF) defines the spectrum amplitude change between two successive frames and is computed by the following equation :

$$SF = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2 \tag{8}$$

where $EN_i(k)$ is the kth DFT coefficient at the current frame defined as following :

$$EN_i(k) = X_i(k)/\sum_{l=1}^{Wf_L} X_i(l) \tag{9}$$

III-A.2.c Spectral Rolloff

The spectral rolloff determines Cth percentile of the power spectral distribution. It defines the frequency value below which the spectrum magnitude distribution is concentrated.

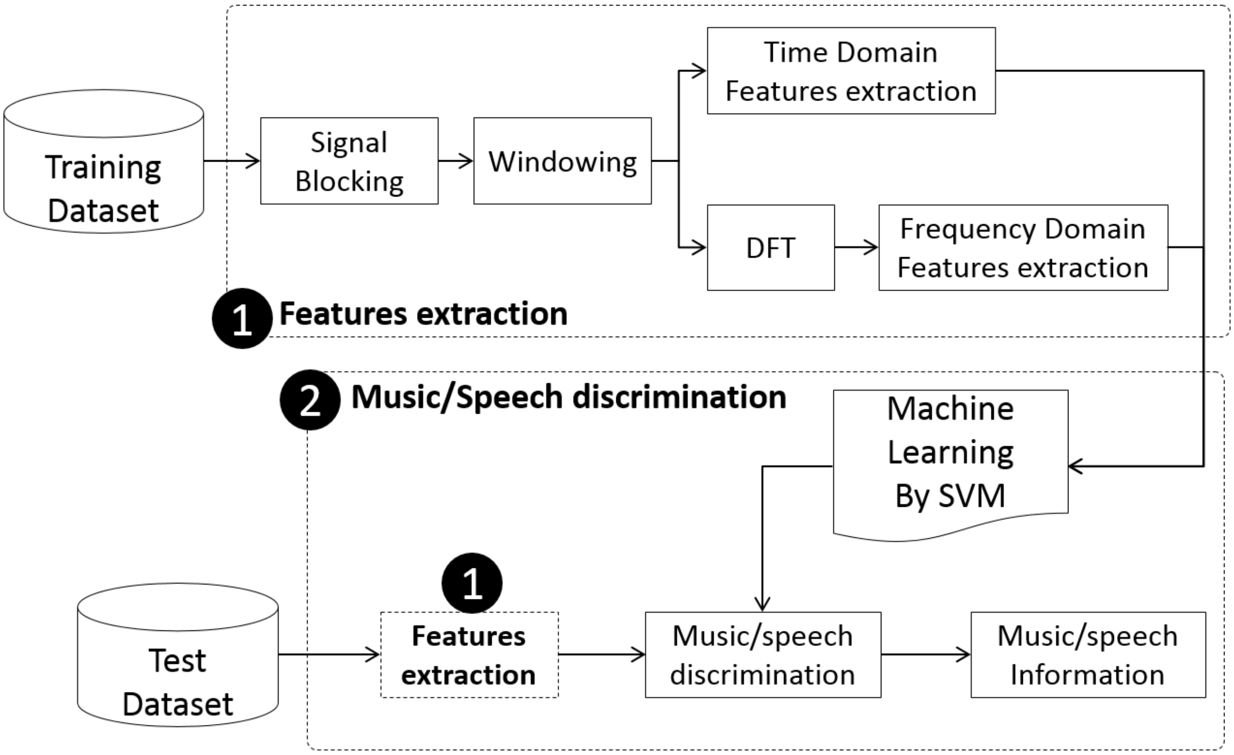


Figure. 1: Music/Speech discriminator scheme

The C parameter value is generally token around the 90%. The mth DFT coefficient of the ith frame is the spectral rolloff which satisfies the following equation:

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{Wf_L} X_i(k) \qquad (10)$$

This feature characterizes the spectral shape. Il was shown it is highly efficient for discrimination between speech and other audio classes.

It can be useful for voiced and unvoiced sounds discrimination and also music genre classification.

III-A.2.d MFCCs

Mel-Frequency Cepstrum Coefficients (MFCCs) is famous feature used especially in speech processing field and also in music/speech discrimination.

MFCCs are computed on the cepstral representation of the signal, where spectral bands are given by the mel-scale.

