Stereo Panning Information for Music Information Retrieval Tasks*

GEORGE TZANETAKIS,[1] *AES Associate Member*, **LUIS GUSTAVO MARTINS,**[2] *AES Associate Member*,
(gtzan@cs.uvic.ca) (lmartins@porto.ucp.pt)

KIRK MCNALLY,[3] *AES Associate Member*, **AND RANDY JONES**[4]
(kmcnally@uvic.ca) (randy@madronalabs.com)

[1]*Department of Computer Science (also cross-listed in ECE, Music), University of Victoria, Victoria, Canada*
[2]*Portuguese Catholic University, Research Center for Science and Technology in the Arts, Porto, Portugal*
[3]*School of Music, University of Victoria, Victoria, Canada*
[4]*Madrona Labs, Seattle, WA, USA*

Recording engineers, mixers, and producers play important yet often overlooked roles in defining the sound of a particular record, artist, or group. The placement of different sound sources in space using stereo panning is an important component of their work. Stereo panning information typically is not utilized in music information retrieval (MIR) tasks such as genre and artist classification. A set of audio features is proposed that can be used to characterize stereo panning information and contrast two different methods of calculation. These features are shown to provide statistically important information for nontrivial audio classification tasks and are compared with the traditional mel-frequency cepstral coefficients for different MIR tasks. They can also be viewed as a first attempt to capture extramusical information related to the production process through music information retrieval techniques.

0 INTRODUCTION

Placing individually recorded tracks within a stereo field using panning tools is one of the main tasks of record producers and engineers. Starting in the 1960s the recording process for rock and popular music moved beyond the convention of recreating as faithfully as possible the illusion of a live performance. Facilitated by technological advances, including multitrack recording, tape editing, and equalization, the creative contributions of record producers became increasingly important in defining the sound of artists and styles [1]. Although not as well known as the artists they worked with, legendary producers including Phil Spector, George Martin, Quincy Jones, and Brian Eno changed the way music was created.

So far, research in music information retrieval [2] has largely ignored information about the recording process, focusing instead on capturing information about pitch, rhythm, and timbre [3]. A common methodology is to extract features (that is, quantifiable attributes of music signals) from recordings and then classify these features into distinct groups using machine learning techniques.

The most common audio features used for these tasks characterize the timbral texture of the audio and how it evolves over time and are based on time–frequency analysis of a monophonic audio signal. For example, some of the most common representations are the mel-frequency cepstral coefficients (MFCCs), which summarize spectral information by taking into account characteristics of the human auditory system. However, given the high impact of the feature extraction stage in the classification performance, and the difficulty involved in selecting the features to extract in order to optimize discrimination, some authors propose automatically generated sound and music descriptors. In [4] a heuristic-based generic approach for extracting automatically high-level music descriptors from acoustic signals is proposed. The approach is based on genetic programming, used to build relevant features as functions of mathematical and signal processing operators. The search of relevant features is guided by specialized heuristics, which embody knowledge about the signal processing functions built by the system. Signal processing patterns are then used in order to control the general processing methods. The MPEG-7 standard also proposes a set of audio descriptors that can be used for sound and music classification (including genre classification) [5]. Fields proposed a system using multiple-feature extraction with statistical models to categorize music by genre [6]. The system uses MPEG-7 feature vectors as well as MFCCs classified through

*Manuscript received 2009 May 19; revised 2010 January 5.

multiple trained hidden Markov models and other statistical methods. The outputs of these models are then compared, and a genre choice is made based on which model gives the best fit. A hierarchical automatic audio signal classification system, based on MPEG-7 and other audio descriptors, is presented in [7]. The audio signals are classified according to audio type, differentiating between three speech classes, 13 musical genres, and background noise.

This two-part process (namely, feature extraction followed by classification) has enabled tasks such as automatic identification of genres, albums, and artists. However, the performance of artist identification systems degrades when music from different albums is used for training and evaluation [8]. The influence of the recording process on automatic classification has been termed the "album effect." Therefore the classification results of such systems are not based entirely on the musical content. Various stages of production of the recorded artifact, including recording, mixing, and mastering, all have the potential to influence classification. This has led to research that attempts to quantify the effects of production on acoustic features. By detecting equalization curves used in album mastering, it is possible to compensate for the effects of mastering so that multiple instances of the same song on different albums can be better compared [9].

We believe that stereo mixing is an important component of understanding modern music and should be incorporated rather than being removed from music information retrieval systems. In fact, stereo information has mainly been utilized for source separation purposes [10]–[14] as well as for upmixing and ambience extraction from audio signals [15]–[17]. In previous work we have shown that it can also be used for automatic music classification [18]. In this engineering report we continue and extend that work by constrasting different ways of stereo feature extraction. We also propose a new stereo-based audio feature that roughly correlates with the number of sources in the sound mixture, and will conduct additional experiments in genre classification. In addition we show results of using stereo-based features in artist identification. Results of evaluating the proposed approach in the MIREX evaluation exchange forum are also reported.

1 STEREO PANNING INFORMATION

In this section we describe the process of calculating stereo panning information for different frequencies based on the short-time Fourier transform (STFT) of the left and right channels. Using the extracted stereo panning representations we propose features that can be used for classification.

