A New Approach to the Automatic Recognition of Musical Recordings*

C. PAPAODYSSEUS, G. ROUSSOPOULOS, D. FRAGOULIS, TH. PANAGOPOULOS, AND C. ALEXIOU

National Technical University of Athens, Department of Electrical and Computer Engineering, Division of Computer Science, GR-15773 Athens, Greece

A new methodology for the automatic recognition of musical recordings is presented. A system has been developed that performs recognition among a set of specific musical recordings. The system claims a high rate of success (greater than 86%), even when the unknown compositions have suffered from up to a medium degree of distortion. It comprises a database of musical characteristics that correspond to a set of model musical recordings. These characteristics are derived by applying novel feature extraction algorithms to every model musical recording selected. In order to determine whether an unknown musical recording corresponds to a piece represented in the database, the same feature extraction algorithm is applied to it, and the characteristics thus derived are compared to the database contents by means of a set of criteria. The system can operate in parallel, essentially in real time, even for a considerable number of model musical recordings, as long as the hardware necessary is available.

0 INTRODUCTION

An important topic in the field of one-dimensional digital signal processing is the realization of a system that recognizes musical recordings automatically. Currently much of the work in this field is oriented toward the development of a system for the automatic transcription of music. The few music transcription systems that exist work with noteworthy success only in the case of monophonic music. In the case of polyphonic music, the performance of such systems is very low, since the variety of musical timbres, harmonic constructions, and transitions impose insurmountable obstacles to the recognition procedure.

There are publications using classical pattern recognition methods in music, which deal with the correlation of short-duration musical parts [4], [5] or with connectionist models of music recognition [6]–[8]. Note that pattern recognition in music is intimately related to the general topics of pattern matching and database searching, where a considerable number of publications exist [9]–[16].

The present paper refers to the automatic recognition of musical recordings that have suffered distortion of up to a medium degree. This category includes most of the musical recordings received by radio. We use the term "musical recording" to describe the recording of a piece of music by an artist or a group of artists. Thus a musical recording may be a modern song, a classical composition, a piece of electronic music, a piece of folk music, and so on. The paper does not examine the problem of polyphonic music transcription. It introduces a new methodology, which can be used only for the realization of a system that can automatically recognize one musical recording among a set of others. Such a system is applicable to automatic broadcast counting and would be a very useful tool for those in the field of intellectual property rights or companies that compile musical data for statistical purposes (such as charts).

The term "medium-degree distortion" or simply "medium distortion" is used here to indicate that:

- There is no audible noise present in the received radio signal.
- The CD recording and the radio broadcast recording are played at the same speed or at almost the same speed.
- The radio station that transmits the musical recording considered does not amplify frequency bands excessively.

A set of approximately 650 broadcast recordings was obtained, and an objective test was performed in order to classify them according to whether or not they fulfill

* Manuscript received 1999 July 19; revised 2000 July 31.

the medium-distortion criterion. The recordings were obtained from 12 different FM radio stations and cover a very extended range of signal strengths. Two of them used a compressed form for each recording they broadcast (for example, transmitting MPEG-layer 3 compressed music). No recordings from AM broadcasts were obtained since only a very limited number of radio stations broadcast in the AM band. To determine the degree of distortion of a specific recording, we applied the following procedure.

First, exact synchronization is achieved between the recordings obtained from radio and from the corresponding CD in the sense that we ensure that both signals are initiated at the same instant of the content of the recording (see Sections 1.2 and 1.9).

Next the degree of difference in playing speed can be computed easily if the displacement of the corresponding peaks in the spectra of the two signals is calculated. Note that the difference between corresponding peaks increases proportionately to the peak position, with the same scale factor SF that represents the calculated playing speed difference. If the percentage difference $|SF - 1| \times 100$ is less than 4%, then we consider that the third criterion of medium distortion is satisfied. Note that a difference in playing speed of less than 2% is not audible to the human ear.

Next the spectra of the two signals are directly scaled according to the scale factor (SF), the spectral envelopes are constructed, and the squared difference (SD) between the two envelopes is computed. If the ratio of SD to the length of the signal tested is smaller than or equal to a specific threshold, then the second criterion of medium distortion holds. The choice of this threshold was made by inspecting many pairs of recordings and making a simultaneous acoustical judgment of the sound quality. This threshold was determined to be $0.95/(128 \times 1024)$.

In our experiments the "unknown" recordings obtained from a variety of radio stations showed that at least 83% of them satisfy the aforementioned restrictions. There is a certain correlation between the recordings that fail and particular radio stations because these criteria characterize the broadcast quality. Note also that this correlation changes, sometimes rapidly, according to a variety of factors (such as the specific radio producer, the sound engineer, or the transmission path used and cannot be predicted.

However, we emphasize once more that a radio signal that satisfies the three restrictions mentioned can still suffer obvious distortion, which presents serious difficulties to the automatic recognition procedure.

1 PRESENTATION OF THE PROPOSED METHODOLOGY

1.1 Problem Description

The problem of automatic recognition of a musical recording has three major aspects. First, we often deal with fast varying signals, that is, the superposition of many signals of a different nature, such as the voice of a singer and the sounds produced by various musical instruments. Note that it is very difficult to identify the type and number of instruments that generated a specific waveform and in particular, to spot characteristics such as pitch, notes, musical scale, tempo, musical intervals, and so on. In addition, performance style plays a critical role in the form of a musical signal.

The second aspect of this problem is the disturbance imposed by radio. In fact, during transmission each signal, which usually comes from digital or analog sources such as CDs, tapes, or even MP3 compressed songs, is distorted. The transmission and reception of the signal can be approximated by two filtering procedures, one taking place in the transmitter and the other in the receiver.

Finally, the third difficulty in automatic recognition lies in the fact that the personnel and the equipment at the transmitting site can create another important type of disturbance, even in the case of digital transmission. A classical example is the disturbance caused by the consoles used by radio stations and DJs.

Therefore it is particularly difficult to find consistently common patterns between a musical signal obtained from a CD and its counterpart obtained from radio.

1.2 Proposed Automatic Recognition Methodology

Suppose we consider a musical recording received by a radio set and therefore having suffered a certain amount of unknown distortion because of the aforementioned reasons. First of all, this distortion changes the quantitative information that can be obtained from time-domain signal analysis drastically. For example, if one considers quantities such as the number of zero crossings per frame of N samples (for example, $N = 2^n$, where $n = 12, 13, 14$, or 15), the relative position of peaks, the average slope per frame, or the relative amplitude of peaks for the model song obtained from a CD and for the very same musical recording received by a radio set, then huge differences will be found between the values of the same quantities for the two signals. Clearly, the greater the distortion, the greater the discrepancy between these values.

