SHAPE-BASED SPECTRAL CONTRAST DESCRIPTOR

Vincent Akkermans
Faculty of Art, Media & Technology
Utrecht School of the Arts
Utrecht, the Netherlands
vincent.akkermans@student-kmt.hku.nl

Joan Serrà, Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
{joan.serraj, perfecto.herrera}@upf.edu

ABSTRACT

Mel-frequency cepstral coefficients are used as an abstract representation of the spectral envelope of a given signal. Although they have been shown to be a powerful descriptor for speech and music signals, more accurate and easily interpretable options can be devised. In this study, we present and evaluate the shape-based spectral contrast descriptor, which is build up from the previously proposed octave-based spectral contrast descriptor. We compare the three aforementioned descriptors with regard to their discriminative power and MP3 compression robustness. Discriminative power is evaluated within a prototypical genre classification task. MP3 compression robustness is measured by determining the descriptor values' change between different encodings. We show that the proposed shape-based spectral contrast descriptor yields a significant increase in accuracy, robustness, and applicability over the octave-based spectral contrast descriptor. Our results also corroborate initial findings regarding the accuracy improvement of the octave-based spectral contrast descriptor over Mel-frequency cepstral coefficients for the genre classification task.

1 Introduction

Music information retrieval (MIR) studies processes, systems and contexts for automatically acquiring information about music from large collections [8]. It plays an increasingly important role in a society that moves towards a freely accessible abundance of recorded music. The main audiences benefiting from MIR research are end-users, industry bodies and academics. Users have easier and personalized access to their collections, the industry employs these methods in the production process from creation to distribution, and researchers are able to discover new patterns in large corpora of data [3].

Content-based MIR methods extract information from the music itself rather than from any supplied meta-data. One of the prototypical tasks in content-based MIR is the automatic classification of a song, in the form of an audio signal, into a music genre [1, 3, 11, 12]. The octave-based spectral contrast (OBSC) is a descriptor specifically designed for this task [6, 14]. The spectral contrast of a sub-band in a signal can be seen as a measure of the signal's difference to white noise [10, 14]. In addition, the concept of spectral contrast is also used to enhance sound for hearing impaired people [16]. Because the spectral contrast of an audio signal is based on its timbre it is related to descriptors like spectral centroid, roll-off, flatness, skewness, spread and Mel-frequency cepstral coefficients (MFCCs) [10, 11, 12]. MFCCs were originally developed for use in speech recognition applications and later on proved to be useful for music information retrieval [10, 12]. They provide good discriminative power but can be hard to interpret [9].

In this study the shape-based spectral contrast (SBSC) descriptor is presented. SBSC yields a significant increase in accuracy, robustness, and applicability over OBSC. For evaluation, SBSC is compared to OBSC and MFCCs in terms of discriminative power and MP3 compression robustness. Discriminative power is evaluated by measuring their accuracy on different combinations of data sets and classifiers for the automatic genre classification task [1, 12]. We study the robustness of the descriptors at different MP3 encodings as it is not so common in the literature to test MP3 robustness [5, 13]. A descriptor is considered to be robust when its values do not change significantly while the audio it describes is encoded at different levels of MP3 compression. Finally, information overlap between MFCC and spectral contrast is briefly investigated.

The structure of this document is as follows. In section 2 we briefly summarize OBSC. Section 3 presents the SBSC descriptor and section 4 the evaluation methodology. In section 5 the results are presented and we conclude our study in section 6.

2 Octave-based Spectral Contrast

The OBSC descriptor was introduced by Jiang et al. in [6]. It describes the ratio between the magnitudes of the peaks and valleys within sub-bands of the frequency spectrum. This way, the relation of harmonic to non-harmonic frequency components of each sub-band is

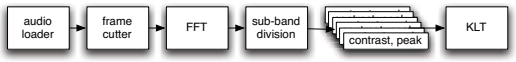

Figure 1. Spectral contrast descriptor general block diagram

reflected.