1.1 Avendano Stereo Panning Spectrum

Avendano [10] describes a frequency-domain source identification system based on a cross-channel metric called the "panning index." We use the same metric as the basis for calculating stereo audio features for classification. For the remainder of this report the term stereo panning spectrum (SPS) is used instead of panning index as we feel it is a more accurate term. The stereo panning spectrum holds the panning values (between -1 and $+1$ with 0 being center) for each frequency bin.

The derivation of the stereo panning spectrum assumes a simplified model of the stereo signal where each sound source is recorded individually and then mixed into a single stereo signal using amplitude panning. If all instruments are recorded live and simultaneously this simplified model is not accurate as it does not take into account time delay and reverberation. However, it still provides a rough approximation of the sound mixing process. One of the advantages of using a machine learning approach is that the results of feature extraction do not need to be perfect in order for them to be useful, so—as we will show later—this simplified model is sufficient to assist in music characterization.

The basic idea behind the stereo panning spectrum is to compare the left and right signals in the time–frequency plane to derive a two-dimensional map that identifies the different panning gains associated with each time–frequency bin. By selecting time–frequency bins with similar panning values it is possible to separate particular sound sources [10]. In this study we utilize the stereo panning spectrum directly as the basis for extracting statistical features without attempting any form of source separation.

If we denote the STFT of the left and right signals $x_{\mathrm{l}}(t)$, $x_{\mathrm{r}}(t)$ by $X_{\mathrm{l}}(k)$, $X_{\mathrm{r}}(k)$, where k is the frequency index, we can define the following similarity measure:

$$\psi(k) = 2 \times \frac{|X_{\mathrm{l}}(k)X_{\mathrm{r}}^{*}(k)|}{|X_{\mathrm{l}}(k)|^{2} + |X_{\mathrm{r}}(k)|^{2}} \tag{1}$$

where $*$ denotes complex conjugation.

We assume that each sound source is amplitude panned by α. For a single source with amplitude panning the similarity function will have a value proportional to the panning coefficient α in those time–frequency regions where the source has energy. More specifically if we assume the sinusoidal energy-preserving panning law $a_{\mathrm{r}} = \sqrt{1 - a_{\mathrm{l}}^{2}}$, then

$$\psi(k) = 2\alpha\sqrt{1 - \alpha^{2}}. \tag{2}$$

If the source is panned to the center (that is, $\alpha = 0.7071$) then the function will attain its maximum value of 1, and if the source is completely panned to either side the function will attain its minimum value of zero. The ambiguity with regard to the latter direction of the source can be resolved using the partial similarity measures

$$\psi_{\mathrm{l}} = \frac{|X_{\mathrm{l}}(k)X_{\mathrm{r}}^{*}(k)|}{|X_{\mathrm{l}}(k)|^{2}}, \qquad \psi_{\mathrm{r}} = \frac{|X_{\mathrm{r}}(k)X_{\mathrm{l}}^{*}(k)|}{|X_{\mathrm{r}}(k)|^{2}} \tag{3}$$

and their difference,

$$\Delta(k) = \psi_{\mathrm{l}} - \psi_{\mathrm{r}} \tag{4}$$

where positive values of $\Delta(k)$ correspond to signals panned toward the left and negative values correspond to signals panned to the right. Thus we can define the following ambiguity-resolving function:

$$\hat{\Delta}(k) = \begin{cases} +1, & \text{if } \Delta(k) > 0 \\ 0, & \text{if } \Delta(k) = 0 \\ -1, & \text{if } \Delta(k) < 0. \end{cases} \tag{5}$$

Shifting and multiplying the similarity function by $\hat{\Delta}(k)$ we obtain the stereo panning spectrum (or panning index) as

$$\text{SPS}(k) = [1 - \psi(k)] \times \hat{\Delta}(k). \tag{6}$$

Fig. 1 shows a visualization of the stereo panning spectrum for the song "Hell's Bells" by ACDC. The visualization is similar to a spectrogram with the *X* axis corresponding to time, measured in number of analysis frames, and the *Y* axis corresponding to frequency bin. No panning is represented by gray, full left panning is represented by black, and full right panning by white. The song starts with four bell sounds that alternate between slight panning to the left and to the right, visible as changes in grey intensity. Near the end of the first 28 seconds a strong electric guitar enters on the right channel, visible as white.

Fig. 2 shows a visualization of the stereo panning spectrum for the song "Supervixen" by Garbage. Several interesting stereo manipulations can be observed in the figure and heard when listening to the song. The song starts with all instruments centered for a brief period and then moves them to the left and right, creating an explosion-like effect. Most of the sound of a fast repetitive hi-hat is panned to the right (the wide dark bar over the narrow horizontal white bar) with a small part of it panned to the left (the narrow horizontal white bar). Near the end of the first 28 seconds the voice enters with a crash cymbal panned to the left, visible as the large black area.

1.2 Azimuth Discrimination and Resynthesis — ADRess

The ADRess algorithm [11] takes advantage of the different mixing parameters used to place each source at discrete locations across the stereo field.[1] Given these mixing parameters, gain scaling will cause one source intensity level to become equal in both the left and right channels, and therefore simple subtraction will cause that particular source to cancel itself out due to phase cancellation. These mixing parameters can be used to isolate a single source within a mixture by implementing appropriate gain scaling and channel substraction.

