Mathematical Characteristics for the Automated Recognition of Musical Recordings

G. ROUSSOPOULOS, D. FRAGOULIS, C. PAPAODYSSEUS, ATH. PANAGOPOULOS, M. EXARHOS
National Technical University of Athens,
School of Electrical and Computer Engineering
9 Heroon Polytechneiou,Athens GR-15773
roussop@cs.ntua.gr

Abstract: - In this paper, a very efficient novel methodology for the automatic recognition of musical recordings is presented. The core of this system employs a set of mathematical characteristics, extracted from a musical recording, whose determination was based on human perception. For the automatic recognition realization a musical signal is sampled, similar features are extracted from it and they are compared with model ones, via proper comparison algorithms. Thus, automatic recognition of musical recordings that may have suffered a very high distortion of an arbitrary type is accomplished, with a considerable success rate.

Key-Words: - : Automated music recognition, audio fingerprinting

1 Introduction

It is worthwhile noticing that human ear is capable of identifying sounds in noisy environments. Systems that perform the same identification task can be used in many applications, like broadcasting monitoring, audio appliances, internet file sharing systems management, legal rights protection etc. Several systems have been proposed that exploit information related to various spectral features. In [1], recognition is performed using Spectral Flatness Measure, while in [2] Mel-Frequency Cepstrum Coefficients are used and in [3] the difference of energy in 33 bark-scaled bands is used. However, these systems, do not offer the same high success rates when noise is present in the unknown signal. Kurth et al. in [4] presented a system that can deal with highly distorted audio material, but the success rate that is offered, in the noisiest case, is not greater than 80%, while processing time is longer.

The methodology presented in this work uses some of the principles introduced in [5], but drastically upgrades the performance capabilities of the recognition system in the presence of noise. This approach is based on the assumption that there exist invariant characteristics in time and spectral domain, which are independent of the kind and degree of distortion and are exploited by the human ear in order to identify a sound and exploits some of the basic mechanisms of human hearing that are related to frequency selectivity and masking. The high efficiency of the presented methodology relies on:

- Application of temporal masking processing, in addition to frequency masking,
- An extended set of spectral features, used for the recognition process that incorporates the spectral frame centroid derivative.
- An extensive study and successful confrontation of the problem of reproduction speed change.
- An octave based division of the audibility domain into bands.
- Determination of a proper frequency range of the masking function.
- Implementation of a novel, fast, very efficient final stage pattern-matching criterion.

A brief description of the musical recordings recognition problem, which is tackled by the proposed methodology follows: 1. A set of recordings considered as model signals is given. 2. A sampled recording is considered as an unknown signal. 3. Find if the unknown signal matches a specific model signal.

2. The Problem Of Musical Recordings Automatic Recognition

The term musical recording (MR) is used to describe a recorded piece of music, performed by an artist or group of artists. MRs may have suffered a very serious distortion, due to reasons like: Analog or digital transmission through radiowaves, playing speed difference, time shift, cropping, volume change by a slightly varying factor, audio coding e.g. MP3, equalization, bandlimiting, dynamic range compression, random noise present in the signal, loudspeaker-microphone transmission etc.

The aforementioned reasons may cause drastic, obvious differences of the sampled unknown MRs from the corresponding model ones, both in time and frequency domain. As a consequence, features used in content based retrieval such as: number of zero crossings, DFT peaks amplitude, time and frequency envelopes, energy distribution in time both in linear and logarithmic scale, etc. [6],

completely fail as related extended experiments performed by the authors confirm.

Moreover, an applicable system for the automatic recognition of musical recordings must be able to recognize one, among many tenths of thousands of others. It is obvious that one must manipulate a huge amount of information, with the best possible time performance.

3 Feature Extraction And Organization

3.1 Division of the audible frequency domain into bands

It is well known that the human hearing system operates obeying a logarithmic rule [7]. Thus, we divide the whole audibility domain into sixty (60) logarithmic bands of the type $\lfloor C_{down}, C_{up} \rfloor$, where:

$C_{down} = F_{note} / \sqrt[24]{2}$, $C_{up} = F_{note} \cdot \sqrt[24]{2}$

and $F_{note} = 110 \cdot i \cdot \sqrt[12]{2}$, $i = 1, 2, ..., 60$.

