Multiscale Fractal Analysis of Musical Instrument Signals With Application to Recognition

Athanasia Zlatintsi, *Student Member, IEEE*, and Petros Maragos, *Fellow, IEEE*

***Abstract*—In this paper, we explore nonlinear methods, inspired by the fractal theory for the analysis of the structure of music signals at multiple time scales, which is of importance both for their modeling and for their automatic computer-based recognition. We propose the multiscale fractal dimension (MFD) profile as a short-time descriptor, useful to quantify the multiscale complexity and fragmentation of the different states of the music waveform. We have experimentally found that this descriptor can discriminate several aspects among different music instruments, which is verified by further analysis on synthesized sinusoidal signals. We compare the descriptiveness of our features against that of Mel frequency cepstral coefficients (MFCCs), using both static and dynamic classifiers such as Gaussian mixture models (GMMs) and hidden Markov models (HMMs). The method and features proposed in this paper appear to be promising for music signal analysis, due to their capability for multiscale analysis of the signals and their applicability in recognition, as they accomplish an error reduction of up to 32%. These results are quite interesting and render the descriptor of direct applicability in large-scale music classification tasks.**

***Index Terms*—Fractals, multiscale analysis, music signals, timbre classification.**

I. Introduction

MUSICAL content and information analysis is of importance in many different contexts and applications, as for instance, music retrieval, audio content analysis for summarization applications or audio thumbnailing, automatic music transcription, indexing of audio and multimedia databases and other. The above mentioned applications require robust solutions to information processing problems, such as automatic musical instrument classification and genre classification [2], [23], [29]. Toward this goal, the development of efficient digital signal processing methods for the analysis of the structure of music signals and the extraction of relevant features becomes essential. Our paper proposes such methods and algorithms, and investigates an alternative feature-set which quantifies fractal-like structures in music signals at multiple time scales. By using the proposed analysis, we seek to explore whether these methods are capable of characterizing musical sounds for recognition tasks and whether it is possible to relate their properties using such measurements.

Both Plato and Aristotle, in many of their treatises, claimed that music fell under the philosophy of "craft of representation" or otherwise *mimesis*. In other words, music imitates nature, human emotions or even properties of certain objects. On the other hand, Mandelbrot [16] has demonstrated how nature contains structures (e.g., mountains, coastlines, the structures of plants), which could be described by fractals[1] and suggested that fractal theory could be used in order to understand the harmony of nature. Fractals can also be found in other natural processes described by time-series measurements (i.e., $1/f$ noises, pitch and loudness variations in music, demographic data and others). He also recognized the widespread existence of $1/f$ in nature. In this paper, inspired by the fact that music somehow imitates the nature, while ideas from the fractal theory are able to describe it, we aspire to scrutinize their relation.

Analysis of musical structure has revealed evidence of both fractal aspects and self-similarity properties in instrument tones and music genres. Voss and Clark [31] investigated $1/f^{\beta}$ aspects in music and speech by estimating the power spectra for slowly varying quantities, such as loudness and frequency. The fractal and multifractal aspects of different genres of music were analyzed in [3], where it was proposed that the use of fractal dimension measurements could benefit the discrimination of musical genres. Su and Wu [27] applied Hurst exponent and Fourier analysis in sequences of musical notes and noted that music shares similar fractal properties with the fractional Brownian motion. Properties of self-similarity, regarding the acoustic frequency of the signals, were observed in [14], where aspects of fractal geometry were studied. Given this previous evidence of fractal properties in music, such as the fractional Brownian motion, the use of fractal and multifractal dimension for genre classification, and evidences of self-similarity properties found on musical tones, we wish to further explore whether multiscale fractal analysis could manifest supplementary facts about the structure of music signals, taking into account that such methods

Manuscript received April 05, 2012; revised August 06, 2012, October 31, 2012; accepted November 13, 2012. Date of publication November 30, 2012; date of current version January 11, 2013. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece (e-mail: nzlat@cs.ntua.gr; maragos@cs.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2231073

[1] The term 'fractal' was coined by Mandelbrot from the Latin word *fractus*, meaning "broken", to describe objects that are irregular (or "fragmented") to fit within the traditional geometry [16]. He defines a set F as fractal when it has a fractal dimension that exceeds its topological dimension. One of the most important characteristics of fractals is that they have similar structure at multiple scales.

have already been employed successfully in speech recognition applications [19].

In addition to fractals, the theory of chaos in nonlinear dynamical systems has contributed several ideas and methods to model complex time-series. In this area, Lyapunov exponents are among the most useful theoretical and computational tools to quantify various aspects of chaotic dynamics in time-series after their embedding in a phase space. For multiscale analysis of non-stationary signals from the viewpoint of nonlinear complex systems, Gao *et al.* [10], [11] have introduced the concept of a scale-dependent Lyapunov exponent (SDLE) and an efficient algorithm to compute it; further, they have applied it to several engineering and scientific problems.

Over the years, various feature sets have been proposed and pattern recognition algorithms have been employed to solve the complex task of recognizing musical instruments. Such feature sets include perception-based, temporal, spectral and timbral features. Cepstral coefficients have been favored a long way back, not only in speech processing but in musical instrument recognition tasks as well. Brown *et al.* [4] used cepstral coefficients, constant Q transform, spectral centroid and autocorrelation coefficients to identify four instruments of the woodwind family. In [5] the performance of several features was compared, including MFCCs, spectral and temporal features, such as amplitude envelope and spectral centroids for instrument recognition. The results favored the MFCC features, which were more accurate in instrument family classification. Experiments on real instrument recordings [21] also favored the MFCCs over harmonic representations.

Various classification techniques have been used to model instruments' sounds as well, sometimes not necessarily as effective in modeling the temporal evolution of the features. For instance, Gaussian mixture models (GMMs) are capable of parameterizing the distribution of observations, although they cannot model the dynamic evolution of the features within a music tone as, for example, hidden Markov models (HMMs) can do. In [6], the feature distribution of MFCCs and delta-MFCCs was modeled with HMMs, while in [23] Variable Duration HMMs were used for classification of musical patterns.

In our work which is an enlarged version of [32], we propose the *Multiscale Fractal Dimension* (MFD) of musical instrument tones through analysis and experimental validation with recognition experiments. The analysis concerns isolated musical instrument tones where signals are taken from the UIOWA database [30]. First, we examine some of the sound characteristics of musical instruments, the structures and sound properties of musical signals, such as timbre and its complexity, and we highlight issues that should be taken under consideration in the analysis that follows (Section II). Section III concerns the description of the proposed algorithm on multiscale fractal dimension, which is based on previous work by Maragos [17]. The analysis of musical instrument tones is performed separately for the attack and the steady state of the tones, while individualities observed for each instrument are pointed out (Section IV). We further examine our observations by experimentally evaluating the MFDs on synthesized sounds composed by one or more sinusoidal signals (Section V). Finally, we investigate the potential of the proposed algorithm with classification experiments using Markov models. Specifically, we compare the descriptiveness of MFDs with MFCCs (Section VI). We report on promising experimental results that could accomplish an error reduction up to 32%.

II. Musical Structures

People are eager to constantly classify the world around them and sound is not an exception. We try to capture each individual sound with its associated characteristics and categorize it according to various aspects, such as natural versus artificial, original or reproduced, transient or steady or according to the means of its production. The last one, which is probably the most significant, holds also for musical instruments which are classified into different families depending on their construction (shape and material) and physical properties. The four main categories or families are: strings (e.g., violin, upright bass), woodwinds (e.g., clarinet, bassoon), brass (e.g., horn, tuba) and percussion (e.g., piano). However, the main attribute that distinguishes musical instruments from each other is timbre. The determination of timbre by the waveform constitutes one of the main relations among sound attributes and relates to our perception of complex sounds. This relation is one of the most difficult to describe (in contrast to i.e., loudness or pitch), since both timbre and waveform are two complex quantities. All complex sounds, such as musical instruments' sounds, are a combination of different frequencies which are multiples of the fundamental frequency f_0 (e.g., $f_0, 2f_0, 3f_0, 4f_0$ and so on). This property is referred to as "harmonicity" and the individual frequencies as harmonics.

Timbre, according to ASA (American Standards Association) [1], is the quality of sound which distinguishes two sounds of the same pitch, loudness and duration and is thus associated with the identification of environmental sound sources. Loosely explained, timbre—also referred as tone color or tone quality—could be defined as the number and relative strength (shaping) of the instrument's harmonics (amplitude distribution) [25], as a consequence of the structural resonances of the instrument. Fletcher [8] showed that this analogy is not that simple since timbre depends on the fundamental frequency and the tonal intensity of a tone as well. In conclusion, timbre depends on the absolute frequencies and relative amplitudes of pure tone components varying in musical instruments from dull or mellow (strong lower harmonics) to sharp and penetrating (strong higher harmonics).

Some of the instruments' sound characteristics are going to be briefly mentioned next, based on Olson's [22] descriptions. **Flute**, in contrast to most musical instruments, has a fundamental frequency which carries a significant amount of the acoustical energy output. Low registers[2] are richer in harmonics, while in high registers the harmonics are practically nonexistent making the tones sound clean and clear. The fact that the fundamental frequency carries this amount of energy results in the distinctive sound of flute which is the thinnest

[2] Fig. 1 shows the frequency ranges of the described instruments.

