Statistical Distribution of Common Audio Features

Encounters in a heavy-tailed universe

Martín Haro Berois

TESI DOCTORAL UPF / 2013

Director de la tesi

Dr. Xavier Serra i Casals,
Department of Information and Communication Technologies

Universitat
Pompeu Fabra
Barcelona

To Ximena

Acknowledgements

More than ten years ago I was living in a country devastated by an economic crisis and, despite I had done everything “right”, the future ahead was far from optimistic. Thus, I decided that it was the perfect time to pursue my craziest unrealistic dreams. Those dreams included moving to Barcelona (more than 10,000 km away) and joining the Music Technology Group (MTG) at the Universitat Pompeu Fabra. Today, I am still amazed by the fact that I am here, finishing my PhD at the MTG.

In these few lines I would like to express my eternal gratitude to some of the wonderful people who helped me throughout this amazing journey (and apologize to those I forgot to mention). First of all, I would like to thank Xavier Serra for his guidance and for giving me the opportunity to join the MTG, first as a master student, and later on as PhD candidate. I would also like to thank Perfecto Herrera for his constant advice, and selfless help. His broad multidisciplinary knowledge always motivated me, and many other students, to go beyond highly transited paths and to look at problems from different perspectives. Furthermore, Perfe’s guidance goes beyond the pure academic field and enters into other areas where he is also an inspiring role-model of humility, enthusiasm, and passion for music. Another very important person I would like to thank is Joan Serrà. Besides being one of the most brilliant researchers I know, Joan is always willing to share his knowledge and collaborate to improve one’s own work. His characteristic mixture of enthusiasm and hard work has been a constant source of inspiration to me. I would also like to mention the remarkable help of Álvaro Corral with respect to heavy-tailed distributions and complex systems. Without the constant support and academic excellency of the above mentioned people this thesis would not be possible.

I would also like to thank Dimitry Bogdanov, Ferdinand Fuhrmann, Emilia Gómez, Piotr Holonowicz, Jordi Janer, Stefan Kersten, Ricard Marxer,

Waldo Nogueira, Gerard Roma, Justin Salamon, Inês Salselas, Mohamed Sordo, and Anna Xambò for all these years of inspiring discussions and clever suggestions. I am also grateful to Lydia Garcia, Cristina Garrido, and Alba Rosado for being always willing to help. I also like to mention Marcos Alonso, Jordi Bonada, Graham Coleman, Daniel Gallardo, Günter Geiger, Enric Guaus, Amaury Hazan, Sergi Jordà, Carles Julià, Martin Kaltenbrunner, Cyril Laurier, Oscar Mayor, Hendrik Purwins, Rafael Ramírez, Nicolas Wack, and the rest of MTG members for their help and inspiration. Moreover, I would specially like to thank all my co-authors from whom I have learned a lot, and Álex Mas for his support during this last year of writing.

Finally, I would like to thank all my family and friends, specially to Ximena for all these amazing years, to my parents Nibia and Màximo for their unconditional love and support, to Pablo and Stella for being always there when you need them, to Lucia and Federico that, together with my undergraduate students, make me trust the future, and to Pocho for his strength and optimistic determination.

Fortunately, my unrealistic dreams are not over yet and now that economic crises are a constant feature of our time, pursuing crazy unrealistic dreams seems to be the only rational thing to do in the future.

This thesis has been carried out at the Music Technology Group, Universitat Pompeu Fabra (UPF) in Barcelona, Spain from September 2008 to August 2013. This work has been supported by an R+D+I scholarship from UPF.

Abstract

In the last few years some Music Information Retrieval (MIR) researchers have spotted important drawbacks in applying standard successful-in-monophonic algorithms to polyphonic music classification and similarity assessment. Noticeably, these so called “Bag-of-Frames” (BoF) algorithms share a common set of assumptions. These assumptions are substantiated in the belief that the numerical descriptions extracted from short-time audio excerpts (or frames) are enough to capture relevant information for the task at hand, that these frame-based audio descriptors are time independent, and that descriptor frames are well described by Gaussian statistics. Thus, if we want to improve current BoF algorithms we could: i) improve current audio descriptors, ii) include temporal information within algorithms working with polyphonic music, and iii) study and characterize the real statistical properties of these frame-based audio descriptors. From a literature review, we have detected that many works focus on the first two improvements, but surprisingly, there is a lack of research in the third one. Therefore, in this thesis we analyze and characterize the statistical distribution of common audio descriptors of timbre, tonal and loudness information. Contrary to what is usually assumed, our work shows that the studied descriptors are heavy-tailed distributed and thus, they do not belong to a Gaussian universe. This new knowledge led us to propose new algorithms that show improvements over the BoF approach in current MIR tasks such as genre classification, instrument detection, and automatic tagging of music. Furthermore, we also address new MIR tasks such as measuring the temporal evolution of Western popular music. Finally, we highlight some promising paths for future audio-content MIR research that will inhabit a heavy-tailed universe.

Resumen

En el campo de la extracción de información musical o *Music Information Retrieval* (MIR), los algoritmos llamados *Bag-of-Frames* (BoF) han sido aplicados con éxito en la clasificación y evaluación de similitud de señales de audio monofónicas. Por otra parte, investigaciones recientes han señalado problemas importantes a la hora de aplicar dichos algoritmos a señales de música polifónica. Estos algoritmos suponen que las descripciones numéricas extraídas de los fragmentos de audio de corta duración (o *frames*) son capaces de capturar la información necesaria para la realización de las tareas planteadas, que el orden temporal de estos fragmentos de audio es irrelevante y que las descripciones extraídas de los segmentos de audio pueden ser correctamente descritas usando estadísticas Gaussianas. Por lo tanto, si se pretende mejorar los algoritmos BoF actuales se podría intentar: i) mejorar los descriptores de audio, ii) incluir información temporal en los algoritmos que trabajan con música polifónica y iii) estudiar y caracterizar las propiedades estadísticas reales de los descriptores de audio. La bibliografía actual sobre el tema refleja la existencia de un número considerable de trabajos centrados en las dos primeras opciones de mejora, pero sorprendentemente, hay una carencia de trabajos de investigación focalizados en la tercera opción. Por lo tanto, esta tesis se centra en el análisis y caracterización de la distribución estadística de descriptores de audio comúnmente utilizados para representar información tímbrica, tonal y de volumen. Al contrario de lo que se asume habitualmente, nuestro trabajo muestra que los descriptores de audio estudiados se distribuyen de acuerdo a una distribución de "cola pesada" y por lo tanto no pertenecen a un universo Gaussiano. Este descubrimiento nos permite proponer nuevos algoritmos que evidencian mejoras importantes sobre los algoritmos BoF actualmente utilizados en diversas tareas de MIR tales como clasificación de género, detección de instrumentos musicales y etiquetado automático de música. También nos permite proponer nuevas tareas tales como la medición de la evolución temporal de la música popular occidental. Finalmente, presentamos algunas prometedoras líneas de investigación para tareas de MIR ubicadas, a partir de ahora, en un universo de "cola pesada".

Resum

En l'àmbit de la extracció de la informació musical o *Music Information Retrieval* (MIR), els algorismes anomenats *Bag-of-Frames* (BoF) han estat aplicats amb èxit en la classificació i avaluació de similitud entre senyals monofòniques. D'altra banda, investigacions recents han assenyalat importants inconvenients a l'hora d'aplicar aquests mateixos algorismes en senyals de música polifònica. Aquests algorismes BoF suposen que les descripcions numèriques extretes dels fragments d'àudio de curta durada (frames) son suficients per capturar la informació rellevant per als algorismes, que els descriptors basats en els fragments son independents del temps i que l'estadística Gaussiana descriu correctament aquests descriptors. Per a millorar els algorismes BoF actuals doncs, es poden i) millorar els descriptors, ii) incorporar informació temporal dins els algorismes que treballen amb música polifònica i iii) estudiar i caracteritzar les propietats estadístiques reals d'aquests descriptors basats en fragments d'àudio. Sorprenentment, de la revisió bibliogràfica es desprèn que la majoria d'investigacions s'han centrat en els dos primers punts de millora mentre que hi ha una mancança quant a la recerca en l'àmbit del tercer punt. És per això que en aquesta tesi, s'analitza i caracteritza la distribució estadística dels descriptors més comuns de timbre, to i volum. El nostre treball mostra que contràriament al què s'assumeix, els descriptors no pertanyen a l'univers Gaussià sinó que es distribueixen segons una distribució de "cua pesada". Aquest descobriment ens permet proposar nous algorismes que evidencien millores importants sobre els algorismes BoF utilitzats actualment en diferents tasques com la classificació del gènere, la detecció d'instruments musicals i l'etiquetatge automàtic de música. Ens permet també proposar noves tasques com la mesura de l'evolució temporal de la música popular occidental. Finalment, presentem algunes prometedores línies d'investigació per a tasques de MIR ubicades a partir d'ara en un univers de "cua pesada".

Contents

List of Figures

List of Tables

Acronyms

BIN: Binary weighting strategy
BoC-W: Bag-of-code-words
BoF: Bag-of-frames
CAL500: Computer audition lab 500-song dataset
CBA: Code bernulli average
CBDC: Class-based distance classifier
EFD: Equal frequency discretization
EM: Expectation maximization
EWD: Equal width discretization
FFT: Fast Fourier transform
GMM-MH: Gaussian mixture models with mixture hierarchies
GMM: Gaussian mixture models
HPCP: Harmonic pitch class profile
IR: Information retrieval
ITA: Information-theoretic algorithms
K-NN: K-nearest neighbors
KS: Kolmogorov-Smirnov
LFD: Last.fm dataset
ML: Maximum likelihood
MFCC: Mel-frequency cepstral coefficients
MIR: Music information retrieval
MIREX: Music information retrieval evaluation exchange
MSD: Million song dataset
MSD-Loudness: Million song dataset loudness descriptor
MSD-Pitch: Million song dataset pitch descriptor
MSD-Timbre: Million song dataset timbre descriptor
SE: Spectral energy
SVM: Support vector machine
TF-IDF: Term frequency-inverse document frequency
TF: Term frequency
VQ: Vector quantization

Chapter 1

Introduction

Music in digital form is nowadays so easily accessible that both, personal collections, and on-line repositories have grown to a size that increasingly poses difficulties for music users who want to navigate throughout this overwhelming amount of data (Casey et al., 2008). Thus, with the aim of fulfilling users' music information needs, a new multidisciplinary field of research known as Music Information Retrieval (MIR) has steadily grown during the last fifteen years (Orio, 2006).

Within the above MIR field description "music users" are considered in a broad sense, that is, music users are not only casual music listeners, but also professional users, such as sound engineers, music critics, musicologists, music teachers, music artists, cognitive scientists, psychologists, etc. Moreover, a myriad of multidisciplinary techniques from music, computer science, signal processing, cognition, and information retrieval are constantly being used and adapted by MIR researchers. Therefore, as stated by Herrera et al. (2009), and Serra et al. (2013) the MIR field is more about *music information research* than just the retrieval of music information as its own name suggests. From our perspective, the term "music" should be also considered in a broad sense including past and present music traditions from around the world, soundscape recordings, sound effects, etc.

Throughout the short MIR history, multiple problems have been addressed such as: Automatic Genre Classification, Automatic Music Recommendation, Automatic Music Transcription, Instrument Recognition, Music Similarity, Automatic Music Tagging, Cover Song Identification, Melody Extraction, Artist Identification, Music Summarization, and Structure Segmentation (Klapuri and Davy, 2006; Li et al., 2011; Müller, 2007).

In order to extract relevant information from music, MIR researchers count with many data sources like: editorial metadata, user-generated context, symbolic representations, and audio content. However, most of the of the proposed techniques and MIR systems rely on audio content solely (Serra et al., 2013). This audio-centric preference within MIR approaches is substantiated on the assumption that, for some of the above mentioned tasks, other data sources are either not suitable, or unreliable, or missing (Orio, 2006). The main strength of content-based systems is their exclusive dependence on the actual music file (i.e. there is no need for other, possibly unreliable, data sources for the system to work). On the other hand, the main weakness of such systems is that extracting high-level concepts that users use to relate with music collections from analyzing the audio signal alone is extremely difficult. This gap between signal-extracted music descriptors and high-level concepts, like evoked emotions and memories, cultural references, etc., has been denoted within MIR literature as the "semantic gap" (Celma et al., 2006).

In order to make music data comparable and algorithmically accessible (i.e. able to be processed by digital computers), instead of working directly with the original audio samples, content-based MIR algorithms start by extracting suitable features that capture relevant key aspects while suppressing irrelevant details or variations (Müller, 2007). Evidently, the distinction between relevant audio descriptors and irrelevant data depends on the task at hand, and is not always easy to determine. For instance, let's consider an audio feature that numerically describes the tonal characteristics (i.e. the harmonic content) of a piece of music. This descriptor would be very relevant for tasks like melody extraction or cover song identification, but it would be considered as irrelevant for an algorithm trying to retrieve songs with similar rhythms.

Fig. 1.1 shows a canonical content-based MIR algorithm for music classification. Starting with the audio file, a set of audio features is computed over consecutive short-time audio segments (or frames). These frame-based features (usually with lengths below 100 ms) can by computed directly from the signal's time domain (Fig. 1.1a) or, more often, from its frequency domain (Fig. 1.1b), usually obtained via the Fast Fourier Transform (Klapuri and Davy, 2006). Thus, from each short-time audio segment a numerical description of the audio frame is obtained (Fig. 1.1c). Moreover, the

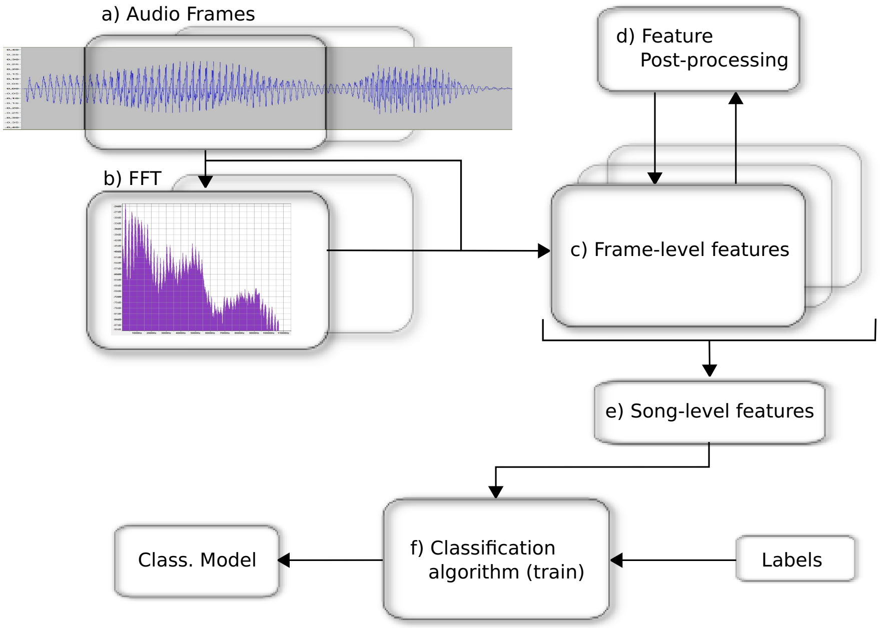

Figure 1.1: Block-diagram of a canonical MIR algorithm for automatic audio classification. Frame-based audio features (c) are computed from the signal's time-domain (a) and/or frequency domain (b). These frame-level featured can be post-processed to obtain new features (d). Next, all frame-level features are aggregated into a song-level feature vector (e). Finally, song-level features from several songs and their corresponding ground truth labels are used to train a classification algorithm (f). This process generates a classification model that is later used to label unseen audio files.

sequence of consecutive frame-based features within a song form a multi-dimensional time series that numerically describes the temporal evolution of relevant characteristics of the song such as its energy, timbre, melody, tonality, tempo, etc. Of course, not all features are extracted in one step (Fig. 1.1d). For instance, a first step could extract the energy of each audio frame, then, a second step could detect those frames that correspond to energy peaks. Finally, these energy peaks could be used to determine note onsets, and these onsets could also be considered as a relevant audio feature by, for instance, a segmentation algorithm.

After the frame-level feature extraction process, a standard approach in song classification is to generate a song-level feature vector that summarizes the frame-level sequences (Fig. 1.1e). This is usually done by computing the first statistical moments such as mean and variance of the frame-level sequences. These song-level feature vectors, together with manually annotated labels that provide information about the class each song belongs to, are thus used as input for training the classification algorithm (Fig. 1.1f). Alternatively, a similarity-based classification strategy can be adopted. This strategy has mainly two approaches. In the first approach a similarity measure (such as cosine or Euclidean distance) between song-level feature vectors of to-be classified songs and song-level feature vectors of previously labeled songs is used to assign the class membership. In the second approach, the between-song distance is computed directly from frame-level feature vectors without the need for a summarization step. Thus, the similarity measure is usually obtained by modeling the distribution of local feature vectors in the feature set with some probability model, such as Gaussian Mixture Models (GMM). Hence, different songs can be compared according to their underlying probability models (Fu et al., 2011).

At the end, all the above described approaches discard the frame-by-frame temporal information, i.e. the frames' temporal ordering. Thus, these algorithms are frequently called "Bag-of-Frames" (BoF) algorithms[1] (Aucouturier et al., 2007; Casey et al., 2008; Marques et al., 2011b; Müller et al., 2011; Quatieri, 2001).

Noticeably, by using audio descriptors, and by discarding temporal information, the BoF algorithms end up working with a rough representation of the audio content. This rough summarization has provided excellent results when working with monophonic audio signals and soundscape recordings, but unfortunately, it seems not-so-adequate when working with polyphonic music (Aucouturier et al., 2007). Among the many causes that could explain this fact, a direct analysis of the canonical BoF process lead us to suspect that there could be three major problems with the BoF approach when working with polyphonic music. The first problem could be that current audio descriptors are not able to properly "describe" polyphonic music (Marques et al., 2010). Thus, we need better music descriptors. The second problem could be that, since we are working with music, and temporal relationships are one of the key ingredients that conforms the music

[1] In information retrieval, those algorithms that discard the temporal relationships between words in text documents are called "Bag-of-Words" methods, thus, the "Bag-of-Frames" term reflects the same behavior but with respect to audio frames.

discourse, it is a mistake to discard this temporal information when working with music (Casey and Slaney, 2006). Hence, we need to somehow include temporal information within BoF algorithms. Finally, the third problem could be that, since we are summarizing frame-level information or computing distance measures within pre-defined feature spaces, it could be the case that we are working with wrong assumptions with respect to this feature space. Thus, we need to better understand where audio features "live" and find more adequate ways to work within these spaces (Marques et al., 2011b). Noticeably, many MIR researchers (including ourselves) have focused on the first and second problem, and we can find a great number of publications proposing new music descriptors and incorporating temporal information within the BoF summarization process (Fuhrmann et al., 2009; Haro and Herrera, 2009; Joder et al., 2009; Lyon et al., 2010; Pachet and Roy, 2009). However, we have found a lack of research efforts in trying to tackle the third problem: understanding the audio feature space. Therefore, the main goal of this thesis will be to characterize the statistical properties of common audio features, and use this information to improve over the standard BoF approach.

1.1 Motivation

When analyzing the underlying assumptions made by BoF algorithms, we observe that these approaches rely on a certain homogeneity in the feature vector space. That is, the multidimensional space of feature values should not have small areas that are extremely populated and, at the same time, extensive depopulated regions. Otherwise, the results obtained from computing statistical moments (such as mean and variance), or from computing distance measures, or modeling with Gaussians, will be highly biased towards the values of those extremely populated areas (i.e. those extremely frequent feature vectors frames).

Interestingly, in other research areas such as natural language processing (Manning and Schütze, 1999) and Web mining (Liu, 2011), the distribution of words and hyperlinks has shown to be heavy-tailed, implying that there are few extremely frequent words/hyperlinks and many rare ones. Knowing the presence of such heavy-tailed distributions has lead to major improvements in technological applications in those areas. For instance, to Web search engines that use the word probability distributions to determine the relevance of a text to a given query (Baeza-Yates, 1999). Recently, these type of text categorization techniques have also been successfully applied in image retrieval (Jiang et al., 2010). Unfortunately, there is a lack of research in the MIR community about the statistical distribution of sound descriptors, and furthermore, the above mentioned homogeneity assumption is adopted without further concerns. This lack of research could be partially substantiated by the fact that sound descriptors do not form discrete units or symbols that can be easily characterized by their frequency of use, as it is the case with, for instance, text or even visual objects.

Meanwhile, in the last few years, MIR researchers have detected some persistent, and possibly related, problems with respect to BoF algorithms. These problems are: the "glass ceiling" (Aucouturier and Pachet, 2004), and the appearance of "hub" songs (Aucouturier and Pachet, 2008; Schnitzer et al., 2012). In the first case, it seems that regardless of the algorithm configuration, empirical classification results are always below an upper-limit performance called the "glass ceiling". That is, despite using different sets of audio features, different machine learning algorithms, and different parameter sets, classification results do not show substantial improvements. Moreover, this upper-bound prevents content-based algorithms to be massively used in commercial applications. In the second case, usually described for distance-based algorithms, some "hub" songs persistently, and irrelevantly, appear in the nearest neighbor lists of other songs. That is, when working with big datasets, some songs are always ranked by the algorithm as being similar to many other songs, but this alleged similarity is not corroborated by human subjects listening to these songs.

In order to overcome these problems some authors have stressed the importance of adding high-level features, using source separation algorithms, or relying on semi-automatic methods (Benetos et al., 2012; Bogdanov et al., 2011). Moreover, other authors have stressed the need for cognitive-based algorithms within MIR (Aucouturier and Bigand, 2013; Wiggins, 2009).

We completely agree that adding more data sources either coming from high-level features, user-provided context, or cognitive models could be of tremendous help and should be further investigated, but we also believe that the knowledge extracted from the systematic analysis of audio features within large music databases could be of great help too. In particular, the work presented in this thesis shows that the study and characterization of the statistical properties of standard audio descriptors helps us to improve current BoF algorithms, and unveils new research paths that exploit the acquired knowledge to go beyond standard MIR approaches.

1.2 Scope and aims of the thesis

In this thesis we exclusively focus on content-based MIR, that is, the proposed analyses and algorithms rely on raw audio files without depending on other data sources. Our work can be also characterized as a “bottom-up” data-driven approach (Casey et al., 2008). In particular, we analyze the statistical distributions of encoded audio descriptors related to three main musical facets such as timbre, pitch, and loudness. We put special emphasis in working with large datasets of real-world polyphonic music. Moreover, we propose new content-based algorithms that take advantage of the found distribution patterns of descriptors to contribute on current MIR tasks such as genre classification, instrument detection, or automatic tagging of songs. Furthermore, we also use the acquired knowledge to address new MIR tasks such as measuring the temporal evolution of Western popular music.

1.3 Main contributions

The main contributions of this thesis can be summarized as follows:

1. We have proposed a simple strategy to encode multidimensional audio descriptors into a dictionary of pre-defined code-words.

2. We have used the proposed encoding strategy to characterize the frequency distributions of common audio descriptors of timbre, chroma, and energy as being heavy-tailed distributed.

3. We have found a simple, parsimonious generative model that could be involved in the generation process of such heavy-tailed distributions.

4. We have proposed new audio features that take advantage of the found distributions.

5. We have evaluated the proposed features within new algorithms that contribute to current content-based MIR tasks and improve over the standard Bag-of-Frames approach.

6. We have illustrated the possibilities of our approach for shedding some light on musicological problems such as describing the temporal evolution of music audio content. In particular, we have studied the temporal evolution of popular Western music from 1955 to 2010 by

measuring the long-term trends within the distribution of the proposed code-words.

7. A moderate contribution to statistical physics is also included as we have unveiled new power-law distributions coming from human-made and natural sounds.

1.4 Outline

In Chapter 2 we introduce the scientific background of this thesis. Next, in Chapter 3 we study and characterize the rank-frequency distribution of audio timbral descriptors for sounds coming from speech, natural sounds, and music from Western and non-Western music traditions. We conclude that timbral audio features follow a power-law distribution. Hence, in Chapter 4 we investigate on plausible generative mechanisms that could be involved in the process of producing the found heavy-tailed distributions. In Chapter 5 we perform a series of experiments for genre and musical instrument detection that provide further evidence that timbral features from individual recordings have the same type of heavy-tailed distribution as found in large-scale databases. In Chapter 6 we characterize the statistical distribution of other key audio features related with timbre, tonal, and energy information as being also heavy-tailed. Moreover, we exploit the found distributions to present new tools for the objective measurement of the evolution of popular Western music from 1955 to 2010. In Chapter 7 we propose new audio features that take advantage of the results presented in previous chapters, and evaluate these new descriptors for the complex task of automatic tagging of music. Finally, Chapters 8 and 9 present a general discussion and promising research directions to continue the here presented work. We also include a series of appendices that present further information regarding the used databases, distribution functions, fitting procedures, and other complementary information for the presented experiments.

Chapter 2

Background

2.1 Heavy-tailed distributions

Many times when we measure human-made and natural phenomena such as air pressure, the height of adult male chimpanzees, sea level, the actual weights of 1 Kg rice bags produced by a particular company, etc. we observe that the measured values vary around some typical number. That is, if we build a histogram with the registered measures we see a characteristic "bell-shaped" (or Gaussian) distribution with the majority of observations clustered around one value. Moreover, we observe that even the largest deviations from this typical value are, not only extremely rare, but also not farther that a factor of two from the majority of observations. In these cases we can describe the main characteristics of the distribution by quoting its mean and standard deviation values (Clauset et al., 2009).

However, not all distributions follow the aforementioned pattern. There are cases when the range of observed values seem to be unbounded, with distribution shapes presenting "heavy-tails". This means that the measured data points are spread over an extremely wide range of possible values, and that there is no typical quantity around which these measurements are centered (Newman, 2005). It also implies that the majority of data points (i.e. the ones in the tail) do not occur frequently (see Appendix B for further information about the heavy-tailed distributions used in this thesis).

A particularly important landmark regarding heavy-tail distributions was the seminal work of Zipf (1949), showing a power-law distribution of word-frequency counts with an exponent α close to 1,

$$z(r) \propto r^{-\alpha}, \tag{2.1}$$

where r corresponds to the rank number ($r = 1$ is assigned to the most frequent word) and $z(r)$ corresponds to the frequency value of the word with rank r. The rank-frequency power-law described by Zipf (Eq. 2.1) also indicates a power-law probability distribution of word frequencies (Adamic and Huberman, 2002),

$$P(z) \propto z^{-\beta}, \tag{2.2}$$

where $P(z)$ is the probability mass function of z and $\beta = 1 + 1/\alpha$.

Remarkably, power-law distributions have been reported in many scientific disciplines such as physics, engineering, computer science, geoscience, biology, economics, linguistics, and social sciences (Adamic and Huberman, 2002; Bak, 1996; Malamud, 2004; Newman, 2005; Zipf, 1949). Moreover, as stated by Clauset et al. (2009), these type of distributions, once regarded as problematic or defective, constitute nowadays one of the most interesting of all scientific observations. Noticeably, power-law distributions have been described in diverse natural and human-made phenomena such as: city sizes (Decker et al., 2007; Simon, 1955), word frequencies (Corominas-Murtra et al., 2011; Ferrer i Cancho and Solé, 2003; Zipf, 1949), Internet file sizes (Reed and Hughes, 2002), the number of visitors on web pages (Adamic and Huberman, 2002), earthquake sizes (Gutenberg and Richter, 1944), moon craters (Neukum and Ivanov, 1994), the numbers of species in biological taxa (Willis and Yule, 1922), rain event size distributions (Peters et al., 2010), and long-term rate adaptations between the inner hair cell and auditory nerve synapse (Zilany et al., 2009).

One of the most interesting behaviors of power-law distributions is the linear relationship that appears when logarithms are applied to both sides of the power-law equation. Thus, when plotting for instance, $z(r)$ vs. r from Eq. 2.1 in logarithmic axes, the graph will show a characteristic straight-line over several orders of magnitude with the negative exponent α depicted in the negative slope of the curve (see Fig. 2.1). This linear relationship reflects the most remarkable attribute of power-laws namely their *scale invariance*, that is, if we multiply the r variable in Eq. 2.1 by a constant factor c this will produce a proportional scaling on the function $z(r)$. Therefore,

$$z(cr) \propto cr^{-\alpha} = c^{-\alpha} z(r) \propto z(r). \tag{2.3}$$

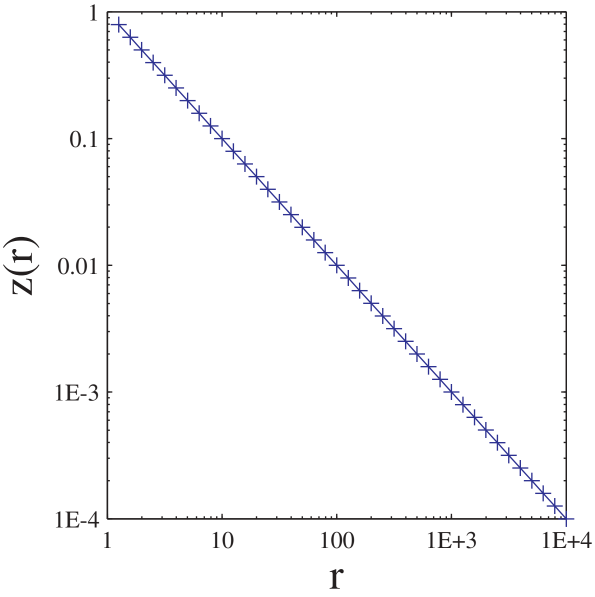

Figure 2.1: Theoretical power-law distribution of word-frequency counts in English texts as reported by Zipf (1949). r corresponds to the word rank number ($r = 1$ is assigned to the most frequent word) and $z(r)$ corresponds to the (normalized) frequency value of the word with rank r. The α exponent is 1, and is reflected by the negative slope of the curve that corresponds with the $-\alpha$ exponent in Eq. 2.1. Empirical data would not show so clean behavior, specially for low rank frequencies.

Hence, given its scale invariance property all power-laws that share the same exponent are equivalent up to constant factors (i.e. each power-law is a scaled version of the others that share the same exponent). This interesting characteristic also links power-law functions with Mandelbrot's fractal geometry (Mandelbrot, 1982).

It is worth to mention here that observing a straight-line in the log log plot is not sufficient condition to claim a power-law behavior. Moreover, even commonly used methods for data analysis, such as least-squares fitting, are error-prone when trying to evaluate if a power-law fits the observed data. Fortunately, the excellent work by Clauset et al. (2009) provides more accurate ways for detecting power-law distributions in empirical data (see also Appendix C for further detail regarding fitting procedures used in this thesis).

Even though a unifying principle that explains why power-law distributions emerge from such a variety of complex systems has not been found yet, major improvements in data analysis and engineering applications have already taken place thanks to the observation and characterization of such heavy-tailed distributions. For instance, as mentioned in Chapter 1, research on statistical analysis of natural languages (Manning and Schütze, 1999) facilitated applications such as text retrieval based on keywords, where the word probability distributions are used to determine the relevance of a text to a given query (Baeza-Yates, 1999). Furthermore, researchers within the image retrieval field have also developed algorithms for detecting images' *regions of interest* by means of exploiting power-law models (Caron et al., 2007).

With respect to music, Zipf himself reported power-law distributions in melodic intervals and distances between note repetitions from selected music scores (Zipf, 1949). Since then, several works have shown heavy-tailed distributions of data extracted from symbolic representations of music such as scores (Hsü and Hsü, 1990, 1991; Levitin et al., 2012; Telesca and Lovallo, 2012) and MIDI files (Beltrán del Río et al., 2008; Manaris et al., 2005; Zanette, 2006)[1]. However, unlike text retrieval, music retrieval has not directly benefited from such observations yet (Zanette, 2008). Indeed, symbolic representations are only available for a small portion of the current and past world's music. Furthermore, they are non-standard and difficult to define for other types of sounds such as human speech, animal vocalizations, and environmental sounds. Hence, it is relevant to work directly with information extracted from the raw audio content. In this line of research, some works can be found describing heavy-tailed distributions of sound amplitudes for crackling noise (Kramer and Lobkovsky, 1996; Sethna et al., 2001), sound amplitudes and pitch (estimated from zero crossing rates) for music and speech signals (Voss and Clarke, 1975), and power spectrum of individual frequency bands for natural sounds, music and speech (Attias and Schreiner, 1997).

[1] MIDI is an industry standard protocol to encode musical information; this protocol does not store sound but information about musical notes, durations, volume level, instrument name, etc.

2.2 Audio descriptors

As mentioned in Chapter 1, one of the cornerstones of MIR is the extraction of relevant audio features or descriptors. These numerical representations of the audio content allow digital computers to work with sound. Hence, in order to fulfill user's information needs, audio descriptors should be linked with the abstractions a listener is able to perceive when listening to music. Thus, relevant descriptors are usually representative of key musical facets such as timbre, harmonic content, perceived energy, rhythm, etc. (Serra et al., 2013).

A great number of MIR descriptors have been proposed or adopted from related fields such as Speech Processing. Usually, audio features are grouped into three main categories according to their level of abstraction (see also Klapuri and Davy (2006), Lesaffre (2006), Müller (2007), and Peeters (2004) for further details regarding audio features). Following Serra et al. (2013), these categories are:

- Low-level features: computed directly from the signal itself. Here we find simple features directly extracted from the signal's time-domain such as zero-crossing rate, or attack time. We also find spectral features extracted from the frequency-domain such as spectral centroid, skewness, and kurtosis, and features based on simple auditory models such as Mel-frequency cepstral coefficients (MFCC), and Bark-band energies.

- Mid-level features: extracted from more complex procedures usually involving task-dependent parameters. These features depart from the audio waveform and aim at describing musical content as a MIDI-like representation including pitches and onset times of individual notes, and melody contours.

- High-level features: represent the highest level of abstraction, and their computation usually involves machine learning algorithms that use low-level and mid-level features as input. These features are related to music users' abstractions and include concepts like genre, instrument, harmony, rhythm, and mood. Automatically computed high-level features are the most unreliable of the three categories, but, on the other hand, offer the most user-friendly information. The development of new and more reliable high-level features is an active

> area of research that constantly tries to bridge the "semantic gap" between low-level features and high-level abstractions as perceived by the listener.

Due to the key role played by low-level features within content-based MIR, and the lack of research regarding their statistical distribution (see Chapter 1), in this thesis, we focus on commonly used low-level features that describe the three main facets related to the perception of short-time audio segments. These facets are: *timbre*, *pitch* and *loudness* (Ball, 2010; Berg and Stork, 1995).

According to the American National Standards Institute (1973), "timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". Moreover, *timbre* mainly correlates with the audio waveform shape and, thus, with the spectro-temporal envelope of the signal (i.e. the temporal evolution of the shape of the power spectrum; Bregman, 1990). Timbre accounts for the sound color, texture, or tone quality, and can be essentially associated with instrument types, recording techniques, and some expressive performance resources.

Pitch is a perceptual attribute that allows sounds to be ordered on a frequency-related scale extending from low to high (Herrera et al., 2006). Pitch basically correlates with the periodicity of air pressure fluctuations (Bregman, 1990). Nevertheless, since there is still no reliable way to extract individual notes from polyphonic music, the most used pitch-related features provide a global description of the harmonic content of an audio segment. Thus, it is more accurate to refer to these descriptors as tonal (or chroma) descriptors[2].

Finally, the sensation of *loudness* is defined by the American National Standards Institute (1973) as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud". Hence, *loudness* correlates with the amplitude of the audio waveform, where sound amplitudes refer to air pressure fluctuations which, when being digitized, are first converted into voltage and then sampled, quantized, and stored in digital format as discrete time series. Thus, loudness descriptors correlate in a non-linear way (because of the particularities of human auditory transduction) with the energy of such digitalized audio signal. Notice

[2] Within this thesis, when referring to audio features we pragmatically use the term pitch as synonym of chroma or tonal descriptors (see Sec. 2.2.2).

that we refer to the intrinsic loudness of a recording, not to the loudness a listener could manipulate by changing the volume control of her audio player.

In the next three sections we provide a detailed explanation about the timbral, tonal and energy descriptors used in this thesis. In particular, we analyze three timbral, two tonal, and two loudness related descriptors. The selection of a particular descriptor within each sound facet will depend on the task at hand, and the availability and type of data sources. For instance, due to copyright issues, public datasets usually consist of collections of pre-computed audio features (see Appendix A). Thus, since we do not have direct access to the original audio files, in these cases we will have to work with the provided audio descriptors. Fortunately, all those descriptors that refer to the same underlying sound facet are expected to provide overlapping information. Hence, as reflected in our experiments, similar results are obtained for descriptors of the same sound facet.

2.2.1 Timbral descriptors

Timbral descriptors aim at describing the spectro-temporal envelope of the audio signal. In particular, it has been shown that "timbre is closely related to the relative level produced at the output of each auditory filter [or critical band of hearing]" (Moore, 2005)[3]. Therefore, timbral descriptors usually characterize timbral sensations by numerically encoding the energy of perceptually motivated frequency bands found in consecutive short-time audio fragments (Müller et al., 2011; Quatieri, 2001). Hence, the timbral content of each audio frame is represented by a multidimensional vector, whereas the timbral information of the full audio file is represented by a multidimensional time series formed by time-ordered frames.

These type of features are of great relevance for several MIR tasks such as genre classification, music recommendation, automatic audio tagging, and instrument identification. Within this thesis, we work with three timbral features namely: Bark-band energies, MFCC, and MSD-Timbre. Next, we provide a detailed explanation of them.

[3] In the auditory filter model, the frequency resolution of the auditory system is approximated by a bank of band-pass filters with overlapping pass-bands.

Bark-band energies

The Bark-band energies descriptor (Zwicker and Terhardt, 1980) rely on the Bark scale (Zwicker, 1961), which is a psychoacoustical scale that subdivides the audible frequency range into critical bands of hearing[4]. This simple descriptor provides a measure of the energy of each Bark-band within the analyzed audio segment (or frame).

The original Bark scale definition considers 24 Bark-bands from 20 to 15,500 Hz (Zwicker, 1961). Nevertheless, the number of used bands for the Bark-band energies descriptor depends on the task at hand. To compute this descriptor, a few steps are need. The first step is to compute the power spectrum of each short-time audio frame by taking the square of the magnitude of the Fast Fourier Transform's (FFT) output. The second step consists of getting the Bark-band energies by adding up the power spectrum values found between two frequency edges as defined by the Bark scale. To map the frequency values f (in Hertz) to the Bark scale B, the following equation can be used:

$$\mathrm{B} = 13\arctan(0.00076f) + 3.5\arctan((f/7500)^2). \tag{2.4}$$

Finally, the Bark-band energy values can be directly used as feature values or, in order to minimize the global energy fluctuations related with the general loudness of each audio frame, these energy values can be normalized by the sum of all energy bands within each temporal frame.

MFCC

The Mel-frequency cepstral coefficients (MFCC) feature is a widely used timbral descriptor that was originally proposed within the Speech Recognition field (Rabiner and Juang, 1993). In this case, the MFCC descriptor is based on the so called Mel scale which is a psychoacoustical scale within which pitches are judged to be equally spaced from one another (Stevens et al., 1937). In order to map frequency values f (in Hertz) to their corresponding Mel (M), the following equation can be used (Herrera et al., 2006):

$$M = 2595\log_{10}\left(1 + \frac{f}{700}\right). \tag{2.5}$$

[4] A critical band is the band of audible frequencies within which a second tone will produce audible interferences to the perception of a first tone placed in the center of the band (Roederer, 2009).

In order to compute the MFCCs, the following steps are needed. Firstly, as in the case of the Bark-band energies descriptor, the power spectrum of each audio frame is computed. Secondly, the energy within each Mel is added up using a triangular band-pass filter. Thirdly, the logarithms of the power at each of the Mel frequencies are taken. Finally, with the aim of separating envelope and pitch information, the resulting Mel log power vector is transformed into the cepstral domain via the Discrete Cosine Transform (DCT) using the following equation:

$$c[n] = 2\sum_{k=0}^{N-1} X_k cos\left(\frac{\pi n(2k+1)}{2N}\right), 0 \leq n \leq N-1, \tag{2.6}$$

where $c[n]$ corresponds to the amplitude of the n^{th} Mel-frequency cepstral coefficient , X_k denotes the magnitude of the FFT bin k and N corresponds to the total amount of bins resulting from a $2N$+1-point FFT.

MSD-Timbre

The MSD-Timbre feature is a multidimensional feature provided within the publicly available Million Song Dataset (MSD) developed by Bertin-Mahieux et al. (2011). This dataset consists of a collection of audio features and metadata for a million popular Western songs (see Appendix A for further details). The audio features that constitute the MSD were provided by The Echo Nest Analyze API[5]. In particular, the MSD-Timbre descriptor represents the timbral characteristics of an audio segment as a 12-dimensional vector. In this case, the spectro-temporal shape of each audio segment is decomposed into 12 bivariate basis functions (i.e. spectro-temporal templates) that capture high-level abstraction with respect to timbral information. For instance, the first template represents the average loudness of the segment, the second its brightness, the third is related with the flatness of a sound, the fourth to sounds with a stronger attack, etc. Hence, the timbral content of the audio segment is described as a weighted linear combination of these 12 basis functions (Jehan, 2010). Fig. 2.2 shows the 12 spectro-temporal templates as depicted in Jehan (2010).