The relation between Melfrequency and linear frequency is given by the following formula:

$$M(f) = 1125log_{10}(1 + \frac{f}{700}) \qquad (11)$$

Where f is the frequency.

B. Audio classification

After audio features extraction step, we move to the classification stage which consists in assigning the considered signal into classes.

We can distinguish between two types of classification algorithms supervised and unsupervised one.

Supervised classification algorithms use labeled training set in order to learn and establish the decision rule. So, classes are defined at the beginning. Whereas unsupervised classification process assign input data into clusters according to their corresponding characteristics vector without having prior knowledge of the classes number.

Support Vector Machines (SVMs) classifiers have been employed in numerous machine learning fields and have shown their efficiency.

In our case, we have two classes which are music and speech. As illustrated in the figure 1, support vector machine (SVM) classifier was employed firstly on the training dataset which was already manually classified.

Support vectors are obtained from SVM machine learning and will be used for the music/speech discrimination task (2).

C. Watermark embedding process

As already mentioned, our watermarking scheme was proposed to characterize the host audio signal at each moment and more precisely will inform about the audio class: music or speech.

Therefore and after detailing the music/speech discriminator bloc, we move to describe the embedding techniques as illustrated in the figure 2.

The original audio signal is first divided into fixed length segment. Each segment is analyzed and classified into speech or music class. Each segment will be then characterized by its corresponding content based watermark.

Recall that before the embedding, the watermark undergoes a pre-processing step by dividing it into 8 bits sets and performing a Hamming coding.

