FACILITATING COMPREHENSIVE BENCHMARKING EXPERIMENTS ON THE MILLION SONG DATASET

Alexander Schindler, Rudolf Mayer, and Andreas Rauber

Institute of Software Technology and Interactive Systems, Vienna University of Technology
{schindler,mayer,rauber}@ifs.tuwien.ac.at

ABSTRACT

The Million Song Dataset (MSD), a collection of one million music pieces, enables a new era of research of Music Information Retrieval methods for large-scale applications. It comes as a collection of meta-data such as the song names, artists and albums, together with a set of features extracted with the The Echo Nest services, such as loudness, tempo, and MFCC-like features.

There is, however, no easily obtainable download for the audio files. Furthermore, labels for supervised machine learning tasks are missing. Researchers thus are currently restricted on working solely with these features provided, limiting the usefulness of MSD. We therefore present in this paper a more comprehensive set of data based on the MSD, allowing its broader use as benchmark collection. Specifically, we provide a wide and growing collection of other well-known features in the MIR domain, as well as ground truth data with a set of recommended training/test splits.

We obtained these features from audio samples provided by 7digital.com, and metadata from the All Music Guide. While copyright prevents re-distribution of the audio snippets per se, the features as well as metadata are publicly available on our website for benchmarking evaluations. In this paper we describe the pre-processing and cleansing steps applied, as well as feature sets and tools made available, together with first baseline classification results.

1. INTRODUCTION

Music Information Retrieval (MIR) research has historically struggled with issues of publicly available benchmark datasets that would allow for evaluation and comparison of methods and algorithms on the same data base. Most of these issues stem from the commercial interest in music by record labels, and therefore imposed rigid copyright issues, that prevent researchers from sharing their music collections with others. Subsequently, only a limited number of data sets has risen to a pseudo benchmark level, i.e. where most of the researchers in the field have access to the same collection.

Another reason identified as a major challenge for providing access to research data in general is the lack of esteem and valuation of these kind of activities. While preparing, maintaining and providing access to massive data collections requires significant investments in terms of system maintenance and data (pre-)processing, it is considered administrative rather than research work (in spite of several even research-affine challenges emerging during such activities), and thus does not gain acceptance in classical research-oriented publication venues. Such lack of career rewards is one of the many factors, next to legal limitations and lack of expertise, limiting sharing of research data [7]. Several initiatives have been started in the Research Infrastructures area to mitigate this problem and foster collaborative research. These areas span across virtually all topical areas, from Astronomy, via meteorology, chemistry to humanities [1] .

A recent effort in the MIR domain has lead to the compilation of the Million Song Dataset [3] (MSD). It provides a database of meta-data for a collection of one million songs, such as the song name, artists and album. In addition, a number of descriptive features extracted with the services from The Echo Nest [2] are provided. These features include tempo, loudness, timings of fade-in and fade-out, and MFCC-like features for a number of segments. Moreover, a range of other meta-data has been published recently, such as song lyrics (for a subset of the collection), or tags associated to the songs from Last.fm [3] .

The MSD enables researchers to test algorithms on a large-scale collection, thus allowing to test them on more real-world like environments. However, there are no easily obtainable audio files available for this dataset, and therefore, researchers are practically restricted to benchmarking on algorithms that work on top of features, such as recommendation of classification, but can not easily develop new or test existing feature sets on this dataset. The availability of just one feature set also does not allow an evaluation across multiple feature sets. As previous studies showed, however, there is no single best feature set, but their performance depends very much on the dataset and the task.

We therefore aim to alleviate these restrictions by pro-

[1] http://ec.europa.eu/research/infrastructures/
[2] http://the.echonest.com
[3] http://www.last.fm

viding a range of features extracted for the Million Song Dataset, such as MFCCs and a set of low-level features extracted with the jAudio feature extraction software as well as the Marsyas framework [13], and the Rhythm Patterns and derived feature sets [9]. To this end, we first obtained audio samples for the MSD by using the content provider 7digital.

A second shortcoming of the MSD is that it does not, at the moment, contain a mapping to categorisation such as genres. Thus, experimental evaluations such as musical genre classification, a popular task in MIR research, are not possible. We therefore further propose a partitioning of the dataset into a set of genres obtained from the All Music Guide [4]. Specifically, we created a partitioning on two levels of detail, with 13 top-level-genres and 25 sub-genres, and propose a number of splits for training and test sets, with different filters, allowing several tasks for evaluation.