Given its intrinsic characteristics, the 12-dimensional vectors can be split into an 11-dimensional timbre component and a unidimensional loudness component. Notice that including the average loudness in the original

[5] http://the.echonest.com/

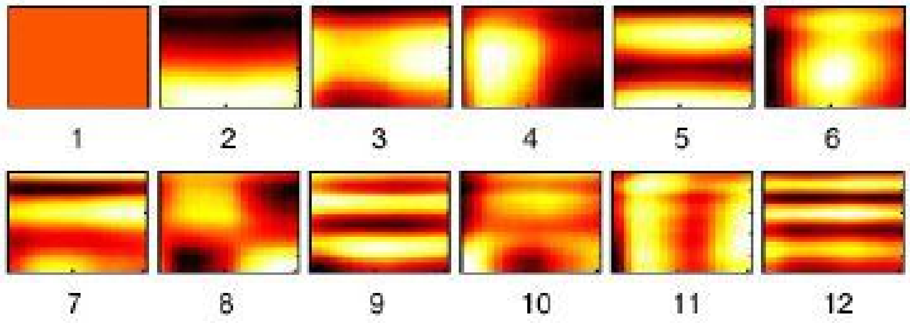

Figure 2.2: Spectro-temporal basis functions used by the MSD-Timbre feature. The x-axis corresponds to time, the y-axis corresponds to frequency, and the z-axis (color) corresponds to amplitude values. Image from Jehan (2010).

timbre representation implies a certain degree of independence of the two components. Since, for perceptual reasons, the frequency resolution of the spectro-temporal representation is intentionally low (Jehan, 2005), the obtained timbre and loudness components can be also assumed to be quite independent of pitch.

2.2.2 Tonal descriptors

Tonal features for short-time audio segments are usually computed as pitch class profiles (PCP) (Fujishima, 1999), where a pitch class is defined as the set of all pitches from the Western music chromatic scale that are a whole number of octaves apart, e.g. notes C1, C2, and C3 all collapse to pitch class C. In particular, as expressed by Serrà (2011), “PCPs are derived from the frequency dependent energy in a given range (typically from 50 to 5,000 Hz) in short-time spectral representations (e.g. 100ms) of audio signals computed in a moving window. This energy is usually mapped into an octave-independent histogram representing the relative intensity of each of the 12 semitones of the Western music chromatic scale (12 pitch classes)”. Thus, the tonal information for each audio frame is represented by a real-valued 12-dimensional vector of pitch class relative energies. For instance, a C Major chord will have high energy values in pitch classes 1, 5, and 8 (i.e. C, E, and G).

This type of features, also called "chroma", have been key in the development of many MIR applications such as the automatic identification of near-duplicate recordings (Serrà et al., 2010), chord/tonality estimation (Gómez, 2006), or music structure segmentation (Paulus et al., 2010). Within this thesis we work with two tonal features, Harmonic Pitch Class Profile (HPCP), and MSD-Pitch. Next, we provide an explanation regarding these features.

HPCP

The Harmonic Pitch Class Profile (HPCP) feature is an enhanced PCP descriptor (Gómez, 2006). HPCPs improve over standard PCP features by diminishing the influence of noisy spectral components. Moreover. HPCPs are tuning independent, and they take into account the presence of harmonic frequencies.

In order to compute the HPCP descriptor, the following steps are needed (see Gómez et al. (2008) for details):

1. Compute the spectrogram of the audio segment via FFT.
2. Take the 30 most prominent spectral peaks between 40 and 5,000 Hz.
3. Estimate a reference tuning frequency by analyzing the deviations of the song's spectral peaks from an equal-tempered chromatic scale with A4 tuned at 440 Hz.
4. Apply a spectral whitening to each peak. In particular, each peak is normalized with respect to the corresponding value of the spectral envelope at the peak's frequency. Since, as mentioned in Sec. 2.2.1, the spectral envelope provide information about timbre, this spectral whitening process aims at obtaining timbre-independent peaks.
5. Take all peaks plus the contributions of 7 harmonics for each peak and build an octave-independent histogram representing the relative intensity of the 12 pitch classes.

As described in Serrà (2011), the computation of the 12-dimensional HPCP vector $h_i = [h_{i,1}, ...h_{i,12}]$ can be expressed mathematically as:

$$h_{i,j} = \sum_{k=1}^{30} \sum_{n=1}^{8} {\alpha_A}^{n-1} \left[\omega \left(j, \frac{f_k}{n} \right) \bar{y_i}^{(f_k)} \right]^2 , \tag{2.7}$$

where i is the audio frame number, j is the pitch class number (from 1 to 12), k is the number of the selected spectral peak, n is the number of the frequency harmonic including the f_k peak, $\bar{y_i}^{(f_k)}$ is the whitened peak magnitude, α_A is a constant, ${\alpha_A}^{n-1}$ is a harmonic weighting term, and $\omega(j, f_n)$ is a cosine weighting function such that

$$\omega(j, f_n) = \begin{cases} cos\left(\frac{\pi}{2}\frac{v(j,fn)}{\alpha_B}\right) & \text{if } |v(j, f_n)| \leq \alpha_B, \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

where α_B is a constant and

$$v(j, f_n) = 12\left[\log_2\left(\frac{f_n}{f_{\text{ref}}2^{\frac{j}{12}}}\right) + \beta\right], \tag{2.9}$$

where β is the integer that minimizes $|v(j, f_n)|$, and f_{ref} is the reference tuning frequency. The constants α_A and α_B are experimentally set to 2/3. Finally the HPCP of a given window is normalized by its maximum value.

The just described HPCP computation, including the mentioned parameters, has delivered excellent results for several MIR tasks such as key estimation, chord extraction, tonal profile determination and version identification (Gómez and Herrera, 2006; Gómez et al., 2006; Serrà, 2011; Serrà et al., 2008; Serrà et al., 2012a).

MSD-Pitch

As mentioned in Sec. 2.2.1 MSD features are delivered within the million song dataset and computed using The Echo Nest API. From the API's documentation we can see that the MSD-Pitch is a standard PCP (or chroma) feature with 12-dimensions corresponding to the 12 pitch classes, and per-segment normalized values as in the case of the just described HPCPs (Jehan, 2010).

2.2.3 Energy descriptors

Energy descriptors are related with our perception of loudness and are used in many MIR tasks such as onset detection, instrument detection, and automatic rhythm description (Bello et al., 2005; Fuhrmann et al., 2009; Gouyon and Dixon, 2005; Haro and Herrera, 2009). Within this thesis we use two energy features namely: Spectral Energy and MSD-Loudness. Next, we offer further details regarding these descriptors.

Spectral Energy

The Spectral energy (SE) feature is computed from the signal's frequency domain obtained via the FFT. Thus, the SE of the frame n is obtained by computing the square of the magnitude of the FFT output. That is

$$SE_n = |X(n)|^2, \tag{2.10}$$

where $X(n)$ is the Fourier transform of the time domain signal $x(n)$.

MSD-Loudness

As described is Sec. 2.2.1, the MSD-Timbre feature provides a spectro-temporal decomposition of the audio segment into 12 bivariate basis function (Jehan, 2010). Since the first basis function corresponds to the energy of the segment, we use the magnitude value of such a first dimension as descriptor of the signal's energy.

2.3 Encoding audio descriptors

As mentioned in Chapter 1, we aim at analyzing the statistical distributions of commonly used audio descriptors. In particular we want to characterize the probability distribution of continuous, and often multidimensional, frame-wise audio features. In order to do that we first need to encode the original continuous features into discrete elements (or events) whose frequency of use can be inferred from a representative sample space. For that we need to encode the (multidimensional) vector space in such way that similar regions of the space are assigned to the same discrete element.

There are many methods used for the discretization of continuous variables (Cios et al., 2007). These methods can be either unsupervised or supervised (Dougherty et al., 1995). Unsupervised methods are the simplest to use and implement. Two of these methods are: equal-width discretization (EWD) and equal-frequency discretization (EFD). In the first case the minimum and maximum values for the continuous variable are computed, and then the interval between both values is divided into n user-provided number of intervals all having the same width. In the case of EFD, the minimum and maximum values are also computed, then all values are sorted in ascending order, and n user-provided intervals are determined in such way that every interval has the same number of values.

Within the supervised methods we find information-theoretic algorithms (ITA), and clustering-based algorithms (Cios et al., 2007). In particular, ITA algorithms need a representative annotated train set where each value of the continuous variable is assigned to a particular element class. Then, by using, for instance, statistical tests (Tay and Shen, 2002) or information entropy (Clarke and Barton, 2000; Fayyad and Irani, 1993), these type of algorithms estimate the best discretization intervals. Unfortunately, regarding audio descriptors, there is no such an annotated dataset since the task of defining a set of suitable classes that can be used to annotate audio descriptors is far from trivial. For instance, what are the classes that should be used to annotate timbre types? Interestingly, recent research in video retrieval has shown that we need less than 3,000 semantic concepts to achieve a suitable description for video images (Hauptmann et al., 2007). Having this type of upper-bound for audio content would be an important starting point towards creating suitable annotations of audio segments (Herrera et al., 2009).

On the other hand, clustering-based algorithms rely on a representative dataset, some pre-defined distance measure, and k user-defined number of discrete classes to build a codebook with k feature prototypes (e.g. class centroids). These class prototypes are later used to discretize new feature vectors by assigning the class membership of the closest prototype. One of the most used cluster-based discretization approaches is the so called Vector Quantization (VQ) algorithm (Linde et al., 1980; Quatieri, 2001). In this case the k prototypes of the codebook correspond to the k centroids estimated from a train set by means of the k-means algorithm (MacQueen, 1967).

Noticeably, some authors have used the VQ algorithm to discretize standard frame-wise audio features into a dictionary of feature code-words. Inspired by text retrieval methods, these code-words are later used to create a Bag-of-Words model which is used as input for content-based MIR tasks such as similarity search (Riley et al., 2008; Seyerlehner et al., 2008), music automatic tagging (Hoffman et al., 2009), genre classification (Marques et al., 2011b) and artist identification (Fu et al., 2011). Unfortunately, these works only use the VQ algorithm as an *ad-hoc* approach and no reports on code-words' probability distributions are made.

With respect to image processing, state-of-the-art video annotation algorithms often use VQ codebooks to build "visual-words" that are later used as input features for classification algorithms (Jiang et al., 2010). Moreover, Mühling et al. (2012) propose a combination of vector quantized "visual code-words" and vector quantized "auditory code-words" for multimodal video concept detection. Finally, Yang et al. (2007) also use a Bag-of-Visual-Words for image retrieval. Interestingly, this work presents graphical information regarding the global distribution of visual code-words. Moreover, the authors characterize the probability distribution of visual code-words as Zipfian. Unfortunately, no formal fitting is presented in the paper.

From our perspective, the use of VQ as suitable tool for inferring the global distribution of short-time audio features poses some problems that lead us to opt for unsupervised discretization approaches. In particular, since we plan to analyze millions of audio frames, the creation of a VQ codebook from such amount of data would not be feasible, and we should pragmatically use random subsamplings. Unfortunately, it is known that the quality of the codebook is much influenced by the random subsampling and the initial selection of cluster centers (Fu et al., 2011). Another problem directly linked to the way VQ works is that every time we run the VQ algorithm a different codebook is created thus affecting the replicability of our experiments. Finally, and more crucially, the selection of a particular distance measure needed to build the VQ clusters is a non trivial decision whose influence over the obtained distribution is not easy to predict. Thus, it would be extremely difficult to determine how the selected distance measure relates with the underlying feature distribution. Furthermore, since the analyzed audio features are mostly multidimensional, the used distance measure could be affected by the so called "curse of dimensionality" that jeopardizes the relevance of such a measure for high-dimensional data (Beyer et al., 1999).

Therefore, after discarding supervised discretization methods, we opt for the fast and simple unsupervised ones. In particular we use EWD for one-dimensional energy features, and EFD for the case of multidimensional timbre and tonal descriptors (see Sec. 2.2). In the first case the range of all values for the energy descriptor is divided into n segments, thus, this energy descriptor is discretized into n code-words.

In the case of multidimensional features (timbre and tonal), each frame-wise feature dimension is discretized into n equal-frequency segments (or "letters"). Thus, a frame-based code-word is constructed from the "letters" coming from each descriptor's dimension in a particular frame. For instance, if we have a 22-dimensional vector (per frame) and we discretize each dimension into 2 equal-frequency segments, we will obtain a codebook of $2^{22} = 4,194,304$ possible code-words. That is, for each audio frame we will have a "word" of 22 letters coming from a binary "alphabet". Noticeably, these unsupervised methods only have one user-provided parameter namely the number of segments used to discretize each feature dimension. Furthermore, these simple discretization methods are akin to encoding methods used, for instance, in automatic audio identification (Haitsma and Kalker, 2002) or in cochlear implant sound processors (Wilson et al., 1991), and roughly resemble the all-or-none behavior of neurons and neuronal ensembles (Bethge et al., 2003).

2.4 Audio classification

As depicted in Fig. 1.1, many MIR algorithms use a set of audio features and manually annotated examples as input for training a classification algorithm. In this training step the classification algorithm "learns" a classification model that will be used to determine the class membership of new (unlabeled) data. For instance, for song-level classification, when a new unlabeled song arrives, the system computes the same set of descriptors used to train the classification model (e.g. a set of aggregated frame-level features). Then these descriptors are used as input for the classification model which will decide on the class membership of the song (e.g. its genre).

From the many classification methods proposed and successfully used within the MIR field (cf. Klapuri and Davy, 2006), in this thesis, we have decided to use the well known support vector machines (SVM) algorithm (Cortes and Vapnik, 1995; Vapnik, 2000). This decision is motivated by several facts. Firstly, SVMs are considered a must try on any machine learning application due to their robustness, generalization capabilities, and accurate results (Wu et al., 2007). Secondly, SVMs have achieved state-of-the-art results in many content-based MIR tasks (Klapuri and Davy, 2006; Mandel and Ellis, 2005). Finally, since we work with large datasets and large sparse feature vectors, SVMs are the option of choice to avoid overfitting (Joachims, 1998).

The SVM algorithm is part of the general class of supervised statistical learning algorithms that use a discriminant analysis function to classify data (Webb, 2002). Hence, the SVM classification model is created by finding the hyperplane with maximum soft-margin for the given training set (Burges, 1998). Once the separating hyperplane $f(x)$ is found, new data instances can be easily classified by evaluating the sign of the function

$f(x)$. If $f(x) > 0$ then the instance belongs to the positive class, otherwise it belongs to the negative class (Webb, 2002). Moreover, if the training data is not linearly separable it can be mapped into a new (hopefully linear) hyperspace. This mapping can be achieved by choosing the correct kernel function where commonly used kernel functions are radial basis functions and polynomials (of various degrees).

For assessing classification performance we opt for standard evaluation measures of information retrieval (IR) namely: precision, recall, and F-measure (Baeza-Yates, 1999).

- Precision: it is the fraction of correctly classified items over the total number of items:

  $$P = \frac{C_a}{D}, \tag{2.11}$$

  where C_a is the number of items correctly classified as "a" and D is the total number of retrieved items.

- Recall: it is the fraction of correctly classified items over the total of items that belong to the class.

  $$R = \frac{C_a}{GT_a}, \tag{2.12}$$

  where GT_a is the total number events that belong to the class "a" (i.e. the ground truth annotations).

- F-measure: it is the weighted harmonic mean of precision and recall, it summarizes a trade-off between both values.

  $$F = \frac{2PR}{P + R} \tag{2.13}$$

The above-described background provides a general overview of the main tools we use in this thesis. These tools will allow us to pursue the proposed thesis' goals, that is, to characterize the statistical distribution of common audio features, and propose improvements over the standard BoF approach. Furthermore, when additional concepts and references are needed, we will include them in the corresponding chapters. For instance, in Chapter 4 we will introduce power-law generative models, and in Chapter 7 we will present several automatic tagging algorithms.

Chapter **3**

Rank-frequency distribution of audio timbral descriptors

Most of the results presented in this chapter were published in Haro et al. (2012a) and Haro et al. (2012b).

3.1 Introduction

In Chapter 2 we have stressed the importance of knowing the frequency distribution of the different elements that constitute a particular area of study. For instance, knowing that word-frequency counts are power-law distributed has had an important impact on several scientific disciplines that work with written text such as information retrieval (Baeza-Yates, 1999) and natural language processing (Manning and Schütze, 1999).

In the case of sound-related areas there are still few publications that address this point. This can be partially explained by the fact that it is non-trivial to determine which are the "natural elements" that constitute an audio signal. Thus, as stated in Sec. 2.1, some authors have focused on symbolic representations of music (such as scores or MIDI files) and reported heavy-tailed distributions on note-related events. There are also some publications characterizing the distribution of sound amplitudes and individual frequency bands directly extracted from raw audio signals. Noticeably, these works also reported heavy-tailed distributions.

Working directly with audio signals would allow us to tackle the problem of relying on arbitrary symbolic notations, most of which are not even defined (e.g. what would be the notation to describe soundscape recordings?). Unfortunately, there is still no automatic algorithm that detects, identifies and segments an audio signal into perceptually-relevant sound objects. Moreover, given the extreme subjectivity and complexity of this cognitive task, it seems that the appearance of such an algorithm is far from imminent. Meanwhile, the sound and music computing field has proven that working with fixed-length audio excerpts (or frames) as "sound units" constitutes an efficient practical approach when designing sound-related applications such as F0 detection, cover song detection, music genre identification, etc. Automatic descriptions of such audio frames constitute the basic low-level elements on which many of those algorithms are constructed (Casey et al., 2008; Müller et al., 2011; Quatieri, 2001).

Consequently, in this chapter, we analyze the rank-frequency distribution of audio frames that account for the timbral characteristics of audio signals. As stated in Sec. 2.2, *timbre* is a key perceptual feature that allows to discriminate between different sounds. Timbral sensations mainly correlate with the audio waveform shape and, thus, with the spectro-temporal envelope of the signal (i.e. the temporal evolution of the shape of the power spectrum) (Berg and Stork, 1995). In order to quantitatively characterize such sensations, the shape of the power spectrum has to be encoded in a way that preserves certain physical and perceptual properties. Therefore, it is common practice to encode short-time power spectra using psychoacoustical frequency scales such as the Bark scale (Zwicker, 1961) or Mel scale (Stevens et al., 1937).

In particular, in the following sections we study and characterize the statistical properties of encoded (i.e. discretized) timbral descriptors extracted on a frame-by-frame basis. For that we focus on two of the most used timbral descriptors namely: normalized Bark-band energies[1], and MFCCs (see Sec. 2.2). We use a simple encoding process which maps each descriptor's frame to a dictionary of more than 4 million binary code-words (see Sec. 2.3). We analyze a large-scale corpus of audio signals consisting of 740 hours of sound coming from disparate sources such as *Speech*, *Western Music*, *non-Western Music*, and *Environmental sounds*. We perform a rank-frequency distribution analysis and show that the frequency distribu-

[1] Within this text we pragmatically refer to the normalized Bark-band energies as Bark-bands

tion of encoded timbral descriptors follows a power-law distribution. This distribution is found independently of sound source, frame size and descriptor type and, since the chosen timbral descriptors are highly related with the signal's spectral shape (or envelope), we hypothesize that the found power-law distribution could also be a general property of short-time spectral envelopes of audio signals. Furthermore, we analyze the inner structure of the most (and least) frequent code-words and provide evidence that a heavy-tailed distribution is also present when analyzing individual recordings (e.g. individual songs). All these findings suggest promising new paths for developing audio-related applications. Some of these paths are started to be walked in the next chapters of this thesis.

3.2 Method

We represent the timbral characteristics of short-time consecutive audio fragments following standard procedures in computational modeling of speech and music (Casey et al., 2008; Müller et al., 2011; Quatieri, 2001). We decided to work with two commonly used timbral descriptors namely: Bark-bands and MFCC (see Sec. 2.2.1). The output of these descriptors is a frame-based multidimensional vector of real numbers. Therefore, as stated in Sec. 2.3, in order to characterize the rank-frequency distribution of such real-valued vectors we first need to quantize (or encode) them in such manner that similar descriptor's values are assigned to the same encoded type. This allows us to count the number of tokens corresponding to each type (i.e. the frequency of use of each encoded type). Ultimately, each of these types can be seen as a code-word assigned from a predefined dictionary of timbres.

In this chapter we consider three perceptually motivated audio fragment sizes namely: 46, 186, and 1,000 ms. The first one (46 ms) is selected because it is extensively used in audio processing algorithms and tries to capture the small-scale nuances of timbral variations (Casey et al., 2008; Müller et al., 2011). The second one (186 ms) corresponds to a perceptual measure for sound grouping called "temporal window integration" (Oceák et al., 2008), usually described as spanning between 170 and 200 ms. Finally, we explore the effects of a relatively long temporal window (1 s) that exceeds the usual duration of speech phonemes and musical notes.

3.2.1 Databases

We analyze 740 hours of real-world sounds grouped into four databases: *Speech*, *Western Music*, *non-Western Music*, and *Sounds of the Elements* (see Appendix A). The *Speech* database is formed by 130 hours of recordings of English speakers from the *Timit* database (Garofolo et al., 1993) (about 5.4 hours), the *Library of Congress* "Music and the brain" podcasts[2] (about 5.1 hours), and 119.5 hours from Nature podcasts[3] from 2005 to April 7th 2011. The *Western Music* database is formed by about 282 hours of music (3,481 full tracks) extracted from commercial CDs accounting for more than 20 musical genres including: rock, pop, jazz, blues, electronic, classical, hip-hop, and soul. The *non-Western Music* database contains 280 hours (3,249 full tracks) of traditional music from Africa, Asia, and Australia extracted from commercial CDs. Finally, in order to create a set that clearly contrasted the other ones, we decided to collect sounds that were not created to convey any message. For that reason we gathered 48 hours of natural sounds produced by natural inanimate processes such as water sounds (rain, streams, waves, melting snow, waterfalls), fire, thunders, wind, and earth sounds (rocks, avalanches, eruptions). This *Sounds of the Elements* database was assembled using files downloaded from *The Freesound Project*[4]. The differences in size among databases try to account for their differences in timbral variations (e.g. the sounds of the elements are less varied, timbrically speaking, than speech and musical sounds; therefore we can properly represent them with a smaller database).

3.2.2 Timbral descriptors

As previously mentioned, we focus on two of the most used timbral descriptors namely Bark-bands and MFCCs (Müller et al., 2011; Quatieri, 2001; Raś, 2010). Both descriptors are computed from perceptually motivated bands of the power spectrum in short-time audio segments, or frames (see Sec. 2.2 for further information).

The normalized Bark-band energies descriptor is obtained by adding up the power spectrum values found between two frequency edges defined by the Bark scale (Zwicker, 1961). Since we want to characterize timbral informa-

[2] http://www.loc.gov/podcasts/musicandthebrain/index.html

[3] http://www.nature.com/nature/podcast/archive.html

[4] http://www.freesound.org

tion regardless of the total energy of the signal, we normalize each Bark-band value by the sum of all energy bands within each temporal frame. The output of this process is a sequence of 22-dimensional vectors that represents the percentage of energy contained in each frequency band between 0 and 9,500 Hz (i.e. the first 22 critical bands of hearing). The used Bark-band frequency edges are: 0, 100, 200, 300, 400, 510, 630, 770, 920, 1,080, 1,270, 1,480, 1,720, 2,000, 2,320, 2,700, 3,150, 3,700, 4,400, 5,300, 6,400, 7,700, and 9,500 Hz (Zwicker, 1961). The 9,500 Hz upper bound is motivated by the fact that most of the perceptually relevant sound energy lie below this threshold (Berg and Stork, 1995) and because adding more bands exponentially multiplies the computational load of our experiments.

The MFCC descriptor is obtained by mapping the short-time power spectrum to the Mel scale (Stevens et al., 1937) which roughly represents the spacing between critical bands of human hearing. The Mel-energy values are then computed using triangular band-pass filters centered on every Mel. The logarithm of every Mel-energy value is taken and the discrete cosine transform (DCT) of the Mel-log powers is computed. The MFCC descriptor corresponds to a real-valued vector of amplitude coefficients of the resulting DCT spectrum. Here, we use the Auditory toolbox MFCC implementation (Slaney, 1998) with 22 coefficients (skipping the DC coefficient). By selecting 22 MFCC coefficients we obtain a good trade-off between the detail of the spectral-envelope description and the computational load of our experiments. Moreover, working with a 22-dimensional vector allows us to use the same encoding strategy for both timbral descriptors (see next section).

3.2.3 Encoding process

Independently of the chosen timbral descriptor we follow the same encoding procedure for every sound file in every database (Fig. 3.1). Starting from the time-domain audio signal (digitally sampled and quantized at 44,100 Hz and 16 bits) we apply an equal-loudness filter. This filter takes into account the sensitivity of the human ear as a function of frequency. Thus, the signal is filtered by an inverted approximation of the equal-loudness curves described by Fletcher and Munson (1933). The filter is implemented as a cascade of a 10th order Yule-Walk filter with a 2nd order Butterworth high-pass filter (Madisetti, 1997).

Next, the signal is segmented into non-overlapping temporal frames of ei-

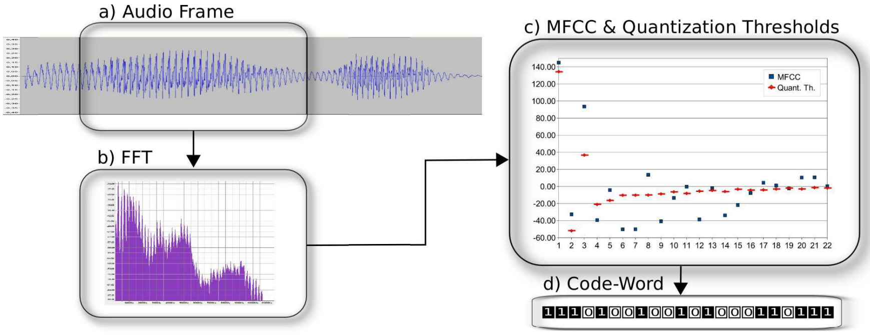

Figure 3.1: Block diagram of the encoding process. a) The audio signal is segmented into non-overlapping frames. b) The power spectrum of each audio frame is obtained. c) Multidimensional timbral descriptor's values (MFCC values in this case) are computed (blue squares) and each vector's dimension is binary-quantized by comparing its value against a pre-computed threshold (red line). d) Each quantized MFCC (or Bark-band) vector forms an MFCC (Bark-band) code-word.

ther 46, 186, or 1,000 ms length (Fig. 3.1a). Then, each audio segment is converted to the frequency domain by taking the Fourier transform (Madisetti, 1997) using a Blackman-Harris temporal window. From the output of the Fourier transform we compute its power spectrum by taking the square of the magnitude (Fig. 3.1b). Next, we compute the corresponding timbral descriptor (i.e. either Bark-bands or MFCCs) obtaining a multidimensional real-valued vector per frame (Fig. 3.1c depicts MFCC values). Finally, we quantize each vector's dimension by comparing its value against a stored threshold (red lines in Fig. 3.1c). In particular, if the dimension's value is smaller than the dimension's threshold we encode this dimension's value as "0", otherwise we encode it as "1" (Fig. 3.1d). Thus, after this quantization process every audio frame is encoded as a sequence of 22 zeros and ones. Therefore, the total amount of possible code-words (i.e. the encoding dictionary) is $2^{22} = 4,194,304$ code-words.

In order to obtain the quantization thresholds we computed both timbral descriptors on a representative database. Then, we stored the median value for each descriptor's dimension and frame size. This way, each dimension is split into two equally populated groups (equal-frequency discretization; see Sec. 2.3). The representative database contains all frame-based descriptor values from the *Sounds of the Elements* database plus a random sample of descriptor values from the *Speech* database that matches in number the ones from the *Sounds of the Elements*. It also includes random selections of *Western Music* and *non-Western Music* matching half of the length of *Sounds of the Elements* each. Thus, our representative database has its descriptor values distributed as one third coming from *Sounds of the Elements*, one third from *Speech*, and one third from *Music* totaling about 20% of the whole analyzed sounds. We constructed 10 of such databases per frame size and, for each dimension, we stored the mean of the median values as quantization threshold.

3.2.4 Fitting procedure

To evaluate if a power-law distribution fits our data we take the frequency of each code-word (i.e. the number of times each code-word is used) as a random variable and apply state-of-the-art methods of fitting and testing goodness-of-fit to this variable (Clauset et al., 2009; Corral et al., 2011). The procedure consists in finding the frequency range $[z_{\min}, z_{\max}]$ for which the best power-law fit is obtained. First, arbitrary values for lower and upper cutoffs $z_{\min}$ and $z_{\max}$ are selected and the power-law exponent β is obtained by maximum-likelihood estimation. Second, the Kolmogorov-Smirnov test quantifies the separation between the resulting fit and the data. Third, the goodness of the fit is evaluated by comparing this separation with the one obtained from synthetic simulated data (with the same range and exponent β) to which the same procedure of maximum-likelihood estimation plus Kolmogorov-Smirnov test is applied. This goodness of the fit yields a p-value as a final result. Then, the procedure selects the values of $z_{\min}$ and $z_{\max}$ which yield the largest log-range $z_{\max}/z_{\min}$ provided that the $p-$value is above a certain threshold (for instance 20%)[5]. See Appendix C for further details.

3.3 Results

3.3.1 Bark-band code-words

In order to illustrate the results of the encoding procedure we show the time-frequency representation (i.e. spectrogram) of a sinusoidal sweep in

[5] In all cases we have obtained that we can take $z_{\max} \to \infty$ and results with finite $z_{\max}$ are not presented here.

Sound Description	**# code-words**
Sine wave 440 Hz	1
Rain	18
1/f (Pink) Noise	26
White Noise	28
Sinusoidal Sweep (0-9,500 Hz)	37
Clarinet solo	97
Female English speaker	128
String Quartet	135
Voice, Drums, Bass & Synth. Strings	140
Philharmonic Orchestra	141
Voice and Electronic Instruments	153

Table 3.1: Number of different Bark-band code-words used to describe each sound. Examples computed from 30 s audio files using a frame size of 186 ms (160 frames in total). Pink and white noise sounds were generated using Audacity (`http://audacity.sourceforge.net`). **String Quartet** corresponds to a rendition of F. Haydn's Op.64 No.5 "The Lark", **Voice, Drums, Bass & Synth. Strings** corresponds to Michael Jackson's *Billie Jean*, **Philharmonic Orchestra** corresponds to a rendition of *The Blue Danube* by J. Strauss II, and **Voice and Electronic Instruments** corresponds to Depeche Mode's *The world in my eyes.*

logarithmic progression over time, ranging from 0 to 9,500 Hz (Fig. 3.2a) and its corresponding Bark-band code-words (Fig. 3.2b). In both plots we can see the sweeping of the sinusoidal sound. Thus, we can observe how the Bark-band code-words form a simplified representation of the spectral content of the signal while preserving the main characteristics of its spectral shape[6]. As a further example, we consider the number of distinct Bark-band code-words used to encode sounds with disparate timbral characteristics, ranging from a simple sinusoidal wave up to multi-instrument polyphonic music (Table 3.1). As expected, we observe a positive correlation between the timbral "richness" of the analyzed sounds and the number of code-words needed to describe them (i.e. as the timbral variability increases, sounds are encoded using a greater number of different code-words).

[6] The difference between both curve shapes is due to the use of different frequency representations; the spectrogram uses a linear frequency representation while the Bark-band code-words are computed using a non-linear scale based on psychoacoustical findings (i.e. the Bark scale).

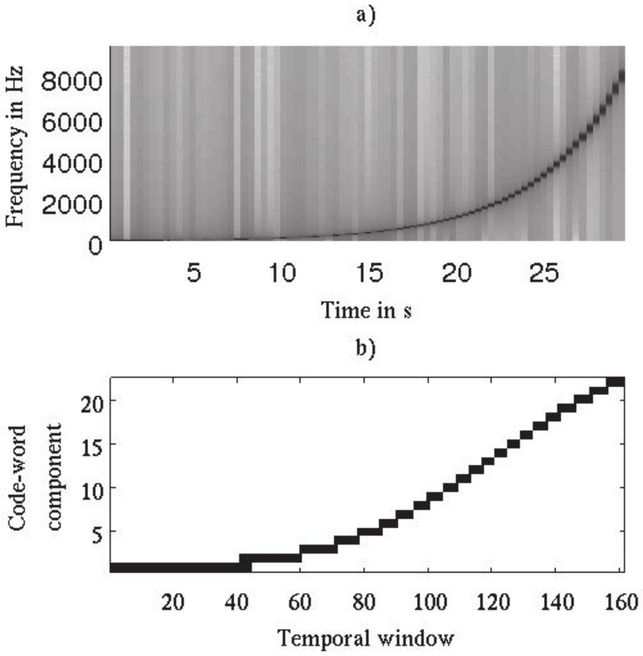

Figure 3.2: Spectrogram vs. Bark-band code-word example. a) Spectrogram representation for a sinusoidal sweep in logarithmic progression over time ranging from 0 to 9,500 Hz. The color intensity represents the energy of the signal (white = no energy, black = maximum energy). This standard representation is obtained by means of the short-time Fourier transform. b) Bark-band code-word representation of the same audio signal. The horizontal axis corresponds to temporal frames of 186 ms and the vertical axis shows the quantized values per Bark-band (black = 1 and white = 0). For instance, in the first 40 temporal frames only the first Bark-band is quantized as one (the first Bark-band corresponds to frequencies between 0 and 100 Hz). A total of 37 different code-words are used to encode this sinusoidal sweep.

Once obtained the Bark-band code-words for all sounds in all databases we count the frequency of use of each code-word within each database (i.e. the number of times each code-word is used in the database) and sort them in decreasing order of frequency (Fig. 3.3a). We find that a few code-words are very frequent while most of them are very rare. In order to evaluate if the found distribution corresponds to a Power-law distribution, instead of working directly with the rank-frequency plots we focus on the equivalent

Frame size	**N words**	$z_{\min}$	β	α
		Speech		
46 ms	494,926	2,000	2.20 ± .05	0.84 ± .04
186 ms	219,595	501	2.22 ± .05	0.82 ± .03
1,000 ms	100,273	79	2.33 ± .05	0.75 ± .03
		Western Music		
46 ms	1,724,245	2,000	2.26 ± .04	0.79 ± .03
186 ms	798,871	794	2.33 ± .06	0.75 ± .03
1,000 ms	240,236	79	2.29 ± .03	0.78 ± .02
		non-Western Music		
46 ms	1,905,444	126	2.17 ± .01	0.85 ± .01
186 ms	947,327	50	2.17 ± .01	0.85 ± .01
1,000 ms	306,682	5	2.17 ± .01	0.86 ± .01
		Sounds of the Elements		
46 ms	125,248	794	1.95 ± .04	1.05 ± .05
186 ms	34,171	20	1.79 ± .02	1.27 ± .03
1,000 ms	10,231	8	1.79 ± .02	1.27 ± .03

Table 3.2: Power-law fitting results for Bark-band code-words per database and frame size. **N words** is the number of used code-words, $z_{\min}$ is the minimum frequency for which the Zipf's law is valid, β is the frequency-distribution exponent (Eq. 2.2), and α corresponds to the Zipf's exponent (Eq. 2.1).

description in terms of the distribution of the frequency (Fig. 3.3b) and apply the fitting procedure described in Sec. 3.2.4. In all cases we obtain that a power-law distribution (see Eq. 2.1) is a good fit beyond a minimum frequency $z_{\min}$. Moreover, consistently with Zipf's findings in text corpora, all the estimated Zipfian exponents are close to one (Table 3.2) therefore, the found distributions can be further described as Zipfian distributions (see Sec. 2.1). The high frequency counts for few Bark-band code-words are particularly surprising given the fact that we used a very large coding dictionary (each temporal window was assigned to one out of more than four million possible code-words).

In the case of text corpora, it has been shown that simple random texts do not produce a Zipfian distribution (Ferrer-i-Cancho and Elvevåg, 2010). In the case of our Bark-band code-words it is clear that it would be very difficult to generate random sequences with Zipf-like rank-frequency distribution. In particular, all Bark-band code-words have the same length (i.e. 22 characters) and are formed by two possible characters ("0" and "1").

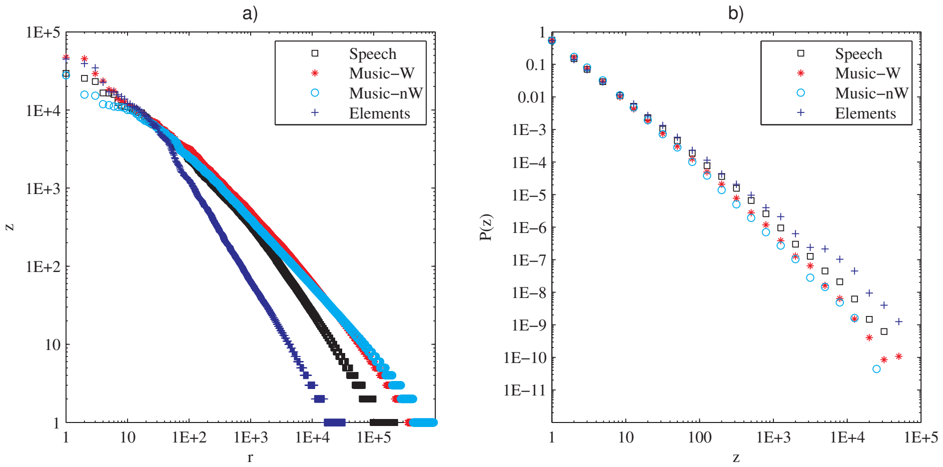

Figure 3.3: Bark-band code-words. a) Rank-frequency distribution of code-words per database (frame size = 186 ms). b) Probability distribution of frequencies for the same Bark-band code-words. Music-W means *Western Music*, Music-nW means *non-Western Music* and Elements means *Sounds of the Elements*.

Given that we opt for a binary equal-frequency discretization (i.e. using representative median values as quantization thresholds), the probability of occurrence of each character in our experiments is close to 0.5. Therefore, if we generate a random sequence of words formed by 22 binary characters having similar probability of occurrence we would observe similar word counts for all generated random words. Thus, the rank-frequency distribution for those random words would be close to a horizontal line (i.e. slope close to zero). Only in extreme cases where the probability of occurrence of one character is much higher than the other we will observe long tailed rank-frequency distributions, but, even in those cases, the distribution will differ from a real Zipfian distribution. In this case, instead of being a straight line in the log-log plot the distribution would present a staircase shape. In the utmost case of one character having probability one, only one word (a sequence of 22 equal characters) will be repeatedly generated producing a delta-shaped rank distribution[7]. Finally, we empirically tested several quantization thresholds, extracted from a sample of different database com-

[7] Note that in our encoding scenario, a delta-shaped rank distribution would be produced if the analyzed database contains only one static sound, like in the case of the sine wave encoded in Table 3.1

binations, without observing any significant change in the rank-frequency plots.

Now we study the robustness of the found Zipfian distributions against several variables used during the encoding process.

Robustness against frame size

Remarkably, changing the frame size by almost one and a half orders of magnitude (from 46 to 1,000 ms) has no practical effect on the estimated exponents. This is especially valid for *Speech* and both *Western* and *non-Western Music* databases. For instance, in Fig. 3.4 we show an example of the probability distribution of frequencies and the estimated power-laws for Bark-band code-words of *non-Western Music* analyzed with the three considered temporal windows or frames (46, 186, and 1,000 ms). The main effect produced by changing the frame size seems to be that the smaller the window, the larger the minimum frequency value from which the power-law is found to be a plausible fit for the data (z_{min} in Table 3.2).

Robustness against frequency bands

Since, to represent timbre, we are describing the spectro-temporal envelopes using a psychoacoustical scale (the Bark scale) and, given that psychoacoustical scales present higher resolution (i.e. small bandwidth) in the low frequency ranges, we now re-compute the code-words using 22 equally-spaced frequency bands (431.8 Hz each). Again, the obtained rank-frequency distribution are very similar to those obtained using Bark-bands (Fig. 3.5). Table 3.3 shows the fitting results for the equally-spaced encodings from the different databases and frame sizes. Noticeably, in this case, all Zipf's exponents are bigger than one and they are stable for the two small temporal frames only (except for *non-Western Music* were all frame sizes share almost the same exponent).

This experiment suggest that similar distributions would be also obtained for other psychoacoustical scales like the Mel scale (Stevens et al., 1937) (see Sec. 3.3.2) or the ERB scale (Moore and Glasberg, 1996).