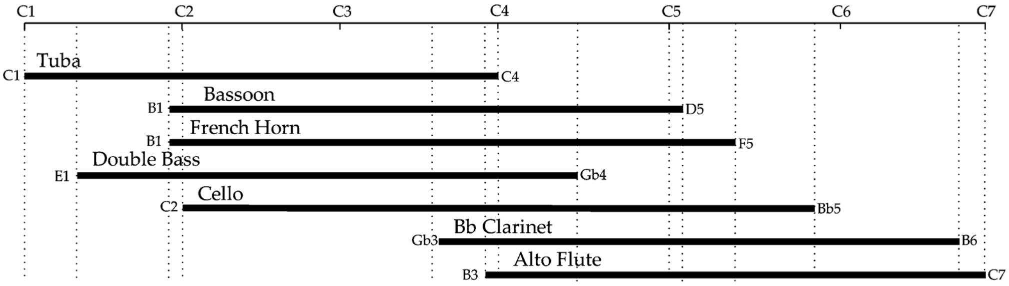

Fig. 1. Frequency ranges of the analysis instruments where the overlap between them can be seen.

and purest of all instruments. In **clarinet** as well, most of the energy resides in the fundamental which makes the sound clear and bright. In lower registers it produces powerful tones, while the even harmonics are suppressed due to the cylindrical pipe which is closed at one end. In contrary, in **bassoon** the fundamental frequency and lower harmonics are low in terms of intensity in low registers, while **tuba** produces large output in the low-frequency region. **Horn**, on the other hand, plays in a higher portion of its harmonics compared to most brass instruments, while its conical bore is assumed to be responsible for its distinctive sound which is often described as "mellow". Finally, the harmonic content of a **double bass** is very high in low register.

Although it is quite easy for people, and especially trained musicians, to recognize the different instruments, this is not the case if only the steady middle state of the note is heard. The difficulty in differentiating timbre lies also in its multidimensionality and the fact that it cannot be represented by 1D scales, which would be used for comparison or ordering [25]. Instrument recognition depends a great deal on hearing the transients of a tone, meaning the beginning (attack) and the ending (release) [13], since they have noise-like properties influencing their subjective quality. For instance, flute with its relatively simple harmonic structure in order to obtain its distinctive sound, it should be preceded by a small "puff" or noise. This is a characteristic sound element that cannot be accomplished by a synthetic sound [20], and it would disappear if only the steady state of the tone would be present. The same applies for trumpet as well, while similarly, it is vital to hear the scrape of the bow on a violin string, or the squeak of a clarinet [13]. Iverson *et al.* [15] compared the timbre contribution of the attacks and steady states of orchestral instruments' tones and concluded that both contributions are roughly comparable, indicating that the salient attributes for complete tones are present in both states. However, it is mentioned that the absence of the attack could negatively affect the determination of whether an instrument is stuck, blown or bowed.

The duration of those transients varies not only among instruments but between higher and lower octave tones as well. Some typical attack durations, Hall [13] reported, are from 20 ms or less for oboe, 30–40 ms for clarinet or trumpet, to 70–90 ms for flute or violin. Additionally, notes above middle C (designated as C4 at ca. 261 Hz) have periods of 2–4 ms, resulting in several dozen vibration periods for the steady state to be established. However, in [9] the duration of the attack is reported as 50 ± 20 ms, independently of the tone or the instrument. Because of such evidence concerning the differences of the tones' transients, we assume that the whole duration of a tone gives vital clues for its identity. Fig. 2 shows the attack, steady state and release for the note A3 of Bb Clarinet.

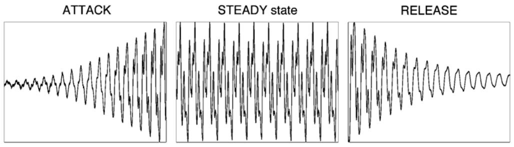

Fig. 2. Attack, steady state and release for Bb Clarinet A3.

In many applications, classification down to the level of instruments families could be sufficient. However, in our approach, we focus on the distinction between individual instruments, pointing out similarities observed for the families. Our main hypothesis is that the multiscale fractal dimension can help distinguish the instruments' timbre by discriminating, not only the steady state of the tones, but the attacks as well.

III. Multiscale Fractal Dimension

Most features extracted from music signals, for classification purposes, are inspired by similar work in speech and so are the fractal features used in this paper. Many speech sounds contain some amounts of turbulence at some time scales. Mandelbrot [16] conjectured that multiscale structures in turbulence can be modeled using fractals. Such ideas motivated Maragos [17] to use the *short-time fractal dimension* of speech sounds, as a feature to approximately quantify the degree of turbulence in them. He also developed in [17], [18] an efficient algorithm to measure it, based on the Minkowski-Bouligand dimension D [7], [18]. This measures the multiscale length of (possibly fragmented) curves F by the creation of a "Minkowski cover", i.e., the covering of F with disks of varying radius s, whose center lies on the curve. The developed algorithm is referred to as the *morphological covering method* and the steps that are followed in this paper as well are:

Step 1: Create the Minkowski cover using *two-dimensional* operations, i.e., morphological set dilation (a.k.a. Minkowski sum) $\oplus$ of the graph F of the signal by multiscale versions $sB = \{sb : b \in B\}$ of a unit-scale convex symmetric planar set B, where $s \geq 0$ is the scale parameter:

$$F \oplus sB = \{z + sb \in \mathbb{R}^2 : z \in F, b \in B\}. \qquad (1)$$

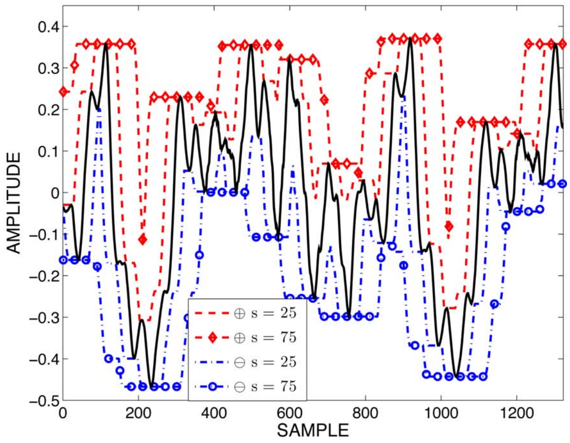

Fig. 3. Double Bass steady state (solid line), its multiscale flat dilations $\oplus$ and erosions $\ominus$ at scales $s = 25, 75$.

Then, compute the cover area $A_B(s) = \text{area}(F \oplus sB)$ of the dilated set at multiple scales. Finally, the following limit of the cover area on a log-log scale yields the fractal dimension:

$$D = \lim_{s \to 0} \frac{\log[A_B(s)/s^2]}{\log(1/s)}. \tag{2}$$

Ideally B is a unit disk. However, D remains invariant as long as B is compact, convex and symmetric [18]. In the discrete case, we select as B an approximation to the disk by a unit-radius convex symmetric subset of $\mathbb{Z}^2$.

Step 2: In [17], [18] Maragos has shown that the above limit for computing D will not change if we approximate $A_B(s)$ with the area of the difference signal between the morphological function dilation $\oplus$ and erosion $\ominus$ of the N-sample discrete signal $F[n]$ by a function $G_s[n]$ that is the upper envelope of the s-scaled discrete set sB:

$$A_B(s) = \sum_{n=0}^{N-1} ((F \oplus G_s) - (F \ominus G_s))[n], \tag{3}$$

for $s = 1, \ldots, s_{\max} \leq N/2$. This greatly reduces the complexity because instead of two-dimensional set operations we perform *one-dimensional* signal operations that are simple nonlinear convolutions. A further reduction of complexity $[O(N^2)$ to $O(N)]$ is accomplished by performing the above signal operations in a *scale-recursive* way:

$$\begin{aligned} F \oplus G[n] &= \max_{-1 \leq k \leq 1} \{F[n+k] + G[k]\}, s = 1 \\ F \ominus G[n] &= \min_{-1 \leq k \leq 1} \{F[n+k] - G[k]\}, s = 1 \\ F \oplus G_{s+1} &= (F \oplus G_s) \oplus G, 2 \leq s \leq s_{\max} \\ F \ominus G_{s+1} &= (F \ominus G_s) \ominus G, 2 \leq s \leq s_{\max}. \end{aligned} \tag{4}$$

where $G = G_1$ and $s = 1, \ldots, s_{\max} \leq N/2$.

The signal dilations and erosions, which are computed in our case, have computational structure similar to convolution and correlation, respectively [18]. They create an area strip as a layer either covering or being peeled off from the graph of the signal at various scales. Fig. 3 shows a special case where B is a 3-sample symmetric horizontal segment with zero height, which implies that $G[n]$ equals zero for $n = -1, 0, 1$ and $-\infty$ elsewhere. This special case yields the *fastest* multiscale covering algorithm, because the corresponding function dilations and erosions simply become local max and min within a moving window; further, the resulting fractal dimensions are invariant to any affine transformation of the signal's range.

