COMPARISON OF DIFFERENT REPRESENTATIONS BASED ON NONLINEAR FEATURES FOR MUSIC GENRE CLASSIFICATION

Athanasia Zlatintsi and Petros Maragos

School of Electr. & Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece

[nzlat,maragos]@cs.ntua.gr

ABSTRACT

In this paper, we examine the descriptiveness and recognition properties of different feature representations for the analysis of musical signals, aiming in the exploration of their micro- and macro-structures, for the task of music genre classification. We explore nonlinear methods, such as the AM-FM model and ideas from fractal theory, so as to model the time-varying harmonic structure of musical signals and the geometrical complexity of the music waveform. The different feature representations' efficacy is compared regarding their recognition properties for the specific task. The proposed features are evaluated against and in combination with Mel frequency cepstral coefficients (MFCC), using both static and dynamic classifiers, accomplishing an error reduction of 28%, illustrating that they can capture important aspects of music.

Index Terms— Music genre classification, AM-FM model, energy separation algorithm, fractals, Bag-of-Words.

1. INTRODUCTION

Genre is the most popular and widespread used term for the description of music both among users and in music industry [1]. It is the main method for organizing databases, music libraries and music stores and basic descriptor in order to find similarities among artists and compositions. Humans while trying to discover new music, they rely on features such as melody, harmony, rhythm, etc. [2], specific emotional content, or they search for music of a particular style and "texture" [3]. Moreover, their music collections are usually sorted according to artist, year, country of origin, but mainly genre. However, the term genre is not considered a reliable descriptor because of the fuzzy boundaries and the overlap among the different genres and sub-genres [4], which makes the task of music genre classification quite complicated. Still, most recent research in the field succeed in achieving very good recognition results.

Over the years, various feature sets have been proposed and pattern recognition algorithms have been employed to solve the complex task of genre classification. Such feature sets include timbral features, e.g., temporal, energy, spectral shape features and others, rhythm and pitch related features. Regarding the classification algorithms hidden Markov models (HMM) [5] and Support Vector Machines (SVM) [6, 7] have been used successfully in various cases. For an overview of features and machine learning techniques see [1,4].

In this paper, we explore and compare five different feature representations based on micro-, macro-structures and Bag-of-Words. For the creation of those representations we use nonlinear features, such as modulations and fractals, which have proven useful for various applications. Section 2 concerns the description of the proposed nonlinear features as well as a "*music*" filterbank for the extraction of the AM-FM features is introduced. In Sec. 3 the five feature representations are presented. We continue with recognition experiments, Sec. 4, where the nonlinear features are also fused, in order to examine their discriminability capabilities regarding the task of genre classification. The results illustrate that they can capture important aspects of musical signals.

2. PROPOSED FEATURES

2.1. AM-FM modulations

Small fluctuations or micro-modulations in frequency occur naturally in both human voice and music [8]. Additionally, the musical signals' temporal micro-structure consists of instantaneous amplitude and frequency modulations of their main resonances, which characterize their waveforms. Inspired by the fact that the AM-FM model has been used successfully for speech processing [9], music instrument classification [10] and audio saliency/event detection [11], we propose the AM-FM modeling of music signals for the task of genre classification. Hence each resonance component is modeled as an amplitude and frequency modulated sinusoid (AM-FM signal) while the whole music signal is modeled as a sum of such AM-FM components $S(t) = \sum_{i=1}^{K} \alpha_i(t) \cos(\phi_i(t))$, where α_i and ϕ_i are the instantaneous amplitude and phase signals of component i.

The AM-FM features investigated in this paper are: the **mean Instantaneous Amplitude** (m-IAM), i.e., the short-time mean of the instantaneous amplitude signal $|\alpha_i(t)|$ for

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co–funded by the European Social Fund (ESF) and Greek National Resources.