This feature has been proven useful for genre classification [6, 14]. The two aforementioned studies explain in detail how the descriptor is calculated. In short, the audio signal is loaded and cut into frames with an overlap of 50% (figure 1). Then, the spectral data resulting from an FFT of every frame is divided into 6 octave-scaled sub-bands (with boundaries at 0, 200Hz, 400Hz, 800Hz, 1.6kHz, 3.2kHz, and 8kHz). For each band k, the magnitudes of the FFT bins x are sorted into descending order and the peak P_k and valley V_k values are subsequently calculated by averaging a percentage of the highest and lowest magnitudes. The ratio between P_k and V_k, defined as contrast C_k, and V_k itself comprise the feature vector $OBSC_k$ of a single frame, with a dimensionality of twice the number of sub-bands:

$$OBSC_k = [C_k, V_k], \tag{1}$$

where

$$C_k = \log \frac{P_k}{V_k}, \tag{2}$$

$$P_k = \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k} x_{k,i}, \tag{3}$$

and

$$V_k = \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k} x_{k,N_k-i+1}. \tag{4}$$

Here α corresponds to the part of all bins in the band to average over ($0 < \alpha \leq 1$), N_k to the total number of bins in the sub-band, and i denotes the sorted bin index. In [6], α is set empirically to 0.02. Finally, the dimensions of the feature vector of OBSC are decorrelated for each frame by a Karhunen-Loève Transform (KLT) to increase the accuracy. The orthogonal base vectors for the KLT are generated from the averaged covariance matrices of all classes involved in the problem [6].

3 Shape-based Spectral Contrast

The SBSC descriptor is a modification of the OBSC descriptor intended to improve accuracy, robustness and applicability. It does so by employing a diferent sub-band division scheme, an improved notion of contrast, and a different use of the KLT transform.

3.1 Accuracy

OBSC calculates a sub-band's contrast by regarding its valley and peak. However, by including information about the shape of the band's sorted spectrum, a better estimation of spectral contrast can be made. Figure 2 shows two possible shapes of sorted sub-bands. Both possibilities would yield the same C_k value, as P_k and V_k would be the same. However, if we consider noise to be the equal presence of all frequencies, figure 2A intuitively corresponds to a more noisy sound than figure 2B. Accordingly, the shape in figure 2B should result in a higher spectral contrast.

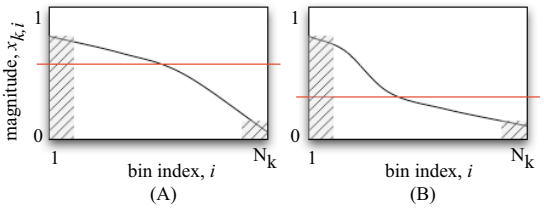

Figure 2. Two possible shapes of sub-bands sorted by magnitude. The horizontal red line indicates the average magnitude of the sub-band.

When we look at the location of the average magnitude in the sub-band relative to the peak and the valley we can better distinguish between both shapes. Accordingly, the contrast C_k (equation 2) could be calculated as:

$$C'_k = \frac{log(P_k/V_k)}{log\mu_k}, \tag{5}$$

where

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_{k,i} \tag{6}$$

Equation 5 is equivalent to $log_{\mu_k}(P_k/V_k)$ and expresses the spectral contrast in base μ_k. Through trial and error, the following equation was determined to have similar characteristics but a slightly better accuracy:

$$C'_k = \left(\frac{P_k}{V_k}\right)^{1/log\mu_k} \tag{7}$$

3.2 Robustness

Because the spectral contrast is calculated from the peaks and valleys, MP3 compression, which eliminates masked frequencies, might have a large effect on the robustness of OBSC [4]. When peaks and valleys are averaged over a

larger number of bins, eliminated frequencies have a smaller impact and the spectral contrast can be expected to be more robust. The neighbourhood ratio α and the total number of bins in a sub-band N_k are therefore looked at for increasing robustness.