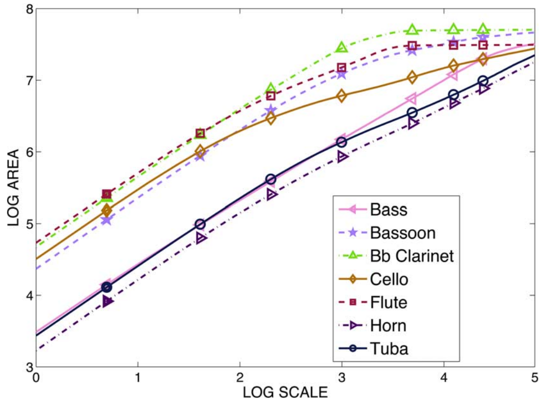

Fig. 4. $\log[A_B(s)]$ vs $\log(s)$ for the seven analyzed instruments for the note C3 except for Bb Clarinet and Flute shown for C5 instead.

Step 3: In practice, D can be estimated by least-squares fitting a straight line to and measuring the slope of the plot of the data $\log[A_B(s)]$ versus $\log(s)$, because

$$\log[A_B(s)] = (2 - D)\log(s) + \text{constant}, \quad \text{as } s \to 0 \tag{5}$$

assuming that $A_B(s) \approx s^{2-D}$ as $s \to 0$. However, real-world signals do not have the same structure over all scales, and hence the exponent in the dominant power s^{2-D} may vary. Thus, we compute the slope of the data $\log[A_B(s)]$ versus $\log(s)$ over a small scale window $\{s, s+1, \ldots s+w\}$ of w scales that can move along the s axis. This process creates a profile of local *multiscale fractal dimensions (MFDs)* $D[s,t]$ at each time location t of the short speech analysis frame. The local slope of this line is an estimate of $2 - D$ and gives us the fractal dimension. Throughout this paper, we have used $w = 10$. Fig. 4 shows a plot of $\log[A_B(s)]$ versus $\log(s)$ for various instruments. Note the difference in the slope for larger scales s. Additionally, D ranges between 1 and 2 for topologically one-dimensional signals (i.e., for continuous functions of one variable); the larger D is, the larger the amount of geometrical fragmentation of the signal graph. D is estimated at the smallest possible discretized time scale as a short-time feature for purposes of audio signal segmentation and event detection. The function $D[s,t]$ can also be called a *fractogram* and can provide information about the degree of turbulence inherent in short-time sounds at multiple scales [17], [19].

The specific algorithm is also significant because of its linear computation complexity, $O(N)$ additions, assuming a N-sample signal, since the required min-max operations are computationally equivalent to additions. Comparing to MFCCs, $O(N \log N)$ multiplications, which throughout the experimental evaluation are used for comparison purposes, we see that the use of MFDs is advantageous since they offer a simple computational solution.

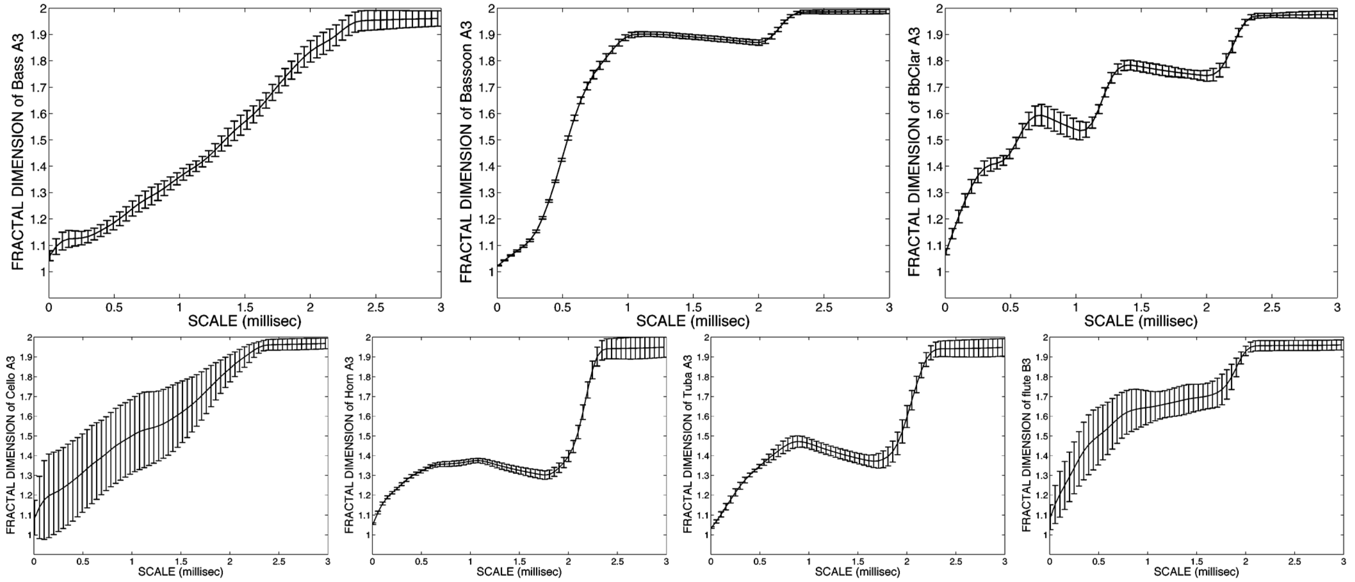

Fig. 5. Mean MFD (middle line) and standard deviation (error bars) of the same note A3 for the instruments Double Bass, Bassoon, Bb Clarinet (first row) and Cello, Horn and Tuba, and the note B3 for Flute (second row) (for 30 ms analysis window, updated every 15 ms).

In general, the short-time fractal dimension at the smallest discrete scale $(s = 1)$ can provide some discrimination among various classes of sounds. At higher scales, the MFD profile can also offer additional information that helps the discrimination among sounds. Actually, the research from [19] and [24] has shown evidence that, such MFD features (in conjunction with other standard features), can provide a modest improvement in recognition performance for certain word-recognition tasks over standard speech databases. In this paper, we have used MFDs as an efficient tool to analyze the structure of music signals at multiple time scales. The results are quite interesting, as we will present further down, by also showing examples of MFDs for music signals from various instruments.

IV. MFD Analysis on Musical Signals

A. MFD on Steady State

Our analysis is not only based on the distinction of different instruments, but on the exploration of the differences between the attack and steady state of the tones as well. We intent to show that the multiscale fractal dimension distribution of the attacks differs enough on different instrument tones, managing to add adequate information in a recognition task.

For the analysis of the steady state we used the whole range of tones from the following instruments: Double Bass, Bassoon, Bb Clarinet, Cello, Flute, French Horn and Tuba. The calculation of the short-time MFDs of the tones was performed using 30 ms segments of the full duration of the tones. However, for the state-specific analysis that follows, only the appropriate segments have been processed. The signals were sampled at 44.1 kHz, and their corresponding profiles of MFD$[s]$ were analyzed for discrete scales $s = 1, \ldots, 133$, corresponding to time scales s_t from 1/(44.1) to 3 ms. Similar results were also obtained from the analysis of 50 ms windows.

Fig. 5 shows the mean MFD and standard deviation (error bars), computed for the note A3 for all analyzed instruments, except Flute which is shown for B3 instead. The MFD profile presented is typical for the following octaves of each instrument (see Fig. 1 for the instruments' frequency ranges and the overlap there is between them): Double Bass for the whole range, Bassoon for octaves 3–5, Bb Clarinet for octaves 3–4, Cello for octaves 2–4, Flute for octaves 3–4 and Horn for octaves 3–5. Fig. 6 shows the MFD profiles for the lower octaves of Bassoon, Tuba and Horn (octaves 1–2), where they appear to have certain similarities, i.e., they get their first peak and higher value D at about $s_t = 0.5$ and then decrease to an intermediate value. Still, they exhibit some important differences; the maximum D is at about 1.8 for Bassoon, while Tuba and Horn share the values of ca. $D = 1.5$. Further, Tuba shows more important deviations of D across the successive analysis frames of each tone than Horn. Regarding the higher octaves of Bb Clarinet and Flute (octaves 5–6) (see Fig. 6 second row), we observe another tendency. The MFD profiles for those ranges get their higher value at around $D = 1.9$ at small time scales, ca. $s_t = 0.8$, and behold it throughout the whole profile. The analysis of Double Bass and Cello has shown more uniform in shape MFD profiles with an increased deviation of D across frames for lower range tones. To conclude, apart from the last two cases, for the rest of the analyzed instruments specific differences are observed between the lower and higher octaves still with unvarying characteristics across the particular octave ranges, as already discussed. Table I presents the averaged values of the instruments' related MFDs for the steady state averaged over the whole range of each instrument (and dynamic range forte) and for specific time scales s_t assumed nodal points after the analysis. In the brackets, the standard deviation is calculated to demonstrate the variability observed for the specific scales. For those measurements, we did not take into account the variability of MFDs through the different octaves as discussed above. The most homogeneous with less variability MFD profiles are noted for Horn, Tuba and Bassoon for smaller scales, and for Bb Clarinet and Flute for larger scales. Analysis of the multiscale fractal dimension on the