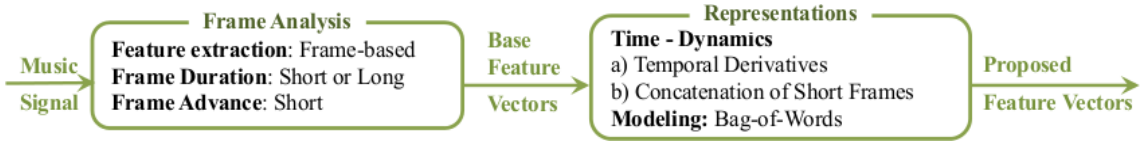

Fig. 1. Different approaches for the analysis of music signals and the extraction of the proposed features.

each resonance component i, parameterizing the resonance amplitudes, the **mean Instantaneous Frequency** (m-IFM), i.e., a short-time weighted mean of the instantaneous frequency $f_i(t)$, providing information about the signal's fine structure taking advantage of the excellent time resolution of the ESA [9], and the **Frequency Modulation Percentage** (FMP) [12], defined as $FMP_i = B_i/F_i$ for each resonance i, where B_i is the mean bandwidth (an amplitude-weighted version of the $f_i(t)$-signal deviation), and F_i is the weighted mean frequency value of resonance i (parameterizing the maximum change from the mean modulation frequency).

Specifically, we use a regularized version of the ESA for the demodulation of the signals, called Gabor-ESA [12], which is a combination of the continuous time ESA and Gabor filtering of the signal, providing smoother instantaneous estimates. In this case the Teager Energy operator $\Psi[x]$ [13] and the bandpass filtering are combined as follows:

$$\Psi[x(t) * g(t)] = \left[x(t) * \frac{dg(t)}{dt}\right]^2 - (x(t) * g(t)) \left[x(t) * \frac{d^2 g(t)}{dt^2}\right], \quad (1)$$

where $x(t)$ is the input signal, and $g(t)$ is the Gabor impulse response. The instantaneous signals are given by $f(t) \approx 1/2\pi\sqrt{\Psi[\dot{x}(t)]/\Psi[x(t)]}$ and $|\alpha(t)| \approx \Psi[x(t)]/\sqrt{\Psi[\dot{x}(t)]}$ where $\Psi[x] = \dot{x}^2 - x\ddot{x}$ and $\dot{x} = dx/dt$.

The filterbank consists of 12 bandpass mel-spaced Gabor filters with 50% overlap of the successive filters. Although this baseline configuration has proven successful for musical instrument recognition [10], in our effort to better control the bandbass filtering of music signals we propose the creation of a *"music"* filterbank, where the center frequency of each filter is determined by the frequency of each musical note. Two different music filterbanks are thus created; the first consisting of 89 filters beginning in the second octave (C2 = 65.4Hz), and the second consisting of 111 filters starting at the first octave (C1 = 32.7Hz). The filters' bandwidth in this case range from the center frequency of the previous filter (i.e., note) to the center frequency of the next filter; thus for a filter with frequency f_i, the bandwidth range is $b_{1i} = [f_{i-1}, f_{i+1}]$.

2.2. Multiscale Fractal Dimension (MFD)

AM-FM features capture one aspect of the nonlinear structure of music signals. Another aspect is their fractal structure, as explored in [14, 15]. Hence, we fuse AM-FM with fractal features to enhance the recognition performance. To compute fractal features we use the method developed in [16]. Specifically, the algorithm is using nonlinear multiscale morphological filters that can create geometrical covers around the graph of a signal. The fractal dimension D can then be found by:

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

In practice, real-world signals do not have the same structure over all scales, hence D can be computed by fitting a line to the log-log data of (2) over a small scale window that moves along the s axis, creating a profile of local *multiscale fractal dimensions (MFDs)* at each time location. In this work, we use the MFDs for the analysis of music's micro-structures.

