MUSICAL INSTRUMENTS SIGNAL ANALYSIS AND RECOGNITION USING FRACTAL FEATURES

Athanasia Zlatintsi and Petros Maragos

School of Electr. & Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece
[nzlat,maragos]@cs.ntua.gr

ABSTRACT

Analyzing the structure of music signals at multiple time scales is of importance both for modeling music signals and their automatic computer-based recognition. In this paper we propose the multiscale fractal dimension profile as a descriptor useful to quantify the multiscale complexity of the music waveform. We have experimentally found that this descriptor can discriminate several aspects among different music instruments. 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 methods and features proposed in this paper are promising for music signal analysis and of direct applicability in large-scale music classification tasks.

1. INTRODUCTION

The analysis of musical content and information is of importance in many different contexts and applications, such as music retrieval, automatic music transcription, indexing of multimedia databases and other. These applications require solutions to information processing problems such as automatic musical instrument classification and genre classification [1, 15, 18]. Toward this goal, it is required to develop efficient digital signal processing techniques for analyzing the structure of music signals and extracting relevant features. Our paper proposes such methods and algorithms to quantify fractal-like structures in music signals at multiple time scales.

Previous analysis of musical structure has revealed evidence of both fractal aspects and self-similarity properties in musical instrument tones and genres. Voss and Clark [19] were the first to investigate $1/f^{\beta}$ aspects in music and speech, using the estimation of power spectra for slowly varying quantities such as loudness and frequency. In [2] the fractal and multifractal aspects of different genres of music were analyzed, using the Variation and the ANAM method and it was proposed that the fractal dimension could help in discrimination of different genres of music. Su and Wu [16] applied Hurst exponent and Fourier spectral analysis in sequences of musical notes, noticing that they share similar fractal properties with the fractional Brownian motion. Aspects of fractal geometry were also studied in [8], where observations of self-similarity properties regarding the acoustic frequency of the signals were made.

None the less, over the years many different feature schemes have been proposed and pattern recognition algorithms have been employed in order to clarify the complex issue of modeling musical instruments. Such feature schemes employ perception-based features, temporal, spectral and timbral features.

Cepstral coefficients were used and have been favored a long way back, not only in speech processing but also in recognition tasks regarding musical instruments, as in Brown et al. [3] where cepstral coefficients, constant Q transform, spectral centroid and autocorrelation coefficients were used on identifying four instruments of the woodwind family. Eronen [4] compared the performance of several features, among them MFCCs, spectral and temporal features such as amplitude envelope and spectral centroids for instrument recognition, using the Karhunen-Loeve transform for decorrelation of the features and k-nearest neighbor (k-NN) for classification. The results favored the MFCC features, which gave the best accuracy in instrument family classification. Experiments on real instrument recordings [13] also favored the MFCCs over Harmonic Representations.

Previous work has used classifiers that are not necessarily effective in modeling the temporal evolution of the features. For instance the Gaussian mixture models (GMMs) are capable of parameterizing the distribution of observations, although, they could not model the dynamic evolution of the features within a note as for example hidden Markov models (HMMs) could do. In [5] the feature distribution of MFCCs and delta-MFCCs was modeled with HMMs while in [15] Variable Duration HMMs were used for classification of musical patterns.

In our work, the analysis concerns isolated musical instrument tones. The signals are derived from UIOWA database with musical instrument samples [14]. We propose new algorithms and features, based on multiscale fractal exponents, which are validated by both static and dynamic classification algorithms, and we compare their descriptiveness with a standard feature set of MFCCs which have been found to be well-performing in musical instrument recognition. For the recognition evaluation, we choose Markov models and report on promising experimental results.

2. MULTISCALE FRACTAL EXPONENTS

Most features extracted from music signals for classification purposes are inspired by similar work in speech. Many speech sounds contain some amounts of turbulence at some time scales. Mandelbrot [9] conjectured that multiscale structures in turbulence can be modeled using fractals. This motivated Maragos [10] 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 [10, 11] an efficient algorithm to measure it using nonlinear multiscale morphological filters that can create geometrical covers around the graph of the speech signal, whose fractal dimension D can then be found by

$$D = \lim_{s \to 0} \frac{\log[\text{Area of dilated graph by disks of radius } s/s^2]}{\log(1/s)} \quad (1)$$

D is between 1 and 2 for one-dimensional signals; 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.

