# Power-Law Distribution in Encoded MFCC Frames of Speech, Music, and Environmental Sound Signals

Martín Haro
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
martin.haro@upf.edu

Joan Serrà
Artificial Intelligence Research
Institute (IIIA-CSIC)
Bellaterra, Spain
jserra@iiia.csic.es

Álvaro Corral
Complex Systems Group
Centre de Recerca
Matemàtica
Bellaterra, Spain
acorral@crm.cat

Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
perfecto.herrera@upf.edu

## ABSTRACT

Many sound-related applications use Mel-Frequency Cepstral Coefficients (MFCC) to describe audio timbral content. Most of the research efforts dealing with MFCCs have been focused on the study of different classification and clustering algorithms, the use of complementary audio descriptors, or the effect of different distance measures. The goal of this paper is to focus on the statistical properties of the MFCC descriptor itself. For that purpose, we use a simple encoding process that maps a short-time MFCC vector to a dictionary of binary code-words. We study and characterize the rank-frequency distribution of such MFCC code-words, considering speech, music, and environmental sound sources. We show that, regardless of the sound source, MFCC code-words follow a shifted power-law distribution. This implies that there are a few code-words that occur very frequently and many that happen rarely. We also observe that the inner structure of the most frequent code-words has characteristic patterns. For instance, close MFCC coefficients tend to have similar quantization values in the case of music signals. Finally, we study the rank-frequency distributions of individual music recordings and show that they present the same type of heavy-tailed distribution as found in the large-scale databases. This fact is exploited in two supervised semantic inference tasks: genre and instrument classification. In particular, we obtain similar classification results as the ones obtained by considering all frames in the recordings by just using 50 (properly selected) frames. Beyond this particular example, we believe that the fact that MFCC frames follow a power-law distribution could potentially have important implications for future audio-based applications.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Methodologies and techniques*

*WWW 2012 Companion*, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

## Keywords

sound retrieval, music information research, timbre, MFCC, power-law, large-scale data

## 1. INTRODUCTION

Many technological applications dealing with audio signals use Mel-Frequency Cepstral Coefficients (MFCC) [11] as main timbral descriptor [30, 21, 6, 27]. It is common practice to compute such MFCC values from consecutive short-time audio frames (usually with lengths below 100 ms). Later on, these frame-based descriptors can be used in a bottom-up audio processing strategy [6]. For instance, in automatic classification tasks, the content of several minutes of audio can be aggregated in a real-valued vector containing the mean values of all MFCC coefficients (and often their variances and covariances). In audio similarity tasks, one can estimate the similarity between two sounds by computing a distance measure between MFCC vectors [21], e.g. by simply using the Euclidean distance or by comparing Gaussian mixture models [2]. Evidently, these types of procedures assume a certain homogeneity in the MFCC vector space (i.e. the multidimensional space of MFCC coefficients should not have small areas that are extremely populated and, at the same time, extensive areas being low-populated). Otherwise, the results obtained from computing statistical moments or some distance measures will be highly biased towards the values of those extremely populated areas (i.e. those extremely frequent MFCC vectors).

In other research areas such as natural language processing [26] and Web mining [23], 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 [3]. Recently, these type of text categorization techniques have been applied with success in image retrieval [20]. Unfortunately, there is a lack of research in the sound retrieval community with regard to the study of the statistical distribution of sound descriptors.

This could be partially substantiated by the fact that low-level descriptors do not form discrete units or symbols that can be easily characterized by their frequency of use, as it is the case with text.

In this paper we study and characterize the probability distribution of encoded (or discretized) MFCC descriptors extracted on a frame-by-frame basis. For that, we employ a simple encoding process which maps a given MFCC frame to a dictionary of more than 4 million binary code-words. 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 analysis and show that encoded MFCC frequencies follow a shifted power-law distribution, a particular type of heavy-tailed distribution. This distribution is found independently of sound source and frame size. 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 music recordings. Finally, we perform two automatic classification tasks that add further evidence to support this last claim.