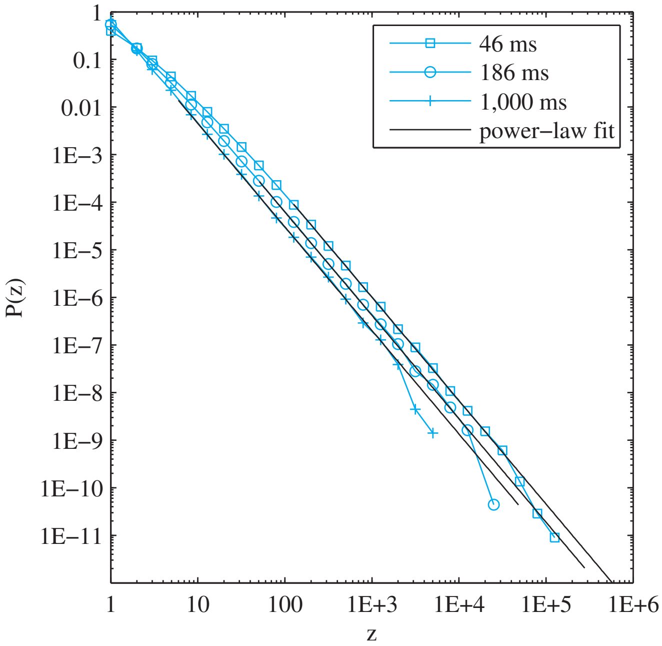

Figure 3.4: Probability distribution of frequencies of timbral code-words for *non-Western Music* analyzed with frame sizes of 46, 186, and 1,000 ms.

Robustness against equal-loudness filtering

As described in Sec. 3.2.3, our original encoding process includes a pre-processing step that in order to emulate the sensitivity of the human ear, filters the signal according to an equal-loudness curve. Thus, we re-compute the whole encoding process without this equal-loudness filter. In this case the obtained results were practically identical to the ones obtained using the equal-loudness filter.

Robustness against audio length

Up to this point all our rank-frequency counts refer to whole databases (i.e. many hours of audio recordings). Now we analyze the distribution of code-words for randomly selected audio segments of up to 6 minutes in length (a duration that includes most of the songs in western popular

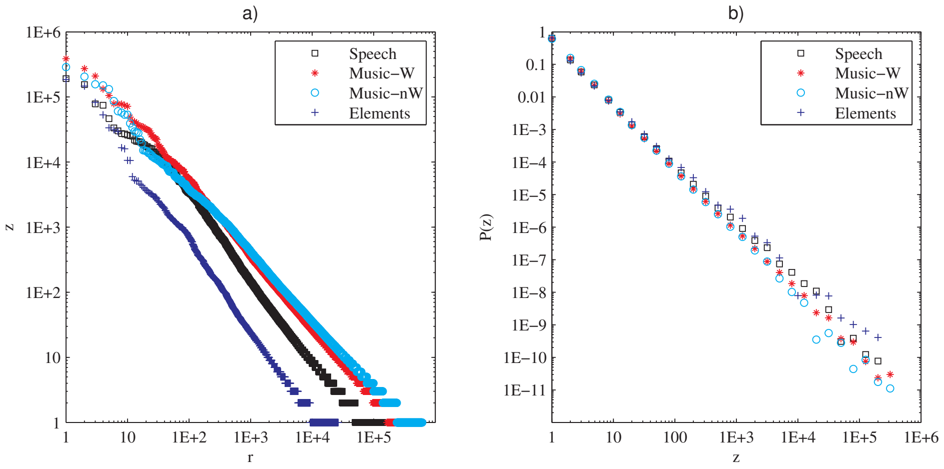

Figure 3.5: Timbral code-words encoded from equally-spaced frequency bands. a) Rank-frequency distribution of timbral code-words encoded from equally-spaced frequency bands (Bandwidth = 431.84 Hz, frame size = 186 ms). b) Probability distribution of frequencies for the same code-words.

music). Noticeably, we find again a similar heavy-tailed distribution as the one found for the whole databases. Fig. 3.6, shows an example of rank-frequency distributions of randomly selected audio excerpts per database. In the case of *Western* and *non-Western Music* databases the excerpts correspond to individual songs. In the case of *Speech* and *Sounds of the Elements* the audio files were cut with arbitrary lengths of up to 6 min. In this experiment we empirically noticed that the rank-frequency exponents of the audio excerpts fluctuate depending on the timbral variety of the excerpts.

The evidence presented so far suggests that the found Zipfian distribution of Bark-band code-words is not the result of a particular type of sound source, sound encoding process, frame size, or sound length. In the next sections we further study the intrinsic characteristics of this encoded timbral descriptor.

Code-Word analysis

Since Bark-bands can be easily traced back to a rough representation of the spectral envelope we now study the specific characteristics of Bark-band

Frame size	**N words**	$z_{\min}$	β	α
		Speech		
46 ms	383,207	200	1.75± .02	1.33 ± .03
186 ms	139,452	79	1.74± .02	1.35 ± .03
1,000 ms	48,717	200	1.95± .06	1.05 ± .06
		Western Music		
46 ms	1,288,416	126	1.91± .01	1.11 ± .01
186 ms	457,575	50	1.88± .01	1.14 ± .02
1,000 ms	103,364	32	1.80± .02	1.26 ± .03
		non-Western Music		
46 ms	1,514,576	50	1.97± .01	1.04 ± .01
186 ms	613,361	20	1.95± .01	1.05 ± .01
1,000 ms	175,518	79	1.98± .03	1.02 ± .03
		Sounds of the Elements		
46 ms	111,593	50	1.77± .02	1.31 ± .03
186 ms	26,557	8	1.74± .02	1.35 ± .03
1,000 ms	5,453	20	1.70± .04	1.44 ± .08

Table 3.3: Power-law fitting results for code-words encoded from equally-spaced frequency bands per database and frame size. **N words** is the number of used code-words, $z_{\min}$ is the minimum frequency for which the Zipf's law is valid, β is the frequency-distribution exponent, and α corresponds to the Zipf's exponent.

code-words. Thus, specific patterns in Bark-band code-words could unveil underlying patterns in spectral envelope shapes.

Noticeably, when we examine the inner structure of Bark-band code-words as ordered by decreasing frequency usage, we find that in all analyzed databases the most frequent code-words present a smoother structure, with close Bark-bands having similar quantization values. Conversely, less frequent elements present a higher band-wise variability (Fig. 3.7).

In order to quantify this observed smoothness, we compute the sum of the absolute values of the differences among consecutive bands of a given code-word. Thus, a code-word smoothness s was computed using

$$s = \frac{c - \sum_{i=1}^{B-1} |b_i - b_{(i-1)}|}{c}, \tag{3.1}$$

where B corresponds to the number of bands per code-word (22 in our case), b_i corresponds to the value of band i and $c = (B-1)(Q-1)$, where

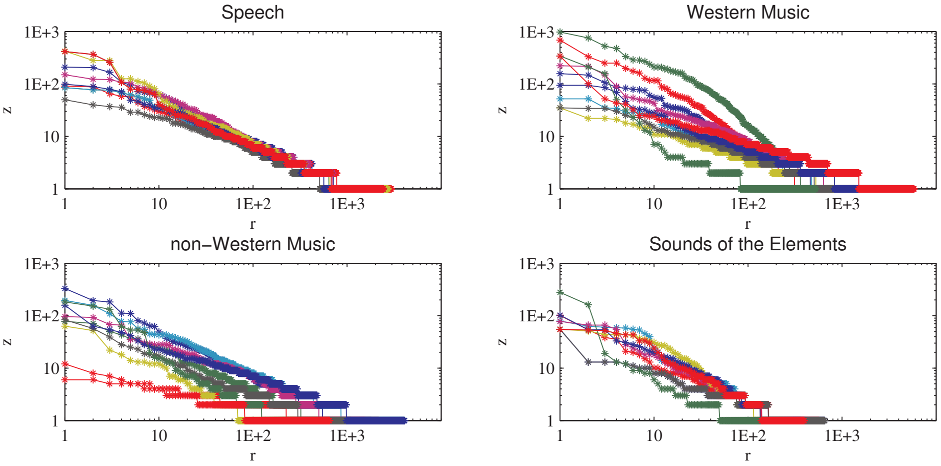

Figure 3.6: Rank-frequency distributions of Bark-band code-words from ten randomly selected audio excerpts per database (frame size = 46 ms). In the case of *Western Music* and *non-Western Music* each line corresponds to one song. In the case of *Speech* and *Sounds of the Elements* each line corresponds to an arbitrary audio segment of up to 6 min in length.

Q corresponds to the number of quantization steps (e.g. $Q = 2$ for binary quantization).

The results show that all databases follow the same behavior, namely, that the most frequent code-words are the smoother ones. Thus, the smoothness value tends to decrease with the rank (see Fig. 3.8).

Next, we analyze the co-occurrence of Bark-band code-words between databases (see also Appendix D). We find that about 80% of the code-words present in the *Sounds of the Elements* database are also present in both *Western* and *non-Western Music* databases. Moreover, 50% of the code-words present in *Sounds of the Elements* are also present in *Speech*. There is also a big overlap of code-words that belong to *Western* and *non-Western Music* simultaneously (about 40%). Regarding the code-words that appear in one database only, we find that about 60% of the code-words from *non-Western Music* belong exclusively to this category. The percentage of database-specific code-words in *Western Music* lies between 30 and 40% (depending on the frame size). In the case of the *Speech* database, this percentage lies between 10 and 30%. Remarkably, the *Sounds of the Elements* database has almost no specific code-words.

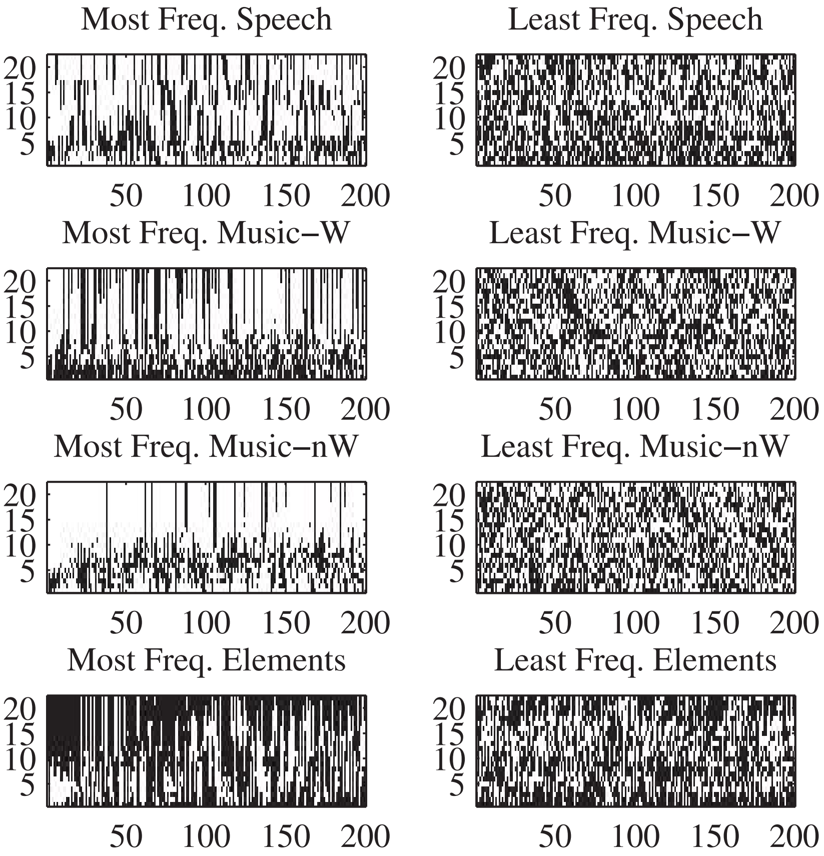

Figure 3.7: Most (left) and least (right) frequent Bark-band code-words per database (frame size = 186 ms). The horizontal axis corresponds to individual code-words (200 most common and a random selection of 200 of the less common). The vertical axis corresponds to quantized values per Bark-band (white = 0, black = 1). Every position in the abscissa represents a particular code-word.

We also find that within each database, the most frequent code-words were temporally spread throughout the database. Therefore, their high frequency values are not due to few localized repetitions. In fact, we observe local repetitions of frequent code-words across the whole database. In Fig. 3.9 an example of the temporal distribution of the most frequent code-words that account for 20% of the *non-Western Music* database is shown. As it can be seen, the code-words are temporally spread throughout the entire

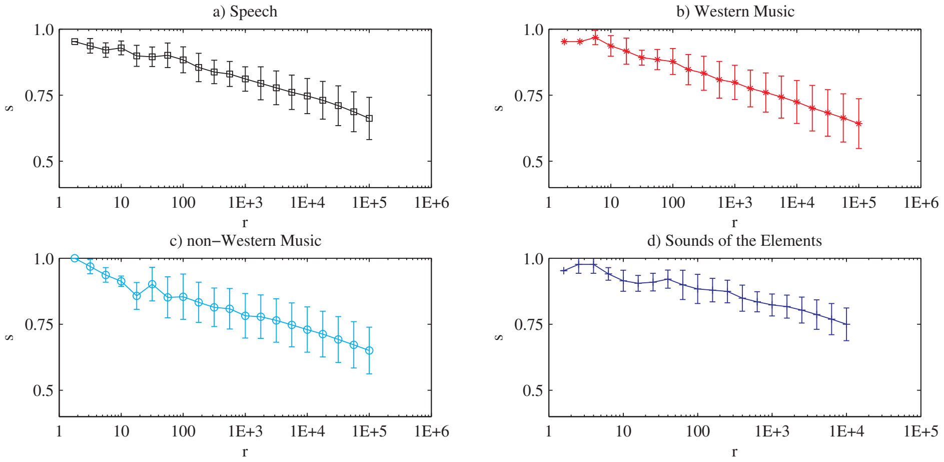

Figure 3.8: Smoothness values (s) per database. For a better visualization we plot the mean and standard deviation of the smoothness value of 20 logarithmically-spaced points per database (frame size= 186 ms).

time axis. The same temporal spreading was observed for the most frequent Bark-band code-words found in the rest of the databases and for different frame sizes.

Finally, we logically find that the smaller the frame size the bigger the number of different code-words obtained after encoding the four databases. In particular, in the case of the 46 ms frame size a total of 2,516,227 different Bark-band code-words were used to encode the four databases. Interestingly, there were 1,678,077 code-words (40% of the dictionary) that were never used (i.e. more than 1.5 million Bark-band combinations that were not present in 740 hours of sound).

3.3.2 MFCC code-words

Following the methodology described in Sec. 3.2.3 we now encode every audio frame into its corresponding MFCC code-word. As done for the Bark-band code-words (Sec. 3.3.1), for each database and frame size, we count the frequency of use of each code-word and sort them by decreasing order of frequency. As shown in Fig. 3.10a, when plotting these rank-frequency counts we observe heavy-tailed distributions for all the analyzed databases.

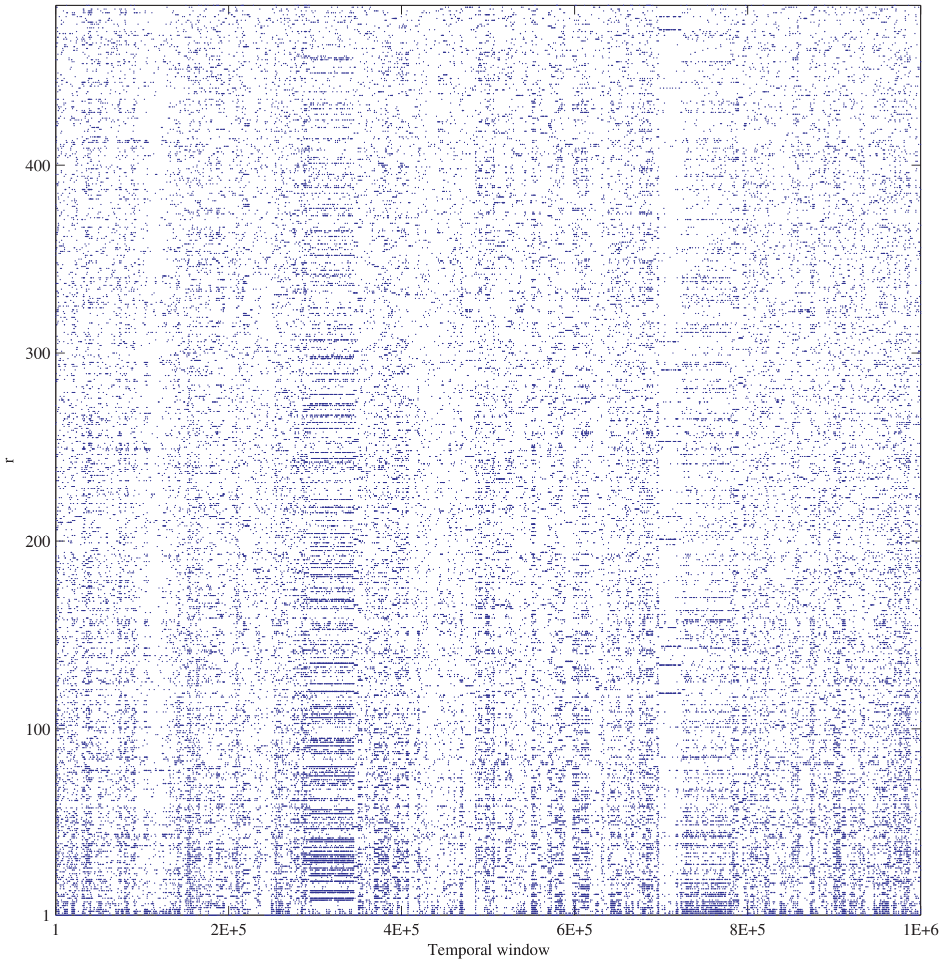

Figure 3.9: Temporal distribution of 485 most frequent code-wods in *non-Western Music* (frame size = 1,000 ms). Each dot indicates the temporal location (x axis) of a particular Bark-band code-word (y axis).

Again, as in the case of Bark-band code-words, these distributions imply that a few code-words are very frequent while most of them are very unusual.

In order to evaluate if the found heavy-tailed distributions of MFCC code-words specifically correspond to power-law distributions we apply the previously described estimation procedure (Sec. 3.2.4). This procedure, instead of working directly with the rank-frequency plots, it focuses on the equiv-

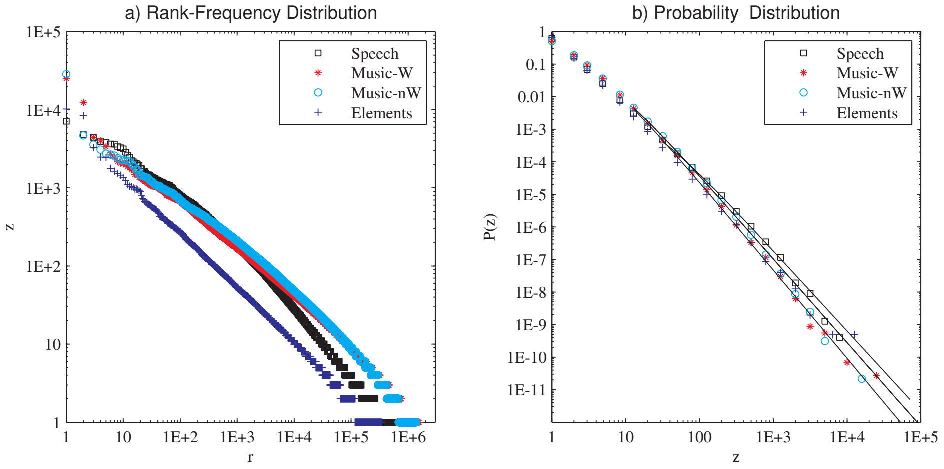

Figure 3.10: a) Rank-frequency distribution of MFCC code-words per database (frame size = 186 ms). b) Probability distribution of frequencies for the same code-words (the black lines correspond to the fitted distribution).

alent description in terms of the distribution of the frequency (Fig. 3.10b). The obtained results reveal that for all analyzed databases and frame sizes, the best fit corresponds to a particular type of power-law distributions called shifted (discrete) power-law. This distribution can be described by the following equation:

$$P(z) \propto (z + c)^{-\beta}, \tag{3.2}$$

where c is a constant value. By adding this constant value to Eq. 2.2 we obtain better fittings, specially in the low z region, whereas for the high z region the distribution tends to a pure power-law. In Table 3.4 a complete list of the fitted parameters can be seen.

From the obtained fitting results we observe that not only all the analyzed databases follow the same distribution type, but also that their exponents are somewhat similar (i.e. all the α exponents lie between 0.45 and 0.81). Regarding the effect of the frame size in the distribution exponent we can see that, for *Speech*, increasing the frame size seems to decrease the rank-frequency exponent α. The opposite effect is observed for *Sounds of the Elements*. Notably, in the case of *Western* and *non-Western Music*, changing the frame size has practically no effect in the distribution exponent. This high stability, also observed in the case of Bark-band code-words, is quite surprising given the fact that we are changing the frame size by almost

Frame size	z_{min}	$\boldsymbol{\beta}$	$\boldsymbol{c}$	$\boldsymbol{\alpha}$
Speech				
46 ms	3.20 (1.93)	2.23 (0.01)	0.76 (0.07)	0.81 (0.01)
186 ms	29.40 (23.43)	2.41 (0.22)	12.98 (12.07)	0.73 (0.12)
1,000 ms	32.00 (0.00)	3.22 (0.00)	36.90 (0.00)	0.45 (0.00)
Western Music				
46 ms	29.90 (21.63)	2.78 (0.08)	8.67 (3.26)	0.56 (0.03)
186 ms	7.50 (4.12)	2.64 (0.06)	1.90 (0.73)	0.61 (0.02)
1,000 ms	4.20 (0.63)	2.61 (0.02)	0.30 (0.10)	0.62 (0.01)
non-Western Music				
46 ms	82.20 (58.94)	2.76 (0.18)	27.85 (35.20)	0.57 (0.05)
186 ms	18.60 (2.95)	2.67 (0.05)	5.38 (1.25)	0.60 (0.02)
1,000 ms	8.50 (6.08)	2.66 (0.13)	1.65 (1.42)	0.61 (0.05)
Sounds of the Elements				
46 ms	8.10 (3.51)	2.70 (0.04)	2.35 (0.49)	0.59 (0.01)
186 ms	3.40 (0.97)	2.42 (0.02)	0.40 (0.07)	0.70 (0.01)
1,000 ms	4.20 (0.63)	2.29 (0.01)	0.15 (0.09)	0.78 (0.01)

Table 3.4: Fitting results for MFCC code-words per database and frame size. Average values from 10 random samples of 300,000 code-words per database and frame size are reported (standard deviation in parenthesis). z_{min} stands for minimum frequency for which the shifted power-law is valid, $\boldsymbol{\beta}$ corresponds to the frequency-distribution exponent, $\boldsymbol{c}$ refers to the constant value of the shifted power-law and $\boldsymbol{\alpha}$ corresponds to the rank-frequency exponent.

one and a half orders of magnitude (from 46 to 1,000 ms) and seems to be a unique feature of timbral music code-words.

To explore the differences between the most and least frequent MFCC code-words we select from each rank-frequency distribution the 200 most frequent and a random sample of 200 of the less frequent code-words per database. As done for Bark-band code-words the white color corresponds to those MFCC values encoded as zero and the black color to those quantized as one (Fig. 3.11). From this exploratory analysis we can clearly see that the most frequent code-words present characteristic structures while the least frequent ones show no detectable patterns. In particular, the most frequent code-words in *Speech* present a very distinctive structure, with some MFCC coefficients mostly quantized as zero (e.g. coefficients 2, 6, 8, and 17) and some others mostly quantized as one (e.g. coefficients 1, 4, 7, and 10). This distinctive pattern in *Speech* is particularly intriguing, specially given the

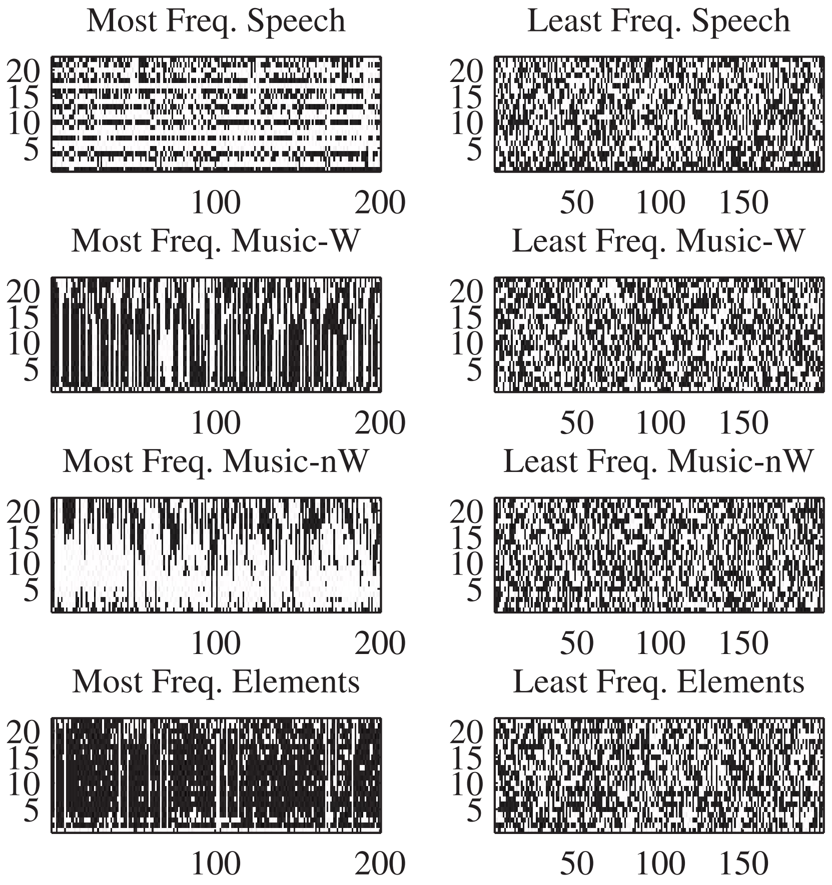

Figure 3.11: Most (left) and least (right) frequent MFCC code-words per database using a frame size of 186 ms. For each plot, the horizontal axis corresponds to individual code-words and the vertical axis corresponds to quantized MFCC coefficients (white = 0, black = 1). Every position in the abscissa represents a particular code-word.

fact that the MFCC descriptor was originally designed to describe speech signals. Furthermore, it is not only that the most frequent code-words of speech are quite different from the ones in the other type of sounds but also, when computing and plotting the smoothness s (see Sec. 3.3.1) we see a different behavior for speech sounds (Fig. 3.12). We leave this issue for future research.

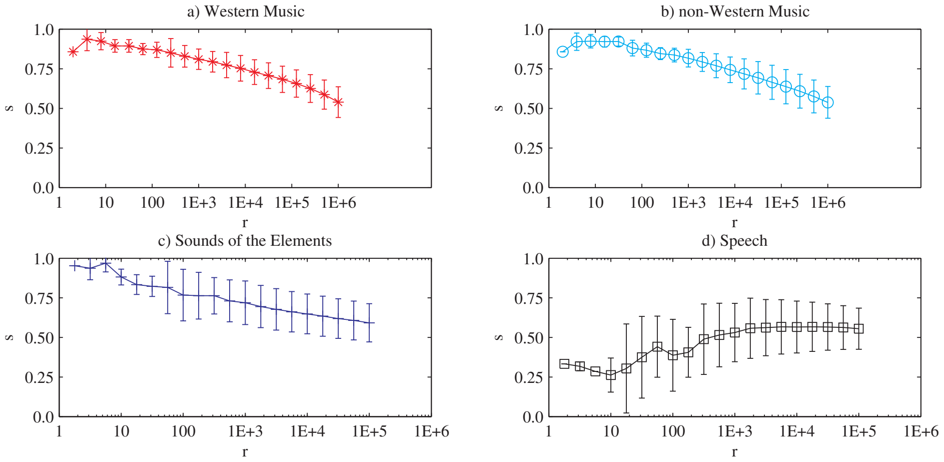

Figure 3.12: Smoothness values (s) per database. Notice that in three "non-Speech" databases the most frequent code-words present a smooth structure, with close/neighboring MFCC coefficients having similar quantization values. In the case of *Speech* the smoothness value seems to somehow stable across all rank values. For a better visualization we plot the mean and standard deviation of the smoothness value of 20 log-spaced points per database (frame size = 186 ms).

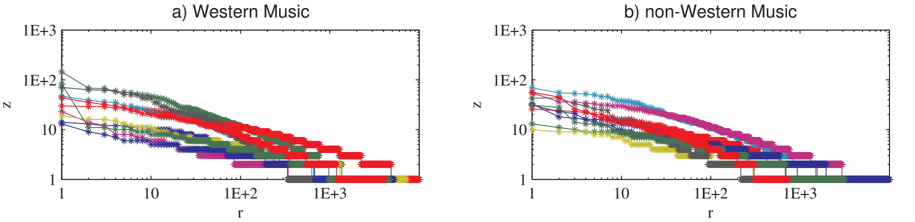

Figure 3.13: Example of rank-frequency distributions of MFCC code-words from 10 randomly selected music recordings per database using a frame size of 46 ms. Each line type corresponds to one recording.

We further investigate the rank-frequency distribution of MFCC code-words of randomly selected audio segments of up to 6 minutes in length. We observe the same behavior as with Bark-band code-words namely, a similar heavy-tailed distribution as the one found for the whole databases. Examples of the obtained distributions, in this case for randomly selected songs from the music databases, can be seen in Fig. 3.13.

3.4 Discussion and conclusion

In this chapter we have analyzed the rank-frequency distribution of encoded timbral descriptors computed out of 740 hours of real-world sounds coming from four sound categories that represent a large portion of the timbral variability perceivable in the world - i.e. *Speech*, *Western Music*, *non-Western Music*, and *Sounds of the Elements*. In this analysis we work with two of the most commonly used timbral descriptor namely Bark-bands and MFCCs. Whilst both descriptors mainly characterize the shape of the spectral envelope of a short-time audio excerpt (or frame), the Bark-band descriptor, due to its straightforward computation, it is easier to relate-back to an envelope shape. This feature motivated us to study the Bark-band code-words in a more exhaustive manner. Noticeably, with respect to their rank-frequency distributions, both descriptor's code-words are power-law distributed in all four databases. In the case of Bark-band code-words this distribution can be further characterized as a Zipfian distribution (i.e. a power-law with exponent close to one) and, in the case of MFCC code-words as a shifted power-law. We also exhaustively analyzed the robustness of the found distributions against several encoding variables such as frame-size, frequency bands, audio length, etc. All robustness experiments showed that the power-law distribution of timbral code-words is very stable and seems to be independent of the type of sounds analyzed and the encoding method. Our results also indicate that regardless of the analyzed database, the most frequent timbral code-words have a more homogeneous structure. This implies that, for frequent code-words, proximate descriptor bands tend to have similar encoded values (except for the case of MFCC code-words of *Speech* where a different pattern is observed for frequent code-words).

Regarding the shared Bark-band code-words among databases we found several interesting patterns. In particular, the presence of database-specific code-words in both speech and music, and the absence of such distinctive code-words for *Sounds of the Elements*. This suggests that these natural sounds have been incorporated, possibly by imitation, within the human-made "palette" of timbres. Noticeably, it has been recognized that human vocal imitation, which is central to the human language capacity, has received insufficient research attention (Hauser et al., 2002). Moreover, a recent work (Assaneo et al., 2011) has suggested a mechanism by which vocal imitation naturally embeds single sounds into more complex speech structures. Thus, onomatopoeic sounds are transformed into the speech elements that minimize their spectral difference within the constraints of the vocal system. In this context, our observations could be taken as supporting the role of imitation within language and music evolution.

The fact that 40% of the Bark-band code-word dictionary remained unused after 740 hours of sounds suggests that this dictionary was big enough to accommodate the different timbral variations present in the databases, but it also poses the question about the reasons for this behavior. It could be that the unused spectral envelopes were unlikely (in physical-acoustical terms) or, perhaps, that animal sounds and urban soundscapes (the two large categories that have not been included in our study) would account for that.

In the light of all these findings, the establishment of a power-law rank-frequency distribution seems to be a physical property of short-time spectral envelopes of sound signals. Thus, all our experiments point towards this intrinsic property of spectral envelopes, where a few spectral shapes are extremely repeated while most of them are very rare and, at the same time, there is no characteristic separation between both groups. All this suggests that, as in the case of text corpora (Ferrer i Cancho and Solé, 2003), the most frequent code-words are also the least informative ones[8]. That is, the highly frequent code-words possess little discriminative power because they are present in many types of sounds. Moreover, we argue that the existence of such scale-invariant distribution should have some influence on the way perception works given that the perceptual-motor system reflects and preserves the scale invariances found in the statistical structure of the world (Chater and Brown, 1999). Following this line of thought, we hypothesize that any auditory system, being natural or artificial, should exploit the here-described distribution and characteristics of short-time spectral envelopes in order to achieve an optimal trade-off between the amount of extracted timbral information and the complexity of the extraction process. Furthermore, the presented evidence could provide an answer to the question posed by Bregman in his seminal book *Auditory Scene Analysis* (Bregman, 1990):

> [...] the auditory system might find some utility in segregating disconnected regions of the spectrum if it were true in some probabilistic way that the spectra that the human cares about

[8] Informative in the sense of information theory's self-information concept, where the self-information (or surprisal) $I(w_n)$ of a code-word w_n is defined as $I(w_n) = -log(P(w_n))$, where $P(w_n)$ is the probability of occurrence of the code-word. Therefore, the bigger the code-word's probability, the smaller its self-information.

> tend to be smoothly continuous rather than bunched into isolated spectral bands.

According to our findings, these smoothly continuous spectra correspond to the highly frequent elements in the power-law distribution. We expect these highly repeated elements to quickly provide general information about the perceived sources (e.g. is it speech or music?). On the other hand, we expect that the rare spectral envelopes will give information about specific characteristics of the sources (e.g. the specific type of guitar that is being perceived).

Since we have found similar distributions for medium-time (i.e. a few minutes) than for long-time (i.e. many hours) code-word sequences, this behavior has direct practical implications that we would like to stress and explore in the following chapters. One practical implication is that when selecting random short-time audio excerpts (using a uniform distribution), the big majority of the selected excerpts will belong to the most frequent code-words. Therefore, the knowledge extracted from such data sample will represent these highly frequent spectral envelopes but not necessary the rest of the elements. This is the case in two recently published papers (Bigand et al., 2011; Plazak and Huron, 2011) where the perception of randomly selected short-time audio excerpts was studied. Moreover, auditory gist perception research (Harding et al., 2008) could also benefit from knowing that spectral envelopes are heavy-tailed distributed. Thus, future gist perception studies can evaluate how fast and accurate we recognize stimuli formed by code-words from different parts of the distribution.

Another area on which the found heavy-tailed distributions will have practical implications is within audio-based technological applications that work with short-time spectral envelope information. For instance, as described in Chapter 1, in automatic audio classification tasks it is common practice to use an aggregated spectral envelope as timbral descriptor. That is, all the short-time spectral envelopes that form an audio file are aggregated into one mean spectral envelope. This mean envelope is then used to represent the full audio file, e.g. one song. Evidently, computing statistical aggregates, like mean, variance, etc. on a set that contains highly frequent elements will be highly biased towards the values of this elements. In audio similarity tasks, the similarity between two sounds is usually estimated by computing a distance measure between sequences of short-time spectral envelope descriptors (Klapuri and Davy, 2006), e.g. by simply using the Euclidean distance. Again, these computations will be highly biased towards those highly frequent elements. Therefore, the influence these biases have on each task should be thoroughly studied in future research. It could be the case that for some applications, considering only the most frequent spectral envelopes is the best solution. But, if we look at other research areas that deal with heavy-tailed data we can see that the information extracted from the distribution's tail is at least, as relevant as the one extracted from the most frequent elements (Liu, 2011; Manning and Schütze, 1999).

Finally, the relationship between the global power-law distribution present in long-time audio sequences, and the local heavy-tailed distributions depicted by medium-time sequences should be also further studied. For instance, in text information retrieval, these type of research has provided improved ways of extracting relevant information (Baeza-Yates, 1999). Therefore, it is logical to hypothesize that this will be also the case for audio-based technological applications.

Chapter 4

A plausible power-law generative model

Most of the results presented in this chapter were published in Haro et al. (2012b).

4.1 Introduction

In Chapter 3 we have found that the frequency distribution of encoded timbral descriptors follows a power-law distribution. In particular, our experiments indicate that, regardless of the sound source, rank-frequency distributions of encoded short-time spectral envelopes show a Zipfian distribution (i.e. a power-law with exponent close to one).

As mentioned in Sec. 2.1, power-laws are highly common in both natural and human-made phenomena. This ubiquitous presence has increasingly attracted research interest over the last decades, specially in trying to find generative mechanisms and unifying principles that link and govern such disparate complex systems. Thus, several models have been proposed including the least effort principle (Ferrer i Cancho and Solé, 2003; Zipf, 1949), preferential attachment (Barabasi et al., 1999; Cattuto et al., 2007; Peterson et al., 2010; Simon, 1955), multiplicative dynamics (Montroll and Shlesinger, 1982), superposition of independent stochastic signals (Eliazar and Klafter, 2009), proportional growth (Saichev et al., 2010), extinction dynamics (Newman and Palmer, 2003), coherent noise (Sneppen and Newman, 1997), self-organized criticality (Bak et al., 1987) or general stochastic systems (Corominas-Murtra and Solé, 2010). See Mitzenmacher (2003), Sornette (2004), and Newman (2005) for excellent reviews on this subject.

In the following sections we provide early evidence that the power-law behavior displayed by encoded short-time spectral envelopes could be generated by a mechanism proposed by Cattuto et al. (2007). This model is a modification of the original "rich-get-richer" model originally described by Simon (1955).

4.2 Method

Given the straightforward mapping between Bark-bands and spectral envelopes, we aim at finding a power-law generative model that matches the distribution of Bark-band code-words as described in Sec. 3.3.1. Finding this generative model could provide more evidence regarding the empirically observed power-law behavior in short-time spectral envelopes of sound signals. In particular, we select from the literature those models that were both simple and applicable to our case.

Thus, we have taken into consideration the following characteristics of our data. First, although a sequence of short-time spectral envelopes constitutes one of the relevant information sources used in the formation of auditory units (Bregman, 1990), individual Bark-band code-words cannot be seen as communication units like in the case of musical notes, phonemes, or words. Second, we have here found the same distribution for processes that involve a sender and a receiver (like in speech and music sounds) and for processes that do not involve an intelligent sender (like inanimate environmental sounds). Therefore, we do not consider generative models that imply a communication paradigm, or any kind of intentionality or information interchange between sender and receiver (e.g. like in the case of the "least effort" model (Ferrer i Cancho and Solé, 2003; Zipf, 1949)).

As for the generative models that are plausible to be applied to our data, we consider two simple preferential attachment models namely: the classical Yule-Simon model (Simon, 1955) and Catutto's model (Cattuto et al., 2007). The Yule-Simon model was originally proposed by Udny Yule (1925) as an explanation of the empirical data on the abundances of biological genera. In 1955 Simon, trying to model the power-law distribution of word-frequencies in text, re-introduced this model in a more mathematically-elegant way (Simon, 1955). A similar behavior is also present in the preferential attachment model proposed by Barabasi et al. (1999) when modeling the growth of networks. The Yule-Simon power-law generative model can be summarized as follows: consider a dictionary of discrete elements, our goal is to form a stream of elements that once examined according to its rank-frequency distribution it produces a power-law behavior. In the original Yule-Simon model, at each time step, a new element (a code-word in our case) is extracted from the dictionary with constant probability q, whereas an existing code-word (an element already present in the stream) is uniformly selected with probability $\overline{q} = 1-q$. This model generates a power-law rank-frequency distribution $P(z) \propto z^{-\beta}$ where $\beta = 1 + 1/\overline{q}$. Since, the power-law is created by favoring the appearance of already used elements, this model is sometimes referred as "rich-get-richer", "cumulative advantage" or "preferential attachment".

Recently Cattuto et al. (2006, 2007) proposed a modification of the original Yule-Simon model. In Cattuto's model a hyperbolic memory kernel is introduced in a way that when selecting an existing code-word, the kernel promotes recently added code-words thus favoring small time gaps between identical elements. That is, instead of choosing uniformly from past code-words (as with the Yule-Simon model), this model selects a past code-word that occurred i time steps behind with a probability K that decays with i as a power-law,

$$K(i) = \frac{C(t)}{\tau + i} \tag{4.1}$$

where τ is a characteristic time-scale over which recent code-words have similar probabilities and $C(t)$ is a time-dependent normalization factor of the form:

$$C(t) = \left(\sum_{i=1}^{i=t} \frac{1}{\tau + i} \right)^{-1} . \tag{4.2}$$

Each realization of the model starts with n_0 code-words and, at each time step t, a new code-word may be introduced with probability q, while with probability $1 - q$ one code-word is copied from the existing stream going back i time steps in the past with a probability $K(i)$.

When evaluating the results produced by both models we consider as output of the model the rank-frequency curve produced after averaging 50 realizations with identical parameters. In all cases we generate a stream of code-words having the same length as each of the databases we aim at modeling. In order to match each model's output against the empirically observed Bark-band distributions we do a grid search over each model's parameters. This grid search aims at minimizing the sum of the squares of the distance (i.e. least-squares fitting) between the data points generated by the model and the one from the empirical distributions. Finally, we define the inter code-word distance as the number of code-words found between two identical and consecutive code-words plus one. Then, we visually compare the histogram of inter code-word distances for the 10 most frequent code-words per database against the inter code-word distance histogram as produced by the best matching model.

4.3 Results

As stated in the previous section we first perform a grid search over Yule-Simon's parameter q and evaluate the obtained distributions against the rank-frequency Bark-band plots for each database. However, no q value produced an even close fit to our data.