In practice, real-world signals do not have the same structure over all scales; hence D is computed by fitting a line to the log-log data of (1) over a small scale window that can move along the s axis and thus create a profile of local *multiscale fractal dimensions (MFDs)* $D(s,t)$ at each time location t of the short speech analysis frame. The function $D(s,t)$ can also be called a *fractogram*

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.

Averaged MFDs (Standard Deviation)						
Time Scale (ms)	$s_t = 1/44$	$s_t = 0.5$	$s_t = 1$	$s_t = 1.5$	$s_t = 2$	$s_t = 2.5$
Double Bass	1.11 (0.050)	1.21 (0.037)	1.31 (0.04)	1.39 (0.04)	1.52 (0.039)	1.61 (0.038)
Bassoon	1.04 (0.004)	1.47 (0.006)	1.75 (0.070)	1.78 (0.08)	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)

Table 1: Averaged MFDs and Standard Deviation for time scale points of the MFD profiles at $s_t = 1/44, 0.5, 1, 1.5, 2, 2.5$ ms.

and can provide information about the degree of turbulence inherent in short-time speech sounds at multiple scales. 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, there is strong evidence from [12] that using such MFDs as features reduces the error in speech recognizers. In this paper, we have used MFDs as an efficient tool to analyze short-time music signal structure at multiple time scales. The results are quite interesting as we discuss next by also showing examples of MFDs for music signals from various instruments.

3. MULTISCALE FRACTAL DIMENSION FOR TIMBRE ANALYSIS

3.1 Timbre characteristics

One of the main relations among sound attributes is the determination of timbre by the waveform. 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. For people and especially trained musicians is quite easy to recognize which instrument is heard, but this is not the case if the part of the note allowed to hear is the steady middle state only. Instrument recognition depends a great deal on hearing the transients of a tone, meaning the beginning (attack) and the ending (release) [7]. It is namely vital to hear even the scrape of the bow on a violin string, or the squeak of a clarinet reed or even the first puff of the air released by a trumpet player.

According to Hall [7], the duration of those transients vary not only among instruments but between higher and lower octave notes too. Some typical attack durations he reported are from 20ms or less for an oboe, 30-40 ms for clarinet or trumpet, to 70-90ms for flute or violin. Additionally, notes in the octaves above middle C (designated as C4 at ca. 261 Hz), have periods of 2 to 4 ms which means several dozen vibrations periods for the steady state to be established. On the other hand, in [6] is reported that the duration of the attack transients is typically 50±20 ms, independent of the note or the instrument. Because of such evidence about the differences of the transients of the tones, we conclude that the whole duration of a tone is important and gives vital clues for its identity. In the following sections our main hypothesis is that the multiscale fractal dimension can help in discrimination of timbre by discriminating not only the steady state of the tones but the attacks too.

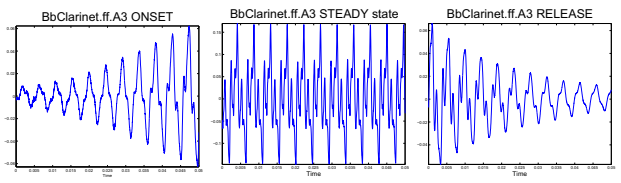

Figure 1: Onset, steady state and release for Bb Clarinet A3, Fs = 44.1Hz

Musical instruments include different instrument families. The four main categories or families are: strings (e.g. violin, upright bass), woodwind (e.g. clarinet, bassoon), brass (e.g. French horn, tuba) and percussion (e.g. piano). In many applications, classification down to the level of instruments families could be sufficient although in our approach we focus more on the distinction of individual instruments, pointing out similarities that are observed for the families.

3.2 MFDs on steady state

We base our analysis not only on the distinction of different instruments classes, but on the exploration of the differences between the attack and steady state of the tones too. We aim to show that the multiscale fractal dimension distribution of the attacks of different instrument tones differs sufficiently in order to add adequate information in a recognition task.

