TOWARDS AN OBJECTIVE COMPARISON OF PANNING FEATURE ALGORITHMS FOR UNSUPERVISED LEARNING

Richard Mitic *

School of Innovation, Design and Engineering
Mälardalens Universitet
Västerås, SE
richard.mitic@mdu.se

Andreas Rossholm

Spotify
Stockholm, SE
arossholm@spotify.com

ABSTRACT

Estimations of panning attributes are an important feature to extract from a piece of recorded music, with downstream uses such as classification, quality assessment, and listening enhancement. While several algorithms exist in the literature, there is currently no comparison between them and no studies to suggest which one is most suitable for any particular task. This paper compares four algorithms for extracting amplitude panning features with respect to their suitability for unsupervised learning. It finds synchronicities between them and analyses their results on a small set of commercial music excerpts chosen for their distinct panning features. The ability of each algorithm to differentiate between the tracks is analysed. The results can be used in future work to either select the most appropriate panning feature algorithm or create a version customized for a particular task.

1. INTRODUCTION

The vast majority of recorded music available today was recorded and produced in stereophonic format. The musicians and producers involved will often pay close attention to the placement of sound objects in the stereo field, and various pieces of equipment such as goniometers, phase scopes, and stereo balance indicators have been standard equipment in recording studios for decades. There is therefore a wealth of spatial information encoded in the final recording. Extracting and disentangling that information has been the subject of research for many years, in various different guises.

A typical use case is classification, where stereo features have been used as a predictor of genre [1, 2, 3], artist/composer/DJ [2, 4], and decade [5]. Additionally, panning features have been used on their own merit (as opposed to predictors of another feature) for tasks such as source separation [6, 7], upmixing [8], and music similarity/recommendation [9, 4]. The classification of panning features themselves has also been studied as a precursor to upmixing, source separation, and audio enhancement [10, 11], giving the ability to select appropriate algorithms based on audio content.

In the aforementioned cases, the audio analysed has either comprised single excerpts or lists of 1000-2000 tracks with labels assigned. However, in recent years much bigger musical datasets have become available, e.g. [12, 13], which contain 55000 and 100000 tracks respectively. These datasets do not contain any metadata regarding panning features, nor any similar information such as recording methods/style. While they do offer genre and date tags, panning features on their own have previously shown low correlation with genre and decade [2, 5], from which we can infer that genre and date cannot be reliably used a proxy for panning features. If one wishes to use these datasets for the aforementioned use cases, there is hence a need to perform unsupervised learning on the extracted panning features, and for this to be successful we must understand the abilities and limits of the extraction method being used.

Several examples of panning feature extractors for stereo audio signals exist in the literature [1, 14, 9, 5], but as yet there has not been a comparison of them and no discussion of what panning features they are able to detect. The aim of this paper is to provide such discussion. It will begin by summarizing the relevant literature in Section 2, describing and drawing links between existing panning feature extraction algorithms. Section 3 presents examples of commercial music chosen for their different uses of amplitude panning, plus the methods used to compare the panning feature algorithms. Section 4 shows the results of the analysis and Section 5 will discuss some overall trends that emerged across all algorithms. Concluding remarks are offered in Section 6.

2. PREVIOUS WORK

[8, 15] are seminal works describing the 'panning index' – a description of amplitude panning for each time-frequency bin in a stereo spectrogram. [8] also presents the 'ambience index' – a measure of inter-channel correlation in time-frequency (TF) bins. While they target re-synthesis and upmix as use cases, the underlying techniques have become the basis for many applications since. [6] is another seminal work which describes a frequency-azimuth plane constructed from phase cancellation of delayed stereo TF planes. Similarly, it targets re-panning and re-synthesis but has since been repurposed. [1, 2] reuse the panning index from [8] (renamed to 'Stereo Panning Spectrum (SPS)') and derive a set of Stereo Panning Features (SPF) that are used for genre classification and Music Information Retrieval (MIR). [14] details a method to estimate the perceptual stereo width of a music recording by forming a Panning Histogram (PH) based on the frequency-azimuth plane from [6]. Another PH method (this time based on the SPS from [1]) is described in [9, 16], specifying the measurement of music similarity as the primary use case. It contributes an extra step of converting the PH into Panning Coefficients (PC) through cepstral analysis, stating that this allows the resulting vector to be used with any generic classification algorithm. In a similar vein (and with classification as a use case) [5] combines cochleagram differences with an analysis of phase differences between

* This work was supported by the KKS Foundation

channels to form two feature sets - the 'Amplitude Stereo Features (ASF)' and 'Phase Stereo Features (PSF)'.

[17, 11, 10] collectively define a new paradigm of classifying regions of audio whereby stereo mixtures are modelled as combinations of centre, panned, and uncorrelated sources. Even though genre classification is used as an example use case, they specifically state that the class representing a combination of all source types (i.e. most music recordings) is very difficult to identify and therefore omitted from the model.

