MUSIC GENRE CLASSIFICATION VIA COMPRESSIVE SAMPLING

Kaichun K. Chang
Department of Computer Science
King's College London
London, United Kingdom
ken.chang@kcl.ac.uk

Jyh-Shing Roger Jang
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
jang@cs.nthu.edu.tw

Costas S. Iliopoulos
Department of Computer Science
King's College London
London, United Kingdom
csi@dcs.kcl.ac.uk

ABSTRACT

Compressive sampling (CS) is a new research topic in signal processing that has piqued the interest of a wide range of researchers in different fields recently. In this paper, we present a CS-based classifier for music genre classification, with two sets of features, including short-time and long-time features of audio music. The proposed classifier generates a compact signature to achieve a significant reduction in the dimensionality of the audio music signals. The experimental results demonstrate that the computation time of the CS-based classifier is only about 20% of SVM on GTZAN dataset, with an accuracy of 92.7%. Several experiments were conducted in this study to illustrate the feasibility and robustness of the proposed methods as compared to other approaches.

1. INTRODUCTION

1.1 Acoustic Features for Audio Music Analysis

In the literature of music information retrieval (MIR), various content-based features have been proposed [1] for applications such as classification, annotation, and retrieval [15]. These features can be categorized into two types, that is, short-time and long-time features. The short-time features are mainly based on spectrum-derived quantity within a short segment (such as a frame). Typical examples include spectral centroids, Mel-frequency cepstral coefficients (MFCC) [1], and octave based spectral contrast (OSC) [2]. In contrast, the long-time features mainly characterize the variation of spectral shape or beat information over a long segment, such as Daubechies wavelet coefficients histogram (DWCH) [3], octave-based modulation spectral contrast (OMSC), low-energy, beat histogram [1], and so on. According to G. Tzanetakis et al. [1], the short and long segments are often referred to as "analysis window" and "texture window", respectively.

Theoretically, both short-time and long-time features should be used together to realize efficient and effective MIR system since they provide different information for the task under consideration. However, in practice, too many features usually degrade the performance since there might be some noises instead of useful cues in the feature set. Moreover, too many features could also entail excessive computation to downgrade the system's efficiency. As a result, we need an effective method for feature selection, extraction, or distillation. CS turns out to be an effective tool for such a purpose.

1.2 Compressive Sampling

CS is firstly proposed by Candès, Romberg, Tao and Donoho, who have showed that a compressible signal can be precisely reconstructed from only a small set of random linear measurements whose number is below the one demanded by the Shannon theorem Nyquist rate. It implies the potential of a dramatic reduction in sampling rates, power consumption, and computation complexity in digital data acquisitions. CS has proved to be very effective in imaging [6] [7], channel estimation [8], face recognition [9], phonetic classification [18], sensor array [19] and motion estimation [20].

In this paper, we propose a CS-based classifier with long-time and short-time features for music genre classification. The remainder of this paper is organized as follows. In section 2, the multiple feature sets used in the proposed method is briefly discussed. In the section 3, we describe multiple feature sets for audio music, and introduce the corresponding CS-based classifier. In section 4, experimental settings and results are detailed to demonstrate the proposed method's feasibility. Finally, conclusions and future work are addressed in the last section.

2. MULTIPLE FEATURE SETS

In the proposed method, multiple feature sets including long-time and short-time features are adopted for genre classification. These acoustic features include timbral texture features, octave-based spectral contrast (OSC), octave-based modulation spectral contrast (OMSC), modulation spectral flatness measure (MSFM), and modulation spectral crest measure (MSCM).

Timbral texture features are frequently used in various music information retrieval system [11]. Some timbral texture features, described in Table 1, were proposed for audio classification [1]. Among them, MFCC, spectral centroid, spectral rolloff, spectral flux, and zero crossings are

Table 1. Timbral texture features

Feature	Description
MFCC	Representation of the spectral characteristics based on Mel-frequency scaling [12]
Spectral centroid	The centroid of amplitude spectrum
Spectral rolloff	The frequency bin below which 85% of the spectral distribution is concentrated.
Spectral flux	The squared difference of successive amplitude spectrum.
Zero crossings	The number of time domain zero crossings of the music signal.
Low-energy	The percentage of analysis windows that have energy less than the average energy across the texture window.

short-time features, thus their statistics are computed over a texture window. The low-energy feature is a long-time feature.