Next, we explore the histogram of inter code-word distances for the 10 most frequent Bark-band code-words per database (Fig. 4.1). From these plots we can see that, in general, the most frequent inter code-word distances correspond to short time gaps. This behavior corresponds with the modification of the Yule-Simon model proposed by Cattuto et al. (2007) where recently occurred code-words have more probability of being selected (see Sec. 4.2). When considering this modified Yule-Simon model a reasonable fitting is observed for all rank-frequency distributions. In this case we perform a grid search for the model's parameters q, τ, and n_0, that correspond to the probability of adding a new code-word, the memory parameter, and the number of initial code-words respectively. In Table 4.1 the resulting model parameters per database can be seen. These parameters produce the output depicted in Fig. 4.2 where we can see the model-produced rank-frequency distributions overlapped with the Bark-band code-word distribution per database.

Finally, we explore the inter code-word distances for the 10 most frequent code-words produced by Cattuto's model per database. In Fig. 4.3 we can see the inter code-word distance histogram for one random realization of Cattuto's model. By comparing the histogram of Fig. 4.3 with the one obtained from the sound databases (Fig. 4.1) we can see that although the histograms are not identical, a similar behavior is present in both cases

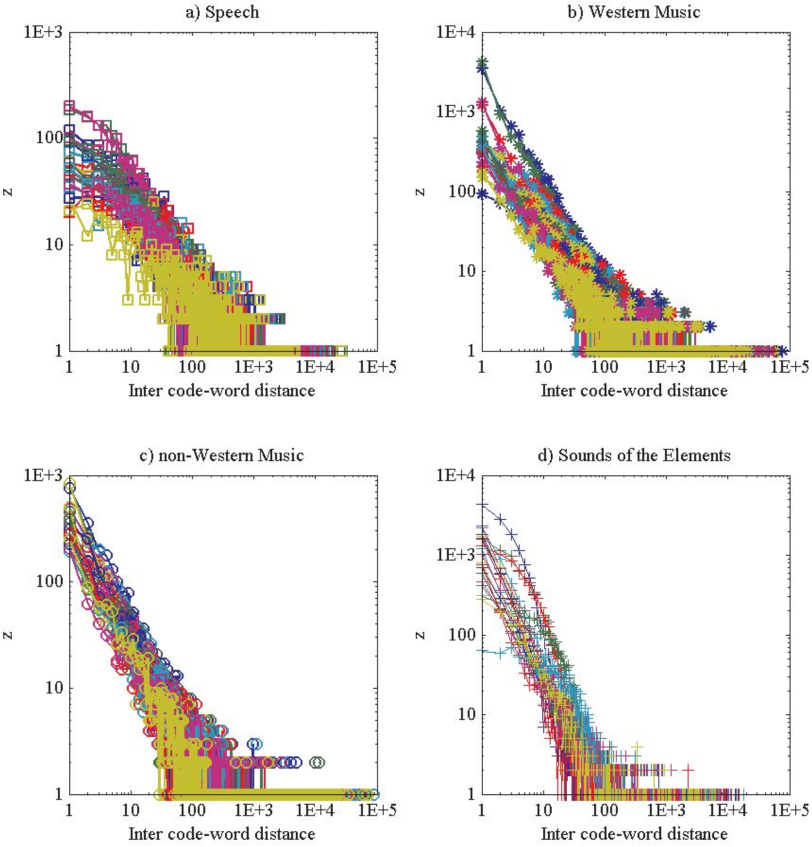

Figure 4.1: Inter Bark-band code-word distance for the 10 most frequent code-words in *Speech* (a), *Western Music* (b), *non-Western Music* (c) and *Sounds of the Elements* (d) databases (frame size = 1,000 ms). Each color represents one particular code-word.

namely a heavy-tailed distribution of inter code-word distances for approximately the same numeric ranges.

4.4 Discussion

The aim of this chapter was to find one power-law generative model that could generate the empirically observed rank-frequency distributions of Bark-band code-words described in Chapter 3. With this goal in mind, we explore

Modeled Database	q	τ	n_0
Speech	0.11	250	200
Western Music	0.095	250	15
non-Western Music	0.12	150	100
Sounds of the Elements	0.05	1,000	50

Table 4.1: Selected parameters in Cattuto's model. The parameters q, τ, and n_0 correspond to the probability of adding a new code-word, the memory parameter, and the number of initial code-words respectively.

and select, from the vast amount of power-law generative models proposed in the literature, two plausible models that could be easily applied when trying to mimic the empirically observed distributions namely: the classic Yule-Simon model and Cattuto's model.

We have found that the modified version of the Yule-Simon model proposed by Cattuto et al. (2007) provides a reasonable quantitative account for the observed distribution of Bark-band code-words. This fact suggests the existence of a common generative framework for all considered sound sources. This model also implies a fundamental role of temporally close events. In our case, this means that when repeating pre-occurred spectral envelopes, those that have occurred recently have more chance to reappear.

This simple generative mechanism could possibly act as universal framework for the generation of timbral features. In particular, we know that the analyzed sounds are formed by mixtures of individual sources (e.g. notes simultaneously played by several musical instruments). Most of these individual sources can be modeled by an excitation-resonance process (Berg and Stork, 1995). That is, an excitative burst (or series of bursts) of decaying energy that goes through biological or human-made structures that impose certain acoustic properties on the original spectrum of the burst (e.g. the spectrum of the burst produced by the vocal folds is modulated/filtered by the shape of the vocal tract). Thus, the intrinsic characteristics of this resonance structure will favor the close reappearance of certain types of spectral envelopes every time the resonance structure is excited. This temporally close reappearance is properly reproduced by Cattuto's model.

As future research, it would be both artistically and scientifically interesting to try to sonify the stream of code-words produced by Cattuto's model. It would be interesting to know if the code-word streams produce sound sequences somehow related with the type of sounds present in the databases

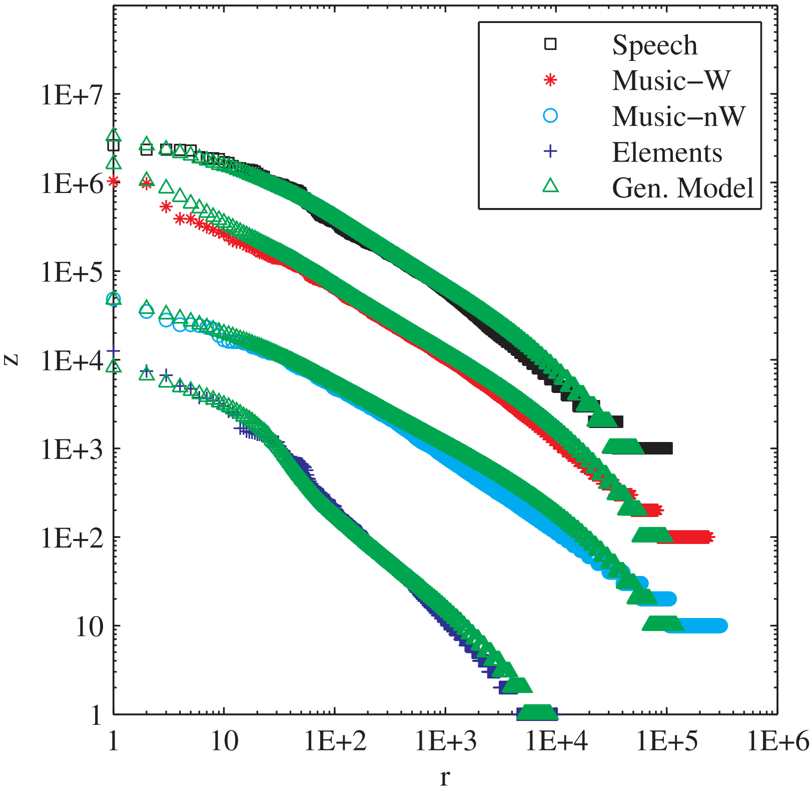

Figure 4.2: Rank-frequency distribution of Bark-band code-words (frame size = 1,000 ms) and Cattuto's model (Cattuto et al., 2007) per database. *Gen. Model* stands for the computed generative model. For clarity's sake the curves for *non-Western Music*, *Western Music*, and *Speech* are shifted up by one, two, and three decades respectively. All model's curves were computed by averaging 50 realizations with identical parameters.

we try to model in the first place. These future experiments could also help us to further explore the impact of power-law distributions as a new tool for algorithmic music composition.

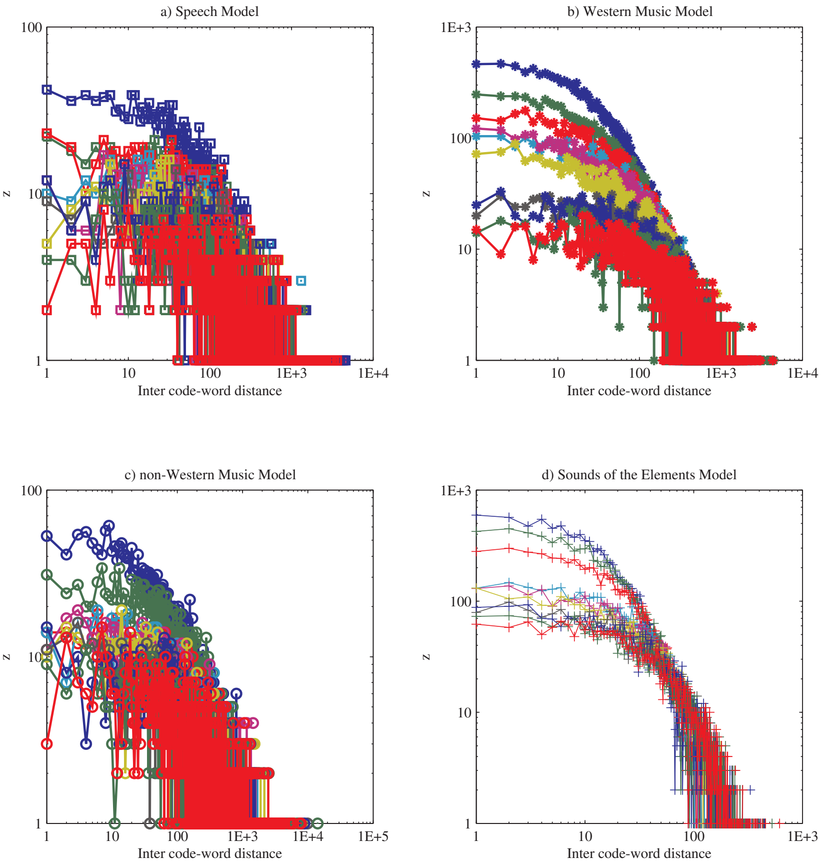

Figure 4.3: Inter code-word distance for the 10 most frequent code-words in one random realization of Cattuto's model of *Speech* (a), *Western Music* (b), *non-Western Music* (c) and *Sounds of the Elements* (d) databases (frame size = 1,000 ms). Each color represents one particular code-word.

Chapter **5**

Song-level distribution assessment

Most of the results presented in this chapter were published in Haro et al. (2012a).

5.1 Introduction

As stated in Chapter 1, it is common practice within the MIR community to build automatic classification algorithms using aggregated descriptor sets (Casey et al., 2008; Klapuri and Davy, 2006). For instance, the content of several minutes of audio (e.g. an entire song) is represented by a real-valued vector containing the mean values of the frame-based audio descriptors (and often their variances and covariances). Moreover, many technological applications dealing with audio signals use Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980) as main timbral descriptor (Casey et al., 2008; Klapuri and Davy, 2006; Müller et al., 2011; Quatieri, 2001).

A standard bottom-up implementation for timbral-based audio classification (e.g. for automatic musical genre recognition) can be implemented as follows: first, the MFCC coefficients from consecutive short-time audio frames (usually with lengths below 100 ms) are computed. Then, the mean value of each MFCC coefficient is computed thus obtaining one mean-MFCC vector per song. Finally, these mean vectors are introduced as features in the machine learning algorithm. This approach is usually referred in the literature as the "Bag-of-Frames" (BoF) approach and is often used by researchers as comparison baseline when evaluating new classification algorithms.

As stated in Chapter 1, these types of procedures assume a certain homogeneity in the MFCC vector space. Otherwise, the results obtained from computing statistical moments, such as mean or variance, will be highly biased towards the values of those extremely populated areas (i.e. those extremely frequent MFCC vectors). Nevertheless, in Chapter 3 we have shown that at the database level, encoded timbral descriptors (including MFCCs) are power-law distributed, thus being unhomogeneous and extremely biased towards certain values. We have also shown that individual recordings (e.g. individual songs) seem to show the same type of rank-frequency distribution (see Fig. 3.13).

In this chapter, we provide additional evidence to support the claim that MFCC vectors from individual music recordings are also heavy-tailed distributed. Our working hypothesis is the following: if a set of MFCC vectors presents a heavy-tailed distribution, then, when computing the mean of such vectors the resulting values will be highly biased towards those few extremely frequent vectors (i.e. those MFCC vectors that belong to the most frequent code-words within the set). Therefore, this bias will imply that computing the aggregated mean vector, used as input for BoF algorithms, from just those few highly frequent MFCC vectors, will yield similar classification results as when computing the mean vector from all frames.

We evaluate this hypothesis with two supervised semantic inference tasks: automatic genre classification and musical instrument identification. Thus, our main goal is to compare the classification results obtained when using all audio frames versus those obtained when using a distribution-based reduced set of selected frames to compute the mean feature vector.

5.2 Method

In order to compare the standard BoF approach against a "Selection-of-Frames" strategy we set up two music classification algorithms namely: automatic musical genre classification and musical instrument identification. Since our goal is not to achieve state-of-the-art performances but, to compare two identical classification strategies that only differ in the frame-selection step, we deliberately choose a simple pattern recognition strategy. In particular, we use support vector machines (SVM) (Cortes and Vapnik, 1995) to classify aggregated feature vectors of 22 MFCC-mean values per audio file (see Sec. 2.2.1 for details regarding the MFCC descriptor and Sec. 2.4 for further information about SVM).

The general framework to compute the "bag-of-frames" approach can be described as follows:

1. Each audio file is segmented into frames of 46 ms (with 50% overlap).
2. For each audio frame 22 MFCC coefficients are computed.
3. A mean MFCC vector is computed by taking the mean of each MFCC coefficient across all frames in the audio file.
4. Each mean MFCC vector is used as song-level descriptor for the SVM classification algorithm.

The framework for the selection-of-frames approach is identical to the bag-of-frames framework except for step number 4. Here, instead of computing a mean MFCC vector with all frames in the audio file, we compute it from a reduced set of pre-selected frames. To select these frames we first encode each MFCC frame into its corresponding MFCC code-word (see Sec. 3.2.3). Next, for each audio file we count the frequency of use of each code-word and sort them by decreasing order of frequency (i.e. we build the song's rank-frequency distribution). Then, we select the N most frequent MFCC code-words of the audio file. Finally, we randomly choose one original MFCC descriptor per code-word. Thus, at the end of this process we have N selected MFCC vectors per audio file that are used to compute the mean MFCC feature vector. Therefore, those selected MFCC vectors belong to the most frequent code-words of the music recording. Finally, we execute this same procedure except that, in this case, we select the N least frequent code-words. In all selection strategies we report mean classification results after running five times each particular framework.

5.2.1 Database

The audio files used in these experiments do not form part of the databases described in Sec. 3.2.1. For the genre classification task we use an in-house collection of 400 full songs extracted from radio recordings. The songs are equally distributed among 8 genres: hip-hop, rhythm & blues, jazz, dance, rock, classical, pop, and speech[1]. The average length of these audio files is 4 min 18 s (9,853 frames). This dataset was defined by musicologists and previously used in Guaus (2009). For the musical instrument identification task we use an in-house dataset of 2,355 audio excerpts extracted from commercial CDs (Fuhrmann, 2012). These excerpts are labeled with one out of 11 possible instrument labels. Each label corresponds to the most salient instrument in the polyphonic audio segment. The audio excerpts are distributed as follows: piano (262), cello (141), flute (162), clarinet (189), violin (182), trumpet (207), saxophone (233), voice (265), organ (239), acoustic guitar (221), and electric guitar (254). The average length for these excerpts is 19 s (828 frames).

5.2.2 Evaluation metrics

To evaluate the classification results we select the best F-measure (see Sec. 2.4) result obtained after performing 10-fold cross-validation in each database. In all cases we keep the best classification result after evaluating four SVM kernels with default parameters[2] (i.e. rbf, linear, and polynomial of degree 2 and 3). Notice that according to each label distribution the F-measure results for a random classification baseline are 2.77% and 1.83% for the genre and instrument datasets respectively.

5.3 Results

The obtained F-measure results for both genre and musical instrument classification can be seen in Fig. 5.1. In both classification tasks we confirm our working hypothesis, i.e. we obtain nearly the same classification results by selecting very few properly selected MFCC vectors than using all frames. In particular, by taking only 50 frames belonging to the 50 most frequent code-words we obtain classification accuracies that are similar to those obtained when using all the frames in the audio file. Importantly, we should notice that 50 frames correspond to just 0.5% of the average song length

[1] The speech audio files consist of radio speaker recordings with and without background music.

[2] We use the LibSVM implementation: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

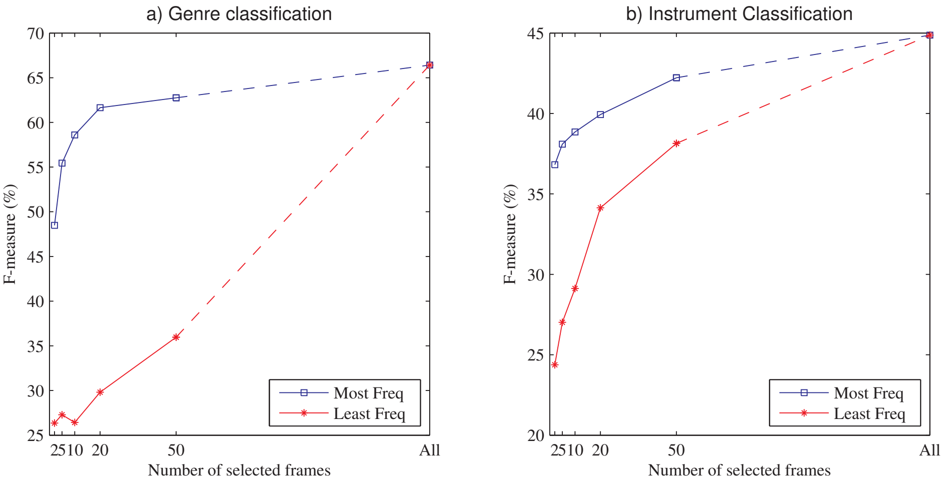

Figure 5.1: Genre (a) and Instrument (b) F-measure classification results (%). These classification results correspond to the selection of either the N most frequent code-words per song (blue line) or the N least frequent code-words per song (red line). The 'ALL' label corresponds to selecting all frames in the song to compute the aggregated mean MFCC vector (i.e. the classic bag-of-frames approach). The selection results correspond to mean F-measure values after running five times each selection framework.

of the genre dataset and 6% of the average sound length of the instrument dataset. The obtained results also show that, in both tasks, selecting the N least frequent code-words delivers systematically poorer results than selecting the N most frequent ones. In particular, the difference between both selection strategies is considerably large in the genre classification task where we obtain, on average, 28.2% worst results when selecting the least frequent code-words (see Table 5.1). In the case of instrument identification we obtain, on average, 8.6% worst results when using this strategy. Notice that in this case we are working with short audio excerpts, which could indicate that the heavy-tailed distribution is not as pronounced as when working with bigger audio segments (e.g. full songs).

Finally, if MFCC frames follow a heavy-tailed distribution, this should imply that when taking N random frames from the bag-of-frames (using uniform distribution) there is a very high probability that those selected frames belong to the most frequent MFCC code-words (because those code-words are very common). Therefore, similar classification results as the ones obtained

	Number of selected frames (*N*)					
Strategy	**2**	**5**	**10**	**20**	**50**	**All**
Genre						
Most Frequent Code-Words	48.49	55.44	58.59	61.65	62.75	66.42
Least Frequent Code-Words	26.36	27.28	26.43	29.81	35.96	66.42
Difference	22.14	28.15	32.16	31.83	26.79	0.00
Instrument						
Most Frequent Code-Words	36.81	38.09	38.85	39.93	42.22	44.87
Least Frequent Code-Words	24.38	27.02	29.12	34.14	38.14	44.87
Difference	12.43	11.07	9.73	5.80	4.08	0.00

Table 5.1: Genre and instrument F-measure classification results (%). We compare two frame selection strategies: taking N MFCC vectors that belong to either the most or less frequent code-words of each audio file. The last column includes the classification result obtained when using the mean of all the frames in the recording. The differences between both classification strategies are also shown. The selection results correspond to mean F-measure values after running five times each selection framework.

	Number of selected frames (*N*)					
Strategy	**2**	**5**	**10**	**20**	**50**	**All**
Genre						
Most Frequent Code-Words	48.49	55.44	58.59	61.65	62.75	66.42
Random Frames	43,99	51,62	56,19	59,26	64,72	66.42
Instrument						
Most Frequent Code-Words	36.81	38.09	38.85	39.93	42.22	44.87
Random Frames	33,57	38,15	40,26	42,23	43,55	44.87

Table 5.2: Genre and instrument F-measure classification results (%) for two frame selection strategies: taking N MFCC vectors that belong to the most frequent code-words of each audio file, and selecting N random MFCC vectors with uniform distribution. The last column includes the classification result obtained when using the mean of all the frames in the recording. The selection results correspond to mean F-measure values after running five times each selection framework.

after selecting the N most frequent code-words should be obtained. Noticeably, once more the experimental results confirm our hypothesis. Table 5.2 shows the obtained classification results when selecting N random MFCC vectors from a uniform distribution, and the aforementioned results after taking the N most frequent code-words. As can be seen in the table both strategies provide similar results.

5.4 Discussion and conclusion

In this chapter we have presented two supervised semantic inference tasks that provide further evidence that MFCC code-words from individual recordings have the same type of heavy-tailed distribution as found in the large-scale databases. Such heavy-tailed distributions allow us to obtain similar classification results when working with just 50 highly-frequent frames per audio file as when using all frames in the file (e.g. reducing the total number of processed frames to 0.5% in the case of full songs). Moreover, since MFCCs are heavy-tailed distributed, when taking N random frames from the bag-of-frames (using uniform distribution) there is a very high probability that those selected frames correspond to the most frequent MFCC code-words. Therefore, high classification results are also achieved by just selecting a few random frames to represent each song. Importantly, this fact could lead to faster classification algorithms that work well with big datasets.

Another area where the presented results could have a major impact is in audio similarity tasks. Here, the highly frequent MFCCs should introduce a tremendous bias in some distance measures and could be the underlying cause of "hub" songs (i.e. songs that appear similar to most of the other songs in a database without having any meaningful perceptual similarity; Flexer et al., 2010, see also Chapter 1). We hypothesize that "hubs" are produced by songs that, according to the distance measure, are close to many other songs because they share a great amount of highly frequent descriptor frames. Whereas, at the same time, these extremely frequent frames are not the ones our perception rely on in order to determine timbral similarities because they are too common (i.e. do not have enough discriminative power). Noticeably, Aucouturier and Pachet (2008) discovered that in typical music databases, hub songs are distributed along a scale-free distribution. From our perspective this scale-free distribution should be some how linked with the underlying heavy-tailed distribution of MFCCs and short-time spectral envelopes. Furthermore, since audio similarity is at the core of audio-based recommender systems, improving the former will also benefit the latter.

Finally, the relationship between global (i.e. database-level) and local (i.e. song-level) distributions should be further considered. For that purpose, we can use the huge amount of mining techniques developed by the text retrieval community. For instance, we could try to remove the highly frequent code-words as found in the global distribution, since these code-words could be considered as analogous to stop words in text processing[3]. We could also try to apply different weights to every frame by using an adaptation of the tf-idf weighting scheme commonly used in text mining tasks (Baeza-Yates, 1999). Later on, these weighted MFCC frames could be used in classification or audio similarity tasks. In Chapter 7 we further explore these ideas.

[3] Stop words are highly frequent words that offer little information when processing natural language text. Some examples of such words are: the, is, at, which, on, etc.

Chapter 6

Measuring the evolution of popular Western music

Most of the results presented in this chapter were published in Serrà et al. (2012b).

6.1 Introduction

In Chapter 3 we have reported our findings regarding rank-frequency distributions of audio timbral descriptors for various types of sound. Our results show that independently of the sound source, encoded timbral descriptors are power-law distributed. In particular, this distribution is remarkably stable for music databases. Furthermore, in Chapter 5 we have presented further evidence that encoded MFCC frames of individual music recordings are also heavy-tailed distributed. Consequently, in this chapter we further concentrate our analysis on musical recordings.

We know that music is a human universal involving perceptually discrete elements displaying organization (Patel, 2007). In other words, as stated by Edgard Varèse, music is organized sound (Roads, 2001). Thus, contemporary popular music may have a well-established set of "rules" that materialize in underlying patterns and regularities such as well established chord sequences, instrument combinations, etc. (Ball, 2010; Honing, 2011; Huron, 2006; Patel, 2007). Some of these "rules" could be inherited from the classical tradition (Lerdahl and Jackendoff, 1983; Levitin et al., 2012; Temperley, 2007). Nevertheless, as an incomparable artistic product for conveying emotions (Juslin and Sloboda, 2001), music must incorporate variation over such patterns in order to play upon people's memories and expectations. Possibly, the "right" combination between known patterns and unexpected variations is what makes music so attractive to listeners (Honing, 2011; Huron, 2006; Levitin et al., 2012)[1]. For the very same reasons, long-term variations of the underlying patterns may also occur across years (Reynolds, 2005). Unfortunately, many of these aspects remain formally unknown or lacking of scientific evidence, specially the latter, which is very often neglected in music-related studies, from musicological analyses to technological applications. Interestingly, current MIR technologies (Casey et al., 2008; Müller et al., 2011) are starting to provide excellent tools to study the evolution of those underlying patterns under objective, empirical, and quantitative premises. Moreover, akin to recent advances in other cultural assets (Michel et al., 2011), they allow for unprecedented large-scale analyses. In resonance with Aucouturier and Bigand (2013), we believe that MIR technologies, if applied correctly, can shed new light on music-related fields such as cognitive psychology, neuroscience, and musicology. Thus, traditional human-based intelligent subjective analysis can be complemented with machine-based "unintelligent" objective large-scale information.

Therefore, in this chapter we study the evolution of popular music under the aforementioned premises and large-scale resources. We take advantage of the tools and concepts described in previous chapters of this thesis to unveil a number of statistical patterns and metrics characterizing the general usage of pitch[2], timbre, and loudness within contemporary Western popular music. Our working hypothesis is that the characterization of the yearly-based changes on the statistical distribution of audio descriptors that account for key musical facets will provide objective, empirical, and quantitative information regarding the long-term temporal evolution of music.

In particular, we take advantage of the existence of a publicly available database of audio descriptors and metadata for approximately one million songs called *The Million Song Dataset* (MSD) (Bertin-Mahieux et al., 2011). We investigate the distribution of the dataset's audio descriptors that ac-

[1] It is worth to mention here the words of the composer Arnold Schöenberg on dissonance (Schönberg, 1983): "Two impulses struggle with each other within man: the demand for repetition of pleasant stimuli, and the opposing desire for variety, for change."

[2] Within this chapter we use the term pitch to refer to tonal information expressed as the per-note energy level of an audio segment collapsed into one octave (i.e. a 12-dimensional chroma feature (Casey et al., 2008)).

count for the three primary and complementary sensations associated with music perception: timbre, pitch and loudness (Ball, 2010). In this case, when choosing our analysis units we have decided that, instead of working with fixed-length audio frames, it would be more relevant to work with audio segments that correspond with rhythmic beats. In particular, rhythmic beats form a discrete time grid that corresponds to the rate at which most people would tap or clap in time with the music (Gouyon and Dixon, 2005). This decision of working with beat-based segments is substantiated in three facts: first, we are now exclusively dealing with music signals, second, the beat interval is probably the most relevant temporal unit in music, specially in Western popular music (Ball, 2010; Honing, 2011) and, third, the MSD already provides, besides the classic frame-based descriptors, beat-based representations. Finally, we use the metadata information that accounts for the year in which each song in the dataset was released to study the temporal evolution of popular Western music from 1955 to 2010.

6.2 Method

6.2.1 Database

The million song dataset (Bertin-Mahieux et al., 2011) is a publicly available collection of audio descriptors and metadata[3] for "a million contemporary popular music tracks"[4] (see also Appendix A for further details). This large-scale collection was made available by Columbia University's LabROSA[5] and the company The Echo Nest[6]. As a whole, it comprises music from 44,745 unique artists and it includes a variety of music genres such as rock, pop, hip-hop, electronic, jazz, or folk. From the million tracks, 515,576 have information on the release year according to MusicBrainz[7], an open music encyclopedia that collects and makes music metadata available. Since there are some duplicate tracks in the original dataset[8] and others that do not have the full audio descriptions, the size of the dataset used here was

[3] Notice that, due to copyright laws, the database contains audio descriptors and metadata but not the actual audio files.

[4] http://labrosa.ee.columbia.edu/millionsong

[5] http://labrosa.ee.columbia.edu

[6] http://the.echonest.com

[7] http://musicbrainz.org

[8] http://labrosa.ee.columbia.edu/millionsong/blog/11-3-15-921810-song-dataset-duplicates

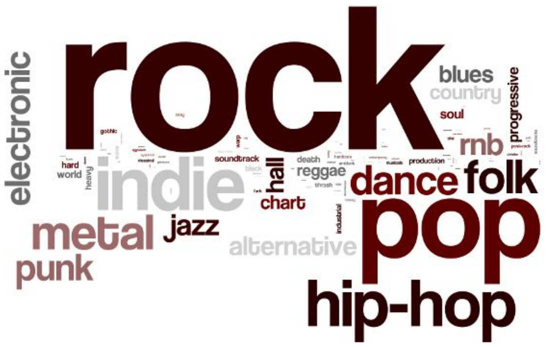

Figure 6.1: Tag cloud of the genres included in the analyzed subset of the million song dataset (using the default MusicBrainz genre information provided in the dataset). The font size represents the logarithm of number of tracks associated with a given annotation or genre tag.

reduced to 465,259 items. Further discarding the years before 1955 due to lack of representativeness, we obtain a working collection of 464,411 distinct music recordings (from 1955 to 2010), which roughly corresponds to more than 1,200 days of continuous listening. A diversity of popular music genres is included in the final subset (Fig. 6.1).

6.2.2 Selected audio descriptors

As previously mentioned, the million song dataset provides state-of-the-art audio descriptors for each beat of a given track (Jehan, 2005, 2010). Therefore, for each track, a beat-based sequence of multi-dimensional values is provided. The most relevant descriptors are related to pitch, timbre, and loudness. These descriptors are psychoacoustically-motivated, and its computation includes several steps to mimic the response of the human ear such as the grouping of energies into perceptually-motivated frequency bands, the consideration of spectro-temporal dynamics, or the application of an outer and middle ear filter (Jehan, 2005).

In this chapter we analyze the statistical properties of three encoded MSD's descriptors namely: MSD-Pitch, MSD-Timbre and MSD-Loudness. As described in Sec. 2.2 MSD-Pitch is a 12-dimensional descriptor that roughly corresponds to the harmonic content of the piece, including its chords, melody, and tonal arrangements. In particular, each descriptor's dimension consists in a real value between 0 and 1 indicating the degree of absence or presence of each of the 12 pitch classes of the chromatic scale (C, C#, D, D#, etc.). MSD-Timbre accounts for the sound color, texture, or tone quality, and can be essentially associated with instrument types, recording techniques, and some expressive performance resources. In particular, each audio segment is decomposed into 12-bivariate spectro-temporal basis that correspond to high level abstractions of the spectral shape. Since the fist dimension of this basis decomposition represents the energy of the signal, we take dimensions 2 to 12 (11 values) as the MSD-Timbral descriptor. Finally, MSD-Loudness basically correlates with our perception of sound amplitude or volume[9] and corresponds with the first dimension of the previously described spectro-temporal basis decomposition (see Sec. 2.2 for further information regarding these descriptors).

6.2.3 Encoding

To identify structural patterns of musical discourse we take advantage of our musical code-words and use them as main analysis units. As described in previous chapters we use a simple encoding process which maps each descriptor's beat-segment to a dictionary of predefined code-words. Fig. 6.2 shows a schematic summary of the encoding process applied, in this case, to MSD-Pitch descriptions. Table 6.1 shows the most important aspects of the followed encoding process.

To facilitate the interpretation of MSD-pitch code-words, we opt for a binary discretization of each MSD-Pitch dimension, therefore only accounting for presence or absence of a given pitch class. This way, this 12-dimensional descriptor can be encoded using $2^{12} = 4,096$ code-words. In particular, we use a single threshold set to 0.5 and map the original pitch vector values to 0 or 1, depending on whether they are below or above the threshold, respectively. The value of 0.5 is near the mean value of the considered vector components and other arbitrary numbers close to it provided no apparent change in the results of our analysis. Before discretization, MSD-Pitch descriptions of each track are automatically transposed to an equivalent main tonality, such that all pitch code-words are considered within the same tonal context or key. For this process we employ a circular shift strategy (Serrà et al., 2008), correlating (shifted) per-track averages to cognitively-inspired

[9] In this case we refer to the intrinsic loudness of a recording, not the loudness a listener could manipulate by changing the volume control of her audio player.

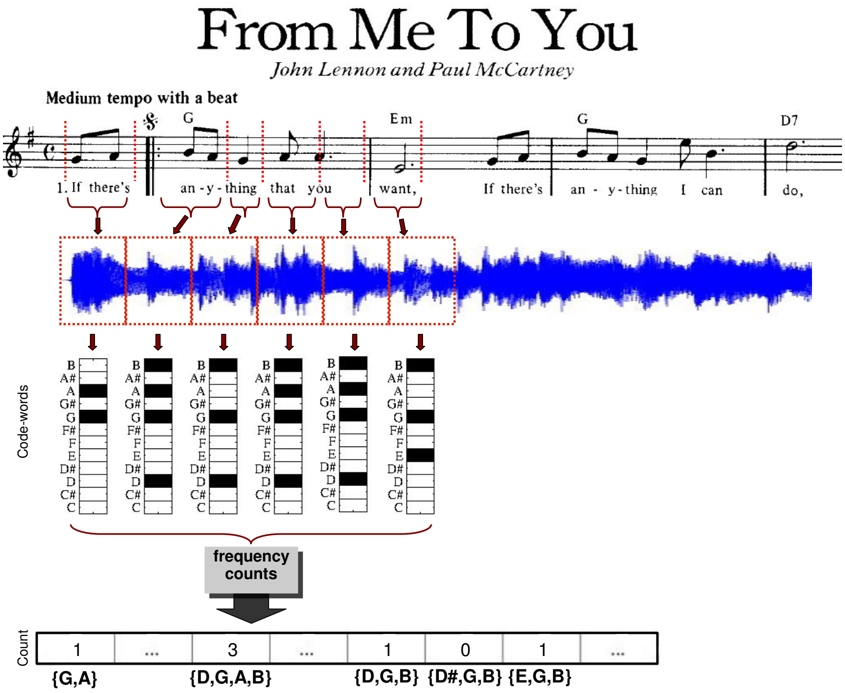
From Me To You
John Lennon and Paul McCartney
Medium tempo with a beat
G
Em
G
D7
1. If there's an - y - thing that you want, If there's an - y - thing I can do,
Code-words
B A# A G# G F# F E D# D C# C
frequency counts
Count
1 ... 3 ... 1 0 1 ...
{G,A}
{D,G,A,B}
{D,G,B}
{D#,G,B}
{E,G,B}

Figure 6.2: Method schematic summary for MSD-Pitch data. The dataset contains the beat-based music descriptors of the audio rendition of a musical piece. For pitch, these descriptions reflect the harmonic content of the piece (Jehan, 2005), and encapsulate all sounding notes of a given time interval into a compact representation (Casey et al., 2008; Müller et al., 2011), independently of their articulation (they consist of the 12 pitch class relative energies). All descriptions are encoded into music code-words, using a binary discretization in the case of pitch. Code-words are then used to perform frequency counts.

tonal profiles (Krumhansl, 1990). This strategy is commonly applied to pitch class descriptions in many music processing contexts (Casey et al., 2008; Müller et al., 2011), specially in the retrieval of versions of the same musical composition (Serrà et al., 2010) and in automatic chord/key estimation (Gómez, 2006).

Compared to pitch, timbre is believed to have a much higher dimensionality, at least perceptually (Bregman, 1990). To account for this, and also in order to better match the underlying distribution of the timbre descriptions provided in the million song dataset, we make use of a ternary, equal-frequency

Musical facet	Pre-processing	Dimensionality	Discretization	Threshold Value(s)
MSD-Pitch	Transposition to the same tonal context.	12 real values (between 0 and 1).	Binary	0.5 (same value for each dimension).
MSD-Timbre	Remove the loudness component and get a sample of beat-based timbre descriptions (see text).	11 real values.	Ternary	33 and 66% quantiles of the extracted sample (different values for each dimension).
MSD-Loudness	Take the loudness component from timbre descriptions and get a sample of beat-based loudness descriptions (see text).	1 real value.	300 steps	Equal-sized steps in the range of the extracted sample.

Table 6.1: Summary of the encoding process for deriving music code-words from the beat-based descriptions provided in the million song dataset. In total we have 4,096 possible MSD-Pitch code-words, 177,147 possible MSD-Timbre code-words, and 300 possible MSD-Loudness code-words.

encoding (Cios et al., 2007), providing a total of $3^{11} = 177,147$ possible MSD-Timbre code-words. Thresholds are set to the 33 and 66% quantiles of a representative sample of beat-based timbre description values[10]. To construct such sample we randomly chose one million MSD-Timbre vectors from the dataset such that a maximum of 8,000 vectors corresponded to the same year. In this way we controlled that no bias towards a certain year was introduced into the sample.

MSD-Loudness values are originally provided in decibels (dB), and limited within a range from 0 to 60 (Jehan, 2005, 2010). To study their distribution we treat these loudness values directly as a random variable (see below). Nonetheless, in order to conform to the standard signal processing criterion (Oppenheim et al., 1999), we subtract the loudness reference of 60 dB used in the million song dataset from them. This yields values $x \in [-60, 0]dB_{FS}$, where dB_{FS} means full-scale decibels. Since this descriptor has only one dimension, in order to encode it we use an unsupervised equal-width discretization (Cios et al., 2007) into 300 equal steps. In preliminary analysis we experimented with other discretizations (e.g. 200 steps, 300 quantiles), obtaining very similar results.

[10] This sample should not be confused with the final sample used for analysis. It is just an initial sample for obtaining the 33 and 66% quantiles that will allow to threshold the music descriptions.

6.2.4 Fitting procedure

As in previous chapters we analyze the rank-frequency distribution of code-words (i.e. the number of times each code-word type appears in a sample). In this case we study the long-term temporal variation (from 1955 to 2010) of the distribution parameters for the three selected audio descriptors (i.e. MSD-pitch, MSD-timbre and MSD-Luodness). For the case of MSD-Pitch and MSD-Timbre we evaluate, as in previous chapters, if a discrete power-law

$$P(z) \propto (z)^{-\beta}, \tag{6.1}$$

or a discrete shifted power-law

$$P(z) \propto (z+c)^{-\beta}, \tag{6.2}$$

fits our data.

In the case of MSD-Loudness we evaluate our data against a truncated reversed log-normal distribution

$$P(z) = \sqrt{\frac{2}{\pi\sigma^2}} \left[\operatorname{erf}\left(\frac{\ln z_{\mathrm{max}} - \mu}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{\ln z_{\mathrm{min}} - \mu}{\sqrt{2}\sigma}\right) \right]^{-1} \ldots \\ \frac{1}{z} \exp\left(-\frac{(\ln z - \mu)^2}{2\sigma^2}\right) \tag{6.3}$$

with $0 \leq z_{\mathrm{min}} \leq z \leq z_{\mathrm{max}}$ and where

$$\operatorname{erf}(y) = 2\pi^{-1/2} \int_0^y e^{-u^2} du \tag{6.4}$$

is the error function (implemented as in Press et al. (1992)). The adjective 'reverse' refers to the fact that, considering x as the MSD-Loudness values, $P(x)$ is the mirror image of the true (truncated) log-normal distribution thus, the variable $z = -x$. See Appendix B for a detailed explanation about the three considered distributions.

The fitting procedure for power-laws and log-normals is based, as in previous chapters of this thesis, on the procedure described by Clauset et al. (2009). See Appendix C for further details.

6.2.5 Variation assessment

To quantify long-term variations of a vocabulary of code-words, we need to obtain samples of it at different periods of time. For that we perform a Monte Carlo sampling in a moving window fashion. In particular, for each year, we sample one million beat-consecutive code-words, considering entire tracks and using a window length of 5 years[11]. This procedure, which is repeated 10 times, guarantees a representative sample with a smooth evolution over the years.

To assess trends in fitting parameters over the years, we perform an ordinary least squares linear regression (Chatterjee and Hadi, 1986; Wasserman, 2003) and report the slope found. Statistical significance is evaluated under the null hypothesis that the slope is different from zero, using a two-tail t-test and $p < 0.01$ (if $p > 0.05$ we deem the slope as not statistically significant).

