PRODUCTION EFFECT: AUDIO FEATURES FOR RECORDING TECHNIQUES DESCRIPTION AND DECADE PREDICTION

Damien Tardieu, Emmanuel Deruty ,Christophe Charbuillet and Geoffroy Peeters*

STMS Lab IRCAM - CNRS - UPMC
1 place Igor Stravinsky
75004 Paris, France
dtardieu, deruty, charbuillet, peeters (at) ircam.fr

ABSTRACT

In this paper we address the problem of the description of music production techniques from the audio signal. Over the past decades sound engineering techniques have changed drastically. New recording technologies, extensive use of compressors and limiters or new stereo techniques have deeply modified the sound of records. We propose three features to describe these evolutions in music production. They are based on the dynamic range of the signal, energy difference between channels and phase spread between channels. We measure the relevance of these features on a task of automatic classification of Pop/Rock songs into decades. In the context of Music Information Retrieval this kind of description could be very useful to better describe the content of a song or to assess the similarity between songs.

1. INTRODUCTION

Recent popular music makes an exhaustive use of studio-based technology. Creative use of the recording studio, referred to as production, exerts a huge influence on the musical content [1]. Sonic aspects of music, as brought by studio technologies, are even considered by some authors to be at the top of the hierarchy of pertinence in contemporary popular music analysis [2]. They can be perceived as more important than rhythm and even than pitch.

Studio techniques may concern many aspects of the musical content. Equalizers modify spectral content, reverberation bring customizable acoustics to the recording, pitch-shifters like Antares Autotune[1] can transform vocals to a point where it becomes the trademark of a song [3]. Double and multiple tracking techniques allow the construction of heavily contrapuntal and spatialized parts from a single original sound source or musician [4]. Dynamic processing used in audio mastering weights so heavily on music perception that it spawns public debate [5].

Studio practices are heavily dependent on equipment: equalizers and dynamic compressors require electronic components, pitch-shifting is impossible to perform without digital processing and recordings have to be made on media whose performance are highly variable across the musical periods. This leads to the hypothesis that some sonic aspects in recorded music are specific to a given period of time.

In the Music Information Retrieval field, this aspect has received few attention. The first work which could be related to production is the Audio Signal Quality Description Scheme in MPEG-7 Audio Amendment 1 [6]. This standard includes a set of audio features describing the characteristics of the support, considered as a transmission channel of a music track: description of BackGroundSoundLevel, RelativeDelay, Balance, Bandwidth. In [7], Tzanetakis uses the Avendano's Panning Index [8] to classify production styles (and then production time). Kim [9] and Scaringella [10] study the effect of remastering on the spectrum of the songs. Their interest in remastering comes from a question that was more debated, the so-called "album effect". This refers to the fact that machine learning algorithms for automatic music classification or music similarity estimation may learn characteristics of the album production instead of general properties such as genre and then be over fitted. Identifying this album effect is still an open problem but we believe that some production aspects do not belong to the album effect and may characterize the period of a song or even its genre. It is thus important to characterize the production effect that is independent of album and relates to more general attributes of a song. In this aim we propose three features describing some aspects of the production of a song. The first feature relates to the temporal variation of the signal amplitude and is described in section 2, the second and the third, detailed in section 3 describe the use of stereo. To assess the accuracy of these features and how they relate to a production period, we use them in a task of automatic classification of songs in decades in section 4. We finally conclude in section 5.

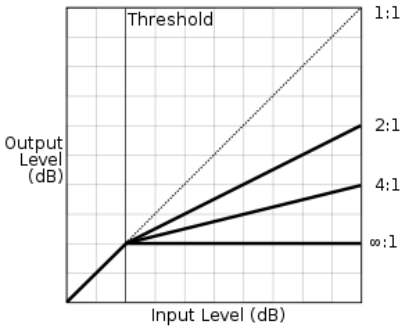

Figure 1: Relationship between input and output level in a compressor for a fixed threshold and various ratios.