Similar to the phase-based features defined in [5], there are a group of algorithms based on the emulation of a studio goniometer; [4, 3] use both a goniometer box-counting technique and inter-channel correlation analysis to derive a time-varying measurement of 'spaciousness'. Interestingly though, [18] found that subjective impressions of spaciousness in various musical styles were largely contradictory to the results of goniometer analysis.

More recently, Machine Learning (ML)-based techniques for disentangling spatial information from audio content have been demonstrated. With a primary use case of upmixing stereo to multichannel, [19] demonstrates that a latent vector formed by running multichannel audio through one half of the learned neural network represents the spatial information contained in the original signal. Similarly, [20] takes a stereo microphone input and regenerates one channel from the other by learning an inter-channel spatial representation. Mirroring our problem statement, [20] specifically mentions a lack of labelled real-world data as motivation for the use of unsupervised learning.

2.1. Selected algorithms

This paper will focus on four algorithms that estimate panning features from stereophonic commercial music recordings and do not require training on a dataset. Specifically, PH [14] PC [9], ASF [5], and SPF [1]. For mathematical definitions the reader is referred to the original papers, but a summary of each is given here.

2.1.1. Panning Histogram (PH) — Sarroff, 2008 [14]

The left and right channels are split into chunks and transformed into the frequency domain using a Fast Fourier Transform (FFT). Each FFT frame is converted into a frequency-azimuth (FA) plane through phase cancellation techniques. The left and right FA planes are concatenated, and a histogram is calculated across all frames and all frequencies. The result is a histogram of energy at each azimuth bin.

Note: The paper continues to collapse the histogram into a single scalar representing overall width, but we will not consider that here since the intent is to compare the PH vector with other high-dimensional representations of panning features.

2.1.2. Panning Coefficients (PC) — Gomez, 2008 [9]

The left and right channels are split into chunks and transformed into magnitude spectrograms using a Short-time Fourier Transform (STFT). The ratio of power spectra is calculated for each frame, resulting in an azimuthal panning factor each frequency bin. These values are warped by a function that accounts for the human perception of the direction of arrival of sounds. Each frame is then converted to an energy-weighted histogram and the mean over time is taken. The histogram is then converted to a compact representation by performing cepstral analysis and truncating the results to length 20.

2.1.3. Amplitude Stereo Features (ASF) — Tardieu, 2011 [5]

The ASF are created from a stereo cochleagram. Left and right audio channels are split into frequency sub-bands using gammatone filters spaced equally on the Equivalent Rectangular Bandwidth (ERB) scale. The amplitude difference of the two cochleagram channels is taken to produce a Cochleagram Difference (CD). The ASFs are created by taking the mean and standard deviation over various axes of the CD. The mathematical definition of a cochleagram is not given in [5], so for this paper we have followed the steps in [21], creating a cochleagram directly from time domain audio as opposed to using an FFT. [5] also does not state the exact form of the resulting vector, so for this paper all features are simply concatenated into a single vector.

2.1.4. Stereo Panning Features (SPF) — Tzanetakis, 2007 [1]

The left and right channels are converted into complex TF regions by chunking, windowing, and the STFT. These two TF regions are converted into an SPS, which is then split into high, mid, and low frequency bands. Each sub-band (along with the full spectrum) are collapsed by calculating the Root Mean Squared (RMS) across the frequency dimension, and smoothed in time by calculating a running mean and standard deviation over M frames. The mean and standard deviation over time are then calculated, resulting in a 16-dimensional vector.

3. METHOD

To assess each algorithm they are run on a selection of musical excerpts exhibiting panning features that it would be useful to differentiate between. As a way of delineating the entire panning feature gamut, we use two 'anchor tracks' - one strongly mono and one strongly stereo. The other tracks each represent an interesting panning feature. For each algorithm, the panning features for all tracks are calculated and considered to be points in an N-dimensional feature space. The Euclidean distances (D) between pairs of tracks are calculated, and the set for each algorithm is normalized so that $D \in [0, 1]$. Specifically, it is hypothesized that

1. The anchor tracks should be located farthest apart in the feature space
2. The panning feature tracks should be well separated and evenly distributed between the anchors

Any pair with low separation (i.e. the algorithm cannot differentiate between them) is further examined to find out why. We define a low separation to be $D < 0.1$.

3.1. Musical excerpts

All excerpts are defined by their title, artist, release year, and timing information. For the rest of the paper they will be referred to by abbreviated title. The panning features they exhibit are also defined here. Anchor tracks are denoted by '*' for mono and '**' for stereo.

- *'Round Midnight* (RM*)*, Miles Davis, 1957, 0:00–0:30. A jazz piece recorded in mono.
- *This Is The Thing (TITT)*, Fink, 2007, 1:00–1:30. An acoustic song with studio reverb added. The vocal switches from single-tracked to multi-tracked with moderate amplitude panning approximately halfway though the excerpt.