This division resembles the octave based note production in western music and offers a very satisfactory recognition success rate.

3.2 A first step towards the band representatives vectors

Suppose that a part of length *ML* of a musical recording is sampled at a frequency *Fs* and a *N* sample frame of this part is kept for processing. On the *N* samples of this "test frame", the maxima of the absolute value of the Discrete Fourier Transform (DFT) are spotted and both their amplitude and position in the test frame are stored in two arrays.

Subsequently, a "first envelope" of the DFT absolute value is obtained by linear interpolation of all points (x_i, P_i), where P_i is an arbitrary peak amplitude and x_i its position in the test frame. Next, the maxima (peaks) of this "first envelope" are once more spotted and both their amplitude and their position are stored in two arrays. If the number *N* of the test frame samples is greater than or equal to 2^{14}, then the aforementioned procedure is repeated. So, eventually two arrays are kept:

One containing the DFT amplitudes first (if $N < 2^{14}$) or second (if $N \geq 2^{14}$) envelope maxima, the other the position of these maxima in the test frame (in samples).

It is well known that the human hearing system is subjected to a so called masking phenomenon, according to which, when a lot of signal energy is present at one frequency, the ear is far less sensitive at nearby frequencies, so that, practically cannot hear them. A similar phenomenon takes place when adjacent frequencies appear with a small difference in time. As it is usually said, the louder frequency masks the softer ones and for this reason it is called the masker (see [8] and [9]).

To incorporate the masking phenomenon in the introduced method, the following procedure is used: the highest amplitude peak of the first or second envelope having a position n_0, in samples, is selected. Then the "masking function", given by Eq. (1), is built around it, where f_0 is the frequency corresponding to the peak position n_0 in samples.

$$F(z) = 15.81 + 7.5*(z + 0.474) - \\ -17.5*\sqrt{1 + (z + 0.474)^2} \quad (1)$$

$$z = 13*\arctan(0.00076(f - f_0)) + \\ + 3.5*\arctan\left(\left(\frac{f - f_0}{7500}\right)^2\right) \quad (2)$$

However, by making a significant deviation from the original masking procedure, the domain of the function $F(f)$ is limited between the frequencies $f_0 - W$ and $f_0 + W$, where f_0 is the masker frequency and *W* a properly chosen positive number. It has been experimentally observed that for the automatic recognition system presented here, the optimal value of *W* is 30 Hz. According to the methodology presented in this paper, if the amplitude of a peak of the first or second DFT envelope in the interval $[f_0 - W, f_0 + W]$ is smaller than the value of $F(f)$, then this peak is removed from the corresponding array of sorted peaks. This procedure is repeated, for all the spotted peaks.

The phenomenon described above refers to simultaneous masking, which takes place in a single time instance, taking into account information only from the frequency domain. A similar phenomenon, observed in the time domain is forward masking. According to it, presence of a loud frequency at one time point affects the adjacent frequencies of the following time points due to human ear adaptation.

In order to take into account this phenomenon, we consider two consecutive "test frames" a and b at time instances t1 and t2 respectively. The n_i^a "masker spectral peaks" of the a frame have been estimated and subsequently their magnitude is reduced by a logarithmic factor, e^k. Then we compute the n_j^b "masker spectral peaks" of the b

frame. To the superimposition of n_i^a and n_j^b the masking function of equation (1) is applied, thus obtaining the n_l^b "final masker spectral peaks" of the b frame.

This procedure is carried on, for all the final masker spectral peaks, and when so, two arrays are constructed: In the first one, the amplitudes of the final masker spectral peaks; we name these max-masked peaks. In the second array the max-masked peaks position is kept. At this point, the max-masked peaks obtained by the above procedure are classified into bands, according to the adopted band division. We define the "band amplitude" for each band as the amplitude of the greater max-masked peak within this band. Subsequently, the L bands with greater amplitude are selected from the 60 bands, and their indices are stored in a separate array, we will name "band representatives vector". The elements of this array will be called "band representatives". The experiments performed indicate that an appropriate choice for L is: $17 \le L \le 25$.