For the analysis of the steady state we used the whole range of tones for the instruments Double Bass, Bassoon, Bb Clarinet, Cello, Flute, French Horn and Tuba. Specifically, we calculated the short-time MFDs of the tones using 30-ms segments from the whole duration of the tones. Although, regarding the state-specific analysis the appropriate segments have been processed. The signals are sampled at 44.1 kHz, and their corresponding profiles of MFD[s] are analyzed for discrete scales $s = 1, ..., 133$. This range of s corresponds to time scales s_t from 1/(44.1) to 3 ms. Similar results were gained for analysis of 50-ms windows.

In Fig. 2, the mean and standard deviation (shown as errorbars) of the MFDs is computed for the note A3 for the analyzed instruments, except Flute which is shown for B3 instead. Regarding the MFDs of each instrument tone, the profile presented is typical for the following octaves of every instrument: Double Bass for the whole range, Bassoon for octave 3-5, Bb Clarinet for octave 3-4, Cello for octave 2-4, Flute for octave 3-4 and Horn for octave 3-5. For the lower octaves of Bassoon, Tuba and Horn (i.e., octaves 1-2) the MFD profiles as shown in Fig. 3 shows some similarities. They get to their higher value D at about $s_t = 0.5$ and then decreases to an intermediate value. Still, they exhibit some important differences; for Bassoon the maximum D is at about 1.8, while Tuba and Horn share the values of ca. $D = 1.5$, thus again Tuba tones show more important deviations of D inside each tone than Horn. About the higher octaves of Bb Clarinet and Flute (octaves 5-6) another tendency was observed, see Fig. 3. The MFD profiles for those ranges gains its higher value around $D = 1.9$ at very small time scales, ca. $s_t = 0.5$, and beholds this value for the whole profile. The analysis of Double Bass and Cello have shown more uniform MFD profiles with an increased deviation of D across frames for lower range tones. Thus, apart from these two cases, for the rest of the analyzed instruments, certain differences are observed between their lower and higher octaves, still with unvarying characteristics across the specific octave tones as discussed in detail before. Table 1 presents the averaged values of the instrument related MFDs, for the steady state averaged over the whole range of each instrument (and dynamic range forte), 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 in each case. 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 and Tuba.