- *Laser Beam (LB)*, Low, 2001, 0:00–0:30. Electric guitar, bass, and vocals with an enveloping sound created mostly by reverberation and multi-miking, but only subtle amplitude panning.
- *Manifesto (M)*, Chilly Gonzales, 2025, 1:00–1:30. Solo upright piano recorded live on stage with close stereo microphones. Individual notes are audibly separated in space even though no studio panning has been applied.
- *Cut And Clicks (CAC)*, Tetsu Inoue, 2000, 0:00–0:30. An electronic piece that is characterized by heavy use of extreme panning, unnatural sounds, and singular audio events with very short durations.
- *Fly Me To The Moon (FMTTM)*, Frank Sinatra, 2008, 1:00–1:30. A full big band recorded live. Most instruments are statically panned to a fairly strong degree.
- *I Saw Her Standing There*** *(ISHST**)*, The Beatles, 1963, 0:00–0:30. A pop song with strong, rudimentary panning. Drums, bass, and lead guitar are panned hard left, while vocals and rhythm guitar are panned hard right.

We define 'natural panning' to mean that recorded instruments that have been amplitude panned to a static position in the final recording. Examples of natural panning are *TITT*, *FMTTM*, and *ISHST***. The counterpart is 'unnatural panning', which we use to refer to sounds that move very fast between left and right. *CAC* is an example of unnatural panning.

All recordings have a sample rate $f_s = 44100$ Hz and were normalized to an average loudness of -23 Loudness Units Full Scale (LUFS) before processing according to [22, 23].

3.2. Algorithm parameters

Each algorithm is defined by a set of input parameters, and the values used for this paper are shown in Table 1. For precise meanings of each one, the reader is directed to the original papers.

Algorithm	Parameters	Dimensions
PH [14]	FFT size = 1024 hop size = 512 azimuth bins = 129	129
PC [9]	FFT size = 1024 hop size = 512 azimuth bins = 129 cepstrum coefficients = 20	20
ASF [5]	Low f_c = 30Hz High f_c = 11050Hz number of bands = 70 chunk length = 20ms hop length = 10ms	272
SPF [1]	FFT size = 1024 hop size = 512 M = 40 low band = 0-250 Hz mid band = 250-2500 Hz high band = 2500-22050 Hz	16

Table 1: *Parameters for panning algorithms*

4. RESULTS

The Euclidean distances between pairs of tracks in each feature space are shown in Fig. 1. Within each feature space, distances have been normalized to the range $D \in [0, 1]$. Cells are coloured in greyscale with white and black representing 0 and 1 respectively, and track titles are abreviated.

Fig. 1 shows that some panning feature algorithms have a clear bias towards one of the anchor tracks. (visible as a much darker section in otherwise pale matrix). A bias towards the mono anchor (i.e. a dark horizontal line on the bottom row) means that all panning feature tracks appear 'very mono' to the algorithm. Conversely, a bias towards the stereo anchor (i.e. a dark vertical line in the left column) means that all panning feature tracks appear as 'very stereo' to the algorithm.

To quantify this, we define a 'bias score' (B) according to Eq. (1). Firstly the mean distances from each anchor track to all panning feature tracks are calculated, then B is defined as the natural logarithm of the ratio of mean mono distance ($\bar{D}_m$) to mean stereo distance ($\bar{D}_s$). A score of $B = 0$ means no bias, a positive score ($B > 0$) means a bias towards the stereo anchor and negative score ($B < 0$) means a bias towards the mono anchor. This is calculated by Eq. (1) where $T = \{t : t \in T\}$ is the set of panning feature tracks, $D_{m,t}$ is the distance between the mono anchor and a panning feature track, and $D_{s,t}$ is the distance between the stereo anchor and a panning feature track.

$$\begin{aligned} \bar{D}_m &= \frac{1}{|T|} \sum_{t \in T} D_{m,t} \\ \bar{D}_s &= \frac{1}{|T|} \sum_{t \in T} D_{s,t} \\ B &= \ln \frac{\bar{D}_m}{\bar{D}_s} \end{aligned} \tag{1}$$

The scores for each algorithm are shown in Table 2 alongside their interpretations.

Algorithm	$\bar{D}_m$	$\bar{D}_s$	B	Interpretation
PH [14]	0.72	0.26	1.02	strong stereo bias
PC [9]	0.91	0.16	1.73	very strong stereo bias
ASF [5]	0.31	0.87	-1.04	strong mono bias
SPF [1]	0.50	0.64	-0.25	slight mono bias

Table 2: *Bias scores for all panning feature algorithms*

From Fig. 1 it is simple to test hypothesis 1 (the anchor tracks should be the furthest apart) by finding the cell with a value of 1. It is also possible to assess hypothesis 2 (the other tracks should be well separated) by finding cells with low values. Each panning feature algorithm will be assessed separately.