6.3 Results

6.3.1 MSD-Timbre

Fig. 6.3a shows examples of the obtained MSD-Timbre density distributions (i.e. the probability distribution of the code-wods frequencies). As in the case of Bark-band and MFCC code-words analyzed in Chapter 3, the distribution of MSD-Timbre code-words across years is also well fitted by a power-law distribution. Once more the robustness of power-law timbral distributions is highlighted since in this experiments we are using a different timbral descriptor which has only 11 dimensions, as opposed to 22, and we are using a ternary, as opposed to binary, quantization. In particular for MSD-Timbre code-words, we find that the distribution of code-word frequencies for a given year nicely fits to a discrete shifted power-law: $P(z) \propto (c + z)^{-\beta}$ for $z > z_{\text{min}}$, where we take z as the random variable (Adamic and Huberman, 2002), $\beta = 1 + 1/\alpha$ as the exponent[12], and c as a constant. As described in previous chapters, a power-law indicates that a few code-words are very frequent while the majority are highly

[11] The window is centered at the corresponding year such that, for instance, for 1994 we sample one million consecutive beats by choosing full tracks whose year annotation is between 1992 and 1996, both included

[12] Where α is the rank-frequency power-law exponent.

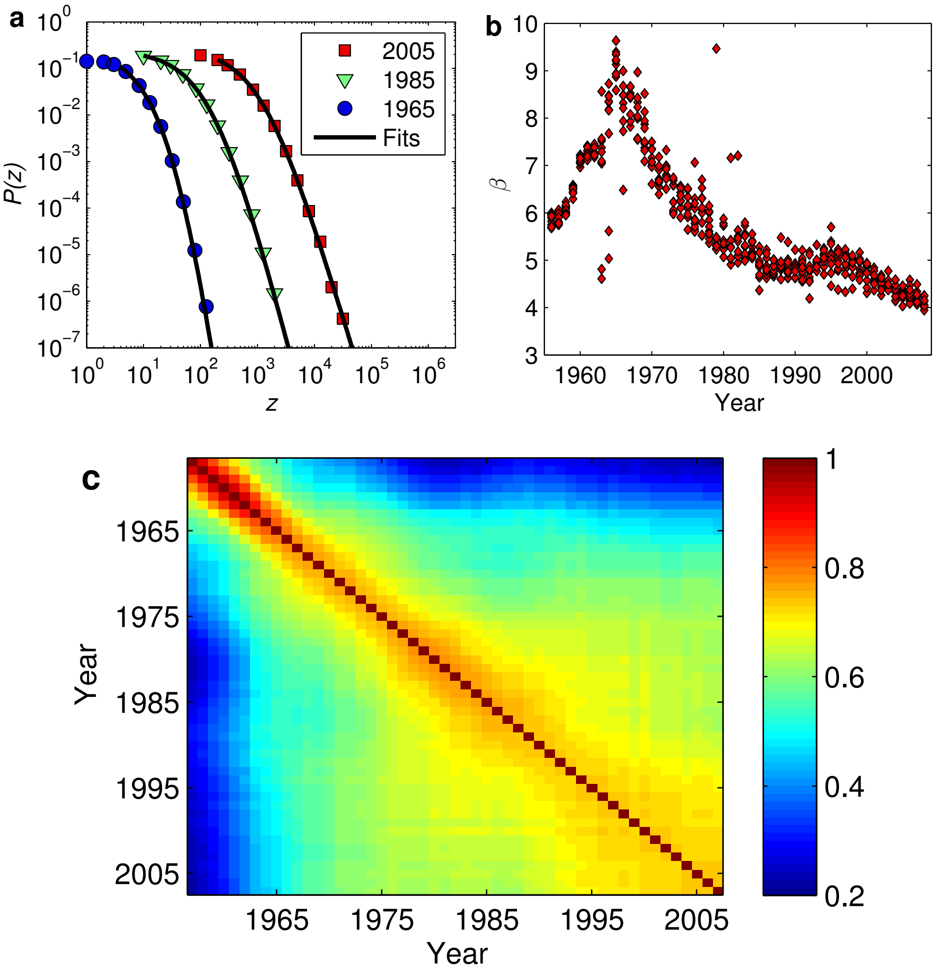

Figure 6.3: MSD-Timbre distributions. (a) Examples of the density values and their fits, taking code-words' frequencies z as the random variable. Curves are chronologically shifted by a factor of 10 in the horizontal axis. (b) Temporal evolution of the fitted exponents β. (c) Spearman's rank correlation coefficient for all pairwise distributions.

infrequent (intuitively, the latter provide the small musical nuances necessary to make a discourse attractive to listeners (Honing, 2011; Huron, 2006; Levitin et al., 2012)). Nonetheless, it also states that there is no characteristic frequency nor rank separating most used code-words from largely unused ones (except for the largest rank values due to the finiteness of the vocabulary). Remarkably, when analyzing the fitted β exponent we can see that since 1965, β constantly decreases to values approaching 4 (Fig. 6.3b). Although such large values of β would imply that other fits could also be acceptable, the power-law provides a simple parameterization to compare the changes over the years (and is not rejected in a likelihood ratio test in front of other alternatives). Smaller values of β indicate less timbral variety: frequent code-words become more frequent, and infrequent ones become even less frequent. This evidences a growing homogenization of the global timbral palette. It also points towards a progressive tendency to follow more fashionable, mainstream sonorities. In order to evaluate if the ranks in the rank-frequency counts were changing across years or not (e.g. a certain code-word was used frequently in 1963 but became mostly unused by 2005) we compute the Spearman's rank correlation coefficients (Hollander and Wolfe, 1999) for all possible year pairs. Interestingly, MSD-Timbre's rank correlation coefficients are generally below 0.7, with an average of 0.57 ± 0.15 (Fig. 6.3c). These rather low rank correlations would act as an attenuator of the sensation that contemporary popular music is becoming more homogeneous, timbrically speaking. The fact that frequent timbres of a certain time period become infrequent after some years could mask global homogeneity trends to listeners. Global timbre properties, like the aforementioned power-law and rankings, are clearly important for music categorization tasks (Ball, 2010; Casey et al., 2008) (one example is genre classification (Scaringella et al., 2006)). Notice however that the evolving characteristics of musical discourse have important implications for artificial or human systems dealing with such tasks. For instance, the homogenization of the timbral palette clearly challenge tasks exploiting this facet.

6.3.2 MSD-Pitch

Fig. 6.4a shows examples of the obtained rank-frequency distributions with illustrations of the most frequent and infrequent code-words. We observe that most used MSD-pitch code-words generally correspond to well-known harmonic items (De Clercq and Temperley, 2011), while unused code-words correspond to rare and dissonant pitch combinations. Noticeably, as in the case of the timbral code-words, the distributions of MSD-Pitch code-words are also well fitted by power-laws. Specifically, as for MSD-Timbre, we find that the distribution of code-word frequencies for a given year fits to a discrete shifted power-law of the form $P(z) \propto (c + z)^{-\beta}$ (Fig. 6.4b). This indicates that a few tonal combinations are very frequent while the big

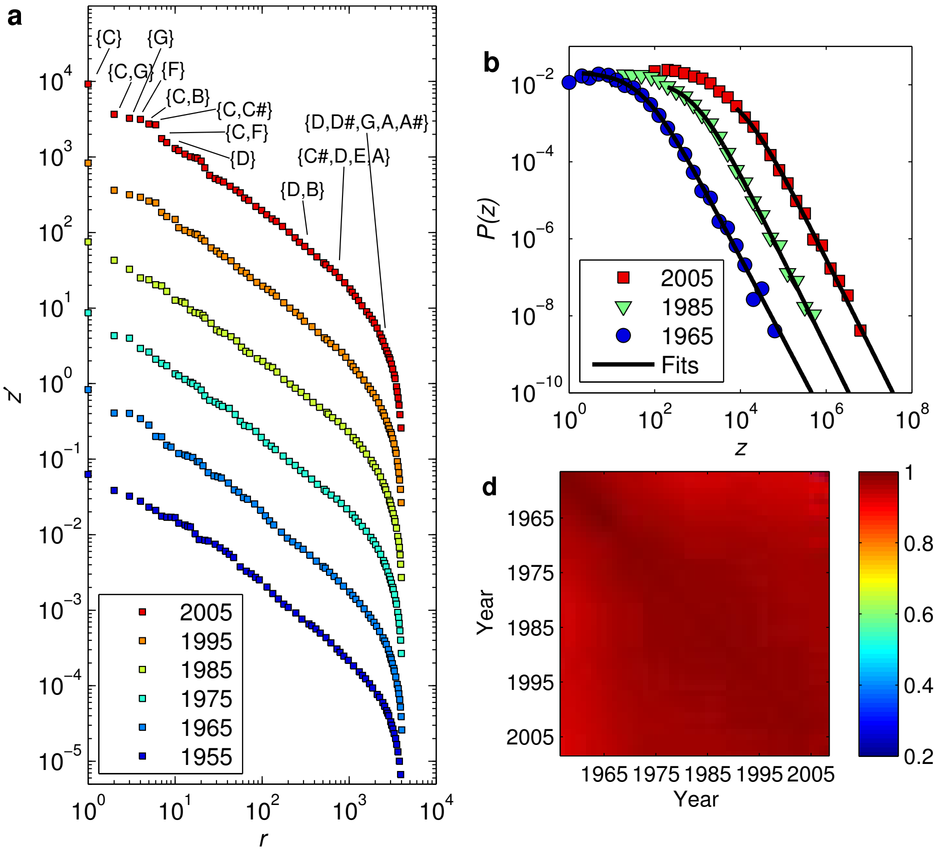

Figure 6.4: MSD-Pitch code-word distributions. (a) Examples of the rank-frequency distribution (relative frequencies z' such that $\sum_r z'_r = 1$). For ease of visualization, curves are chronologically shifted by a factor of 10 in the vertical axis. Some frequent and infrequent code-words are shown. (b) Examples of the density values and their fits, taking z as the random variable. Curves are chronologically shifted by a factor of 10 in the horizontal axis. (c) Spearman's rank correlation coefficient for all pairwise distributions. As mentioned in the text, correlations are all above 0.92 (we use the same color bar as in Fig. 6.3c for the sake of comparison).

majority are extremely rare. Again, we hypothesize that this power-law behavior provides an optimum balance between expected and unexpected note combinations that make a musical discourse attractive to listeners (Honing, 2011; Huron, 2006; Levitin et al., 2012). As previously mentioned, it also implies that there is no characteristic frequency nor rank separating most used code-words from largely unused ones. Another non-trivial consequence of power-law behavior is that when $\alpha \leq 2$, extreme events (i.e. very rare code-words) will certainly show up in a continuous discourse providing the listening time is sufficient and the pre-arranged dictionary of musical elements is big enough.

Importantly, we find this power-law behavior to be invariant across years, with practically the same fit parameters. In particular, the exponent β remains close to an average of 2.18 ± 0.06 (corresponding to α around 0.85), which is similar to Zipf's law in linguistic text corpora (Zipf, 1949) and contrasts with the exponents found in previous small-scale, symbolic-based music studies (Beltrán del Río et al., 2008; Zanette, 2006). The slope of the least squares linear regression of β as a function of the year is negligible within statistical significance ($p > 0.05$, t-test)[13]. This makes a high stability of the distribution of MSD-Pitch code-word frequencies across more than 50 years of music evident. However, it could well be that, even though the distribution is the same for all years, code-word rankings were changing across years. As in MSD-Timbre, we evaluate this possibility by computing the Spearman's rank correlation coefficients for all possible year pairs and find that they are all extremely high, with an average of 0.97 ± 0.02 and a minimum above 0.91. These high correlations indicate that code-word rankings practically do not vary with years.

6.3.3 MSD-Loudness

MSD-Loudness distributions are generally well-fitted by a reversed log-normal function (Fig. 6.5a). Plotting them provides a visual account of the so-called loudness race (or loudness war), a terminology that is used to describe the apparent competition to release recordings with increasing loudness (Deruty, 2011; Milner, 2009), perhaps with the aim of catching potential customers' attention in a music broadcast (from our point of view, loudness changes are not only the result of technological developments but, in part, also the result of conscious decisions made by musicians and producers in the musical creation process, cf. Milner (2009)). The empiric median of the MSD-Loudness values x grows from -22 dB$_{\mathrm{FS}}$ to -13 dB$_{\mathrm{FS}}$ (Fig. 6.5b), with a least squares linear regression yielding a slope of 0.13 dB/year ($p < 0.01$, t-test)[14]. In contrast, the absolute dif-

[13] The specific linear regression values for the β parameter are: Slope = 0.002, p-value = 0.097, t-statistic = 1.66, $R^2 = 0.005$.

[14] The specific linear regression values for median(x) are: Slope = 0.13, p-value = $2.4 \cdot 10^{-96}$, t-statistic = 25.84, $R^2 = 0.554$.

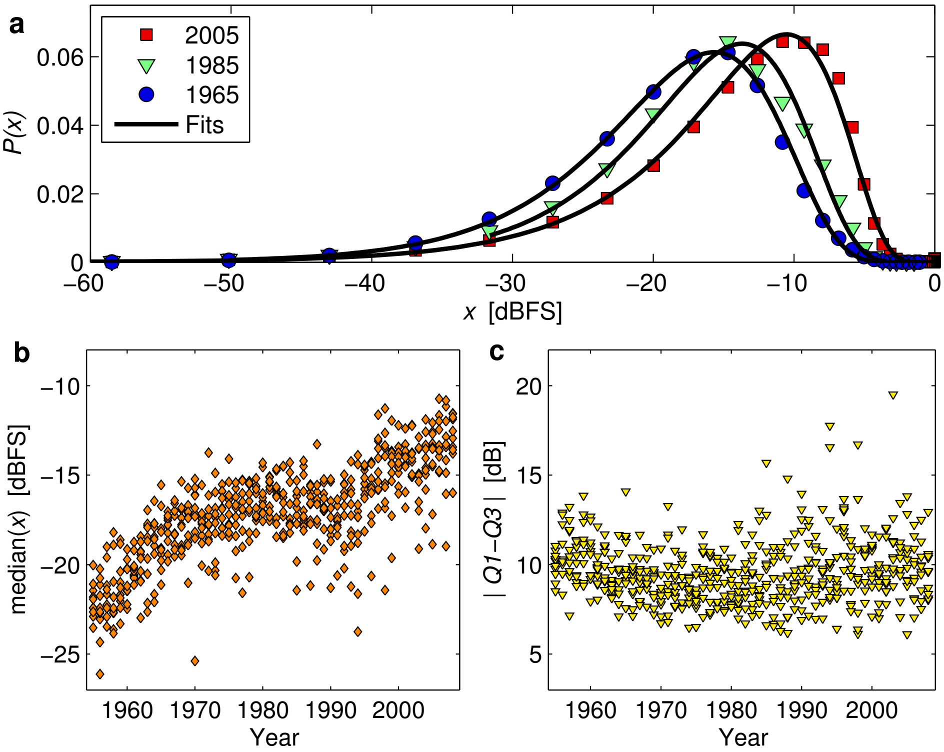

Figure 6.5: MSD-Loudness distributions. (a) Examples of the density values and fits of the loudness variable x. (b) Empiric distribution medians. (c) Dynamic variability, expressed as absolute loudness differences between the first and third quartiles of x, $|Q_1 - Q_3|$.

ference between the first and third quartiles of x remains constant around 9.5 dB (Fig. 6.5c), with a regression slope that is not statistically significant ($p > 0.05$, t-test)[15]. This shows that, although tracks become louder year after year, their absolute dynamic variability has been conserved, understanding dynamic variability as the range between higher and lower loudness passages of a recording (Deruty, 2011). However, and perhaps most importantly, one should notice that digital media cannot output signals over 0 $\mathrm{dB}_{\mathrm{FS}}$ (Oppenheim et al., 1999), which severely restricts the possibilities for maintaining the dynamic variability if the median continues to grow.

[15] $|Q_1(x) - Q_3(x)|$ linear regression values: Slope = 0.002, p-value = 0.321, t-statistic = 0.99, R^2 = 0.002.

6.4 Discussion and conclusion

Beyond the specific outcomes discussed above, we now focus on the evolution of musical discourse. Much of the gathered evidence points towards an important degree of conventionalism, in the sense of blockage or no-transformation, in the creation and production of contemporary Western popular music. Thus, from a global perspective, popular Western music would have no clear trends and show no considerable changes in more than fifty years. MSD-Pitch code-word frequencies are found to be always under the same underlying pattern: a power-law with the same exponent and fitting parameters. Moreover, frequency-based rankings of MSD-pitch code-words are practically identical. Frequency distributions for MSD-Timbre and MSD-Loudness also fall under a universal pattern: a power-law and a reversed log-normal distribution, respectively. However, these distributions' parameters do substantially change with years.

In Serrà et al. (2012b) we also studied the characteristics of the yearly-based evolution of the transition networks formed by code-word successions, where each node represents a code-word and each link represents a transition. Again, several of the computed network metrics for MSD-Pitch, MSD-Timbre, and MSD-Loudness remain immutable across years. Remarkably, the yearly-based evolution of the MSD-Pitch networks showed a reduction in the variety of pitch transitions from 1955 to 2010.

Thus, beyond the global perspective, we observe a number of trends in the evolution of contemporary popular music. These point towards a consistent homogenization of the timbral palette (although with timbral popularity varying across years), towards a standardized pitch usage and less varied pitch transitions, and towards louder and, in the end, potentially poorer volume dynamics.

Each of us has a perception of what is new and what is not in popular Western music. According to our findings, this perception should be largely rooted on well known pitch sequences, the usage of relatively novel timbral mixtures that are in agreement with the current tendencies, and the exploitation of modern recording techniques that allow for louder volumes. This brings us to conjecture that an old popular music piece would be perceived as novel by essentially following these guidelines. In fact, it is informally known that a "safe" way for contemporizing popular music tracks is to record a new version of an existing piece with current means, but without altering the main "semantics" of the discourse.

Some of the conclusions reported here have historically remained as conjectures, based on anecdotal evidence, or rather framed under subjective, qualitative, and non-systematic premises. Noticeably, by taking advantage of the previously proposed feature encodings and distribution analysis, we have explored a promising new way to acquire formal, quantitative, and systematic empirical evidence throughout the analysis of large-scale music collections. Thus, we encourage the development of further historical databases to be able to quantify the major transitions in the history of music, and to start looking at more subtle evolving characteristics of particular genres or artists, without forgetting the whole wealth of cultures and music styles present in the world.

Chapter 7

Music autotagging

7.1 Introduction

In Chapters 1 and 5 we have described the standard “Bag-of-frames” (BoF) approach used within the MIR community. This approach is often used to build automatic classification algorithms by means of aggregated descriptor sets (Casey et al., 2008; Klapuri and Davy, 2006). In Chapter 5 we compare the BoF algorithm against a “Selection-of-frames” strategy where only the most frequent frames within a song (i.e. those frames that belong to the most frequent code-words) where used to compute the aggregate feature vector. We obtained similar classification results, for genre and musical instrument classification tasks, for both approaches (i.e. using just 50 highly-frequent frames per audio file or using all frames in the song as in the BoF approach). This behavior can be explained by code-word frequencies being heavy-tailed distributed within songs. Thus, according to the findings reported in this thesis, encoded audio descriptors for timbre, chroma and loudness are heavy-tailed distributed whether they be at database-level or at song-level. This fact is akin to what happens when analyzing text documents and it has be thoroughly exploited by the IR community (Baeza-Yates, 1999).

Noticeably, recent works in video tag classification use a combination of code-word encoding and text-retrieval techniques to outperform state-of-the-art tag classification algorithms (Jiang et al., 2010). In these cases the code-words are obtained by the well known vector quantization (VQ) algorithm. In particular, one very recent paper applies this technique to multimodal video concept detection with promising results (Mühling et al., 2012). In this work "visual code-words" are complemented with "auditory code-words" which are computed from vector-quantized MFCC frames. Regarding audio classification Fu et al. (2011), also inspired by text-retrieval methods, proposed, as an alternative to the BoF approach, a bag-of-features approach. In this case, MFCC code-words are obtained via VQ. Then, a feature vector is computed by taking the frequency of use of each code-word within the song. Promising results are reported by using SVM as classification algorithm in two public datasets of about 1,000 songs. The evaluated classification tasks were genre classification and artist identification. Unfortunately, the paper does not report on rank-frequency distribution of VQ code-words. Finally, due to the intrinsic problems of VQ, where the performance of the algorithms are much influenced by the quality of the codebook (that depends on the initial sample, chosen distance measure, etc.), the authors propose a more complicated multi-codebook approach.

Motivated by the similarities between the distributions of our code-words and words in text documents (both heavy-tailed distributed), and the promising results in video tag classification when using text-retrieval techniques, in this chapter we explore the use of text-retrieval approaches to automatic audio classification. In particular, we explore the use of frequency weighted code-words for automatic tagging of songs. We call this strategy as "Bag-of-Code-Words" (BoC-W) and we compare classification results against the BoF approach and several other state-of-art automatic tagging systems.

We perform two autotagging experiments using two completely different public databases. In the first experiment we use the Million Song Dataset (MSD) to evaluate our algorithm against the BoF approach using the same audio descriptors and classification algorithm. Our train set consists on 259,552 full tracks, and our test set has 35,811 tracks, totaling 295,363 tracks. From the labels provided by last.fm[1] users we have selected 54 tags that appear in at least 10,000 songs of the dataset. In the second experiment we use the *Computer Audition Lab 500-Song* (CAL500) dataset (Turnbull et al., 2007) to compare classification results from a small and well-annotated dataset that has been used to evaluate and compare several autotagging algorithms in the literature (see also Appendix A).

[1] www.last.fm

7.2 Background on automatic tagging of music

Music automatic tagging (or autotagging) refers to the task of automatically attaching meaningful semantic labels (or tags) to a novel piece of music (Turnbull et al., 2008). These meaningful tags are related to very different music descriptions such as evoked emotion, presence of musical instruments, genre, music usage, etc. (Marques et al., 2011a). Usually, these tags are later on used as input for music recommendation systems.

Current commercial systems assign tags to music pieces by either expert-generated descriptions, which manually annotate songs with a rich vocabulary of musical terms, or by combinations of collaborative filtering and analysis of user-supplied tags for artists, albums and tracks (Levy and Sandler, 2009). Two paradigmatic examples are Pandora[2] for the first case, and Last.fm[3] for the second. However, these approaches have their own drawbacks, for instance, expert-based manual annotations are expensive with respect to both money and time, and more importantly, they are not scalable (i.e. it would be impossible to manually-annotate the huge amount of new music that is generated every day). On the other hand, user-based systems that exploit the user-generated context of each track suffer from the so called "cold-start" problem (Celma, 2008). This problem appears when the system does not have enough information from users or tracks in order to assign relevant tags and recommend new music. Another drawback of these systems is that track information is highly unbalanced and often follows a power-law distribution (i.e. a small set of extremely popular songs gets the majority of user-generated labels whereas the majority of songs are mostly unknown and therefore gets very few labels). Thus, new or not-so-popular songs that lay in the "long tail" of the distribution are not recommended by the system (Celma, 2008).

To tackle these problems many researchers have proposed automatic tagging algorithms that, using audio descriptors and machine learning algorithms, can directly exploit the audio content of the analyzed tracks (Sordo, 2011). Moreover, it has been shown that combining context and content provides better results than using one approach alone (Knees et al., 2009; Turnbull et al., 2009).

Unfortunately, the proposed content-based systems have been tested mostly on relatively small or private corpora. Moreover, some of the methods

[2] www.pandora.com

[3] www.last.fm

do not scale to work with massive datasets (Sordo, 2011). Thus, there is a lack of research on content-based algorithms dealing with large public datasets while, at the same time, having those systems could provide a major aid for solving the aforementioned problems of music recommendation systems (Casey et al., 2008).

Content-based autotagging algorithms started as a logic expansion of previous works in genre, instrument, and artist classification (Marques et al., 2011a). However, this is a much more difficult problem that involves multi-class classification with a greater number of tags covering many musical facets. Moreover, tags are not always clearly defined and can have multiple meanings (polysemy). Furthermore, autotagging is also a multi-label problem, that is, multiple pertinent tags are associated to the same song (e.g. a song can have many tags such as: "happy","party", "dance", "piano", etc.).

Next, we describe state-of-the-art autotagging algorithms that are used within this chapter as basis for comparison. We also refer the interested reader to Sordo (2011) for an excellent review on the subject.

Turnbull et al. (2008) consider the autotagging task as one supervised multi-class, multi-label problem. The authors propose to model the joint probability of audio features and words. For that they train a Gaussian mixture model (GMM) over a timbral audio feature space. They estimate the parameters of the GMM using the weighted Mixture Hierarchies expectation maximization algorithm which is computationally less expensive than traditional parameter estimation techniques.

In Bertin-Mahieux et al. (2008) the authors propose a set of 360 classifiers trained using the on-line ensemble learning algorithm FilterBoost. They evaluate aggregated feature vectors of MFCCs, and multi-descriptor sets including autocorrelation coefficients of an onset trace, and spectrogram coefficients sampled by constant-Q (or log-scaled) frequency.

Hoffman et al. (2009) propose a Code Bernulli Average (CBA) probabilistic model that attempts to predict the probability that a tag applies to a song based on a vector-quantized (VQ) representation of MFCCs delta features. The CBA's model parameters are estimated with Maximum Likelihood estimation using the Expectation Maximization algorithm.

Finally, Sordo (2011) proposes several distance-based autotagging algorithms. In this case instead of learning from train set observations beforehand (as in the case of GMM, Boosting methods or SVMs) the proposed method propagates tags to unlabeled songs from closest tagged songs in some pre-defined acoustic space. A set of multi-faceted musical descriptors are used including low-level (i.e. signal-based), rhythm, tonal, and high-level (i.e. semantic tags inferred from SVM classifiers). Besides K-NN tag propagation with or without feature selection, Sordo (2011) also propose a Class-based Distance Classifier (CBDC) where instead of looking at the nearest songs, it focuses on the nearest tags. In particular, the CBDC algorithm computes, form a training set of tagged feature vectors, a cluster-based representation (i.e. a centroid) for every tag. Then, when a new song arrives, all tags whose centroids are closer than a pre-defined threshold are attached to the song. Noticeably, two of the proposed distance-based autotagging algorithms provided state-of-the-art results within the Music Information Retrieval Evaluation eXchange (MIREX[4]) 2011 Audio Tag Classification task. MIREX is an annual evaluation contest for Music Information Retrieval algorithms. Regarding tag classification the MIREX competition compares algorithm performances against two datasets. One collection has 1,400 different tracks with 45 diverse tags not related with mood (i.e. tags from genre, instrument, sound characteristics, etc.). The other dataset has 18 mood tag groups and 3,469 unique songs.

7.3 General method

In the following experiments we use the same general method to compute, on one side, the standard BoF approach and, on the other side, the proposed BoC-W algorithm. In both cases we start from short-time audio descriptors of timbre, chroma, and loudness and generate a feature vector that is used as input by an SVM classification algorithm.

In the case of the BoF approach, we decided that, in order to have a more challenging algorithm to compare with, we should compute not only the mean feature values but also its covariance, delta-mean and delta-covariance, where delta denotes the difference between descriptor values of two consecutive audio frames and covariance refers to the upper triangle of the covariance matrix computed as:

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \tag{7.1}$$

where $\mu_i = \mathrm{E}(X_i)$ is the expected value of ith entry in the random vector X (i.e. the descriptor's values). For instance, for a frame-level descriptor of

[4] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Musical Facet	Descriptor	Quantization	Codebook size
Autotag experiment I: Autotagging the MSD			
Timbre	MSD-Timbre (11 dim)	Ternary	177,147 (3^{11})
Harmonic Cont.	MSD-Pitch (12 dim)	Binary	4,096 (2^{12})
Loundness	MSD-Loudness (1 dim)	300 steps	300
Autotag experiment II: Autotagging the CAL500 dataset			
Timbre	MFCC (11 dim)	Ternary	177,147 (3^{11})
Harmonic Cont.	HPCP (12 dim)	Binary	4,096 (2^{12})
Loundness	SE (1 dim)	300 steps	300

Table 7.1: Summary of encoding strategies for the BoC-W approach.

12 dimensions we obtain a feature vector with 180 values: 12 mean values, 78 covariance values, 12 delta mean values, and 78 delta-covariance values.

For the BoC-W algorithm we first encode each frame-level descriptor into its corresponding code-word (see Sec. 3.2.3) then we count the frequency of use of each code-word within the audio file. Following text-retrieval techniques we apply a weighting strategy to each frequency value and then use these weighted values as feature vector (see Sec. 7.3.1).

It is important to notice that in the case of the BoF approach, regardless of the song content, we always obtain a fixed-sized feature vector as summarized representation of the song (e.g. the 180 values of the above example for a descriptor of 12 dimensions). Nevertheless, in the case of the BoC-W approach we obtain a very sparse feature vector that will vary according to the content of the song. For instance, taking a binary-quantized 12-dimensional descriptor, it could be the case that an extremely repetitive song whose frames are encoded into only 10 different code-words the resulting feature vector will have only 10 non-zero values (out of a codebook of 2^{12} possible values). These sparse representations are also obtained when classifying text documents.

As mentioned in Sec. 7.1 we perform two autotagging experiments, one for the MSD, and one for a small and well-annotated dataset called CAL500. Table 7.1 shows a summary of the encoding strategies for each experiment.

7.3.1 Weighting strategies

In the proposed experiments we use common weighting strategies from text information retrieval namely: Binary (BIN), Term Frequency (TF), and Term Frequency - Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988).

The BIN weighting strategy just looks for the presence or absence of a code-word within a document (or song in our case). Thus, the BIN weight of a code-word will be "1" if the song contains such a code-word and "0" otherwise.

In the case of TF we count the number of times a code-word is used within a song and divided it by the total number of code-words present in the song:

$$tf_i = \frac{c_{i,j}}{\sum c_j}, \tag{7.2}$$

where tf_i is the TF of code-word i, $c_{i,j}$ is the number of times the code-word i appears in song j, and $\sum c_j$ corresponds to the total amount of code-words within song j. Thus, the weight of each code-word can be seen as the probability of occurrence of such code-word within the song. The main advantage of normalizing by the song's total number of code-words is that the length of the song does not interfere with the resulting weights of its code-words (which will certainly do if we just count code-words' frequency of use).

If we consider that highly frequent code-words as found in the "universe" of songs (i.e. code-words with low ranks in the global distribution) are not very informative (like stop-words in document retrieval) we could try to "reduce" their weights within the song's feature vector. Following the same line of thought, we could try to "amplify" the original weight of rare code-words as found in the global distribution. Thus, we multiply the TF weights by an inverse-document-frequency (IDF) factor that tries to compensate for global frequency distributions. Variations of this so called TF-IDF algorithm have proven to be very effective within the IR filed (Manning et al., 2008).

In our case the TF-IDF ($tfidf$) weight for a code-word i is computed as follows:

$$tfidf_i = tf_i * idf_i, \tag{7.3}$$

where tf_i corresponds with the previously defined TF weight of the code-word i and idf_i is the inverse document frequency for code-word i computed as:

$$idf_i = \log(\frac{N}{dc_i}), \tag{7.4}$$

where N is the total number of documents (i.e. songs) in the train set and dc_i corresponds to the number of documents of the train set where the code-word i was used.

7.3.2 Classification

Regarding classification algorithms we opt for the well known support vector machines (SVM) classifier (see Sec. 2.4 for further information about SVMs). This decision is motivated by the fact that SVMs are widely used in MIR in general and in autotagging algorithms in particular (Sordo, 2011). Furthermore, SVMs are also of choice to avoid overfitting when having sparse feature vectors as in our BoC-W model (Joachims, 1998).

In all cases we train binary classifiers using class-weights to compensate for imbalance data (He and Garcia, 2009; Tang et al., 2009). Given the big amount of data we are working with we opt for the large scale implementation of linear SVMs called LIBLINEAR (Fan et al., 2008). We also perform a grid search for several complexity parameters (C) to find the best classification results in every autotag experiment.

7.4 Autotag experiment I: autotagging the MSD

7.4.1 Database

For this experiment we use the same large-scale public database as in Chapter 6 namely: the Million Song Database (MSD) (Bertin-Mahieux et al., 2011). Fortunately, the authors of the dataset also provide song-level tags for about 500,000 songs within the MSD. These tags correspond with user-provided labels from the on-line music service last.fm[5] that were matched with songs in the MSD. This MSD subset is called the Last.fm dataset[6] (LFD). In this experiment we use the provided train / test splitting for the LFD. This splitting was made in such way that artists belong into one set only (i.e. if an artist is present in the train set is not present in the test set and *vice versa*).

Firstly, we pre-process the provided tags by manually stemming similar tags. For instance, *Progressive rock*, and *prog rock* tags are merged into the tag *Progressive rock*. Next, from the 522,366 unique tags, we select those used in at least 10,000 songs (after the stemming process) obtaining a final list of 54 tags to work with. The majority of the selected tags are related with musical genre and mood (see Appendix E for a complete list of stemming

[5] www.last.fm

[6] http://labrosa.ee.columbia.edu/millionsong/lastfm

words and selected tags). Finally, following LFD authors' recommendations we deleted a set of duplicate tracks[7], and a set of matching errors between song names and last.fm tags[8].

After this selection and correction process the final dataset that we call *MSD-Tag*, contains 259,552 tracks in the train set and 35,811 tracks in the test set totaling 295,363 songs with at least 1 of the 54 selected tags.

7.4.2 Method

We use the general method described in Sec. 7.3 to compare the proposed BoC-W algorithm against the standard BoF approach. We use the same MSD descriptors as in Chapter 6 namely: MSD-Timbre, MSD-Pitch, and MSD-Loudness (see also Sec. 2.2). However, in this case instead of using a beat-based temporal segmentation, we use the more detailed temporal resolution as provided by the MSD's *segments*[9].

In the case of BoF, for each song we compute an aggregate feature vector of mean, covariance, delta mean and delta covariance values. Whereas, for the BoC-W approach we use the same encoding strategy as in Chapter 6 (see also Table 7.1). Thus, each of the 11 dimensions of the MSD-Timbre descriptor is quantized into one out of three possible values (ternary quantization) providing a total of $3^{11} = 177,147$ possible MSD-Timbre code-words.

In order to estimate the quantization thresholds for MSD-Timbre we randomly select 167,754 songs from MSD-Tag. This corresponds to 155,938,102 frame-size segments. Then, we use as future quantization thresholds the values that correspond to the 33 and 66% quantiles of the distribution of each dimension's values.

In the case of MSD-Pitch we binary-quantize each of the 12 descriptor's dimension using a threshold of 0.5. Thus, we obtain a total of $2^{12} = 4,096$ possible MSD-Pitch code-words.

[7] http://labrosa.ee.columbia.edu/millionsong/blog/11-3-15-921810-song-dataset-duplicates

[8] http://labrosa.ee.columbia.edu/millionsong/blog/12-2-12-fixing-matching-errors

[9] According to MSD's on-line documentation *segments* are the smallest musically relevant elements, and are defined as: "a set of sound entities (typically under a second) each relatively uniform in timbre and harmony. Segments are characterized by their perceptual onsets and duration in seconds, loudness (dB), pitch and timbral content (from http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf)

The MSD-Loundess descriptor is quantized by using 300 equal-sized steps extracted from the same database sample used to estimate the MSD-Timbre's quantization thresholds.

When computing the sparse feature vector of BoC-W we evaluate the three weighting strategies described above namely: BIN, TF and TF-IDF (see Sec. 7.3.1).

Regarding classification, as described in Sec. 7.3, we use the same SVM classification algorithm to automatically label tracks from the MSD-Tag test-set for both BoF and BoC-W feature vectors. In this case we perform a grid search of the C parameter within the linear kernel[10] using the following values: C=1, C=10, and C=100.

7.4.3 Evaluation

Due to the imbalance nature of this classification task (i.e. some tags are much more frequent than others), instead of reporting per-song results, we compute per-tag F-measure results. Thus, in order to obtain global classification results, following Marques et al. (2011a) recommendations, we compute the mean of all tags' F-measures (meanF), and we also compute a global F-measure (globalF). In this case, instead of taking the mean of all F-measure values, we take the mean values for both Precision and Recall and then use these mean values to compute the global F-measure. Therefore, our evaluation measures are:

$$meanF = \frac{\sum(Fmeasure_{tag})}{numTags}, \tag{7.5}$$

where $Fmeasure_{tag}$ is the F-measure value for a particular tag computed as $F = 2*Precision*Recall/(Precision+Recall)$ and $numTags$ corresponds with the total number of tags (i.e. 54 in this experiment), and

$$globalF = \frac{2*meanPrecision*meanRecall}{meanPrecision+meanRecall}, \tag{7.6}$$

where meanPrecision and meanRecall are the mean values for Precision and Recall from all Tags (see also Sec. 2.4 for further information).

We use the same evaluation measures for both BoF and BoC-W algorithms.

[10] Due to the size of the database it was not possible to evaluate other SMV kernels. For instance, using the LiBSVM (Chang and Lin, 2011) implementation with rbf kernel, it took about ten hours to train and test only one label. Moreover, the results obtained for the few calculated labels were worst than those obtained using a linear kernel.

			BoC-W		BoF
Descriptor	**Dataset**	**Codebook**	**Mean**	**Std. Dev.**	**Length**
MSD-Timbre	train	177,147	776,38	386,81	154
MSD-Timbre	test	177,147	802,17	418,10	154
MSD-Pitch	train	4,096	231,34	103,62	180
MSD-Pitch	test	4,096	235,47	107,29	180
MSD-Loudness	train	300	104,68	32,16	4
MSD-Loudness	test	300	105,36	32,68	4

Table 7.2: BoC-W and BoF feature vector lengths for train and test datasets. In the case of BoC-W only non-zero values are counted. Codebook's size, mean, and standard deviation values are shown for BoC-W. In the case of BoF all feature vectors have the same length.

7.4.4 Results

Table 7.2 shows the mean and standard deviation values of the number of different code-words in BoC-W feature vectors (i.e. the number of different code-words per song). For illustration purpose we also show the size of the encoding codebook and the dimensionality of each fixed-length BoF feature vector. As can be seen from the table, the BoC-W's feature vectors are sparse and their lengths (i.e. non-zero values) are, on average, larger than the BoF feature vectors.

Tag classification

Fig. 7.1 shows meanF classification results for the BoF approach and the different weighting strategies of the BoC-W approach. All results correspond to the best classification values from grid search of the C parameter. Detailed results can be seen in Table 7.3, including meanF, globalF and best C parameters.

The classification results show that for both meanF and globalF the BoC-W approach outperforms the standard BoF approach. In particular, comparing BoF and best BoC-W results per descriptor type we always obtain classification results that at least double the ones obtained with the BoF approach. Furthermore, the MSD-Timbre descriptor using the BoC-W strategy with TF-IDF weighting produces the best overall classification results (27,91% and 31.13% for meanF and globalF respectively). These results are more than 2.4 times the ones obtained with the best BoF approach.

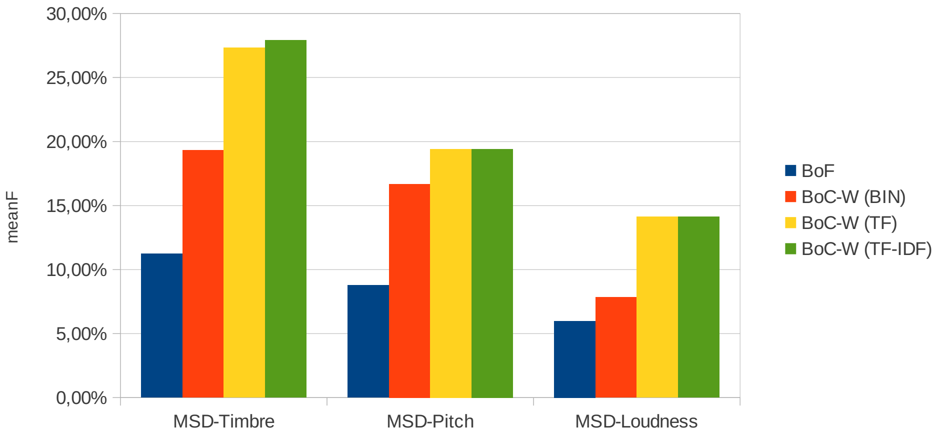

Figure 7.1: MSD-Tags' meanF classification results (y-axis) per descriptor type (x-axis). BoF corresponds with the Bag-of-Frames approach, BoC-W means Bag-of-Code-Words with either BIN, TF, or TF-IDF weighting strategy (in parenthesis). Classification results are shown in percentage.

Approach	**MSD-Timbre**	**MSD-Pitch**	**MSD-Loudness**
	MeanF values		
BoF	11,23 (C=1)	8,76 (C=10)	5,98 (C=10)
BoC-W (BIN)	19,32 (C=1)	16,69 (C=1)	7,83 (C=1)
BoC-W (TF)	27,34 (C=10)	19,42 (C=10)	14,13 (C=10)
BoC-W (TF-IDF)	27,91 (C=10)	19,42 (C=10)	14,13 (C=100)
	GlobalF values		
BoF	12,70 (C=1)	10,00 (C=10)	7,42 (C=10)
BoC-W (BIN)	22,83 (C=1)	21,05 (C=1)	15,09 (C=1)
BoC-W (TF)	29,52 (C=10)	19,97 (C=10)	14,33 (C=10)
BoC-W (TF-IDF)	31,13 (C=10)	20,03 (C=10)	14,28 (C=100)

Table 7.3: F-measure results for MSD-Tag dataset. F values are expressed as percentage, and best C values from grid search are shown in parenthesis.