3.3 Taking into account playing speed differences

The "band representatives" extraction introduced above leads to an identification procedure with excellent results for MRs that have suffered an even very high distortion in the frequency domain. This procedure, however, degrades in the case that the musical recording has suffered a considerable distortion in playing speed, more than one and a half percent (1.5%) approximately. In order to deal with this additional type of distortion, the method introduced in the present section has been developed.

Thus, consider once more, the *N* samples test frame and the two arrays of the corresponding DFT absolute value maxima amplitudes and position. To take into account the distortion in playing speed, the peaks positions are multiplied by a properly chosen factor $\lambda_i, i \in N$, called "the i th stretch factor". The range of λ_i values depends on the expected speed distortion; so, for decisively most cases of radio or TV broadcasted musical recordings, one may safely state that λ_i belongs to the interval $[0.88,1.12]$.

So, for a specific λ_i, one considers that every peak position x_k of the DFT absolute value is moved to the position $\lfloor x_k * \lambda_i \rfloor$, where $\lfloor y \rfloor$ stands for the integer part of $y \in \mathbf{R}$. In other words, a new peaks position vector is obtained:

$$SP_{(i)} = \left\| \lfloor x_1 * \lambda_i \rfloor, \lfloor x_2 * \lambda_i \rfloor, \lfloor x_3 * \lambda_i \rfloor, ..., \lfloor x_p * \lambda_i \rfloor \right\|, \quad (4)$$

which is considered to represent the MR DFT absolute value peaks played with a different speed than the model one, corresponding to the specific stretch factor. Extended experiments show that in order to cover the playing speed range of radio/TV stations, it suffices to consider a limited number of λ_i, differing by a step of the order of 0.007 to 0.012. Thus, by choosing a specific $STEP$, we obtain a sequence of recognition stretch factors:

$$\lambda_i = \begin{cases} 1 + (\frac{i+1}{2}) * STEP, & \text{if i odd} \\ 1, \quad if \quad i = 0 \qquad i = 1,2,...,S \\ 1 - (i/2) * STEP, & \text{if i even} \end{cases} \quad (5)$$

that give rise to a corresponding sequence of shifted (stretched) position vectors.

Next, for each such shifted position vector $SP_{(i)}$, the procedure described in sub-section 3.2 is repeated, thus finally obtaining a "i-stretch band representative vector" having as elements L "i-stretch band representatives", where again $17 \le L \le 25$. In this way one obtains the general class of band representatives vectors for the specific frame.

3.4 The set of compact spectral centroid derivative vector

Band representative vectors, as defined above, can be considered as a very precise sound fingerprint. However, such large information cannot offer a fast indexing in an audio database. However, we can use a more compact, lossy, form that enables very fast indexing within a database. Using frames of length *N*, as above, band-pass filtering is applied between $110\,\text{Hz}$ and $110 \cdot 60 \cdot \sqrt[12]{2}\,\text{Hz}$, the mean value of spectral centroid is estimated and its value is assigned to one of the 60 audibility domain bands. Next, the band index difference between successive frames is calculated with a step of K samples and the sign of this difference is stored using a single bit. Thus, a sequence of (song length)/K-N bits is obtained.

From this sequence we extract all possible KL-bit subsequences, which we call compact spectral centroid derivative vectors, and store them in a database along with their exact position. In this way it is possible to locate the exact sample of the model MR that corresponds to each compact spectral

centroid derivative vector. This information does not offer such a high accuracy as the band representative vectors but on the other hand can be used to drastically speed up the recognition process.

3.5 The sets of band representatives vectors for the model musical recordings

To each model signal a procedure analogous to the one described in sub-section 3.2 is applied. Thus, a "model set of band representative vectors" is obtained, for all samples of the model signal, without applying any stretching procedure, like the one described in sub-section 3.3. Notice that, very frequently, two or more frames starting at consecutive time samples correspond to identical model band representatives vectors. So, to each such vector we attach the number of starting time samples for which it remains the same, called "repetitions number of the vector". In the following, when we refer to a band representatives vector we consider that a number of repetitions is attached to it.

The creation of the model set of band representatives vectors would require a considerable amount of computational complexity, since it involves a N-samples FFT computation for each sample of the model signal in hand. For this reason, an adaptive FFT computation algorithm is used ([10], [5]). In order to obtain a more efficient coding of the band representatives vectors, we have used the ideas presented in [5], slightly modified to take into account the different division in 60 frequency bands.