For flexible adjustment of N_k and the number of sub-bands K, a different sub-band division scheme is used in SBSC instead of the original octave-based scheme. A portion of the bins of the FFT analysis is distributed equally among sub-bands while the rest is distributed exponentially:

$$B_k = \left(\frac{(1-s)U}{L}\right)^{k/K} L + \frac{(U-L)sk}{K}, \qquad (8)$$

where B_k denotes the upper boundary of the k-th band, s the portion of the spectrum to distribute equally ($0 \leq s \leq 1$), and U and L the upper and lower bound (in Hz) of the spectrum, respectively. This way, all sub-bands contain enough bins to be stable and the distribution still mimics the non-linear frequency response of the human ear [7]. In all SBSC tests in this study, $s = 0.15$ and $L = 20Hz$ (see section 4.1). For $U = 11kHz$ and $K = 6$, boundaries are 20 Hz, 330 Hz, 704 Hz, 1256 Hz, 2303 Hz, 4729 Hz, and 11 kHz. Parameters K, and U vary from test to test and are set manually.

3.3 Applicability

Instead of applying a KLT based on the averaged covariance matrices of all classes [6] (see section 2), SBSC applies the KLT based on the covariance matrix of each individual song. This does not perform significantly different and has two additional advantages: (a) when either the instances in the data set or the number of categories changes, the average covariance matrix will not have to be recalculated and (b) the descriptor can be used in other applications where no training data set is required (e.g., song similarity, audio fingerprinting, etc.).

4 Evaluation Methodology

We here detail the methods employed for determining genre classification accuracy, evaluating MP3 compression robustness, and studying MFCC and SBSC information overlap.

4.1 Genre Classification Accuracy

The discriminative power of both OBSC and SBSC is compared by measuring their accuracy in a genre classification problem. In addition, we evaluate MFCC (12 coefficients with the zeroth coefficient included) as a baseline. The means and variances of the feature vectors of all frames are used for classification.

Three distinct classifiers are trained and evaluated on two data sets. The classifiers used are decision trees, support vector machines (SVM), and linear logistic regression models (LLR) as implemented by the WEKA software [1] [15]. All classification results are computed on a 10-fold cross-validation scheme and averaged over 10 runs. The 2 data sets used are the following. Data set A is the same data set that was used by Tzanetakis in [12]. It consists of 10 genres with 100 song-excerpts per genre and has a sample rate of 22050 Hz. The genres include blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Data set B was developed in-house, has a sample rate of 44100 Hz, and consists of 55 full songs for each of 8 genres, which are classical, dance, hip-hop, jazz, pop, rhythm and blues, rock, and speech.

The data set employed by Jiang et al. in [6] had a frequency range of 8 kHz. To check the effect of a bigger frequency range, OBSC is compared to MFCCs at two different frequency ranges, 8 kHz and 11 kHz. Because MFCCs are more accurate at a frequency range of 11 kHz, SBSC is only tested at this frequency range. The SBSC descriptor is also tested with the spectrum divided into both 6 and 9 bands. Finally, MFCCs, OBSC, and SBSC are tested with the delta coefficients included. The descriptors are evaluated at the settings summarized in table 1. For all SBSC tests, $\alpha = 0.4$ and for all OBSC tests $\alpha = 0.1$ (the α value in [6] is 0.02, and is said to perform the same as $\alpha = 0.1$).

Test	Descriptor	Freq. range	Bands
MFCC-A	MFCC	8 kHz	12
MFCC-B	MFCC	11 kHz	12
MFCC-C	MFCC + ΔMFCC	11 kHz	12
OBSC-A	OBSC	8 kHz	6
OBSC-B	OBSC	11 kHz	6
OBSC-C	OBSC + ΔOBSC	11 kHz	6
SBSC-A	SBSC	11 kHz	6
SBSC-B	SBSC	11 kHz	9
SBSC-C	SBSC + ΔSBSC	11 kHz	6

Table 1. Test settings for descriptors.

4.2 MP3 Compression Robustness

The two spectral contrast descriptors and MFCCs are tested for MP3 robustness on a song by song basis. OBSC+ΔOBSC is not tested for robustness due to the expectancy of it to be highly unstable and the considerable time it takes to run the test. The wave files of data set A are compressed at two different compression rates, 192 kb and 64 kb by the Lame MP3 encoder [2] . The descriptors for all three versions of all songs in the data set are extracted. In order to make the descriptors from the compressed and original data comparable in terms of distributions and ranges, we adapted Box-Cox's transformation [2]