Regarding BoC-W weighting strategies we observe that both TF and TF-IDF produce better classification results than BIN (except for globalF MSD-Pitch and globalF MSD-Loudness). Moreover, we observe that TF and

TF-IDF strategies mostly produce comparable results for both meanF and globalF.

Following Marques et al. (2011a) recommendations, we also analyze how individual tag performances are related to the a-priori tag frequencies. As can be seen in Table 7.4 classification results are not strongly correlated with tag frequencies. When computing the correlation coefficient for all tags and their corresponding total frequencies we also obtain a moderate correlation value of 0.4683 (see Appendix E for the complete list of per tag results, and a scatter plot of classification results vs. tag-frequencies). This seems to indicate that there are sufficient examples in the database for the machine learning algorithm to learn from. Therefore, since classification differences can not be explained by tag frequencies, it seems that those differences are related with inherent tag properties which facilitate or not the learning process. Moreover, nine of the ten best classified tags correspond with musical genres being "female vocalists" the only non-genre tag within this group. In the case of the 10 worst classified tags we found genre tags (indie pop, progressive rock, new wave, pop rock, folk), mood tags (sad, fun, happy), and the tags "party" and "soundtrack". Not surprisingly, most of these tags are ill-defined or too complex to be captured by nowadays audio features.

In the next sections we further evaluate the best classification strategy namely: BoC-W for MSD-Timbre with TF-IDF weighting.

Code-word selection

In this experiment we analyze the impact of removing (or keeping) the most frequent code-words as found in the MSD-Tag train set. This procedure is akin to "stop word removal" in IR where the most frequent terms found in texts (e.g. words like *a, the, and, to..*) can be removed to increase search performance. The idea behind this experiment is to assess the importance of these extremely frequent code-words within the tag classification task[11]. For that we apply the same method used to classify tags for the MSD-Tag dataset. In particular, we use the best approach found in the previous experiment namely: BoC-W with MSD-Timbre and TF-IDF weighting, and classified with a linear SVM with $C = 10$. The only difference is that we remove (or keep) from each song's feature vector those code-words that

[11] Note that these highly frequent code-words correspond with low-rank code-words within the power-law rank-frequency distribution described in previous chapters.

		Tag Frequency		
Tag	**F-measure**	**Train**	**Test**	**Total**
10 best classified tags				
hip-hop	58.70	14,694	2,282	16,976
metal	56.90	23,948	3,292	27,240
rap	50.49	7,918	1,220	9,138
electronic	49.71	31,266	4,734	36,000
jazz	49.51	21,264	3,585	24,849
rock	48.85	73,233	8,284	81,517
pop	41.31	47,031	5,790	52,821
female vocalists	40.42	33,388	4,344	37,732
trance	39.93	6,208	972	7,180
indie	37.70	36,105	4,588	40,693
10 worst classified tags				
indie pop	15.25	10,062	1,257	11,319
progressive rock	13.86	9,638	1,593	11,231
sad	13.50	9,062	1,084	10,146
party	13.22	8,745	1,166	9,911
new wave	12.32	6,994	721	7,715
pop rock	12.01	11,295	1,279	12,574
folk	11.71	17,402	2,306	19,708
fun	9.32	7,964	897	8,861
soundtrack	7.87	8,773	849	9,622
happy	7.75	8,797	948	9,745

Table 7.4: 10 best and 10 worst tag classification results for BoC-W (MSD-Timbre with TF-IDF weighting) and their corresponding tag frequencies. F-measure results in percentage.

belong to the N most frequent code-words in the train set. We analyze logarithmically spaced values of N from 1 to 100,000. Fig. 7.2 shows the meanF classification results obtained after removing (or keeping) the N most frequent code-words. Table 7.5 also shows meanF and globalF results.

Interestingly, if we compare the obtained results with the ones obtained when using all code-words (i.e. 27,91% and 31.13% for meanF and globalF respectively) we observe that we can safely remove up to 10,000 of the most frequent code-words without affecting the classification results. On the other hand, classification results obtained after using only those N most frequent code-words are comparable to the previous ones only when we keep the 100,000 most frequent code-words.

The previous experiments suggest that we can discard the 10,000 most frequent code-words and, at the same time, that the 100,000 most frequent

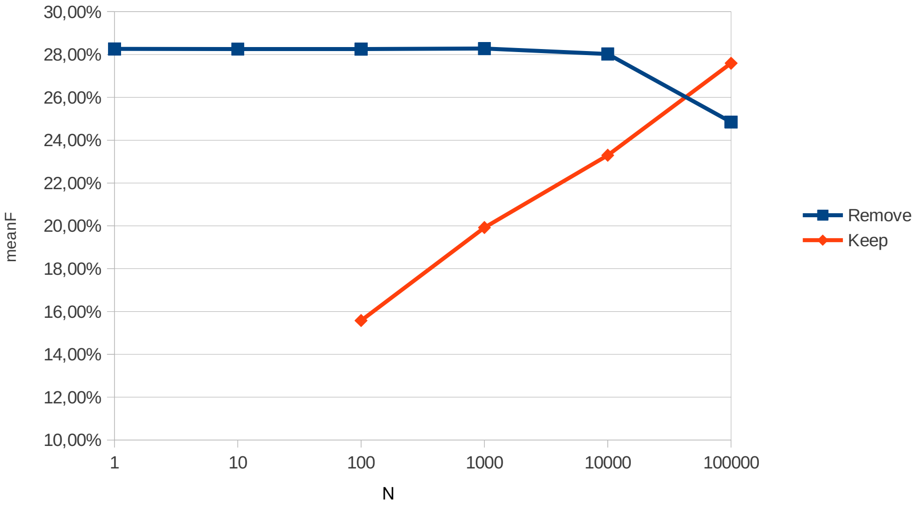

Figure 7.2: MeanF classification results after removing (blue) or keeping (red) the N most frequent code-words. F-measure results in percentage.

	MeanF		**GlobalF**	
N	**Remove**	**Keep**	**Remove**	**Keep**
1	28.26		31.77	
10	28.26		31.77	
100	28.25	15.58	31.75	15.86
1,000	28.28	19.92	31.76	20.48
10,000	28.03	23.30	31.46	24.35
100,000	24.85	27.59	27.40	30.48

Table 7.5: MeanF and globalF classification results for MSD-Tag dataset after removing or keeping the N most frequent code-words. F-measure results in percentage.

code-words contain enough information to classify the test set with the same F-measure values as when using all code-words. Thus, we also evaluate the classification results obtained after using only those code-words with $10,000 \leq N \leq 100,000$.

The meanF and globalF results for code-words with N between 10,000 and

Selection Strategy	**MeanF**	**GlobalF**
All	27,91	31.13
Remove 10k	28.03	31.46
Keep 100k	27.59	30.48
Keep btw 10k-100k	26,13	28,61

Table 7.6: MeanF and globalF classification results for MSD-Tag dataset per code-word selection strategy. "All" corresponds to the original results using all code-words, "Remove 10k" means removing the 10,000 most frequent code-words, "Keep 100K" means keeping the 100,000 most frequent code-words, and "Keep btw 10k-100k" corresponds to keeping code-words whose frequency lies between 10,000 and 100,000 (both included). F-measure results in percentage.

100,000 are 26,13% and 28,61% respectively (see Table 7.6). This result further suggests that like in the case of text classification, the most "informative" code-words are neither the most frequent nor the most rare ones (Ferrer i Cancho and Solé, 2003).

Regarding computation times, Table 7.7 depicts the required time to train and test all 54 labels for the MSD-Tag using MSD-Timbre descriptor. These results correspond with the output of the linux `time` command executed on an Intel®Core^{TM}2 Duo CPU E8200 @ 2.66GHz x2 with 6 GB of RAM running Ubuntu 12.04. As can be seen from the table, computation times are drastically reduced when discarding the most frequent code-words. In particular, the total CPU time (i.e. Usr+Sys times) is 62m06.021s for code-words between 10,000 and 100,000 whilst the total CPU time for BoC-W using all code-words is 127m01,913s, and for BoF is 301m52,888s. Notice that due to the reduction of train and test file sizes the BoC-W algorithm that uses code-words with frequencies between 10,000 and 100,000 also drastically reduces the total (Real) computation time from 670m9.053s (BoC-W with all code-words) to 93m38.779s (about 14% of the original time).

Train set reduction

Finally, we evaluate the classification results while reducing the size of the MSD-Tag training set for MSD-Timbre, MSD-Pitch and MSD-Loudness in the proposed BoC-W algorithm ($C = 10$ and TF-IDF weighting). We randomly select a percentage of the original train set and perform the classification as in the previous experiments. We execute this procedure 3 times (i.e. performing 3 random selections) and analyze the mean and standard

	Computation times		
Algorithm	**Real**	**Usr**	**Sys**
BoF	302m46.638s	299m53.201s	1m59.687s
BoC-W	670m9.053s	106m37.132s	20m24.781s
BoC-W Removed 10,000	180m3.199s	73m51.865s	9m28.208s
BoC-W Between 10,000-100,000	93m38.779s	56m34.624s	5m31.397s

Table 7.7: Computation time required to train and test with LibLinear all tags using the MSD-Timbre descriptor. All BoC-W algorithms use the TF-IDF weighting. **Real**: corresponds to the total amount of time between invocation and termination of the classification process. **Usr**: corresponds to user CPU time and **Sys**: corresponds to system's CPU time.

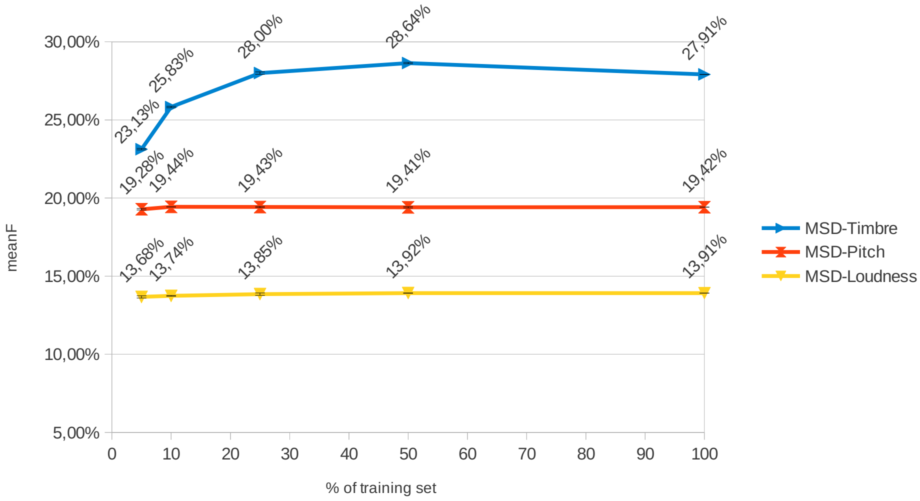

Figure 7.3: MeanF classification results after randomly selecting a percentage of the MSD-Tag train set. Selection percentages are: 100% (i.e. the original train set), 50%, 25%, 10%, and 5%. The random selection process is performed 3 times. Here, mean and standard deviation (error bars) of the selection rounds are depicted.

deviation of the meanF and globalF values. Fig. 7.3 shows the meanF classification results for train set reduction. The selection percentages are: 100% (i.e. the full train set), 50%, 25%, 10%, and 5%. These results show that the MSD-Tag train set can be safely reduced up to one quarter of its original size without affecting the classification results. The same behavior is shown when evaluating globalF results.

7.4.5 Conclusions for autotag experiment I

In this section we have evaluated the proposed BoC-W algorithm for automatic tag classification within a large-scale dataset. Comparing the BoC-W against the standard BoF approach it seems clear that the BoC-W provides much better results than the BoF approach in all evaluated descriptors. In particular, we have found that the MSD-Timbre descriptor with TF-IDF weighting provides the best classification results. These BoC-W results are more than 2.4 times the ones obtained with the best BoF approach.

Code-word selection experiments suggest that the most informative code-words are neither the most frequent nor the most rare ones of the global distribution. We hypothesize that those extremely frequent code-words have not enough discriminative power with respect to tags because they are present in almost all tracks regardless of their attached tags. On the other hand, rare code-words do not have enough generalization power to be considered by the machine learning algorithm as characteristic of a particular tag. Thus, according to our code-word selection experiments, those code-words whose frequency ranks lie between 10,000 and 100,000 offer a good trade-off between generalization and discrimination power. Furthermore, working with this reduced set of code-words not only produces comparative classification results as working with all code-words, but also provides significant reduction in computation time specially when working with large-scale datasets.

Finally, our experiments also reveal that the MSD-Tag train set can be reduced down to 25% of its original size (from 259,552 original tracks to 64,888 tracks) without damaging the overall classification results. Both train set reduction and code-word selection experiments provide good heuristics for drastically reducing the size of the datasets, and therefore the computation time of the classification process, specially when working in tag classification of datasets containing thousands of songs.

To the best of our knowledge there are no publications reporting tag classification for the *Million Song Dataset*. Therefore, since this is a public dataset, the here presented results can be also used as baseline for future research.

7.5 Autotag experiment II: autotagging the CAL500

7.5.1 Database

In this experiment we use the *Computer Audition Lab 500-Song* (CAL500) Dataset (Turnbull et al., 2007). This small and clean dataset consists of 500 musical tracks from 500 artists with 174 manually annotated tags related to genre, mood, instrumentation, solo instrument, music usage, and vocal characteristics (see Appendix E for a complete list of CAL500's tags). In our case, the main advantages of using this public dataset are: i) CAL500's tags are well annotated (i.e annotated by a minimum of three listeners), and ii) there are several algorithms in the literature that have reported tag classification results from this dataset. Therefore, our results can be put in perspective with respect to state-of-the-art tag classification algorithms.

As in the case of the MSD, the audio files of the dataset are not distributed by the authors due to copyright reasons. Nevertheless, the reduced number of songs that constitute this dataset allow us to manually grab from our in-house music collection the 500 tracks that form the CAL500 dataset.

7.5.2 Method

Here we use the same BoC-W algorithm as in *autotag experiment I*, but in this case, since we have the actual audio files, we compute the audio descriptors related with timbral, tonal, and energy information. In particular, as timbral descriptor we compute MFCC coefficients, for tonal information we use the Harmonic Pitch Class Profile (HPCP) descriptor, and for energy we compute the Spectral Energy (SE) descriptor (see Sec. 2.2 for further information). The implementation details of each descriptor are the following ones:

- **MFCC**: FFT frameSize=2,048 samples, FFT Hop Size=1,024 samples, FFT windowType= BlackmanHarris window with 62dB rolloff, number of triangular band-pass filters=40, low frequency bound=0 Hz, high frequency bound= 11,000 Hz. In order to be consistent with the MSD-Timbre descriptor we select the first 11 MFCC coefficients (skipping the DC coefficient). As in Chapter 3 we use the Auditory toolbox MFCC implementation (Slaney, 1998).

- **HPCP**: FFT frame size=4,096 smples, FFT hop size=2,048 samples, FFT window type=BlackmanHarris window with 62dB rolloff. In this case we use the 12 HPCP coefficients following the implementation described in Gómez (2006).
- **SE**: FFT frame size=2,048 samples, FFT hop size=1,024 samples, FFT window type=BlackmanHarris window with 62dB rolloff.

As in Chapter 3 we pre-process each audio file by applying an equal-loudness filter. This filter implements an inverted approximation of the equal-loudness curves described by Fletcher and Munson (1933).

When generating the corresponding code-words the 11 MFCC coefficients are ternary quantized, the 12 HPCP coefficients are binary quantized (with threshold = 0.5), and the SE descriptor is quantized by using 300 equal-sized steps. The threshold values used to quantize the MFCCs (33 and 66% quantiles) and SE descriptors were computed from a public medium-size dataset called CAL10k dataset (Tingle et al., 2010). As in the case of the CAL500 dataset we match 7,065 files from the CAL10k dataset to our in-house collection. Afterwards, we extract MFCC and SE descriptors using the parameters listed above. As result we obtained 14,344,343 descriptor-frames that were used to estimate the quantization thresholds.

In this experiment we work with BoC-W with TF-IDF weighting. Thus, the CAL10K dataset was also used to determine the number of documents and the code-word document frequencies that form the IDF part of the TF-IDF weighting strategy.

The aim of this experiment is to compare the BoC-W (TF-IDF) algorithm against state-of-the-art autotagging algorithms. Nevertheless, for the sake of completeness, we also included the results of the BoF approach. That is, from the computed descriptors we generate an aggregated feature vector of mean, covariance, delta mean, and delta covariance values.

For classification, we use the same linear SVM (libLinear) algorithm as in *autotag experiment I* after grid search of the C parameter (C= [0.5,1,10,100]). We use the same configuration for the BoF strategy.

This time, instead of classifying with one descriptor at a time (like in *autotag experiment I*), and since according to the previous results, Timbral descriptors seem to be more suitable for the task, we decided to evaluate the descriptors' synergies with respect to classification results by progressively adding to the MFCC feature vector the HPCP and SE feature vectors.

Finally, since most autotagging algorithms evaluated against the CAL500 dataset follow the evaluation strategy proposed by Turnbull et al. (2008), in this experiment we also adopt the same evaluation criteria. This evaluation reports per-Tag results after predicting ten tags per song (i.e. the ten most reliable tags from the output of the machine learning algorithm). In the case of BoF and BoC-W we take the ten most probable tags from the SVM probability outputs. The reported metrics are: mean Precision, mean Recall, and F-measure (i.e. globalF). All BoF and BoC-W results are means and standard errors computed from five-times tenfold cross-validation (i.e. 450-song for training, and 50-songs for testing). For the sake of completeness we also report meanF results. Moreover, besides reporting tag results for best 10-tags as in Turnbull et al. (2008) we also compute meanF and globalF measures for all predicted tags (regardless of their number) as in the case of *autotag experiment I.*

7.5.3 Results

Table 7.8 (adapted from Sordo (2011)) shows the different categories of acoustic description used by several state-of-the-art algorithms together with the BoF, and the proposed BoC-W. Interestingly, all algorithms include timbral descriptors (either alone or together with other audio features). This fact goes in concordance with our findings in *autotag experiment I* where MSD-timbre produced the best classification results when evaluating timbre, tonal, or energy descriptors separately.

Table 7.9 shows classification results for several autotagging algorithms including: three baseline models from Turnbull et al. (2008) (i.e. Random, Upper-bound, and Human), the state-of-the-art algorithms from Table 7.8[12], BoF, and the proposed BoC-W. In particular, the "Random" (lower-bound) baseline is obtained by sampling tags (without replacement) from a multinomial distribution parameterized by the tag's prior distribution computed from the ground-truth tags' frequencies. The "Upper-Bound" model is computed directly taking ten ground-truth tags per song[13]. However, since CAL500 songs have different number of tags (26 on average), and we are annotating only ten tags per song, this upper-bound is less than 100%. The "Human" model is created by comparing the annotations of one

[12] All state-of-the-art results are taken as published by their corresponding authors. Thus, we did not re-run nor re-implement these algorithms.

[13] If the song has less than 10 tags, random tags are selected until reaching 10 tags

Algorithm	Spectral	Timbral	Tonal	Rhythmic	High-Level	Autocorrelation
GMM-MH (Turnbull et al., 2008)		Yes				
Boost MFCC (Bertin-Mahieux et al., 2008)		Yes				
Boost afeats exp. (Bertin-Mahieux et al., 2008)	Yes	Yes				Yes
CBA (Hoffman et al., 2009)		Yes				
K-NN MFCC (Sordo, 2011)		Yes				
K-NN PCA (Sordo, 2011)	Yes	Yes	Yes	Yes	Yes	
CBDC (Sordo, 2011)	Yes	Yes	Yes	Yes	Yes	
BoF MFCC		Yes				
BoF MFCC+HPCP		Yes	Yes			
BoF MFCC+HPCP+SE	Yes	Yes	Yes			
BoC-W MFCC		Yes				
BoC-W MFCC+HPCP		Yes	Yes			
BoC-W MFCC+HPCP+SE	Yes	Yes	Yes			

Table 7.8: Categories of audio descriptors used by state-of-the-art algorithms (adapted from Sordo (2011). We also include the different descriptors sets for BoF and BoC-W evaluated in this experiment (see text for details). GMM-MH corresponds to Mixture Hierarchy Gaussian Mixture Model, CBA refers to the Code Bernulli Average algorithm, “Boost MFCC” and “Boost afeat exp.” correspond to FilterBoost algorithms usng MFCCs or a set of multiple features. Finally, CBDC refers to the Class-Based Distance Classification algorithm (see Sec. 7.2 for further details).

Algorithm	MeanP	MeanR	GlobalF	MeanF
Random	14.40 (0.40)	6.40 (0.20)	8.86	N.A.
Upper-Bound	71.20 (0.70)	37.50 (0.60)	49.13	N.A.
Human	29.60 (0.80)	14.50 (0.30)	19.46	N.A.
GMM-MH (Turnbull et al., 2008)	26.50 (0.70)	15.80 (0.60)	19.80	N.A.
Boost MFCC (Bertin-Mahieux et al., 2008)	28.10 (6.60)	13.10 (1.90)	17.87	N.A.
Boost afeats exp. (Bertin-Mahieux et al., 2008)	31.20 (6.00)	15.30 (1.50)	20.53	N.A.
CBA (Hoffman et al., 2009)	28.60 (0.50)	16.20 (0.40)	20.68	N.A.
K-NN MFCC (Sordo, 2011)	20.50 (0.80)	9.00 (0.40)	12.51	N.A.
K-NN PCA (Sordo, 2011)	23.10 (0.80)	10.10 (0.40)	14.05	N.A.
CBDC (Sordo, 2011)	27.70 (0.80)	19.20 (0.70)	22.68	N.A.
BoF MFCC	13.27 (0.23)	17.10 (0.31)	14.93	4.60 (0.10)
BoF MFCC+HPCP	13.36 (0.24)	17.27 (0.32)	15.07	4.87 (0.09)
BoF MFCC+HPCP+SE	13.65 (0.22)	16.90 (0.30)	15.10	4.78 (0.09)
BoC-W MFCC	21.42 (0.30)	21.47 (0.27)	21.44	11.66 (0.16)
BoC-W MFCC+HPCP	21.49 (0.29)	20.81 (0.21)	21.14	11.93 (0.13)
BoC-W MFCC+HPCP+SE	21.21 (0.30)	21.19 (0.28)	21.20	12.34 (0.18)

Table 7.9: Autottaging classification results for BoC-W, BoF, state-of-the-art algorithms, and baseline models (random, upper-bound, and human). All results are presented as percentage. Standard Deviations are depicted in parenthesis. GlobalF corresponds to F-measure computed from mean Precision and mean Recall. BoF and BoC-W results correspond to best C parameter after grid search with $C = 0.5$ for BoF MFCC and BoF MFCC+HPCP+SE, $C = 10$ for BoC-W MFCC and BoC-W MFCC+HPCP, and $C = 100$ for BoF MFCC+HPCP and BoC-W MFCC+HPCP+SE. GMM-MH corresponds to Mixture Hierarchy Gaussian Mixture Model, CBA refers to the Code Bernulli Average algorithm, “Boost MFCC” and “Boost afeat exp.” correspond to FilterBoost algorithms usng MFCCs or a set of multiple features. Finally, CBDC refers to the Class-Based Distance Classification algorithm (see Sec. 7.2 for further details).

Algorithm	MeanP	MeanR	GlobalF	MeanF
BoF MFCC	22,75 (1,42)	32,57 (2,26)	26,79	25,46 (1,49)
BoF MFCC+HPCP	23,61 (1,67)	26,90 (1,91)	25,15	24,06 (1,51)
BoF MFCC+HPCP+SE	23,44 (1,67)	27,46 (2,04)	25,29	24,25 (1,57)
BoC-W MFCC	21,09 (0,98)	55,44 (3,33)	30,55	26,55 (1,24)
BoC-W MFCC+HPCP	21,68 (1,06)	53,86 (3,03)	30,92	26,62 (1,29)
BoC-W MFCC+HPCP+SE	21,34 (1,15)	51,28 (3,08)	30,14	25,59 (1,36)

Table 7.10: Autottaging classification results for BoC-W, and BoF for all predicted tags. All results are presented as percentage. Standard Deviations are depicted in parenthesis. All C parameters are equal to 0.5.

subject against a ground truth formed by the annotations of at least other four individuals.

Noticeably, results from Table 7.9 show that the proposed BoC-W approach outperforms all state-of-the-art algorithms but the CBDC algorithm (which is 1.24 percentage points above). Moreover, whilst the CBDC algorithm uses multiple descriptor categories (see Table 7.8), our BoC-W does not benefit from adding tonal and energy descriptors to BoC-W MFCC. Therefore, the BoC-W approach offers a simple approach that relies only in one timbral descriptor and offers excellent classification results when compared with more complex state-of-the-art algorithms. Interestingly, GMM-MH, Boost afeats exp., CBA, CBDC, and BoC-W algorithms provide globalF results that are better than the "Human" baseline.

If we now focus on results from the BoF approach we can see that despite using the same descriptors as the BoC-W, the obtained classification results are far below the ones obtained with the BoC-W algorithm (and other state-of-the-art algorithms such as CBDC or CBA). Regarding meanF results we observe an important difference of approximately 10 percentage points below the corresponding globalF results (much more than the difference between MSD-Tag's meanF and globalF results). This difference could be linked to the fact of taking only 10-tags and the relatively small size of the dataset, but unfortunately, meanF results were not reported for the other approaches and we do not know if this difference is also present in the other algorithms.

Now, instead of forcing the per-song predictions to be equal to ten labels we take, as in *autotag experiment I*, all tags predicted by the SVM, regardless of their number. In this case, both BoF and BoC-W algorithms offer better results (see Table 7.10) and the difference between globalF and meanF is smaller. In particular, the best results for BoC-W are 30.92% and 26.62% for globalF and meanF (BoC-W MFCC+HPCP), and the best BoF results are 26.79% and 25,46% (globalF and meanF) for BoF MFCC. Furthermore, we observe that the improvement in F-measure results for the BoC-W approach is due to an increment in recall values which are above 50% in all BoC-W approaches[14].

Finally, also as in *autotag experiment I*, we analyze how individual tag performances are related to the a priori tag frequencies. Thus, Table 7.11 shows F-measure classification results for the 10 best and 10 worst classified tags for the BoC-W MFCC algorithm. The table shows that besides the intrinsic characteristics of each tag, there is a strong correlation between tag frequency and classification results (i.e. low frequent tags are also poorly classified and *vice versa*). The correlation coefficient between all F-measure and tag frequency values also indicates a strong correlation of 0.9101 (see Appendix E for the complete list of per tag results, and a scatter plot of classification results vs. tag-frequencies). This fact stresses the importance of having enough examples in the dataset for all to-be-classified elements, one of the main weakness of working with small datasets.

7.5.4 Conclusion for autotag experiment II

In this experiment we have evaluated the proposed BoC-W algorithm for automatic tag classification within a small and well-annotated public dataset (i.e. the CAL500 dataset). After comparing the performance of the BoC-W approach (with TF-IDF weighting) against the BoF approach and several state-of-the-art algorithms we observe that the simple BoC-W MFCC approach outperforms all other algorithms except for the more sophisticated CBDC algorithm proposed by Sordo (2011) that uses multiple descriptors. Once more, the BoC-W approach offers significantly better results than the standard BoF approach computed from the same descriptors and same classification algorithm. With respect to the CAL500 dataset, we have spotted the performance differences between computing meanF and globalF results. Finally, we have also stressed the importance of having enough tag examples in the dataset to achieve good classification results. Otherwise, the classification results are affected by those underrepresented tags.

[14] Recall values reflect the fraction of correctly classified items over the total of items that belong to the class.

F-measure	Category	Tag	Tag Frequency
		10 best classified tags	
74.31	Song	Texture-Electric	326
72.85	Song	Recorded	444
69.81	Song	High-Energy	231
68.38	Song	Quality	287
66.86	NOT-Emotion-Tender	Soft	206
66.05	Instrument	Drum-Set	275
65.71	Instrument	Male-Lead-Vocals	339
63.70	NOT-Emotion-Touching	Loving	219
62.92	NOT-Emotion-Angry	Aggressive	319
61.91	NOT-Emotion	Sad	221
		10 worst classified tags	
1.82	Genre-Best	Soul	5
1.64	Instrument	Acoustic-Guitar-Solo	6
1.63	Usage	Waking-up	8
1.11	Genre	Bebop	6
1.11	Genre	Roots-Rock	8
0.95	Instrument	Female-Lead-Vocals-Solo	7
0.00	Genre	Contemporary-Blues	7
0.00	Genre	Country-Blues	6
0.00	Usage	With-the-family	5
0.00	Instrument	Trumpet-Solo	6

Table 7.11: 10 best and 10 worst tag classification results for BoC-W MFCC. F-measure results in percentage. Category and tag names correspond to the original CAL500 definition.

7.6 Discussion and conclusion

In this chapter we have proposed and evaluated a new algorithm for automatic tagging of audio files. This algorithm is inspired by the similarities between text distributions and the distribution results of the code-word encoding proposed in previous chapters of this thesis. Thus, we apply text-IR classification techniques to our encoded audio descriptors.

The proposed BoC-W algorithm provides sate-of-the-art results a in small and clean dataset such as CAL500 and, more importantly, scale well to real-life large and noisy datasets such as the Million Song Dataset. In all cases the obtained results for the BoC-W approach are much better that using the classic BoF approach. Moreover, the fact that we use a simple encoding strategy that provides a very large dictionary leading to a very sparse feature vector with similar characteristics as the bag-of-words features used in document classification, enables us to use fast IR algorithms specially developed to work with this type of features (like the SVM LibLinear implementation). Furthermore, results from code-word selection and train set reduction experiments can be exploited as a way to reduce the computation time when classifying datasets containing thousands of instances.

Finally, since in the conducted experiments, we use two public datasets, the here presented results provide an important comparative landmark for future research in autotagging. Specially in the case of the MSD where, up to our knowledge, this are the first autotagging results presented for this corpus.

Chapter 8

Final conclusions

As stated in Chapter 1 the main goals of this thesis were: a) to analyze the statistical distributions of commonly used MIR audio-content descriptors as found in large datasets of real-world polyphonic music; and b) to use the acquired knowledge regarding the distribution patterns of descriptors to contribute to current MIR tasks and, if possible, address new tasks. On the light of the results presented in this work we can say that these goals have been fully accomplished. On the other hand, we feel that this thesis constitutes a promising starting point and, at the present time, we envision many future applications that could take profit of the here described approach (see Chapter 9). Therefore, there is still much work to be done in the near future.

We started this manuscript by providing a general description of current popular content-based MIR algorithms, together with their main recognized problems, and possible solutions as suggested by some MIR researches (Chapter 1). Then, we highlighted some undergoing assumptions regarding the statistical distribution of frame-wise audio features and, at the same time, we stressed the lack of research in trying to characterize such distributions. Moreover, since this characterization effort has proven its validity in related research fields we incorporated this aim as one of the thesis goals.

In Chapter 2 we presented background information regarding the main topics addressed in the thesis. Hence, we provide a review on heavy-tail distributions (Sec. 2.1), audio descriptors (Sec. 2.2), encoding methods (Sec. 2.3), and audio classification (Sec. 2.4).

Next, in Chapter 3, we proposed an unsupervised encoding strategy to study the rank-frequency distribution of timbral code-words as found in large databases of Western and non-Western music, speech, and natural sounds. After conducting several experiments we concluded that the distribution of the studied timbral features (i.e. Bark-band energies and MFCCs) for all four databases is well fitted by a power-law function. Moreover, a detailed study of the Bark-band energies descriptor led us to conclude that short-time spectral envelopes are also heavy-tailed distributed. Furthermore, our experiments also showed that short-time spectral envelopes present characteristic morphological differences between their most and least frequent elements.

Then, we wondered if a common mechanism that generates the observed power-law distributions of encoded Bark-band energies could be found. In order to answer this question, in Chapter 4, we reviewed the most suitable power-law generative models described in the literature. After selecting two plausible model candidates we tried to recreate the empirically observed power-law distributions by performing a grid search on each model's parameters. From this process we found that a modified version of the traditional Yule-Simon model was able to produce quite similar distribution results. Remarkably, with respect to our data, this model implies that when music, speech or natural sound signals are being recorded or analyzed, we can expect that recently occurred spectral envelopes have a great chance of reappearing in the close next future.

Up to this point we had focused on the distribution of timbral code-words as found in full databases. However, when plotting the rank-frequency distribution of code-words for individual songs we had also observed similar heavy-tailed distributions as the ones depicted in the full databases. Hence, in Chapter 5 we provided further evidence to support the fact that timbral features from individual songs are also heavy-tailed distributed. For that we performed frame selection experiments for two classification tasks namely: genre classification and musical instrument detection. We demonstrated, using the standard BoF approach, that working with a reduced set of most frequent MFCC frames, either obtained by using code-words' frequencies or by simply selecting random frames, produces similar classification results as working with all song's frames. This fact can be perfectly explained if the song-wise distribution of MFCC frames follows a heavy-tail distribution.

In Chapter 6 we decided to use the proposed encoding algorithm for a twofold purpose. The first purpose was to analyze the distribution of timbral, tonal and energy descriptors as found in a very large music dataset. The second purpose was to measure the evolution of popular Western music by characterizing the year-based patterns of features' distributions. For that we used the publicly available *Million Song Dataset*, and analyzed more than 460,000 songs with yearly annotations between 1955 and 2010. Regarding the distribution of the database's timbral feature we found that, as in the previous experiments, its distribution can be well fitted by a power-law. With respect to the tonal descriptor we found that its distribution also follows a power-law. In particular, the best fit corresponds to a shifted power-law distribution. Finally, the energy descriptor follows a reversed log-normal distribution. On the other hand, the year-based evolution of popular Western music showed that the distribution patterns of the tonal descriptor remained practically unaltered for all the analyzed time period (1955-2010). The timbral content showed a consistent homogenization of the timbral palette with timbral popularity varying across years. Finally the energy descriptor depicted a characteristic trend towards ever louder recording. The proposed historical analysis of music content offers a new perspective and new empirical evidence through a formal, quantitative, and systematic analysis of large-scale music collections. Thus, this new source of information could be used to complement other musicological research.

Finally, inspired by text retrieval algorithms, in Chapter 7 we proposed new audio descriptors that take advantage of the proposed encoding and the found distributions. We evaluated these new "Bag-of-Code-Words" features in the complex task of automatic tagging of songs. Noticeably, the proposed features provided state-of-the-art results when used to classify a small and well-annotated database. Moreover, since the proposed algorithm scale well to very large datasets, we were able to report, for the first time, on tagging results for the Million Song Dataset. In this case, using the "Bag-of-Code-Words" features offered much better results than using the standard "Bag-of-Frames" approach. Furthermore, we conducted code-word selection, and train set reduction experiments that can be exploited to reduce the computational time when classifying datasets containing thousands of instances.

From a global perspective this work shows that, contrary to what is usually assumed, common frame-wise audio features are heavy-tailed distributed. Noticeably, this type of distributions are also observed for image and text documents. This new knowledge led us to propose new strategies that exploit this fact and provide better results than the standard Bag-of-Frames approach. Moreover, the proposed approach is able to address common MIR tasks such as genre classification, instrument detection and automatic audio tagging, and, at the same time, can be used to objectively characterize the evolution of audio content coming from very large datasets.

The heavy-tail distribution of audio features, together with the discovered evidence that short-time spectral envelopes are also heavy-tailed distributed, lead us to wonder on the implications with respect to standard MIR algorithms. In particular, current feature summarization methods and distance measures within the feature space should be re-assessed from this new perspective. Furthermore, this here-discovered property could be one of the causes of known MIR problems such as the existence of a "glass ceiling" in audio classification and the presence of "hub" songs. Thus, we urge researchers to do more work in this direction. Moreover, in the next chapter we introduce some promising paths for future research.

Finally, according to the well known "efficient coding hypothesis" there is a quantitative link between the statistical properties of the world and the structure of the perceptual system (Geisler, 2008; Simoncelli, 2003). Hence, since heavy-tailed distribution patterns seem to be a core statistical property of short-time audio segments, we expect that this fact should be also exploited by future "artificial perception" algorithms involved in sound processing, sound creation, and machine listening tasks.

Chapter 9

Future work

During the creation process of this thesis many ideas have appeared in the form of interesting new tasks and experiments. Unfortunately, time is finite and one has to prioritize what experiments to conduct and what others to include in the future-research bag. Nevertheless, moved more by curiosity than by strict experimental design, we have performed a set of exploratory experiments on some of the to-do-later ideas. Hence, in this chapter we briefly report on some of these informal experiments whose outcomes depicted promising results. Therefore, we encourage future researches to further investigate on these subjects.

9.1 BoC-W for artist identification

The automatic identification of the performer/s of a song (i.e. artist identification) from the song's audio content has attracted increasing interest from MIR researches during the last few years (Ellis, 2007; Kim et al., 2006). The interest on this task is not only on trying to label songs with missing artist names, or identifying forgeries or false attributions, but also on trying to understand how this task is so easily performed by humans (e.g. what are the audio features that help most in achieving this task?).

In this case we perform an exploratory study regarding electronic music artist identification. These experiments are included in the master thesis of Melidis (2012). The goal of the master's thesis was to report on the challenging task of automatic artist classification within a particular musical

Approach	**MFCC**	**HPCP**	**SE**
Essentia	19.42%	8.48%	2.25%
MIRToolbox	4.74%	3.39%	6.66%
BoC-W	21.37%	8.62%	8.17%

Table 9.1: Electronic artist identification results (F-measure) from Melidis (2012). The Essentia, and MIRToolbox approaches correspond to the standard BoF approach with features extracted using either Essentia or the MIRToolbox libraries. MIRToolbox results indicate hold-out set for testing.

genre. This task is more difficult than traditional artist identification because, since in this experiments, all artist belong to the same musical genre, their music would much probably share several stylistic features. Thus, the classification task becomes more difficult, even for human listeners. In particular, Melidis (2012) gathered a collection of 111 electronic music artists, each of them contributing with 5 albums. The training of the algorithms was done using 3 albums whereas the remaining 2 were used for testing. Moreover, train and test sets were interleaved along time, e.g. artist X having albums from 2000 (train), 2001 (test), 2003 (train), 2005 (test), 2006 (train).

This database was analyzed by several types of audio features (from Essentia[1] and the MIRToolbox[2]), and classification algorithms. The best classification result, 24.8% F-measure, was obtained with a set of 315 features classified with SVMs in 5-fold cross-validation. This result, even not very high, is far from the random classification baseline of 1.1% F-measure. Noticeably, if instead of the full set of descriptors, an aggregated feature vector of mean and covariance MFCCs is used, the F-measure performance stayed relatively high at 19.4%. This fact points towards the relevance of timbral information for artist identification. At this point we performed an exploratory evaluation of our BoC-W descriptor for MFCCs with TF-IDF weighting (see Chapter 7). Interestingly, we obtained better classification results than the standard MFCCs approach (BoF). In particular, the final F-measure result for the BoC-W was 21.4% (see Table 9.1).

Thus, automatic artist identification seems to be a MIR task that could also benefit from our BoC-W approach. In future research, we would like to evaluate the BoC-W from different feature combinations, and extend this

[1] http://essentia.upf.edu/

[2] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

within-genre artist identification evaluation to bigger datasets containing artist from different musical genres.

9.2 Tonal "unexpectedness" descriptor

In this experiment we propose a new audio descriptor that aims at describing unexpected tonal segments of music. Since the TF-IDF weighting of each audio frame (see Sec. 7.3.1) provides high values when the code-word occurs many times within a small number of songs; mid values when the code-word is not-so-frequent in a song, or occurs in many songs; and low values when the code-word is present in virtually all songs, we propose to use this weighting strategy, computed from tonal code-words, as a measure of tonal "unexpectedness" of the frames within a song. Thus, high tonal "unexpectedness" values will correspond to pitch-class combinations that are used many times in a given song but are rare in most of other songs, etc.

In particular, the tonal "unexpectedness" descriptor is computed as follows:

1. Frame-based HPCP features are computed (see Sec. 2.2).
2. HPCP frames are transposed to a common key (see Sec. 6.2).
3. Each HPCP frame is encoded into a 12 dimensional tonal code-word (using binary EFD; see Sec. 6.2).
4. For each tonal code-word we define the "unexpectedness" value as the TF-IDF weight of the frame computed as in Sec. 7.3.1. In this case we used the CAL10k dataset (Tingle et al., 2010) to estimate the global distributions.
5. The tonal "unexpectedness" descriptor corresponds to the time series of code-words "unexpectedness" values.

In order to evaluate the proposed descriptor we take advantage of the audio content used as stimuli in Koelsch et al. (2008)[3]. In particular, Koelsch et al. (2008) investigated neural correlates of music processing by recording event-related brain potentials, skin conductance responses and heart rate

[3] Downloaded from `http://www.stefan-koelsch.de/stimulus_repository/Koelsch+_2008/Repository6.html`

while listening to authentic music stimuli. The authors created a database with 28 excerpts from classical piano sonatas performed by professional pianists and recorded in MIDI format. These excerpts contain music-syntactic irregularities in the form of “unexpected” irregular ending chords as originally arranged by the composers. Then, by manipulating the recorded MIDI notes, the authors created two more sets namely: one set where the irregular chords were transformed into “expected” regular endings, and another set with “very unexpected” endings. Moreover, by further manipulation of the MIDI notes the authors created a second dataset with the three aforementioned sets but, in this case, played without musical expression (i.e. without variations in tempo and loudness). The study shows a number of physiological responses from the played stimuli for unexpected endings independently of the emotional qualities the stimuli.