However, since the true mixing parameters are usually unknown when taking a stereo audio track, the ADRess implementation starts by creating a frequency-azimuth plane where all the possible mixing parameters are used to construct a full representation of the stereo field. At each discrete azimuth position on this plane, gain scaling and short-time Fourier transform spectral subtraction are applied, causing local nulls to appear at the points where phase cancellation occurs. These local nulls indicate the presence of frequencies associated with a source at the corresponding azimuth position, allowing to estimate an azimuth or panning position index for each spectral component of the stereo signal. A detailed description of the ADRess algorithm can be found in [11], and an improved implementation was presented in [19].

1.3 Stereo Panning Spectrum Features

In this section we describe a set of features that summarize the information contained in the stereo panning spectrum, which can be used for audio classification tasks. The main idea is to capture the amount and distribution

[1] In the context of ADRess, the term azimuth is used to refer to the position in the stereo field.

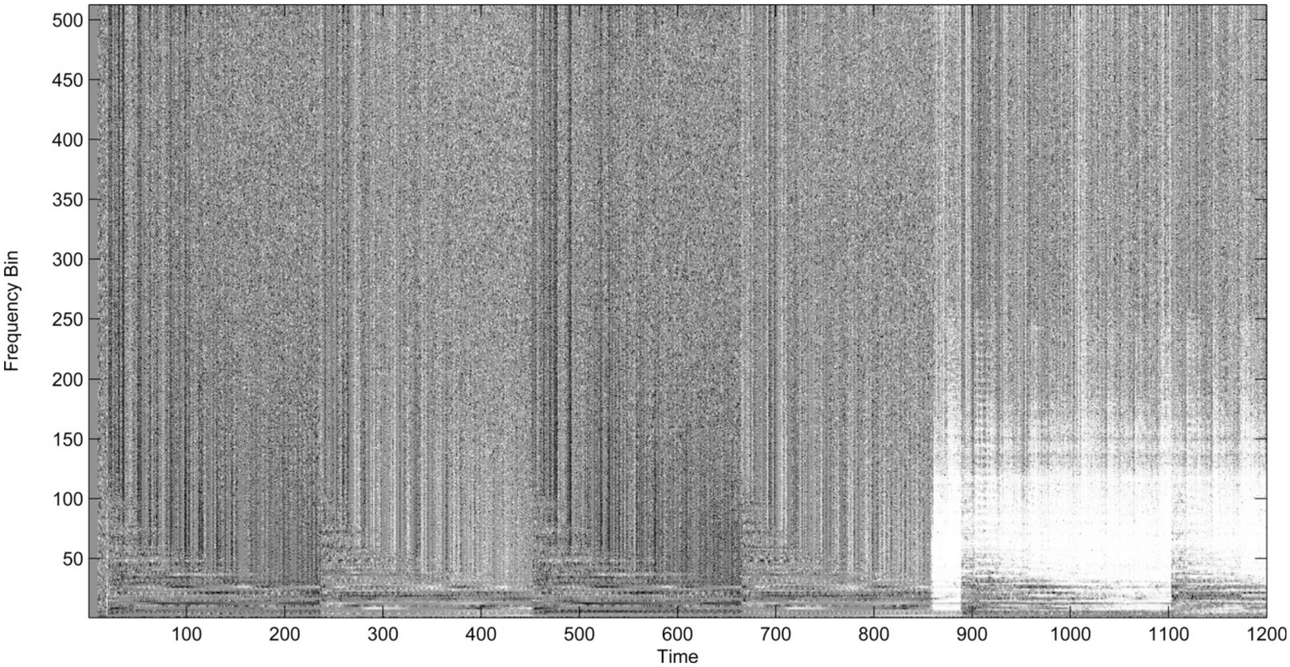


Fig. 1. Stereo panning spectrum of "Hell's Bells" by ACDC (approximately 28 seconds).

of panning in different frequency bands as well as its dynamic evolution over time. We define the panning root mean square for a particular frequency band as

$$P_{\mathrm{l,h}} = \sqrt{\frac{1}{h-l}\sum_{k=l}^{h}[\mathrm{SPS}(k)]^2} \tag{7}$$

where l is the lower frequency and h is the highest frequency of the band. In addition to the features proposed in [18] we also consider a feature that roughly correlates with the number of sources that have been stereo panned in distinct ways. If we sort the stereo panning spectrum values, we expect to find a stair-shaped curve, where each step corresponds to the components at the same azimuth. As a simple estimate of the number of sources we count the number of peaks of the first derivative of the sorted signal. We then consider the following five-dimensional feature vector corresponding to an analysis window t:

$$\Phi(t) = [P_{\mathrm{total}}(t), P_{\mathrm{low}}(t), P_{\mathrm{medium}}(t), P_{\mathrm{high}}(t), S(t)]. \tag{8}$$

The P values correspond to overall panning (0–22 050 Hz) and to panning for low (0–250 Hz), medium (250–2500 Hz), and high frequencies (2500–22 050 Hz). The choice of three frequency bands was motivated by the common informal division into low, medium, and high frequencies utilized by recording engineers when discussing panning, and the corresponding ranges were chosen to be consistent with common perception of these bands. S is the estimated number of sources in the stereo field. To capture the dynamics of panning information we compute a running mean and standard deviation over the past M frames,

$$m\Phi(t) = \mathrm{mean}[\Phi(t-M+1), \ldots, \Phi(t)] \tag{9}$$

$$s\Phi(t) = \mathrm{std}[\Phi(t-M+1), \ldots, \Phi(t)]. \tag{10}$$

This results in a ten-dimensional feature vector at the same rate as the original five-dimensional feature vector. For the experiments described in the next section M is set to 40, corresponding to approximately 0.5 second. To avoid any duration effects on classification we only consider approximately the first 30 seconds of each track, which results in a sequence of 1000 ten-dimensional feature vectors for each track. The majority of existing data sets for MIR consists of 30-second clips frequently due to processing time and copyright restrictions. The tracks are stereo, 16-bit, 44 100-Hz sampling rate audio files and the STFT window size is set to 1024 samples. The sequence of feature vectors is collapsed to a single feature vector representing the entire track by taking again the mean and the standard deviation across the first 30 seconds, which results in the final twenty-dimensional feature vector for each track.