* Part of this work was made as an independant consultant

[1] http://www.antarestech.com/

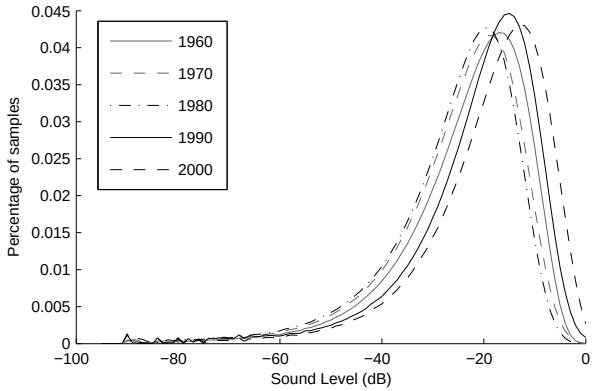

Figure 2: Mean decibel amplitude histogram for five decades from 1960 to 2000

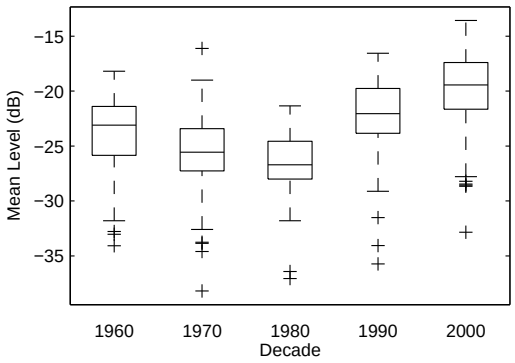

Figure 3: Mean of the signal amplitude absolute value in dB for five different decades. The center horizontal line represents the median. The middle two horizontal lines represent the upper and lower limits of the inter-quartile range. The outer whiskers represent the highest and lowest values that are not outliers. Outliers, represented by '+' signs, are values that are more than 1.5 times the inter-quartile range.

2. COMPRESSION AND LIMITING

2.1. A growing use of compressors and limiters over the years

The first techniques we study are compression and limiting. Their aim is to alter the amplitude of the signal in order to reduce its dynamic range (ie. the ratio between the loudest and the weakest parts of the signal's power). They can be used to deal with technical limitations of the recording system, or to improve the audibility of the signal for aesthetic reasons. A compressor applies a non-linear transformation to the sound level across time (see Fig. 1). It applies a negative gain to the signal whenever the amplitude exceeds a user-set threshold. Another way to deal with dynamic compression is to consider that one applies an input gain to the signal, which increases its power, while the signal's peaks must not get over a given threshold under in any circumstance. This is the principle of limiting. Intensive usage of limiting results in signals with many samples very close to 0 dB Full Scale (the maximum possible level on digital media). From the beginning of the 90s, this technique has been increasingly used to make songs sound "louder" while peaking at the same level . Each music recording company wanting to make records that sound "louder" than the ones from the competitor, this degenerated into a so-called "loudness war" (see for instance [11]). To describe these effects we propose a feature based on the amplitude of the signal in dB FS (Decibel Full-Scale).

2.2. Signal description of compression and limiting effects

2.2.1. Dynamic histogram

This feature corresponds to the histogram of the peak normalized signal level represented in dB. Let $s(n)$ be the audio signal with $s(t) \in [-1, 1]$.

$$s_{dB}(n) = 20 * log_{10}(|x(n)|) \quad (1)$$

The bins of the histogram are 1dB wide and the centers go from -95.5 dB to -.5dB. These values are chosen considering the 96 dB dynamic of a 16-bit signal. The histogram is normalized to represent percentage values.
We use the signal amplitude instead of any energy estimate to be able to precisely detect the effect of limiting. Indeed, both limiters and compressors apply a gain directly on the signal. As a result, the effect of the limiter, as it has been used recently, will be the presence of many samples with an amplitude very close to 0 dB. If we were using an energy estimate, such as RMS, these peak values would be smoothed and then less visible.
Fig. 2 shows the mean amplitude histogram computed on 1042 Pop/Rock songs (see section 4 for details) for five different decades ranging from 1960 to 2010. First we notice the progressive displacement of the histogram toward the right, ie. toward high sound level value, from the 80s to the 00s. This is typically the effect of a higher compression rate over decades. Looking at Fig. 3 representing the mean of the signal amplitude absolute value in dB for the five decades, we can confirm the increase of the sound level from the 80s to the 00s, but we also see that this value decreases from the 60s to the 80s. This diminution of the mean sound level can be explained by the increasing bandwidth of the recording media from less than 75dB in the 60s [12] to the 96 dB of the audio CD. Indeed, if the bandwidth increases and the peak value stays constant, the mean decreases. The second noticeable observation on these histograms appears on the high sound level bins, particularly in the [-1dB,0dB] bin. Indeed we see that the height of these bins increase with the decade. This is an effect of the intensive use of limiters, and justifies the use of amplitude instead of energy estimation. This effect is more visible on Fig. 4 that shows the percentage of samples between -1 and 0 dB (ie. the height of the rightest bin of the histogram) for the five decades. We see that this value does not vary much in the three first decades and starts growing in the nineties to reach a top value in the 00s.