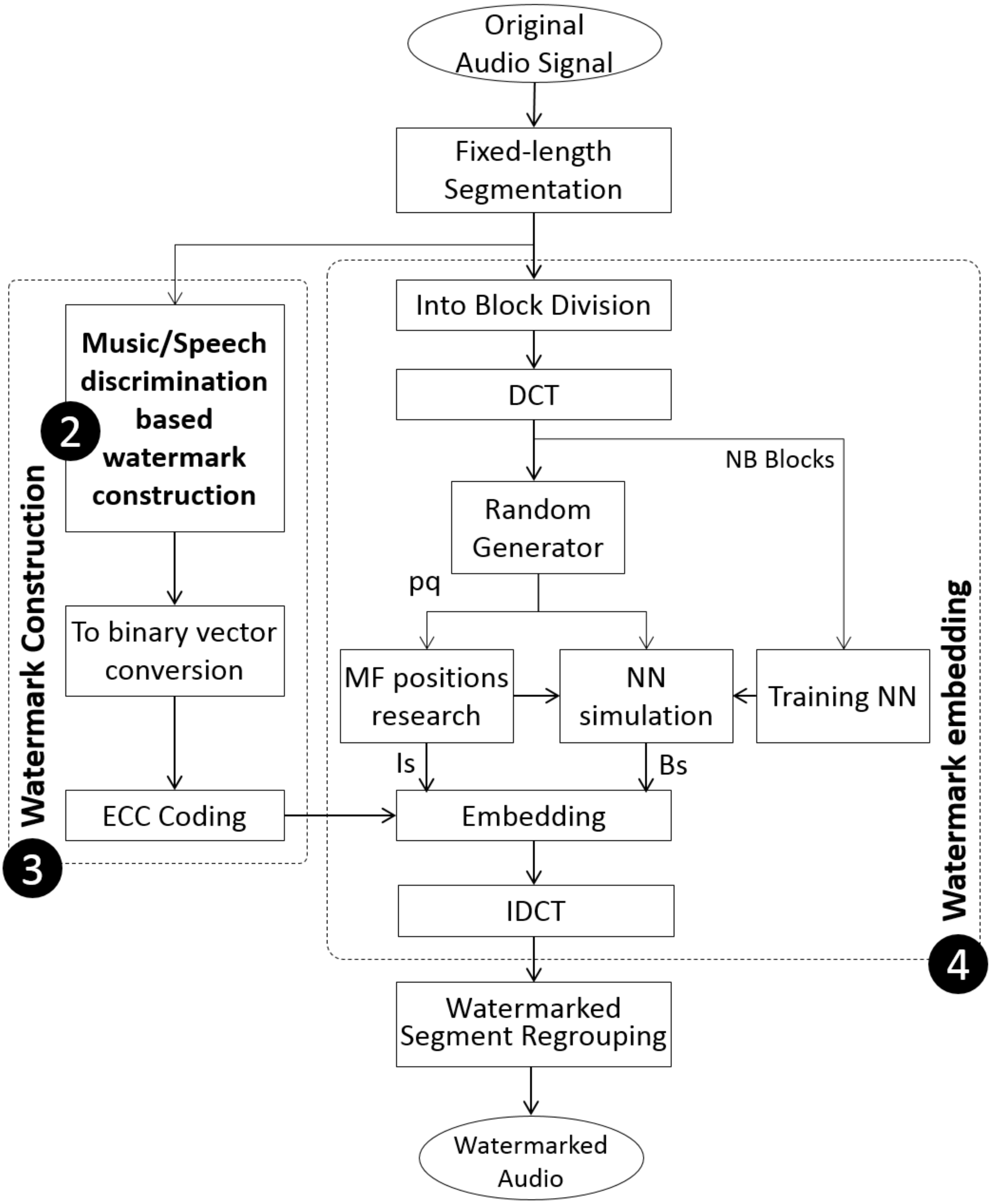


Figure. 2: Watermark embedding process

So, after signal segmentation, each segment is itself divided into 512 samples length blocks. After that, DCT coefficients are computed for each block.
A stage of training using neural network (NN) is performed. A set of synaptic weights, which characterize the behavior of the trained NN, are obtained and will be used in the NN simulation step.
In order to ensure the transparency of our watermarking scheme, low frequencies can't be considered as embedding region since human auditory system (HAS) is more sensitive to this band.
High frequencies are also inappropriate to hide our mark since this band will be considerably altered by compression. Therefore, middle frequency (MF) band ranging from 4 to 10 kHz seems to be the best embedding region. [12]
Each watermark bit is inserted by altering the frequency sample of the selected block located in the middle frequency band.
So, the watermark insertion is performed by comparing the NN output value (Bs) and the original central sample value at the calculated position (Is).
Finally, the watermarked segment is obtained by performing an inverse DCT transform. And the watermarked audio signal is given by regrouping the totality of the watermarked segments.

IV. Experimental Results

In this paper, reported Experimentations uses public dataset: GTZAN for the training and for testing stage, music and speech corpora downloaded from the net. The training data consisted of speech and music audio recordings got from the publicly available GTZAN music/speech dataset.
The GTZAN corpus was collected for music/speech discrimination purposes and consists of 64 speech tracks and 4 music excerpts, each 30 seconds long. The tracks are sampled at 22050Hz and are mono 16-bit audio wav files.
This dataset is diversified and contains different music styles as well as speech which was recorded in various conditions. Many previous studies was carried on this corpus [18, 19].
The performance of our classification method was tested on different dataset. Music dataset downloaded from [1] holding various music genre (rock, pop, jazz, etc). The tracks are sampled at 44100Hz and converted from MP3-128k compressed files to wav format.

A. Classification evaluation

In order to assess the performance of our music/speech discriminator, recall and precision metrics are used.
The recall Re(i) determines the fraction between data with true class i by correctly classified to class. It is computed as following :

$$Re(i) = \frac{CM(i,i)}{\sum\limits_{m=1}^{N} CM(i,m)} \tag{12}$$

Where CM(i, m) is the number of all samples belonging to class i.
The precision Pr(i) is proportion of classified data. In other words, it defines the accuracy of classification system by dividing correctly assigned samples by total classified ones to the class i and is computed as :

$$Pr(i) = \frac{CM(i,i)}{\sum\limits_{m=1}^{N} CM(m,i)} \tag{13}$$

The table IV-A confirms the efficiency of the proposed classification scheme by high precision and recall values around 90%.

Table 1: Classification Results

Dataset	Recall (%)	Precision (%)
Speech corpus	96.48	93.56
Music corpus	91.36	95.85
Global	93.92	94.70

B. Watermarking evaluation

After exhibiting discriminator system evaluation, we pass to present some experimental results done a set of the training dataset GTZAN. Four audio files were selected from each class. And corresponding results of inaudibility and robustness was given in next subsections.