3. DIFFERENT FEATURE REPRESENTATIONS

In this paper we have examined different approaches for the analysis of music signals, aiming in the exploration of their micro- and macro-structures, see Fig. 1. The basic feature set, used in all evaluation cases, is based on modulations and consists of log(m-IAM), m-IFM and FMP. However, it is often augmented with MFDs and/or MFCCs (which are mainly used for comparison). The examined methods and Feature Representations (denoted as FR) are explained next:

FR 1: Short time analysis for the recognition of the signals' micro-structures, where the proposed instantaneous features are calculated with the baseline mel-spaced Gabor filterbank, consisting of 12 filters, and bandwidth overlap of adjacent filters equal to 50%. The mean features are calculated using 30 ms frames with 50% overlap.

FR 2: Short time analysis for the recognition of the signals' micro-structures, where the AM-FM features are calculated with the "music" filterbank of 89 or 101 Gabor filters.

FR 3: Recognition of the signal's macro-structures using analysis frames of 125 or 200 ms with 80% overlap for the calculation of the mean instantaneous features, using the baseline Gabor filterbank.

FR 4: Creation of an augmented feature vector by concatenating the features (FR1) from a number of short time frames around an observation vector, for the recognition of the signals' temporal information and macro-structures. We experimented by combining successive frames with a total duration of 1/8, 1/4, 1/2 and 1 sec which corresponds to 8, 15, 33 and 65 frames, respectively. For the reduction of the feature space PCA analysis was used and experimention was conducted so as to find the optimal number of principal components. This approach, using HLDA (Heteroskedastic Linear Discriminant Analysis) for dimensionality reduction, has been used in speech applications with successful results [17].

FR 5: Bag-of-Words (BoW) modeling for classification of the music signals. BoW representations were originally

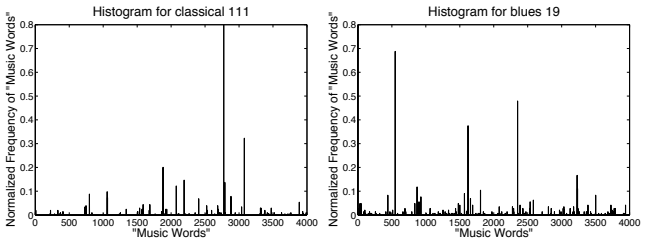

Fig. 2. Bag-of-Words representations for classical and blues.

proposed for text analysis [18], but then became one of the most popular methods in computer vision and applications such as object classification, video scene analysis and action recognition [19, 20].

The process required in order to create the BoW representations is summarized in the following steps: (a) feature extraction from the music signals (in our case the modulation features from FR1 are used after feature selection with a sequential forward selection algorithm). (b) Clustering of the feature vectors, with K-means, in order to generate "*music words*" and create the *music dictionary*. (c) Calculation of the frequency of music words (histograms) in each music signal for the creation of BoW representations. In other words, considering a set of data $D = \{d_1, \ldots, d_N\}$, where d_i are the features of N music signals, then K-means clusters the features D into a fixed number of K centers, creating the music dictionary $W = \{w_1, \ldots, w_K\}$, represented by K words. Subsequently, each music piece can be represented by a $K \times N$ co-occurrence table of counts $N_{ij} = n(w_i, d_j)$ where $n(w_i, d_j)$ is the frequency of the word w_i in a music piece d_j [21]. The result of the BoW representation is usually a sparse feature vector with dimension defined by the number of centers, see Fig. 2. The advantage of this procedure is the reduced computational complexity since the recognition problem is simplified in finding similarities between the music pieces. After experimentation we found that clustering in 4000 centers achieved the best accuracy.

For all representations, except FR4, the feature vectors were incremented with the first and second temporal derivatives. In order to create robust descriptors the dimensionality was reduced through PCA analysis, hence, creating feature sets that were uncorrelated while exhibiting the maximum variance among them.

4. RECOGNITION EXPERIMENTS

In this section, we investigate the recognition properties of the proposed features, using the GTZAN database [22], which consists of 30 sec excerpts from ten different genres. To reduce complexity during feature extraction (note that the instantaneous signals are calculated in the entire audio signal) and training, we divide each music excerpt into three 10 sec segments. This way, we also achieve the expansion of the database (1000×3), considering the triplets of each excerpt as a "different version" of the same music track. However, for the random selection of training and test sets we take notice that the triplets of the same audio file are included in either the train or the test set, since a completely random selection and mixing would result in much better recognition accuracy. In all experiments the training data is randomly selected to be 90% of the available audio signals and the results presented have emerged after 5-fold cross validation.

