Improving Audio Retrieval through Content and Metadata Categorization

Sanjeel Parekh

Master's Thesis UPF / 2015

DIRECTOR DE LA TESI

Dr. Xavier Serra and Dr. Frederic Font

Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Audio content sharing on online platforms has become increasingly popular. This necessitates development of techniques to better organize and retrieve this data. In this thesis we look to improve audio retrieval through content and metadata categorization in the context of Freesound. For content, we focus on organization through morphological description. In particular, we propose a taxonomy and thresholding-based classification approach for loudness profiles. The approach can be generalized to structure information about the temporal evolution of other sound attributes. To this end, we also discuss our preliminary findings from extension of this methodology to pitch profiles. On the other hand, metadata systematization has been approached through a topic model known as the Latent Dirichlet Allocation (LDA). Herein automatic clustering of tag information is performed to achieve a higher level representation of each audio file in terms of 'topics'.

We evaluate our approach for both the tasks through several experiments conducted over two datasets. This thesis finds immediate application in online audio sharing platforms and opens up several interesting future research avenues. Specifically, evaluation indicates that our methods can be immediately applied to improve Freesound's similarity and context-based search. Moreover, we believe our work on content categorization makes it possible to include an advanced content-based search facility in Freesound.

Acknowledgements

I would like to thank my supervisors Xavier Serra and Frederic Font for invaluable guidance, regular reviews, immense support and patience. A special thanks to all Freesound team members for sharing with me their knowledge and experience and providing key suggestions on my work. I want to extend my gratitude to all SMC professors and researchers at MTG for vital insights, knowledge and regular discussions regarding sound and music computing. I want to thank all my SMC classmates and friends for discussions and indispensable help during the master's course.

Finally, my sincerest thanks go to my family and UPF for making this unforgettable journey possible!

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

There is an exponential increase in the amount of annotated multimedia content on the internet. This requires development of sophisticated techniques to classify, index and retrieve this content for better navigation and storage. In this thesis we focus on *Freesound*[1] which is an online database of sounds where audio clips are shared by users under the creative commons license [Font et al., 2013]. This database, which is now used by more than four million users continues to expand each day. Each uploaded sound in the database is accompanied by a set of tags and description. Currently, users can browse through sounds using tags and content-based audio similarity search (or query-by-example). Freesound allows for similarity search through content descriptors, however it is not optimized. Some mid/high level descriptors are missing such as those representing the evolution of sound dynamics/pitch etc. and the tag information is not utilized. As a consequence, the search results are often not relevant. This makes conducting structured content search a persisting problem.

1.1 Motivation

The general motivation for this thesis arises from the need to structure content and metadata present in Freesound. In particular, we consider the following problems:

1. The current retrieval system *does not utilize any perceptual criteria or information about sound's temporal evolution*. As a result we are faced with two primary difficulties:

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

   - Audio search engines, like Freesound contain many abstract sounds which are difficult to access through a text-based search. For instance, for the class of sound effects, where a sound might lack a source or is not adequately described in words, retrieval would be difficult without a perception based advanced search. In general, some sound engineers or artist also tend to have a template of ”how the sound should be ?” in terms of its loudness or pitch profile. We believe that giving people the ability to filter sounds based on such criteria would help improve the retrieval results and user experience.

   - Users are often presented with irrelevant similarity search results because it disregards the temporal evolution of several perceptually relevant features. Refining the results of this search based on such criteria could help improve the retrieval quality.

2. The *tags are noisy and unstructured*. This is a common problem faced by manual tagging systems like that of Freesound. In this direction, we wish to categorize tags ’meaningfully’ by determining a concise representation and subsequently evaluate its use for retrieval.

1.2 Aim of this thesis

In this thesis our primary objective is to organize unstructured data in Freesound through audio content and metadata categorization for better retrieval. Specifically, for organizing content we focus on a taxonomy based on morphological description and similarly, look for higher level semantic representation for the associated metadata (refer to Fig. 1.1). For the former task, we focus on analyzing the sound effects (SFX) class in Freesound. The SFX class includes sounds from a very broad range, for example, digitally generated glitches, sirens, foley sounds, modified instrument samples and ambient sounds. We believe this would provide us with a heterogeneous subset of sounds which has a potential for being better characterized by their own internal characteristics than by their source of generation.

1.3 Thesis Structure

The thesis is organized as follows:

- Chapter 2 - Review of literature on relevant topics such as morphological description, content-based audio retrieval systems and topic models

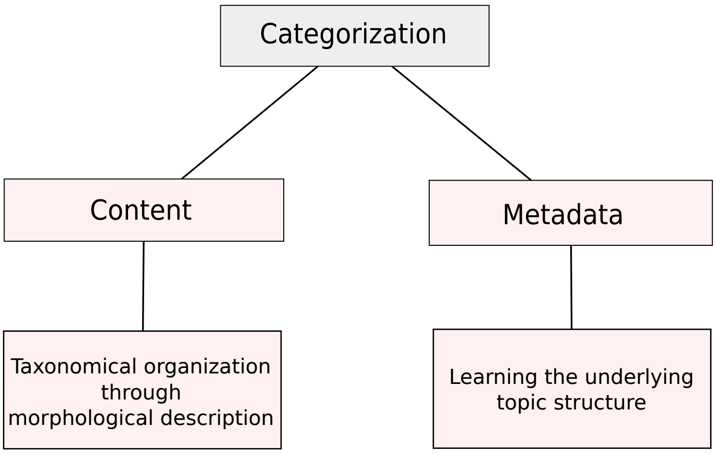

Categorization
Content
Metadata
Taxonomical organization through morphological description
Learning the underlying topic structure

Figure 1.1: Thesis Outline

- Chapter 3 - (i) Methodology for morphological description of content based on Schaeffer's typo-morphology [Schaeffer, 1966] (ii) Related experiments and discussion

- Chapter 4 - Provides technical details for Latent Dirichlet Allocation together with the evaluation experiments and their discussion

- Chapter 5 - Summary and Future work

Chapter 2

STATE OF THE ART

In this chapter we review literature on concepts and algorithms relevant to this thesis. In the sections that follow, we investigate morphological description, content-based audio retrieval systems and topic models.