4.1. Panning Histogram [24]

Fig. 1 shows that this is the only algorithm for which hypothesis 1 is not satisfied. Instead of the two anchor tracks (*RM** and *ISHST***) having the largest distance, *RM** and *CAC* are in fact the two most

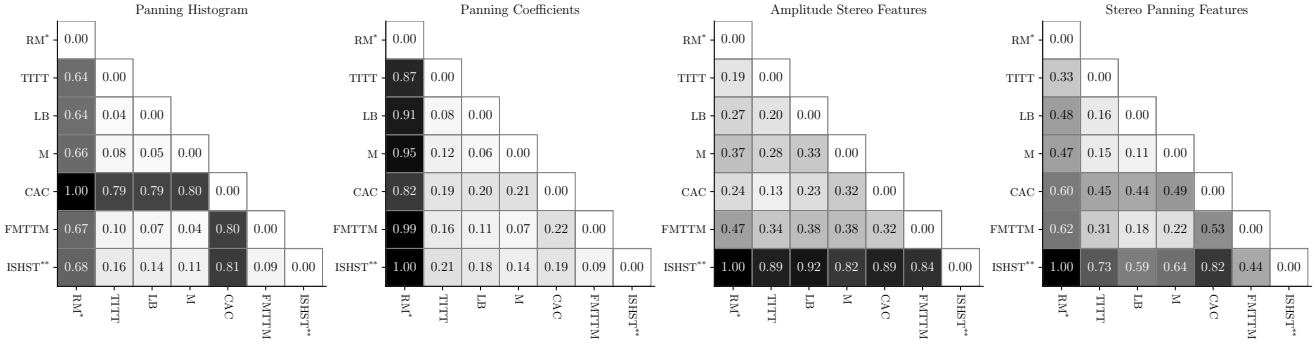

Figure 1: *Distance between pairs of tracks in the feature spaces of the four panning feature algorithms.*

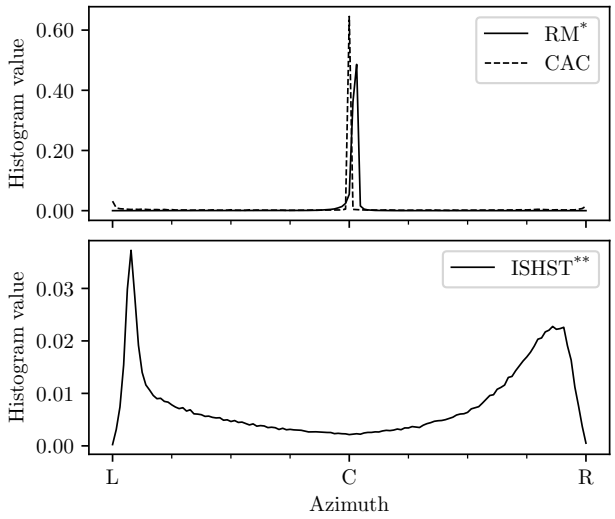

Figure 2: *PHs of RM*, CAC and ISHST**, demonstrating that CAC has been wrongly detected as monophonic.*

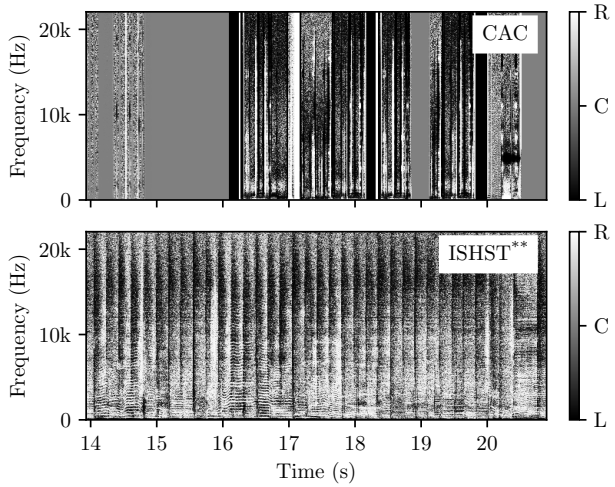

Figure 3: *Stereo Panning Spectrum of a short section of CAC and ISHST**.

separated tracks. To determine why this was the case, the PHs of all three tracks are plotted in Fig. 2.

From this diagram we can observe two important phenomena. Firstly, *CAC* (which is characterized by extreme and unnatural panning) actually appears close to mono in its PH. By inspecting the panning spectrogram (a section of which is shown in Fig. 3) we see that when hard panning occurs, time-frequency frames panned hard left and right are roughly evenly distributed. Hence, when the mean over time is taken, the hard-panned frames cancel each other out and all energy appears to be located in the centre. This differs from e.g. *ISHST*** which consistently has energy panned to both channels simultaneously and hence has two histogram peaks on the left and right.