FR1	LMF_{72}	12log(m-IAM)+12m-IFM (+12Δ+12$\Delta\Delta$).
	LMFiPf_{52}	39 AMFM+13 FMP features after PCA.
	MFC_{39}	13 MFCC + 13 Δ + 13 $\Delta\Delta$.
FR2	$\text{LMF89}b1_{198}$	Feature sets emerged using 89 or 101 filters. b denotes the bandwidth. The final number of features have emerged after PCA analysis of the AM-FM features + Δs.
	$\text{LMF89}b1_{240}$	
	$\text{LMF101}b1_{265}$	
FR4	$\text{LMFPC}_{392(8F)}$	Feature sets emerged by concatenating 8 short time frames (8F) using the features: 12log(m-IAM)+12m-IFM+12FMP+13MFCC +58MFD, denoted by the letters LM,F,P,C,D respectively. The final number of features have emerged after PCA analysis. 8FH denotes concatenation with 50% overlap.
	$\text{LMFPCD}_{214(8F)}$	
	$\text{LMFPCD}_{214(8FH)}$	
	$\text{LMFPCD}_{428(8F)}$	
	$\text{LMFPCD}_{428(8FH)}$	
FR5	LMF_{58}	58 AMFM features selected from LMF_{72}.
	MFC_{21}	21 MFCC selected from MFC_{39}.
	LMF-MFC_{74}	74 features selected from LMF_{72}+MFC_{39}.
	LMFiD-MFC_{66}	66 features selected from $\text{LMFiD}_{40}\text{MFC}_{39}$ which has emerged after PCA analysis of LFM_{72}+MFD[$s = 1$]+MFC_{39}.

Table 1. List of proposed feature sets (the subscript in the name of each feature set shows the total number of features).

The proposed features (expect for the BoWs) were evaluated with HMMs, using the HTK [23] system, varying the number of Gaussian mixtures $M = [1 - 16]$ and states $N = 5$ and/or 7, 9. Their performance was compared with a standard set of 13 MFCCs plus their derivatives. Moreover, multi-stream modeling was conducted, where the different features were combined by changing the weight of each stream.

It is common practice for BoW representations to be evaluated using SVMs. Thus, we used nonlinear SVMs [24] (one-against-all) with a generalized Gaussian kernel χ^2 (chi-squared), where $\chi^2(Hi, Hj)$ is the distance between two histograms with K centers [19]. The χ^2 distance for comparing two histograms H_i and H_j is defined as: $\chi^2(H_i, H_j) = \frac{1}{2}\sum_{k=1}^{K} \frac{[h_{ik} - h_{jk}]^2}{h_{ik} + h_{jk}}$, where K is the number of music words (clusters).

4.1. Results

Based on the recognition results of the different feature representations, which are listed in Table 1, we show that the AM-FM features achieve good recognition of the ten music genres and better performance than the MFCCs. The MFDs were only evaluated fused with AM-FM features and results are presented for the cases they enhanced the recognition. In the following results the number of M mixtures that achieved the best recognition accuracy is shown in brackets.