2.2.2. Summary features

To obtain a more compact representation, we also compute the four first moment of the distribution of s_{dB}, ie. the mean, the variance, the skewness and the kurtosis, as well as the median and inter-quartile range. In the following we call these features, together with the histogram bin amplitude, the *Dynamic Features*. A higher compression rate should be materialized by a higher mean

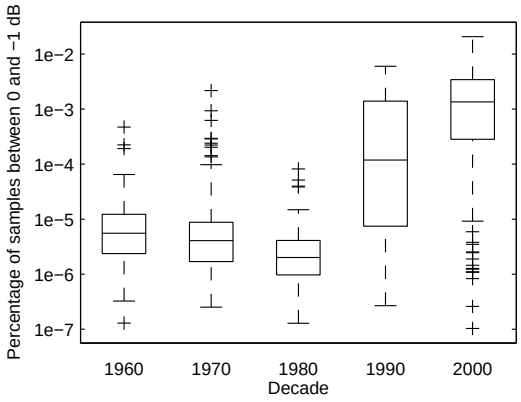

Figure 4: Percentage of samples between -1 and 0 dB for five different decades.

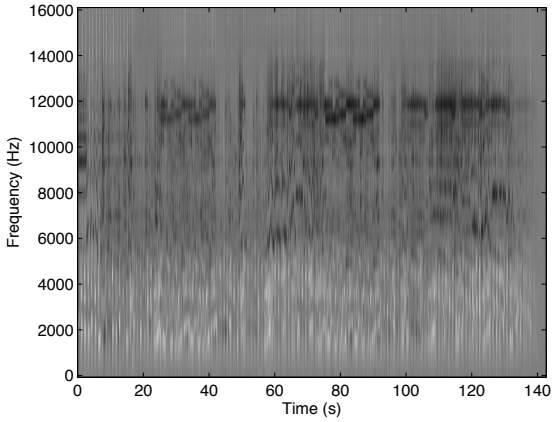

Figure 5: Cochleagram difference for the *While My Guitar Gently Weeps* from *The Beatles*. Color ranges from -.3 (white) representing right channel to .3 (black) representing left channel.

or median and also by a lower skewness (the mass of the distribution is concentrated on the high values). Fig. 3 shows the mean of the distribution over decades. The observation is the same as on the histogram, showing an increasing use of compression from the 90s.

3. STEREO AND PANNING

The second group of techniques that we study relates to the differences that are observed between the left and right channels of a stereo recording. This panel of techniques results in a variety of signals, that can range from mono ones, for which there is no difference between the two channels, to complex stereo images produced by using amplitude and phase differences. We present two measures that intent to describe the differences between the two channels.

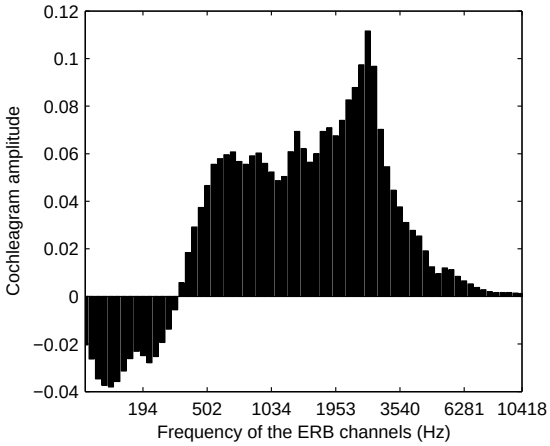

Figure 6: Mean of cochleagram difference for the song *While My Guitar Gently Weeps* from *The Beatles*. Negative values indicate right channel, positive values indicate left channel.