2.1 Morphological Description

In this section, we discuss the perceptual attributes and their taxonomical organization in the context of Schaeffer's seminal work Traité des objets musicaux [Schaeffer, 1966]. We begin with a brief overview of the primary attributes of sound: loudness, pitch and timbre. This is followed by a discussion on taxonomical organization for characterizing aforementioned perceptual attributes.

2.1.1 Sound Attributes

1. **Loudness** - American National Standards Institute (ANSI) defines loudness as *that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud*. This percept is dependent upon both level and frequency. Loudness has been studied in terms of the equal loudness curves where a 1kHz tone is used as a reference. A 1kHz tone at 40dB SPL is said to produce a loudness level of 40 *phons*. Measurements are made through loudness matching experiments. Thus, any sound perceived as loud as the reference tone also has a loudness level of 40 phons.

   Loudness' dependence on critical bandwidth is a crucial one and can be observed using complex sounds. Herein the loudness increases additively only when the energy is spread across critical bands. Keeping this in view, many loudness computation methods follow the steps expressed in Fig. 2.1 [Timoney et al., 2004]. Here the frequency decomposition would be in relation with the critical bands. Perhaps the most well-known model is the one proposed by Zwicker [Moore and Glasberg, 1996]. In this work we use a similar approach for loudness computation (discussed in chapter 3).

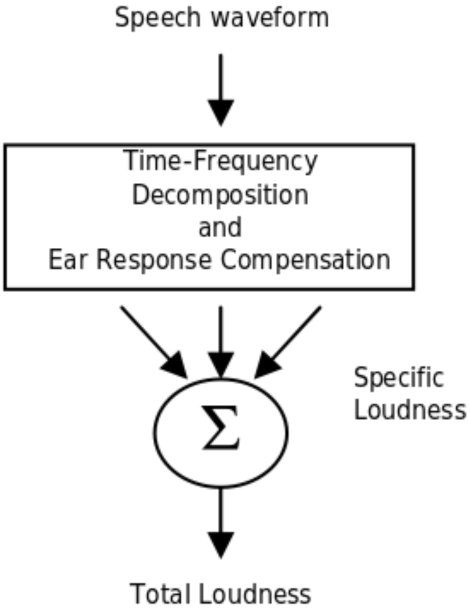


Figure 2.1: Loudness Computation [Timoney et al., 2004]

2. **Pitch** - According to ANSI, pitch can be defined as *that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high*. The physical correlate for pitch is the fundamental frequency denoted by f_0. Hartman relates the two when he says *sound has certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude* [Hartmann, 1996]. However, the 'missing fundamental' phenomena is very well known, where pitch is perceived even in the absence of the fundamental.

   Various methods have been proposed for the computation of fundamental frequency, both in time and frequency domain. Algorithms for monophonic pitch estimation rely on the zero-crossing rate, autocorrelation function (in time or frequency), cepstrum, spectral pattern matching (two-way mismatch) and auditory model based computations. An alternative of the autocorrelation method, popularly known as YIN was introduced in 2002 [De Cheveigné and Kawahara, 2002]. The problem of f_0 estimation in polyphonic signals is especially hard. However, several iterative and joint multipitch estimation methods have been proposed [Klapuri, 2005].

Despite the existence of many algorithms the problem poses difficulty due to several factors such as quasi-periodicity, noise, presence of temporal variations and ambiguous events.

3. **Timbre** - Timbre is the most difficult of these attributes to define and characterize primarily because of its dependence on multiple factors. ANSI defines timbre as, *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*.

   To study features which affect timbre early studies used dis-similarity judgement between pairs of sound and multidimensional scaling analysis to propose timbre spaces [Herrera-Boyer et al., 2003, McAdams et al., 1995]. From Fig. 2.2 we see that the space comprises of rise time, spectral centroid and spectral flux. However, its perception is dependent upon the joint effect of a wide range of attributes namely shape of the temporal and spectral envelope, onset dynamics, time variation of the spectrum [Shamma, 2003, Schedl et al., 2014]. Researchers have attempted to capture these dependencies by extracting various features [Schedl et al., 2014, Peeters et al., 2011, Herrera-Boyer et al., 2003]. List of commonly extracted features includes spectral centroid (correlated with the brightness of sound), Mel frequency cepstral coefficients (MFCC), spectral flux, log-attack time, temporal centroid. For a detailed list of timbre descriptors please refer to [Schedl et al., 2014, Peeters et al., 2011].

2.1.2 Taxonomical Organization

A sound can be described in terms of the above stated perceptual characteristics or based on its source of generation. Though the source-centric description is important, online search engines host a large number of sounds which are either digitally generated or have no clearly identifiable source. Thus a classification in terms of perceptual traits (or morphological characteristics) would provide a more generic description for any kind of sound. For instance, the class of sound effects would be better characterized through perceptual attributes than by their