Secondly, even though *RM** and *CAC* have similarly-shaped PHs, their Euclidean distance is very large. This is because the central peaks of each track appear in different azimuth bins. Since each bin is a separate dimension in the feature space, these two tracks effectively appear as two vectors with large amplitudes pointing in different directions. From this we can determine that the Euclidean distance is not a good measure of similarity for Panning Histograms, and we will therefore not assess hypothesis 2.

4.2. Panning Coefficients [9]

Inspection of Fig. 1 reveals that, as hypothesized, *RM** and *ISHST*** are the two tracks with the most separation. However, as Table 2 shows, it also has a strong bias towards the stereo anchor. The tracks *TITT*, *LB*, *M*, and *FMTTM* all appear in close pairs, implying that PC might have trouble differentiating between tracks that use moderate, but different, amplitude panning techniques. Reassuringly though, tracks that include extreme panning are separated from moderately-panned tracks, with $D \approx 0.2$.

The PC algorithm is based on a Panning Histogram that is quantized by truncation in the cepstral domain. To assess the effect of this, we can reverse the cepstral process using Eq. (2) where $\boldsymbol{c}$ is the vector of cepstral coefficients that has been truncated to length L. First, a pseudo-spectrum $\boldsymbol{s}$ is created by taking the exponential of the real part of the Discrete Fourier Transform (DFT) of $\boldsymbol{c}$. $\boldsymbol{s}$ is then truncated to half its length to recreate a histogram, and normalized so that the total area is 1. The result is a Panning Histogram $\boldsymbol{h}$ with $L/2$ azimuth bins.

$$\begin{aligned} \boldsymbol{c} &= (c_1, c_2, \ldots, c_L) \\ \boldsymbol{s} &= \exp(\mathrm{Re}(\mathrm{DFT}(\boldsymbol{c}))) \\ \boldsymbol{h} &= \frac{\boldsymbol{s}_{1:L/2}}{\frac{L}{2}\sum \boldsymbol{s}_{1:L/2}} \end{aligned} \tag{2}$$

Fig. 4 shows the quantized and original PHs for *ISHST*** and its three closest neighbours. From these diagrams it is clear to see why the PC algorithm is biased towards the stereo anchor - the anchor itself has had its panning energy moved from the side lobes towards the centre, and hence the detected panning features are more akin to the moderately-panned tracks (which are not so affected by the quantization).

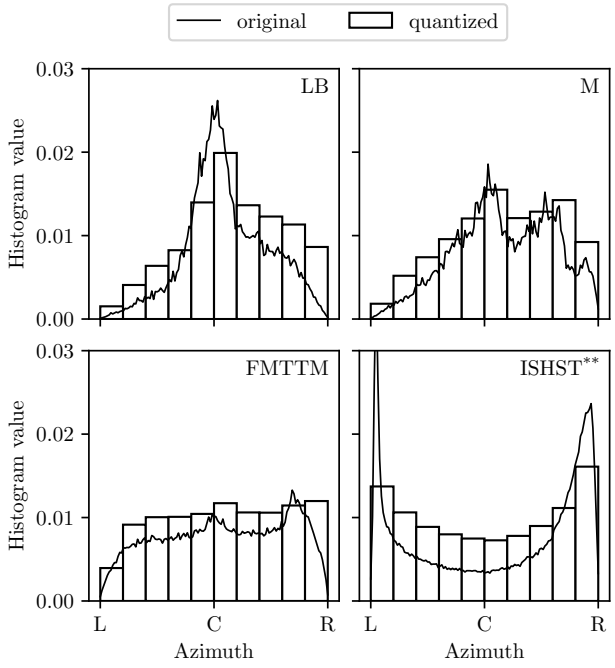

Figure 4: *PHs of ISHST** and its three closest neighbours before and after cepstral quantization.*

4.3. Amplitude Stereo Features [5]

Similarly to PC, the ASF distance matrix (Fig. 1) shows that the two anchor tracks are indeed the furthest separated. Conversely though, Table 2 shows that ASF has a bias towards the mono anchor. The separation of all tracks except *ISHST** is, however, quite adequate, with a range of $D \in [0.13, 0.47]$.

Curiously, *FMTTM* and *ISHST** have a high separation ($D = 0.84$) despite both exhibiting strong natural panning. Fig. 5 shows a visual representation of ASF for these two tracks. The panning features 'mean over time' and 'standard deviation over time' (see [5]) for each cochlea band are depicted as dots with error bars. From this diagram it is clear the two tracks have roughly opposite panning along the frequency axis. For example, *FMTTM* has a double bass panned to the right, while *ISHST** has a bass guitar panned to the left. This difference in panning direction is certainly

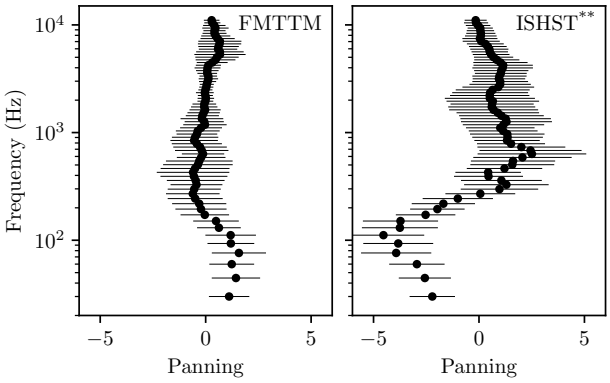

Figure 5: *ASF for FMTTM and ISHST** showing similar frequency bands panned in opposite directions.*

a contributor to the high separation between these two tracks, and hence the overall mono bias of ASF.