3.1. Amplitude panning

3.1.1. Cochleagram differences

Amplitude panning consists in distributing the sound of each sources on each channel. Avendano [8] proposes a method to measure the differences between left and right channels. He computes a normalized similarity measure between left and right channel spectrograms. We use a slightly different measure based on channel cochleagrams. The cochleagram represents the excitation pattern of the basilar membrane. We use this method to obtain a more perceptually meaningful representation of the sound. The cochleagram is computed using a gammatone filterbank whose center frequencies follows the ERB scale [13]. The ERB scale is computed as follows:

$$ERB_n = 21.4 log_{10}(0.00437f + 1); \quad (2)$$

where f represent the frequency.
We use a filterbank of 70 filters with frequency centers between 30 Hz and 11025 Hz. To measure the spectral difference between channels over time, we compute the difference between both channel cochleagram. We call this representation *Cochleagram Difference* (CD). Fig. 5 shows the Cochleagram Difference of the song *While My Guitar Gently Weeps* from *The Beatles*. We can clearly see the guitar and the organ (in black) between 500 Hz and 3 kHz that are almost fully panned on the left. In the low frequency we notice (in white) the bass and the drums that are much louder on the right channel. The remaining green color is mainly due to voices that are in the center.

3.1.2. Summary features

To summarize the information contained in this representation we use four features that we will call *Amplitude Stereo Features* (ASF) in the following.

- The global mean over frequency and time of the absolute value. This feature indicate the global amount of panning in the song,
- The standard deviation over frequency of the mean over time of the absolute value. This feature is an indication of the amount of panning variation across time,
- The mean over time which measure the mean panning across frequencies,
- The mean over time of the absolute value which gives the same indication but ignoring the panning direction (left/right),
- The standard deviation over time.

As an illustration, Fig. 6 shows the mean over time of the cochleagram difference for the same song as in Fig. 5. This figure shows that, over the song, bass frequencies are panned on the right (indicated by negative values), while medium and high frequencies are panned on the left (indicated by positive values).

3.2. Phase stereo

In the last two decades, sound engineers have been broadly using mixing techniques based on slight differences between the left and right channel that give a sense of "wideness" to the sources. We will group these techniques under the designation of "phase stereo" as opposed to "amplitude stereo", of which panning is an example. There exist at least three of these techniques. The simplest one is based on an inversion of phase between the two channels. Another one is based on a single original track, that is being panned as it is on one channel, and panned with a short delay (between 10 and 30ms) on the other channel. A third one, sometimes called "double-tracking", consists in recording at least twice the same musical phrase played on the same instrument and to pan each take on a different channel. This method is widely used by heavy metal producers on guitar parts, in order to provide an impression of a "huge" guitar sound. Such techniques are made easy to implement by the precision of track synchronization brought by reliable multi-track recorders, as well as the abundance of available tracks provided by digital recording systems. As a consequence of the use of these mixing techniques, recordings with very few panning can still give a sense of space. To describe these effects we propose a new representation inspired by the phase meters used by sound engineers.

3.2.1. Spectral Stereo Phase Spread (SSPS)

We denote by $s_L(n)$ and $s_R(n)$ the left and right channel of a stereo audio signal over sample n. The common tools used in music production to analyze the stereo de-phasing of an audio signal is named the "phase-meter". It displays over a 2D representations the values $y(n) = s_L(n) - s_R(n)$ (on the ordinate) and $x(n) = s_L(n) + s_R(n)$ (on the abscissa). When the channels L and R are "in phase", $y(n)$ cancels, when they are in phase opposition, $x(n)$ cancels. This is illustrated in Fig. 7. We therefore use the following σ_{LR} to measure the spread of the audio signal due to de-phasing

$$\sigma_{LR} = \frac{\sigma\left(s_L(n) - s_R(n)\right)}{\sigma\left(s_L(n) + s_R(n)\right)} \tag{3}$$

where $\sigma(x)$ denotes the standard deviation of the values x.

As for the Cochleagram Difference (which measures stereo spread in frequency due to amplitude panning), we propose a formulation of the stereo phase spread in the frequency domain. The goal is to obtain a spectral location of the use of de-phasing techniques. For this a Short Time Fourier Transform analysis is first performed using a Blackman window of length 40ms with a 20ms hop size. We denote by $S_L(f_k, m)$ and $S_R(f_k, m)$ the respective short time Fourier complex spectrum at frame m and frequency f_k. The phase components, $\Phi_L(f_k, m)$ and $\Phi_R(f_k, m)$ represents the phase of each cosinusoidal component at frequency f_k and at the beginning of the frame. The phase components $\Phi_L(f_k, m)$ and $\Phi_R(f_k, m)$ over frame m can therefore be considered as an equivalent of $s_L(n)$ and $s_R(n)$. We can therefore compute the same measures $Y(f_k, m) = S'_L(f_k, m) - S'_R(f_k, m)$ and $X(f_k, m) = S'_L(f_k, m) + S'_R(f_k, m)$ using