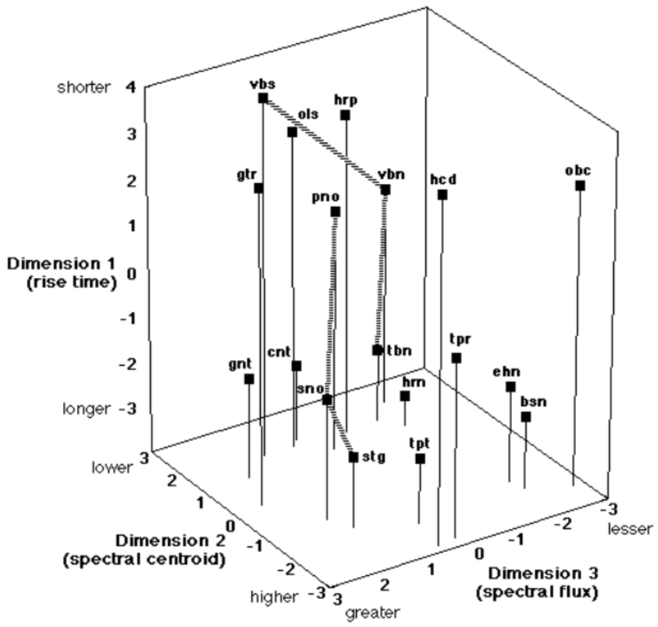


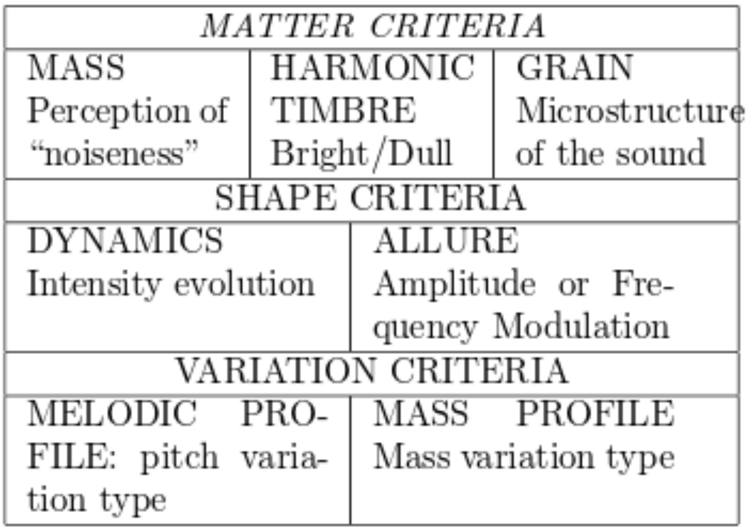
Figure 2.2: Timbre Space [McAdams et al., 1995]

source. In this context, we discuss Schaeffer's work [Schaeffer, 1966] on typomorphology which has been utilized to build taxonomies for sound indexing and retrieval [Ricard and Herrera, 2004, Peeters and Deruty, 2010].

Schaeffer defines causal, semantic and reduced listening as three perspectives for describing a sound. *Causal* refers to recognition of the sound's source, *semantic* to identifying the meaning attached to a sound and *reduced* points to the description of a sound regardless of its cause or meaning. From the latter comes the concept of a *sound object* which is defined as a sound unit perceived in its material, its particular texture, its own qualities and perceptual dimensions [Chion, 1983]. Schaeffer proposes to describe these *objects* using seven morphological components, grouped into three 'criteria'. Fig. 2.3 concisely presents this with a short description of each of them. As indicated by the three clusters in the figure, the matter/form (or shape) pair are central to Schaeffer's morphological taxonomy. As described in [Chion, 1983],

> *Matter* is what persists almost unchanged throughout its duration, it is what could be isolated if it were immobilized, so that we could hear what it is at a given moment of listening.
> *Form* is the course which shapes this matter in duration, and perhaps makes it evolve.

The *variation criteria* comes about when both form and matter vary. Several descriptors have been explored in literature for quantification of these constructs and the given morphological components. Table 2.1 gives a summary of the suggested scheme from two major works in this area. However, a complete list of such descriptors and their representation for various models still remains to be studied in detail.



MATTER CRITERIA		
MASS Perception of "noiseness"	HARMONIC TIMBRE Bright/Dull	GRAIN Microstructure of the sound
SHAPE CRITERIA		
DYNAMICS Intensity evolution	ALLURE Amplitude or Frequency Modulation	
VARIATION CRITERIA		
MELODIC PROFILE: pitch variation type	MASS PROFILE Mass variation type	

Figure 2.3: Schaeffer's morphological criteria [Cano et al., 2004]

We must emphasize that the concept of a sound object is an abstract one. It is difficult to give a precise definition. For completeness, we also mention here Gaver's taxonomy [Gaver, 1993] which is better suited for the class of environmental sounds. His classification is based on production of sounds which can be categorized as aerodynamic sounds, liquid sounds and sounds due to vibrating objects. This organization is from the ecological and physical perspective. For any non-ecological sound such an organization would pose difficulties.

For the purpose of this thesis, our primary interest is the exploration of attributes which support reduced listening. In the present work, we base our taxonomical organization on loudness (dynamics) and pitch profiles. We use [Peeters and Deruty, 2010] and [Ricard, 2004a, Ricard and Herrera, 2004] as key references. We propose a few new features to capture certain additional characteristics of these profiles. Chapter 3 provides the details of our implementation and approach.

Morphological Criteria	Types [Ricard and Herrera, 2004]	Types [Peeters and Deruty, 2010]
Dynamic Profiles	Unvarying Varying: -Impulse -Iterative (several transients) -Crescendo -Decrescendo -Delta (Crescendo- Decrescendo) -Other	stable impulsive - ascending descending ascending/descending -
Pitchness	Pitched Complex Noisy	-
Pitchness Profile	Unvarying Varying	
Melodic Profiles	Unvarying Varying: -Continuous -Stepped (several transients)	stable up down up/down down/up
Complex Iterative	-	Non-Iterative Iterative: -Grain -Repetition

Table 2.1: Morphological Description Scheme as given in [Peeters and Deruty, 2010] and [Ricard and Herrera, 2004]

2.2 Content-based Audio Retrieval Systems

We discuss next some important content-based audio retrieval systems proposed in literature. In the past, researchers have tackled content-based audio retrieval through both, purely content-based and hybrid (content+tags) approaches. In subsequent sections we investigate various algorithms proposed under both of these approaches. Table 2.2 provides a brief overview of audio retrieval systems.

2.2.1 Content-based Approaches

The general methodology followed by approaches under this category is to first extract and process relevant features from audio and then define a similarity measure over them to retrieve sounds acoustically similar to a query sound.