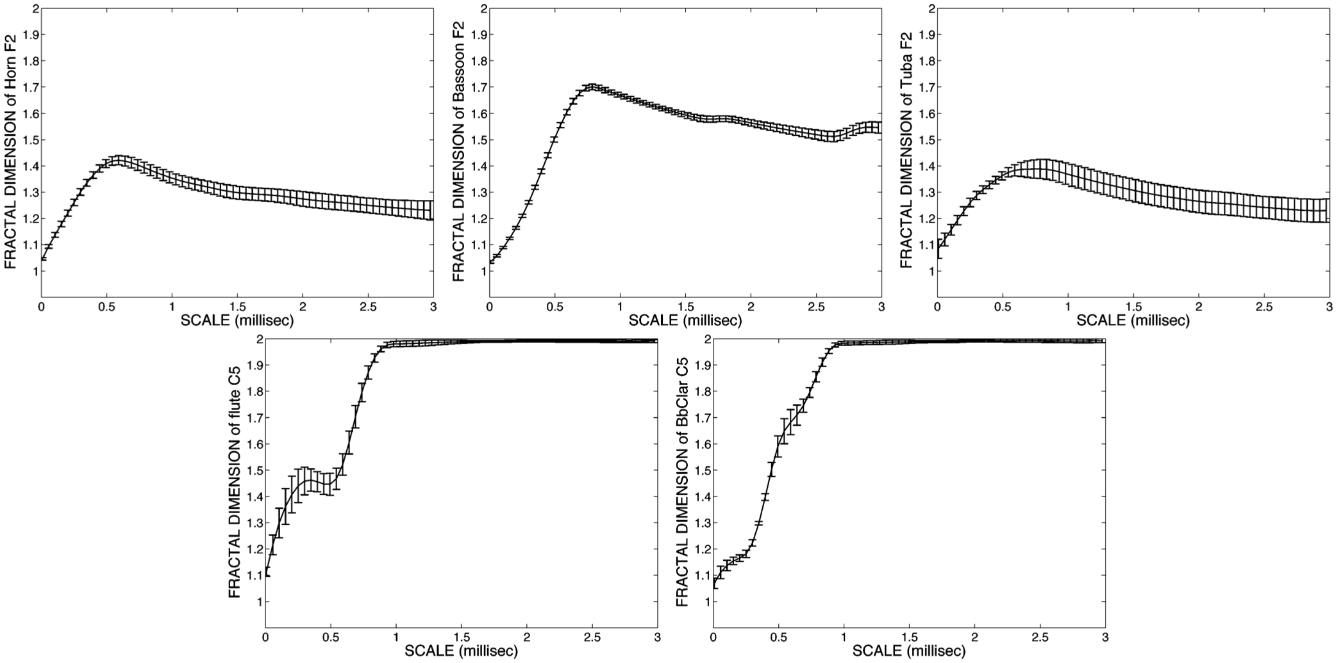

Fig. 6. Mean MFD and standard deviation (error bars) for the note F2 for the instruments Horn, Bassoon and Tuba (first row) and the note C5 for Flute and Bb Clarinet (second row). The MFD profiles shown are typical for the lower octaves of the three first row instruments, respectively for the higher octaves of the two second row instruments (30 ms analysis window, updated every 15 ms).

TABLE I
AVERAGED MFD AND STANDARD DEVIATION FOR VARIOUS TIME SCALE POINTS s_t OF THE MFD PROFILES

	Averaged MFDs (Standard Deviation)					
Time Scale (ms)	$s_t = 1/44$ (ms)	$s_t = 0.5$	$s_t = 1$	$s_t = 1.5$	$s_t = 2$	$s_t = 2.5$
Double Bass	1.11 (0.050)	1.21 (0.037)	1.31 (0.040)	1.39 (0.040)	1.52 (0.039)	1.61 (0.038)
Bassoon	1.04 (0.004)	1.47 (0.006)	1.75 (0.070)	1.78 (0.080)	1.80 (0.090)	1.83 (0.010)
Cello	1.12 (0.017)	1.47 (0.066)	1.63 (0.076)	1.73 (0.077)	1.80 (0.067)	1.85 (0.058)
Bb Clarinet	1.14 (0.035)	1.69 (0.033)	1.84 (0.035)	1.90 (0.027)	1.95 (0.021)	1.96 (0.017)
Flute	1.13 (0.018)	1.77 (0.036)	1.90 (0.037)	1.95 (0.021)	1.98 (0.010)	1.98 (0.010)
French Horn	1.06 (0.002)	1.38 (0.006)	1.49 (0.009)	1.54 (0.019)	1.59 (0.022)	1.64 (0.024)
Tuba	1.10 (0.026)	1.35 (0.013)	1.40 (0.120)	1.36 (0.015)	1.38 (0.017)	1.42 (0.022)

steady state of the instruments' tones reinforces the claims that the MFDs convey information that is instrument related. Even for the cases of instruments belonging in the same family or the same frequency range showing similar tendencies, specific differences can be observed regarding the dimension D, the scale s_t, or the deviation of D across scales. Finally, we notice a dependence of the MFD on the acoustical frequency of the sound, which will be further explained in Section IV.C.

B. MFD on Attack

Acoustic characteristics of an instrument's attack may be uniquely important in order to determine whether it is struck, blown, or bowed [15]. Continuing the analysis, we perform an analogous study on the attacks of the instruments' tones to explore possible alterations. The configuration is similar to the prior one and the process takes place after considerations of the individualities presented on the attack of each instrument, e.g., the duration. The MFD profiles for the attack present similar tendencies as the steady state of the tones. However, some of the differences observed are the following: they have higher D for small scales s_t, and they present more fragmentation in comparison to the steady state. Those two alterations could be possibly explained by noise-like factors in the beginning of the tones as discussed in Section II and the fragmentation of the waveform. Fig. 8 shows the average MFDs for the attack for the whole range of the analyzed instruments (dynamic range forte). In this case, we notice an increased value of $D(s = 1)$ and a quite clear distinction of D among some of the analyzed instruments. In conclusion, the analysis of the attack has shown certain differences, both between attack and steady state of the same tone and among the instruments as well. This is of significant importance since it could mark the transition from attack to steady state, while it simultaneously carries instrument-specific information. Fig. 7 shows examples of the attack and steady state for the notes A3 for Cello and F4 for Flute. For Cello a higher $D(s = 1)$ and more fragmented profile is observed on the attack, while for Flute the two states present more similarities, however, the attack has its own individualities.

C. MFD Variability for Each Instrument

An important finding of our study concerns the analysis of the MFDs for individual tones of the same instrument. Fig. 9 shows the MFD profiles for the tones C4-B4 of Bb Clarinet over one octave, with frequencies between ca. 260–493 Hz, which confirm the preceding evidence of this study that there is a dependency of the MFD profile on the acoustical frequency of the

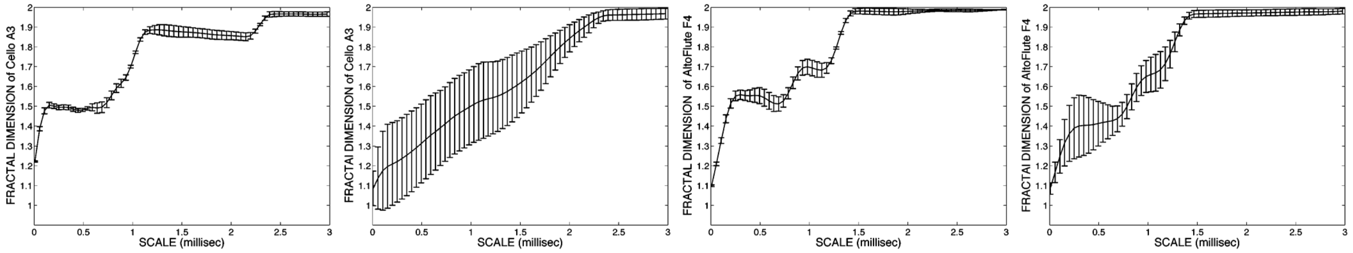

Fig. 7. Mean MFD and standard deviation of the attack and steady state of A3 for Cello (left images) and F4 for Flute (right images).

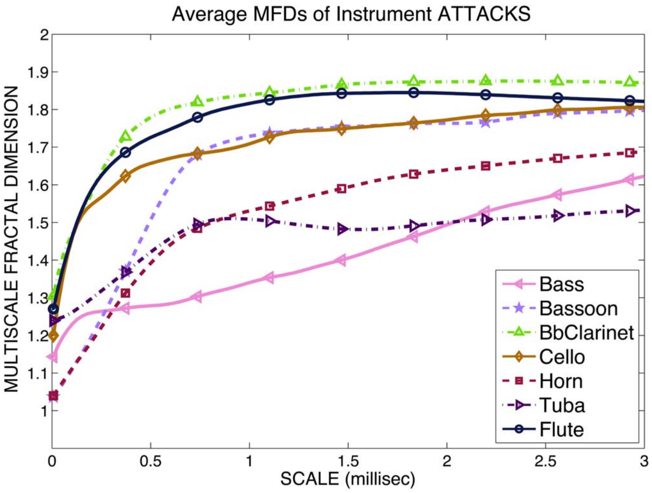

Fig. 8. MFDs estimated for the 7 analyzed instruments attacks, averaged over the whole range (using 30 ms analysis windows). (Please see color version for better visibility).