In our case, we use the above dataset, and information regarding the trigger time of the final chord[4], to evaluate how the proposed tonal ”unexpectedness” descriptor behaves with such dataset. Hence, figures 9.1 and 9.2 show the mean “unexpectedness” values for the three sets of 28 excerpts with and without musical expression respectively. Noticeably, these encouraging results clearly show that our descriptor reacts to unexpected and very unexpected chords regardless of the musical expression[5]. Remarkably, this behavior is similar to the physiological responses reported by Koelsch et al. (2008).

In future research we plan to evaluate the usefulness of the proposed descriptor within MIR tasks such as automatic mood classification and musical surprise detection. We also believe that this new audio feature could help in the automatic segmentation of music.

9.3 BoC-W weights for detecting regions-of-interest

Following the previous idea of the tonal unexpectedness descriptor we conducted a series of informal tests with other encoded descriptors (e.g. MFCC, and SE). Interestingly, the time series of TF-IDF weightings ramped in song segments that were somehow different from the rest of the song. Depending on the used descriptor these region-of-interest (ROI) zones were “unexpected” with respect to timbre, energy, etc. Thus, a systematic evaluation

[4] Information generously provided by Dr. Koelsch and collaborators in personal communication.

[5] Quartile information shows a similar behavior than mean values.

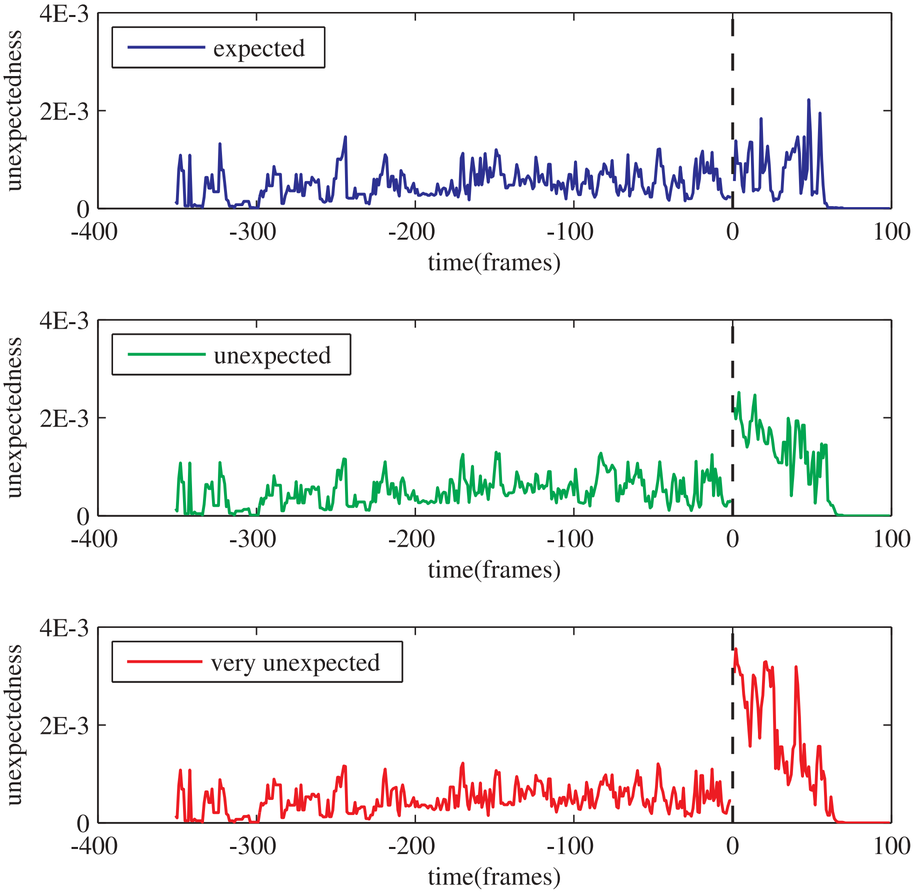

Figure 9.1: Mean values for the tonal "unexpectedness" descriptor (y-axis) from 28 stimuli **with** musical expression as reported in Koelsch et al. (2008). The x-axis corresponds to time in frames of 2,048 samples. At time = 0 the final (expected, unexpected, or very unexpected) chord is played.

of this approach is needed in the future. This task will require an annotated database of ROI for songs. Unfortunately, up to our knowledge there is not such public dataset available.

Noticeably, Caron et al. (2007) use power-law models to detect ROI in image data. In this work ROIs are detected from automatic descriptions of the image's real rank-frequency distribution and its fitted power-law. We

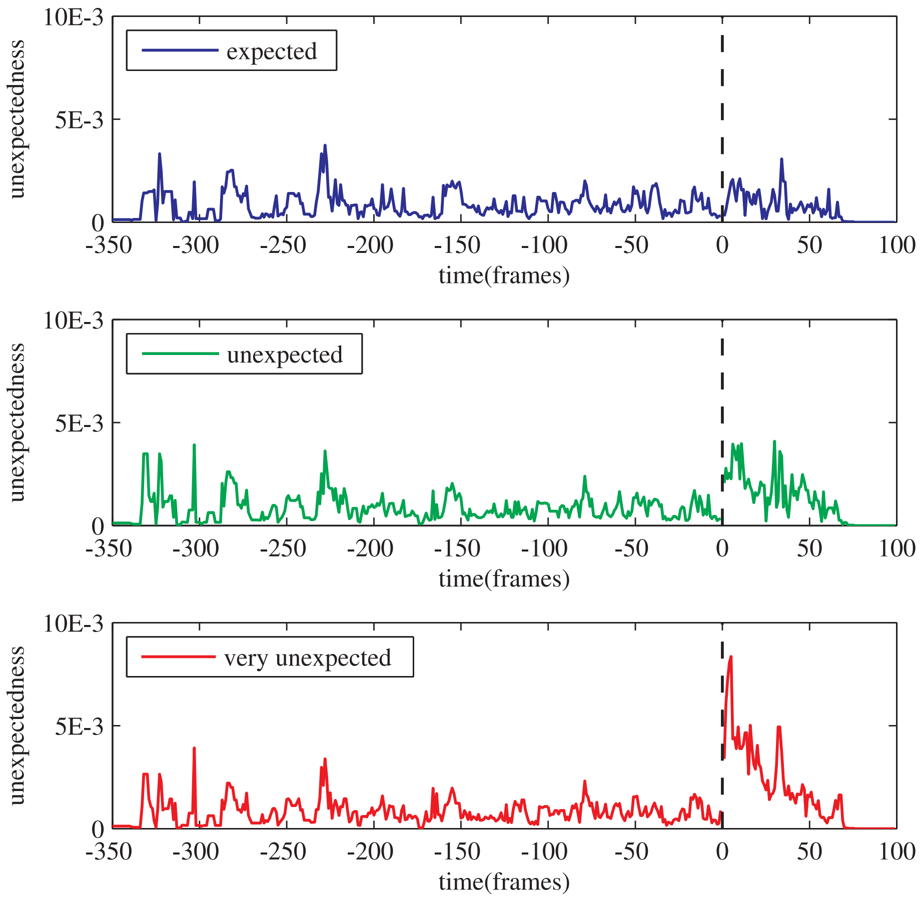

Figure 9.2: Mean values for the tonal "unexpectedness" descriptor (y-axis) from 28 stimuli **without** musical expression as reported in Koelsch et al. (2008). The x-axis corresponds to time in frames of 2,048 samples. At time = 0 the final (expected, unexpected, or very unexpected) chord is played.

believe that a similar approach can be also applied to our code-words.

Finally, if frame-wise code-word weightings are proven to be useful for detecting ROIs in audio files, we believe that a song-level "interestingness" descriptor could provide useful information for many classification tasks. This "interestingness" descriptor can be computed as the number of ROI frames divided by the total number of frames of the audio file, where ROI frames are those frames with TF-IDF weights above some pre-defined threshold.

9.4 Concluding thoughts

In the first chapter of this thesis we highlighted three main problems of BoF algorithms working with polyphonic music namely: the need for better audio descriptors, the need for incorporating temporal information, and the need to better understand the feature vector space. Now, at the end of this work, we see these three problems as highly interlinked. Thus, we have started the thesis by studying the feature vector space (problem number three). Then, the unexpected results we have discovered -i.e. that the feature vector space is distributed in a heavy-tailed manner- led us to propose new audio descriptors thus, moving to problem number one. Finally, in this chapter we have proposed new frame-level features that can be used to detect interesting regions of songs. This approach can be easily adapted to incorporate more temporal information within our descriptor sets. For instance, we could compute one set of features from ROI segments and another set for non-ROI ones. Moreover, looking once more at the IR community, it would be very interesting to work with code-word n-grams for classification. That is, to work with sequences of n code-words thus reflecting temporal relationships between frames. All this approaches are directly linked with problem number two. Finally, we expect that studying n-gram distributions and statistics from code-word networks (as we started to do in Serrà et al. (2012b)) will provide further solutions to tackle problem number three and thus, this virtuous circle will start again.

Bibliography

The numbers at the end of each bibliographic entry indicate the pages in which it is cited.

L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143–150, 2002. 10, 79

American National Standards Institute. *Psychoacoustical terminology S3.20.* ANSI/ASA, 1973. 14

M. F. Assaneo, J. I. Nichols, and M. A. Trevisan. The anatomy of onomatopoeia. *PLoS ONE*, 6(12):e28317, 2011. doi: 10.1371/journal.pone.0028317. 50

H. Attias and C. E. Schreiner. Temporal low-order statistics of natural sounds. In *NIPS*, pages 27–33. MIT Press, 1997. 12

J. J. Aucouturier and E. Bigand. Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, pages 1–15, 2013. Article in Press. 6, 72

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. 6

J. J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *Journal of the Acoustical Society of America*, 122(2):881–891, 2007. 4

J. J. Aucouturier and F. Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, 41(1): 272–284, 2008. ISSN 0031-3203. doi: 10.1016/j.patcog.2007.04.012. 6, 69

R. Baeza-Yates. *Modern information retrieval.* ACM Press, Addison-Wesley, 1999. ISBN 9780201398298. 5, 12, 25, 27, 53, 70, 87

P. Bak. *How nature works: the science of self-organized criticality.* Copernicus, New York, 1996. ISBN 038798738X. 10

P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters*, 59(4):381, 1987. doi: 10.1103/PhysRevLett.59.381. 56

P. Ball. *The Music Instinct: How Music Works and Why We Can't Do Without It.* Oxford University Press, USA, September 2010. ISBN 0199754276. 14, 71, 73, 81

A. L. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187, 1999. 55, 57

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, 13(5):1035–1047, September 2005. ISSN 1063-6676. doi: 10.1109/TSA.2005.851998. 20

M. Beltrán del Río, G. Cocho, and G. G. Naumis. Universality in the tail of musical note rank distribution. *Physica A*, 387(22):5552–5560, 2008. ISSN 0378-4371. doi: 10.1016/j.physa.2008.05.031. 12, 83

E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: Breaking the glass ceiling. pages 379–384, Porto, Portugal, 2012. FEUP Edies. 6

R. E. Berg and D. G. Stork. *The physics of sound.* Prentice Hall, 2 edition, 1995. ISBN 9780131830479. 14, 28, 31, 60

T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: a model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research, special issue: "From genres to tags: Music Information Retrieval in the era of folksonomies."*, 37(2): 115–135, 2008. 90, 108, 109

T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 591–596, 2011. 17, 72, 73, 94, 145

M. Bethge, D. Rotermund, and K. Pawelzik. Second order phase transition in neural rate coding: binary encoding is optimal for rapid signal transmission. *Phys Rev Lett*, 90(8):088104, 2003. doi: 10.1103/PhysRevLett.90.088104. 24

K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "Nearest neighbor" meaningful? In *Proceedings of the 7th International Conference on Database Theory*, pages 217–235. Springer-Verlag, 1999. ISBN 3-540-65452-6. 23

E. Bigand, C. Delbé, Y. Gérard, and B. Tillmann. Categorization of extremely brief auditory stimuli: Domain-Specific or Domain-General processes? *PLoS ONE*, 6(10):e27024, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0027024. 52

D. Bogdanov, J. Serra, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4):687–701, 2011. ISSN 1520-9210. doi: 10.1109/TMM.2011.2125784. 6

A. S. Bregman. *Auditory scene analysis: the perceptual organization of sound*. The MIT Press, 1990. ISBN 0262022974. 14, 51, 56, 76

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121167, 1998. 24

Y. Caron, P. Makris, and N. Vincent. Use of power law models in detecting region of interest. *Pattern Recogn.*, 40(9):25212529, 2007. ISSN 0031-3203. doi: 10.1016/j.patcog.2007.01.004. 12, 123

M. Casey and M. Slaney. The importance of sequences in musical similarity. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, pages V–5–V–8, Toulouse, France, 2006. doi: 10.1109/ICASSP.2006.1661198. 5

M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008. ISSN 0018-9219. doi: 10.1109/JPROC.2008.916370. xxiv, 1, 4, 7, 28, 29, 63, 72, 76, 81, 87, 90

C. Cattuto, V. Loreto, and V. D. P. Servedio. A yule-simon process with memory. *Europhysics Letters (EPL)*, 76(2):208–214, 2006. ISSN 0295-5075. doi: 10.1209/epl/i2006-10263-9. 57

C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proc Natl Acad Sci USA*, 104(5):1461 –1464, 2007. doi: 10.1073/pnas.0610487104. xxiv, 55, 56, 57, 58, 60, 61

O. Celma. *Music Recommendation and Discovery in the Long Tail.* PhD thesis, Universitat Pompeu Fabra, 2008. 89

O. Celma, P. Herrera, and X. Serra. Bridging the music semantic gap. In *Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, volume 187, Budva, Montenegro, 2006. CEUR. 2

C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 96

N. Chater and G. D. A. Brown. Scale-invariance as a unifying psychological principle. *Cognition*, 69(3):B17–B24, 1999. ISSN 0010-0277. doi: 10.1016/S0010-0277(98)00066-3. 51

S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1:379–416, 1986. 79

K. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan. *Data mining : a knowledge discovery approach.* Springer, New York, 2007. ISBN 9780387333335. 21, 22, 77

E. J. Clarke and B. A. Barton. Entropy and MDL discretization of continuous variables for bayesian belief networks. *International Journal of Intelligent Systems*, 15(1):6192, 2000. ISSN 1098-111X. doi: 10.1002/(SICI)1098-111X(200001)15:1⟨61::AID-INT4⟩3.0.CO;2-O. 22

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661, 2009. ISSN 00361445. doi: 10.1137/070710111. 9, 10, 11, 33, 78, 151, 152, 153

B. Corominas-Murtra and R. V. Solé. Universality of Zipf's law. *Phys Rev E*, 82(1):011102, 2010. doi: doi:10.1103/PhysRevE.82.011102. 56

B. Corominas-Murtra, J. Fortuny, and R. V. Solé. Emergence of zipf's law in the evolution of communication. *Phys. Rev. E*, 83:036115, Mar 2011. doi: 10.1103/PhysRevE.83.036115. 10

A. Corral, A. Ossó, and J. E. Llebot. Scaling of tropical-cyclone dissipation. *Nature Physics*, 6:693–696, 2010. 154

A. Corral, F. Font, and J. Camacho. Non-characteristic half-lives in radioactive decay. *Phys Rev E*, 83:066103, 2011. 33, 153

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, September 1995. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00994018. 24, 65

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357– 366, 1980. ISSN 0096-3518. doi: 10.1109/TASSP.1980.1163420. 63

T. De Clercq and D. Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(1):47–70, 2011. 81

E. H. Decker, A. J. Kerkhoff, and M. E. Moses. Global patterns of city size distributions and their fundamental drivers. *PLoS ONE*, 2(9):e934, 2007. doi: 10.1371/journal.pone.0000934. 10

E. Deruty. 'Dynamic range' and the loudness war. *Sound on Sound – September 2011*, pages 22–24, 2011. 83, 84

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *ICML*, page 194202, 1995. 21

I. Eliazar and J. Klafter. A unified and universal explanation for Levy laws and 1/f noises. *Proc Natl Acad Sci USA*, 106(30):12251–12254, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0900299106. 55

D. Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, pages 339–340, 2007. 119

R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008. ISSN 1532-4435. 94

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In R. Bajcsy, editor, *IJCAI*, pages 1022–1029, Chambéry, France, 1993. Morgan Kaufmann. 22

R. Ferrer-i-Cancho and B. Elvevåg. Random texts do not exhibit the real zipf's law-like rank distribution. *PLoS ONE*, 5(3):e9411, 2010. doi: 10.1371/journal.pone.0009411. 36

R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proc Natl Acad Sci USA*, 100(3):788 –791, 2003. doi: 10.1073/pnas.0335980100. 10, 51, 55, 56, 102

H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *J Acoust Soc Am*, 5(2):82, 1933. ISSN 00014966. doi: 10.1121/1.1915637. 31, 106

A. Flexer, D. Schnitzer, M. Gasser, and T. Pohle. Combining features reduces hubness in audio similarity. In *ISMIR*, pages 171–176, 2010. 69

Z. Fu, G. Lu, K. M. Ting, and D. Zhang. Music classification via the bag-of-features approach. *Pattern Recognition Letters*, 32(14):1768–1777, October 2011. ISSN 0167-8655. doi: 10.1016/j.patrec.2011.06.026. 4, 22, 23, 88

F. Fuhrmann. *Automatic musical instrument recognition from polyphonic music audio signals.* PhD thesis, Universitat Pompeu Fabra, 2012. 66

F. Fuhrmann, M. Haro, and P. Herrera. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. In *ISMIR*, pages 321–326, Kobe, Japan, 2009. 5, 20

T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. *International Computer Music Conference Proceedings*, pages 464–467, 1999. 18

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. *TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.* Linguistic Data Consortium,, [Philadelphia, Pa.] :, 1993. ISBN 9781585630196. 30, 146

W. S. Geisler. Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59(1):167–192, 2008. doi: 10.1146/annurev.psych.58.110405.085632. PMID: 17705683. 118

E. Gómez. *Tonal description of music audio signals.* PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006. 19, 76, 106

E. Gómez and P. Herrera. The song remains the same identifying versions of the same piece using tonal descriptors. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pages 180–185, 2006. 20

E. Gómez, B. Ong, and P. Herrera. Automatic tonal analysis from music summaries for version identification. In *AES 121th Convention*, page 6902, 2006. 20

E. Gómez, P. Herrera, P. Cano, J. Janer, J. Serrà, J. Bonada, S. El-Hajj, T. Aussenac, and G. Holmberg. Music similarity systems and methods using descriptors. 2008. patent num. WO 2009/001202. 19

E. Gómez, M. Haro, and P. Herrera. Music and geography: content description of musical audio from different parts of the world. In *10th International Society for Music Information Retrieval Conference*, pages 753–758, Kobe, Japan, 2009. 146, 147

F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005. 20, 73

E. Guaus. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.* PhD thesis, Universitat Pompeu Fabra, 2009. 66

B. Gutenberg and C. F. Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944. ISSN 0037-1106, 1943-3573. 10

J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proceedings of the 3rd Conference on Music Information Retrieval (ISMIR)*, page 107–115, 2002. 24

S. Harding, M. Cooke, and P. König. Auditory gist perception: an alternative to attentional selection of auditory streams? Attention in Cognitive Systems, Lecture Notes in Artificial Intelligence,. Springer-Verlag, 2008. ISBN 4840:399-416. 52

M. Haro and P. Herrera. From low-level to song-level percussion descriptors of polyphonic music. In *10th International Society for Music Information Retrieval Conference*, pages 243–248, Kobe, Japan, 2009. 5, 20

M. Haro, J. Serrà, A. Corral, and P. Herrera. Power-law distribution in encoded MFCC frames of speech, music, and environmental sound signals. In *21st International World Wide Web Conference (WWW 2012): 4th International Workshop on Advances in Music Information Research (AdMIRe 2012)*, pages 895–902, Lyon, April 2012a. ACM. 27, 63

M. Haro, J. Serrà, P. Herrera, and A. Corral. Zipf's law in short-time timbral codings of speech, music, and environmental sound signals. *PLoS ONE*, 7(3):e33993, 2012b. doi: 10.1371/journal.pone.0033993. 27, 55

A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 627–634, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-733-9. doi: 10.1145/1282280.1282369. 22

M. D. Hauser, N. Chomsky, and W. T. Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569 –1579, 2002. doi: 10.1126/science.298.5598.1569. 50

H. He and E. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.239. 94

P. Herrera, A. Klapuri, and M. Davy. Automatic classification of pitched musical instrument sounds. Signal Processing Methods for Music Transcription, pages 163–200. Springer, New York, 2006. ISBN 0-387-30667-6. 14, 16

P. Herrera, J. Serrà, C. Laurier, E. Guaus, E. Gómez, and X. Serra. *The Discipline formerly known as MIR.* Kobe, Japan, 2009. 1, 22

M. D. Hoffman, D. M. Blei, and P. R. Cook. Easy as cba: A simple probabilistic model for tagging music. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 369–374, 2009. 22, 90, 108, 109

M. Hollander and D. A. Wolfe. *Nonparametric statistical methods.* Wiley, New York, USA, 2nd edition, 1999. ISBN 0471190454. 81

H. Honing. *Musical cognition: a science of listening.* Transaction Publishers, Piscataway, USA, 2011. ISBN 141284228X. 71, 72, 73, 80, 82

K. J. Hsü and A. J. Hsü. Fractal geometry of music. *Proc Natl Acad Sci USA*, 87(3):938 –941, 1990. 12

K. J. Hsü and A. J. Hsü. Self-similarity of the "1/f noise" called music. *Proc Natl Acad Sci USA*, 88(8):3507 –3509, 1991. 12

D. Huron. *Sweet anticipation : music and the psychology of expectation.* MIT Press, Cambridge Mass., 2006. ISBN 9780262083454. 71, 72, 80, 82

T. Jehan. *Creating Music by Listening.* PhD, MIT, 2005. xxiv, 18, 74, 76, 77

T. Jehan. The echo nest analyze documentation. Technical report, 2010. URL `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`. xxi, 17, 18, 20, 21, 74, 77

Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, January 2010. ISSN 1520-9210, 1941-0077. doi: 10.1109/TMM.2009.2036235. 6, 23, 87

T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nedellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142. Springer, 1998. 24, 94

C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):174–186, 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2007613. 5

P. Juslin and J. A. Sloboda. *Music and emotion: theory and research.* Oxford University Press, Oxford, UK, 2001. ISBN 0192631888. 72

Y. E. Kim, D. S. Williamson, and S. Pilli. Towards quantifying the album-effect in artist classification. In *Proceedings of the International Symposium on Music Information Retrieval*, 2006. 119

A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription.* Springer, 1 edition, 2006. ISBN 0-387-30667-6. 2, 13, 24, 52, 63, 87

P. Knees, T. Pohle, M. Schedl, D. Schnitzer, K. Seyerlehner, and G. Widmer. Augmenting text-based music retrieval with audio similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 579–584, Kobe, Japan, 2009. 89

S. Koelsch, S. Kilches, N. Steinbeis, and S. Schelinski. Effects of unexpected chords and of performer's expression on brain responses and electrodermal activity. *PLoS ONE*, 3(7):e2631, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002631. xxvi, 121, 122, 123, 124

E. M. Kramer and A. E. Lobkovsky. Universal power law in the noise from a crumpled elastic sheet. *Phys Rev E*, 53(2):1465, 1996. doi: 10.1103/PhysRevE.53.1465. 12

C. L. Krumhansl. *Cognitive foundations of musical pitch*. Oxford University Press, Oxford, UK, 1990. ISBN 0195148363. 76

F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*. MIT Press, Cambridge, USA, 1983. ISBN 026262107X. 71

M. Lesaffre. *Music information retrieval : conceptual framework, annotation and user behaviour*. PhD thesis, Ghent University, 2006. 13

D. J. Levitin, P. Chordia, and V. Menon. Musical rhythm spectra from bach to joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences*, February 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1113828109. 12, 71, 72, 80, 82

M. Levy and M. Sandler. Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009. ISSN 1520-9210. doi: 10.1109/TMM.2009.2012913. 89

T. Li, M. Ogihara, and G. Tzanetakis, editors. *Music Data Mining*. CRC Press, 1 edition, July 2011. ISBN 1439835527. 2

Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28(1):84–95, 1980. 22

B. Liu. *Web data mining : exploring hyperlinks, contents, and usage data*. Springer, New York, 2nd edition, 2011. ISBN 9783642194597. 5, 53

R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, 22(9):2390–2416, September 2010. ISSN 1530-888X. doi: 10.1162/NECO_a_00011. PMID: 20569181. 5

J. B. MacQueen. Some methods for classification and analysis of Multi-Variate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967. 22

V. Madisetti. *The digital signal processing handbook*. CRC Press, 1997. ISBN 0849385725. 31, 32

B. D. Malamud. Tails of natural hazards. *Phys World*, 17 (8):31–35, 2004. 10

B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, and R. B. Davis. Zipf's law, music classification, and aesthetics. *Computer Music Journal*, 29:55–69, 2005. ISSN 0148-9267. doi: 10.1162/comj.2005.29.1.55. 12

M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *ISMIR*, pages 594–599, 2005. 24

B. B. Mandelbrot. *The fractal geometry of nature.* W.H. Freeman, San Francisco, 1982. ISBN 0716711869. 11

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing.* The MIT Press, 1 edition, 1999. ISBN 0262133601. 5, 12, 27, 53

C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, New York, 2008. ISBN 9780521865715. 93

G. Marques, M. Lopes, M. Sordo, T. Langlois, and F. Gouyon. Additional evidence that common low-level features of individual audio frames are not representative of music genre. In *7th Sound and Music Computing Conference, 2010, Barcelona, Spain*, pages 134–139, Barcelona, 2010. 4

G. Marques, M. A. Domingues, T. Langlois, and F. Gouyon. Three current issues in music autotagging. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 795–800, Miami (Florida), USA, 2011a. 89, 90, 96, 99

G. Marques, T. Langlois, F. Gouyon, M. Lopes, and M. Sordo. Short-term feature space and music genre classification. *Journal of New Music Research*, 40(2):127–137, 2011b. ISSN 0929-8215. doi: 10.1080/09298215.2011.573563. 4, 5, 22

P. Melidis. Electronic music artist identification. Master thesis, Universitat Pompeu Fabra, 2012. xxx, 119, 120

J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1199644. 72

G. Milner. *Perfecting sound forever: an aural history of recorded music.* Faber and Faber, London, UK, 2009. ISBN 0865479380. 83

M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003. ISSN 1542-7951. 56

E. W. Montroll and M. F. Shlesinger. On 1/f noise and other distributions with long tails. *Proc Natl Acad Sci USA*, 79:3380–3383, 1982. 55

B. C. J. Moore. Loudness, pitch and timbre. In *Blackwell handbook of sensation and perception*. Blackwell Pub., 2005. ISBN 9780631206842. doi: 10.1002/9780470753477.ch13. 15

B. C. J. Moore and B. R. Glasberg. A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82:335–345, 1996. 38

M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben. Multimodal video concept detection via bag of auditory words and multiple kernel learning. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, K. Schoeffmann, B. Merialdo, A. G. Hauptmann, C.-W. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, *Advances in Multimedia Modeling*, volume 7131, pages 40–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27354-4, 978-3-642-27355-1. 23, 88

M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1088 –1110, 2011. ISSN 1932-4553. doi: 10.1109/JSTSP.2011.2112333. xxiv, 4, 15, 28, 29, 30, 63, 72, 76

M. Müller. *Information Retrieval for Music and Motion*. Springer, 2007. ISBN 9783540740483. 2, 13

G. Neukum and B. A. Ivanov. Crater size distributions and impact probabilities on earth from lunar, terrestrial planeta, and asteroid cratering data. In T. Gehrels, editor, *Hazards Due to Comets and Asteroids*, pages 359–416. University of Arizona Press, Tucson, Arizona, 1 edition, 1994. ISBN 0816515050. 10

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323, 2005. doi: 10.1080/00107510500052444. 9, 10, 56

M. E. J. Newman and R. G. Palmer. *Modeling Extinction*. Oxford University Press, USA, 2003. ISBN 0195159454. 56

A. Oceák, I. Winkler, and E. Sussman. Units of sound representation and temporal integration: A mismatch negativity study. *Neurosci Lett*, 436 (1):85 – 89, 2008. ISSN 0304-3940. doi: 10.1016/j.neulet.2008.02.066. 29

A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time signal processing*. Prentice-Hall, Upper Saddle River, USA, 2nd edition, 1999. ISBN 0137549202. 77, 84

N. Orio. Music retrieval: a tutorial and review. *Found. Trends Inf. Retr.*, 1(1):196, 2006. ISSN 1554-0669. doi: 10.1561/1500000002. 1, 2

F. Pachet and P. Roy. Analytical features: A knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:153017, 2009. ISSN 1687-4722. doi: 10.1155/2009/153017. 5

A. D. Patel. *Music, Language, and the Brain*. Oxford University Press, USA, 1 edition, December 2007. ISBN 0195123751. 71

J. Paulus, M. Mller, and A. Klapuri. Audio-based music structure analysis. In *11th International Society for Music Information Retrieval Conference*, pages 625–636, Utrecht, 2010. 19

G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, CUIDADO I.S.T. project at Ircam, 2004. URL `http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf`. 13

O. Peters, A. Deluca, A. Corral, J. D. Neelin, and C. E. Holloway. Universality of rain event size distributions. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(11):P11030, 2010. ISSN 1742-5468. doi: 10.1088/1742-5468/2010/11/P11030. 10

G. J. Peterson, S. Presse, and K. A. Dill. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proc Natl Acad Sci USA*, 107(37):16023–16027, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1010757107. 55

J. Plazak and D. Huron. The first three seconds. *Musicae Scientiae*, 15(1): 29–44, 2011. doi: 10.1177/1029864910391455. 52

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in FORTRAN*. Cambridge University Press, Cambridge, 2 edition, 1992. ISBN 052143064X. 78, 150, 152

T. F. Quatieri. *Discrete-time speech signal processing: principles and practice*. Prentice Hall, 1 edition, 2001. ISBN 9780132429429. 4, 15, 22, 28, 29, 30, 63

L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1 edition, April 1993. ISBN 9780130151575. 16

Z. Raś. *Advances in music information retrieval*. Springer Verlag, Berlin, 2010. ISBN 9783642116735. 30

W. J. Reed and B. D. Hughes. From gene families and genera to incomes and internet file sizes: why power laws are so common in nature. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 66(6 Pt 2): 067103, 2002. ISSN 1539-3755. PMID: 12513446. 10

R. Reynolds. The evolution of sensibility. *Nature*, 434(7031):316–319, 2005. 72

M. Riley, E. Heinen, and J. Ghosh. A text retrieval approach to content-based audio hashing. In *Proceedings of the 9th Conference on Music Information Retrieval (ISMIR)*, pages 295–300, 2008. 22

C. Roads. *Microsound*. Cambridge, Mass. : MIT Press, c2001., 2001. ISBN 0262182157. 71

J. G. Roederer. *The physics and psychophysics of music*. Springer New York, 2009. ISBN 978-0-387-09470-0. 16

A. Saichev, Y. Malevergne, and D. Sornette. Continuous gibrats law and gabaixs derivation of zipfs law. In *Theory of Zipf's Law and Beyond*, volume 632 of *Lecture Notes in Economics and Mathematical Systems*, pages 9–18. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-02945-5. doi: 10.1007/978-3-642-02946-2_2. 55

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513523, 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. 93

N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *Signal Processing Magazine*, 23(2):133–141, 2006. 81

D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer. Local and global scaling reduce hubs in space. *J. Mach. Learn. Res.*, 13(1):28712902, 2012. ISSN 1532-4435. 6

A. Schönberg. *Theory of Harmony*. University of California Press, 1983. ISBN 9780520049444. 72

J. Serrà. *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2011. 18, 19, 20

J. Serrà, E. Gómez, and P. Herrera. Transposing chroma representations to a common key. In *IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects*, pages 45–48, August 2008. ISBN 88-7595-010-5. 75

J. Serrà, E. Gómez, P. Herrera, and X. Serra. Statistical analysis of chroma features in western music predicts human judgments of tonality. *Journal of New Music Research*, 37:299–309, 2008. doi: 10.1080/09298210902894085. 20

J. Serrà, E. Gómez, and P. Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In Z. W. Raś and A. A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, volume 274 of *Studies in Computational Intelligence*, chapter 14, pages 307–332. Springer, Berlin, Germany, 2010. 19, 76

J. Serrà, H. Kantz, X. Serra, and R. Andrzejak. Predictability of music descriptor time series and its application to cover song detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):514–525, 2012a. ISSN 1558-7916. doi: 10.1109/TASL.2011.2162321. 20

J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos. Measuring the evolution of contemporary western popular music. *Scientific Reports*, 2, 2012b. ISSN 2045-2322. doi: 10.1038/srep00521. 71, 85, 125

X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gmez, F. Gouyon, P. Herrera, S. Jorda, O. Paytuvi, G. Peeters, J. Schlter, H. Vinet, and G. Widmer. *Roadmap for Music Information ReSearch*. 1.0.0 edition, 2013. ISBN 978-2-9540351-1-6. 1, 2, 13

J. P. Sethna, K. A. Dahmen, and C. R. Myers. Crackling noise. *Nature*, 410(6825):242–250, 2001. ISSN 0028-0836. doi: 10.1038/35065675. 12

K. Seyerlehner, G. Widmer, and P. Knees. Frame level audio similarity - a codebook approach. In *In: Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pages 349–356, Espoo, Finland, 2008. 22

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42 (3-4):425 –440, 1955. doi: 10.1093/biomet/42.3-4.425. 10, 55, 56, 57

E. P. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149, 2003. ISSN 0959-4388. doi: 10.1016/S0959-4388(03)00047-3. 118

M. Slaney. Auditory toolbox v2. Technical Report 1998-010, 1998. URL `https://engineering.purdue.edu/~malcolm/interval/1998-010/`. 31, 105

K. Sneppen and M. Newman. Coherent noise, scale invariance and intermittency in large systems. *Physica D: Nonlinear Phenomena*, 110(34): 209–222, 1997. ISSN 0167-2789. doi: 10.1016/S0167-2789(97)00128-0. 56

M. Sordo. *Semantic Annotation of Music Collections: A Computational Approach*. PhD thesis, Universitat Pompeu Fabra, 2011. xxix, 89, 90, 91, 94, 107, 108, 109, 111

D. Sornette. *Critical phenomena in natural sciences*. Springer, Berlin, 2nd edition, 2004. ISBN 3540308822. 56

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *J Acoust Soc Am*, 8(3):185–190, 1937. doi: 10.1121/1.1915893. 16, 28, 31, 38

Y. Tang, Y.-Q. Zhang, N. Chawla, and S. Krasser. SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, February 2009. ISSN 1083-4419. doi: 10.1109/TSMCB.2008.2002909. 94

F. Tay and L. Shen. A modified chi2 algorithm for discretization. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):666–670, 2002. ISSN 1041-4347. doi: 10.1109/TKDE.2002.1000349. 22

L. Telesca and M. Lovallo. Analysis of temporal fluctuations in bach's sinfonias. *Physica A: Statistical Mechanics and its Applications*, 391(11): 3247–3256, 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2012.01.036. 12

D. Temperley. *Music and probability*. MIT Press, Cambridge Mass., 2007. ISBN 9780262201667. 71

D. Tingle, Y. Kim, and D. Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the international conference on Multimedia information retrieval*, pages 55–62. ACM, 2010. 106, 121, 147

D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008. ISSN 1558-7916. doi: 10.1109/TASL.2007.913750. 89, 90, 107, 108, 109

D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the CAL500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 439–446, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277817. 88, 105, 147

D. R. Turnbull, L. Barrington, G. Lanckriet, and M. Yazdani. Combining audio content and social context for semantic music discovery. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 387–394, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572009. 89

G. Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Royal Society of London Philosophical Transactions Series B*, 213:21–87, 1925. ISSN 0962-8436. 56

V. Vapnik. *The nature of statistical learning theory*. Springer, New York, 2nd ed. edition, 2000. ISBN 9780387987804. 24

L. Vepstas. An efficient algorithm for accelerating the convergence of oscillatory series, useful for computing the polylogarithm and Hurwitz zeta functions. *Numerical Algorithms*, 47(3):211–252, 2008. 149

R. F. Voss and J. Clarke. 1/f noise in music and speech. *Nature*, 258(5533): 317–318, 1975. doi: 10.1038/258317a0. 12

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer, Berlin, Germany, 2003. ISBN 0387402721. 79

A. Webb. *Statistical Pattern Recognition*. 2nd edition, 2002. ISBN 0470845139. 24, 25

G. A. Wiggins. Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music. In *2009 11th IEEE International Symposium on Multimedia*, pages 477–482, San Diego, California, USA, December 2009. doi: 10.1109/ISM.2009.36. 6

J. C. Willis and G. U. Yule. Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, 109 (2728):177–179, 1922. ISSN 0028-0836. doi: 10.1038/109177a0. 10

B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz. Better speech recognition with cochlear implants. *Nature*, 352(6332):236–238, 1991. doi: 10.1038/352236a0. 24

X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):137, 2007. ISSN 0219-1377. doi: 10.1007/s10115-007-0114-2. 24

J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07, pages 197–206, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-778-0. doi: 10.1145/1290082.1290111. 23

D. H. Zanette. Zipf's law and the creation of musical context. *Musicae Scientiae*, 10(1):3–18, 2006. 12, 83

D. H. Zanette. Playing by numbers. *Nature*, 453(7198):988–989, 2008. ISSN 0028-0836. doi: 10.1038/453988a. 12

M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America*, 126:2390, 2009. ISSN 00014966. doi: 10.1121/1.3238250. 10

G. K. Zipf. *Human behavior and the principle of least effort.* Addison-Wesley, 1949. ISBN 49007787. xxi, 9, 10, 11, 12, 55, 56, 83

E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *J Acoust Soc Am*, 33(2):248, 1961. ISSN 00014966. doi: 10.1121/1.1908630. 16, 28, 30, 31

E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J Acoust Soc Am*, 68 (5):1523, 1980. ISSN 00014966. doi: 10.1121/1.385079. 16

Appendix **A**

Used databases

A.1 The Million Song Dataset

The *Million Song Dataset* (Bertin-Mahieux et al., 2011) is a publicly available dataset of audio descriptors and metadata for "a million contemporary popular music tracks"[1]. This dataset was made available by Columbia University's LabROSA[2] and the company The Echo Nest[3].

The dataset contains tracks from 44,745 unique artists from a variety of popular Western music genres such as rock, pop, hip-hop, electronic, jazz, or folk. It also includes metadata information like the name of the song, the artist and, in the case of 515,576 tracks, it also includes the release year (from 1922 to 2010).

The audio features for each track were computed by The Echo Nest Analyze API and include descriptors of timbral, tonal, loudness, and rhythmic content. Furthermore, the provided audio features were computed at the segment level (i.e. small audio excerpts that mainly correspond with note onsets), but the rhythmic information can be also used to obtain, for instance, beat-synchronous features.

[1] http://labrosa.ee.columbia.edu/millionsong

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

[3] http://the.echonest.com

A.2 Speech database

The *Speech* dataset is an in-house collection of 130 hours of English speaker audio files gathered from the following places:

- 5.48 hours from the TIMIT dataset (Garofolo et al., 1993), which is a corpus of phonemically and lexically transcribed speech of English speakers.
- 5.11 hours from the *Library of Congress* "Music and the brain" podcasts[4], which corresponds to a set of 15 interviews made to scientists related with the music and cognition fields.
- 119.43 hours from Nature podcasts[5] from 2005 to April 7th 2011. These podcasts are audio shows that feature highlighted content about Nature published works. In order to skip music content from the opening and closing of the audio show, we have removed the first and last 2 minutes of sound in every file.

A.3 Western Music database

The *Western Music* database is an in-house database built with approximately 282 hours of Western music extracted from commercial CDs. This collection has a total of 3,481 full tracks accounting for more than 20 musical genres including: rock, pop, jazz, blues, electronic, classical, hip-hop, and soul. A subset of this collection (2,720 tracks) was previously used in Gómez et al. (2009).

A.4 Non-Western Music database

The *non-Western Music* database contains 280 hours (3,249 recordings) of traditional music distributed by geographical regions, as defined by UNESCO[6] that corresponds to countries from the Pacific, Central Asia, Asia, Arab States and Africa. The recordings were extracted from commercial CD

[4] http://www.loc.gov/podcasts/musicandthebrain/index.html

[5] http://www.nature.com/nature/podcast/archive.html

[6] http://portal.unesco.org/geography

collections used for ethnomusicological studies (field recordings and compilations of traditional music) discarding those having some Western influence (e.g. equal-tempered instruments). Once more, a great part of this collection (3,185 tracks) was previously used in Gómez et al. (2009).