An early work in this area is that of Foote, who proposes a template matching approach using tree-vector quantizer [Foote, 1997]. Though, the method scales well its major drawback is that it does not take into account any perceptual criteria. Moreover, the experiments are carried out over a limited category of sounds. Given the variety in the sound effects category such methods would not perform well. Following from this basic pattern matching paradigm, a content-based audio retrieval system known as *Soundspotter* was proposed by Spevak et al. [Spevak and Favreau, 2002]. The system allows the user to search for similar instances of any specific part of an audio clip. The general scheme of the system is as shown in Fig. 2.4. The system relies on pattern matching algorithms for retrieval. Several methods such as the dynamic time warping, histogram matching, string matching and trajectory matching using self organized maps (SOM) have been discussed. While dynamic time warping can handle vectors of different length that is not the case for trajectory matching.

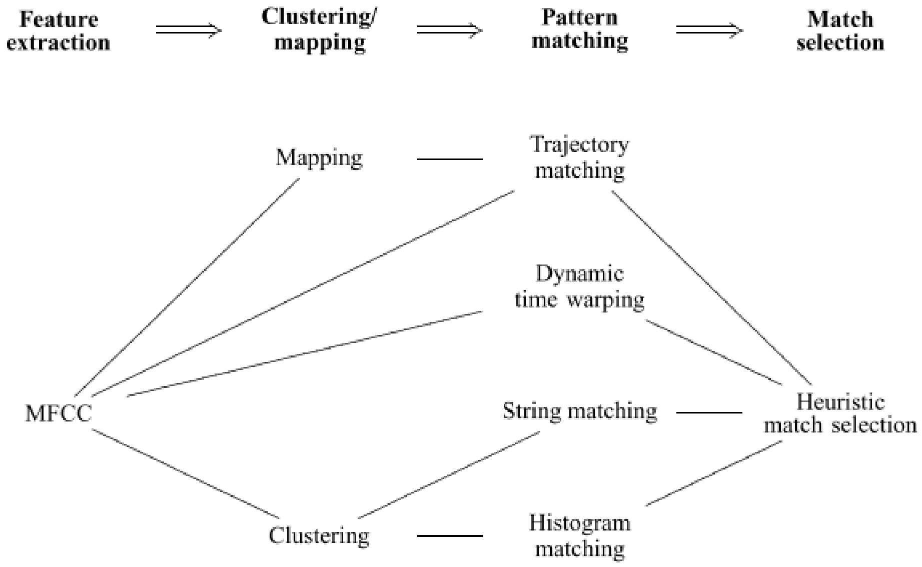


Figure 2.4: Soundspotter [Spevak and Favreau, 2002]: System overview

In 2005, a dominant feature vector method for audio similarity was proposed wherein a new similarity measure was shown to be better than the conventional KL divergence, L_2 norm and Bhattacharya distance [Gu et al., 2005]. This method essentially computes the eigen-decomposition of the feature vector covariance matrix. They report better results than those obtained when using mean and variance of features. The similarity is computed as the weighted sum of the similarity between their dominant feature vectors.

Helen and Lahti [Helen and Lahti, 2006] propose a HMM and feature histogram based approach. In the former, separate models for the example and background are made. Subsequently, the likelihood of sounds belonging to each of those models is computed. Evidently, the ones with the higher likelihood for the example model are considered to be similar. The primary problem with this approach is the lack of instances for training the example model. For feature histogram-based approach, first, the quantization levels are estimated based on the features calculated from the whole dataset. Each audio clip is then represented as a feature histogram and distances are computed based on a threshold. The major downside of this method is that it disregards the temporal variations. Moreover, each of these algorithms rely on a certain set of pre-defined features which might adversely affect their applicability to any general audio sample which is not characterized by the feature set. To overcome this problem perceptual coding and compression based similarity measure was proposed by Helen and Virtanen [Helén and Virtanen, 2007b]. An overview of this unique approach is as given in Fig. 2.5. Herein the normalized compression distance is used as a similarity measure where the idea is to first compare the compression ratios achieved on compressing two files separately and jointly. It is a method with practical utility, however for more specific problems one might lose information which is otherwise gained from specific feature vectors.

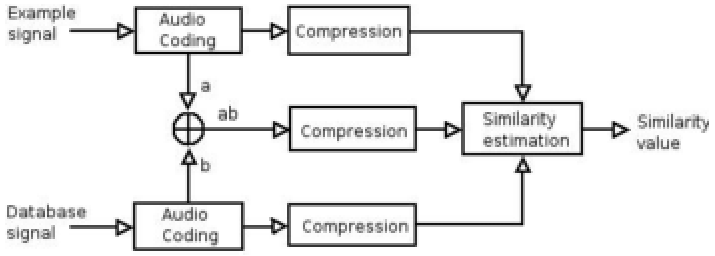


Figure 2.5: Similarity measure based on perceptual coding and compression [Helén and Virtanen, 2007b]: System overview

Another approach is to model feature vectors extracted from an audio probabilistically and defining similarity metric as the distance between feature distributions. Helen et al. [Helén and Virtanen, 2007a] compute euclidean distance between gaussian mixture models (GMMs) trained over the feature vectors from two audio clips. Probability-based similarity measures have also been propounded [Virtanen and Helen, 2007]. The techniques are shown to be successful for audio content. Cross-likelihood ratio tests are introduced for comparing the distances based on probabilistic models.

In this thesis, we work in the context of *freesound.org* which also provides a content-based similarity search facility. Here a kNN search is performed over PCA features extracted using the *Essentia*[1] [Bogdanov et al., 2013] framework. The euclidean similarity measure is used. The features consist of statistics computed over various low-level features. Thus the information embedded in temporal variations is lost.

2.2.2 Hybrid Approaches

Approaches utilizing both content and textual information tackle the problem of audio retrieval and auto-tagging simultaneously. The general scheme has been to define two spaces, namely acoustic and semantic and then discover relations or correspondences between the two. This also enables such approaches to retrieve acoustically similar sounds through text searches. Most hybrid approaches have primarily tackled this issue.