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://lame.sourceforge.net/

for obtaining uniform distributions for each descriptor attribute between 0 and 1. The transformation is calculated from the uncompressed descriptors and applied to both the uncompressed and compressed descriptors. For each dimension of each song, the absolute difference between the values of the uncompressed and compressed descriptor sets is calculated:

$$D_{h,d,e} = |R_{h,d,e} - Q_{h,d}|\,, \tag{9}$$

where $D_{h,d,e}$ is the robustness value of the h-th song for the d-th dimension of the e-th encoding, and R and Q denote the descriptor values of the encoded version and the uncompressed version, respectively. We say that the descriptor fails for that particular song, dimension and encoding when $D_{h,d,e} > 0.1$. The percentage of songs that fail per dimension is calculated and the average and maximum are kept as the final robustness error measures.

4.3 MFCC and SBSC Information Overlap

We follow different paths with the aim to obtain converging evidence of the amount of overlap of information between MFCC and SBSC. First, we use several attribute selection techniques on the combined feature vectors MFCC-B and SBSC-A of data set A in order to check the ranking of features or the near-optimal subset that is chosen. The selection methods used are correlation based feature subset selection (CFS), SVM ranking, and individual attribute rankers based on the most frequently preferred indices (infogain, gain ratio, and chi-square test) [15]. The percentage of SBSC attributes, as opposed to MFCC attributes, present in the first and last quartile of the combined ranked attribute list is used as a measure of information overlap. If there is no clear majority of one of the features in the quartiles, this can be interpreted as evidence of overlap.

Secondly, we perform a one-way analysis of variance (ANOVA) on the ranks obtained by the feature selection indices mentioned above, in order to check if there is an effect of the subset type (MFCC versus SBSC) on the index value. If the subset proves not to be significant, then we can consider it as evidence of overlap between them.

5 Results

5.1 Genre Classification Accuracy

Our results of analyzing and testing the OBSC descriptor corroborate the findings of Jiang et al. [6] and West and Cox [14]. In our tests (table 2), for both data sets and all three classifiers, the OBSC descriptor performs better than MFCCs, although the increase in accuracy is not as high as previously reported in [6]. There, on a different data set and using Gaussian mixture models, OBSC performs at 82.3% and MFCCs at 74.1%. The accuracy of our MFCC implementation is tested with a naive Bayes single Gaussian classifier (47.6% for 13 coefficients) and is similar to the one achieved on the same data set in [12] (47% for 10 coefficients).

We can see in table 2 that SBSC's accuracy is higher than that of OBSC and MFCCs for SVM and LLR. For these two classifiers the average relative increase for both data sets is 6.5%. Only with trees OBSC performs better than SBSC, but these accuracies are significantly lower than those achieved with SVM or LLR.

We also see that an increased frequency range has a small and slightly irregular effect on the accuracy (MFCC-A,B and OBSC-A,B, table 2). It raises MFCCs' accuracy for almost every combination of classifier and data set, while OBSC's accuracy only increases for LLR on data set A and trees on data set B.

Test	Data set A			Data set B		
	Trees	SVM	LLR	Trees	SVM	LLR
MFCC-A	41.3	60.0	**61.3**	56.6	77.6	**78.6**
MFCC-B	41.9	60.6	**63.6**	59.7	76.8	**78.6**
MFCC-C	48.2	**71.6**	71.1	63.1	**81.4**	80.1
OBSC-A	47.4	61.6	**62.4**	58.7	82.8	**83.8**
OBSC-B	46.4	61.4	**64.4**	64.2	**81.4**	81.0
OBSC-C	49.0	67.3	**69.0**	61.4	**82.2**	80.7
SBSC-A	45.5	67.3	**68.1**	63.1	**85.7**	85.5
SBSC-B	48.5	67.0	**68.9**	62.3	**86.8**	85.5
SBSC-C	49.9	72.5	**72.7**	65.1	**86.2**	**86.2**

Table 2. Genre classification accuracy (%) for the descriptors tested.

Looking at the SBSC-A and SBSC-B tests we can see that using more than six sub-bands for SBSC does on average only provide a slightly better performance. Including the delta coefficients for SBSC increases accuracy but does not provide the same improvement as the delta coefficients do for MFCCs (an average of 9.8% for MFCC+ΔMFCC and 4.8% for SBSC+ΔSBSC). Including the delta coefficients for OBSC provides a modest average accuracy increase of 3.1%.