1) Transparency results

In order to validate the transparency of our watermarking scheme, we compute the objective metric, Signal to Noise Ratio (SNR) between the original signal and the watermarked one.
In fact, the imperceptibility of the watermark can be confirmed by high SNR values. The SNR formula is the following :

$$SNR = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} \bar{s}(n)^2}{\sum_{n=0}^{N-1} (\bar{s}(n) - s(n)^2}\right), \tag{14}$$

Where: s(n) and $\bar{s}(n)$ are respectively the original audio signal and the watermarked one.
According to the recommendation of IFPI (International Federation of the Phonographic Industry), watermarking scheme transparency is ensured if SNR values is higher than 20dB. [21]
The figure 3 presents transparency results for the considered audio files, and high SNR values (around 45dB) confirm the inaudibility of our watermarking scheme.

2) Robustness results

The robustness of our watermarking schemes is validated by testing the persistence of the watermark after several attacks. We begin by compression attacks and stirmark audio benchmark.
Therefore, we compute the Normalized Correlation (NC) between the embedded watermark "bin" and the extracted watermark "bin" to estimate the rate of correctly detected bit.
NC values ranges between 0 and 1. Closer NC value to 1

[1] http://www-ai.cs.uni-dortmund.de/audio.html

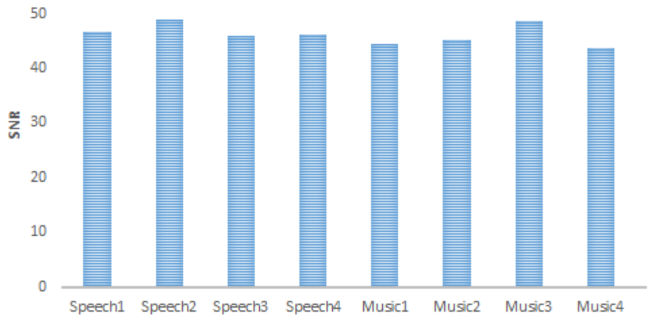


Figure. 3: Transparency results

allow to confirm the good detection and so high robustness against the considered attack.

$$NC = \frac{\sum_{i,j} bin_{i,j} * bin'_{i,j}}{\sqrt{\sum_{i,j}(bin_{i,j})^2 * \sum_{i,j}(bin'_{i,j})^2}}, \qquad (15)$$

As compression is a very common operation that may be applied to the watermarked signal. That's why we should verify whether the mark can be correctly detected even if the watermarked signal was compressed.

Thus, robustness against the most popular encoder which is MP3 was evaluated with the usual bit rate 128k and 96k. NC values are given in the figure 4 informing about the rate of correct detected bit. The retrieved values very close to 1, confirm the performance of our watermarking schemes.

Although our watermarking scheme was proposed for non security purposes, it's recommended to validate the robustness against some attacks. We use Stirmark attacks benchmark [20].

The figure 5 presents the NC values for each audio signal after different attacks. Some alteration is remarked after add noise attack, echo and compressor attacks. But, NC values still very close to 1.

V. Conclusion

In this paper, a new music/speech discrimination-based characterization using audio watermarking scheme.

The hided watermark encloses music or speech class information in time. Audio or video browsing and monitoring systems can benefit from beforehand analyzed content just by extracting the mark.

Experimentations have shown good performance of our system at the classification level and also at the level of watermarking. Thus, we can confirm that this work offers a promising foundation for further work. The watermark can be enriched by embedding more information about the audio content such as the music genre, the speaker, etc.

This work can be also extended by applying this system to audio stream of digital video. Moreover, the music/speech discrimination approach may be enhanced by adding other audio features.

References

[1] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis. A computationally efficient speech/music discriminator for radio recordings. In *ISMIR*, p. 107–110, 2006.

[2] J. Vavrek, E. Vozarikova, M. Pleva, and J. Juhar. Broadcast news audio classification using svm binary trees. In *35th IEEE International Conference on Telecommunications and Signal Processing (TSP)*, p. 469–473, 2012.

[3] G. Sell and P. Clark. Music tonality features for speech/music discrimination. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2489–2493, 2014.

[4] H. Zhou, A. Sadka, and R. M. Jiang. Feature extraction for speech and music discrimination. In *International Workshop on Content-Based Multimedia Indexing CBMI*, p. 170–173, 2008.

[5] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP00*, p. 2445–2448, 2000.

[6] J. E. Munoz-Exposito, S. Garcia-Galan, N. Ruiz-Reyes, P. Vera-Candeas, and F. Rivas-Pena. Speech/music discrimination using a single warped lpc-based feature. In *Proc. ISMIR, vol. 5*, p. 16–25, 2005.