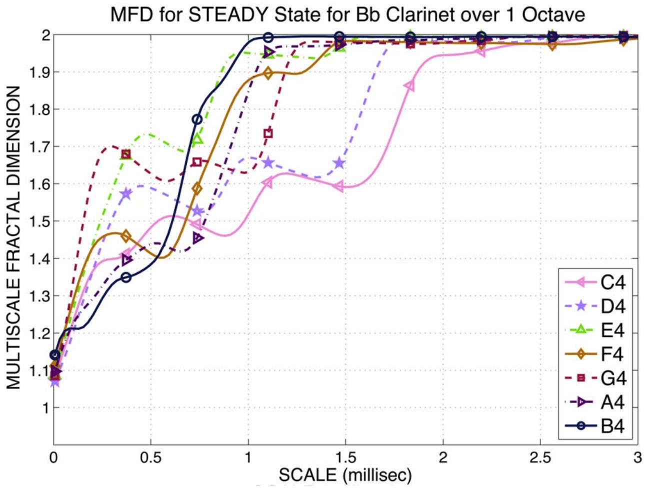

Fig. 9. MFD of Bb Clarinet steady state notes, over one octave for one 30 ms analysis window. (Please see color version for better visibility).

sound. We notice that the profile increases hastily for higher frequency sounds (i.e., gets its first peak and then its highest D for smaller scales s_t). Still, the instrument's specific MFD profile beholds the shape (most of the bends and sharp edges) observed for the specific octave ranges, as discussed in Section IV.A. This phenomenon with instrument specific variabilities is observed mostly in woodwinds and brass instruments, while it starts developing at about the frequency range shown in Fig. 9 (i.e., C4 and above).

In the next section, supplementary analysis about the frequency dependency along with other characteristics already discussed will be further explored using synthesized signals, pure and complex tones, composed by sinusoidal signals. However, these last findings give us evidence that the MFDs could be useful not only for the discrimination of different instrument classes but possibly for a proximate interpretation of the acoustical frequency distribution of the tone as well.

V. MFD Analysis on Synthesized Signals

We apply the MFD algorithm to smaller and more manageable synthesized signals, such as simple or complex sinusoidal signals, in order to evaluate observations made in our previous analysis, e.g., the MFD deviation across analysis frames of individual notes and the variability of MFD profiles for the same instrument, but different frequency ranges. In this experimental analysis, we isolate and vary individual parameters of the sinusoids while holding all the others constant. The examined cases are: (i) Simple sinusoidal tones of different frequencies. (ii) Composite sinusoidal tones where sinusoids of higher frequencies are added while keeping constant or reducing the amplitude, and simultaneously keeping constant or varying the phase. (iii) Simulation of a "tone" of certain frequency while adding sinusoids of frequencies equal to its harmonics, and finally, (iv) simulation of a "tone" while individual harmonics are missing in order to imitate instruments, such as the clarinet, which generally plays only the odd members of the harmonic series, e.g., $f_0, 3f_0, 5f_0$ etc. The configuration used for the experimentation is similar to the previous one.

Single Sinusoids: Fig. 10 shows the mean and standard deviation (error bars) of the MFD profiles for the simplest case of single sinusoidal signals using different frequencies. The frequencies used are 5, 100, 300, and 500 Hz. The amplitude and phase are constant equal to 1 and 3/4 of a cycle, respectively. We observe that the MFD profile shows a dependency on the frequency of the signal. Specifically, the first peak realizes at half the period. Note on the last figure where the frequency is 500 Hz, the first peak of the MFD profile is at about 1 ms.

Complex Signals With Sinusoids of Double Frequency: Fig. 11 shows the mean MFD and standard deviation of sinusoidal signals, when successively adding sinusoids of double frequency to the initial sine of frequency 50 Hz. The frequencies of the added sinusoids are: 100, 200, 400, 800 Hz. The amplitude and phase remains constant. Here, we notice that the structure of the MFD profile shows more bends while the number of sinusoids increases.

Complex Signals With Sinusoids of Different Frequencies, Amplitudes and Randomly Chosen Phases: In Fig. 12 sinusoids

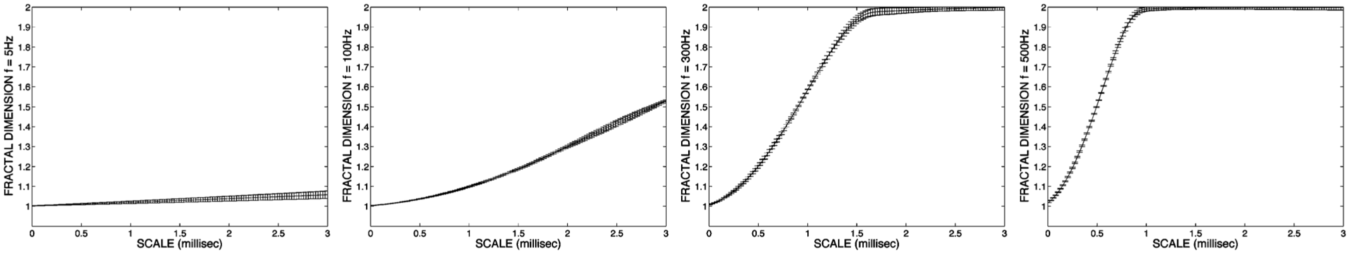

Fig. 10. Mean MFD and standard deviation (error bars) of simple sinusoidal signals with frequencies 5, 100, 300, and 500 Hz.

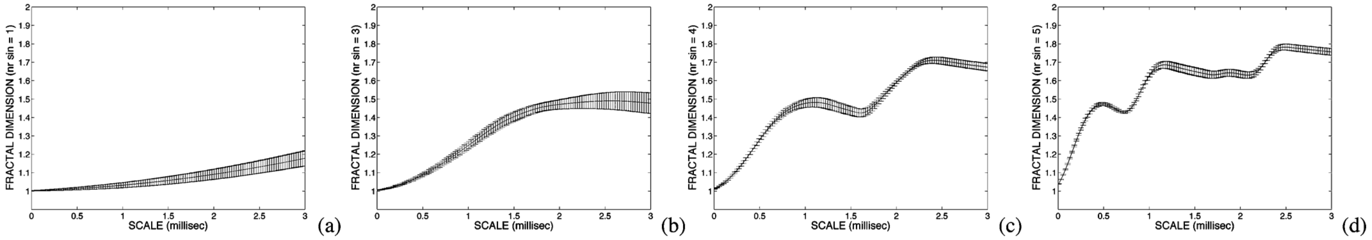

Fig. 11. Mean MFD and standard deviation of synthesized sinusoidal signals. (a) Initial sine $x_0 = x_{50}$ (50 Hz), (b) $x = x_{50} + x_{100} + x_{200}$, (c) $x = x_{50} + x_{100} + x_{200} + x_{400}$, and (d) $x = x_{50} + x_{100} + x_{200} + x_{400} + x_{800}$.

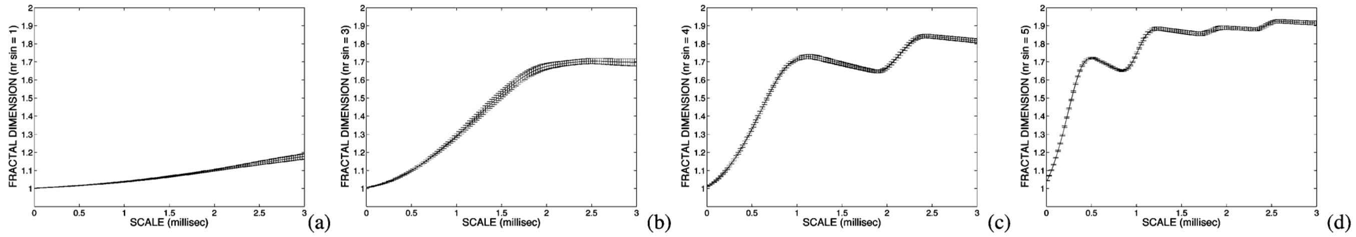

Fig. 12. Mean MFD and standard deviation of synthesized sinusoidal signals while sines of double frequency and geometrically reduced amplitude are added. (a) Initial sine $x_0 = x_{50,1}$ (where 50 in Hz and 1 the amplitude), (b) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4}$, (c) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4} + x_{400,1/8}$, and (d) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4} + x_{400,1/8} + x_{800,1/16}$. The phase offset is randomly varied.

of different frequency, amplitude and randomly chosen phase $[0, 2\pi]$ are added to the initial signal x_0 of 50 Hz and amplitude equal to 1. The frequencies and amplitudes of the added sinusoids are: $x_1 = 100, 1/2, x_2 = 200, 1/4, x_3 = 400, 1/8$, and $x_4 = 800, 1/16$. Here, we observe that the reduced amplitudes do not really affect the profile, while the randomly chosen phases increase somehow the overall MFD for scales greater than $D(s = 1)$. However, the shape and structure of the profile still develop similarly. Our observations seem to be consistent with the fact that the phase does not contribute to the perception of timbre, but produces only small changes in the sensation produced upon the listener [8], [25].