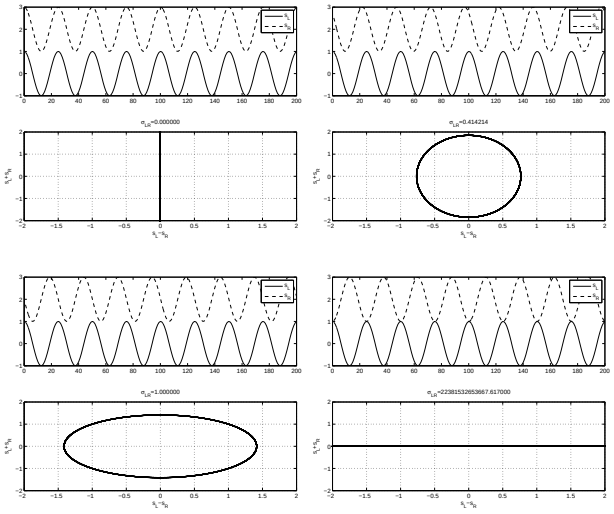

Figure 7: Representation of $s_L(n)$;$s_R(n)$ (top parts) and $y(n)$;$x(n)$ for various case of de-phasing between left and right channels. [top-left]: $\phi = 0$, [top-right]: $\phi = \pi/4$, [bottom-left]: $\phi = \pi/2$ and [bottom-right]: $\phi = \pi$

$$\begin{aligned} S'_L(f_k, m) &= \cos(2\pi f_k/sr + \Phi_L(f_k, m)) \\ S'_R(f_k, m) &= \cos(2\pi f_k/sr + \Phi_R(f_k, m)) \end{aligned} \tag{4}$$

$$\sigma_{LR}(k) = \frac{\sigma\left(S'_L(f_k, m) - S'_R(f_k, m)\right)}{\sigma\left(S'_L(f_k, m) + S'_R(f_k, m)\right)} \tag{5}$$

In order to derive a perceptual measure from $\sigma_{LR}(k)$, we group the values over frequencies f_k into ERB bands.

$$\sigma_{LR}(b) = \sum_{f_k \in \{B\}_k} \sigma_{LR}(k) \tag{6}$$

where $\{B\}_k$ denotes the set of frequency of the b^{th} ERB bands. A further refinement is to weight each value of $\sigma_{LR}(k)$ by the amplitude of the corresponding frequency bin f_k

$$\sigma'_{LR}(b) = \sum_{f_k \in \{B\}_k} A(k)\sigma_{LR}(k) \tag{7}$$

where $A(k)$ is the mean of the contribution of the modulus (amplitude spectrum) $|S_L(f_k, m)|$ and $|S_R(f_k, m)|$.

In Fig. 8, we illustrate the computation of $S'_L(f_k,m) - S'_L(f_k,m)$ and $S'_L(f_k,m) + S'_L(f_k,m)$ for five frequency bands and dephasing of $\phi = 0$, $\phi = \pi$, $\phi = \pi/2$, $\phi = \pi/4$ and $\phi = 0$ in each band.

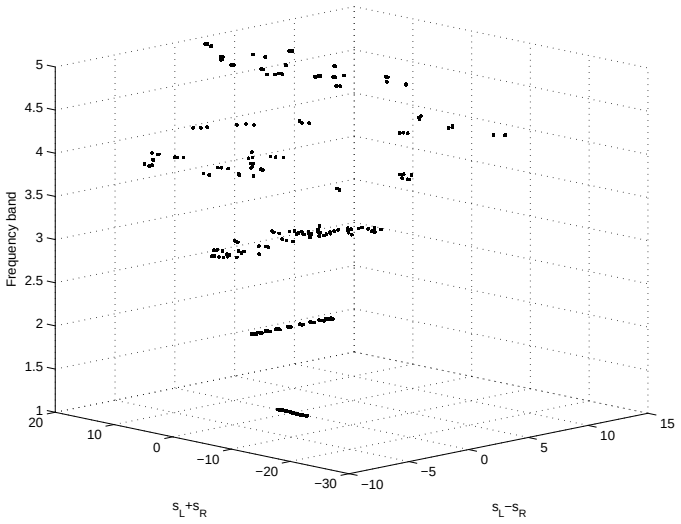

Figure 8: Computation of $\sigma_{LR}(k)$ in the frequency domain.