In the next subsection, an overview on heavy-tailed distributions is given. In Section 2, a description of the used methodology is presented, including descriptions of the analyzed databases, encoding process, and power-law estimation method. Section 2.3 reports on the MFCC distributions. In Section 4, the two classification experiments are presented. Finally, Section 5 concludes the paper.

## 1.1 Heavy-tailed distributions

When studying the statistical properties of data coming from several scientific disciplines, researchers often report heavy-tailed distributions [1, 4, 24, 28, 36]. This means that the measured data points are spread over an extremely wide range of possible values and that there is no typical value around which these measurements are centered [28]. It also implies that the majority of data points do not occur frequently (i.e. the ones in the tail).

A particularly important landmark in the study of heavy-tailed distributions was the seminal work of Zipf [36], showing a power-law distribution of word-frequency counts with an exponent $\alpha$ close to 1,

$$z(r) \propto r^{-\alpha}, \tag{1}$$

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 $r$. Such power-law behaviour implies that a few words occur very frequently and many happen rarely, without a characteristic separation between them. Zipf's power-law (Eq. 1) also indicates a power-law probability distribution of word frequencies [1],

$$P(z) \propto z^{-\beta}, \tag{2}$$

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

Pioneering also the study of the statistical properties of music-related data, Zipf himself reported power-law distributions in melodic intervals and distances between note repetitions from a reduced set of music scores [36]. In the last decades, other researchers have reported heavy-tailed distributions of data extracted from music scores [18, 19] and MIDI files [5, 25, 35]. Regarding audio-based descriptors, few works can be found showing heavy-tailed distributions. These works have mainly focused on sound amplitudes of music, speech, and crackling noise signals [22, 31, 34]. Nonetheless, we recently found evidence for a power-law (Zipfian) distribution of encoded short-time spectral envelopes [17], where the spectral envelopes were characterized by the energy found in Bark-bands of the power spectrum [37]. Since, as mentioned, MFCC descriptors are the primary source of information for many audio classification and retrieval tasks, we now expand and improve our previous study by focusing on the distribution of this descriptor and by providing a specific example of one of the consequences of such distribution.

# 2. METHODOLOGY

## 2.1 Databases

In this work we analyze 740 hours of real-world sounds. These sounds are grouped into four databases: *Speech*, *Western Music*, *non-Western Music*, and *Sounds of the Elements* (i.e. sounds of natural phenomena such as rain, wind, and fire). The *Speech* database contains 130 hours of recordings of English speakers from the Timit database [15] (about 5.4 hours), the Library of Congress podcasts[1] (about 5.1 hours), and 119.5 hours from Nature podcasts[2] from 2005 to April 7th 2011 (the first and last 2 minutes of sound were removed to skip potential musical contents). The *Western Music* database contains 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, we gathered 48 hours of sounds produced by natural inanimate processes such as water (rain, streams, waves, melting snow, waterfalls), fire, thunders, wind, and earth (rocks, avalanches, eruptions). This *Sounds of the Elements* database was assembled using files downloaded from *The Freesound Project*[3]. The differences in size among databases try to account for 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).

## 2.2 Encoding process

A block-diagram of the encoding process can be seen in Fig. 1. Starting from the raw audio signal (44,100 Hz, 16 bits) we first apply an equal-loudness filter consisting of an inverted approximation of the equal-loudness curves described by Fletcher and Munson [12]. Then, we cut the audio signal into non-overlapping temporal frames (Fig. 1a). In this study we consider three perceptually motivated frame sizes, namely 46, 186, and 1,000 ms. The 46 ms frame size is extensively used in audio processing algorithms [6, 27]. The 186 ms frame corresponds to a perceptual measure of sound grouping called "temporal window integration" [29], usually described between 170 and 200 ms. Finally, we study a

[1] "Music and the brain" podcasts: `http://www.loc.gov/podcasts/musicandthebrain/index.html`