Both the feature sets and the partitioning into genres are available from our website [5]. The features are stored in the WEKA Attribute-Relation File Format (ARFF) [14], with one attribute being the unique identifier of the song in the MSD. We further provide a set of scripts to match the features with the genre mapping so that they can be used for classification experiments.

The remainder of this paper is structured as follows. Section 2 introduces the dataset and the properties of the audio samples, while Section 3 describes the sets of features extracted from them. Section 4 gives details on the genre assignment obtained, and in Section 5, we describe benchmark partitions and how we aim to facilitate exchange between researchers. Finally, we provide conclusions in Section 6

2. THE DATASET

The Million Song Dataset contains in the meta-data a unique identifier for an audio sample at the content provider 7digital [6]. For some songs no sample could be downloaded, as the identifier was unknown to 7digital. We thus obtained a total of 994,960 audio samples, i.e. a coverage of 99.5% of the dataset; the list of missing audio samples is available on the website. This points to an important issue related to the use of external on-line resources for scientific experimentation. Especially when the provider is not genuinely interested in the actual research performed, there is little motivation to maintain the data accessible in unmodified manners, and is thus susceptible to changes and removal. Thus, maintaining a copy of a fixed set of data is essential in benchmarking to allow the evaluation of newly developed feature sets, and for acoustic evaluation of the results.

In total, the audio files account for approximately 625 gigabyte of data. The audio samples do not adhere to a common encoding quality scheme, i.e. they differ in length and quality provided. Figure 1 shows a plot of the sample lengths; please note that the scale is logarithmic. It can

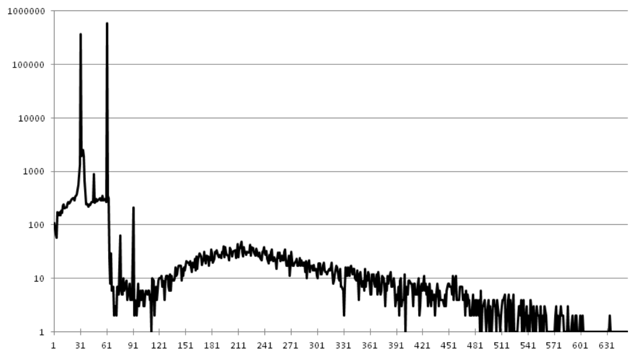

Figure 1: Distribution of sample length

be observed that there are two peaks at sample lengths of 30 and 60 seconds with 366,130 and 596,630 samples, respectively, for a total of 96,76% of all the samples. These shorter snippets normally contain a section in the middle of the song. Many other well-known collections in the MIR domain as well contain only 30 second snippets, and feature extraction algorithms normally deal with this.

Table 1: Audio properties of 7Digital Samples

Samplerate		
22	768,710	77,26%
44	226,169	22,73%
other	81	0,01%
Bitrate		
128	646,120	64,94%
64	343,344	34,51%
other (VBR)	5,494	0,55%
Channels		
Mono	6,342	0.64%
Stereo	150,779	15.15%
Joint stereo / dual channel	837,839	84.21%

Table 1 gives an overview on the audio quality of the samples. The majority, more than three quarters, of the audio snippets have a sample rate of 22khz, the rest has a sample rate of 44khz (with the exception of 81 songs, of which the majority have 24 and 16khz). Regarding the bitrate, approximately two-thirds of the songs are encoded with 128kbit, the majority of the rest with 64kbit; only about half a percent of the songs come with higher (192 or 320kbps) or variable bitrates (anywhere between 32 and 275kpbs). Almost all samples are provided in some form of stereo encoding (stereo, joint stereo or dual channels) – only 0.6% of them have only one channel. As these characteristics, specifically the sample rate, may have significant impact on the performance of the data analysis algorithms, we must consider these for stratification purpose when designing the benchmark splits.