Even in the frequency domain, many quantitative characteristics differ drastically, and the discrete Fourier transform (DFT) of the musical recording of a CD and that of the very same musical recording signal received by radio are examples. Such parameters are the number of peaks per frame of N samples, the amplitude of actual DFT peaks, the energy for various bands, the order of the higher peaks, and so on. A manifestation of the drastic differences between two such DFT signals is presented in Fig. 1.

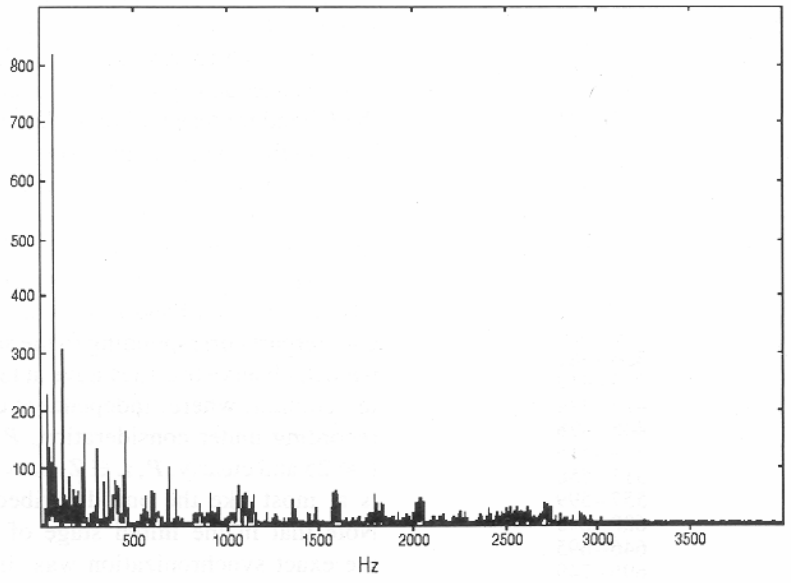
However, we have spotted some critical similarities between the CD signal and the radio signal of the very same musical recording in both the frequency and the time domains. In particular, we focused our research on the absolute values of the spectrum. After extensive related experiments we concluded that the musical information existing in a time frame of a musical recording

is intimately connected to the position of the spectral peaks of that frame.

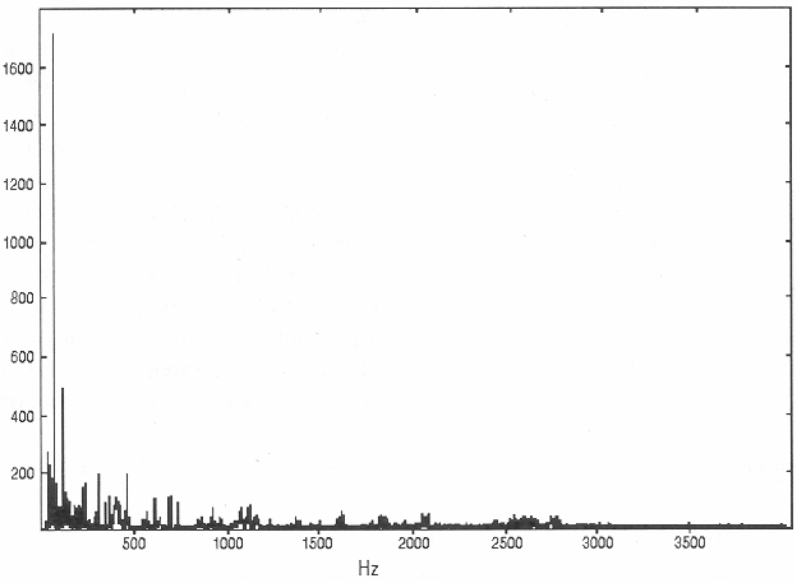
Therefore if for a musical recording we retain the information concerning the position of the spectral peaks for a sufficient number of frames starting at various instants of time, then we will realize the recording identification. But since it is impossible to store such a large amount of information for just one musical recording, we attempted a reduction of the storage capacity necessary by dividing the frequency domain into bands. The widths of the bands were chosen to be increasing almost exponentially in order to imitate the frequency selectivity of the human ear, that is, the experimentally verified shape of the auditory filter [17]–[19]. Based on the bandwidth used we decided that 48 bands were adequate to obtain the necessary information. Note that similar results can be obtained by choosing a slightly different number of bands. The selected widths of the individual bands are shown in Table 1. Of course, the final criterion for the correctness of such a choice is the efficiency of recognition validated by the experiment.

We would like to emphasize that in the proposed musical recording identification procedure we did not consider the bands that do not include a DFT peak, even if the energy of these bands was very high. On the contrary, we chose only the bands that included a spectral peak of a magnitude greater than an experimentally chosen cutoff. For this reason we did not adopt an approach using a filter bank.

Positioning one of the dominant frequencies into the



(a)



(b)

Fig. 1. Absolute DFT values of two signals corresponding to exactly the same piece of music. Both signals are of 2-second duration and their exact synchronization was achieved acoustically and then verified and fine-tuned with close inspection of the signals' plots in the time domain. (a) Part of CD musical recording. (b) Same part received by radio set.

proper band is clearly essential for the identification of musical recordings. Since it is well known that the DFT resolution, that is, the accuracy of the correspondence between samples and actual frequencies, is proportional to the DFT length, it is clear that one may expect a more representative band selection as the DFT length increases. For example, since the width of the lower band is narrow, a wrong placement of the peaks in these bands can occur easily if a small DFT length is chosen.

Table 1. Division of audibility domain into 48 bands.

Band	Band Frequency Interval (Hz)
0	0–50
1	51–99
2	100–107
3	108–116
4	117–125
5	126–134
6	135–145
7	146–156
8	157–168
9	169–181
10	182–195
11	196–210
12	211–227
13	228–244
14	245–263
15	264–284
16	285–306
17	307–329
18	330–355
19	356–382
20	383–412
21	413–444
22	445–478
23	479–516
24	517–556
25	557–599
26	600–645
27	646–695
28	696–749
29	750–807
30	808–869
31	870–937
32	938–1009
33	1010–1088
34	1089–1172
35	1173–1263
36	1264–1360
37	1361–1466
38	1467–1579
39	1580–1702
40	1703–1834
41	1835–1976
42	1977–2129
43	2130–2293
44	2294–2471
45	2472–2663
46	2664–2869
47	2870–11025

Moreover, if a peak occurs near the border of two arbitrary adjacent bands, low DFT resolution may once more result in a wrong peak classification. Since, as it has already been stated, we have observed experimentally that the accurate placement of each peak in the proper frequency band is fundamental for the identification of musical recordings, it follows that the greater the DFT length, the better the results of the identification procedure. However, after a certain length of the DFT the observed improvement in identification no longer justifies the considerable additional computational effort. Consequently a DFT length of 8192 ($= 8 \times 1024$) samples has been chosen as a good tradeoff between proper peak placement and computational effort. Therefore a very satisfactory and efficient procedure for achieving the identification of musical recordings can be based on the following remarks.