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

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

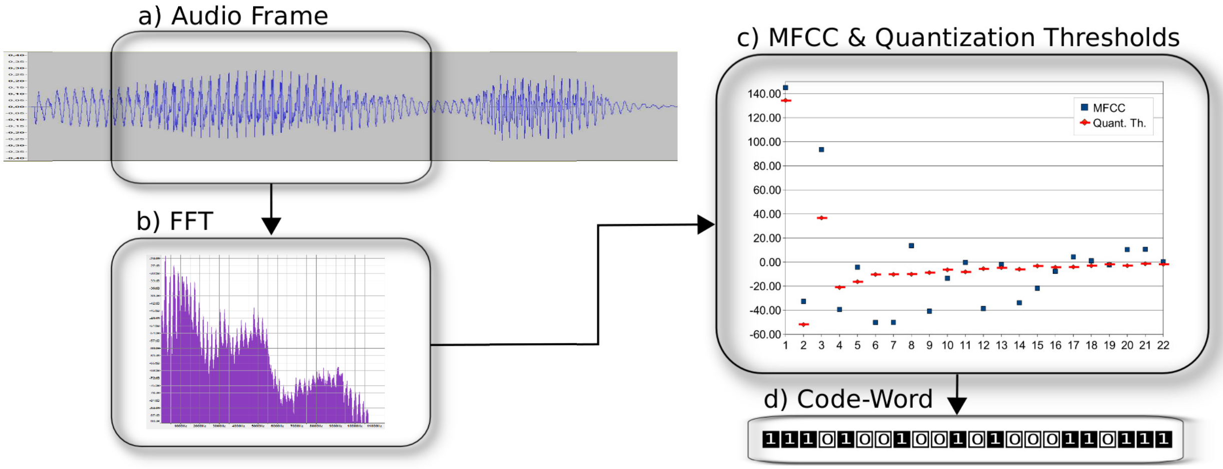

**Figure 1: Block diagram of the encoding process. a) The audio signal is segmented into non-overlapping frames. b) The power spectrum of each frame is obtained. c) MFCC coefficients (blue squares) are computed and each coefficient is binary-quantized by comparing its value against a pre-computed threshold (red line). d) Each quantized MFCC vector forms an MFCC code-word.**

relatively long temporal frame (1 s) that exceeds the usual duration of musical notes and speech phonemes.

After frame cutting, the signal of each frame is converted to the frequency domain by taking its Fourier transform using a Blackman-Harris window. From the output of the Fourier transform we compute its power spectrum, taking the square of the magnitude values (Fig. 1b). The MFCC descriptor is obtained by mapping the short-time power spectrum to the Mel scale [33]. 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 [32] 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.

In order to be able to account for the rank-frequency distribution of MFCC frames we first need to discretize the multidimensional MFCC vector space in such way that similar regions are assigned to the same discrete point (or code-word). Since we are dealing with a 22 dimensional vector space, discretizing each dimension into just two values already produces millions of possible code-words. Thus, we opt for the simple, unsupervised equal-frequency discretization approach [7] that allows us to work with such big dictionaries. It is worth noting here that the use of more elaborated coding techniques, like vector quantization [30], would rely on predefined distance measures, and would require a high computational load to infer millions of code-words.

To obtain an MFCC code-word, we quantize each MFCC coefficient by assigning all values below a stored threshold to 0 and those being equal or higher than the threshold to 1 (Fig. 1c). These quantization thresholds are different for each MFCC coefficient and correspond to the median values found in a representative dataset (i.e. the value that splits the distribution of the coefficient into two equally populated groups). The representative dataset we used to compute the median values contained all MFCC frames from the *Sounds of the Elements* database plus a random sample of MFCC frames from the *Speech* database that match in number the ones from the *Sounds of the Elements*. It also included random selections of *Western Music* and *non-Western Music* matching half of the length of *Sounds of the Elements* each. Thus, the dataset had its MFCC frames distributed as one third coming from *Sounds of the Elements*, one third from *Speech* and one third from *Music*. We constructed 10 of such datasets per frame size and stored the mean of the median values as the quantization threshold. After this binary encoding process, every audio frame is mapped into one of the $2^{22} = 4,194,304$ possible MFCC code-words (Fig. 1d).