Besides these features, OSC and OMSC features are also considered. OSC considers the spectral peak, spectral valley, and spectral contrast in each subband [2]. The spectrum is first divided into octave-based subband (as explained next). Then spectral peaks and spectral valleys are estimated by averaging across the small neighborhood around maximum and minimum values of the amplitude spectrum respectively. OMSC [1] is extracted using long-time modulation spectrum analysis [13].

In this paper, the amplitude spectrum of a music signal is divided into octave-based subbands of 0-100Hz, 100Hz-200Hz, 200Hz-400Hz, 400Hz-800Hz, 800Hz-1600Hz, 1600Hz-3200Hz, 3200Hz-8000Hz, 8000Hz-22050Hz. Within each subband, the amplitude spectrum is summed. Then for each subband, the modulation spectrum is obtained by applying the discrete Fourier transform (DFT) on the sequence of the sum of amplitude spectrum.

OMSC is obtained from spectral peaks and spectral contrasts of the modulation spectrum. MSFM and MSCM are obtained from a texture window [4] using the long-time modulation spectrum [13] that can describe the time-varying behavior of the subband energy. These features are also considered as parts of our multiple feature sets.

3. COMPRESSIVE SAMPLING BASED CLASSIFIER

As inspired by CS and the sparse signal representation theory, here we shall propose a CS-based classifier for genre classification. First of all, we shall cover the basics of the CS theory [5].

In Figure 1, consider a signal x (length N) that is K-sparse in sparse basis matrix Ψ, and consider also an $M \times N$ measurement basis matrix Φ, $M << N$(M is far less than N), where the rows of Φ are incoherent with the columns of Ψ. In term of matrix notation, we have $x = \Psi\theta$, in

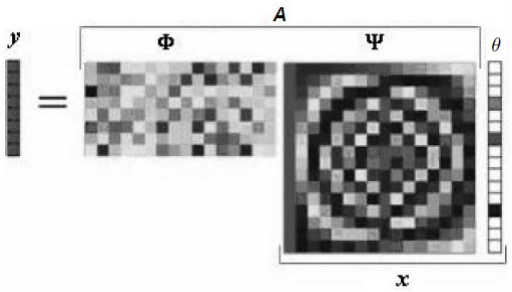

Figure 1. The measurement of Compressive Sampling

which θ can be approximated using only $K << N$ non-zero entries. The CS theory states that such a signal x can be reconstructed by taking only $M = O(K \log N)$ linear, non-adaptive measurement as follows:

$$y = \Phi \cdot x = \Phi \cdot \Psi \cdot \theta = A \cdot \theta, \quad (1)$$

where y represents an $M \times 1$ sampled vector, $A = \Phi\Psi$ is an $M \times N$ matrix. The reconstruction is equivalent to finding the signal's sparse coefficient vectors θ, which can be cast into a ℓ_0 optimization problem.

$$min \left\|\theta\right\|_0 \;\; s.t.\; y = \Phi \cdot x = A \cdot \theta \quad (2)$$

Unfortunately (2) is in general NP-hard, and an optimization ℓ_1 is used to replace the above ℓ_0 optimization [10].

$$min \left\|\theta\right\|_1 \;\; s.t.\; y = \Phi \cdot x = A \cdot \theta \quad (3)$$

Let the dimension of the extracted feature be denoted as m, and the extracted feature vector of the j-th music in the i-th class as $\nu_{i,j} \in R^m$. Moreover, let us assume there are sufficient training samples for the i-th class $A_i = [\nu_{i,1}, ...\nu_{i,n_i}] \in R^{m \times n_i}$. Then any new (test) sample $y \in R^m$ (i.e, the extracted feature of the test music) from the same class will approximately lie in the linear span of the training samples associated with object i:

$$y = \sum_{i=1}^{n_i} \alpha_{i,n_i} \, \nu_{i,n_i}, \quad (4)$$

for some scalars $\alpha_{i,j}(j = 1, .., n_i)$. Since the membership i (or the label) of the test sample is initially unknown, we define a new matrix $\mathbf{A}$ for the entire training set as the concatenation of the n training samples of all k classes: $\mathbf{A} = [A_1, ...A_k]$ Then the linear representation of y can be rewritten in terms of all training samples as:

$$y = \mathbf{A}x_0 \in R, \quad (5)$$

where $x_0 = [0, .., 0, \alpha_{i,1}, ..., \alpha_{i,n}, ..., 0, .., 0]^T \in R^n$ is a coefficient vector whose entries are zero except those associated with the i-th class. As the entries of the vector x_0 encode the identity of the test sample y, it is tempting to obtain it by solving the equation (4). This is called a sparse representation based classifier (SRC) [9].