It is also notable that the panning variation for both tracks shown in Fig. 5 is quite high, when we can hear in the original audio that instruments are panned statically. This is most likely caused by overlapping frequency components in the left and right spectra – a common issue discussed in previous literature [8, 6, 25]. Certainly the phenomenon is not unique to the ASF algorithm but nevertheless is illustrated particularly well by this diagram.

While not too close by our standards, *CAC* and *TITT* ($D = 0.13$) are the closest pair. This is surprising since one would imagine that the moderate, natural panning of *TITT* would be well separated from the extreme, unnatural panning of *CAC*. Examining their ASF plots (see Fig. 6) reveals that while the panning variation of each track is quite different, they both exhibit mean panning that is roughly centred, mirroring the PH error described in Section 4.1. One possible explanation is that the fast panning of *CAC* manifests more as phase difference than amplitude difference, which would match the dispersion introduced by the reverberation in *TITT*. While panning variation for *CAC* is much higher in the lower frequencies, it could be the case that the high number of dimensions in ASF is acting to quash those differences.

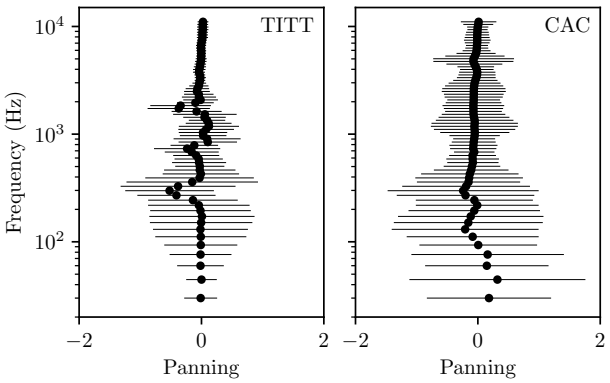

Figure 6: *ASF for TITT and CAC, both exhibiting average central panning.*

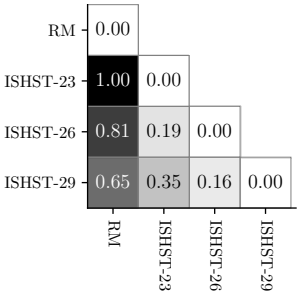

Figure 7: *Distance matrix comparing RM* (at -23 LUFS) to ISHST** at -23, -26, and -29 LUFS.*

One important factor in the use of ASF is that its calculation (defined in [5]) does not include any step to account for the overall loudness of the audio. To demonstrate the effect of this, we can calculate another distance matrix comparing *RM** to several versions of *ISHST*** that are normalized to different loudness levels. Fig. 7 visualizes this and clearly shows that the relative loudness of two tracks affects their distance in the feature space. All tracks analysed in this paper were normalized to equal loudness to account for this, but that might not be possible in a real-world use case, e.g. analysing streaming audio in real time.

4.4. Stereo Panning Features [1]

The distance matrix (Fig. 1) for SPF reveals that the two anchor tracks are indeed the furthest separated, and Table 2 shows the lowest bias. There are also no track pairs with a low separation. However, the distance matrix does show a generally high separation for *CAC*. This implies that SPF is able to differentiate between natural and unnatural panning quite well. In particular, *CAC* is maximally separated from *ISHST*** even though both contain extreme panning.

The definition of SPF includes estimations of both 'short-term panning' (the running standard deviation over M frames) and 'long-term panning' (the running mean over M frames). These are calculated over four frequency bands (fullband, lows, mids, and highs), and the global means and standard deviations of these comprise the final feature vector. We can therefore visualize the SPF as a scatter plot with the vertical position of a point denoting the amount of panning in each frequency band and error bars denoting panning variation, with short- and long-term values separated.

Fig. 8 shows the SPF plots for *CAC* alongside the two anchor tracks. It reveals that indeed, the unnatural panning of *CAC* is visible as large error bars in all dimensions whereas *ISHST*** has much higher average panning and relatively low variation, thus demonstrating how SPF is able to differentiate between natural and unnatural panning.

*RM** has low, but non-negligible values in all dimensions. Inspection of the original formula [1] shows that a truly mono signal would produce an SPF vector of all 0s, revealing that *RM** is, in fact, not entirely mono. By listening to the difference between the left and right channels, we can hear that the original mono recording has had a decorrelation effect applied at the mastering stage. We can therefore conclude that SPF is sensitive to this kind of effect.