## 2.3 Power-Law Estimation

To evaluate if a power-law distribution fits our data we take the frequency count of each MFCC 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 [8, 9]. We now give a brief overview of the process. For more details we refer to the references above or to [17].

The procedure consists of finding the minimum frequency $z_{\min}$ for which an acceptable power-law fit is obtained. First, arbitrary values for the lower cutoff $z_{\min}$ are selected and the power-law exponent $\beta$ is obtained by maximum-likelihood estimation of the distribution of frequencies. Next, the Kolmogorov-Smirnov test quantifies the separation between the resulting fit and the data. 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) 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. Fi-

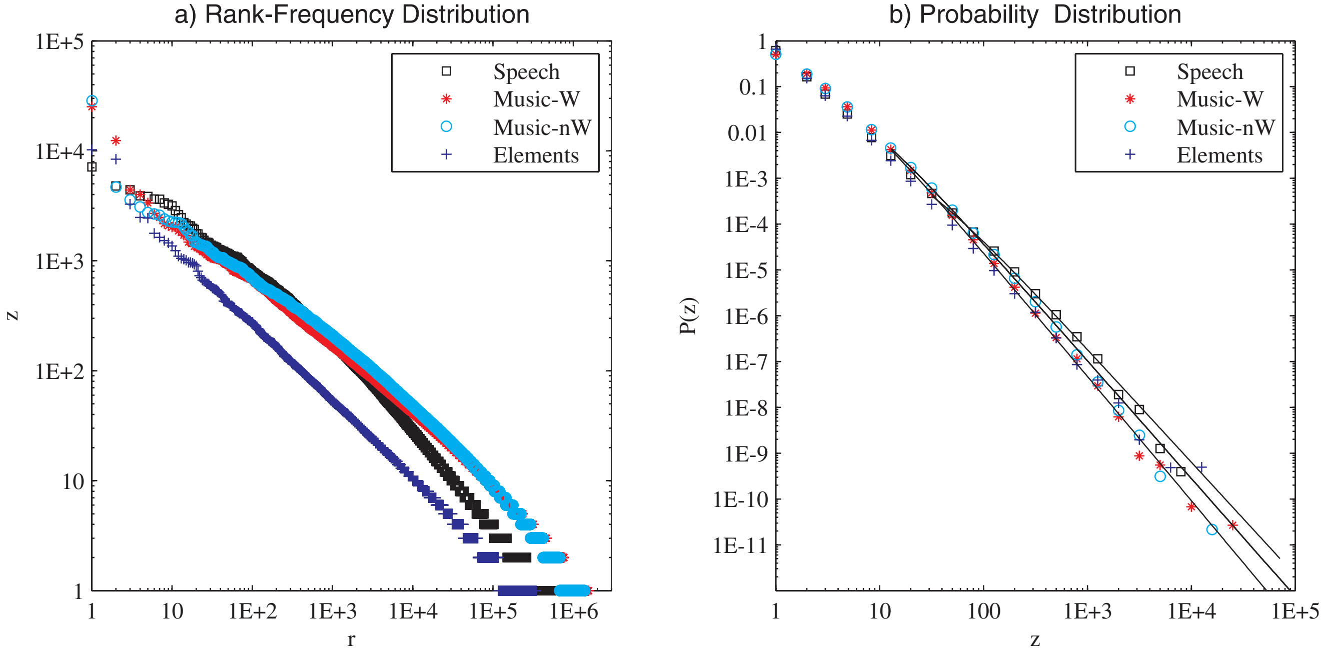

**Figure 2: a) Rank-frequency distribution of MFCC code-words per database (frame size = 186 ms). b) Probability distribution of frequencies for the same code-words (the black lines correspond to the fitted distribution).**

nally, the procedure selects the value of $z_{\min}$ which yields the largest power-law range (i.e., the smallest $z_{\min}$) provided that the $p$-value is above a certain threshold (for instance 20%). We apply this fitting procedure to 10 random samples of 300,000 code-words per database and frame size.