Turnbull et al. deal with the problem of music retrieval through text search by determining a GMM-based acoustic space distribution over each word in the vocabulary [Turnbull et al., 2008]. This helps them retrieve and annotate songs with meaningful words. They propose the use of weighted mixture hierarchies expectation maximization for parameter estimation, which is a scalable approach. In order to show that their method is general, they perform this task specifically for the case of sound effects. They use the bag-of-feature vector representation for the audio. This work does not take into account the temporal dependencies of the audio data and uses the same set of features to represent all types of audio. Moreover, the problem of sparse and noisy tagging like in real world data has not been tackled.

Slaney proposed a clustering based GMM approach where, for each query, probability over clusters in the acoustic space is computed [Slaney, 2002]. The proposed approach is not scalable and its use of GMM disregards the temporal

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

variations. Hoffman et al. use a probabilistic generative model called codeword bernoulli average to determine the joint probability distribution function for tags and content [Hoffman et al., 2009]. The model assumes that the tags are conditioned on the audio content. Here, unlike other models, the bag-of-codewords approach is used for feature representation. This method is specifically tested for music auto-tagging and retrieval.

Some works have utilized simpler schemes such as using k-NN to perform audio annotation and classification [Cano and Koppenberger, 2004, Cano et al., 2005]. One nearest-neighbour decision rule is used where any unknown sound borrows its tags from its nearest neighbour. Normalized manhattan distance of the features is used as a similarity measure. It is reported in [Cano et al., 2005] that standard deviation normalized euclidean distance does not work well in this setting. It is also illustrated that contextual information helps disambiguate the retrieval results as two recordings from different sources might sound the same perceptually. These works consider unambiguously labelled datasets, however in the case of Freesound the context information is noisy. In addition, scalability of such a system to a large dataset is also an issue.

Recently, Mesaros et al. proposed an integrated similarity measure i.e. a weighted linear combination of similarities in content and text space [Mesaros et al., 2013]. The measure given by $C = (1-w)A + wS$ where A, S and w represent content similarity matrix, semantic similarity matrix and weighting coefficient respectively. Though the objective evaluation metrics seem to give satisfactory retrieval results, the method overestimates it's recall metric. Also, no perceptual(or listening) tests have been carried out. This metric has also been applied to Freesound's similarity search [Dimitriou, 2014].

2.2.3 Feature Selection and Representation

We have already discussed the proposed frameworks for the task of audio retrieval. Using any model requires us to determine appropriate description and representation for both content and text. In this section we delineate the features and approaches researchers have taken to represent sounds and annotations. This essential step is often overlooked in systems where thousands of descriptors are computed and fed to a dimensionality reduction algorithm like *Principal Component Analysis*. While this approach might work for certain class of problems, in sounds, where perception plays a crucial role, careful feature selection is necessary.

1. **Content**

The MFCCs were used as feature vectors by all the works discussed previously. The following additional features have also been tested:

- Spectral Descriptors - spectral spread, spectral flux, spectral centroid, spectral flatness, harmonic ratio, spectral kurtosis, skewness, barkband energy
- Temporal Descriptors - zero crossing rate, energy, power variance

In order to take into account the temporal variations of the feature vectors even their mean and the variance are computed. The representation used by each work varies depending upon their use of a particular framework. Some works, like those of [Foote, 1997, Slaney, 2002, Helen and Lahti, 2006, Helén and Virtanen, 2007a, Turnbull et al., 2008, Hoffman et al., 2009, Mesaros et al., 2013] disregard the temporal variation due to their modelling based on feature histograms, bag-of-feature-vectors, bag-of-codewords and GMM-based probability distributions.

2. **Tags and Description**

Gathering reliable and non-sparse semantic data is another persisting problem. So far, well-labeled datasets have been utilized by researchers. However, in a real world case, like that of freesound.org this problem must be dealt with. WordNet has emerged as a popular tool for systematic semantic analysis [Cano et al., 2005, Mesaros et al., 2013] where the hierarchy provides a means to extend the tags meaningfully. Other techniques include tag propagation and expansion as proposed in [Sordo, 2012, Font et al., 2014].

For context the usual approach is to use a vector space model where each element of the vector is indicative of the presence of a particular tag. The 'presence' could be defined using 'soft' (weighted) or 'hard'(binary) constraints. Such representations are particularly important for approaches requiring to learn joint distributions over multimodal data. Hoffman et al. use binary representation for annotations given by $y_{j,w} \in \{0,1\}$ which indicates the presence of a tag w for a song j in their database. Another approach is to use an annotation vector for each song given by $y = (y_1, y_2, \ldots, y_n)$ where y_i for $i \in \{1, \ldots, n\}$ represents the semantic weights attached with each word [Turnbull et al., 2008]. Semantic weights are computed using an average statistic of votes given by the annotators.

Finally, a commonly used representation utilizes the $tf - idf$ statistic computation for each word.

	Study	Framework/Model	Similarity Measure
Content-based Approaches	[Foote, 1997]	Template Matching	L_2 norm, cosine distance
	[Spevak and Favreau, 2002]	Pattern Matching Algorithms	L_2 norm, Edit distance
	[Gu et al., 2005]	Dominant Feature Vectors	Weighted Inner product Sum
	[Helen and Lahti, 2006]	HMM	Likelihood
		Likelihood Ratio Test	Likelihood
		Feature Histogram	L_1, L_2, L_{inf} norms, KL Divergence
	[Helén and Virtanen, 2007b]	Perceptual Coding and Compression	Normalized Compression Distance
	[Helén and Virtanen, 2007a]	GMM	L_2 norm
	[Virtanen and Helen, 2007]	HMM and GMM	Cross-Likelihood Tests
Hybrid Approaches	[Slaney, 2002]	MPESAR	Probability-based (Model-dependent)
	[Cano et al., 2005]	k-Nearest Neighbour	Normalized Manhattan Distance
	[Turnbull et al., 2008]	GMM	Probability-based (Model-dependent)
	[Hoffman et al., 2009]	Codeword Bernoulli Average	Probability-based (Model-dependent)
	[Mesaros et al., 2013]	GMM	Linear Weighted Combination

Table 2.2: Overview of Audio Retrieval Systems

2.3 Topic Models

As stated in [Blei, 2012]:

> *Topic models* are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.