In each band, all peaks of the absolute DFT value are spotted. Then the peak with the greatest amplitude is chosen as the representative peak of the band in question. Subsequently the band indices that correspond to the L band representative peaks of greater amplitude are kept, in the sense that they are stored in a separate array. We will derive this array "band representative vector" or "band representative array." The elements of this array will simply be called "band representatives."

If we compare a band representative vector of the CD signal and a band representative vector of its radio counterpart corresponding to the same time sample, then we will observe that they have at least P of the L elements in common, where, independent of the specific musical recording under consideration, $P \geqslant 0.58L$ when $16 \leqslant L \leqslant 25$ and clearly, $P, L \in \mathbf{Z}$, provided that the distortion is at most like the one described in the Introduction. Note that in the initial stage of system development, the exact synchronization was, in each case, achieved acoustically and then verified and fine-tuned with close inspection of the two signal plots in the time domain. However, now the system presented here (see Section 1.9), while identifying the unknown recording, gives the exact time sample where the unknown recording and its counterpart from the CD match. The experimental results showing the correlation between P and L are displayed in Table 2.

Therefore one might consider that a criterion for recognizing a specific musical recording would be the presence of $0.58L$ elements common to two band representative vectors, one computed in a single frame of the radio signal and the other in a single frame of a CD signal. However, experiments have shown that quite frequently, two N = sample signal parts coming from completely

Table 2. Minimum number of common elements P between two band representative vectors of length L for 10 different values of L.*

L	16	17	18	19	20	21	22	23	24	25
P_{min}	10	10	11	12	12	13	13	14	14	15
Ratio P/L	0.625	0.588	0.611	0.632	0.600	0.619	0.591	0.609	0.583	0.600

* This table resulted as follows: for each value L extensive experiments were performed to obtain the minimum number of common elements P between two band representative vectors when matching occurs. One can express the results presented in this table with the concise formula $P \geqslant 0.58L$ when $16 \leqslant L \leqslant 25$.

different musical recordings may also satisfy this criterion. In addition, the DFT transform of a single frame does not contain information about evolution of the signal in time, which is obviously very important in human perception and musical recording recognition. Hence we decided to enhance the aforementioned criterion by introducing one more constraint related to the time domain. Hence the first matching criterion applied, which is simply necessary but not yet sufficient for the recognition of musical recordings, is as follows.

If two signals correspond to the same musical recording, then M properly chosen frames of length N samples must have band representative vectors that share at least $0.58L$ common elements simultaneously. In addition, the mean value of the number of common band representatives must be at least $0.71L$. These results were obtained experimentally for the values of L given in Table 2. For example, for the case $L = 20$ this requirement means that although some of the band representative vectors may have 12 elements in common, the mean value of the common elements of the band representative vectors is at least 14.2 when matching occurs. The analysis in Section 1.7 will clarify this criterion further.

1.3 Building a Set of Band Representative Vectors for the Unknown Musical Recording

To derive a group of band representative vectors for a part of an unknown musical recording the following steps are taken.

1.3.1 First Frame

1.3.1.1 We sample at random a part of the unknown musical recording, of a length of BL samples, and we transform it into a form suitable for processing by a computer, preferably in ".wav" format, with a sampling frequency $F_s = 22\,050$ Hz. In the subsequent analysis we will refer to it as the "radio signal part," although the unknown musical recording can be obtained from CD, television, tape, or any other related source.

1.3.1.2 At the beginning of this signal we pick a first frame of N samples (say, 8×1024) and apply the DFT to the frame.

1.3.1.3 We calculate the absolute value of the DFT and divide the N samples into 48 bands, according to Table 1. In each band we spot the peaks, if any exist, and we keep only the peak of maximum amplitude. If such a maximum peak exists, we assign its amplitude to the band at hand, and hereafter we consider it as the amplitude of the band. Otherwise we assign a value of zero to the band.

1.3.1.4 We find L bands of greater amplitude, and store their corresponding indices in a vector, in descending order. In this way we obtain the first band representative vector, which corresponds to the first frame, with values being the aforementioned index numbers.

1.3.2 Second Frame

1.3.2.1 We choose a second frame of N samples, N being exactly the same as in step 1.3.1.2, at a fixed distance of l time samples from the first sample of the radio signal, and we apply the DFT to that frame. Next we repeat steps 1.3.1.2, 1.3.1.3, and 1.3.1.4.

1.3.2.2 We choose an Mth frame of N samples at a fixed distance of l time samples from the first sample of the $(M - 1)$th frame, and we apply the DFT to the frame. Next we repeat steps 1.3.1.2, 1.3.1.3, and 1.3.1.4 for this Mth frame. In this way we obtain M band representative vectors, corresponding to the M chosen frames (see Fig. 2).

1.4 Building a Set of Band Representative Vectors for a Model Musical Recording

We apply a similar procedure to each signal obtained from a CD, calling it the CD signal. Thus we take N samples starting at the first sample of the CD signal and repeat steps 1.3.1.2, 1.3.1.3, and 1.3.1.4. In this way

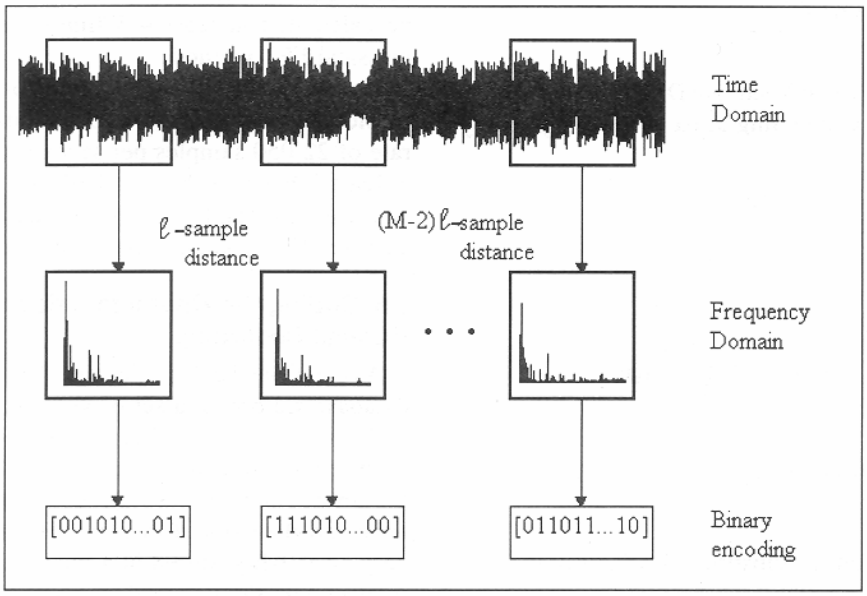


Fig. 2. Construction of band representative vectors for unknown part of musical recording.

we create a vector of L elements. We do the same for every sample of the CD signal, thus finally obtaining a set of vectors, each consisting of L elements. We will call this set the model set of band representative vectors. Note that it is very common that two or more consecutive time samples correspond to identical band representative vectors. Therefore we attach to each such vector the number of time samples for which it remains identical, and we call this the "repetitions number of the vector."