2 EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of the proposed features we considered two nontrivial tasks for which we thought stereo information would be useful. As a side note, using the proposed features it is trivial (although quite useful) to detect mono recordings that have been directly converted to stereo without remastering.

The first classification task is distinguishing two collections of rock music, one from the 1960s and another one from the 1990s. In genre terms these can be loosely categorized as “garage” and “grunge.” Both these styles would be classified into the top-level genre of rock. To isolate the effects of recording production, we only included albums that had as their main instrumentation the standard rock ensemble of electric guitar, electric bass, drums, and vocals. Albums with an excess of keyboards or experimental studio

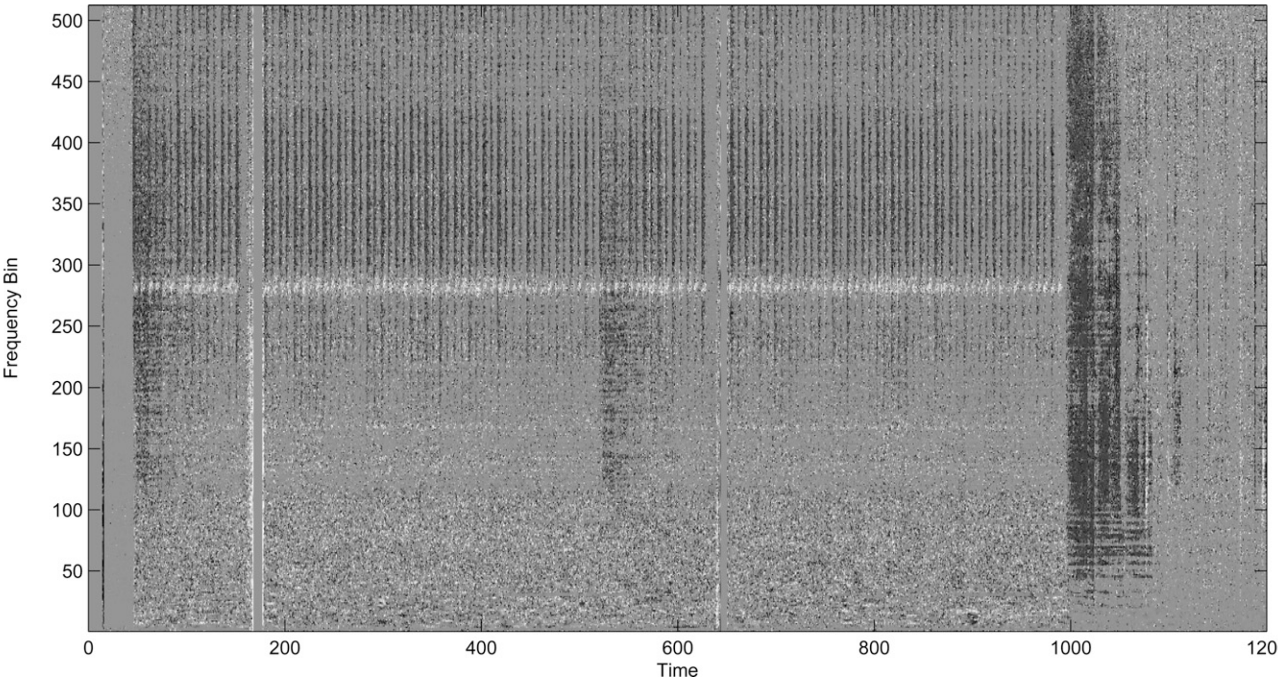


Fig. 2. Stereo panning spectrum of “Supervixen” by Garbage (approximately 28 seconds).

techniques, late 1960s Beatles, for example, were excluded. We used 227 tracks from the 1960s and 176 tracks from the 1990s. Example "garage" groups include The Byrds, The Kinks, and Buddy Holly. Example "grunge" groups include Nirvana, Pearl Jam, and Radiohead.

The second classification task we consider is distinguishing electric jazz from acoustic jazz. Both styles would be classified into the top-level genre of jazz. Acoustic jazz tends to have relatively pronounced panning of the solo instruments (saxophone and trumpet), which does not vary over time. We used 175 electric jazz tracks and 184 acoustic jazz tracks. Example electric jazz groups include Weather Report, Return to Forever, Medeski, Martin and Wood, and Mahavishnu Orchestra. Example acoustic jazz groups led by artists include Miles Davis, John Coltrane, Lee Morgan, and Branford Marsalis.