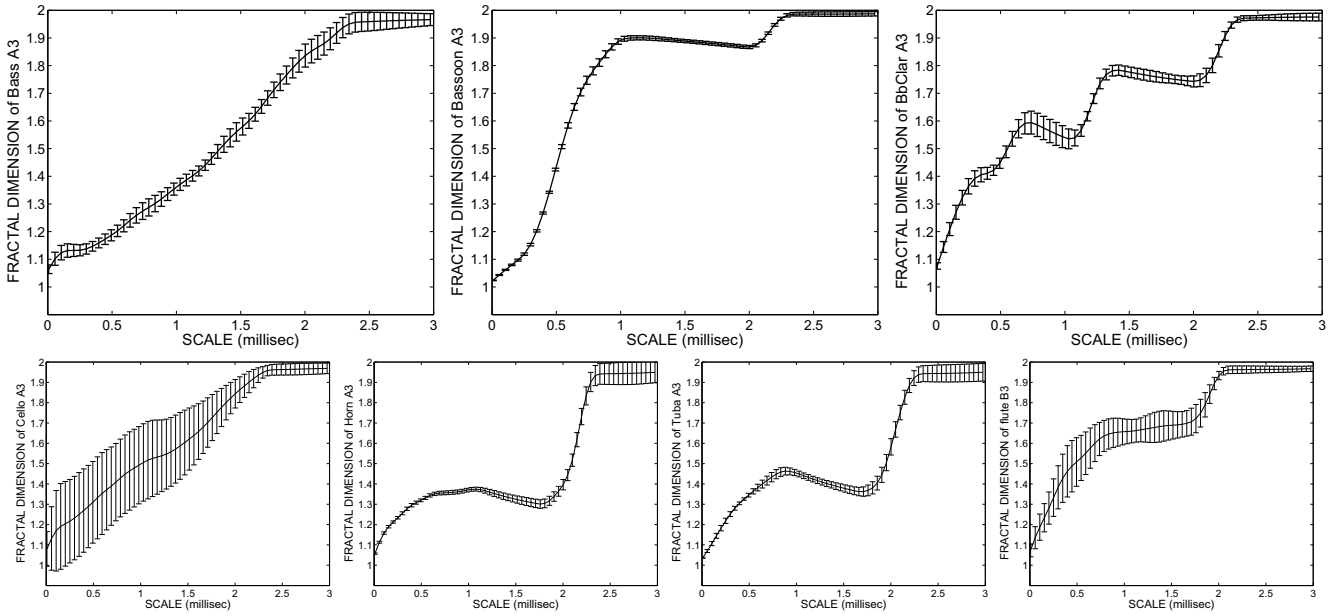

Figure 2: Mean and standard deviation (error bars) of the multiscale fractal dimension distribution of the same note A3 for the instruments Double Bass, Bassoon, Bb Clarinet (first row) and Cello, Horn and Tuba, and the note B3 for Flute (second row) (for 30-ms analysis window, updated every 15 ms).

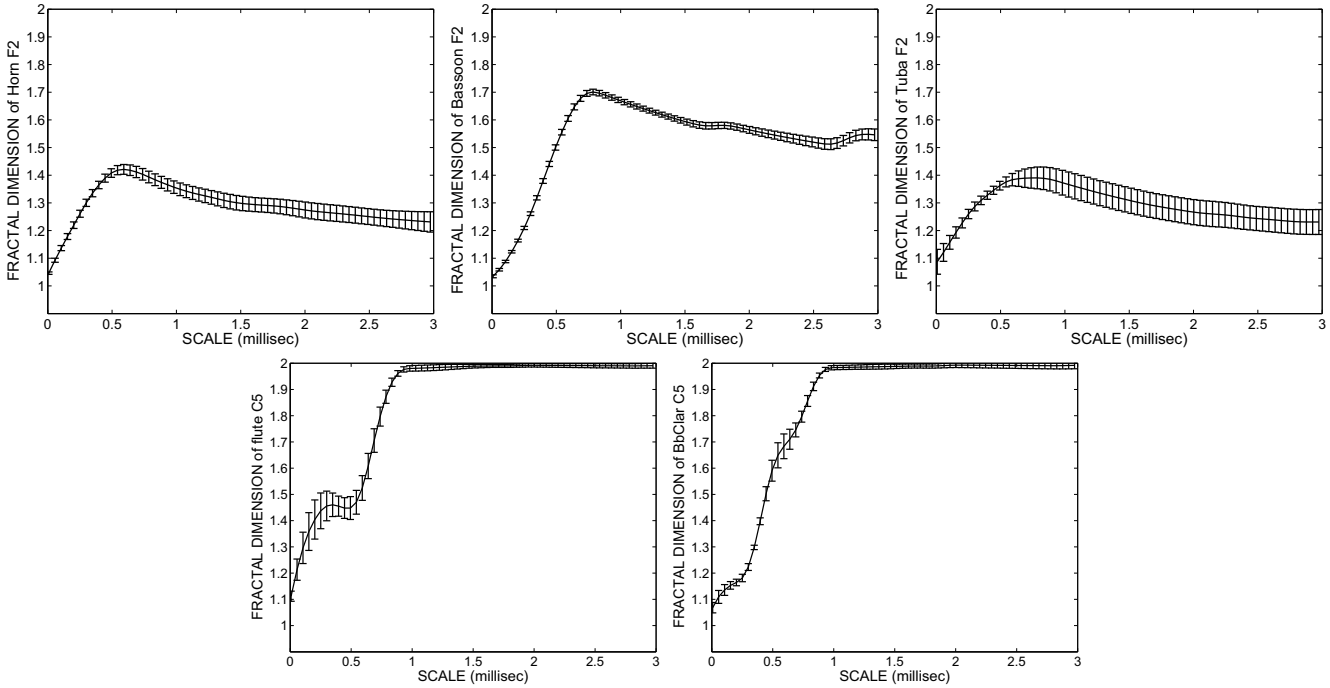

Figure 3: Mean and standard deviation (error bars) of the multiscale fractal dimension distribution for the note F2 for the instruments Horn, Bassoon and Tuba (first row) and the note C5 for Flute and Bb Clarinet (second row). The MFD profiles shown are typical for the lower octaves of the three first row instruments, respectively for the higher octaves of the two second row instruments, (30-ms analysis window, updated every 15 ms).