Accuracy Results %				
FR	Features		HMM	
		# States :	$N=5$	$N=7$
FR1	LMF_{72}	-	76.46 (16)	76.46 (16)
	LMFiPf_{52}	-	77.86 (16)	**79.87 (15)**
	MFC_{39}	-	**78.06** (14)	78.35 (14)
FR2	$\text{LMF89}b1_{198}$	-	81.68 (14)	-
	$\text{LMF89}b1_{240}$	-	81.81 (14)	-
	$\text{LMF101}b1_{265}$	-	**83.22** (14)	-
Multi-Stream Cases				
	Features	Weights	$N=5$	$N=7$
FR1	$\text{LMFi}_{39}\text{MFC}_{39}$	0.3 - 0.7	81.40 (16)	81.47 (12)
		0.5 - 0.5	**82.81** (16)	82.27 (14)
	$\text{LMFiPf}_{52}\text{MFC}_{39}$	0.3 - 0.7	81.54 (16)	81.87 (12)
		0.5 - 0.5	82.08 (12)	**82.61** (14)
FR2	$\text{LMF89b1}_{198}\text{MFC}_{39}$	0.3 - 0.7	**84.22** (16)	-
		0.5 - 0.5	83.50 (12)	-
		0.7 - 0.3	83.13 (16)	-
	$\text{LMF89b1}_{240}\text{MFC}_{39}$	0.3 - 0.7	83.48 (15)	-
		0.5 - 0.5	**84.15** (11)	-
		0.7 - 0.3	82.74 (15)	-
	$\text{LMF101b1}_{265}\text{MFC}_{39}$	0.3 - 0.7	**84.41** (14)	-
		0.5 - 0.5	83.68 (14)	-
		0.7 - 0.3	83.28 (14)	-

Table 2. Recognition average results for 10 music genres with HMM for **FR1** and **FR2**, where N the # of states. In brackets the # of mixtures M for which the best accuracy was obtained can be seen.

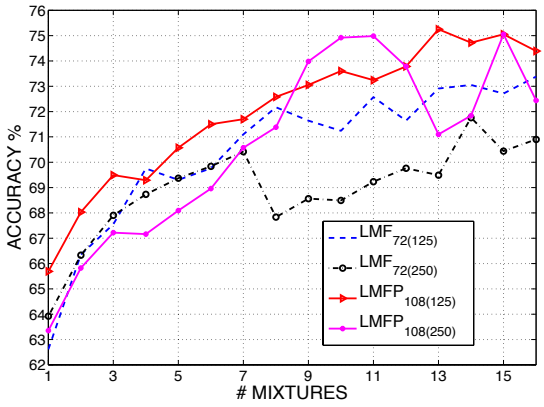

Fig. 3. Recognition average results for 10 music genres with HMM for **FR3**, where $N=5$ states and $M=[1-16]$ mixtures.

Accuracy Results %			
	HMM		
Features	$N=5$	$N=7$	$N=9$
$\text{LMFPC}_{392(8F)}$	82.61 (15)	81.20 (8)	81.20 (12)
$\text{LMFPCD}_{214(8F)}$	82.28 (14)	80.88 (15)	81.94 (11)
$\text{LMFPCD}_{214(8FH)}$	**82.88** (10)	82.34 (15)	82.01 (11)
$\text{LMFPCD}_{428(8F)}$	82.48 (16)	81.20 (15)	80.74 (10)
$\text{LMFPCD}_{428(8FH)}$	81.68 (16)	**82.35** (14)	**82.61** (13)

Table 3. Recognition average results for 10 music genres with HMM for **FR4**, where $N=5,7,9$ states.

The experimental results for **FR1** (see Table 2 for classification accuracy (%) and Table 1 for feature specific information) showed that the combination of the proposed features with the MFCCs proved out to be better than the MFCCs in all cases. Even though the AM-FM features alone exhibit lower recognition at a rate of about 0.5–2% for $N=5$, they achieve better recognition when $N=7$. Specifically, we observe that the feature set LMFiPf_{52} shows an error rate reduction (ERR) of 7% for $N=7$, while for the multi-stream experiments and the best combination, i.e., $\text{LMFi}_{39}\text{MFC}_{39}$ an ERR of 20% is obtained, for HMMs with $N=5$ and equal weights for the two streams $s_{1,2}=0.5$. Regarding the AM-FM features we note that the addition of FMPs in the mean instantaneous features, followed by PCA, see LMFiPf_{52}, reduces the error compared to LMF_{72} ca. 5% for $N=5$ and 14% for $N=7$.