4. The Developed Matching Test

The introduction of the band representative compact and spectral centroid derivative vectors, provides the ability to successfully and rapidly check for matching between an unknown and a model musical recording via the use of the comparison methods and the matching criteria described below.

4.1 Spectral centroid derivative vectors comparison

Consider two recordings of the same musical composition, an unknown and a model MR, starting at a sample corresponding to exactly the same piece of music. If one applies the procedure described in section 3.4 and obtains the spectral centroid derivative vectors for both MRs, he will observe that there exists a stretch factor λ_i for which these vectors are identical, in almost every case. A specific vector may exist in more than one MR or more than once in the same MR. Therefore this information can not be used as an autonomous matching criterion. However, it drastically reduces the required recognition time since we only need to explore further a very small subset of the band representative vectors.

Notice that in the case of very high distortion, two corresponding spectral centroid derivative vectors may differ by one or two bits. To diminish the error probability we compare more than one vectors.

4.2 Single frames comparison

Consider two recordings of the same musical composition, an unknown and a model MR and suppose at first that they have both been played with the same speed. In addition, consider that one arbitrarily selects two N samples frames of these signals, corresponding to exactly the same piece of music, and computes the band representatives vectors of these two frames. All entries of these vectors cannot be identical due to the existent distortion. However, extended experiments performed by the authors show that, even in the case the unknown musical recording has suffered a particularly heavy distortion of the type described in section 2, then, still, the two band representatives vectors have at least $0.49 * L$ common elements, independently of the position the frame in hand is chosen and of the kind of musical composition.

In the case the unknown MR has been played with a different speed than the model one, then the performed experiments clearly indicate that there is a stretch factor λ_i for which the corresponding i-stretch band representatives vector has at least $0.49 * L$ elements in common with the band representatives vector of the counterpart model frame.

Therefore, one might consider that a first criterion for deciding if two musical recordings match, would be the demand that there are two frames of these two signals whose band representatives vectors share at least $0.49 * L$ elements. Such a criterion, however, is not sufficient for automatic recognition, since: First, there may be considerably more unknown MRs satisfying the above criterion and second this criterion does not take into consideration the signal evolution in time. Hence, an efficient matching criterion described in the subsequent sub-section has been adopted.

4.3 The final stage identification criterion

Let $\mathbf{V_{k,i}}$ be the kth band representatives vector of the unknown part being calculated at a N samples frame starting at sample n_u; the second index i of

this vector corresponds to the stretch factor λ_i, as defined in Eq. (5). Moreover, let $\mathbf{U_{k,n,R}}$ be the k th band representatives vector of the R th model MR in the database; the second index **n** of this vector corresponds to the starting time sample of the model frame-window from which the representatives vector in hand has been generated. In the sampled broadcasted signal part, we select M frames of N sample length, where two consecutive frames have an ℓ-samples distance and we construct the band representative vectors for each such frame and all S stretch factors λ_i. In this way one obtains S sequences of M vectors. $\left[\mathbf{V_{1,i}}, \mathbf{V_{2,i}}, ..., \mathbf{V_{M,i}}\right]$.

Next, for an arbitrary musical recording of the model database, one retrieves the band representatives vectors $\left[\mathbf{U_{1,n,1}}, \mathbf{U_{2,[n+\ell*\lambda_i],1}}, ..., \mathbf{U_{M,[n+(M-1)*\ell*\lambda_i],i}}\right]$ corresponding to M samples of the model recording, where the first sample is n and all consecutive samples have a distance $\ell * \lambda i$. Subsequently, one compares the pairs of band representative vectors $\left(\mathbf{V_{1,i}}, \mathbf{U_{1,n,1}}\right), \left(\mathbf{V_{2,i}}, \mathbf{U_{2,n+\ell,1}}\right), ..., \left(\mathbf{V_{M,i}}, \mathbf{U_{M,n+(M-1)*\ell,1}}\right)$ and then performs the following steps:

1). One checks if the number of common elements between each pair members $\mathbf{V_{k,i}}$ and $\mathbf{U_{k,(n+(k-1)*\ell),1}}$, $\mathbf{k = 1,2,..., M-1}$ is greater or equal than $0.49 * L$.