This is extremely useful in organizing datasets for better retrieval and navigation. Keeping in view the scope of this thesis, in this section we discuss the general idea behind topic models, give an introduction to latent dirichlet allocation (LDA) and discuss several applications to retrieval and data navigation.

The latent dirichlet allocation is the most popular and the simplest topic model [Blei et al., 2003]. The basic intuition behind LDA and topic models in general is best explained through Fig. 2.6. LDA assumes that any document is generated from a set of topics. For instance, a news article might be composed of several topics namely sports, politics and entertainment. More formally, LDA models documents as a mixture of topics where each topic is represented as a distribution over words. It is important to emphasize here that these models discover these underlying topics *automatically*. LDA appeared in 2003 as an improvement over the probabilistic latent semantic indexing (pLSI) [Hofmann, 1999]. It improved over the following drawbacks of pLSI [Blei et al., 2003]:

- pLSI does not have a natural way of assigning topic probabilities to unseen documents
- pLSI suffers from overfitting

Hence, given that the corpus was generated through the probabilistic generative approach defined by LDA the aim is to determine the model parameters and utilize the posterior distribution for various tasks such as topic visualization, document visualization, similarity search and text-based search [Blei and Lafferty, 2009]. Topic visualization would involve exploring the topic-word matrix and similarly documents can be explored through the topic-document matrix. The application to visualization has been explored for wikipedia articles [Chaney and Blei, 2012]. They also provide a framework that could be used for various other datatypes. For similarity search one could utilize the topic-document matrix as feature vectors for determining similar documents.

LDA has served as the basis for many topic models and several extensions like correlated and dynamic topic models [Blei and Lafferty, 2009] that appeared in the last decade. Several non-probabilistic methods have also been proposed, for example the seminal work on latent semantic indexing (LSI) [Deerwester et al., 1990]

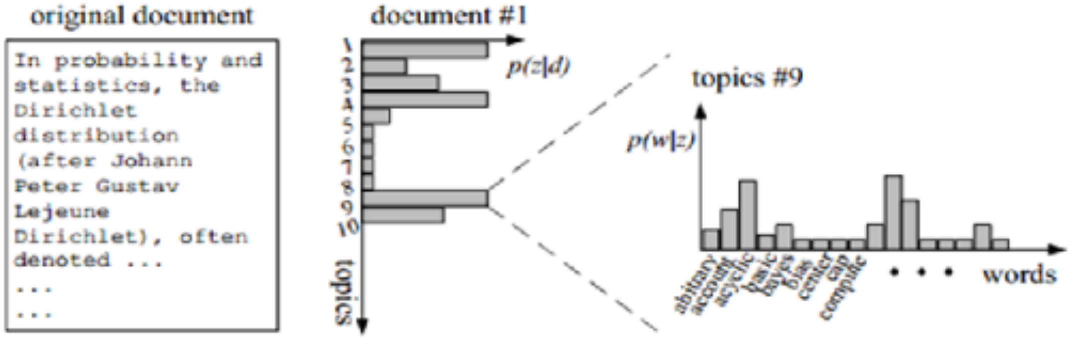


Figure 2.6: Elements of Latent Dirichlet Allocation

and non-negative matrix factorization[Lee and Seung, 2001]. However the modularity and extendability to new data makes generative approaches advantageous.

Chapter 3

CONTENT CATEGORIZATION BASED ON MORPHOLOGICAL DESCRIPTION

In this chapter we establish our framework for sound categorization based on the temporal evolution of its attributes such as loudness and pitch. We test our approach through several experiments including a subjective evaluation.

3.1 Loudness Profiles

Loudness profile can be defined as the temporal evolution of a sound's loudness. Its categorization schema is as shown in Fig. 3.1. We proceed with first classifying the dataset into complex and single events. As indicated in the figure, for single events the loudness curve could belong to one of the following categories:

1. Impulsive
2. Stable
3. Increasing
4. Decreasing
5. Increasing-Decreasing (or Delta)
6. Others

A sound is said to be *impulsive* if it has either a sharp attack or is of a very short duration. It is considered to be of *stable* class if the loudness of the sound does not vary much. It is said to be of *increasing* (or decreasing) category if the loudness increases (or decreases) for significant portion of the sound's duration. Similarly, the sound would be of *delta* class if it is perceived of first increasing and then decreasing loudness. The *others* class would contain sounds which lie in the "confusion" areas. Some of them are even difficult to categorize perceptually. We discuss the definition of each of these classes more formally in Sec. 3.1.2. To categorize audio based on the schema shown in Fig. 3.1 we discuss next our loudness profile modeling and classification methodologies.