The creation of those vectors requires a considerable amount of computational complexity, since it involves an FFT computation of N samples for each sample of the CD signal at hand, except for the $N - 1$ last. Considering that a monophonic musical recording of 3.5-min duration, sampled at a rate of 22 050 samples per second, consists of approximately $3.5 \times 60 \times 22\,050 = 4\,630\,500$ samples, it is clear that the creation of these vectors may require many hours of computation, even on a 500-MHz Pentium III processor. Clearly, for a longer composition, say a classical piece of 30-min duration, the creation of these vectors may require several days if the typical FFT method is used. For this reason we applied a sliding FFT computation algorithm, which performs at a considerably reduced overall computation time [20]–[23].

1.5 Adaptive FFT Computation

Suppose that we computed the FFT of a signal $x[n]$ of N samples, starting at sample α in the time domain and ending at sample $(\alpha + N - 1)$. Next suppose that we want to calculate the N-sample FFT of the same signal $x[n]$ starting at sample $(\alpha + 1)$ in the time domain and ending at sample $(\alpha + N)$. This second FFT calculation can be performed adaptively, that is, by taking into account the information of the first FFT, as described next.

Note first that the N-sample DFT of $x[n]$, starting at sample α and ending at $(\alpha + N - 1)$, is given by

$$X[k] = \sum_{n=0}^{N-1} x[n + \alpha]W^{kn}$$

where $W = e^{-j(2\pi/N)}$. The N-sample DFT of $x[n]$, starting at sample $(\alpha + 1)$ and ending at $(\alpha + N)$, is given by

$$X_s[k] = \sum_{n=0}^{N-1} x[n + \alpha + 1]W^{kn}.$$

Thus

$$X_s[k] = (-x[\alpha] + x[\alpha])W^{-k} + \left(\sum_{n=0}^{N-2} x[n + \alpha + 1]W^{k(n+1)}\right)W^{-k} + x_s[\alpha + N]W^{k(N-1)}.$$

In the last summation we substitute the dummy variable

$$i = n + 1$$

which implies that

$$X_s[k] = -x[\alpha]W^{-k} + x[\alpha]W^{-k} + \left(\sum_{i=1}^{N-1} x[i + \alpha]W^{ki}\right)W^{-k} + x_s[\alpha + N]W^{k(N-1)}$$

and hence,

$$X_s[k] = -x[\alpha]W^{-k} + \left(\sum_{i=0}^{N-1} x[i + \alpha]W^{ki}\right)W^{-k} + x_s[\alpha + N]W^{k(N-1)}.$$

Therefore we obtain

$$X_s[k] = x[\alpha]W^{-k} + X[k]W^{-k} + x_s[\alpha + N]W^{k(N-1)}$$

which is the recursive-adaptive FFT computation we desired.

Note that the computational complexity of a standard FFT is $(N/2)\log_2 N$ complex multiplications, that is, $2N \log_2 N$ simple real-number multiplications, whereas the adaptive FFT computation algorithm, introduced here, requires $8N$ real multiplications. Therefore the latter algorithm is many times faster than the standard FFT for a window length N, with $N > 16$ samples. In our case it is desirable to have as high a resolution and accuracy as possible, with the only restriction being the limited processing time. Therefore we found that FFT windows of length $N = 8 \times 1024$, $N = 16 \times 1024$, and $N = 32 \times 1024$ offer very good results, as shown in Table 3. It is clear that for these values of N with the adaptive FFT algorithm introduced here there are 3.25 to 3.75 times fewer multiplications required than with the standard FFT. Considering furthermore the memory allocations and the additions involved, it can safely be stated that using the proposed adaptive FFT method, the set of L-element vectors of the entire musical recording can be calculated at least 4.5 times faster than using the classical FFT method.

In order to obtain this set of vectors for a monophonic musical recording of 3.5-min of duration, sampled at a rate of 22 050 samples per second, the time required is at least 7 hours if the standard FFT algorithm is used, whereas less than 1.5 hours is needed when the proposed adaptive FFT computation is applied.

1.6 Coding the Characteristics of a Model Musical Recording

As mentioned before, for each CD signal in the system database we obtain a set of vectors, each consisting of L elements, and a number of repetitions associated with each vector. Note that for each model musical recording, the set of band representative vectors is computed only once and then stored in a file. However, these sets are very large when stored in a file with the ASCII format, or even with the standard binary format. For this reason we developed a storing protocol that enables us to store

the data into encoded files. According to this protocol, we assign to each vector of the CD signal a 48-element binary array. Each element of this binary array represents one of the 48 bands into which we divided the entire audibility domain, in ascending order. A value of 1 is assigned to an array element where the corresponding band representative vector belongs to the L greater ones in amplitude and has amplitude greater than 0, whereas a value of 0 is assigned otherwise. Note that in each binary array most L bits can be set to 1.

We will use an example to illustrate this coding technique. Suppose that for the time domain sample n we keep the $L = 16$ greater band representatives that form the following vector:

[47 43 39 34 33 30 29 25 22 20 17 14 11 9 5 2].

In order to reduce the storage need, we apply the following technique. We create a binary array of 48 binary digits in which all elements are zero except for those with a position corresponding to a band index of the above vector. Thus we obtain the following array:

[001001000101001001001010010001100110000100010001].

Through bitwise operations we store this array as a binary sequence. Since it is quite common that the same sequence appears more than once in consecutive samples, it is not necessary to store the complete sequence for each time sample. On the contrary, we store the bit sequence only once, together with a binary number corresponding to the number of repetitions of this vector, that is, the number of consecutive time samples for which this specific sequence appears. In this way, each band representative vector occupies 10 to 12 bytes, according to the number of repetitions.