2). If one pair of vectors $\left(\mathbf{V_{k,i}}, \mathbf{U_{k,(n+(k-1)*\ell),1}}\right)$ has a number of common elements less than $0.49 * L$, then one decides that the sampled signal does not match to the sample n of model MR in hand, for the specific stretch factor λ_i. Hence, one proceeds to comparing the sequence of vectors $\left[\mathbf{V_{1,i+1}}, \mathbf{V_{2,i+1}}, ..., \mathbf{V_{M,i+1}}\right]$, corresponding to the next stretch factor λ_{i+1}.

3). If comparison 1) fails for all stretch factors λ_i, then, one proceeds to comparing the sequence of vectors $\left[\mathbf{V_{1,i}}, \mathbf{V_{2,i}}, ..., \mathbf{V_{M,i}}\right]$ with the sequence of model band representatives vectors $\left[\mathbf{U_{1,n+1,1}}, \mathbf{U_{2,[n+1+\ell*\lambda_i],1}}, ..., \mathbf{U_{M,[n+1+(M-1)*\ell*\lambda_i],i}}\right]$ corresponding to a set of M samples, where the first sample is, now, $(n+1)$ and, once more, all consecutive samples have a distance $\ell * \lambda i$.

4) If the condition described in (1) above is satisfied, then the algorithm proceeds to the comparison between the mean values of common elements between all the previous pairs of band representatives vectors: $\left(\mathbf{V_{1,i}}, \mathbf{U_{1,n,1}}\right), \left(\mathbf{V_{2,i}}, \mathbf{U_{2,n+\ell,1}}\right), ..., \left(\mathbf{V_{M,i}}, \mathbf{U_{M,n+(M-1)*\ell,1}}\right)$.

5) If this mean value is greater or equal than $0.72 * L$ and if only at most $\lceil L/3 \rceil$ of the above pairs of vectors had common elements in the interval $[0.49 * L, 0.72 * L]$, then the first matching criterion is satisfied and the system proceeds to the final stage criterion described in the next section.

6) Otherwise, if the mean value is smaller than $0.72 * L$ or if more than $\lceil L/3 \rceil$ of the above pairs of vectors had common elements in the interval $[0.49 * L, 0.72 * L]$, then the matching criterion is not satisfied. The algorithm considers that no matching exists and it proceeds to the comparison of the two new sequences of vectors $\left[\mathbf{U_{1,n+1,1}}, \mathbf{U_{2,[n+1+\ell*\lambda_i],1}}, ..., \mathbf{U_{M,[n+1+(M-1)*\ell*\lambda_i],i}}\right]$.

7) If all sequences of vectors $\left[\mathbf{U_{1,n,1}}, \mathbf{U_{2,[n+\ell*\lambda_i],1}}, ..., \mathbf{U_{M,[n+(M-1)*\ell*\lambda_i],i}}\right]$ of the model musical recording in hand have been compared to the sequence $\left[\mathbf{V_{1,i}}, \mathbf{V_{2,i}}, ..., \mathbf{V_{M,i}}\right]$ without successful matching, then the algorithm decides that the unknown signal part does not correspond to the specific model recording.

Notice that the above algorithm is executed very fast, since there is a minimal number of comparisons for each couple $\left(\mathbf{V_{k,i}}, \mathbf{U_{k,(n+(k-1)*\ell),1}}\right)$ as we fully exploit the stored encoded information pertinent to the specific model recording. In fact, two vectors, $\mathbf{U_{k,n,R}}$ $\mathbf{U_{k,n+1,R}}$ corresponding to two consecutive samples n and n+1, in practice, are identical or in most cases differ by 1 band representative or by two or rarely by 3 and in some extreme cases by more than three. The different band representatives between any two consecutive samples, as well as the number of repetitions for any vector $\mathbf{U_{k,n,R}}$ are encoded [5] and stored in a database. Now, clearly if $\mathbf{U_{k,n,R}}$ and $\mathbf{U_{k,n+1,R}}$ are identical no comparison is necessary, if they differ by a single band representative only two comparisons are needed etc.

If the whole database of possible candidate model set of band representatives has been exhausted and no matching is reported, then one deduces that the specific unknown musical recording does not correspond to any model musical recording of the database. If this matching criterion is proved successful then the current value of the shift factor, say λ_m, is stored as well as the exact samples numbers where fulfillment of the first matching

criterion occurs and then, we proceed to the final criterion described in sub-section 4.4.