The recognition results for **FR2**, namely the feature sets extracted with the music filterbank, see Table 2, are higher, with ERR at about 24% for the $\text{LMF101}b1_{265}$ compared to LMFiPf_{52} and MFC_{39}, while their combination with MFC_{39} yields an ERR of 28% compared to MFC_{39} and about 8% compared to the best combination of FR1 i.e., $\text{LMFi}_{39}\text{MFC}_{39}$ for $N=5$. Observe also that in some cases the best accuracy is gained when the weight of AM-FM features is equal to the MFCCs, which strengthens the fact that the modulations contribute notably to the specific task.

Figure 3 shows the classification accuracy for **FR3**, thus the mean AM-FM features extracted using analysis frames of 125 and 250 ms respectively. In general, we observe that this type of analysis does not perform equally good. The best recognition rate of 75.3% is achieved by $\text{LMFP}_{108(125)}$, consisting of LMF_{72} plus 36 FMP using analysis windows of 125 ms and for $M=13$ mixtures. The worst results are presented for the features extracted with analysis windows of 250 ms, except for $\text{LMFP}_{108(250)}$ for $M=[9-11,15]$, which achieved a maximum recognition of 75%. Though, we note that the addition of FMPs to the mean instantaneous features enhanced the recognition and reduced the error up to 7% and 12% for $\text{LMFP}_{108(125)}$ and $\text{LMFP}_{108(250)}$ respectively.

Table 3 shows the accuracy for the best sets of **FR4**, thus the concatenation of short time frames, with HMMs for $N=5,7$, and 9 and $M=[1-16]$. The set $\text{LMFPCD}_{214(8FH)}$ showed the greatest recognition ability 82.9% for $N=5$ and $M=10$. Nevertheless, most feature sets showed competitive properties not only because of the good recognition, but also for the short training times. Note that despite the sizable number of features after the PCA analysis, due to the reduction of the time dimension (i.e., the number of total frames) the obtained representations were quite compact. It is also important to emphasize that the specific feature representations achieved comparable to the best accuracy results with the use of only a few Gaussian mixtures. Finally, we notice that in this case the MFDs contribute to the recognition performance. Although we created sets concatenating successive frames of varying duration the evaluation showed that best recognition was achieved when only 8 frames were concatenated.

Table 4 shows accuracy results for **FR5** with SVMs. We note that BoW representations, using the AM-FM features, achieved better recognition than the MFCCs, decreasing the error about 11% (LMF_{58}) compared to MFCCs and 16% when combined, for the set LMFiD-MFC_{66}, which also in-

Features	**Accuracy Results %**
LMF_{58}	82.62
$LMFP_{50}$	82.42
MFC_{21}	80.54
LMF-MFC_{74}	82.88
$LMF_{58}MFC_{21}$	82.56
LMF-MFC-D_{60}	81.48
LMFiD-MFC_{66}	**83.56**

Table 4. Recognition average results for 10 music genres and **FR5** (BoW representations) evaluated with SVM.

	Accuracy Results %				
Music Genre	$LMF101b1_{265}$	$LMF101b1_{265}$ MFC_{39}	LMFPCD $214(8FH)$	$LMFi_{39}$ MFC_{39}	MFCC
Blues	89.3	90.0	87.3	**92.0**	88.7
Classical	95.3	**96.0**	93.4	95.3	93.3
Country	**85.8**	84.5	85.1	81.8	83.8
Disco	75.8	74.5	**78.0**	71.8	69.7
Hip-hop	74.0	**83.3**	78.7	80.0	77.3
Jazz	**94.0**	90.7	90.7	89.3	76.7
Metal	91.3	**92.7**	88.7	89.2	89.3
Pop	85.3	**94.7**	90.0	92.7	84.0
Reggae	73.3	74.7	**75.3**	72.7	69.3
Rock	**67.8**	63.1	61.8	61.1	51.0

Table 5. Recognition results for individual genres for the four best feature combinations compared to MFCCs, with HMM.

cludes the fractal dimension. Generally, we observed that the histograms for the different genres were quite dense, which we suppose it points at the less successful recognition.