## 3. DISTRIBUTION RESULTS

Following the methodology described in the previous section we encode every audio frame into its corresponding code-word. Next, for each database and frame size, we count the frequency of use of each MFCC code-word (i.e. the number of times a code-word appeared in the database) and we sort them by decreasing order of frequency. As it can be seen in Fig. 2a, when plotting these rank-frequency counts we observe heavy-tailed distributions for all the analyzed databases. These distributions imply that a few MFCC code-words are very frequent while most of them are very unusual [28].

Next, in order to evaluate if the found heavy-tailed distributions specifically correspond to power-law distributions we apply the previously described estimation procedure which, instead of working directly with the rank-frequency plots, it focuses on the equivalent description in terms of the distribution of the frequency (Fig. 2b). The obtained results reveal that for all analyzed databases and frame sizes, the best fit corresponds to a shifted (discrete) power-law

$$P(z) \propto (z+c)^{-\beta}, \tag{3}$$

where $c$ is a constant value. By adding this constant value to Eq. 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 (see Table 1 for a complete list of the fitted parameters).

From the fitting results of Table 1 we observe that not only all analyzed databases correspond to the same distribution type, but also their exponents are somewhat similar (i.e. all the $\alpha$ 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 $\alpha$. 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 is quite surprising given the fact that we are changing the frame size by almost one and a half orders of magnitude (from 46 to 1,000 ms) and seems to be a unique feature of music-derived 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 (note that due to the heavy-tailed distribution there are thousands of code-words with frequency one; see Fig. 2a). Since each code-word corresponds to a 22-dimensional vector of zeros and ones, we can easily visualize them by assigning the white color to those values equal to zero and the black color to those quantized as one (Fig. 3). 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 fact that the MFCC descriptor was orig-

**Table 1: Fitting results. Average values from 10 random samples of 300,000 code-words per database and frame size are reported (standard deviation in parenthesis).**

| Database/frame size | $z_{min}$ | $\beta$ | $c$ | $\alpha$ |
|---|---|---|---|---|
| *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) |

inally designed to describe speech signals. Furthermore, it turns out that the most frequent code-words of speech are quite different from the ones in the other type of sounds. We leave this issue for future research. Notice that in the other databases the most frequent code-words present a smooth structure, with close/neighboring MFCC coefficients having similar quantization values.

We further investigate the rank-frequency distribution of MFCC code-words for individual songs found in both *Western* and *non-Western Music* databases. Noticeably, these individual songs show a heavy-tailed distribution similar to that observed in the full databases. Examples of the obtained distributions can be seen in Fig. 4.

## 4. CLASSIFICATION EXPERIMENTS

In the previous section we have shown that encoded short-time MFCC vectors follow a shifted power-law distribution, where the most copied code-words have characteristic patterns. We have also shown that individual music recordings seem to present the same type of distribution. In this section, we provide additional evidence to support the claim that MFCC vectors from individual music recordings are also heavy-tailed. 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 using just those few highly frequent MFCC vectors as input for an automatic classification task will yield similar results as selecting all frames and taking the mean (i.e. the classic bag-of-frames approach).

We evaluate this hypothesis with two supervised semantic inference tasks: automatic genre classification and musical instrument identification. In both tasks we deliberately use a simple pattern recognition strategy. Specifically, we use support vector machines (SVM) [10] to classify aggregated feature vectors of 22 MFCC means per audio file. Our main goal is to compare the classification results obtained when using all audio frames versus using a reduced set of selected frames to compute the mean feature vector. To select these frames we first encode each audio frame into its corresponding MFCC code-word. 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 rank-frequency distribution). Then, we select the $N$ most (or least) frequent