4.4 The final stage identification criterion

Suppose that the previous matching criterion is satisfied at a specific time sample q of the kth model musical recording for the shift factor λ_m. In order to verify that this specific model musical recording is the exact counterpart of the unknown recording in hand, the final stage criterion is applied. This criterion is essentially based on the comparison of a large number of band representatives vectors between the model and the unknown MR. Therefore, a number P of N sample frames are selected from the unknown part, where $32 \le P \le 50$ and the starting sample of the ith frame has a distance l samples from the starting sample of the $(i-1)$th frame. The procedure described in sub-sections 3.2 and 3.3 with the stretch factor λ_m is applied to each one of these frames and in this way P band representative vectors $\mathbf{W}_{\mathbf{k,m}}$, $\mathbf{k} = \mathbf{1,2,...,P}$ are obtained. Next, the pairs of vectors:

$$(\mathbf{W}_{\mathbf{1,m}}, \mathbf{U}_{\mathbf{1,q,k}}), (\mathbf{W}_{\mathbf{2,m}}, \mathbf{U}_{\mathbf{2,[q+\ell_1*\lambda_m],k}}),$$
$$..., (\mathbf{W}_{\mathbf{P,m}}, \mathbf{U}_{\mathbf{P,[q+(P-1)*\ell_{P-1}*\lambda_m],k}})$$

are compared as described in section 4.2. If the mean value of common elements between all the above pairs of band representatives vectors is greater or equal than $0.73*L$ and if only at most $[0.175*L]$ of the above pairs of vectors had common elements less than $0.73*L$, then the final matching criterion is satisfied. Otherwise, the final matching criterion is not satisfied. In this case, the criterion considers that no matching exists at the specific time sample of the model musical recording in hand. Both the first stage criterion and mainly the final stage one are very powerful since they incorporate essential information from the frequency domain well distributed in the time domain.

5. Developing and Testing a System

5.1 System Description

Based on the methodology presented above we have developed a system that can operate in common PCs and performs automatic recognition of broadcasted musical recordings. The developed system consists of three modules, namely:

1) a database of the band representative vectors,
2) the acquisition of signals module that samples signals through various sources
3) the recognition module

In fact, due to the extraordinary large number of band representative vector combinations, it is not possible to perform identification using any known indexing method. In order to drastically speed up the identification process the system exploits the additional information stored in the compact spectral centroid derivative vectors. This information provides a fast index, and diminishes the set of model band representative vectors under comparison. The system proceeds to accomplishing the matching procedure via the criteria described in section 4, as follows:

(a) The no-stretch case, ($\lambda_0 = 1$) is treated first, with M=10. So, the band representative vectors $\mathbf{V}_{\mathbf{k,0}}$, $k = 1,2,...,10$ of $L = 20$ elements are computed in ten windows of length $N = 8*1024$, where the first starts at the first sample of the unknown signal, while each subsequent window starts at a distance of $\ell = 21000$ samples from the previous one. In addition, the compact spectral centroid derivative vectors are calculated.

(b) Then, the system uses a value $STEP = 0.011$, to obtain a sequence of stretch factors. The upper and lower limits of this sequence are defined by the user, according to the expected maximum degree of playing speed distortion. Our observations show that a practically absolute upper limit is 1.12, while a corresponding lower one is 0.88; however, one can cover almost 99,5% percent of cases by choosing these limits to be 1.06 and 0.94 correspondingly. Hence, the S sequences of the band representatives $\mathbf{V}_{\mathbf{k,i}}$, $k = 1,2,...,10$ are computed for the unknown signal. Next, for each stretch factor, we obtain the compact spectral centroid derivative vectors.

(c) For each of the compact spectral centroid derivative vectors that have been computed, we perform a query in the database to find the corresponding MRs and sample positions. Thus, we obtain a set of model MRs that is the candidate set for recognition.