Tables 1 and 2 show the classification accuracy results for the stereo panning spectrum features and compare them with the results obtained from stereo mel-frequency cepstrum coefficients (MFCCs) (basically the MFCC of the left and right channels concatenated) as well as their combination for the two tasks. MFCCs are the most common feature front end for evaluating timbral similarity [20]. We believe that the experimental results would be similar if other timbral feature front ends such as MPEG-7 audio descriptors were utilized [5]. The accuracies are in percentages and are computed using stratified ten-fold cross validation. The ZeroR classifier is a simple baseline, NBS corresponds to a simple Naive Bayes classifier, SMO corresponds to a linear Support Vector Machine trained with sequential minimal optimization with $\varepsilon = 10^{-12}$ and $\gamma = 0.01$, and J48 is a decision tree with a confidence factor of 0.25. More information about these representative classifiers can be found in [21] or any pattern recognition textbook. As can be seen the stereo panning spectrum features (SPSFs) perform well and for the acoustic versus electric jazz task achieve almost perfect classification. As a side note the classification accuracy of mono MFCC was almost identical to that of stereo MFCC and therefore the forever was not included in the tables. The performance of the Avendano (AVESPSF) and ADDress (ADRSPSF) methods of calculation is also compared. Avendano performs better and is also significantly more efficient to compute, and therefore it will be used for the remaining experimental results. For the remainder of this section all the classification results are based on a linear support vector machine classifier with $\varepsilon = 10^{-12}$ and $\gamma = 0.01$.

Table 1. Classification accuracies for garage/grunge.

Garage/Grunge	ZeroR	NBC	SMO	J48
AVESPSF	57.9	80.4	80.7	84.1
ADRSPSF	57.9	78.2	76.8	82.1
SMFCC	57.9	76.5	77.7	71.3
AVESPSF/SMFCC	57.9	84.0	85.2	85.2

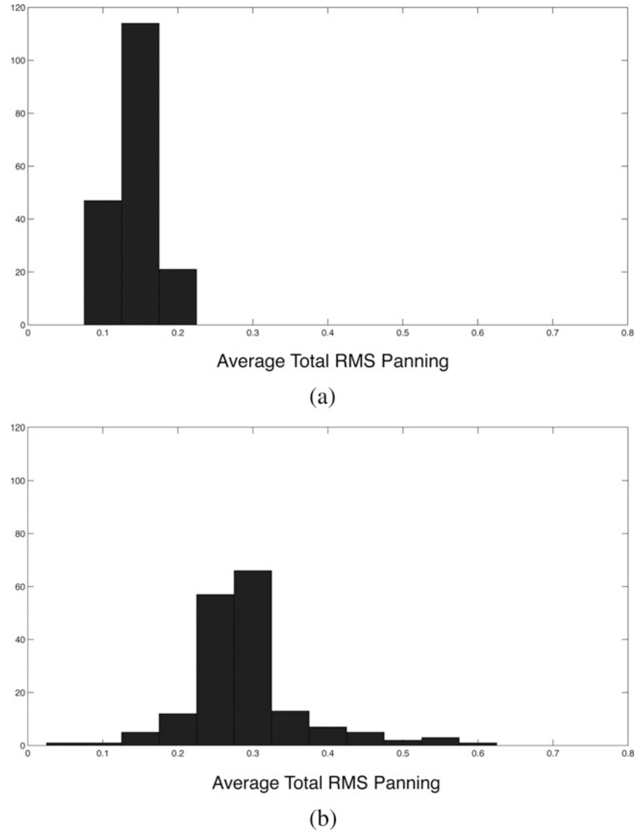
Table 2. Classification accuracies for acoustic/electric jazz.

Acoustic/Electric	ZeroR	NBC	SMO	J48
AVESPSF	51.3	99.4	99.7	99.1
ADRSPSF	51.3	97.7	98.3	99.7
SMFCC	51.3	71.8	79.4	68.4
AVESPSF/SMFCC	51.3	98.5	99.1	99.1

It is important to note that the proposed features only capture stereo information and are not influenced by any spectral content or amplitude dynamics. For example, applying any amplitude change to both channels does not change the SPSF values. Therefore the spectrum could be completely altered without changing the features as long as the changes are proportional to the panning coefficients.

As a way to interepret the performance of the proposed features Fig. 3 shows the histograms of a single feature: the mean total rms panning for acoustic jazz and for electric jazz. As can be seen, acoustic jazz has lower but more consistent panning values whereas electric jazz has more pronounced and spread panning values.

Table 3 shows the classification results for all four styles combined. Although somewhat artificial as a task, this provides information about the robustness of the proposed features as well as the value of combining the



(a)

(b)

Fig. 3. Histogram of mean overall panning. (a) Acoustic jazz. (b) Electric jazz.

standard MFCC features with the proposed stereo panning spectrum features.

Encouraged by the results of these experiments we decided to explore the use of stereo features for the more general and difficult task of artist identification. For that we used the artist20 database, compiled by Dan Ellis and his group.[2] The artist20 is a database of six albums by each of 20 artists, making a total of 1413 tracks. Each artist is represented by six regular studio albums. A six-fold jacknife train/test scheme, where each fold consists of training on five albums per artist, and testing on the remaining one was used for the experiments presented in Table 4. It is also shown that the new feature, which estimates the number of sources, improves performance compared to the features proposed in [18] (the numbers after the slashes).