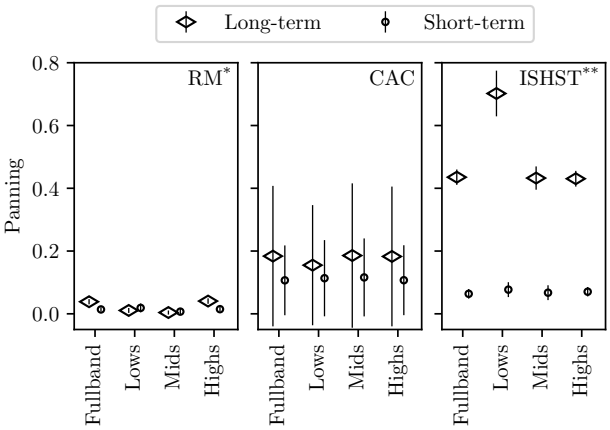

Figure 8: *SPF plots for RM*, CAC and ISHST** showing clear separation between all three.*

5. DISCUSSION

Table 3 shows a summary of each algorithm. For the purposes of unsupervised learning on commercial music, SPF has shown the most promise as it has the lowest bias score and good separation between all the panning feature tracks. The features related to panning variation over time appear to be important as they can help avoid issues such as the misidentification of unnatural panning as monophonic audio (see Section 4.1).

Each algorithm explicitly chooses whether to represent the panning *direction* or just the panning *amount*. ASF even includes features for both. Likewise, some algorithms act on frequency sub-bands whereas others process the entire spectrum. When the intention is to use panning features for unsupervised learning, the importance of both panning direction and frequency sub-bands is application-specific; any of the algorithms could be modified to include or exclude these two features.

The computational complexity of the algorithms was not addressed in this paper, but varied significantly among them. SPS-based descriptors ([1, 9]) are reliant on an STFT for their TF representation and hence are efficient but sometimes difficult to reconcile with human perception, whereas algorithms based on cochleagrams ([5]) produce more perceptually relevant results but require very large banks of time-domain filters.

An additional variable not covered by this paper is the duration of analysis. For the purposes of this experiment, 30s was chosen to roughly match [1], although [9] mentions that any duration between 2s and the entire recording is acceptable and [5] shows examples of entire songs.

All algorithms differ greatly in the number of dimensions represented in the final panning features. When calculating Euclidean distances in N-dimensional space, it has been proven that the inclusion of variables which provide no additional accuracy both increases the distance between points and raises the likelihood of attributing significance to spurious measurements [26]. A balance must therefore be struck between the resolution of the panning features and the ease of differentiating between tracks. Again, each algorithm can be modified to increase or decrease the dimensionality.

Algorithm	Frequency bands	Variation over time	Panning direction	Azimuthal resolution	Loudness invariant	Notes
PH [14]	Fullband	No	Yes	High	Yes	Cannot use Euclidean distance. Azimuthal resolution configurable.
PC [9]	Fullband	No	Yes	Low	Yes	Strong stereo details are lost. Azimuthal resolution configurable.
ASF [5]	ERB, 70 bands	Yes	Yes	High	No	Strong mono bias. Very large feature space.
SPF [1]	Lows, Mids, Highs	Yes	No	High	Yes	Relatively low bias. Separates natural/unnatural panning well. Sensitive to decorrelation effects.

Table 3: *Summary of properties for all panning feature algorithms.*

6. CONCLUSION

Four panning feature extraction algorithms were tested with respect to their use for unsupervised learning. They were run on five tracks that exhibit interesting panning features, plus two 'strongly mono' and 'strongly stereo' anchor tracks. The algorithms' results for individual tracks were analysed to assess what panning features each algorithm is able to differentiate between. Each algorithm was assigned a 'bias score' based on how the panning feature tracks were distributed within the anchor tracks. SPF from [1] showed the most promise, although it lacks differentiation between left and right and has relatively low frequency resolution.

In future work using panning features for unsupervised learning we recommend setting precise requirements for relevant panning features and choosing an algorithm accordingly based on the results given here. Alternatively, algorithms may be modified or combined to suit a specific need. The inclusion of other spatial descriptors such as those based on phase differences or generated via ML might also prove fruitful. To that end, a much larger survey of spatial analysis techniques would be beneficial.

In the cases where the original literature performed experiments on a commercial music collection [1, 5], it is not possible to fully reproduce their results since the exact tracks are unspecified. Analysis of modern, open datasets (e.g. [12, 13]) would therefore be advantageous and would strengthen the results obtained in the current experiments.

7. REFERENCES

[1] G. Tzanetakis, Randy Jones, and K. McNally, "Stereo Panning Features for Classifying Recording Production Style," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2007, pp. 441–444.