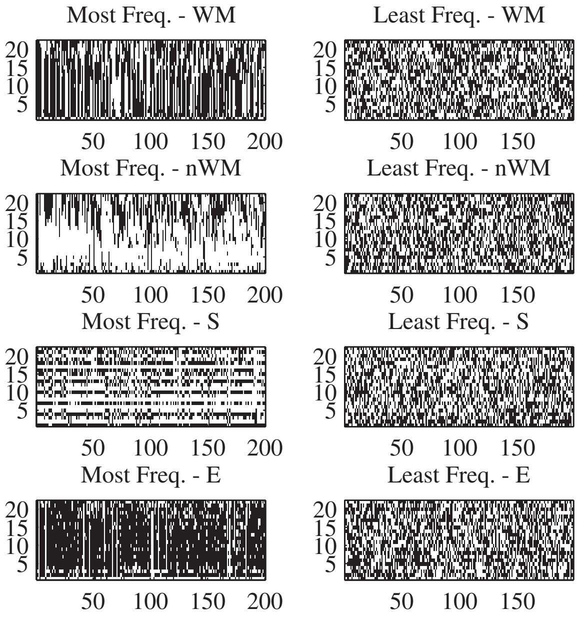

**Figure 3: Most (left) and least (right) frequent MFCC code-words per database using a frame size of 186 ms. For each plot, the horizontal axis corresponds to individual code-words and the vertical axis corresponds to quantized MFCC coefficients (white = 0, black = 1). Every position in the abscissa represents a particular code-word. From top to bottom we plot code-words for *Western Music* (WM), *non-Western Music* (nWM), *Speech* (S), and *Sound of the Elements* (E) databases.**

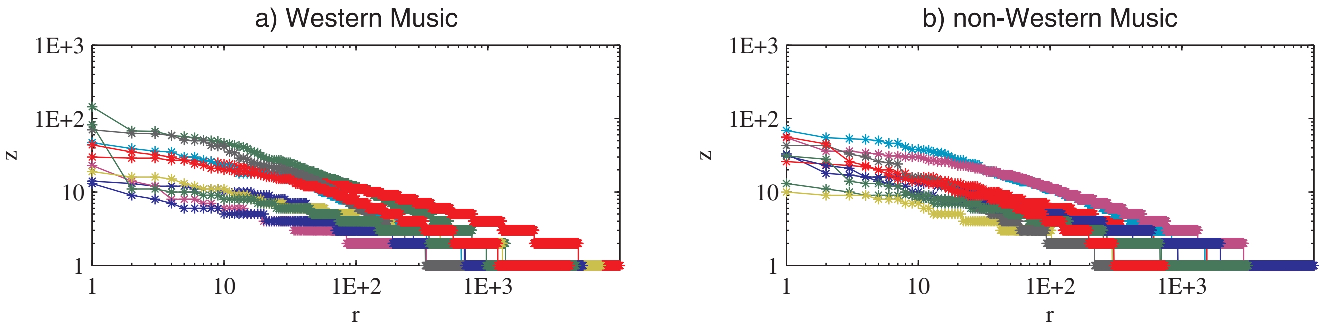

**Figure 4: Example of rank-frequency distributions of MFCC code-words from 10 randomly selected music recordings per database using a frame size of 46 ms. Each line type corresponds to one recording.**

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 (or least) frequent code-words of the music recording.

The audio files used in these experiments do not form part of the databases described in Section 2. 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[4]. 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 [16]. For the musical instrument identification task we use an in-house dataset of 2,355 audio excerpts extracted from commercial CDs [14]. 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). In both tasks, for the extraction of MFCC descriptors we use a frame size of 46 ms with 50% overlap. We select the best F-measure[5] classification result after evaluating four SVM kernels with default parameters[6] (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.

The obtained F-measures can be seen in Table 2. 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 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. 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).

## 5. CONCLUSION AND FUTURE WORK

In the present work we have analyzed the rank-frequency distribution of encoded MFCC vectors. We study a large database of sounds coming from disparate sources such as speech, music, and environmental sounds. This database represents a large portion of the timbral variability perceivable in the world. We have found that the corresponding frequency distributions can be described by a shifted power-law with similar exponents. This distribution is found regardless of the analyzed sound source and frame size, and suggests that it is a general property of the MFCC descriptor (and possibly of the underlying sound generation process or the musical facet the MFCC accounts for). Noticeably, the fitting results have shown almost identical exponents for both *Western* and *non-Western Music* databases and across different frame sizes. A further study of the inner structure of MFCC code-words reveals that the most copied code-words have characteristic patterns in all analyzed sound sources. In particular, the most frequent code-words in *Western Music*, *non-Western Music*, and *Sounds of the Elements* present a smooth structure where close/neighboring MFCC coefficients tend to have similar quantization values. In the case of *Speech*, we observe a different pattern where some coefficients of the most copied code-words tend to be quantized as zero while other coefficients tend to be quantized as one.