Analysis of the multiscale fractal dimension on the steady state of the instrument's tones reinforce the claims that the MFDs conveys information that is instrument related. Even for the cases of instruments that belong to the same family or the same frequency range and show similar tendencies, specific differences can be observed regarding either the dimension D, the scale s_t, or the deviation of D across scales. Additionally, we notice a dependence on the acoustical frequency of the sound and the MFDs, which will be further explained in 3.4.

3.3 MFDs for attack detection

For the analysis of the onset the same configuration was used as before and the process took 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. Although, they have higher D for small

scales s_t, and individually for each tone they present more fragmentation in comparison to the steady state, which we assume depends on the fragmentation of the waveform. Figure 4 shows the average MFDs for the onset of the whole range of the analyzed instruments (dynamic range forte). In this case, we notice the increased value of $D(s=1)$ and a quite clear distinction of the D among some of the analyzed instruments. Thus, the analysis of the attack have shown certain differences both between attack and steady state of the same tone, and among the instruments too.

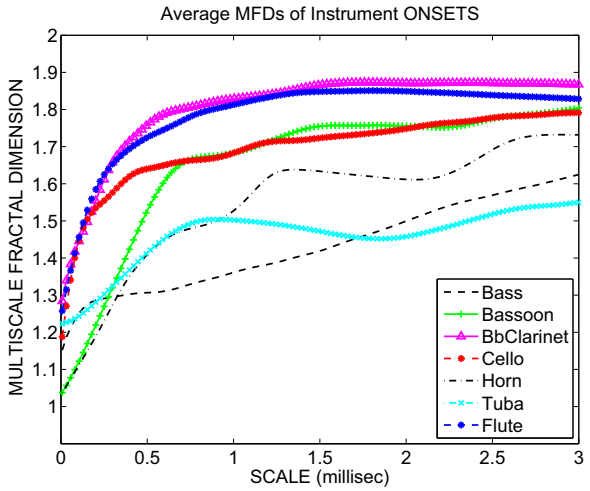

Figure 4: MFDs estimated for the 7 analyzed instruments onsets (attacks), averaged over the whole range (using 30-ms analysis windows). (Please see color version for better visibility.)

3.4 MFD variability for each instrument

Another important observation concerns the analysis of individual notes of the same instrument over one octave, Fig. 5. The notes used for the analysis (C4-B4) from Bb Clarinet, range between ca. 260-493Hz. The MFDs confirm the preceding evidence of our study that there is a dependence on the acoustical frequency of the sound and the multiscale fractal profile that increases rapidly for higher frequency sounds (i.e., higher D for smaller scales s_t). Still, the instrument's specific MFD profile beholds the shape observed for the specific octave ranges as discussed in 3.2. The same phenomenon with instrument specific variabilities has been observed for all analyzed instruments. These last observations give us evidence that the MFDs could be useful not only for the discrimination of different instrument classes but possibly for estimation of the acoustical frequency of the tone too.

4. RECOGNITION EXPERIMENTS

In order to evaluate and confirm the results of our previous analysis, we continue with recognition experiments. The experiments discussed in this section were carried out using 1331 notes from 7 different instruments, which are Double Bass, Bassoon, Cello, Bb Clarinet, Flute, Horn and Tuba. The collection consists of the instrument's full range and cover the dynamic range from piano to forte. Five different cases of feature sets or feature set combinations were evaluated, using static (GMMs) and dynamic (HMMs) classifiers with diverse combinations of N states and/or M mixtures. Dimensionality reduction of the MFD feature space was conducted using PCA, in order to decorrelate the data, and to obtain the optimal number of features that accounts for the maximal variance. In this case the principal components proved to be six. The performance of the selected features was compared with a standard feature set of 13 MFCCs, which are chosen both for their good performance and the acceptance they have gained for instrument recognition tasks. The analysis of the MFCCs is performed in 30