[2] G. Tzanetakis, L. Martins, K. McNally, and Randy Jones, "Stereo Panning Information for Music Information Retrieval Tasks," *J. Audio Eng. Soc.*, June 2010.

[3] Tim Ziemer, "Goniometers are a Powerful Acoustic Feature for Music Information Retrieval Tasks," Feb. 2023.

[4] Tim Ziemer, Pattararat Kiattipadungkul, and Tanyarin Karuchit, "Acoustic features from the recording studio for Music Information Retrieval Tasks," *Proceedings of Meetings on Acoustics*, vol. 42, no. 1, pp. 035004, Feb. 2021.

[5] Damien Tardieu, Emmanuel Deruty, Christophe Charbuillet, and Geoffroy Peeters, "Production Effect: Audio Features For Recording Techniques Description And Decade Prediction," in *Proc. 14th Int. Conf. Digit. Audio Eff. DAFx-11*, Paris, FR, Sept. 2011.

[6] Dan Barry and Robert Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis," in *Proc. 7th Int. Conf. Digit. Audio Eff. DAFx04*, Naples, IT, 2004.

[7] Shim Hwan, Jonathan S. Abel, and Sung Koeng-Mo, "Stereo Music Source Separation for 3D Upmixing," *J. Audio Eng. Soc.*, vol. 7938, Oct. 2009.

[8] Carlos Avendano and Jot Jean-Marc, "Frequency Domain Techniques For Stereo To Multichannel Upmix," in *J. Audio Eng. Soc.*, Espoo, FI, June 2002, vol. 000251.

[9] Emilia Gomez, Perfecto Herrera, Pedro Cano Vila, Jordi Janer, Joan Serra, Jordi Bonada, Shadi Walid El-Hajj, Thomas Etienne Aussenac, and Gunnar Nils Holmberg, "Music similarity systems and methods using descriptors," US Patent US20080300702A1, Dec. 2008.

[10] Aki Härmä, "Stereo audio classification for audio enhancement," *2011 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, pp. 457–460, May 2011.

[11] Aki Härmä, "Classification of time-frequency regions in stereo audio," *J. Audio Eng. Soc.*, vol. 59, no. 10, pp. 707–720, 2011.

[12] D. Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," in *International Conference on Machine Learning*, June 2019.

[13] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A dataset for music analysis," in *18th Int. Soc. Music Inf. Retr. Conf. ISMIR*, 2017.

[14] Andy M. Sarroff and J. Bello, "Measurements of spaciousness for stereophonic music," in *J. Audio Eng. Soc.*, Oct. 2008, vol. 7539.

[15] Carlos Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," *2003 IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 55–58, 2003.

[16] Enric Guaus, *Audio Content Processing for Automatic Music Genre Classification: Descriptors, Databases, and Classifiers*, Ph.D. thesis, Universitat Pompeu Fabra, Barcleona, 2009.

[17] Aki Härmä, "Classification of time-frequency regions in stereo audio," *128th Audio Eng. Soc. Conv. 2010*, pp. 204–215, 2010.

[18] Claudia Stirnat and Tim Ziemer, "Spaciousness in music: The tonmeister's intention and the listener's perception," in *KLG 2017 Klingt Gut 2017 – Int. Symp. Sound*, Philipp Kessling and Thomas Görne, Eds. 2019, vol. 1 of *EPiC Series in Technology*, pp. 42–51, EasyChair.

[19] Haici Yang, Sanna Wager, Spencer Russell, Mike Luo, Minje Kim, and Wontak Kim, "Upmixing Via Style Transfer: A Variational Autoencoder for Disentangling Spatial Images And Musical Content," in *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, May 2022, pp. 426–430.

[20] Bing Yang and Xiaofei Li, "Self-Supervised Learning of Spatial Acoustic Representation With Cross-Channel Signal Reconstruction and Multi-Channel Conformer," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 4211–4225, 2024.

[21] Ryo Isobe and Takashi Nishi, "Relationships between interaural cross-correlation function and spatial impression of sound reproduced with five loudspeakers," *Acoust. Sci. & Tech.*, vol. 33, no. 3, pp. 193–196, 2012.

[22] "Loudness Normalisation And Permitted Maximum Level Of Audio Signals," EBU Recommendation R128, Geneva, CH, 2023.

[23] "Algorithms to measure audio programme loudness and true-peak audio level," ITU Standard BS.1770, Geneva, CH, Oct. 2015.

[24] Andy M. Sarroff, "Spaciousness In Recorded Music: Human Perception, Objective Measurement, And Machine Prediction," M.S. thesis, New York University, New York, 2009.

[25] Derry FitzGerald and Dan Barry, "On inpainting the Adress algorithm," in *IET Ir. Signals Syst. Conf. ISSC 2012*, June 2012, pp. 1–6.

[26] Benjamin Thirey and Randal Hickman, "Distribution of Euclidean Distances Between Randomly Distributed Gaussian Points in n-Space," Aug. 2015.