A.5 Sounds of the Elements database

The *Sounds of the Elements* is an in-house database that contains 47.8 hours of sounds of natural inanimate phenomena such as water (rain, water streams, waves, melting snow, waterfalls, etc), fire, thunders, wind and earth sounds (rocks, rumbles, volcanic eruptions, etc.). This database consists of 1,141 files manually gathered from *The Freesound Project*[7] and labeled by the site's users as "field-recording".

A.6 The CAL500 database

The *Computer Audition Lab 500-Song*[8] (CAL500) dataset (Turnbull et al., 2007) is a public dataset that consists of 500 musical tracks from 500 artists annotated by 3 non-expert undergraduate students using 174 tags related to genre, mood, instrumentation, solo instrument, music usage, and vocal characteristics. The database's tags correspond to those tags annotated by at least three human annotators (see Appendix E for a list of possible tags).

A.7 The CAL10k database

The CAL10k database (Tingle et al., 2010) contains 10,870 songs that were weakly-labeled (i.e. the absence of a tag does not mean that the tag does not apply to the song) with 475 acoustic tags and 153 genre tags. These tags were gathered from Pandora's[9] website and, given the characteristics of the site's annotations, the tags were annotated by expert musicologists involved with the Music Genome Project.

[7] `http://www.freesound.org`

[8] `http://cosmal.ucsd.edu/cal/projects/AnnRet/`

[9] `http://www.pandora.com`

Appendix B

Distribution functions

As stated in the main text of this thesis, three different types of heavy-tailed distributions are reported: discrete (pure) power-laws, shifted discrete power-laws, and truncated reversed log-normals. For the discrete cases, the random variable takes only integer values, which represent, in our case, the frequency of a codeword. Then, $P(z)$ is the probability mass function, and gives the probability that the random variable takes the value z.

For the discrete power-law, and discrete shifted power-law $P(z)$ is given by

$$P(z) = \frac{1}{\zeta(\beta, c + z_{min})(c+z)^{\beta}} \tag{B.1}$$

with $z = z_{\mathrm{min}}, z_{\mathrm{min}}+1, \dots$, where c and β are parameters ($\beta \geq 1$), and z_{min} is the minimum value of the variable for which the fit holds. We note that z_{min} takes integer values and that fulfills $c + z_{min} > 0$. The discrete (pure) power-law case is recovered by setting $c = 0$.

The bivariate function $\zeta(\beta, q)$ is the Hurwitz zeta function,

$$\zeta(\beta, q) = \sum_{n=0}^{\infty} \frac{1}{(q+n)^{\beta}}, \tag{B.2}$$

which yields the Riemann zeta function for $q = 1$, i.e. $\zeta(\beta, 1) = \zeta(\beta)$. At several points the fitting procedure will require the computation of the Hurwitz zeta function, which is done by means of an algorithm based on the Euler-Maclaurin series (Vepstas, 2008).

For the MSD-loudness distribution values of Chapter 6 (denoted by x), z is a real variable, defined as $z = -x$, as well as its minimum and maximum values $z_{\rm min}$ and $z_{\rm max}$. Although we use the same notation as for discrete variables, for a continuous variable the function $P(z)$ will not be the probability mass function but the probability density, given in this case by a truncated log-normal,

$$P(z) = \sqrt{\frac{2}{\pi\sigma^2}} \left[\mathrm{erf}\left(\frac{\ln z_{\rm max} - \mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{\ln z_{\rm min} - \mu}{\sqrt{2}\sigma}\right)\right]^{-1} \dots \\ \frac{1}{z} \exp\left(-\frac{(\ln z - \mu)^2}{2\sigma^2}\right) \tag{B.3}$$

with $0 \leq z_{\rm min} \leq z \leq z_{\rm max}$ and where

$$\mathrm{erf}(y) = 2\pi^{-1/2} \int_0^y e^{-u^2} du \tag{B.4}$$

is the error function (implemented as in Press et al. (1992)). The adjective 'reverse' used in the main text refers to the fact that $P(x)$ is the mirror image of the true (truncated) log-normal distribution for the variable $z = -x$. Note that μ and σ do not correspond to the mean and standard deviation of the data, but to those of the underlying non-truncated normal distribution.

Appendix C

Power-law fit

As mentioned in the main text, to visualize a straight-line in the log-log plot it is not sufficient condition to claim that the observed data is well fitted by a power-law. Moreover, commonly used methods for data analysis, such as least-squares fitting, are also prone to errors when trying to evaluate if a power-law fits our data. Hence, following the recommendations made by Clauset et al. (2009) we use maximum likelihood (ML) estimation to perform the fits. In particular, we evaluate if frequency distributions are well fitted by power-laws. The reason to work with frequency distributions is that the frequency can be considered as a random variable, whereas the rank is not.

For a continuous random variable Z, following a power-law distribution given by the probability density,

$$f(z) = \frac{\beta - 1}{1 - j^{\beta-1}} \frac{a^{\beta-1}}{z^{\beta}} \propto \frac{1}{z^{\beta}},$$

defined in the range $a \leq Z \leq b$ and with $j = a/b$ (note that a and b correspond to $z_{\min}$ and $z_{\max}$ in the main text), the ML estimator of the exponent β is given by the maximization of the log-likelihood as a function of β,

$$\frac{\ln L}{N_{ab}} = \ln(\beta - 1) - \ln(1 - j^{\beta-1}) - (\beta - 1) \ln \frac{G_{ab}}{a} - \ln G_{ab},$$

where N_{ab} is the number of code-word types comprised between a and b and G_{ab} is the geometric mean of the frequencies on that range.

For a discrete random variable Z taking values $a, a+1, \dots b-1, b$, it is easy to show, following Clauset et al. (2009), that the log-likelihood in the power-law case is well approximated, in the limit of large a, by

$$\frac{\ln L}{N_{ab}} = \ln(\beta-1) - \ln(1-k^{\beta-1}) - (\beta-1)\ln\frac{G_{ab}}{a-1/2} - \ln G_{ab},$$

where $k = (a-1/2)/(b+1/2)$. This means that a discrete power-law distribution between a and b can be replaced by a continuous one between $a-1/2$ and $b+1/2$ if a is large enough, which in practice is usually achieved by $a \geq 5$ (Clauset et al., 2009) if b is much larger than a. For a power-law with no upper limit ($b \to \infty$), the previous formula is still valid just taking $k=0$ and therefore a closed formula can be obtained for β, which is given by $\beta = 1 + 1/[\ln G_{ab} - \ln(a-1/2)]$.

For the error ε of the exponent β, we approximate the formula for the continuous case,

$$\sqrt{N_{ab}}\,\varepsilon = \left[\frac{1}{(\beta-1)^2} - \frac{k^{\beta-1}\ln^2 k}{(1-k^{\beta-1})^2}\right]^{-1/2},$$

which corresponds to one standard deviation of the distribution of β when N_{ab} is large. For $b \to \infty$, the limit $k=0$ yields $\varepsilon = (\beta-1)/\sqrt{N_{ab}}$.

A maximization of the likelihood does not guarantee a good fit if the probabilistic (power-law) model is not appropriate. Thus, it is necessary then to test the goodness of the fit. In the same way as Clauset et al. (2009) (and this choice is a matter of taste) we use the Kolmogorov-Smirnov (KS) test (Press et al., 1992). This is defined by the KS statistic, or KS distance, which is the maximum difference between the empirical cumulative distribution and the theoretical cumulative distribution corresponding to the ML fit, i.e.,

$$d_{KS} = \max_{\forall z_i}\left[S(z_i) - \frac{i}{N_{ab}}\right],$$

where i denotes the number of data equal or above z_i, z_i corresponds to a value taken by the variable Z, and $S(z_i)$, the survivor function of Z, is well approximated (for large a) by

$$S(z) = \frac{1}{1-k^{\beta-1}}\left[\left(\frac{a-1/2}{z-1/2}\right)^{\beta-1} - k^{\beta-1}\right].$$

The value of the KS distance does not suffice to characterize the fit as good or bad, thus, we need a scale in order to compare it. This scale is obtained by computer simulations of the resulting fitted ML power-law distribution, approximated, for reasonably large a, by

$$z = \left\lfloor \frac{a - 1/2}{[1 - u(1 - k^{\beta-1})]^{1/(\beta-1)}} + \frac{1}{2} \right\rfloor,$$

where u is a continuous uniform random number between 0 and 1 and $\lfloor ... \rfloor$ denotes the integer part of its argument (which therefore, with the term $+1/2$ inside, rounds the other term to the nearest integer). For a large number of synthetic data sets, with N_{ab} elements each, the same procedure as for the empirical data is repeated: ML estimation of the β exponent plus the calculation of the KS distance between each synthetic distribution and its fit. In this way, a distribution of KS distances is obtained under the null hypothesis that the data come from a power-law distribution. The $p-$value is then defined as the probability that for true power-law distributed data, as the synthetic sets we have generated, the KS distance is above the empirical value; this is computed as the number of synthetic data sets for which their KS distance is larger than the empirical one divided by the total number of synthetic data sets.

In principle, for fixed values of a and b, we obtain the ML value of the exponent β and an associated $p-$value. In practice, however, a and b are not known, and one needs a criterion to select the optimum ones. At this point we depart from the recipe provided by Clauset et al. (2009) since that algorithm was shown to reject the power-law hypothesis for power-law simulated data in some specific cases (Corral et al., 2011). We repeat the previous procedure for many different values of a and b and select the ones which maximize the log-range of the data, b/a, provided that the corresponding $p-$value is high enough. We usually use a threshold value equal to 20%. It is important to realize that the $p-$value of the whole procedure is not the one corresponding to the selected values of a and b. Computer simulations tell us that the former is a factor 2 or 3 smaller than the latter. Nevertheless, the precise calculation of the $p-$value is not relevant for our purposes.

It turns out that for the data analyzed in this thesis the resulting values of b are always larger than the maximum value taken by the variable (i.e. no data are outside the power-law range from the right side) and therefore it is simpler to assume $k = 0$ in the previous formulas and just work with a non-upper truncated power-law.

Additional details regarding the fitting procedure can be found in the supplementary information of Corral et al. (2010).

Appendix D

Co-occurrence tables for Bark-band code-words

Tables D.1, D.2, D.3 and D.4 show the number of co-occurring Bark-band code-words as obtained with the 186 ms frame. These tables account for co-occurrence of code-words that describe 20, 50, 80 and 100% of each database, respectively.

20%			Music-nW	¬(Music-nW)
Music-W	Speech	Elements	0	0
		¬(Elements)	34	1
	¬(Speech)	Elements	5	0
		¬(Elements)	175	91
¬(Music-W)	Speech	Elements	0	0
		¬(Elements)	18	16
	¬(Speech)	Elements	1	0
		¬(Elements)	323	—

Table D.1: Co-occurrence of code-words that account for 20% of each database (frame size = 186 ms). The symbol ¬() denotes the negation of the proposition inside the parentheses, e.g. ¬(Speech) stands for timbral code-words that do not belong to the Speech database.

50%			Music-nW	¬(Music-nW)
Music-W	Speech	Elements	33	0
		¬(Elements)	523	6
	¬(Speech)	Elements	9	0
		¬(Elements)	4,434	2,253
¬(Music-W)	Speech	Elements	1	0
		¬(Elements)	328	154
	¬(Speech)	Elements	0	0
		¬(Elements)	11,784	—

Table D.2: Co-occurrence of code-words that account for 50% of each database (frame size = 186 ms).

80%			Music-nW	¬(Music-nW)
Music-W	Speech	Elements	438	1
		¬(Elements)	8,906	280
	¬(Speech)	Elements	68	0
		¬(Elements)	60,471	41,477
¬(Music-W)	Speech	Elements	1	0
		¬(Elements)	2,992	847
	¬(Speech)	Elements	0	0
		¬(Elements)	121,154	—

Table D.3: Co-occurrence of code-words that account for 80% of each database (frame size = 186 ms).

100%			Music-nW	¬(Music-nW)
Music-W	Speech	Elements	20,495	741
		¬(Elements)	93,674	21,912
	¬(Speech)	Elements	8,344	1,729
		¬(Elements)	291,497	360,478
¬(Music-W)	Speech	Elements	363	166
		¬(Elements)	38,101	44,142
	¬(Speech)	Elements	1,055	1,277
		¬(Elements)	493,797	—

Table D.4: Co-occurrence of code-words that account for 100% of each database (frame size = 186 ms).

Appendix E

Autotagging experiments

E.1 Experiment I: detailed results

The list of stemming tags is shown in Table E.1 (the first tag in every line is the one used as final tag).

Table E.2 shows the final list of selected tags and their frequencies within the MSD-Tag dataset.

Table E.3 shows the per-tag F-measure results for the MSD-Tag dataset. The table presents results from both BoF and BoC-W approaches using the three selected audio features namely MSD-Timbre, MSD-Pitch, and MSD-Loudness.

Fig. E.1 shows per-tag F-measure results for the BoC-W approach using the MSD-Timbre descriptor vs. the total tag frequency of the MSD-Tag dataset. As can be seen in the figure, there is no strong correlation between both variables.

E.2 Experiment II: detailed results

Tables E.4, E.5, and E.6 show the original CAL500 tag-categories, tag names and tag frequencies. Moreover, the tables show per-tag F-measure classification results for the BoC-W MFCC, and BoF MFCC approaches.

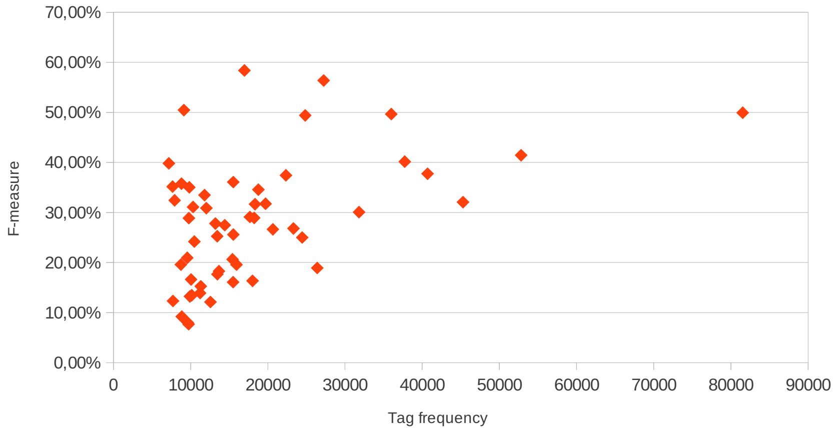

Figure E.1: F-measure results for the BoC-W approach using MSD-Timbre (y-axis, in percentage) vs. total tag frequency (x-axis) for the MSD-Tag dataset.

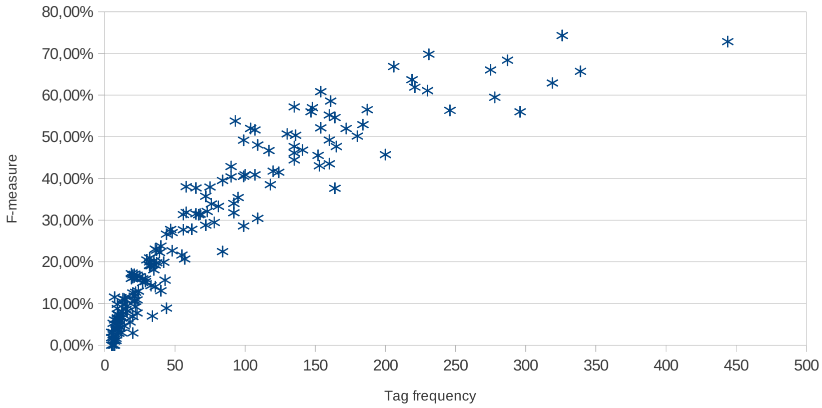

Figure E.2: F-measure results for the BoC-W approach using MFCCs (y-axis, in percentage) vs. total tag frequency (x-axis) for the CAL500 dataset.

Fig. E.2 shows per-tag F-measure results for the BoC-W approach using the MFCC descriptor vs. the total tag frequency of the CAL500 dataset. As can be seen in the figure, there is a strong correlation between both variables.

rock	80s rock , 70s rock , 60s rock , 2000s rock , 00s rock.
pop	My pop music , 80s Pop , 60s pop , Pop Music , pop singles , 90s pop , general pop , 70s pop , Good Pop , 00s pop , pop favorites.
alternative	00s alternative , Alternative In The 2000s , Alternative In The 1990s.
indie	extraordinary indie , favorite indie , mainstream-indie , indie hits.
electronic	electronic music , electronic top , -electronic- , greatest electronic.
female vocalists	female vocalist , female vocals , female vocal , Female Voices , female singers , female voice , female singer-songwriter , femalesinger , female singer , vocals female , magic female voice , sexy female vocals , female singer songwriter , female vocal group , female singer-songwriters , female lead singer , female-vocalists , female vocalist , f singer-songwriter.
dance	dance dance dance , 80dance , dance music , 90s dance , dance dance , dancemusic , dance top , dance musik , Dance 90s.
alternative rock	alternative-rock , rock alternative.
jazz	pure jazz , Jazzz , Great Jazz , greatest jazz.
singer-songwriter	singer songwriter , singersongwriter , singer-songwriters , Singer/Songwriter.
metal	heavy metal , classic metal , 80s metal , true metal , classic heavy metal , 80s Heavy Metal , metal top , 90s metal , 90s heavy metal , traditional heavy metal , 70s heavy metal , Metal Gods , Metal songs , heavy.
chillout	chill , chilled , chill-out , chill music , chilly , Chilled Out.
male vocalists	male vocalist , m singer-songwriter , male vocals , malesinger , male vocal , vocals male , Male Singers , magic male voice , male voice , male-vocalist , male singer songwriter , a distinctive male lead vocal , malevoice , male singer.
classic rock	rock classics , classic rock favorites , Classic Rock , classicrock , the best of classic rock.
soul	soul tag , 80s soul , favouritesoul , 70s soul , soulsongs , 90s soul , 60s soul , the very best of soul , soul music.
indie rock	Indie-Rock , indierock , indie rock favs.
instrumental	instrumentals , Instrumental music.
punk	Punk Favorites , top punk songs , 1970s-punk , 1970s punk , 80s punk , 70s punk.
oldies	golden oldies , Oldies Tag , oldie , oldiess , oldies but goldies , oldies favorites , golden oldie.
blues	Blues Tag , Blues Blues Blues.
hard rock	80s hard rock , 90s hard rock , hardrock , 70s hard rock , 60s hard rock.
guitar	guitar virtuoso , acoustic guitar , Guitar Solo , Guitar Gods , electric guitar , guitar riffs and solos , guitar god , an electric guitar solo , guitars , instrumental guitar , great guitar solo , 100 Greatest Guitar Solos.
Hip-Hop	hip hop , hiphop , True Hip Hop , Real hip-hop , hip hop tag , real hip hop , hip-hop favorites , greatest hip hop.
party	party music , it is party time , party time , party songs , Party Mix.
country	country legends , country music , 90s country , Real Country , Country Favorites , Country Songs.
funk	funky , Funk Tag , favouritefunk.
Progressive rock	70s progressive rock , prog rock.
rnb	r&b , rythm and blues , r-n-b , r & b , rhythm and blues , rhythum and blues tag , rhythm-blues , rhythm & blues.
indie pop	indiepop , Fave Indie Pop , indie-pop , indie pop favs.
Soundtrack	Soundtracks , movie soundtrack , movie soundtracks.
sad	sad songs , sad song , sadness , so sad , Mood: Sad.
House	house music.
happy	makes me happy , songs that make me happy , Happy Music , get happy , Happy songs , Make you happy , happy happy.
punk rock	punkrock , Punk-Rock.
piano	solo piano , piano solo.
psychedelic	psych , psy , psychadelic , psychodelic , psyhdelic.
pop rock	RockPop , Pop/Rock , Rock Pop , pop - rock , rock-pop , Rock/Pop , Rock Pop , Pop-Rock , poprock.
downtempo	down tempo.
trance	favorite trance , Anthem Trance.
melancholy	melancholic , melancholia , -melancholic- , melancholie , Melancholisch , melancoholic , melancolia , melancolic.
techno	Tecno.
relax	relaxing , relaxed , relaxing mood , relaxation.
new wave	80s New Wave , newwave.

Table E.1: Stemming Tags. The first column shows the selected tag name.

		Tag frequency		
Tag	**F-measure**	**Train**	**Test**	**Total**
hip-hop	58.70	14,694	2,282	16,976
metal	56.90	23,948	3,292	27,240
rap	50.49	7,918	1,220	9,138
electronic	49.71	31,266	4,734	36,000
jazz	49.51	21,264	3,585	24,849
rock	48.85	73,233	8,284	81,517
pop	41.31	47,031	5,790	52,821
female vocalists	40.42	33,388	4,344	37,732
trance	39.93	6,208	972	7,180
indie	37.70	36,105	4,588	40,693
dance	37.59	19,565	2,781	22,346
ambient	36.18	13,446	2,079	15,525
reggae	35.93	7,906	918	8,824
hardcore	35.24	8,700	1,136	9,836
techno	35.23	6,579	1,089	7,668
instrumental	34.51	16,199	2,579	18,778
country	33.41	10,305	1,494	11,799
house	32.44	6,905	1,021	7,926
alternative	31.82	40,751	4,533	45,284
electronica	31.77	16,040	2,306	18,346
electro	31.18	8,861	1,458	10,319
rnb	30.98	10,596	1,443	12,039
chillout	30.07	28,083	3,734	31,817
punk	29.12	16,159	1,516	17,675
soul	28.93	16,177	2,053	18,230
piano	28.89	8,576	1,204	9,780
oldies	27.69	11,634	1,581	13,215
acoustic	27.50	12,935	1,483	14,418
singer-songwriter	26.85	21,210	2,113	23,323
indie rock	26.67	18,442	2,215	20,657
blues	25.56	14,087	1,448	15,535
funk	25.31	11,857	1,591	13,448
alternative rock	24.97	22,052	2,407	24,459
punk rock	24.23	9,528	962	10,490
downtempo	20.95	8,459	1,102	9,561
experimental	20.63	13,566	1,861	15,427
lounge	19.65	7,669	1,080	8,749
hard rock	19.55	14,731	1,219	15,950
male vocalists	18.64	23,602	2,800	26,402
relax	18.32	12,000	1,674	13,674
melancholy	17.82	12,022	1,449	13,471
classic rock	16.62	16,644	1,378	18,022
psychedelic	16.53	8,909	1,151	10,060
guitar	16.20	14,048	1,458	15,506
indie pop	15.25	10,062	1,257	11,319
progressive rock	13.86	9,638	1,593	11,231
sad	13.50	9,062	1,084	10,146
party	13.22	8,745	1,166	9,911
new wave	12.32	6,994	721	7,715
pop rock	12.01	11,295	1,279	12,574
folk	11.71	17,402	2,306	19,708
fun	9.32	7,964	897	8,861
soundtrack	7.87	8,773	849	9,622
happy	7.75	8,797	948	9,745
Total		882,030	111,507	993,537

Table E.2: Selected Tags and their frequencies for the train and test subsets; also F-measure results (in percentage) for BoC-W (MDS-Timbre with TF-IDF weighting) are shown.

	MSD-Timbre		MSD-Pitch		MSD-Loudness	
Tag	**BoF**	**BoC-W**	**BoF**	**BoC-W**	**BoF**	**BoC-W**
acoustic	25.99	27.50	0.13	16.80	11.60	11.15
alternative	29.73	31.82	10.58	29.15	26.94	25.77
alternative rock	8.68	24.97	1.77	20.35	6.90	19.03
ambient	28.16	36.18	28.74	21.43	18.74	18.06
blues	11.73	25.56	25.90	16.64	10.21	10.94
chillout	17.40	30.07	11.44	25.62	12.50	23.88
classic rock	12.06	16.62	0.22	13.03	5.67	8.89
country	26.46	33.41	9.46	20.71	6.20	10.15
dance	29.73	37.59	22.76	28.01	12.96	17.73
downtempo	3.59	20.95	0.98	10.62	2.07	9.09
electro	16.99	31.18	7.98	20.34	8.26	11.44
electronic	19.88	49.71	20.20	40.56	7.74	28.46
electronica	4.27	31.77	5.01	23.94	2.03	15.54
experimental	7.63	20.63	4.96	14.55	4.69	12.07
female vocalists	21.91	40.42	9.58	26.24	8.93	15.91
folk	11.54	11.71	4.36	24.35	6.33	16.57
fun	4.41	9.32	0.00	6.97	1.53	6.18
funk	12.21	25.31	1.49	13.94	6.71	12.06
guitar	4.78	16.20	0.00	10.80	3.08	8.57
happy	1.79	7.75	0.00	7.35	0.92	6.09
hard rock	8.93	19.55	8.46	15.64	5.99	11.50
hardcore	22.45	35.24	22.46	21.12	12.38	15.54
hip-hop	39.55	58.70	30.39	30.86	16.79	17.53
house	8.88	32.44	9.25	19.91	4.08	9.51
indie	11.01	37.70	11.93	31.75	8.53	25.60
indie pop	2.11	15.25	0.74	11.60	0.87	8.36
indie rock	3.14	26.67	1.80	18.25	2.04	16.74
instrumental	12.89	34.51	9.08	21.43	8.34	19.86
jazz	25.54	49.51	37.00	38.90	16.97	29.80
lounge	1.01	19.65	0.00	10.52	0.70	9.03
male vocalists	6.02	18.64	0.00	17.63	2.71	11.90
melancholy	1.85	17.82	3.00	15.83	1.00	9.93
metal	30.51	56.90	36.49	42.68	20.11	34.10
new wave	3.80	12.32	0.00	6.12	0.00	4.36
oldies	9.78	27.69	7.36	18.77	3.04	12.15
party	1.36	13.22	5.63	10.19	0.82	8.08
piano	4.15	28.89	10.56	14.46	2.30	12.62
pop	10.92	41.31	9.32	35.06	6.10	26.41
pop rock	0.79	12.01	0.00	10.79	0.36	8.69
progressive rock	6.69	13.86	6.42	13.15	4.43	9.27
psychedelic	4.41	16.53	2.74	9.80	1.84	7.08
punk	9.40	29.12	12.09	19.95	5.98	16.45
punk rock	4.32	24.23	6.09	14.56	2.61	11.85
rap	9.33	50.49	6.22	20.23	2.57	10.65
reggae	24.00	35.93	26.98	15.00	8.61	7.53
relax	0.47	18.32	0.00	15.08	0.29	12.65
rnb	4.85	30.98	1.61	12.86	2.89	10.29
rock	9.31	48.85	9.55	45.67	7.36	38.48
sad	0.54	13.50	0.00	12.27	0.21	7.19
singer-songwriter	2.04	26.85	2.13	19.11	1.11	12.06
soul	4.81	28.93	3.61	17.20	1.90	14.61
soundtrack	4.19	7.87	0.00	6.08	1.80	6.83
techno	8.44	35.23	12.08	21.70	4.25	9.93
trance	9.78	39.93	14.63	22.98	0.00	8.96
Mean F	11.23	27.91	8.76	19.42	5.98	14.13
Global F	12.70	31.13	10.00	20.03	7.42	14.28

Table E.3: Per-tag autotagging results (F-measure in percentage) for best BoF and BoC-W approaches for MSD-Timbre, MSD-Pitch, and MSD-Loudness.

Category	Tag	Tag Frequency	BoC-W MFCC	BoF MFCC
Emotion-Angry	Aggressive	48	27.02	35.09
NOT-Emotion-Angry	Aggressive	319	62.92	72.46
Emotion-Arousing	Awakening	154	52.15	46.84
NOT-Emotion-Arousing	Awakening	99	49.18	49.87
Emotion-Bizarre	Weird	22	9.18	4.32
NOT-Emotion-Bizarre	Weird	296	56.00	59.83
Emotion-Calming	Soothing	148	56.98	56.96
NOT-Emotion-Calming	Soothing	154	60.84	55.91
Emotion-Carefree	Lighthearted	109	30.43	32.45
NOT-Emotion-Carefree	Lighthearted	152	45.54	39.39
Emotion-Cheerful	Festive	107	40.89	37.16
NOT-Emotion-Cheerful	Festive	180	50.17	51.33
Emotion-Emotional	Passionate	160	43.54	39.84
NOT-Emotion-Emotional	Passionate	73	32.09	29.67
Emotion-Exciting	Thrilling	117	46.70	41.93
NOT-Emotion-Exciting	Thrilling	147	56.02	56.06
Emotion	Happy	135	44.45	41.13
NOT-Emotion-Happy	Happy	135	46.15	43.11
Emotion-Laid-back	Mellow	109	48.04	50.20
NOT-Emotion-Laid-back	Mellow	161	58.57	52.97
Emotion-Light	Playful	92	31.77	25.76
NOT-Emotion-Light	Playful	187	56.52	50.94
Emotion-Loving	Romantic	76	33.89	28.08
NOT-Emotion-Loving	Romantic	230	61.09	56.92
Emotion-Pleasant	Comfortable	184	52.92	48.55
NOT-Emotion-Pleasant	Comfortable	67	31.30	35.42
Emotion-Positive	Optimistic	120	41.75	39.20
NOT-Emotion-Positive	Optimistic	118	38.53	38.09
Emotion-Powerful	Strong	160	49.23	43.40
NOT-Emotion-Powerful	Strong	65	31.50	19.95
Emotion	Sad	58	31.86	28.67
NOT-Emotion	Sad	221	61.91	56.46
Emotion-Tender	Soft	104	52.00	50.00
NOT-Emotion-Tender	Soft	206	66.86	59.18
Emotion-Touching	Loving	75	37.94	33.28
NOT-Emotion-Touching	Loving	219	63.70	56.32
Genre	Alternative	100	40.89	38.28
Genre	Alternative-Folk	8	3.92	0.00
Genre	Bebop	6	1.11	0.00
Genre	Brit-Pop	9	6.41	2.86
Genre	Classic-Rock	90	40.40	35.78
Genre	Contemporary-Blues	7	0.00	0.00
Genre	Contemporary-R&B	16	8.21	0.00
Genre	Cool-Jazz	13	9.68	10.19
Genre	Country-Blues	6	0.00	0.00
Genre	Dance-Pop	21	10.45	4.72
Genre	Electric-Blues	9	9.80	0.00
Genre	Funk	11	5.65	5.00
Genre	Gospel	7	11.54	0.00
Genre	Metal/Hard-Rock	32	19.17	26.18
Genre	Punk	23	16.81	20.05
Genre	Roots-Rock	8	1.11	0.00
Genre	Singer-Songwriter	25	16.47	11.95
Genre	Soft-Rock	48	22.67	26.44
Genre	Soul	29	15.26	4.58
Genre	Swing	5	3.10	0.00
Genre	Bluegrass	10	7.25	0.00
Genre	Blues	24	12.95	4.22
Genre	Country	33	19.04	18.67
Genre	Electronica	56	27.70	37.74
Genre	Folk	30	20.47	17.62
Genre	Hip-Hop/Rap	21	17.01	61.86
Genre	Jazz	32	19.96	19.10
Genre	Pop	72	35.73	34.09
Genre	R&B	36	19.21	10.27
Genre	Rock	136	50.36	52.68

Table E.4: CAL500 Tag-category, Tag name, tag frequencies, F-measure results for BoC-W MFCC, and F-measure results for BoF MFCC (part I).

Category	Tag	Tag Frequency	BoC-W MFCC	BoF MFCC
Genre	World	21	11.79	9.00
Instrument	Acoustic-Guitar	58	38.02	39.51
Instrument	Ambient-Sounds	32	20.96	14.95
Instrument	Backing-vocals	153	43.05	47.46
Instrument	Bass	164	54.63	46.75
Instrument	Drum-Machine	44	26.63	24.32
Instrument	Drum-Set	275	66.05	73.10
Instrument	Electric-Guitar-(clean)	124	41.49	36.63
Instrument	Electric-Guitar-(distorted)	65	37.72	37.51
Instrument	Female-Lead-Vocals	90	42.87	55.28
Instrument	Hand-Drums	12	6.48	0.00
Instrument	Harmonica	10	6.22	9.00
Instrument	Horn-Section	18	5.58	2.00
Instrument	Male-Lead-Vocals	339	65.71	85.24
Instrument	Organ	6	5.24	0.00
Instrument	Piano	84	39.51	31.12
Instrument	Samples	32	19.09	18.57
Instrument	Saxophone	23	7.71	6.54
Instrument	Sequencer	39	22.24	20.42
Instrument	String-Ensemble	20	16.22	21.86
Instrument	Synthesizer	99	40.52	35.38
Instrument	Tambourine	12	2.97	12.86
Instrument	Trombone	8	4.26	0.00
Instrument	Trumpet	21	16.35	4.04
Instrument	Violin/Fiddle	9	7.48	0.00
Song	Catchy/Memorable	165	47.66	49.28
NOT-Song	Catchy/Memorable	81	33.31	24.81
Song	Changing-Energy-Level	36	14.04	6.95
NOT-Song	Changing-Energy-Level	200	45.73	46.77
Song	Fast-Tempo	135	47.67	45.81
NOT-Song	Fast-Tempo	135	57.18	57.80
Song	Heav-Beat	130	50.70	42.19
NOT-Song	Heavy-Beat	107	51.65	53.33
Song	High-Energy	231	69.81	67.17
NOT-Song	High-Energy	93	53.80	49.56
Song	Like	164	37.66	35.27
NOT-Song	Like	55	21.62	18.79
Song	Positive-Feelings	172	51.99	43.60
NOT-Song	Positive-Feelings	78	29.45	30.82
Song	Quality	287	68.38	59.98
NOT-Song	Quality	13	11.16	13.33
Song	Recommend	84	22.46	18.92
NOT-Song	Recommend	99	28.59	27.67
Song	Recorded	444	72.85	81.73
NOT-Song	Recorded	9	8.79	6.67
Song	Texture-Acoustic	278	59.48	69.63
Song	Texture-Electric	326	74.31	75.81
Song	Texture-Synthesized	160	55.23	47.07
Song	Tonality	92	34.01	30.51
NOT-Song	Tonality	44	8.87	9.29
Song	Very-Danceable	68	31.46	30.97
NOT-Song	Very-Danceable	246	56.34	60.45
Usage	At-a-party	62	27.82	36.12
Usage	At-work	14	4.58	0.00
Usage	Cleaning-the-house	43	15.61	8.44
Usage	Driving	141	46.82	38.79
Usage	Exercising	22	12.70	13.83
Usage	Getting-ready-to-go-out	29	15.86	8.53
Usage	Going-to-sleep	56	31.34	32.82
Usage	Hanging-with-friends	34	6.99	14.75
Usage	Intensely-Listening	20	2.92	7.86
Usage	Reading	20	12.54	12.73
Usage	Romancing	19	17.16	5.56
Usage	Sleeping	8	5.02	13.33
Usage	Studying	33	14.28	13.33
Usage	Waking-up	8	1.63	0.00
Usage	With-the-family	5	0.00	0.00

Table E.5: CAL500 Tag-category, Tag name, tag frequencies, F-measure results for BoC-W MFCC, and F-measure results for BoF MFCC (part II).

Category	Tag	Tag Frequency	BoC-W MFCC	BoF MFCC
Vocals	Aggressive	37	22.81	31.65
Vocals	Altered-with-Effects	35	18.08	10.30
Vocals	Breathy	20	6.85	4.68
Vocals	Call-Response	15	8.62	13.33
Vocals	Duet	6	3.06	0.00
Vocals	Emotional	95	35.42	34.91
Vocals	Falsetto	11	4.25	2.50
Vocals	Gravelly	12	8.04	9.00
Vocals	High-pitched	35	20.23	8.61
Vocals	Low-pitched	27	15.02	15.89
Vocals	Monotone	6	3.08	0.00
Vocals	Rapping	20	16.82	62.17
Vocals	Screaming	15	10.90	10.83
Vocals	Spoken	10	2.84	0.00
Vocals	Strong	72	28.80	23.85
Vocals	Vocal-Harmonies	57	20.69	20.04
Genre-Best	Alternative	42	19.90	17.21
Genre-Best	Classic-Rock	47	27.76	18.15
Genre-Best	Metal/Hard-Rock	10	7.15	4.00
Genre-Best	Punk	9	6.38	0.00
Genre-Best	Soft-Rock	27	15.94	20.85
Genre-Best	Soul	5	1.82	0.00
Genre-Best	Blues	8	4.17	0.00
Genre-Best	Country	9	3.72	0.00
Genre-Best	Electronica	40	23.79	31.15
Genre-Best	Folk	7	6.04	0.00
Genre-Best	Hip-Hop/Rap	19	16.01	68.83
Genre-Best	Jazz	9	6.57	5.00
Genre-Best	Pop	23	10.87	1.67
Genre-Best	R&B	13	7.28	11.86
Genre-Best	Rock	37	19.79	16.15
Genre-Best	World	16	9.39	3.33
Instrument	Acoustic-Guitar-Solo	6	1.64	0.00
Instrument	Electric-Guitar-(clean)-Solo	21	10.80	6.04
Instrument	Electric-Guitar-(distorted)-Solo	36	23.08	12.44
Instrument	Female-Lead-Vocals-Solo	7	0.95	0.00
Instrument	Harmonica-Solo	6	2.50	0.00
Instrument	Male-Lead-Vocals-Solo	40	13.05	11.81
Instrument	Piano-Solo	14	11.05	2.86
Instrument	Saxophone-Solo	10	4.63	0.00
Instrument	Trumpet-Solo	6	0	0.00

Table E.6: CAL500 Tag-category, Tag name, tag frequencies, F-measure results for BoC-W MFCC, and F-measure results for BoF MFCC (part III).

Appendix **F**

Author's publications

F.1 ISI-indexed peer-reviewed journals

Bogdanov D., Haro M., Fuhrmann F., Xambó A., Gómez E., and Herrera P. Semantic audio content-based music recommendation and visualization based on user preference examples, *Information Processing & Management*, Volume 49, Issue 1, 13-33, ISSN 0306-4573, 2013.

Serrà J., Corral A., Boguñá M., Haro M., and Arcos J. L. Measuring the evolution of contemporary western popular music. *Scientific Reports*. 2(521), doi:10.1038/srep00521, 2012.

Haro M., Serrà J., Herrera P., and Corral A. Zipf's Law in Short-Time Timbral Codings of Speech, Music, and Environmental Sound Signals. *PLoS ONE*. 7(3), e33993, doi:10.1371/journal.pone.0033993, 2012.

F.2 Book chapters

Van Balen J., Serrà J., and Haro M. Sample identification in hip-hop music. *Lecture Notes in Computer Science, Springer, Volume 7900*, In press.

F.3 Peer-reviewed conference publications

Van Balen J., Haro M., and Serrà J. Automatic Identification of Samples in Hip Hop Music. *9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012.

Haro M., Serrà J., Corral A., and Herrera P. Power-Law Distribution in Encoded MFCC Frames of Speech, Music, and Environmental Sound Signals. *21st International World Wide Web Conference (WWW 2012): 4th International Workshop on Advances in Music Information Research (AdMIRe 2012).* 895-902, 2012.

Bogdanov D., Haro M., Fuhrmann F., Xambó A., Gómez E., and Herrera P. A Content-based System for Music Recommendation and Visualization of User Preferences Working on Semantic Notions. *9th International Workshop on Content-based Multimedia Indexing*, 2011.

Molina P., Haro M., and Jordà S. BeatJockey: A new tool for enhancing DJ skills. *New Interfaces for Musical Expression (NIME).* 288-291, 2011.

Bogdanov D., Haro M., Fuhrmann F., Gómez E., and Herrera P. Content-based music recommendation based on user preference examples. *The 4th ACM Conference on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010)*, 2010.

Haro M., Xambó A., Fuhrmann F., Bogdanov D., Gómez E., and Herrera P. The Musical Avatar - A visualization of musical preferences by means of audio content description. *5th Audio Mostly Conference: A Conference on Interaction with Sound*, 2010.

Gómez E., Haro M., and Herrera P. Music and geography: content description of musical audio from different parts of the world. *10th International Society for Music Information Retrieval Conference*, 2009

Fuhrmann F., Haro M., and Herrera P. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. *Conference of the International Society for Music Information Retrieval (ISMIR)*, 2009.

Janer J., Haro M., Roma G., Fujishima T., and Kojima N. Sound Object Classification for Symbolic Audio Mosaicing: A Proof-of-Concept. (Gouyon, F., Barbosa, A., Serra, X., Ed.), *Sound and Music Computing Conference.* 297-302, 2009.

Haro M., and Herrera P. From Low-level to Song-level Percussion Descriptors of Polyphonic Music. *10th International Society for Music Information Retrieval Conference*, 2009.

F.4 Other conferences

Serrà J., Corral A., Boguñá M., Haro M., and Arcos J. L. Quantifying the evolution of popular music. *NoLineal 2012*, Zaragoza, Spain, 2012.

Nogueira W., Haro M., Herrera P., and Serra X. (2011). Music Perception with Current Signal Processing Strategies for Cochlear Implants. *Isabel*, Barcelona, Spain, 2011.

F.5 Theses

Haro M. Detecting and Describing Percussive Events in Polyphonic Music. *Master Thesis*, Universitat Pompeu Fabra, 2008.

F.6 Patents

Uemura N., Usui J., Kamiya Y., Arimoto K., Janer J., Haro M., and Roma G. Musical Composition Processor and Program Application, Priority number: JP20090037564 20090220, Publication date: 09/02/2010, Holder Entity: YAMAHA Corp.

Additional and up-to-date information about the author may be found at the authors web page[1].

[1] http://martinharo.weebly.com