[7] J. M. Exposito, S. G. Galan, N. R. Reyes, P. V. Candeas, and F. Pena. Expert system for intelligent audio codification based in speech/music discrimination. In *International Symposium on Evolving Fuzzy Systems*, p. 318–322, 2006.

[8] F. Chaabane, M. Charfeddine, W. Puech, and C. Ben Amar. A qr-code based audio watermarking technique for tracing traitors. In *23rd European Signal Processing Conference (EUSIPCO)*, p. 51–55, 2015.

[9] E. Mezghani, M. Charfeddine, H. Nicolas, and C. Ben Amar. Audiovisual video characterization using audio watermarking scheme. In *ISDA, 2015 15th International Conference on Intelligent Systems Design and Applications*, p. 213–218, 2015.

[10] G. Tzanetakis. Music information retrieval: theory and applications. In *the 17th ACM international conference on Multimedia ACM*, p. 915–916, 2009.

[11] M. Charfeddine, E. Mezghani, and C. Ben Amar. Modified video watermarking scheme using audio silence deletion. In *ELMAR, 2013 55th International Symposium*, p. 203–206, 2013.

[12] E. Mezghani, M. Charfeddine, and C. Ben Amar. Audio silence deletion before and after mpeg video compression. In *EUROCON*, p. 1625–1629, 2013.

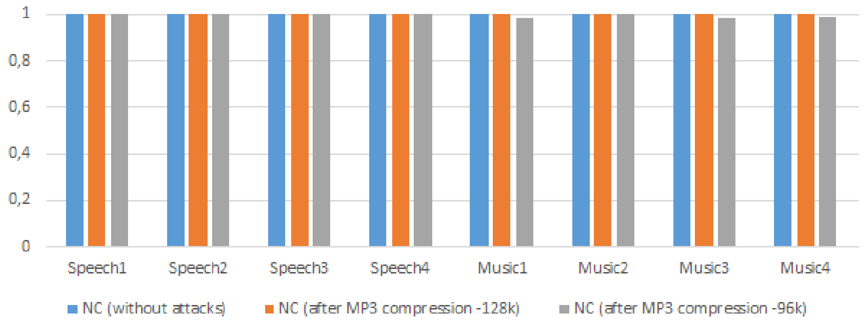


Figure. 4: Robustness against MP3 compression

[13] A. Al-Haj, L. Bata, and A. Mohammad. Audio watermarking using wavelets. In *First International Conference on Networked Digital Technologies NDT09* p. 398–403, 2009.

[14] S. Vongpraphip and M. Ketcham. An intelligence audio watermarking based on dwt-svd using ats. In *WRI Global Congress on Intelligent Systems GCIS09*, p. 150–154, 2009.

[15] L. Wu, J. Zhang, W. Deng, and D. He. Arnold transformation algorithm and anti-arnold transformation algorithm. In *1st International Conference on Information Science and Engineering (ICISE)*, p. 1164–1167, 2009.

[16] V. Veena, G. Jyothish Lal, S. Vishnu Prabhu, S. Sachin Kumar, and K. Soman. A robust watermarking method based on compressed sensing and arnold scrambling. In *International Conference on Machine Vision and Image Processing (MVIP)*, p. 105–108, 2012.

[17] F. Chaabane, M. Charfeddine and C.B. Amar. The impact of error correcting coding in audio watermarking. In *3rd International Conference on Next Generation Networks and Services*, p. 90–95, 2011.

[18] G. Tzanetakis and P. Cook. A framework for audio analysis based on classification and temporal segmentation. In *25th EUROMICRO Conference*, p. 61–67, 1999.

[19] G. Tzanetakis and P. Cook. Sound analysis using mpeg compressed audio. In *International Conference on A-coustics, Speech, and Signal Processing ICASSP00*, p. II761–II764, 2000.

[20] A. Lang, J. Dittmann, R. Spring, and C. Vielhauer. Audio watermark attacks: from single to profile attacks. In *Proceedings of the 7th workshop on Multimedia and security ACM*, p. 39–50, 2005.

[21] M. Charfeddine, M. Elarbi, and C. B. Amar. A new dct audio watermarking scheme based on preliminary mp3 study, *Multimedia tools and applications*, vol. 70, no. 3, pp. 1521-1557, 2014.