1.7 First-Stage Pattern Matching Algorithm

The recognition procedure is essentially a pattern matching between two sets of vectors, each vector being a band representative vector. The first set corresponds to the part of the unknown musical recording, the second to an arbitrary model musical recording of the database. Note that any two vectors thus compared cannot in practice be identical, even if they correspond to exactly the same piece of music because of existing distortion. Therefore in order to achieve musical recording recognition it is absolutely necessary to employ a pattern matching algorithm that allows for a successful matching between two sets of vectors, even if the vector pairs being compared have a considerable number of different elements.

Thus a pattern matching algorithm has been developed in which each band representative vector of the unknown composition is considered an independent state. Transition to the mth state is allowed if and only if all restrictions imposed in the previous $m - 1$ states are satisfied. For example, we compare the ith band representative vector $\boldsymbol{V}_i$ of the unknown part with the corresponding ith band representative vector $\boldsymbol{U}_{i,n}$ of a model musical recording, where n expresses the starting time sample of the window that has generated the model representative vector in question. We then proceed as follows:

1) If the number of common elements is less than $0.58L$, then we restart the comparison of vector $\boldsymbol{V}_1$ with the next band representative vector $\boldsymbol{U}_{1,(n+1)}$ of the model set.

2) If the number of common elements is greater than or equal to $0.58L$, then and only then we proceed to the comparison between the second band representative vector $\boldsymbol{V}_{i+1}$ of the unknown composition and the vector $\boldsymbol{U}_{i+1,(n+l)}$ of the model set of band representatives corresponding to the time sample l.

3) If the number of common elements between $\boldsymbol{V}_M$ and $\boldsymbol{U}_{M,[n+(M-1)l]}$ is less than $0.58L$, then we do not continue the comparison any further. We consider that the matching criterion has failed for this sample of the model musical recording and we restart the comparison of vector $\boldsymbol{V}_1$ with the next band representative vector $\boldsymbol{U}_{1,(n+1)}$ of the model set, just as in step 1).

4) If the number of common elements between V_M and $\boldsymbol{U}_{M,[n+(M-1)l]}$ is greater than or equal to $0.58L$, then and only then we proceed to the comparison between the mean values of common elements and all the previous pairs of band representative vectors.

5) If this mean value is greater than or equal to $0.71L$, then the matching criterion is satisfied. Otherwise, if the mean value is smaller than $0.71L$, then we consider that the matching criterion has failed for this sample of the model musical recording and we restart the comparison as discussed in step 1).

Note that $\boldsymbol{U}_{i,n}$ can be identical to $\boldsymbol{U}_{i,(n+1)}$, $i = 1, 2, \ldots, M$, in which case no comparison is performed at that specific stage, but instead the value of the previous comparison is used.

Table 3. Necessary real-number multiplications of standard FFT and proposed adaptive FFT for three window lengths.

	Window Length		
	$N = 8 \times 1024$	$N = 16 \times 1024$	$N = 32 \times 1024$
Standard FFT	212 992	458 752	983 040
Adaptive FFT	65 536	131 072	262 144

In the case that all comparisons prove successful, the system offers the exact time-sample position at which first-stage matching occurs and proceeds to the final criterion, which will be described in Section 1.9. Otherwise the whole procedure stops when all the model band representative sets have been examined, in which case the system decides that the part of the specific unknown musical recording does not correspond to any of the model musical recordings of the database. The state diagram shown in Fig. 3 represents this algorithm.

1.8 Vector Comparison

The algorithm described in Section 1.7 requires a considerable number of vector comparisons. These comparisons are the main time-consuming operations of the algorithm. We have therefore developed a more efficient method for the calculation of the number of common elements between two vectors, reducing the time needed to perform a two-vector comparison. This method, which is based on the encoded representation (see Section 1.6) of the band representative vectors, uses a number of binary (logical) operations. We apply a bitwise AND to the two encoded vectors, which are in binary format, and then count the bits that are set to 1.

For example, consider the following two band representative vectors:

$$V_1 = [47 \quad 43 \quad 39 \quad 34 \quad 33 \quad 30 \quad 29 \quad 26 \quad 22 \quad 20 \quad 17 \quad 14 \quad 11 \quad 9 \quad 5 \quad 2]$$

$$V_2 = [47 \quad 44 \quad 41 \quad 39 \quad 37 \quad 33 \quad 30 \quad 28 \quad 25 \quad 20 \quad 17 \quad 14 \quad 11 \quad 9 \quad 6 \quad 2]$$

which are encoded as

$$V_{1,\mathrm{enc}} = [0010010000101001001001010001001100110000100010001]$$

$$V_{2,\mathrm{enc}} = [0010000100101001001001000010010100100010101001001] .$$

then

$$V_{1,\mathrm{enc}} \text{ AND } V_{2,\mathrm{enc}} = [0010000001010010010010000000001001000001000000001] .$$

By counting the bits that are set to 1 we find that the number of common elements between these vectors is 10. This method, when executed in a 32-bit microprocessor, requires two logic operations and 96 operations to count the bits.

1.9 Final-Stage Pattern Matching Algorithm—Envelope Matching in the Time Domain

After application of the first matching criterion a time-sample point of matching has been obtained and a final-stage criterion is necessary in order to verify whether the current model recording corresponds exactly to the unknown recording.

Suppose that we sample a part of a radio signal of a duration of, say, T_R seconds, which has suffered an up to medium distortion. Suppose moreover that we pick a part of a CD signal of the same duration. We create the time-domain envelopes of these two signal parts as follows:

1) We find all positive maxima of both signal parts and interpolate them linearly.

2) We find all negative minima of both signal parts and interpolate them linearly too.

In this way we obtain a "first positive" and a "first negative" envelope of each signal part. However, these two envelopes follow the two signals very closely, a fact that is not satisfactory for the present application. Therefore we repeat steps 1) and 2) of the procedure in the sense that:

- We find all positive maxima of the first positive envelope of both signal parts and interpolate them linearly.
- We find all negative minima of the first negative envelope of both signal parts and interpolate them linearly.

Thus we obtain a "second positive" and a "second negative" envelope of each signal part, which however, are still unsatisfactorily close to the two signal parts for the present application. We repeat once more steps 1) and 2) of the procedure and get a "third positive" and a "third negative" envelope of each signal part. In the subsequent analysis we will refer to them by "positive envelope" and "negative envelope." A typical example of a positive envelope derived according to the preceding description is presented in Fig. 4.

Next we compute the integral of each such envelope and normalize the envelopes by dividing all their values by the corresponding integral. Then we calculate the squared differences between the two normalized positive envelopes and between the two normalized negative envelopes. It has been observed that if the two parts (CD and radio) correspond to the same part of the musical recording of duration T_R seconds, then their squared difference has a value smaller than a specific corresponding threshold. To obtain such a threshold value, we examined many (more than 200) isolated parts of model recordings with their corresponding radio recordings (synchronized in time). Those experiments have shown that for $T_R = 10$ seconds, the threshold value is 0.85.