3. FEATURE SETS

We extracted a wide range of audio features from the samples provided, namely features provided by the jAudio feature extraction software (which is a part of the jMIR pack-

[4] http://allmusic.com
[5] http://www.ifs.tuwien.ac.at/mir/msd/
[6] http://www.7digital.com

Table 2: Overview on features extracted from the MSD samples. *Dim.* denotes the dimensionality, *Deriv.* derivatives computed from the base features

#	Feature Set	Extractor	Dim	Deriv.
1	MFCCs [12]	MARSAYS	52	
2	Chroma [6]	MARSAYS	48	
3	timbral [13]	MARSAYS	124	
4	MFCCs [12]	jAudio	26	156
5	Low-level spectral features [11] (Spectral Centroid, Spectral Rolloff Point, Spectral Flux, Compactness, and Spectral Variability, Root Mean Square, Zero Crossings, and Fraction of Low Energy Windows)	jAudio	16	96
6	Method of Moments [11]	jAudio	10	60
7	Area Method of Moments [11]	jAudio	20	120
8	Linear Predictive Coding [11]	jAudio	20	120
9	Rhythm Patterns [9]	rp_extract	1440	
10	Statistical Spectrum Descriptors [9]	rp_extract	168	
11	Rhythm Histograms [9]	rp_extract	60	
12	Modulation Frequency Variance Descriptor [10]	rp_extract	420	
13	Temporal Statistical Spectrum Descriptors [10]	rp_extract	1176	
14	Temporal Rhythm Histograms [10]	rp_extract	420	

age [11]), the MARSYAS feature extractor [13], and the Rhythm Patterns family of feature sets [9]. An overview on these features is given in Table 2.

The jAudio software provides a range of 28 features associated with both the frequency and time domains. It includes several intermediate-level musical features, mainly related to rhythm, as well as lower-level signal processing-oriented features. It also provides an implementation of MFCC features [12], using 13 coefficients. jAudio computes in general mean and standard deviations over the sequence of frames, and provides for most measures also derivatives, i.e. additional statistical moments over the basic measures. For the extraction, we utilised jAudio as bundled in the jMIR 2.4 release [7].

A very popular audio feature extraction system is MARSAYS, one of the first comprehensive software packages to be available to MIR researchers. A very popular set from this audio extractor is the so-called "timbral" set, which is composed of 13 MFCC coefficients, and the twelve chroma features and the average and minimum chroma value, and the four low-level features zero crossings, and rolloff, flux and centroid of the spectrum. For these 31 values, four statistical moments are computed, resulting in a 124 dimensional vector. For the extraction, we utilised MARSYAS version 0.4.5 [8].

The Rhythm Patterns and related features sets are extracted from a spectral representation, partitioned into segments of 6 sec. Features are extracted segment-wise, and then aggregated for a piece of music computing the median (Rhythm Patterns, Rhythm Histograms) or mean (Statistical Spectrum Descriptors, Modulation Frequency Variance Descriptor) from features of multiple segments. For the extraction, we employed the Matlab-based implementation, version 0.6411 [9].

It is intentional that we provide two different versions of the MFCCs features, as this will allow for interesting insights in how these implementations differ on various MIR tasks.

3.1 Publication of Feature Sets

All features described above are available on our website for download, encoded in the WEKA Attribute-Relation File Format (ARFF) [14]. The features are available under the Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic License [10].

To allow high flexibility when using them, we provide one ARFF file for each type of features; these can then be combined in any particular way when performing experimental evaluations. A set of scripts is provided as well on the website for this. In total, the feature files account for approximately 40 gigabyte of uncompressed text files. The feature files contain the numeric values for each feature, and additionally the unique identifier assigned in the MSD. This way, it is possible to generate various feature files with different ground truth assignments; we again provide scripts for this. The proposed assignment into genres for genre classification tasks is described in Section 4.

Further feature sets to be extracted and provided will include e.g. MIRtoolbox [8] or M2k [5].

4. ALLMUSIC DATASETS

The All Music Guide (AMG) [4] was initiated by an archivist in 1991 and emerged 1995 from its book form into a database which can be accessed through the popular commercial Web page allmusic.com. The Web page offers a wide range of music information, including album reviews, artist biographies, discographies well as classification of

[7] available from `http://jmir.sourceforge.net/`
[8] available from `http://sourceforge.net/projects/marsyas/`
[9] available from `http://www.ifs.tuwien.ac.at/mir/downloads.html`
[10] `http://creativecommons.org/licenses/by-nc-sa/2.0/`

albums according to genres, styles, moods and themes. Information is provided and curated by experts.