Concluding, when it comes to the accuracy results for each individual genre, see Table 5, we mark that classical music has the most successful recognition, followed by pop, metal and country, while worst results were obtained for rock in all feature sets. We should also point out that the combination of AM-FM features performed better than the MFCCs for all genres.

5. CONCLUSIONS

We have examined five different approaches for the creation of feature representations, based on nonlinear features. The proposed methods were applied on classification experiments illustrating that they can capture important aspects of music, such as the micro-variations of their structure. Regarding the various representations, we conclude that short time analysis with the introduced "music" filterbank for the extraction of AM-FM appears to be really promising achieving the best recognition accuracy. Descriptors based on macro-structures, through the concatenation of short time frames, may reduce the complexity of the classification system, since satisfactory results can be achieved using simpler statistical models while the compact representations results in shorter training duration. With the Bag-of-Words we have introduced an alternative feature extraction method for music signals that creates compact representations resulting in reduced computational complexity and higher recognition compared to MFCCs. Finally, regarding the use of MFDs we note that they can benefit and enhance the recognition performance, which could thus reveal evidence of fractal aspects on musical sounds, but only when fused with other features. In our ongoing research we intent to find why some genres, such as rock, show lower performance and try to enhance their recognition.

REFERENCES

[1] J.J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Jour. New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[2] R.O. Gjerdingen and D. Perrott, "Scanning the dial: The rapid recognition of music genres," *Jour. New Music Research*, vol. 37, 2008.

[3] D. Huron and B. Aarden, "Cognitive issues and approaches in music information retrieval," 2002 [Online-unpublished]. Available: http://www.musicog.ohio-state.edu/Huron/publications.html.

[4] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content a survey," *IEEE Signal Process. Mag.*, 2006.

[5] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden Markov models," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, 2006.

[6] T. Lidy and A. Rauber, "Combined fluctuation features for music genre classification," in *Proc. ISMIR*, 2005.

[7] E. Guaus and P. Herrera, "A basic system for music genre classification," in *Proc. MIREX*, 2007.

[8] A. S. Bregman, *Auditory Scene analysis, The Perceptual Organization of Sound*, MIT Press: Cambridge, MA, 1990.

[9] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. on Signal Process.*, vol. 41, pp. 3024–3051, Oct. 1993.

[10] A. Zlatintsi and P. Maragos, "AM-FM modulation features for music instrument signal analysis and recogntion," in *Proc. EUSIPCO*, 2012.

[11] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in *Proc. EUSIPCO*, 2012.

[12] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Letters*, 2005.

[13] H.M. Teager and S.M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*. Boston, MA:Kluwer, 1989.

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

[15] A. Zlatintsi and P. Maragos, "Multiscale fractal analysis of musical instrument signals with application to recognition," *IEEE Trans. Audio, Speech and Language Process.*, vol. 21, no. 4, pp. 737–748, 2013.

[16] P. Maragos, "Fractal signal analysis using mathematical morphology," *Adv. in Electronics and electron physics, Acad. Press*, vol. 88, 1994.

[17] N. Jakovljevic, D. Miskovic, M. Janev, M. Secujski, and V. Delic, "Comparison of linear discriminant analysis approaches in automatic speech recognition," *Electronics and Electr. Enginr.*, vol. 19, 2013.

[18] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008.

[20] C.-F. Tsai, "Bag-of-Words representation in image annotation: A review," *ISRN Artificial Intelligence*, vol. 2012, 2012.

[21] A. Bosch, X. Munoz, and R. Martí, "A review: Which is the best way to organize/classify images by content?," *Image and Vis. Comp.*, 2006.

[22] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Process.*, vol. 10, 2002.

[23] S. Young et al., *The HTK Book (for HTK Version 3.4)*, Copyright © 2001-2009 Cambridge Univ. Engineer. Department, 2009.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Sys. and Tech.*, vol. 2, 2011.