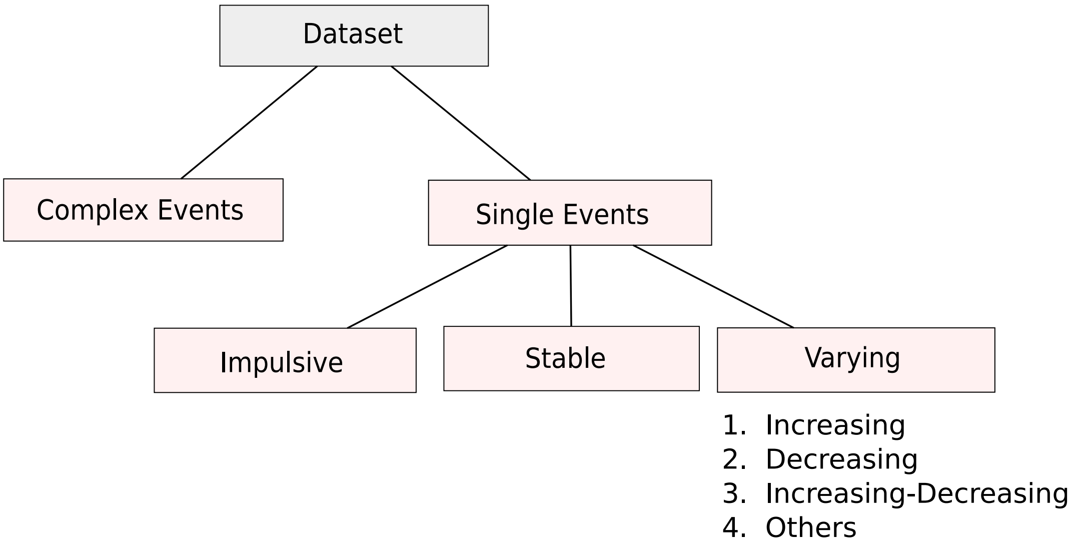


Figure 3.1: Content categorization scheme based on loudness profiles

3.1.1 Modeling Methodology

Each of the different steps involved in profile computation and feature extraction are delineated below. Excluding the addition of step 2, modeling is carried out as in [Peeters and Deruty, 2010].

1. **Loudness Computation** - For each windowed frame of the signal the spectrum is computed and outer-mid ear filtering is performed [Kabal, 2002]. The transfer function for ear filtering is given by eqn. 3.1.

$$A_{db}(f) = -2.184\left(\frac{f}{1000}\right)^{-0.8} + 6.5e^{-0.6(f/1000-3.3)^2} - 0.001\left(\frac{f}{1000}\right)^{3.6} \tag{3.1}$$

   Next, the energy in each bark band, denoted by $E(z,t)$ is obtained. The loudness is then computed as

$$l(t) = \sum_{z} l'(z,t) \text{ where } l'(z,t) = E(z,t)^{0.23} \tag{3.2}$$

In order to smooth the signal, $l(t)$ is lowpass filtered. Its maximum value, l_m is determined and the part of $l(t)$ over 10% of l_m is considered for subsequent stages. The cut-off of the filter was set at 2 Hz after determining its effects on certain post-processing issues. Since we are interested in relative measures, the time axis is normalized for all the sounds.

2. **Complex and Single Event Classification** - Clearly, the profile description would apply to only single sound events. Since the sound effects dataset we use also contains complex events, we describe here our approach for automatically separating complex and single events using the loudness curve computed in the previous step. The process is summarized in Fig. 3.2.

   First, an onset detection function is constructed from the derivative of the loudness profile and subsequently, peaks of this function are detected using a running mean threshold. Any sound with more than one peak is classified as complex. Though we do not cater to cases with soft onsets, from the loudness profile characterization perspective we believe this approach suffices. Onsets are detected using essentia's $Onsets$ function. Hereafter, we only consider the loudness curves for single events.

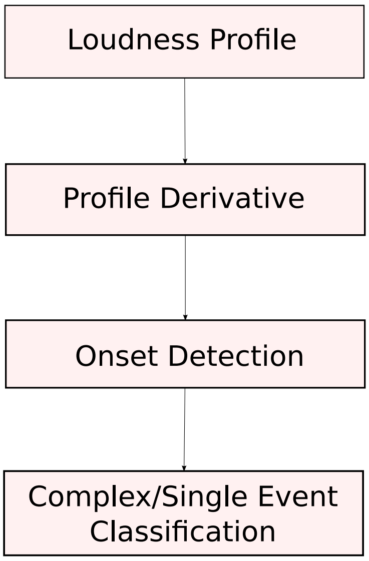


Figure 3.2: Complex-Single event classification process

3. **B-Spline Modeling** - In order to extract meaningful descriptors for our classification we obtain a first-order B-spline approximation (or straight line approximation) for the loudness curve that is continuous at l_m. Since the decay of sounds can be modeled using eqn. 3.3 [Peeters and Deruty, 2010], for a straight line approximation we must express the loudness function in the log-scale.

   $$l(t) = Ae^{-\alpha(t-t_m)} \tag{3.3}$$

   Next, we consider three knot points with co-ordinates $[(t_s, l_s), (t_m, l_m), (t_e, l_e)]$ where the subscripts s and e represent the start and the end of the thresholded loudness curve respectively. At the end of this step we have a straight line approximation for the filtered, thresholded log-scale loudness curve.

4. **Extracted Features**

   - As shown in Fig. 3.4, we extract the following slope and relative duration features from this representation:

     - RD1 - Relative duration given by $t_m - t_s$
     - RD2 - Relative duration given by $t_e - t_m$ or 1-RD1
     - S1 - Slope of the approximation from where if begins (t_s) to the maximum (t_m)
     - S2 - Slope of the approximation between maximum (t_m) and the end point (t_e)

   - We also compute the absolute effective duration at 10% (ED10) and 40% (ED40) i.e. the duration for which the profile is above 10% and 40% of its maximum, respectively. Also, the relative (normalized time axis) effective duration at 80% (ED80) is computed. These features help us classify impulsive and stable sounds.