Genre information is very coarse, provided as a single tag for each album. Further the two main categories Pop and Rock are combined into a single genre 'Pop/Rock'. Additionally to genre labels, style tags are provided allowing for a more specific classification. Usually multiple style tags are applied for each album, but unfortunately no weighting scheme can be identified and in many cases only one tag is provided. Style tags also tend to be even more generic than genre labels. Especially non-American music is frequently tagged with labels describing country or region as well as the language of the performing artist. Instrumentation, situational descriptions (e.g. *Christmas*, *Halloween*, *Holiday*, etc.) as well as confessional or gender attributes (e.g. *Christian*, *Jewish*, *Female*, etc.) are also provided. Unfortunately these meta-descriptive attributes are not used as isolated synonyms, but are concatenated with conventional style information (e.g. *Japanese Rock*, *Christian Punk*, *Classic Female Blues*, etc.).

Allmusic.com assembles styles to meta-styles which can be interpreted as sub genres used to diversify the major genre labels. Meta-styles are not distinctive and are used overlapping in many meta-styles (e.g. *Indie Electronic* is contained in the meta-styles *Indie Rock*, *Indie Pop* and *Alternative/Indie Rock*).

4.1 Data Collection

Data was collected automatically from Allmusic.com using direct string matching to query for artist-release combinations. From the resulting Album Web page genre and style tags were collected.

We were able to collect 21 genre labels for 62,257 albums which initially provided genre tags for 433,714 tracks. Style tags were extracted attributing only 42,970 albums resulting in 307,790 labeled tracks. An average of 3.25 tags out of a total of 905 styles were applied to each album, but 5,742 releases were only tagged with a single style label. The most popular genre with 32,696 tagged albums, was *Pop/Rock* - this is 10% more as the sum of all other genres. Referring to tracks the difference rises to 30%. Further, the granularity of Rock is very scarce, including Heavy Metal, Punk, etc. A similar predominating position of this genre as well as was also reported by [2]. The most popular style tag is *Alternative/Indie Rock* applied to 12,739 albums, which is more than twice as much as the second popular style *Alternative Pop/Rock*. About 120 tags describe the country of the performing artist or the language of the interpretation - the most common among them is *Italian Music* which has been applied to 610 albums.

4.2 Allmusic Genre Dataset

The Allmusic Genre Dataset is provided as an unoptimized expert annotated ground truth dataset for music genre classification. We provide two partitions of this set. The *MSD Allmusic Genre Dataset (MAGD)* assembles all collected genres including generic and small classes.

Table 3: MSD Allmusic Genre Dataset (MAGD) - upper part represents the MSD Allmusic Top Genre Dataset (Top-MAGD)

Genre Name	Number of Songs
Pop/Rock	238,786
Electronic	41,075
Rap	20,939
Jazz	17,836
Latin	17,590
R&B	14,335
International	14,242
Country	11,772
Reggae	6,946
Blues	6,836
Vocal	6,195
Folk	5,865
New Age	4,010
Religious	8814
Comedy/Spoken	2067
Stage	1614
Easy Listening	1545
Avant-Garde	1014
Classical	556
Childrens	477
Holiday	200
Total	422,714

The second partition - *MSD Allmusic Top Genre Dataset (Top-MAGD)* - consists of 13 genres - the 10 major genres of Allmusic.com (Pop/Rock, Jazz, R&B, Rap, Country, Blues, Electronic, Latin, Reggae, International) including the three additional genres Vocal, Folk, New Age (see Table 3). Generic genres as well as classes with less than 1% of the number of tracks of the biggest class Pop/Rock are removed. Due to the low number of tracks, the Classical genre is also removed from the Top Genre dataset.