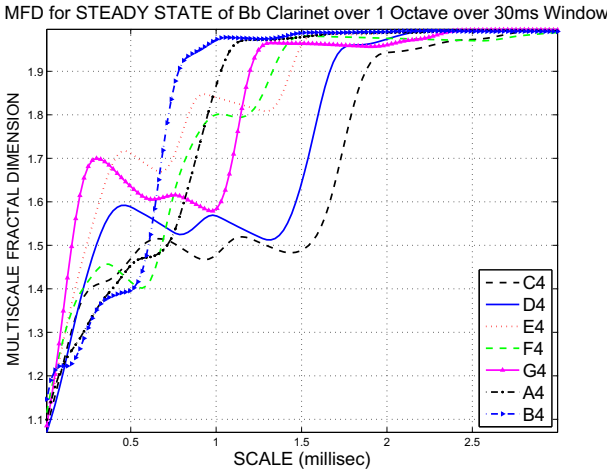

Figure 5: MFD for steady state of Bb Clarinet notes, over one octave for one 30ms analysis window. (Please see color version for better visibility)

ms windowed frames with a 15 ms overlap, and with 24 triangular bandpass filters. For the implementation of the Markov models the HTK [17] HMM-recognition system was used. In all cases, the train sets were randomly selected to be the 80% of the available tones, and the results presented are after a five-fold cross validation.

4.1 Experimental configuration

In our experiments we evaluate the performance of fixed sets of features, which are listed in Table 2. The first set of experiments,

1	6 MFDs after PCA decorelation (MFDPC)
2	13 MFDs logarithmicaly sampled (MFDLG)
3	13 MFCCs
4	6 MFDPCs + 13 MFCCs
5	13 MFDLGs + 13 MFCCs

Table 2: List of feature sets.

employs Gaussian Mixture Models (GMMs) up to 3 mixtures. A GMM is a probability density function represented as a weighted sum of Gaussian component densities, parameterized by mean vectors, covariance matrices and mixture weights from all component densities. In the second set of experiments we aimed to model the temporal characteristics of the signals using Hidden Markov Models. HMMs are dynamic models which in that case allows the modeling of the structure of a musical tone. They are statistical processes used to model a series of unobserved states which produce an output with a specific probability for each state. Taking into consideration the structure of the instruments' tones, as discussed in the previous sections, we adopt a left-right topology for the modeling. Each subset of features was trained in a different stream and then fused employing different stream weights for experimentation purposes, by EM estimation using the Viterbi algorithm. The experimental methods consisted also of the variation of the number of states N [3-9] and the number of mixtures M [1-3].

4.2 Results

The obtained accuracy scores of the recognition results for the different cases of feature sets were quite promising and the more representative are reported in Table 3.

The combination of the proposed features with the MFCCs proved out to yield slightly better results than the MFCCs alone for most cases (even those not presented here), although the MFDs alone show lower discriminability. Additionally, we see that HMMs achieved greater results, since they imply the temporal information

Mean Accuracy				
Feature Set	Weights	GMM	HMM	
	MFD-MFCC	$M=3$	$N=3$ $M=3$	$N=5$ $M=3$
MFDPC-MFCC	0.2 - 0.8	86.01	93.01	94.68
	0.5 - 1	85.78	92.31	94.22
	0.5 - 0.5	86.01	92.78	94.68
MFDLG-MFCC	0.2 - 0.8	85.25	92.93	94.98
	0.5 - 1	85.78	93.16	94.91
	0.5 - 0.5	85.40	92.47	94.45
MFCC	-	87.07	92.93	94.40
MFDPC	-	69.35	75.59	76.51
MFDLG	-	64.95	71.79	72.11

Table 3: Recognition Results, where *N* denotes the number of states and *M* the number of mixtures. For feature set specific information, see Table 2.

of the tones too. The disadvantage of the MFDs for those experiments is the low discriminability between Bb Clarinet and Flute which yield the lower results among the investigated instruments (ca 55% recognition each). Our analysis, has already shown the similarities of their MFD profiles for the higher frequency tones, and this is possibly the consequence of the low accuracy rates. We calculate the median average for this reason which for the best case of MFDPC ($N=5$, $M=3$) adjusts to 79.66% recognition rate and for the best case of MFDLG ($N=5$, $M=3$) to 75.87%. Nevertheless, we have to point out that Tuba, Bassoon and Double Bass were among the best recognized instruments regarding the MFDs, in accordance to our expectancies after the analysis.