"Simulation" of C3: In Fig. 13 a simulation of the tone C3 is attempted by using sinusoids of frequencies equal to the fundamental frequency $f_0 = 131$ Hz and the harmonics of C3. The amplitude and phase remains constant. The frequencies of the successively added sinusoids are integer multiples of f_0, i.e., $f = 262, 393, 524, 655, 789, 917, 1046$ Hz. Note an increased variation of D across different analysis frames for the signals that are composed of six or more sinusoids.

"Simulation" of C3 Using Frequencies Equal to the Odd Harmonics: In Fig. 14, the mean MFD and standard deviation of the simulation of the tone C3 can be seen, while adding sinusoids of frequencies equal only to the odd harmonics of the tone. Here, we attempt to imitate instruments such as the clarinet while trying to determine whether such characteristics of the instruments' harmonic content could be visible in the shape of the MFD profile. The sinusoids used for this experiment have frequencies equal to $f = 131, 393, 655, 917$ Hz. The amplitude and phase remains constant. In this case, note how the MFD profiles differentiate when individual frequencies are missing; higher multiscale fractal dimensions and more complex structures are observed. In the case where the amplitudes of the even harmonics were just lowered to half, certain changes were observed, although not as significant.

After the analysis of the MFDs on synthesized signals, we can conclude that there is a dependency between the MFD profile and the frequency of the signal and this is manifested both for simple and more complex signals. Additionally, the number of the sinusoids added to the initial signal affects the shape of the MFD profile, which becomes more complicated in structure while an increased variation of D across the analysis frames may be observed as well. Finally, the short-time fractal dimension at the smallest discrete scale $D(s = 1)$ gets higher when a random signal is added to the initial signal.

Even though there is no direct comparison of the previous synthesized signals to the more complex instrument tones, we believe that we gain significant insight concerning some of the differences observed among the analyzed instruments, the MFD profiles for tones of the same instrument and even across a single tone. For instance, the fact that the attack of some instrument tones shows a higher fractal dimension at the smallest scale $D(s = 1)$ could possibly imply the existence of noise-like factors. The increased deviation of D for lower octaves, as in Tuba,

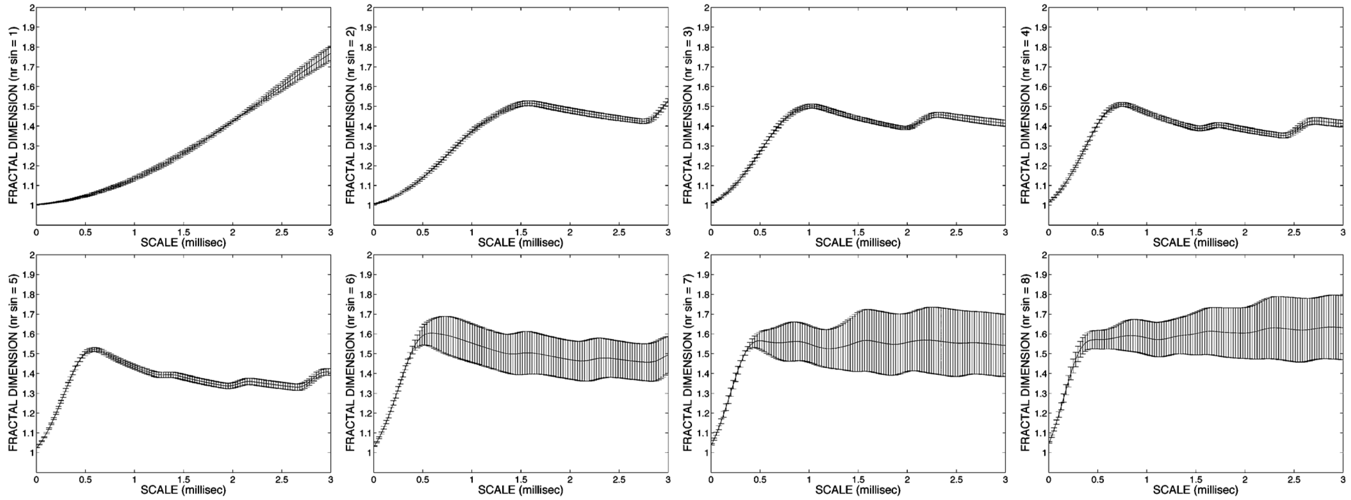

Fig. 13. Mean MFD and standard deviation of sinusoidal signals while adding sinusoids of frequency equal to the harmonics of C3 (131 Hz).

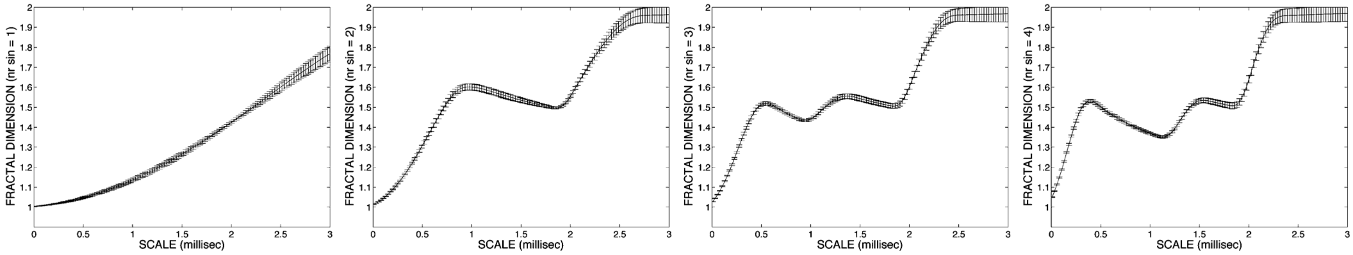

Fig. 14. Mean MFD and standard deviation of sinusoidal signals while adding sinusoids of frequency equal to the odd harmonics of C3.

could point towards a richer harmonic content. The fact that the MFD profiles differ when the frequency content of the tone changes (e.g., for higher frequencies) could give us an indication of the relative position of a tone on the musical scale and an intuitive approximation of the actual frequency distribution of the signal. Although synthesized tones, consisting of steady component frequencies, could not really simulate tones of real instruments, since such a synthesis cannot produce the dynamic variations of the instrument's envelope characteristics, these experiments gave us a somehow better understanding about the perception of real musical instrument sounds.

VI. Recognition Experiments

A. Data

It has been demonstrated that the multiscale fractal dimension could be used to distinguish different musical instruments. In this section, we attempt to incorporate the MFDs to recognition experiments in order to evaluate the results of our previous analysis. The experiments were carried out using 1331 notes, the full range from 7 different instruments, which are Double Bass, Bassoon, Cello, Bb Clarinet, Flute, Horn and Tuba; and they cover the dynamic range from piano to forte. The analysis was performed in 30 ms frames with a 15 ms overlap.

To efficiently succeed, it is essential that the incorporated features contain information that is relevant to the classification task, and that the dimensionality of the final feature set is small enough to accomplish the best possible computational performance. To achieve this, dimensionality reduction of the MFD feature space was conducted using PCA analysis, so as to decorrelate the data and obtain the optimal number of features that accounts for the maximal variance. Additionally, other dense and non-redundant feature sets emerged after sampling of the feature space (logarithmically or by observation). The final feature sets and feature set combinations were evaluated using static Gaussian mixture models (GMMs) and dynamic hidden Markov models (HMMs), to model the temporal characteristics of the signals, with diverse combinations of N states and/or M mixtures. The performance of the selected features was compared to a standard feature set of 12 MFCCs plus the energy, separately or enhanced with their first and second temporal derivatives. MFCCs were chosen both for their good performance and the acceptance they have gained in instrument recognition tasks. The analysis of the MFCCs was performed in 30 ms windowed frames with a 15 ms overlap, using 24 triangular bandpass filters. For the implementation of the Markov models the HTK [28] HMM-recognition system was used, by EM estimation using the Viterbi algorithm. In all cases, the train sets were randomly selected to be the 70% of the available tones, and the results presented are after a five-fold cross validation.