In addition to the results presented in table 2, we also studied the effect of different KLT application variants (sections 2 and 3). Decorrelating the dimensions of the feature vector using KLT for each song separately results in a small decrease in accuracy of 0.3%. Not applying KLT results in a significant performance drop (e.g. from 67.3% to 59.7% for SBSC when using SVM).

5.2 MP3 Compression Robustness

In table 3 it can be seen how unstable OBSC is, and how robust the MFCC implementation is in comparison. MFCCs only fail for 0.7% of the songs, while OBSC fails for 13.6% (MFCC-B and OBSC-B, 64 kb, table 3). A

higher neighbourhood ratio α, together with the application of equation 8 for sub-band division, results in an increased robustness for the SBSC in respect to OBSC, as it only fails for 1.6% of the songs. The robustness of the most unstable dimensions of SBSC improves significantly from 61.4% to 12.4% of failed songs when compared to OBSC (SBSC-A, 64 kb, table 3).

When the frequency spectrum is divided into 9 bands, SBSC is more unstable (SBSC-B, table 3). This can be attributed to a smaller number of bins in each sub-band and thus a smaller number of bins over which the peaks and valleys are averaged. Apart from the results shown in table 3, it is worth to mention that the delta coefficients for both SBSC and the MFCCs are very unstable with an average failure rate of 75% for 192 kb. One might hypothesize that this is due to the large effect that small changes in the audio have on the delta coefficients. Also, we find that the valley dimensions are more unstable than the contrast dimensions as these are affected most by the MP3 compression.

Test	Error rate at 192kb		Error rate at 64kb	
	mean	max	mean	max
MFCC-A	0.7	8.7	**0.4**	**4.9**
MFCC-B	**0.4**	**3.0**	0.7	6.7
MFCC-C	22.1	87.9	21.8	83.0
OBSC-A	6.5	45.3	9.9	58.6
OBSC-B	5.4	23.2	13.6	61.4
SBSC-A	1.4	4.9	1.6	12.4
SBSC-B	2.2	13.5	3.1	13.4
SBSC-C	19.7	83.7	19.7	79.9

Table 3. MP3 compression robustness error rates (%) for the descriptors tested.

5.3 MFCC and SBSC Information Overlap

The highest scoring subset of SBSC and MFCC attributes is achieved with CFS and has the same accuracy as a subset of only SBSC attributes: 67.3% for SVM (both subsets consist of 24 attributes). We also find that SBSC attributes are predominant in the first quartile of all ranked attributes and that this pattern is reversed in the last quartile, where the predominant attributes are MFCCs (table 4). Table 5 shows that for every ranking method the average SBSC rank is higher than the average MFCC rank. According to ANOVA, SBSC's attributes also rank statistically significantly higher than MFCCs' attributes.

The possible synergistic effect of the combination of both sets is addressed in figure 3, which shows the step by step increase in accuracy when adding more attributes to the selected subset for SVM classification (the order of addition is provided by the chi-square test ranking). MFCCs' accuracy quickly climbs when adding more attributes but increases only slightly after the 10th attribute. SBSC's accuracy increases at more regular intervals and is higher than MFCCs' when using 10 or more dimensions. Combining both sets and using 10 or more coefficients increases the performance above that of MFCCs, but never reaches that of the SBSC attributes. This can be taken as evidence that the description provided by SBSC subsumes and improves classification over that of MFCCs.

With all this accumulated evidence, it seems safe to conclude that the two subsets of attributes are capturing similar aspects of the sound spectra and, as their combination does not increase the classification performance beyond the level attainable by SBSC attributes alone, we should prefer them over MFCCs. However, we cannot say there is a clear overlap of information as SBSC is preferred by all attribute selection methods.

Rank method	SBSC presence in first quartile	SBSC presence in last quartile
SVM	75%	25%
Chi-square	67%	17%
info gain	75%	17%
gain ratio	75%	17%

Table 4. Percentual presence of SBSC attributes in the first and last quartile of a ranked list containing MFCC and SBSC attributes.