The annual music information retrieval evaluation exchange (MIREX)[3] is a forum for comparing different algorithms on a variety of tasks. Even though we did not manage to include stereo information features in our submission to MIREX 2007, the organizers allowed us to perform experiments using the approach described in this study after the completion date. The results on three audio classification tasks are summarized in Table 5 and Fig. 4 and were calculated by the MIREX organizers as the audio collections used were not available. The artist identification task is based on 3150 clips from 105 artists, the genre classification task consists of 7000 clips drawn from 10 genres, and the classical composer identification task consists of 2772 clips organized into 11 "composers." More details can be found at the MIREX Web page. The results are very encouraging, showing that stereo information is useful even in the unexpected task of classical composer identification. After discussing it with the MIREX organizers who had access to the data we realized that it was due to the fact that all the recordings of each composer come from the same group/set of recordings.

In MIREX 2008 we were prepared and submitted both a system using traditional features on the mono signal as well as one that also included the stereo panning features. Table 6 and Fig. 5 compare the raw classification accuracy results between the mono and stereo configurations. The accuracy is averaged over three train/test folds. Album filtering is used for the test and training splits, that is, training and test sets contain tracks from different albums. The full results, including those from

Table 3. Classification accuracies for four styles.

Gr/Ga/Aj/Ej	ZeroR	NBC	SMO	J48
SPSF	29.8	73.6	81	76.5
SMFCC	29.8	56.4	65.9	52.3
SPSF+SMFCC	29.8	75.2	87.4	79.9

Table 4. Classification accuracies for artist identification in the artist20 dataset using stereo information.

Artist20Id	ZeroR	SPSF	SMFCC	SPSF/SMFCC
fold0	6.7	20.7/19.4	32.2	34.3/31.9
fold1	5.9	21.4/16.9	31.05	30.6/28.7
fold2	4.6	25.9/20.0	35.1	43.0/28.7
fold3	5.2	14.4/13.9	37.5	36.2/36.6
fold4	5.1	21.6/20.4	35.4	37.2/36.6
fold5	5.0	12.9/12.5	17.6	18.8/18.8
Average	5.4	19.5/17.1	31.4	33.35/32.6

Table 5. Overall classification accuracy for MIREX 2007 tasks (artist identification, genre classification, classical composer identification) Using stereo information.

	Artist	Genre	Composer
SPSF	14.22	41.41	26.19
SMFCC	36.50	61.79	45.13
SPSF/SMFCC	45.16	64.13	49.89
Best MIREX07*	48.14	68.29	53.72

*Highest task classification accuracy achieved in 2007.

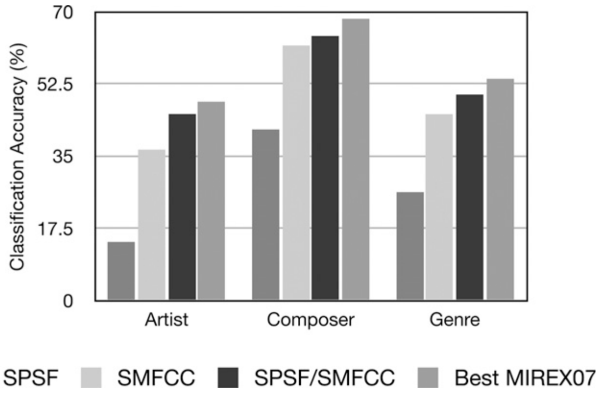


Fig. 4. Overall classification accuracy for MIREX 2007 tasks (artist identification, classical composer identification and genre classification) using stereo information (see Table 5). Best MIREX07 is highest task classification accuracy achieved in 2007.

Table 6. Overall classification accuracy for MIREX 2008 tasks (artist identification, classical composer identification, genre classification, latin genre classification, mood classification) Using stereo information.

	Artist	Composer	Genre	Latin	Mood
Mono-Feat	35.27	43.81	65.62	53.67	55
Stereo-Feat	43.47	45.82	66.41	53.79	52.5
Best MIREX08*	47.65	53.25	66.41	65.17	63.67

*Highest task classification accuracy achieved in 2008.

[2] http://labrosa.ee.columbia.edu/projects/artistid/

[3] http://www.music-ir.org/mirexwiki/index.php/Main_Page

all the other submissions to the conference, are available on the MIREX Web site. There are several interesting observations. Our system that included the stereo panning features achieved the best classification accuracy for genre classification. The biggest improvement due to the addition of the stereo panning features was in artist identification, and the performance of mood classification actually was made worse in the stereo case (see Table 6 and Fig. 5). Although we do not have access to the data used, our conjecture is that moods span diverse artists and genres and therefore do not have a consistent profile related to their record production. By analyzing the results it is also possible to find for which artists the stereo panning information did provide the most improvement in performance. Table 7 shows artists for which the inclusion of features based on stereo panning improved classification accuracy by more than 20%. Most of these artists are early jazz pioneers. Although timbrally and rhythmically they sound similar to each other, the way they were placed in the recording sessions was unique enough to identify them in a rough statistical sense.

Researchers interested in replicating these experiments can obtain the complete lists of tracks and albums for both of these tasks by contacting the authors via email. The code for the calculation of the stereo panning spectrum features has been integrated into Marsyas,[4] an open source framework for audio processing with specific emphasis on music information retrieval. The machine learning part of the experiments was conducted using Weka[5] [21].

3 CONCLUSIONS AND FUTURE WORK

A new feature set based on the stereo panning spectrum has been proposed and is shown to be effective for a variety of nontrivial audio classification tasks. It has been argued that the approach of modeling timbral similarity using MFCC has reached a "glass ceiling" [20]. We believe that information related to the recording process such as the stereo panning information used in this engineering report can help future audio MIR systems escape this ceiling. More detailed features related to stereo information than the ones proposed here can be envisioned, such as features capturing the amount of reverberation applied to different tracks in the original mix.