4.3 Allmusic Style Dataset

The Allmusic Style Dataset attempts to more distinctively separate the collected data into different sub-genres, alleviating predominating classes. For the compilation of the dataset genre labels were omitted and solely style tags were used. In a first step metastyle description as presented on the Allmusic.com Web site were used to map multiple style tags to a single genre name - in this case we used the metastyle name. This simple aggregation approach generated a total of 210 genre labels many of them highly generic or hierarchical specializing (e.g. *Electric Blues* and *Electric Chicago Blues*. The *MSD Allmuisc Metastyle Dataset - Multiclass (MAMD)* was derived from these 210 resulting metaclasses. Each track was matched to one or more metaclasses according to its style tags. In a second step we removed from the initial set of 905 style tags all confessional, situational and language specific entries. Regional tags were discarded if they do not refer to a specific traditional cultural music style (e.g. *African Folk*). Popular music attributed with regional information was dis-

Table 4: The MSD Allmusic Style Dataset (MASD)

Genre Name	Number of Songs
Big Band	3,115
Blues Contemporary	6,874
Country Traditional	11,164
Dance	15,114
Electronica	10,987
Experimental	12,139
Folk International	9,849
Gospel	6,974
Grunge Emo	6,256
Hip Hop Rap	16,100
Jazz Classic	10,024
Metal Alternative	14,009
Metal Death	9,851
Metal Heavy	10,784
Pop Contemporary	13,624
Pop Indie	18,138
Pop Latin	7,699
Punk	9,610
Reggae	5,232
RnB Soul	6,238
Rock Alternative	12,717
Rock College	16,575
Rock Contemporary	16,530
Rock Hard	13,276
Rock Neo Psychedelia	11,057
Total	273,936

carded due to extensive genre overlaps (e.g. *Italian Pop* ranges from Hip-Hop to Hard-Rock). Finally, we successively merged these genres into general descriptive classes until we finalized the dataset into the *MSD Allmusic Style Dataset (MASD)* presented in Table 5. For completeness we also provide the *MSD Allmuisc Style Dataset - Multiclass (Multi-MASD)*. This set contains the pure track-style mapping as collected from Allmusic.com.

5. BENCHMARK PARTITIONS

Influenced by the tremendous experience in the text classification domain, specifically with the landmark Reuters-21578 corpus, we provide a number of benchmark partitions that researcher can use in their future studies, in order to facilitate repeatability of experiments with the MSD beyond x-fold cross validation. We also encourage and provide a platform for exchange of results obtained and new partitions created via our website.

We provide the following categories of splits:

- Splits with all the tow ground truth assignments into genre and style classes, described in Section 4.
- Splits with just the majority classes from these two ground truth assignments.
- Splits considering the sample rate of the files, i.e. only the 22khz samples, only the 44khz samples, and a set with all audio files.

Table 5: Classification results on MSD Allmusic Guide Style dataset (MASD), 66% training set split

Dataset	NB	SVM	k-NN	DT	RF
MFCC (4)	**15.04**	20.61	*24.13*	14.21	18.90
Spectral (5)	*14.03*	17.91	13.84	12.81	17.21
Spectral Derivates (5)	11.69	*21.98*	16.14	*14.09*	*19.03*
MethodOfMoments (6)	13.26	16.42	12.77	11.57	14.80
LPC (8)	13.41	17.92	15.94	11.97	16.19
SSD (10)	13.76	**27.41**	**27.07**	**15.06**	**20.06**
RH (11)	12.38	17.23	12.46	10.30	13.41

In particular, we provide the following size partitions:

- "Traditional" splits into training and test sets, with 90%, 80%, 66% and 50% size of the training set, applying stratification of the sampling to ensure having the same percentage of training data per class, which is important for minority classes.
- A split with a fixed number of training samples, equally sized for each class, with 2,000 and 1,000 samples per class for the genre and style data sets, respectively. This excludes minority classes with less than the required number of samples.

Finally, we apply stratification on other criteria than just the ground truth class, namely:

- Splits into training and test sets with an artist filter, i.e. avoiding to have the same artist in both the training and test set; both stratified and non-stratified sets are provided
- As above, but with an album filter, i.e. no songs from the same album appear in both training and test set, to account for more immediate production effects
- As above, but with a time filter, i.e. for each genre using the earlier songs in the training set, and the later releases in the test set.

Full details on the results for predictions for the different tasks outlined above are available on our website. In this paper, we discuss the results of a musical genre classification experiment on the MSD Allmusic Guide Style Dataset (MASD) with a frequently-used 2/3 training and 1/3 test set split.

Table 5 shows classification accuracies obtained with five different classifiers using the WEKA Machine Learning Toolkit [14], version 3.6.6. Specifically, we employed Naïve Bayes, Support Vector Machines (polynomial kernel with exponent 1), k-nearest Neighbours with k=1, a J48 Decision Tree, and Random Forests, with the default settings. Due to space limitations, we selected the most interesting of the feature sets. The number in parentheses after the feature set name corresponds to the number given in Table 2. Bold print indicates the best, italics the second best result per feature set (column-wise).