In SRC, for a new test sample y from one of the classes in the training set, we first compute its sparse representation $\hat{x}$ via (2). Ideally, the nonzero entries in the estimate $\hat{x}$ will all be associated with the columns of $\mathbf{A}$ from a single object class i, and we can easily assign the test sample y to that class. To better harness such linear structure, we instead classify y based on how well the coefficients associated with all training samples of each object reproduce y. For each class i, let $\delta_i : R^n \rightarrow R^n$ be the characteristic function which selects the coefficients associated with the i-th class[12]. For $x \in R^n$, $\delta_i(x) \in R^n$ is a new vector whose only nonzero entries are the entries in x that are associated with class i. Using only the coefficients associated with the i-th class, one can approximate the given test sample y as

$$\hat{y}_i = A\delta_i(\hat{x}) \tag{6}$$

We then classified y based on these approximations by assigning it to the object class that minimizes the residual between y and $\hat{y}_i$:

$$\min r_i(y) = \|y_i - A\delta(\hat{x})\|_2 \tag{7}$$

The proposed CS-based classifier is based on the principle of SRC, with an additional random measurement on the extracted features to reduce the dimension of the input. According to the CS theory, this reduction can capture the structure of the features and automatically remove possible redundancy. The realization of the algorithm is summarized in Table 2.

It should be noted that SRC is a sparse representation based classifier, without the dimension reduction over the input signals. Here the random measurement of compressive sampling is used to perform a dimension reduction and feature extraction. So the classification complexity of CS based method is remarkably lower than that of SRC. Moreover, the multiple features will also improve the classification accuracy. The sparse representation is one part of compressive sampling. Taking the training samples matrix as the transform matrix will be helpful to the classification. We will find that the procedure of CS based classifier are very different from the classical methods, because steps 3 to 6 are all based on compressive sampling. Currently many non-linear dimensionality reduction methods have been proposed, such as Local Coordinates Alignment(LCA) and Non-Negative Matrix Factorization(NMF). Compressive sampling theory provides a random measurement of signals, and proves to be able to keep the information of the signals under the condition of enough number of measurement and incoherence between the measurement matrix and the transform matrix.

Consequently, it is a natural compressive process of signals, which can also be regarded as the process of dimension reduction. CS is different from LCA and NMF due to that fact that it is a near method, which lend itself to efficient implementation.

Table 2. CS-based Classification

Algorithm: CS-based Classification

Step 1:
Perform a feature extraction on the music samples for k classes.
Step 2:
Perform a feature extraction (described in section 2) on the training songs to obtain a matrix of training samples $\mathbf{A} = [A_1, ...A_k]$ and calculate the feature y of the test sample.
Step 3:
Perform a random measurement (the measurement matrix is a Gaussian random matrix) on the features of the training samples and the test sample feature to obtain $\mathbf{A}' = H \cdot \mathbf{A}$ and $y' = H \cdot y$ respectively.
Step 4:
Normalize the columns of $\mathbf{A}'$ to have unit ℓ_2 norm and solve the ℓ_1-minimization problem:
$\min \|x\|_1 \ \ s.t.\ y' = \mathbf{A}' \cdot x$
Step 5:
Compute the residuals
$\min r_i(y') = \left\|y' - \mathbf{A}'\delta_i(\hat{x})\right\|_2$
Step 6:
Output: $identity(y) = \arg\min r_i(y')$

Table 3. Classification accuracies achieved by various methods on GTZAN dataset.

Method	Dataset	Accuracy	Feature dimensions
MF + CSC (Ours)	GTZAN	92.7	64
TPNTF + SRC	GTZAN	93.7	135
NTF + SRC	GTZAN	92.0	135
MPCA + SRC	GTZAN	89.7	216
GTDA + SRC	GTZAN	92.1	216

4. EXPERIMENTAL RESULTS

The experiments are divided into three parts. Section 4.1 details our experiment with music genre classification. Section 4.2 explores multiple features and dimension reduction. Section 4.3 investigates the feature extractor in an noisy environment.