Motivated by the extreme stability of the shifted power-law in both music databases we have also analyzed the rank-

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

[5] Where F-measure=2*Precision*Recall/(Precision+Recall).

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

**Table 2: 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. In the last column we include the classification result obtained when using all the frames of the recording. The differences between both classification strategies are also shown.**

| | Number of selected frames ($N$) | | | | | |
|---|---|---|---|---|---|---|
| **Task / 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 |

frequency distributions of individual music recordings. By visualizing several randomly selected recordings of both music databases we discovered that in most of the cases their distributions were also power-law shaped. Finally, we presented two supervised semantic inference tasks providing 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 selected 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).

Since current technological applications do not take into account that the MFCC descriptor follows a shifted power-law distribution, the implications of the results presented here for future applications should be thoughtfully considered and go beyond the scope of this paper. In the near future we plan to further explore these implications. For instance, as shown in our experiments, taking very few highly-frequent MFCC vectors provides similar classification results as compared to taking all vectors in a song. Moreover, assuming a descriptor's power-law distribution, one could speculate that when taking $X$ random frames from a bag-of-frames (using uniform distribution) there is a very high probability that those selected frames belong to the most copied MFCC code-words (because those code-words are very frequent). Therefore, high classification results should be also achieved using just this random selection strategy. Importantly, this 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 have a tremendous impact 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) [13]. 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. recording-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. 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 [3]. Later on, these weighted MFCC frames could be used in classification or audio similarity tasks.

## 6. ACKNOWLEDGMENTS

This work has been supported by the following projects: FIS2009-09508, 2009SGR-164, the European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711, JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas and 2009-SGR-1434 from Generalitat de Catalunya.

## 7. REFERENCES

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

[2] J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the 3rd International Symposium on Music Information Retrieval*, pages 157–163, Paris, France, 2002.

[3] R. Baeza-Yates. *Modern information retrieval.* ACM Press, Addison-Wesley,New York, 1999.

[4] P. Bak. *How nature works: the science of self-organized criticality.* Copernicus, New York, 1996.

[5] 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.

[6] 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.

[7] K. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan. *Data mining: a knowledge discovery approach.* Springer, New York, 2007.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661, 2009.

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

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sept. 1995.

[11] 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.

[12] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *J Acoust Soc Am*, 5(2):82, 1933.

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

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

[15] 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, Philadelphia, 1993.

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

[17] M. Haro, J. Serrá, P. Herrera, and A. Corral. Zipf's law in short-time timbral codings of speech, music, and environmental sound signals. *PLoS ONE*, 2012. In press.

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

[19] 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.

[20] Y. Jiang, J. Yang, C. Ngo, and A. Hauptmann. Representations of Keypoint-Based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, Jan. 2010.

[21] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, New York, 1 edition, 2006.

[22] 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.

[23] B. Liu. *Web data mining : exploring hyperlinks, contents, and usage data*. Springer, New York, 2nd edition, 2011.

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

[25] 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.

[26] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. The MIT Press, 1 edition, 1999.

[27] 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.

[28] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323, 2005.

[29] 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.

[30] T. F. Quatieri. *Discrete-time speech signal processing: principles and practice*. Prentice Hall, New Jersey, 1 edition, 2001.

[31] J. P. Sethna, K. A. Dahmen, and C. R. Myers. Crackling noise. *Nature*, 410(6825):242–250, 2001.

[32] M. Slaney. Auditory toolbox v2. Technical Report 1998-010, 1998.

[33] 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.

[34] R. F. Voss and J. Clarke. 1/f noise in music and speech. *Nature*, 258(5533):317–318, 1975.

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

[36] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, 1949.

[37] 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.