Rank method	F	Probability	Average MFCC rank	Average SBSC rank
SVM	5.8	0.0204	29.1	19.9
Chi-square	7.0	0.0110	29.5	19.5
info gain	8.8	0.0048	30.0	19.0
gain ratio	10.4	0.0023	30.5	18.5

Table 5. ANOVA results of attributes' rank number and origin, degrees of freedom is always 1.

6 Conclusion

In this paper, the shape-based spectral contrast descriptor is presented and evaluated. It is based on the octave-based spectral contrast, but takes the mean magnitude of a band into account in order to calculate a more descriptive measure of spectral contrast. Also, it divides the spectrum and applies KLT differently for increased robustness and applicability. SBSC is compared to both OBSC and MFCCs in terms of genre classification accuracy, to test discriminative power, and MP3 compression robustness. OBSC is shown to achieve higher accuracies than MFCCs in the genre classification task, corroborating initial findings of Jiang et al. Moreover, SBSC's outperforms OBSC and MFCCs. When it comes to MP3 compression robustness,

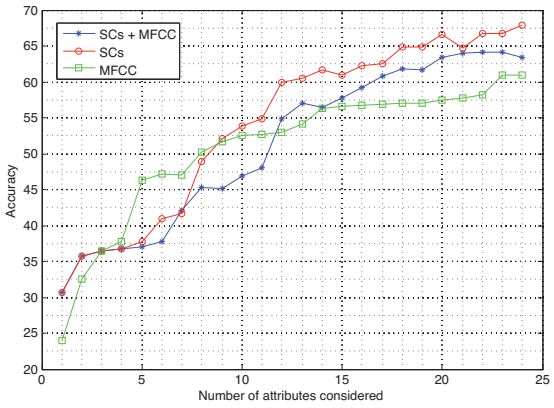

Figure 3. Increase in accuracy while considering an increasing number of attributes.

MFCCs provide the most robust option. However, SBSC represents a significant increase in robustness over OBSC. Including the delta coefficients results in higher accuracies for all descriptors but yields very unstable descriptors. Results obtained from testing information overlap between SBSC and MFCC indicate they capture similar aspects of the sound spectra. However, we found no clear evidence of overlapping information.

7 Acknowledgements

The authors want to thank their colleagues and staff at the Music Technology Group (UPF) for their support and work, specially O. Meyers and N. Wack. The authors also thank G. Tzanetakis for kindly sharing his music genre database. This research has been partially funded by the EU-IP project PHAROS [3] (IST-2006-045035).

8 References

[1] J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[2] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the royal Statistical Society*, 26(2):211–252, 1964.

[3] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. In *Proceedings of the IEEE*, volume 96, pages 668–696, April 2008.

[4] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. In *Proceedings of the IEEE*, volume 81, pages 1385–1422, 1993.

[5] J.H. Jensen, M.G. Christensen, D.P.W. Ellis, and S.H. Jensen. Quantitative analysis of a common audio similarity measure. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 17, pages 693–703, May 2009.

[6] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai. Music type classification by spectral contrast feature. *Proceedings of the IEEE International conference on Multimedia and Expo (ICME)*, 1:113–116, 2002.

[7] Brian C. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press, April 2003.

[8] Nicola Orio. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, 2006.

[9] D.D. O'Shaughnessy. Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[10] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Technischen Universität Wien, 2006.

[11] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, March 2006.

[12] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[13] S. Wegener, M. Haller, J.J. Burred, T. Sikora, S. Essid, and G. Richard. On the robustness of audio features for musical instrument classification. In *Proceedings of the European Signal Processing Conference*, 2008.

[14] K. West and S. Cox. Features and classifiers for the automatic classification of musical audio signals. In *Proceedings of the ISMIR conference*, page 531, 2004.

[15] I.H. Witten and E. Frank. *Data Mining, Practical Machine Learning Tools and Techniques, 2nd edition*. Elsevier, 2 edition, 2005.

[16] J. Yang, F. Luo, and A. Nehorai. Spectral contrast enhancement: algorithms and comparisons. *Speech Communication*, 39(1-2):33–46, 2003.

[3] http://www.pharos-audiovisual-search.eu