4.1 Music Genre Classification

Our experiments of music genre classification are performed on GTZAN dataset, which are widely used in the literature [16]. GTZAN consists of the following ten genre classes: Classical, Country, Disco, Hip-Hop, Jazz, Rock, Blues, Reggae, Pop, and Metal. Each genre class contains 100 audio recordings of 30 seconds, with sampling rate of 44.1kHz and resolution of 16 bits.

To evaluate the proposed method for genre classification, we set up all the experimental parameters to be as close as possible to those used in [18]. In particular, the recognition rate is obtained from 10-fold cross validation. Table 3 is a comparison table which lists several other ex-

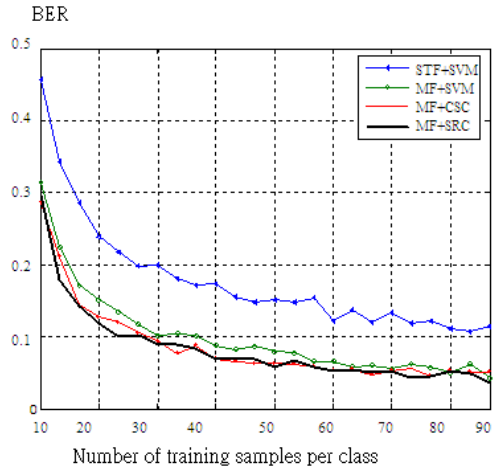

Figure 2. Genre classification result.(CSC is ours)

isting methods together with their recognition rates, such as Topology Preserving Non-Negative Matrix Factorization (TPNMF), Non-Negative Tensor Factorization (NTF), Multilinear Principal Component Analysis (MPCA), and General Tensor Discriminant Analysis (GTDA) [17]. As can be seen from the table, the proposed method (MF + CSC) outperforms all the state-of-the-art SRC-based approaches except one. Moreover, the feature dimension of the proposed approach is considerably lower than those of the SRC-based approaches, demonstrating the effective of CS in extracting features with discriminating power.

This experiment addresses the problem of genre classification using Compressive Sampling. A CS recovery is applied on short-term and long-term features that are relevant for genre classification. The measurement vectors are trained on labeled sets, then the classification is performed by computing the approximation of unknown samples with each class-specific features.

Figure 2 plots the recognition rates of the four methods with respect to no. of training samples per class. (The no. of training samples were randomly selected from each class, while the test samples stayed the same.) The figure demonstrates that multiple features indeed improve the classification accuracy. Moreover, CSC and SRC do have consistent higher accuracy than SVM classifier. More importantly, these two methods do not require the long training process of SVM. In Figure 3 , the computation time of MF+SRC and MF+CSC is only 30% and 20%, respectively, of SVM, due to dimension reduction in compressive sampling.

Table 4 shows the confusion matrix of the CS-based classifier [1]. The columns stands for the actual genre and the rows for the predicted genre. It can be seen that the recognition rate of each class is almost evenly distributed.

4.2 Multiple Features Dimension

In this experiment, we combine feature sets (long-time features and short-time features, and short-time features only) and different classifiers (SVM [14], SRC and the proposed classifier) to investigated their joint effects. The descrip-

Table 4. Confusion matrix of the proposed method

	cl	co	di	hi	ja	ro	bl	re	po	me
cl	**96**	0	0	3	1	0	0	0	0	0
co	0	**92**	4	0	2	0	0	1	0	1
di	0	4	**93**	0	0	1	0	1	0	1
hi	3	0	0	**94**	0	1	1	1	0	0
ja	1	2	0	0	**93**	0	3	1	0	0
ro	0	0	1	1	0	**89**	2	3	3	1
bl	0	0	0	1	3	2	**90**	1	3	0
re	0	1	1	1	1	3	1	**92**	0	0
po	0	0	0	0	0	3	3	0	**94**	0
me	0	1	1	0	0	1	0	0	0	**97**