In conclusion, we can say that the squared difference between the envelopes of two signal parts is another

criterion for determining whether these two parts correspond to the same musical recording. At the same time this value represents the degree of distortion of the radio signal.

It should be emphasized that after successful application of both the first- and the final-stage criteria the exact time sample where matching between radio and model recordings occurs is found.

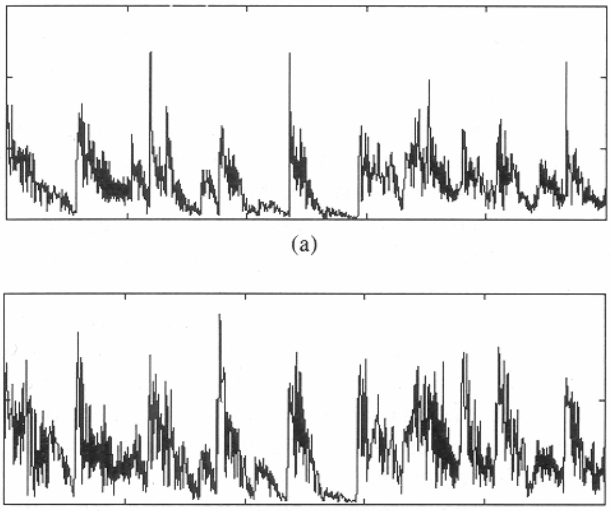
2 APPLICATION OF THE METHODOLOGY

On the basis of the aforementioned analysis, a system for the automatic recognition of musical recordings has been developed which performs the task as described next.

2.1 Implementation of First-Stage Recognition

1) The system samples at random a part of the unknown musical recording (the radio signal) of approximately 10-second duration. Next the procedure described in Section 1.3 is applied to that part, with frames of length $N = 2^{13}$ samples. In this way 13 vectors, each consisting of $L = 16$ elements, are obtained. Each vector has been calculated at a sample of the time domain having a distance of 20 000 samples from its subsequent and or its previous vector.

2) For each model set of band representative vectors of the system database, we check whether it includes elements that have a distance in the time domain of 20 000 samples too, and match it with the 13 vectors obtained from the unknown musical recording, according to the first matching criterion. If the counterpart musical recording of the radio signal exists in the system database, then some band representative vectors of this model musical recording satisfy the first matching criterion. In the case where there is more than one musical recording that satisfies this criterion, the final-stage cri-



(a)

(b)

Fig. 4. Typical example of positive envelope for the same piece of music. (a) Part of a model musical recording. (b) Part of an unknown musical recording received by radio and having suffered medium distortion. Similarities between both positive envelopes described in text are quite obvious.

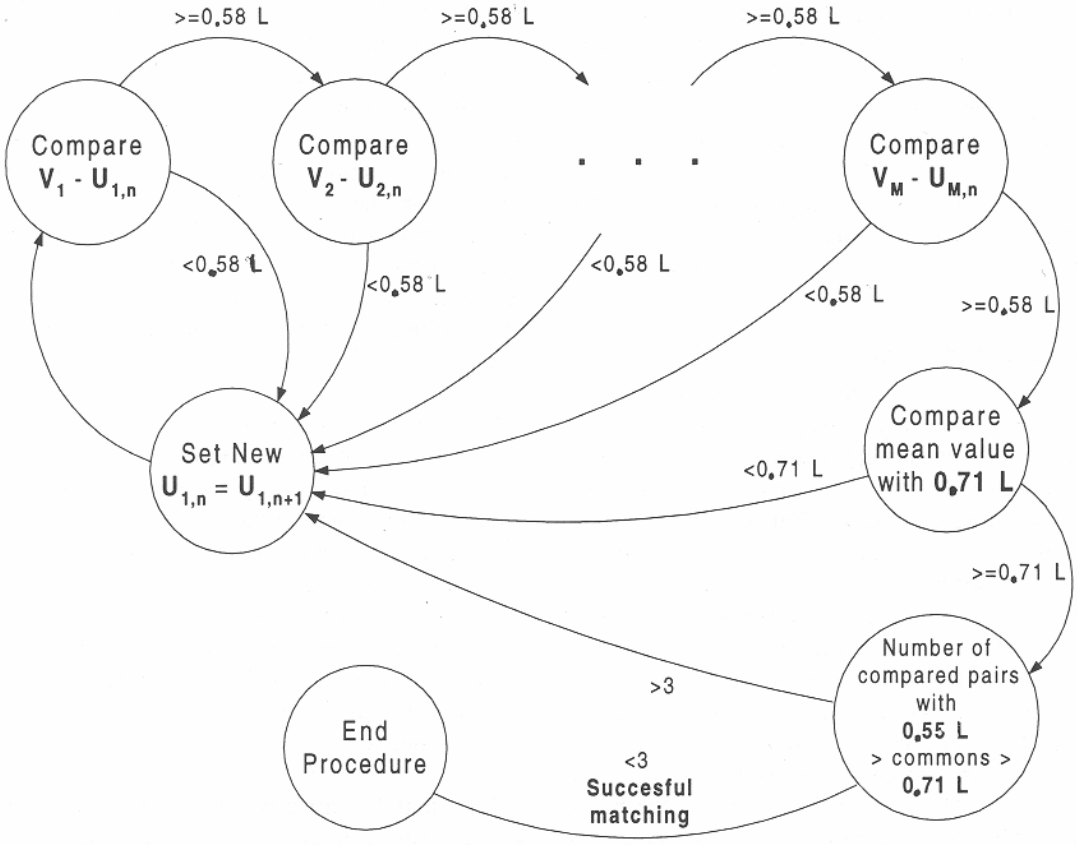


Fig. 3. State diagram for one musical recording.

terion defines which one corresponds to the radio signal. According to our experiments, multiple matching was found to occur only in the 4% of the radio signals that satisfy the requirement of medium distortion.

2.2 Implementation of Final-Stage Recognition

The final time-domain envelope matching criterion is applied to a frame of 220 500 samples. To reduce the time necessary for the application of the final-stage criterion, the positive and negative envelopes for each CD signal of the database are created and stored in a binary file. Actually only the position and value of all positive maxima and negative minima are kept in that file, thus achieving a vast reduction in storage need. We will refer to those files as "time envelope files."

As mentioned before, the final criterion is applied to define which specific musical recording corresponds to the radio signal, in the case where more than one musical recording has satisfied the first criterion. However, application of the first matching criterion offers the exact time samples at which a CD musical recording satisfies it, too. Thus the final criterion has to be applied only for the time samples of the CD signals suggested by the first matching criterion.

