STEREO PANNING FEATURES FOR CLASSIFYING RECORDING PRODUCTION STYLE

George Tzanetakis, Randy Jones, and Kirk McNally
University of Victoria
Computer Science
gtzan@cs.uvic.ca, rj@csc.uvic.ca, kmcnally@uvic.ca

ABSTRACT

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 information is an important component of the production process. Audio classification systems typically convert stereo signals to mono and to the best of our knowledge have not utilized information related to stereo panning. In this paper we propose a set of audio features that can be used to capture stereo information. These features are shown to provide statistically important information for a non-trivial audio classification task and are compared with the traditional Mel-Frequency Cepstral Coefficients. The proposed features can be viewed as a first attempt to capture extra-musical information related to the production process through music information retrieval techniques.

1 INTRODUCTION

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 multi-track recording, tape editing, equalization and compression, the creative contributions of record producers became increasingly important in defining the sound of artists, groups, and styles [4]. Although not as well known as the artists they worked with, legendary producers including Phil Spector, George Martin, Brian Eno and Quincy Jones changed the way music was created.

So far, research in music information retrieval has largely ignored information about the recording process, focusing instead on capturing information about pitch, rhythm and timbre. A common methodology is to extract features, quantifiable attributes of music signals, from recordings, then to classify these features into distinct groups using machine learning techniques. This two-part process has enabled tasks such as automatic identification of genres, albums and artists.

The influence of the recording process on automatic classification has been acknowledged and termed the *album effect.* The performance of artist identification systems degrades when music from different albums is used for training and evaluation [7]. 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 which 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 [3]. We believe that other information related to the recording process, specifically mixing, is an important component of understanding modern pop and rock music and should be incorporated rather than being removed from music information retrieval systems.

The goal of this paper is to explore stereo panning information as an aspect of the recording and production process. Stereo information has been frequently utilized for source separation purposes [2, 9]. However, to the best of our knowledge, it has not been used in music classification systems for audio signals. In this paper we show that stereo panning information is indeed useful for automatic music classification.

2 STEREO PANNING INFORMATION EXTRACTION

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 Spectrum we propose features that can be used for classification.

2.1 Stereo Panning Spectrum

Avendano [2] 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 the paper the term *Stereo Panning Spectrum (SPS)* is used instead of the *panning index* as we feel it is a more accurate term. The SPS holds the panning values (between -1 and +1 with 0 being center) for each

frequency bin.

The derivation of the SPS assumes a simplified model of the stereo signal. In this model each sound source is recorded individually and then mixed into a single stereo signal by amplitude panning. Stereo reverberation is then added artificially to the mix. The basic idea behind the SPS 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 [2]. In this paper we utilize the SPS directly as the basis for extracting statistical features without attempting any form of source separation. Our SPS definition directly follows Avendano [2].

If we denote the STFT of the left,right signals $x_l(t), x_r(t)$ for a particular analysis window as $X_l(k), X_r(k)$, where k is the frequency index we can define the following similarity measure:

$$\psi(k) = 2 * \frac{|X_l(k)X_r^*(k)|}{|X_l(k)|^2 + |X_r(k)|^2} \quad (1)$$

where $*$ denotes complex conjugation. For a single amplitude panned source this 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: $a_r = \sqrt{1 - a_l^2}$ then:

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

If the source is panned to the center (i.e $\alpha = 0.7071$) 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 regards to the later direction of the source can be resolved using the partial similarity measures:

$$\psi_l = \frac{|X_l(k)X_r^*(k)|}{|X_l(k)|^2}, \psi_r = \frac{|X_r(k)X_l^*(k)|}{|X_r(k)|^2} \quad (3)$$

and their difference:

$$\Delta(k) = \psi_l - \psi_r \quad (4)$$

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

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

Shifting and multiplying the similarity function by $\hat{\Delta}(k)$ we obtain the Stereo Panning Spectrum (or panning index) as:

$$SPS(k) = [1 - \psi(k)] * \hat{\Delta}(k) \quad (6)$$

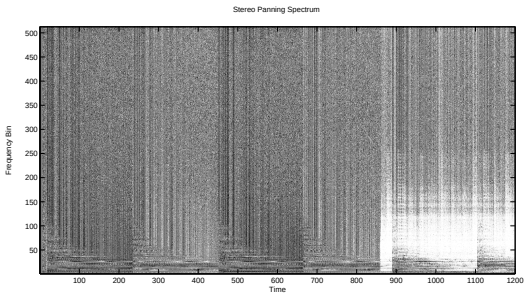

Figure 1. *Stereo Panning Spectrum of "Hell's Bells" by ACDC (approximately 28 seconds).*

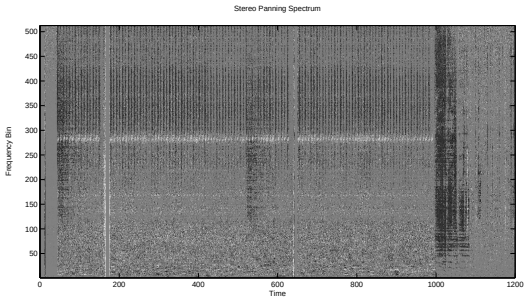

Figure 2. *Stereo Panning Spectrum of "Supervixen" by Garbage (approximately 28 seconds)*

Figure 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 by black and full right panning by white. The songs 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. Figure 2 shows a visualization of the SPS 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 the a crash cymbal panned to the left, visible as the large black area.

2.2 Stereo Panning Spectrum Features

In this section we describe a set of features that summarize the information contained in the Stereo Panning Spectrum. The main idea is to capture the amount of panning in different frequency bands as well as how it changes over time. We define the Panning Root Mean Square for a par-

ticular frequency band as:

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

where l is the lower frequency of the band, h is the high frequency of the band, and N is the number of frequency bins. By using RMS we only consider the amount of panning without taking into account whether it is to the left or right. We consider the following 4-dimensional feature vector corresponding to an analysis window t:

$$\Phi(t) = [P_{total}(t), P_{low}(t), P_{medium}(t), P_{high}(t)] \quad (8)$$

The PRMS values correspond to overall panning (0–22050 Hz), and panning for low (0–250 Hz), medium (250–2500 Hz) and high frequencies (2500–22050 Hz) respectively.

To capture the dynamics of panning information we compute a running mean and standard deviation over the past M frames:

$$m\Phi(t) = mean[\Phi(t-M+1), .., \Phi(t)] \quad (9)$$
$$s\Phi(t) = std[\Phi(t-M+1), .., \Phi(t)] \quad (10)$$

This results in a 8-dimensional feature vector at the same rate as the original 4-dimensional feature vector. For the experiments described in the next section M is set to 40 corresponding to approximately 0.5 seconds. To avoid any duration effects on classification we only consider approximately the first 30 seconds of each track, resulting in a sequence of 1000 8-dimensional feature vectors for each track. The tracks are stereo, 16-bit, 44100 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 standard deviation across the first 30 seconds resulting in the final 16-dimensional feature vector for each track.

3 EXPERIMENTS

In order to evaluate the effectiveness of the proposed features we considered two non-trivial tasks. As a sidenote, using the proposed features it is trivial (although quite useful) to detect mono recordings directly converted to stereo without remastering.

The first classification task we consider is distinguishing two collections of rock music, one from the 1960s and another from the 1990s. In genre terms, these can be loosely categorized as 'garage' and 'grunge.' Both of these styles would be classified to the top-level genre of rock. To isolate the effects of recording production, we only included albums which 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 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.