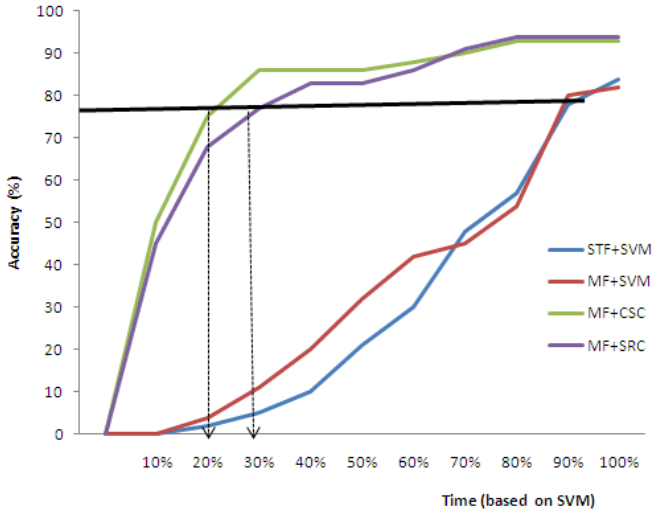

Figure 3. Genre classification time analysis.(CSC is ours)

tions of these methods and their parameter settings are shown in Table 5.

All the samples are digitized 44.1 kHz, 16-bit, and mono signal in preprocessing. The 30-seconds of audio clips after initial 12 seconds are used. The length of the analysis window was set to 93ms, and 50% overlap was used for feature extraction. The length of texture window was set to 3 second, thus a texture window contains 63 analysis windows. The 13-dimensional MFCCs are computed in an analysis window, mean and variance of each dimension are computed in a texture window.

Table 6 shows the multiple features set and dimension. As mentioned in section 2, eight octave subbands were used to compute the OSC, OMSC, MSFM, and MSCM. They are computed based on octave subband. Thus, the dimensions of the features are dependent on the number of octave subband (eight subbands were used in this experiment). The dimensions of the OSC, the OMSC, the MSFM, and the MSCM are respectively 32, 32, 8 and 8.

4.3 Under Noise Environment

In Figure 2, sparse representation based classifier and CS-based classifier have similar performance in music genre classification. The robustness of the system is tested under

Table 5. Methods used in the experiment

Method	Description	Parameters
STF+SVM	Short-time feature only and followed by a SVM classifier	SVM is used and α takes between 0 and 1. The optimal value is chosen experientially.
MF+SVM	Multiple feature and followed by a SVM classifier	As above
MF+CSC	Multiple feature and followed by a compressive sampling based classifier	The sampling rate takes 67% and the optimization algorithm is basis pursuit algorithm.
MF+SRC	Multiple feature and followed by a sparse representation based classifier	The optimization algorithm is basis pursuit algorithm.

Table 6. Multiple features set and dimension

Feature	Set	Dimension
OMSC	Long-time feature	32
Low-energy	Long-time feature	1
OSC	Short-time feature	32
MFCCs	Short-time feature	26
MSFM	Short-time feature	8
Spectral centroid	Short-time feature	2
Spectral rolloff	Short-time feature	2
Spectral flux	Short-time feature	2
Zero crossings	Short-time feature	2

the following conditions.

- Additive white uniform noise (AWUN)
- Additive white Gaussian noise (AWGN)
- Linear speed change (LSC)
- Band-pass filter (BPF)

The robustness of these two methods was compared, as shown in Table 7. We can find the average BER of the CSC system is lower than SRC. CSC has better performance under the conditions of linear speed change, band-pass filter, and additive white uniform noise.

Figure 4 shows the classification results of different methods when the Gaussian noise with different variance are added to the music. From the figure, we can see that the proposed method is quite immune to noise.

5. CONCLUSIONS

In this study, we have proposed a CS-based classifier and verified its performance by a common dataset for music genre classification. Moreover, we have also explored the possibility of using multiple feature sets for improving the performance of genre classification. The experiments demonstrates that the proposed CS-based classification together with the use of multiple feature sets outperform quite a few state-of-the-art approaches for music genre classification. The success of the proposed CS-based classifier is attributed to CS's superb capability in feature extraction for generating parsimonious representation of the original signals.

For immediate future work, we will focus on the possibility of porting the proposed CS-based classifier for other MIR tasks, such as onset detection, beat tracking, and tempo estimation.