For the combined feature set case, we can see in Table 4 the results obtained by HMMs ($N=5, M=3$), for each individual instrument class in comparison with the MFCCs. We observe that the combined feature sets enhance the discriminability of the Bassoon, Bb Clarinet and Horn while they decrease the accuracy observed by the MFCCs for Cello and Flute. Finally, Double Bass and Tuba beholds the already good performance of the MFCCs.

Mean Accuracy			
Instrument Classes	MFDPC + MFCC	MFDLG + MFCC	MFCC
Double Bass	100	100	100
Bassoon	93.32	95.84	88.52
Bb Clarinet	78.54	77.07	72.25
Cello	93.64	93.94	96.73
Horn	97.9	100	92.08
Tuba	100	100	100
Flute	96.02	95.58	97.25

Table 4: Recognition Results per instrument class for the two best combined feature sets in comparison with MFCCs.

5. CONCLUSION

In this study, we propose a multiscale fractal method for structure analysis of musical instrument tones motivated from similar successful ideas used for speech recognition tasks. Based on our experimental hypothesis and recognition evaluation there is strong evidence that musical instruments has structure and properties that could be emphasized by the use of multiscale fractal methods (MFDs) as an analysis tool of their characteristics. We have shown that they can provide information about different properties of the notes and the instruments, while the recognition experiments have shown to be promising in most cases if not winning.

In our ongoing research on music signal processing we are also working to enhance these aspects of multiscale fractal methods using different feature parameterizations and modeling techniques, such as parameterization of the actual shape of the MFD profile and decision level fusion. Additional performance improvements could be achieved with a more careful choice of these parameters. The relation of such ideas with the physics of the instruments is also in our future intensions to explore. Furthermore, we are inquiring the usage of MFDs for genre recognition. Our initial experiments gives evidence that MFDs could prove promising. Some first observations focuses on the fact that $D(s=1)$ estimated at the smallest time scale, differs significantly for some genres, besides the variations presented in the genre related MFD profiles.

Acknowledgment

The authors would like to thank S. Theodorakis and V. Pitsikalis for their help regarding the HTK classification toolkit, and for the constructive comments that helped to improve this paper.

REFERENCES

[1] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non[negative matrix factorization algorithms," in *in Proc. Acoustics, Speech & Signal Processing, (ICASSP-06)*, 2006.

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

[3] J. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. SOc. Amer.*, vol. 109(3), pp. 1064 –1072, Mar. 2001.

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

[5] ——, "Musical instrument recognition using ica-based transform of features and discriminatively trained hmms," in *in Proc. Signal Processing & Its Applications*, vol. 2, 2003, pp. 133–136.

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

[7] D. E. Hall, *Musical Acoustics*, 3rd ed. Brooks/Cole, 2002.

[8] K. J. Hsu and A. J. Hsu, "Fractal geometry of music," in *Proc. Natl. Acad. Sci. USA Physics*, vol. 87, February 1990, pp. 938 – 941.

[9] B. Mandelbrot, *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco, 1982.

[10] P. Maragos, "Fractal aspects of speech signals: Dimension and interpolation," in *Proc. ICASSP-91*, May 1991.

[11] ——, "Fractal signal analysis using mathematical morphology," *Advances in electronics and electron physics, Academic Press*, vol. 88, pp. 199 – 246, 1994.

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

[13] A. Nielsen, S. Sigurdsson, L. Hansen, and J. Arenas-Garcia, "On the relevance of spectral features for instrument classification," in *in Proc. Acoustics, Speech and Signal Processing (ICASSP-07)*, 2007.

[14] U. of Iowa Musical Instrument Sample Database. [Online]. Available: http://theremin.music.uiowa.edu/index.html.

[15] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden markov models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), pp. 1795–1807, Sept. 2006.

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

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

[18] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, p. 293V302, July 2002.

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