Fig 9 shows the CD (on the top) and the SSPS (on the bottom) of the song *Gangsta's Paradise* by *Coolio*. Compared to the previous *Beatles*' song, this song presents very few amplitude panning as shown by the almost uniform green color of the CD. In contrast the SSPS shows some very strong variations. The lighter areas of the SSPS corresponds to points in time and frequency where the phase difference between channels is higher. These segments correspond to the entrance of the choir which as been mixed using the *double tracking* technique.

3.2.2. Summary features

To summarize the information contained in SSPS we use four features that we will call *Phase Stereo Features* (PSF) in the following.

- The global mean over frequency and time,
- The standard deviation over frequency of the mean over time,
- The mean over time,
- The standard deviation over time.

4. CLASSIFYING SONGS INTO DECADES

As a proof of concept of our features we propose to automatically classify songs into decades. Since the proposed features are designed to describe production characteristics of the records, and since these characteristics have changed over time, our features should allow to guess the period of production. This kind of classification could be very interesting for measuring the similarity between songs or for automatically generating playlists.

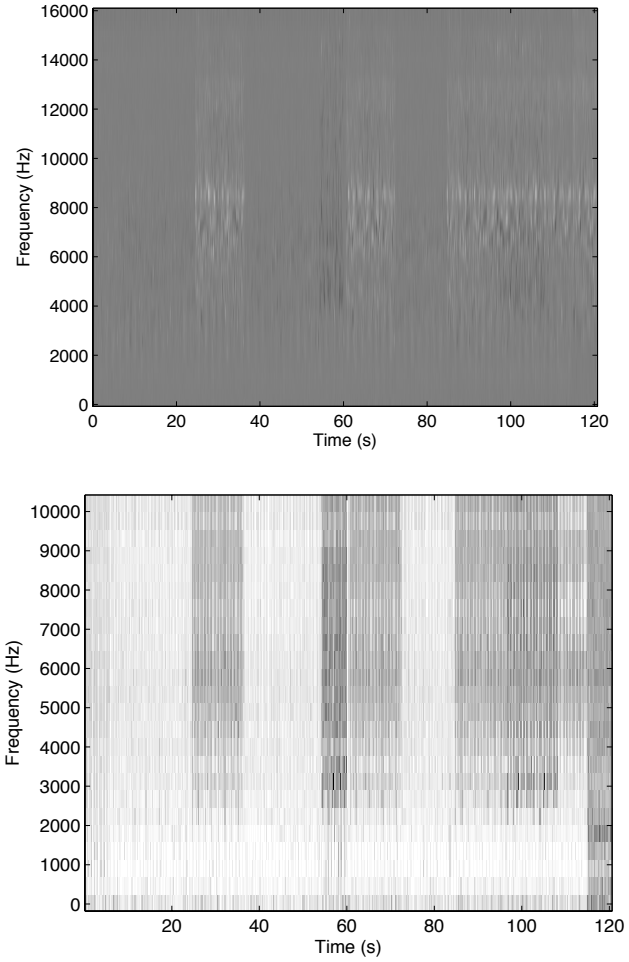

Figure 9: Cochleagram Difference (Top) and Spectral Stereo Phase Spread (Bottom) for the song *Gangsta's Paradise* by *Coolio*. In the Cochleagram Difference, Color ranges from -.3 (white) representing right channel to .3 (black) representing left channel. In the Spectral Stereo Phase Spread lighter colors represent higher phase spread.

4.1. Sound set

We use a set of 1980 Pop/Rock songs by 181 different artists. The set contains 396 songs for each decade. The year were obtained from a metadata database. The set is divided into a train set of 1042 songs and a test set of 938 songs. To avoid over-fitting of the models due to the album effect, the train and test sets contains different artists.