B. Experimental Configuration

Aside from the five sets of features that were evaluated during previous experiments, see [32], all feature sets were enhanced with their first and second temporal derivatives. Table II shows the feature sets with Δ's which are going to be discussed next, however, lots of further examination preceded the final feature selection. In the case of the sampled feature sets two different configurations were considered: a) the logarithmically sampled feature set (consisting of thirteen sample points), example of which can be seen in Fig. 15, augmented with its first and second temporal derivatives MFDLG$_{\Delta}$ and b) an enhanced

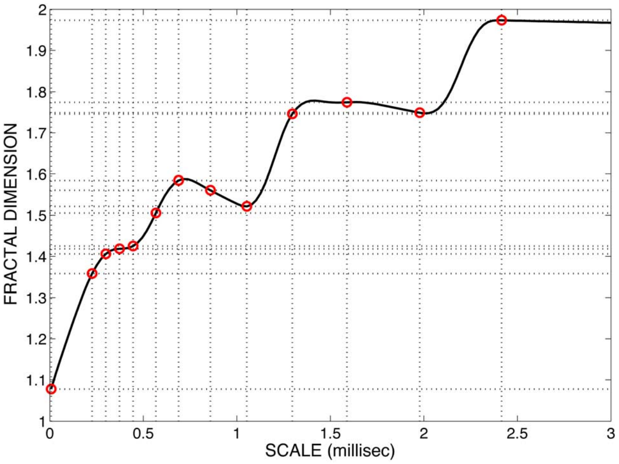

Fig. 15. Example of the 13 logarithmically sampled points of the MFD, for Bb Clarinet (A3), forming the MFDLG feature vector.

TABLE II
LIST OF ENHANCED FEATURE SETS WITH ΔS. MFDPC$_{\Delta i}$ DENOTES PCA ANALYSIS ON THE INDIVIDUAL FEATURE SETS, WHILE MFDPC$_{\Delta f}$ ON THE FULL CONCATENATED FEATURE SET

	Feature Sets
1	13 MFDPCs + 13 Δ + 13 $\Delta\Delta$ (MFDPC$_\Delta$)
2	13 MFDLGs + 13 Δ + 13 $\Delta\Delta$ (MFDLG$_\Delta$)
3	13 MFCCs + 13 Δ + 13 $\Delta\Delta$ (MFCC$_\Delta$)
	Multi-stream cases
4	39 MFDPC$_{\Delta i}$ + 39 MFCC$_\Delta$
5	39 MFDPC$_{\Delta f}$ + 39 MFCC$_\Delta$
6	39 MFDLG$_\Delta$ + 39 MFCC$_\Delta$

feature vector (MFDLGOB) consisting of twenty-four sample points, namely, the MFDLG plus eleven more points carefully chosen after observation. The MFDLG feature vector consisted of D, where $s = 1, 10, 13, 16, 19, 24, 29, 36, 44, 54, 66, 82$, and 100, while the MFDLGOB was augmented with D at sample points $s = 3, 5, 6, 39, 48, 57, 64, 74, 91, 115, 122$. Both sets included the fractal dimension at the smallest scale $D(s = 1)$. Experimentation was also carried out concerning the sets where the PCA analysis would be applied. The two cases considered were: i) PCA on the concatenated feature set of the MFDs with Δ's or ii) on the three individual features sets: the MFD feature vector, its first, and its second temporal derivatives.

After several evaluations of the features and since the MFDLGOB$_\Delta$ gained comparable results with the MFDLG$_\Delta$, we only report results for the MFDLG$_\Delta$. Regarding the PCA applied on the concatenated or the individual feature sets, we notice that when applied to the individual feature vectors, resulting in a 13-dimensional vector from each set (in total 39 features), there is in general an increase in recognition. However, good results are also gained by applying PCA in the concatenated feature vectors consisting of 30, 32 or 39 principal components, some of which are reported next.

The evaluation employed the variation of the number of states N [3–9] and the number of mixtures M [1–5] using GMMs up to 5 mixtures and HMMs up to 9 states. Considering the structure of the instruments' tones, as discussed in previous sections, we adopted a left-right topology for the modeling. In addition, we used multi-stream modeling to separately model the two different sets of features (i.e., MFD versus MFCC) using different stream weights to indicate the reliability of each stream. Stream weights can either be fixed by hand to some values that reflect the relative confidence on one stream or they can be estimated [12], [26]. In this paper, the optimization of the weights was performed on a hold-out set, which was selected from the initial train set (the 70% of the initial train set was split and 60% was used for training and 10% formed the hold-out set). For the experimentation, we assumed that the two stream exponents w_1, w_2 satisfied the constraints $0 \leq w_1, w_2 \leq 1$ and $w_1 + w_2 = 1$. The stream weights that maximized the accuracy on the hold-out set were selected and applied to the actual test set. Fig. 16 shows the accuracy obtained on the hold-out data after five-fold cross validation, while total accuracy results on the test set, which are going to be discussed next, are shown in Table III.

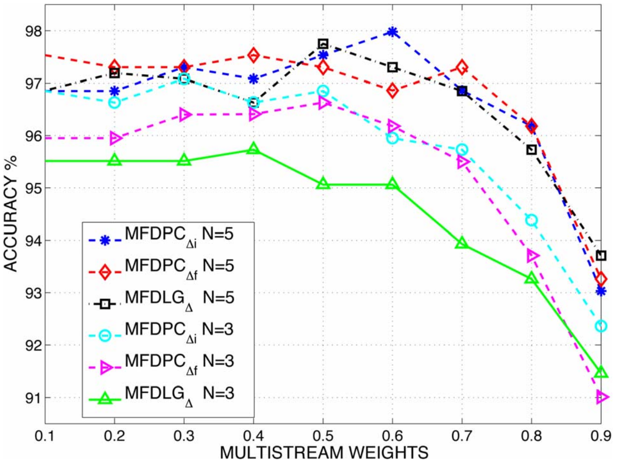

Fig. 16. Weight optimization for multistream cases for HMMs for $N = 3, 5$ and $M = 5$. X-axis shows the stream weight w_1 for the MFDs (where $w_1 + w_2 = 1$).

TABLE III
RECOGNITION RESULTS, WHERE N DENOTES THE NUMBER OF STATES AND M THE NUMBER OF MIXTURES. FOR FEATURE SET SPECIFIC INFORMATION, SEE TABLE II

Mean Accuracy					
Feature Set	Weights	GMM		HMM	
	MFD-MFCC	$M = 3$	$M = 5$	$N = 3$ $M = 5$	$N = 5$ $M = 5$
MFDPC$_{\Delta i}$ MFCC$_\Delta$	0.2 - 0.8	84.36	**94.70**	97.52	**97.93**
	0.4 - 0.6	84.80	94.65	**98.03**	97.67
	0.5 - 0.5	**85.75**	94.59	97.67	97.83
MFDPC$_{\Delta f}$ MFCC$_\Delta$	0.2 - 0.8	**86.75**	94.29	**97.63**	**98.18**
	0.4 - 0.6	85.20	**94.54**	97.42	97.93
	0.5 - 0.5	84.50	93.88	97.17	97.93
MFDLG$_\Delta$ MFCC$_\Delta$	0.2 - 0.8	83.74	**93.79**	**96.82**	**97.58**
	0.4 - 0.6	83.84	92.73	96.77	97.42
	0.5 - 0.5	**83.94**	91.82	96.72	97.43
MFCC$_\Delta$	-	83.64	93.23	96.41	97.32
MFDPC$_{\Delta i}$	-	70.10	77.88	85.91	88.08
MFDPC$_{\Delta f}$	-	67.42	75.35	85.66	88.53
MFDLG$_\Delta$	-	68.08	75.25	86.41	87.43

C. Results

The obtained accuracy scores of the recognition results for the various cases of featurs sets were quite promising and the most

TABLE IV
RECOGNITION RESULTS PER INSTRUMENT CLASS FOR THE THREE BEST COMBINED FEATURE SETS, $\mathrm{MFDPC}_{\Delta\mathrm{i}}$, $\mathrm{MFDPC}_{\Delta\mathrm{f}}$, MFDLG_{Δ}, COMPARED TO MFCC_{Δ} (FOR $N = 5, M = 5$, EXCEPT FOR $\mathrm{MFDPC}_{\Delta\mathrm{i}}$ WHICH IS SHOWN FOR $N = 3, M = 5$)

	Correct Percentage			
Instrument Classes	$\mathrm{MFDPC}_{\Delta\mathrm{i}}$ MFCC_{Δ} 0.4-0.6	$\mathrm{MFDPC}_{\Delta\mathrm{f}}$ MFCC_{Δ} 0.2-0.8	MFDLG_{Δ} MFCC_{Δ} 0.2-0.8	MFCC_{Δ}
Double Bass	99.76	99.52	95.52	**100**
Bassoon	96.66	**97.20**	96.08	94.98
Bb Clarinet	**94.64**	88.78	88.80	91.68
Cello	97.88	94.24	93.10	**98.26**
Horn	**99.28**	87.84	86.42	94.30
Tuba	**100**	99.40	98.18	**100**
Flute	**97.34**	89.70	90.26	97.06

representative are reported next. Fig. 16 shows the accuracy obtained on the hold-out set for the three different MFD sets fused with the MFCCs. We notice that the assignment of higher or equal stream weight on the MFCCs, i.e., between 0.5–0.8, results in most cases on better scores for either three or five states and five mixtures. Therefore, we choose three cases, which are $w_2 = 0.8, 0.6, 0.5$ for MFCCs, and in Table III, we present the results on the final test sets for the various feature sets with Δ's. For most cases (even those not presented here), the combination of the proposed features with the MFCCs proves out to yield slightly better results than the MFCCs alone, although the MFDs alone show lower discriminability. Since we noted an absolute increase of almost 10% for GMMs with $M = 5$, in comparison to $M = 3$, the scores and the discussion that follows regarding both classification methods concerns the cases where $M = 5$ mixtures. Our first remark is about the error reduction, which is up to 26% for $\mathrm{MFDPC}_{\Delta\mathrm{i}}(N = 3, M = 5)$ and up to 32% and 10% for $\mathrm{MFDPC}_{\Delta\mathrm{f}}$ and $\mathrm{MFDLG}_{\Delta}(N = 5, M = 5)$, respectively. Furthermore, comparing with previous experiments (see [32]), we observe that the addition of Δ's on the MFDLG achieves an error reduction in recognition up to ca 50%, while on the MFDPC up to 35%. HMMs acquire greater results, since they also imply the temporal information of the tones.