Another interesting direction for future work is to explore the characterization of individual sound sources in the stereo panning spectrum. By clustering bins with similar panning indexes it is possible to identify candidate sound sources [10], [11] and then individually characterize their panning and spectral characteristics for the purpose of genre classification and artist identification.

We are also interested in exploring other aspects of the studio production process for MIR purposes. Examples include equalization, compression, and effects including reverberation and delay. We are collaborating with professional studio recording engineers who also teach recording. We are planning to develop visualization and editing tools that can help reverse-engineer the stero mixing of audio recordings for pedagogical purposes. Another interesting direction is to perform automatic stereo panning based on audio feature extraction.

Engineers communicate about mixing with a particular lexicon of qualitative terms. A good example comes from an interview with *Mix Magazine*, where Dave Pensado

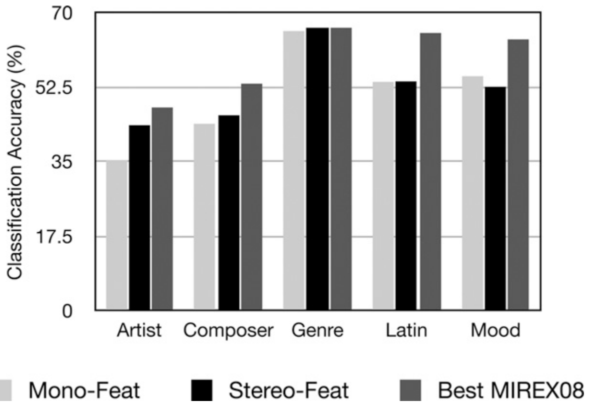


Fig. 5. Overall classification accuracy for MIREX 2008 tasks (artist identification, classical composer identification, genre classification, latin genre classification, mood classification) using stereo information (see Table 6). Best MIREX08 is highest task classification accuracy achieved in 2008.

Table 7. Classification for individual artists sorted by largest improvement in classification accuracy using the features based on stereo panning information (MIREX 2008).

Artist	Stereo-Mono	Mono Feature	Stereo Feature
Chick Webb	0.50	0.27	0.77
Chick Webb	0.50	0.20	0.70
Louis Armstrong	0.37	0.00	0.37
Fat Swaller	0.37	0.30	0.67
Benny Carter	0.37	0.30	0.67
Jimmie Noone	0.33	0.40	0.73
Fletcher Henderson	0.33	0.10	0.43
Nat King Cole	0.30	0.33	0.63
Artie Shaw	0.30	0.40	0.70
Woody Herman	0.27	0.03	0.33
George Jones	0.23	0.53	0.77
Benny Goodman	0.23	0.13	0.37
Billie Holiday	0.23	0.40	0.63
Benny Moten	0.23	0.47	0.70
Tom Jones	0.20	0.77	0.97
Roxette	0.20	0.17	0.37

[4] http://marsyas.sourceforge.net

[5] http://www.cs.waikato.ac.nz/ml/weka/

describes one of his mixes as having "massive club bottom, hip hop sensibility in the middle, and this real smoothed-out, classy, Quincy Jones-type top" [22]. Our hope is to eventually be able to translate this type of discussion into a more quantitative domain.

4 ACKNOWLEDGMENT

The authors would like to thank the National Sciences and Engineering Research Council (NSERC) and the Social Sciences and Humanities Research Council (SSHRC) of Canada, and the Science and Inovation 2010 Operational Program (POCI 2010), cofunded by the Portuguese Government and the European Union (through the European Fund for the Regional Development (FEDER) and the Portuguese Science and Technology Foundation (FCT)), for funding this work. The authors would also like to acknowledge Perry Cook for suggesting the idea of using stereo a long time ago, and C. Avendano and D. Barry for describing their methods with sufficient clarity and detail to be reimplemented. Randy Jones and Kirk McNally helped with the initial implementation and design of this work.

5 REFERENCES

[1] V. Moorefield, *The Producer as Composer* (MIT Press, Cambridge, MA, 2005).

[2] F. Rumsey, "Searching, Analyzing, and Recommending Audio Content," *J. Audio Eng. Soc. (Features)*, vol. 57, pp. 166–169 (2009 Mar.).

[3] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 293–302 (2002).

[4] A. Zils and F. Pachet, "Automatic Extraction of High-Level Music Descriptors from Acoustic Signals Using EDS," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 52, p. 814 (2004 July/Aug.), convention paper 6127.

[5] P. Herrera, X. Serra, and G. Peeters, "Audio Descriptors and Descriptors Schemes in the Context of MPEG-7," in *Proc. Int. Computer Music Conf.* (ICMC, 1999).

[6] B. Fields, "Using Multiple Feature Extraction with Statistical Models to Categorize Music by Genre," presented at the 122nd Convention of the Audio Engineering Society, (Abstracts) www.aes.org/events/122/122ndWrapUp.pdf, (2007 May), convention paper 7015.

[7] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *J. Audio Eng. Soc.*, vol. 52, pp. 724–739 (2004 July/Aug.).

[8] B. Whitman, G. Flake, and S. Lawrence, "Artist Detection in Music with Minnowmatch," in *Proc. IEEE Workshop on Neural Networks for Signal Processing* (2001), pp. 559–568.