4.2. Classification method

As a classifier, we use support vector machines (SVM) with a Gaussian radial basis function kernel. We set $\gamma = 1/d$ [14] where d is the dimension of the feature set and $C = 1$. The implementation is the one of LIBSVM [15]. To make a multi-class classifier from the 2-class SVM we use the one versus all method. We train a classifier for each class versus all the remaining classes. To make a decision we compare the posterior probabilities provided

Features				Performance
DF	ASF	PSF	MFCC	Accuracy
		×		0,39
	×			0,46
			×	0,47
×				0,47
	×	×		0,51
×	×	×		0,61
×	×	×	×	0,64

Table 1: Classification accuracy for various feature combinations. DF=Dynamic Features, ASF=Amplitude Stereo Features, PSF=Phase Stereo Features, MFCC=Mel Frequency Cepstral Coefficients

		Classified as					
		1960	**1970**	**1980**	**1990**	**2000**	recall
real class	**1960**	125	23	5	1	0	0,81
	1970	28	111	78	12	7	0,47
	1980	1	30	152	5	1	0,80
	1990	16	36	43	125	24	0,51
	2000	5	13	11	29	161	0,74
	precision	0,71	0,52	0,53	0,73	0,83	

Figure 10: Confusion matrix for the classification with all the features

by LIBSVM and affect the class with the highest probability to the incoming data.
We compare the results of the proposed features either separately or grouped. For comparison purposes we added the Mel Frequency Cepstral Coefficients (MFCC) that are widely used features for spectral envelope description.

4.3. Results

Tab. 1 shows the classification results for various feature combinations. First, we see that every features carry information about decade, the best one being the dynamic features with an accuracy of 0.47. An interesting observation is that the two kind of stereo features (amplitude and phase) perform better when used together (.51) than separately (respectively .39 and .45), showing that they carry different kind of information. When all the features are used in conjunction we obtain a score of .64. Tab. 10 shows the confusion matrix of this last case. As expected, the main confusions occurs between adjacent decades. The 00s obtain the best recognition rate (.83) followed by the 90s and 60s (resp. .73 and .71). Confusion occurs more often between 70s and 80s.

5. CONCLUSION

In this paper, we presented three innovative audio features to describe the characteristics of the music production effect. These features are related to dynamic range and stereo mixing. Dynamic features pointed out the increasing use of compressors and limiters across decades. Stereo features were shown to be able to characterize both amplitude panning and phase stereo. The relevance of the features was tested in a task of automatic decade classification of music tracks. An accuracy of 60% on a five decade task was reached using our features. While such classification can be useful for automatic song tagging or for music similarity, it could be interesting to try regression methods to estimate more precisely the within decade period of production. Also, since the production techniques can vary across genres, further research should focus on possible variations of our features across genres.

6. ACKNOWLEDGEMENT

This work was supported by French Oseo Project QUAERO.

7. REFERENCES

[1] Virgil Moorefield, *The Producer as Composer: Shaping the Sounds of Popular Music*, The MIT Press, 2005.

[2] François Delalande, *Le son des musiques entre technologie et esthétique*, INA-Buchet/Chastel, Paris, 2001.

[3] Sue Sillitoe, "Recording Cher's 'Believe'," *Sound on Sound magazine*, Feb. 1999.

[4] Emmanuel Deruty, "Archetypal vocal setups in studio-based popular music," in *EIMAS*, Rio de Janeiro, Brazil, 2010.

[5] Etan Smith, "Even Heavy-Metal Fans Complain That Today's Music Is Too Loud!!!," *The Wall Street Journal*, Sept. 29, 2008.

[6] MPEG-7-Audio-Amendment-1, "Information Technology - Multimedia Content Description Interface - Part 4: Audio," ISO/IEC FDIS 15938-4/A1, ISO/IEC JTC 1/SC 29.

[7] George Tzanetakis, Randy Jones, and Kirk Mc Nally, "Stereo panning features for classifying recording production style," in *International Symposium on Music Information Retrieval*, Vienna, Austria, 2007.

[8] Carlos Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2003, pp. 55–58, IEEE.

[9] Youngmoo E. Kim, Donald S. Williamson, and Sridhar Pilli, "Towards quantifying the album effect in artist identification," in *ISMIR*, Victoria, Canada, 2006, pp. 393–394.

[10] Nicolas Scaringella, *On the Design of Audio Features Robust to the Album-Effect for Music Information Retrieval*, Ph.D. thesis, 2009.

[11] Sarah Jones, "Dynamics are Dead, Long Live Dynamics-Mastering Engineers Debate Music's Loudness Wars,," 2005.

[12] John M. Eargle, *Handbook of Recording Engineering*, Springer, 1996.

[13] Brian R. Glasberg and Brian C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[14] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik, "Extracting Support Data for a Given Task," in *Proceedings ot the First International Conference on Knowledge Discovery & Data Mining*. 1995, pp. 252–257, AAAI Press.

[15] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1—-27:27, 2011.