Garage/Grunge	ZeroR	NBC	SMO	J48
SPSF	56.4	77.2	81	84.2
SMFCC	56.4	74. 6	76.7	71.6
SPSF+SMFCC	56.4	82.7	83.7	83.2

Table 1. Classification accuracies for Garage/Grunge

Acoustic/Electric	ZeroR	NBC	SMO	J48
SPSF	51.3	99.4	99.7	99.1
SMFCC	51.3	71.8	79.4	68.4
SPSF+SMFCC	51.3	98.5	99.1	99.1

Table 2. Classification accuracies for Acoustic/Electric Jazz

The second classification task we consider is distinguishing electric jazz from acoustic jazz. Both of these styles would be classified to the top-level genre of jazz. Acoustic jazz tends to have relatively pronounced panning of the solo instruments (saxophone and trumpet) that doesn't 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, 2 show the classification accuracy results for the Stereo Panning Spectrum Features and compares them with the results obtained from stereo Mel-Frequency Cepstrum Coefficients (MFCC) (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 [1]. The accuracies are in percentages and are computed using stratified 10-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 and J48 is a decision tree. More information about these representative classifiers can be found in [8] or any pattern recognition textbook. As can be seen the Stereo Panning Spectrum Features (SPSF) perform well and for the acoustic vs electric jazz task achieve almost perfect classification. As a sidenote the classification accuracy of mono MFCC were almost identical to the stereo MFCC therefore were not included in the Tables.

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 changes to both channels doesn't change their values and the spectrum could be completely altered without changing the features as long as the changes are proportional to the panning coefficients.

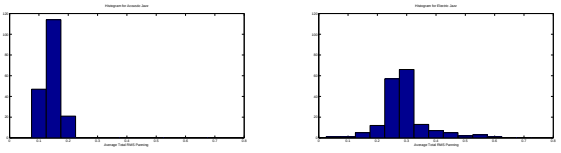

Figure 3. *Histogram of mean overall panning for Acoustic Jazz (Left) and Electric Jazz (Right)*

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 3. Classification accuracies for four styles

Figure 3 shows the histograms of a single feature: the mean total RMS panning for acoustic jazz (left) and electric jazz (right). 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 standard MFCC features with the proposed Stereo Panning Spectrum Features.

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 SPS features has been integrated into *Marsyas* [1] [6], an open source framework for audio processing with specific emphasis on Music Information Retrieval. The machine learning part of the experiments were conducted using Weka [2] [8].

4 CONCLUSIONS AND FUTURE WORK

A new feature set based on the Stereo Panning Spectrum was proposed and shown to be effective for two non-trivial audio classification tasks. It has been argued that the approach of modeling timbral similarity using MFCC has reached a "glass ceiling" [1]. We believe that information related to the recording process such as the stereo panning information used in this paper can help future audio MIR systems escape this ceiling. More detailed features related to stereo information than the ones proposed in this paper can be envisioned. For example, by clustering the panning values it might be possible to determine how many tracks were used in the mix.

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. One of the authors is a professional studio recording engineer who teaches recording techniques. We are planning to develop visualization and editing tools that can help reverse-engineer the stereo mixing of audio recordings for pedagogical purposes.

Engineers communicate about mixing with a particular lexicon of qualitative terms. A good example comes from an interview with Mix Magazine where Dave Pensado 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." [5]. Our hope is to eventually be able to translate this type of discussion into a more quantitative domain.

[1] `http://marsyas.sourceforge.net`

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

Acknowledgments

The authors would like to thank the National Sciences and Engineering Research Council (NSERC) and Social Sciences and Humanities Research Council (SSHRC) of Canada for funding this work, Perry Cook for suggesting the idea of using stereo a long time ago, and Carlos Avendano for describing his method with sufficient clarity and detail to be re-implemented.

5 REFERENCES

[1] J.J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] 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)*, pages 55–58, 2003.

[3] 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)*, pages 393–394, 2006.

[4] V Moorefield. *The Producer as Composer*. MIT Press, 2005.

[5] Dave "Hard Drive" Pensado. Interview. Mix Magazine, September 2001.

[6] G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(3), 2000.

[7] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with minnowmatch. In *Proc. IEEE Workshop on Neural Networks for Signal Processing, 2001*, pages 559–568, 2001.

[8] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[9] 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.