[9] Y. E. Kim, D. S. Williamson, and S. Pilli, "Towards Quantifying the 'Album Effect' in Artist Identification," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)* (2006), pp. 393–394.

[10] C. Avendano, "Frequency-Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression and Re-panning Applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2003), pp. 55–58.

[11] D. Barry and B. Lawlor, "Sound Source Separation: Azimuth Discrimation and Resynthesis," in *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04)* (Naples, Italy, 2004).

[12] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing Stereo Music with Score-Informed Source Separation," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)* (2006).

[13] M. Vinyes, J. Bonada, and A. Loscos, "Demixing Commercial Music Productions via Human-Assisted Time–Frequency Masking," presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 54, p. 680 (2006 July/Aug.), convention paper 6719.

[14] M. Cobos and J. J. Lopez, "Resynthesis of Sound Scenes on Wave-Field Synthesis from Stereo Mixtures Using Sound Source Separation Algorithms," *J. Audio Eng. Soc.*, vol. 57, pp. 91–110 (2009 Mar.).

[15] C. Avendano and J. M. Jot, "Ambience Extraction and Synthesis from Stereo Signals for Multi-Channel Audio Up-Mix," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2 (2002), pp. 1957–1960.

[16] C. Avendano and J. M. Jot, "Frequency Domain Techniques for Stereo to Multichannel Upmix," in *Proc. AES 22nd Conf. on Vitual, Synthetic, and Entertainment Audio* (Espoo, Finland, 2002 June 15–17), pp. 121–130.

[17] J. Merimaa, M. Goodwin, and J. M. Jot, "Correlation-Based Ambience Extraction from Stereo Recordings," presented at the 123rd Convention of the Audio Engineering Society, (Abstracts) www.aes.org/events/123/123rdWrapUp.pdf, (2007 Oct.), convention paper 7282.

[18] G. Tzanetakis, R. Jones, and K. McNally, "Stereo Panning Features for Classifying Recording Production Style," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)* (2007).

[19] R. Cooney, N. Cahill, and R. Lawlor, "An Enhanced Implementation of ADR (Azimuth Discrimination and Resynthesis) Music Source Separation Algorithm," presented at the 121st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 54, pp. 1287, 1288 (2006 Dec.), convention paper 6984.

[20] J. J. Aucouturier and F. Pachet, "Improving Timbre Similarity: How High Is the Sky?," *J. Negative Results in Speech Audio Sci.*, vol. 1, no. 1 (2004).

[21] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. (Morgan Kaufmann, San Francisco, CA, 2005).

[22] D. "Hard Drive" Pensado, "Interview," *Mix Mag.* (2001 Sept.).

THE AUTHORS

G. Tzanetakis

L. G. Martins

K. McNally

R. Jones

George Tzanetakis received a Ph.D. degree in computer science from Princeton University, Princeton, NJ, in 2002 and was a postdoctoral fellow at Carnegie Mellon University in 2002–2003. He is currently an assistant professor in the Department of Computer Science with crosslisted appointments in ECE and Music at the University of Victoria, Victoria, Canada.

His research spans all stages of audio content analysis, such as feature extraction, segmentation, and classification with specific emphasis on music information retrieval. He is also the primary designer and developer of Marsyas, an open source framework for audio processing with specific emphasis on music information retrieval applications. His pioneering work in musical genre classification received an IEEE Signal Processing Society young author award and is frequently cited. More recently he has been exploring new interfaces for musical expression, music robotics, computational ethnomusicology, and computer-assisted music instrument tutoring. These interdisciplinary activities combine ideas from signal processing, perception, machine learning, sensors, actuators and human–computer interaction with the connecting theme of making computers better understand music to create more effective interactions with musicians and listeners.

●

Luís Gustavo Martins received a Ph.D. degree in electrical and computer engineering from the University of Porto, Portugal, in 2009, with a thesis on the topic of sound segregation in music signals. He is currently a professor in the Sound and Image Department of the School of the Arts, Portuguese Catholic University, and a researcher at the Research Center for Science and Technology in the Arts (CITAR), Porto.

His research is mainly focused on audio content analysis, sound processing, and synthesis. His research interests include signal processing, machine learning, perception and cognition, and software development. More recently he has been exploring the use of tangible and multitouch interfaces for sound and music exploration and interaction. He is an active developer of the open source Marsyas audio processing software framework.

●

Kirk McNally studied music and sound recording at McGill University, Montreal, Canada, and received a degree in B.Mus 1998 and an M.Mus degree in 2000. His teachers included Gerald Danovitch, Wieslaw Wosczcyk, George Massenburg, and Steven Epstein. As a recording engineer he has worked with the Boston Symphony Orchestra in Tanglewood, Reaction Studios in Toronto, The Warehouse Studios and Amoury Studios, both in Vancouver. His recording credits include R.E.M, Bryan Adams, The National Youth Orchestra of Canada, The Aventa Ensemble, live sound and recording work for CBC radio. He is the recording engineer/audio specialist for the School of Music at the University of Victoria, Victoria, Canada, where he teaches as part of the joint major program in music and computer science.

●

Randy Jones received a B.S.C.S. degree from the University of Wisconsin–Madison in 1992. In 2002 he was a cocreator of Cycling 74's Jitter software. In 2008 he received an M.S.C.S. degree from the University of Victoria, Victoria, Canada, under George Tzanetakis. He currently lives in Seattle, WA, where he is starting up Madrona Labs LLC to develop new computer-based musical instruments.