[22] B. K. Khonglah and S. M. Prasanna. Speech/music classification using speech-specific features, *Digital Signal Processing*, vol. 48, pp. 71-83, 2016.

[23] E. Didiot, I. Illina, D. Fohr, and O. Mella. A wavelet based parameterization for speech/music discrimination, *Computer Speech and Language*, vol. 24, no. 2, pp. 341-357, 2010.

[24] L. Girin, A. Liutkus, G. Richard, and R. Badeau. Procede et dispositif de formation dun signal mixe numerique audio, procede et dispositif de separation de signaux, et signal correspondant, 2010.

[25] P. Bas, J. Lienard, J. Chassery, D. Beautemps, and G. Bailly. Artus: animation realiste par tatouage audiovisuel ‘a lusage des sourds, *J3eA*, vol. 3, pp. 016, 2004.

[26] O. T. Chen and W.-C. Wu. Highly robust, secure, and perceptual-quality echo hiding scheme, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 629-638, 2008.

[27] S. Xiang and J. Huang. Histogram-based audio watermarking against time-scale modification and cropping attacks. *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1357-1372, 2007.

[28] M. Akhaee, M. J. Saberian, S. Feizi, F. Marvasti et al. Robust audio data hiding using correlated quantization with histogram-based detector, *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 834–842, 2009.

Author Biographies

Eya Mezghani was born in Ariana (Tunisia) in 1985.
She received the Telecom Engineer Diploma in 2009 from the National Engineering School of Tunis (ENIT), Tunisia.
She obtained the Master Diploma in computer sciences in

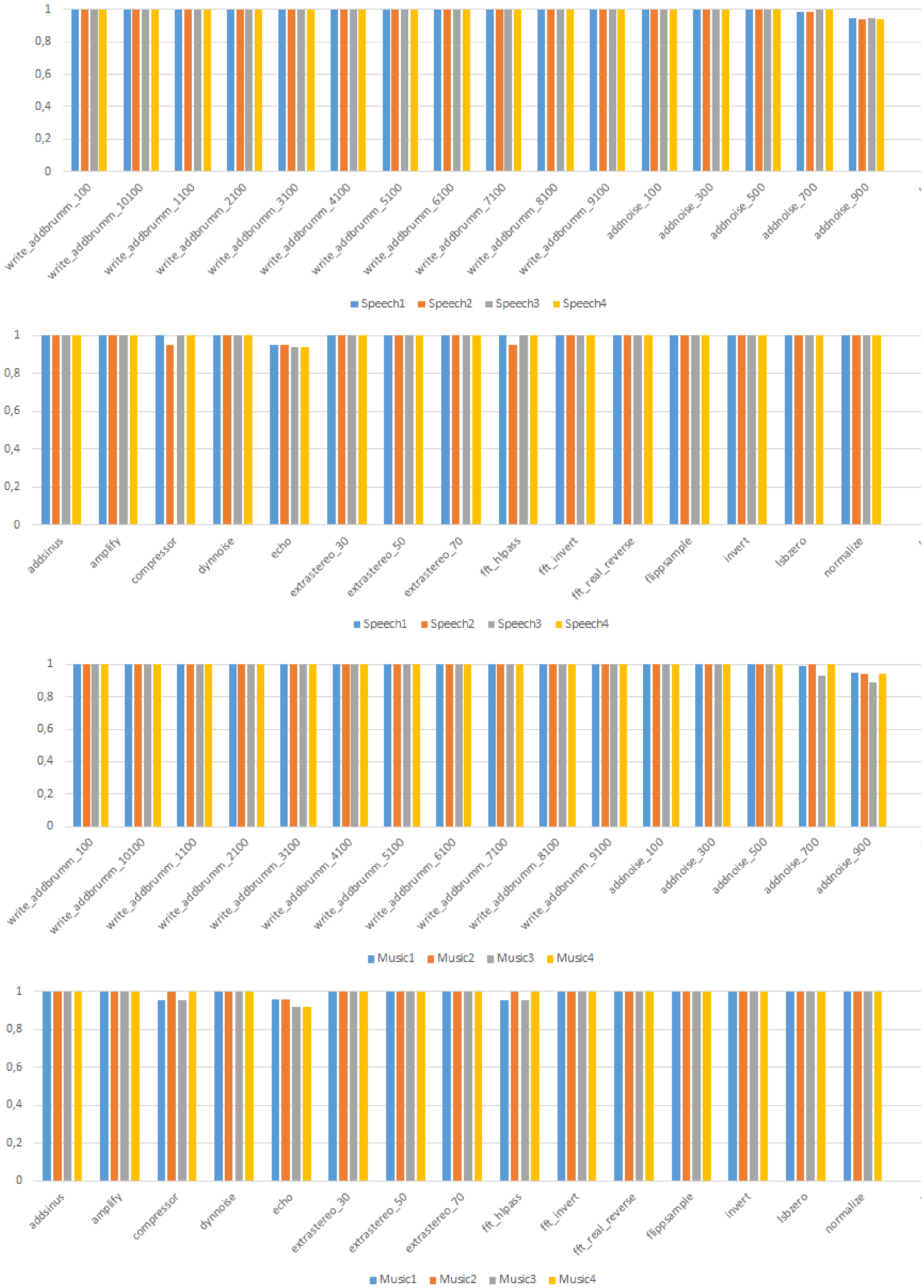
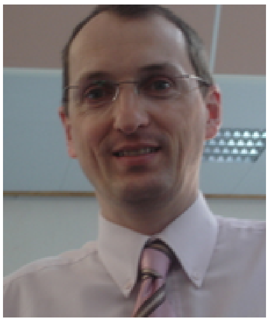