(d) Subsequently, the matching criterion is applied for these S groups of vectors, on the candidate set for recognition. If this criterion is satisfied at a specific sample n_m of the kth model MR for the shift factor λ_m, then the system proceeds to the final stage criterion, employing $P = 42$ vectors. The first vector is computed in the window starting at the first sample of the unknown signal, while the other vectors are computed in windows each one starting

at a distance of $[BL * \lambda_m / \mathrm{P} - 1]$ samples from its previous one. If the final stage criterion is satisfied too, then the system decides that the specific model recording is indeed the counterpart of the unknown MR in hand, and proceeds to the next sampling.

If the candidate model set of band representatives have been exhausted, without the final stage criterion being satisfied, then the system decides that the unknown recording in hand does not correspond to any one of the model MRs.

5.2 System Performance

We have used a model database consisting of twelve thousand one hundred and eighty six (12186) MRs obtained from CDs and the overall system performance for more than 100.000 pseudosamplings was ninety eight point eight per cent (98.8 %). No erroneous matching of an unknown MR with a non-corresponding model one has been signaled at all.

Another important test for the system is the one described below performed in concert with three radio and one TV stations, where the system has operated in actual, practical conditions. In fact, the system has monitored the four stations in predefined time intervals for seven (7) days. During these hours the stations personnel kept record of the transmitted songs, so that a direct evaluation of the system performance could be realized. Eventually, the system has reported one thousand nine hundred seventy two (1972) identifications which all have been confirmed by the two stations playlists. Moreover, we emphasize that, all other musical recording transmitted by the two stations in the predefined time intervals, did not have a counterpart in the system model database. In other words, the system had a hundred per cent success in identifying musical recordings stored in the model database, while at the same time, it never generated an erroneous identification.

Perhaps the most important test is the one performed with five different cellular phones. A mobile phone is placed near a sound source and the audio signal is transmitted, recorded and used as an unknown MR. One hundred and sixty (160) different pieces of MRs have been recorded. The system showed an average recognition performance of more than 92%, while no false recognition has been reported.

6. Conclusion

In this paper, a methodology for the automatic recognition of musical recordings is presented. For the automatic recognition realization, proper patterns are extracted from a set of selected model musical recordings based on human perception. A similar set of patterns is extracted from an unknown musical recording and they are compared with the whole database in two stages, via proper comparison algorithms. Thus, automatic recognition of musical recordings that suffered even a very high distortion in both time and frequency domain is accomplished, with an average success rate of more than 97%, without any erroneous identification.

References:

[1] E. Allamanche, J. Herre, O. Helmuth, B. Froba, T. Kasten, and M. Cremer, *Content-based identification of audio material using mpeg-7 low level description*, in Proc. of the Int. Symp. of Music Information Retrieval, Indiana, USA, Oct. 2002.

[2] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied, *Robust sound modeling for song detection in broadcast audio*, in Proc. AES 112th Int. Conv., Munich, Germany, May 2002.

[3] J. Haitsma, T. Kalker, and J. Oostveen, *Robust audio hashing for content identification*, in Proc. of the Content-Based Multimedia Indexing, Firenze, Italy, Sept. 2001.

[4] F. Kurth, A. Ribbrock, and M. Clausen, *Identification of highly distorted audio material for querying large scale databases*, in Proc.AES 112th Int. Conv., Munich, Germany, May 2002.

[5] D. Fragoulis, G. Roussopoulos, Th. Panagopoulos, C. Alexiou, C. Papaodysseus, *On the automated recognition of seriously distorted musical recordings*, IEEE Transactions on Signal Processing, Vol. 49, No 4, pp. 898-908, April 2001.

[6] Yao Wang, Zhu Liu and Jin-Cheng Huang, *Multimedia Content Analysis using both audio and visual clues*, IEEE Signal Processing Magazine Vol. 17 (6), pp. 12-36, Nov 2000.

[7] Moore B.C.J., *An Introduction to the Psychology of Hearing*, Academic Press (1997).

[8] Brandenburg K., Stoll G., *ISO-MPEG-1 Audio: a generic standard for coding of high quality digital audio*, Journal of the Audio Engineering Society 42(10) (1994) 780-792.

[9] Bosi M. et al., *ISO/IEC MPEG-2 Advanced Audio Coding*, Journal of the Audio Engineering Society 10 (1997) 789-813.

[10] Allen J.B. and Rabiner L.R., *A Unified Approach to Short-time Fourier Analysis and Synthesis*, Proc. IEEE 65 (1977) 1558-1564.