For the experiments of features without the Δs, the main disadvantage of the MFDs was the low discriminability between Bb Clarinet and Flute which yield the lower results among the investigated instruments (ca 55% recognition each). Our analysis has pointed out some of the similarities of their MFD profiles for the higher frequency tones and that was possibly the main drawback of the method and the consequence for the low accuracy scores. Table IV shows the percentage of correct recognition per instrument obtained by HMMs for the fused feature set cases in comparison to MFCC_{Δ}. By reviewing these results, we note an improvement in recognition of all instruments and especially in the discrimination of Bb Clarinet and Flute. Note that Double Bass and Tuba are again among the best recognized instruments regarding the MFDs, in accordance with our expectancies after the analysis.

Additionally, for the first set of experiments (see [32]), we noticed that the combination of MFDs with MFCCs enhanced the discriminability of the Bassoon, Bb Clarinet and Horn while they decreased the accuracy obtained by the MFCCs for Cello and Flute. Double Bass and Tuba kept the already good performance of the MFCCs. Again, after inspection of the latter set of experiments, we mark an increase in recognition for most analyzed instruments, although there are cases of some of the MFD feature sets where the use of the derivatives decreases individual instruments' good results, as for Cello.

Finally, regarding the MFDPC_{Δ} and MFDLG_{Δ}, we note that the logarithmically sampled features are almost as good if not better in specific evaluation cases as the PCA acquired features, something that signifies the fact that there is practically no need for further processing of the features and thus decreased calculation burden.

VII. CONCLUSION

In this paper, we employ fractal dimension measurements and propose the use of a multiscale fractal feature for structure analysis of musical instrument tones motivated from similar successful ideas used for speech recognition tasks. Our goal is to gain insight about the instruments' characteristics and achieve better discrimination in tasks such as instrument classification. Experiments were conducted, where the proposed features (MFDs) were evaluated against the baseline MFCC features. The results show that the MFDs can improve the recognition accuracy when fused with the MFCCs, accomplishing an error reduction up to ca. 32%. Even though the specific fractal features have lower discriminability than the MFCCs as far as the resulting accuracy is concerned, yet they acquire high discriminability in some of the analyzed instruments. With the MFD analysis on synthesized sounds we managed to get a higher level of intuition regarding the different phenomena observed on the MFD profiles of the instruments. To conclude, based on our experimental hypothesis and recognition evaluation, there is strong evidence that musical instruments have structure and properties that could be emphasized by the use of multiscale fractal methods as an analysis tool of their characteristics. We have shown that they can provide information about different properties of the tones and the instruments, while the recognition experiments have shown to be promising in most cases.

For our future research we intend to enhance the usage of multiscale methods for music analysis by relating such ideas with the physics of the instruments. Additionally, we are inquiring the usage of multiscale fractal dimension for genre classification. Some initial experimental evaluation gave us evidence that MFDs could prove promising. It remains to investigate whether the MFDs can be applied in other audio signals and for other purposes as well.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for all their suggestions for improving this paper.

REFERENCES

[1] *Acoustical Terminology*, American Standard Association, 1960, N.Y..

[2] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 221–224.

[3] M. Bigerelle and A. Iost, "Fractal dimension and classification of music," *Chaos, Solitons, Fractals*, vol. 11, pp. 2179–2192, 2000.

[4] J. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1064–1072, 2001.

[5] A. Eronen, "Comparison of features for musical instrument recognition," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 19–22.

[6] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. Signal Process. and Its Applicat.*, 2003, vol. 2.

[7] K. Falconer, *Fractal Geometry, Mathematical Foundations and Applications*, 2nd ed. New York: Wiley, 2003.

[8] H. Fletcher, "Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure," *J. Acoust. Soc. Amer.*, vol. 6, no. 2, pp. 59–69, 1934.

[9] N. H. Fletcher and T. Rossing, *The Physics of Musical Instruments*, 2nd ed. New York: Springer, 1998.

[10] J. Gao, J. Cao, W. Tung, and J. Hu, *Multiscale Analysis of Complex Time Series—Integration of Chaos and Random Fractal Theory, and Beyond*. New York: Wiley-Interscience, 2007.

[11] J. Gao, J. Hu, W. Tung, and Y. Zheng, "Distinguishing chaos from noise by scale-dependent Lyapunov exponent," *Phys. Rev. E*, vol. 74, 2006, 066204.

[12] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, 2002.

[13] D. E. Hall, *Musical Acoustics*, 3rd ed. Independence, KY: Brooks/Cole, 2002.

[14] K. Hsu and A. Hsu, "Fractal geometry of music," in *Proc. Nat.. Acad. Sci.*, 1990, vol. 87.

[15] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2595–2603, 1993.

[16] B. Mandelbrot, *The Fractal Geometry of Nature*. San Francisco, CA: Freeman, 1982.

[17] P. Maragos, "Fractal aspects of speech signals: Dimension and interpolation," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, 1991, pp. 417–420.

[18] P. Maragos, "Fractal signal analysis using mathematical morphology," in *Advances in Electronics and Electron Physics*. New York: Academic, 1994, vol. 88, pp. 199–246.

[19] P. Maragos and A. Potamianos, "Fractal dimension of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1925–1932, 1999.

[20] B. Moore, *Psychology of Hearing*, 5th ed. New York: Academic, 2003.

[21] A. Nielsen, S. Sigurdsson, L. Hansen, and J. Arenas-Garcia, "On the relevance of spectral features for instrument classification," in *Proc. Int. Conf. Acoust., Speech, Signal Process*, 2007, pp. 485–488.

[22] H. F. Olson, *Music, Physics and Engineering*. : Dover, 1967.

[23] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden Markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1795–1807, Sep. 2006.

[24] V. Pitsikalis and P. Maragos, "Filtered dynamics and fractal dimensions for noisy speech recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 11, pp. 711–713, Nov. 2006.

[25] R. Plomp, *The Intelligent Ear: On the Nature of Sound Perception*, 1st ed. New York: Psychology Press, 2001.

[26] G. Potamianos and H. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 3733–3736.

[27] Z. Su and T. Wu, "Music walk, fractal geometry in music," *Physica A*, vol. 380, pp. 418–428, 2007.

[28] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, Dec. 2002 [Online]. Available: http://htk.eng.cam.ac.uk/, The HTK Book, Revised for HTK Version 3.2, Cambridge Research Lab.

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[30] [Online]. Available: http://theremin.music.uiowa.edu/, Univ. of Iowa Musical Instrum. Sample Database

[31] R. F. Voss and J. Clarke, "1/f noise in music and speech," *Nature*, vol. 258, pp. 317–318, Nov. 1975.

[32] A. Zlatintsi and P. Maragos, "Musical instruments signal analysis and recognition using fractal features," in *Proc. EUSIPCO-11*, 2011.

Athanasia Zlatintsi (S'12) received the Master of Science in media technology from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2006. Since 2007 she has been a research assistant at the Computer Vision, Speech Communication, and Signal Processing Group, NTUA, participating in research projects while she is currently working towards her Ph.D. degree. Her research interests lie in the areas of music and audio signal processing and include analysis and recognition.

Petros Maragos (F'96) received the EE Diploma from NTUA in 1980 and the M.Sc. and Ph.D. from Georgia Tech in 1982 and 1985. He has worked as ECE professor at Harvard University (1985–1993), at Georgia Tech (1993–1997), and at NTUA (1998–present). His research interests include signals and systems, pattern recognition, image processing and computer vision, audio, speech and language processing, cognition, and robotics. He has served as an associate editor for IEEE Transactions and other journals; as co-organizer of several conferences and workshops; and as a member of three IEEE SPS committees. He is the recipient or co-recipient of several awards, including a 1987 NSF PYIA, the 1988 IEEE SPS Young Author Paper Award, 1994 IEEE SPS Senior Award, 1995 IEEE W.R.G. Baker Award, 1996 Pattern Recognition Honorable Mention Award, 2011 CVPR Gesture Workshop best paper award, and the 2007 EURASIP Technical Achievement Award. He is a fellow of IEEE and of EURASIP.