For this classification task, we have 25 categories, for which the biggest "Pop Indie" accounts for 6.60% of the songs, which is thus the lowest baseline for our classifiers. It can be noted from the results that the jMIR MFFC features provide the best results on the Naïve Bayes classifier, followed by the jMIR low-level spectral features. However, all results on this classifier are just roughly twice as good as the baseline identified above, and low in absolute terms. Better results have been achieved with Support Vector Machines and k-NN classifiers, on both the Statistical Spectrum Descriptors achieve more than 27% accuracy. Also on the other two classifiers, Random Forests and Decision Trees, the SSD feature set is the best, followed by either the derivatives of the jMIR spectral features, or the jMIR MFFC implementation.

6. CONCLUSION AND FUTURE WORK

Benchmarking is an important aspect in experimental sciences – results reported by individual research groups need to be comparable. Important aspects of these are common platforms to exchange these results, and datasets that can be easily shared among researchers, together with a set of defined tasks. The MIR community has traditionally suffered from only few (and small) data collections being available, also complicated by stringent copyright laws on music. Recently, the publication of the Million Song Dataset has aimed at alleviate these issues. The dataset comes with associated metadata and a basic set of features extracted from the audio. Other modalities such as lyrics have subsequently been provided for (parts of the) collection.

To increase the usefulness of the dataset, we presented a wide range of other features extracted from the audio signals, and enabled musical genre classification tasks by providing a ground-truth annotation to a significant part of the dataset. To foster exchange between different researchers, we defined a number of tasks by providing standardised splits between training and test data.

Our goal is to create a collaborative research environment for sharing data and adding new features (by inviting other researchers to submit their algorithms), also for other data sets besides the MSD. We will extend the features provided also by features for each short segment of the audio analysed, similar to the Echonest features currently available for the MSD, which will allow for time-based analysis over a song. The platform shall also allow sharing of results. This is an important aspect in experimental research, as researchers normally know well how to tune their own algorithms and to optimise parameters to achieve better results – but when they utilise other algorithms for comparison, we often simply apply the default parameter settings, which does not create a realistic baseline Such a collaborative platform will allow fairer comparisons, relieving researchers from the need to run all permutations of feature extractions and settings, and will enable moving towards evaluation platforms as described in [1].

7. REFERENCES

[1] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proc. of the ACM Conf. on Information and Knowledge Management*, CIKM '09, New York, USA, 2009.

[2] James Bergstra, Re Lacoste, and Douglas Eck. Predicting genre labels for artists using freedb. In *Proc. of the Int. Conf. on Music Information Retrieval*, 2006.

[3] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proc. of the Int. Conf. on Music Information Retrieval*, 2011.

[4] Dave Datta. Managing metadata. In *Proc. of the Int. Conf. on Music Information Retrieval*, Paris, France, October 2002.

[5] J. Stephen Downie, Joe Futrelle, and David K. Tcheng. The international music information retrieval systems evaluation laboratory: Governance, access and security. In *Proc. of the Int. Conf. on Music Information Retrieval*, Barcelona, Spain, October 10-14 2004.

[6] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. on Audio, Speech & Language Processing*, 14(5):1783–1794, 2006.

[7] Michel Jubb, Alma Swan, and Sheridan Brown. To share or not to share. Publication and Quality Assurance of Research Data Outputs. Technical report, Research Information Network, 2008.

[8] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. pages 261–268. 2008.

[9] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. of the Int. Conf. on Music Information Retrieval*, pages 34–41, London, UK, September 11-15 2005.

[10] Thomas Lidy, Carlos N. Silla, Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso A. A. Kaestner, and Alessandro L. Koerich. On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing, structuring and accessing non-western and ethnic music collections. *Signal Processing*, 90(4):1032–1048, 2010.

[11] Cory Mckay. *Automatic Music Classification with jMIR*. PhD thesis, McGill University, Canada, 2010.

[12] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[13] George Tzanetakis and Perry Cook. Marsyas: a framework for audio analysis. *Organized Sound*, 4(3):169–175, December 1999.

[14] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations, 1999.