Table 7. The comparison result about robustness

	CSC		SRC	
	Rate(%)	BER	Rate (%)	BER
AWUN	73.8	0.262	73.5	0.265
AWGN	76.6	0.234	78.8	0.212
LSC	81.7	0.183	64.8	0.352
BPF	71.2	0.288	65.8	0.342

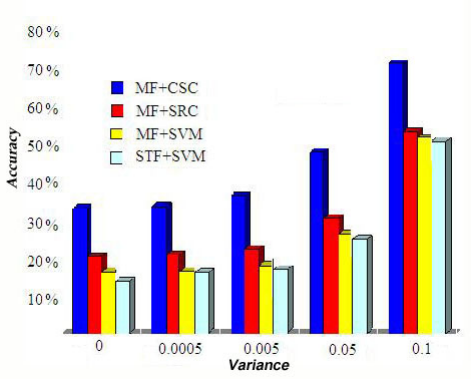

Figure 4. Genre classification result under noise.

6. REFERENCES

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

[2] D. N. Jiang, L. Lu, H. J. Zhang, J. II. Tao, and L. II. Cai: "Music type classification by spectral contrast feature,"*Proc. ICME 02*, vol. I, pp. 113-116, 2002.

[3] T. Li , M. Ogihara, and Q. Li: "A comparative study on content based music genre classification,"*Proc. ACM Con! on Research and Development in Information Retrieval*, pp. 282-289, 2003.

[4] D. Jang and C. Yoo: "Music information retrieval using novel features and a weighted voting method,"*IEEE International Symposium on Industrial Electronics (ISlE 2009) Seoul Olympic Parktel*, Seoul, Korea July 5-8, 2009.

[5] D. Donoho: "Compressed sensing,"*IEEE Transactions on Information. Theory*, 52(4), pp. 1289-1306, Apr. 2006.

[6] J. Romberg: "Imaging via compressive sampling,"*IEEE Signal Processing Magazine*, 25(2), pp. 14 - 20, March 2008.

[7] Z. Chen, J. Wen, Y. Han, J. Villasenor, S. Yang: "A compressive sensing image compression algorithm using quantized DCT and noiselet information,"*ICASSP*, 2010.

[8] W. Bajwa, J. Haupt, A. Sayeed and R. Nowak: "Joint source-channel communication for distributed estimation in sensor networks,"*IEEE Transactions on Signal Processing*, 53(10), pp. 3629-3653, October 2007.

[9] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma: "Robust face recognition via sparse representation,"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), pp. 210-227, February 2009.

[10] E. Candes, M. Wakin, and S. Boyd: "Enhancing sparsity by reweighed ℓ_1 Minimization,"*Journal of Fourier Analysis and Applications*, 14(5), pp. 877-905, October 2008.

[11] L. Lu, D. Liu, and H-J. Zhang: "Automatic mood detection and tracking of music audio signals,"*IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, Jan. 2006.

[12] X. Huang, A. Acero and H.-W. Hon: "Spoken Language Processing,"*Prentice Hall PTR*, 2001.

[13] S. Sukittanon, L. E. Atlas, and 1. W. Pitton: "Modulation-scale analysis for content identification,"*IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 3023-3035, Oct., 2004.

[14] C.-C. Chang and C.-J. Lin: "LIBSVM - A Library for Support Vector Machines,"*[Online] Available: www.csie.ntu.edu.tw/.*

[15] I. Karydis , A. Nanopoulos , A. Papadopoulos , Y. Manolopoulos: "Audio Indexing for Efficient Music Information Retrieval,"*Proceedings of the 11th International Multimedia Modeling Conference*, p.22-29, January 12-14, 2005.

[16] D. Jang; M. Jin; C. Yoo: "Music genre classification using novel features and a weighted voting method,"*In Proceeding of the 2008 IEEE International Conference on Multimedia and Expo*, pp. 1377–1380, April 2008.

[17] Y. Panagakis, C. Kotropoulos: "Music genre classification via topology preserving non-negative tensor factorization and sparse representations,"*ICASSP*, 2010.

[18] T. Sainath, A. Carmi, D. Kanevsky, B. Ramabhadran: "Bayesian compressive sensing for phonetic classification,"*ICASSP*, 2010.

[19] Y. Yoon, M. Amin: "Through-the-wall radar imaging using compressive sensing along temporal frequency domain,"*ICASSP*, 2010.

[20] N. Jacobs, S. Schuh, R. Pless: "Compressive sensing and differential image motion estimation,"*ICASSP*, 2010.