Figure. 5: Robustness against Stirmark Audio Attacks

2012 from the National Engineering School of Sfax (ENIS), Sfax (Tunisia).
She is currently pursuing the Doctor degree in the National Engineering School of Sfax.
She is a research member in the Research Group of Intelligent Machine (REGIM), (ENIS), Sfax (Tunisia).
Her research interests include digital audio and video watermarking, data hiding and signal processing.

Maha Charfeddine was born in Sfax (Tunisia) in 1981.
She received the engineering diploma in computer science from the Tunisian engineering ENIS-SFAX school in June 2005.
In addition, she received the M.Sc. degree in 2007 and the Ph.D degree in Computer Science in 2013 both from the ENIS school.
Nowadays, she is an assistant professor in the Computer-Engineering-and-Applied-Mathematics-Department (ENIS), Sfax (Tunisia) and a research member in the Research Group of Intelligent Machine (REGIM), (ENIS), Sfax (Tunisia).
Her research interests include digital audio and video watermarking, data hiding, traceability, indexation and MPEG video and audio compression.



Henri Nicolas obtained his engineering degree from INSA of Rennes in 1988. He received his Ph-D degree in computer science from the University of Rennes in 1992. For his PhD, he worked at IRISA/INRIA of Rennes (France). From 1993 to 1996, he worked at the 'Swiss Federal Institute of Technology of Lausanne' (EPFL, Switzerland) as the leader of a research group of the Signal Processing Laboratory and as a research team leader for CRAY Research Switzerland. From 1996 to 2005, he works as a senior researcher at INRIA of Rennes. Since September 2005, he is a full professor at the University of Bordeaux 1.

Chokri Ben Amar received the B.S. degree in Electrical Engineering from the National Engineering School of Sfax (E-NIS) in 1989, the M.S. and PhD degrees in Computer Engineering from the National Institute of Applied Sciences in Lyon, France, in 1990 and 1994, respectively. He spent one year at the University of Haute Savoie (France) as a teaching assistant and researcher before joining the higher School of Sciences and Techniques of Tunis as Assistant Professor in 1995. In 1999, he joined the Sfax University (USS), where he is currently a professor in the Department of Electrical Engineering of the National Engineering School of Sfax (ENIS), and the Vice director of the REsearch Group on Intelligent Machines (REGIM). His research interests include Computer Vision and Image and video analysis. These research activities are centered on Wavelets and Wavelet networks and their applications to data Classification and approximation, Pattern Recognition and image and video coding, indexing and watermarking. He is a senior member of IEEE, and the chair of the IEEE SPS Tunisia Chapter since 2009. He was the chair of the IEEE NGNS2011 (IEEE Third International Conference on Next Generation Networks and Services) and the Workshop on Intelligent Machines: Theories & Applications (WIMTA 2008) and the chairman of the organizing committees of the Traitement et Analyse de lInformation: Methodes et Applications (TAIMA 2009) conference, International Conference on Machine Intelligence ACIDCA-ICMI2005 and International Conference on Signals, Circuits and Systems SCS2004.