Before the final criterion is applied to compare a radio signal and a CD signal, the corresponding envelope file of the CD signal has to be loaded. The positive and negative envelopes of the CD signal can then be reconstructed using linear interpolation. The positive and negative envelopes of the radio signal are derived according to the procedure described in Section 1.9. Subsequently the time envelopes of the radio signal are compared with those of the CD signal that start at the time samples dictated by the first criterion. If the squared difference between the time envelopes is smaller than 0.85 for both positive and negative envelopes, then the system decides irrevocably that the specific CD signal corresponds to the radio signal at hand.

Note that we do not use time-domain envelope matching as a first-stage criterion, because it is much more time-consuming than the band representative matching criterion.

2.3 Experimental Results

On the basis of the methodology introduced in this paper we have developed a system for automatic recognition of radio broadcast musical recordings.

We have tested this system in connection with 350 CD musical recordings, as listed in Table 4, and approximately 650 musical recordings recorded from a variety of radio stations. For each of the radio signals we have chosen parts of 10-second duration at thousands of starting samples. Therefore we made "pseudo samplings" for each recorded radio signal in the sense that we simulated a sampling process thousands of times. As mentioned in the Introduction, at least 83% of the radio broadcast musical recordings we recorded suffer from medium distortion. The system developed on the basis of the aforementioned methodology automatically recognizes successfully the medium-grade distorted musical recordings with 100% accuracy for the thousands of random pseudo samplings performed in each composition.

The remaining 17% of the radio broadcast musical recordings suffer from more serious distortion, ranging from high to extrahigh. This kind of distortion is caused by some basic factors that have also been mentioned in the Introduction, namely:

- The CD and radio broadcast musical recordings were played at essentially different speeds so that the recorded composition spectrum was seriously distorted.
- The radio station that transmitted a musical recording under consideration greatly amplified the frequency bands.

The developed system fails to recognize compositions played at essentially different speeds. However, in the case of having just a strong frequency band amplification, the system succeeds in automatically recognizing the song at hand with 11 to 90% accuracy for the cases of the pseudo samplings in it. The exact percentage of successful recognition of pseudo samplings depends on the degree of distortion and the exact signal characteristics in both the time and the frequency domains.

We must emphasize that if we consider as an unknown musical recording any CD signal or any specific musical recording obtained from another source with minimal distortion, then the developed system recognizes it with 100% accuracy provided that the counterpart of the unknown musical recording is included in the database.

2.4 Time Required for Automatic Recognition

Suppose that part of a specific musical recording has been obtained from radio and stored in a ".wav" format file with a sampling frequency $F_s = 22\ 050$. It is difficult to give an exact estimate of the time required for the automatic recognition of the radio signal, since this time depends on the exact size of the database of musical characteristics as well as on the peculiarities of the radio signal at hand. However, extensive experiments show that when the underlying hardware is a Pentium II at 500 MHz, with a 256-MB RAM, and the operating system is UNIX, then:

- The typical maximum time required for deciding whether or not a radio signal at hand corresponds to a specific CD musical recording whose musical characteristics are stored in the system database is somewhat less than 2.5 min for a database containing 350 model musical recordings.
- The average time required for deciding whether or not a radio signal at hand corresponds to a specific CD musical recording of the system database is approximately 1 min for the same database containing 350 model musical recordings.

Note that the entire system has been developed so that it can operate in parallel. In this way if one uses K processors, then the time required for the automatic rec-

ognition of a radio signal when the model database consists of S songs is approximately divided by the number of processors K. Therefore one can achieve real-time automatic recognition even when the database contains many thousands of musical recordings, the only restriction being the hardware availability.

2.5 Time Required for Feature Extraction from CD Reproduced Musical Recordings

Once more, the processing time necessary for extracting a set of musical characteristics from a single model musical recording depends to a large extent on the exact size of the specific composition. Experiments show that when the aforementioned hardware and operating system are used, then for a model musical recording of 3.5-min duration, the necessary time is 2 hours. However, we must stress that this procedure takes place only once for each model musical recording, so that the file of its musical characteristics thus obtained is inserted into the database. Note that this procedure operates in parallel too.

3 CONCLUSIONS

A powerful system that performs automatic recognition of musical recordings has been presented. The system comprises a database of musical characteristics corresponding to a set of model musical recordings, derived by applying a novel feature extraction algorithm on every model musical recording. It has been asserted, with numerous experiments, that this system offers a very high rate of successful recognition, even for a musical recording that has suffered from medium distortion. It can operate in parallel, essentially in real time, for many thousands of compositions with the appropriate underlying hardware.

Table 4. List of composers and number of their compositions used to form system database.

Name of Artist or Group	Number of Compositions in Model Database	Name of Artist or Group	Number of Compositions in Model Database
1. Abba	9	48. Iggy Pop	2
2. Alan Parson's Project	2	49. Kate Bush	2
3. Animals	4	50. Ksilina Spathia	2
4. Aretha Franklin	2	51. Led Zeppelin	5
5. Asia	2	52. Lipps Inc.	1
6. Barkley-James-Harvest	2	53. Lou Reed	4
7. Beatles (The)	19	54. Luis Armstrong	2
8. Bee Gees	10	55. Madonna	4
9. Beethoven	5	56. Michael Jackson	4
10. Berlin	1	57. Mitropanos (Greek artist)	2
11. Boney M	2	58. Moody Blues	1
12. Bonnie Tyler	3	59. Motorhead	2
13. Boomtown Rats	1	60. Mozart (Wolfang Amadeus)	5
14. Bruce Springsteen	5	61. Nazareth	1
15. Bryan Adams	2	62. Oasis	3
16. Bryan Ferry	7	63. Peridis Orefeas (Greek artist)	4
17. Celine Dion	2	64. Phil Collins	5
18. Chopin	3	65. Pink Floyd	5
19. Chris de Burgh	3	66. Police	7
20. Christopher Cross	3	67. Prince	3
21. City	2	68. Queen	5
22. Cockney Rebel	2	69. R.E.M.	3
23. Creedence Clear Water Revival	4	70. Rainbow	4
24. Dalaras Giorgos (Greek musician)	2	71. Reo Speedwagon	3
25. David Bowie	4	72. Robert Palmer	4
26. Deep Purple	6	73. Rolling Stones (The)	12
27. Demis Roussos	6	74. Roy Orbinson	6
28. Diana Ross	4	75. Santana	3
29. Dire Straits	8	76. Scorpions	5
30. Donna Summer	4	77. Shirley Bassey	3
31. Doors	4	78. Stefka Sabotinova	1
32. Eagles	3	79. Steve Miller Band	4
33. Elo	3	80. Styx	1
34. Elton John	10	81. Supertramp	4
35. Elvis Presley	3	82. Talking Heads	1
36. Eric Clapton	3	83. Tchaikovsky	3
37. Eros Ramazzoti	3	84. The Trammps	1
38. Errol Brown	1	85. Theodorakis (Mikis)	5
39. Eurythmics	1	86. Tina Turner	3
40. Foreigner	6	87. Tom Jones	6
41. Frank Sinatra	2	88. Toto	4
42. Galani Dimitra (Greek artist)	3	89. U2	3
43. Gary Newman	3	90. Uriah Heep	2
44. Gloria Gaynor	2	91. Vangelis	4
45. Grieg	4	92. Village People	1
46. Hatzidakis (Manos)	4	93. Vivaldi	3
47. Hot Chocolate	1	94. Webber (Andrew Lloyd)	5

4 REFERENCES

[1] M. Mongeau and D. Sankoff, "Comparison of Musical Sequences," *Comput. and Humanities*, vol. 24, pp. 161–175 (1990).

[2] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Towards the Digital Music Library: Tune Retrieval from Acoustic Input," in *Proc. ACM Digital Libraries*, pp. 11–18 (1996).

[3] R. J. McNab, L. A. Smith, D. Bainbridge, and I. H. Witten, "The New Zealand Digital Library MELody inDEX," *D-Lib Mag.* (1997 May).

[4] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Recognition of Isolated Musical Patterns in the Context of Greek Traditional Music," in *Proc. IEEE Conf. on Electronics Circuits and Systems* (ICECS) (1997).

[5] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Recognition of Isolated Musical Patterns Using Discrete Observation Hidden Markov Models," in *Proc. Eur. Signal Processing Conf. (EUSIPCO)* (1998).

[6] C. J. Stevens and C. R. Latimer, "Music Recognition: An Illustrative Application of a Connectionist Model," *Psychol. of Music*, vol. 25, pp. 161–185 (1997).

[7] C. J. Stevens and C. R. Latimer, "A Comparison of Connectionist Models of Music Recognition and Human Performance," *Minds Machines*, vol. 2, pp. 279–400 (1992).

[8] C. J. Stevens and C. R. Latimer, "Recognition of Short Tonal Compositions by Connectionist Models and Listeners: Effects of Feature Manipulation and Training," *Musikometrika*, vol. 5, pp. 197–224 (1993).

[9] A. V. Aho, J. E. Hopcroft, and J. D. Ulmann, *Data Structures and Algorithms* (Addison-Wesley, Reading, MA, 1983).

[10] R. M. Karp and M. O. Rabin, "Efficient Randomize Pattern-Matching Algorithms," *IBM J. Research Develop.*, vol. 31, pp. 249–260 (1987).

[11] Y. Matias and U. Vishkin, "On Parallel Hashing and Integer Sorting," *J. Algorithms*, vol. 12, pp. 573–606 (1991).

[12] D. E. Knuth, J. H. Morris, and V. R. Pratt, "Fast Pattern Matching in Strings," *SIAM J. Comp.*, vol. 6, pp. 323–350 (1977).

[13] D. Harel and R. E. Tarjan, "Fast Algorithms for Finding Nearest Common Ancestor," *Comput. Sys. Sc.*, vol. 13, pp. 338–355 (1984).

[14] D. E. Knuth, *The Art of Computing Programming*, vol. 3: *Sorting and Searching* (Addison-Wesley, Reading, MA, 1973).

[15] P. Weiner, "Linear Pattern Matching Algorithm," in *Proc. 14th IEEE Symp. on Switching and Automata Theory* (1973), pp. 1–11.

[16] B. Baker, "A Theory of Parameterized Pattern Matching: Algorithms and Applications," in *Proc. 25th Ann. ACM Symp. on Theory of Computing* (1993), pp. 71–80.

[17] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, p. 248 (1961).

[18] B. C. J. Moore and B. R. Glasberg, "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753 (1983).

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

[20] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[21] M. E. Frerking, *Digital Signal Processing in Communication Systems* (Kluwer Academic, Dordrecht, 1994).

[22] M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-24, pp. 243–248 (1976).

[23] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proc. IEEE*, vol. 65, pp. 1558–1564 (1977).

THE AUTHORS

Constantin Papaodysseus was born in Athens, Greece. He received a Diploma in electrical and computer engineering from the National Technical University of Athens (NTUA) and an M.Sc. degree from Manchester University, U.K. He received a Ph.D. degree in computer engineering from NTUA. Since 1996 he has been assistant professor in the Department of Electrical and Computer Engineering at NTUA.

His research interests include music and speech processing and automatic recognition, image processing, applied mathematics, algorithm robustness, quantization error analysis, adaptive algorithms, and biomedical engineering. He has published more than 25 publications in international journals and has presented many papers at international conferences on these subjects.

•

George Roussopoulos was born in Athens, Greece, in 1971. He received a Diploma in computer and software engineering in 1994 from the Technical University of Patras and a Ph.D. in computer engineering from the National Technical University of Athens in 2000.

His research interests and recent work include the following subjects: music and speech processing and automatic recognition, image processing, pattern recognition, algorithm robustness, and algorithms for echo cancellation. He has six publications in international journals and has presented 10 papers at international conferences on these subjects.

•

Dimitrios Fragoulis was born in Athens, Greece, in 1973. He received a Diploma and M.Sc. degree in electrical and computer engineering from the National Technical University of Athens in 1996. He is currently a Ph.D. student in computer engineering at the same university.

His research interests and recent work include the following subjects: music and speech processing and automatic recognition, and the study of psycho-

C. Papaodysseus

G. Roussopoulos

D. Fragoulis

Th. Panagopoulos

C. Alexiou

logical and perceptual aspects of sound. He has four publications in international journals and has presented eight papers at international conferences on these subjects.

•

Athanasios Panagopoulos was born in Athens, Greece, in 1973. He received a Diploma and M.Sc. degree in electrical and computer engineering from the National Technical University of Athens in 1996. He is currently a Ph.D. student in computer engineering at the same university.

His research interests and recent work include the following subjects: music and speech processing and automatic recognition, image processing, pattern recognition, and algorithms for echo cancellation. He has three publications in international journals and has presented five papers at international conferences on these subjects.

•

Constantin Alexiou was born in Igoumenitsa, Greece, in 1973. He received a Diploma and M.Sc. degree in electrical and computer engineering from the National Technical University of Athens in 1996. He is currently a Ph.D. student in computer engineering at the same university.

His research interests and recent work include the following subjects: music and speech processing and automatic recognition, algorithm robustness, algorithms for echo cancellation, and biomedical engineering. He has three publications in international journals and has presented seven papers at international conferences on these subjects.