We propose two additional descriptors in order to take into account the loudness modulations. Both stable and varying class can contain sounds of this nature. In order to extract these features the computation is performed over $m(t)$ which is the signal obtained after subtracting the running mean from the profile curve. Thus, the characterization is done in terms of the modulation rate and extent:

- **Modulation Rate** - The zero crossing rate of $m(t)$ provides a good quantification of this parameter

- **Modulation Extent** - The mean of the standard deviation of $m(t)$

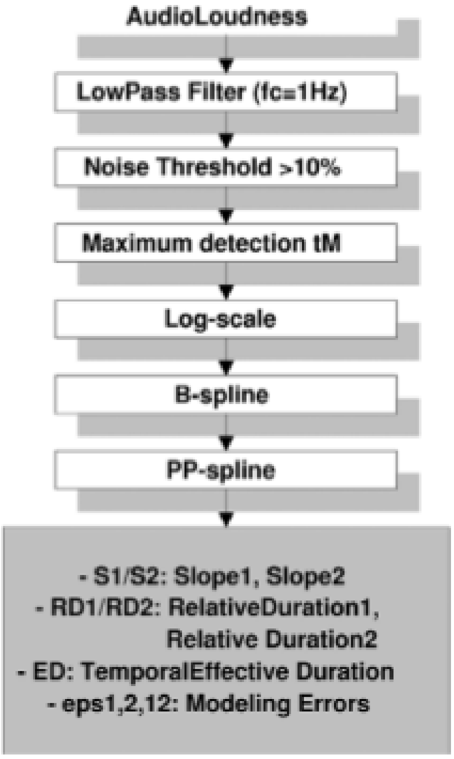


Figure 3.3: Profile Computation Methodology [Peeters and Deruty, 2010]

3.1.2 Thresholding-based Profile Classification

Approaches

In order to achieve the proposed classification one could either look at machine learning schemes or a process of manual classification by defining each class based on criteria one believes would hold true. We opt for the latter, which we call the *thresholding* approach, for the following reasons:

- Several "confusion areas" exist and not all sounds would fall into one of the classes. Hence, we would like the system to be flexible enough to incorporate such cases.

- For assisting users in retrieval tasks we would like to provide them with a facility to control some *meaningful* parameters.

Though we discuss our experiment with a decision tree based rule learning algorithm (in Sec. 3.2), the machine learning approach has the following limitations:

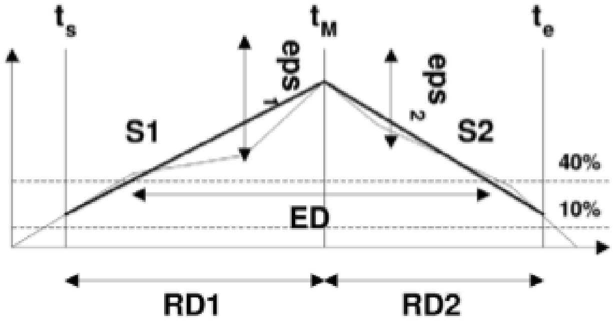


Figure 3.4: Loudness Profile Descriptors [Peeters and Deruty, 2010]

- It is very difficult to create large labeled datasets representative of each class from very diverse unstructured databases like that of Freesound
- Classification by unsupervised learning schemes, like clustering, might result into centroid positions or parameters that seem non-meaningful from the problem's perspective. Moreover, interpreting the results of such an approach will also be difficult

Category Definitions for Classification

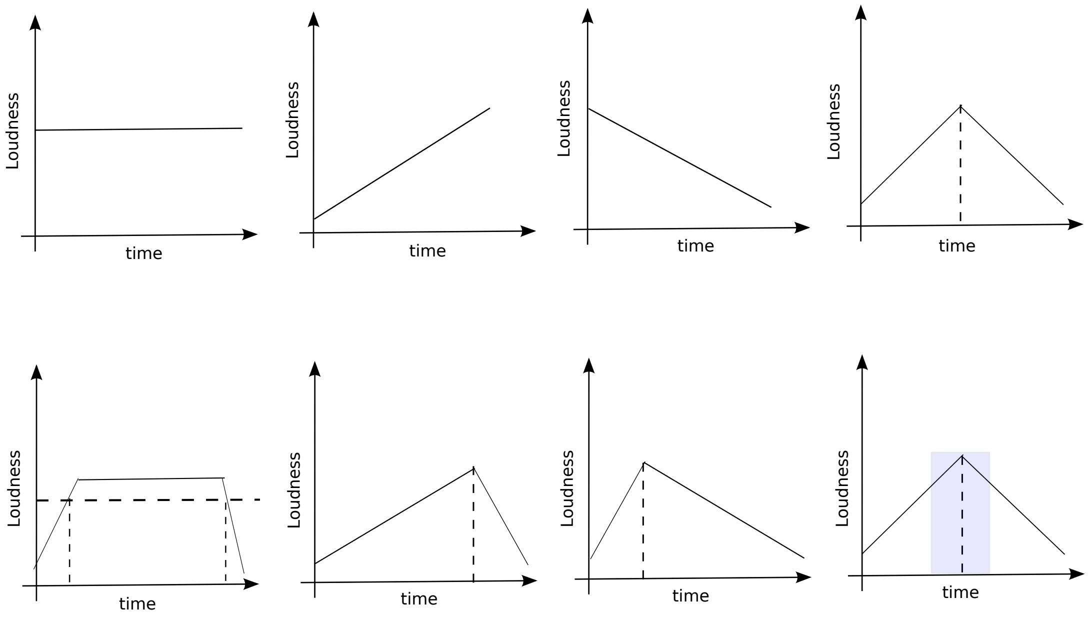
We now discuss the definitions for each of the loudness profile classes from the viewpoint of applying a *threshold*. Fig. 3.5 provides a good visual representation and "rationale" for our approach. Consider for instance the increasing class (Fig. 3.5 B), ideally the sound would increase in loudness for all its duration (first row), however, in reality, the sound's loudness would fall after rising for a 'significant' part of the sound's total duration. Thus, with the thresholding approach we say that a sound would belong to the increasing category if it rises for atleast 70% of its total duration (denoted by a dashed line). In this case, we have set a threshold on the duration for which the sound must rise to be classified as increasing. The other profiles can be understood similarly.

Stated more formally, our aim is to determine the deviation, δ of relevant features from the "ideal case" (Fig. 3.5) such that all the sounds with values above (or below) δ clearly belong to a particular category. For simplicity in defining these thresholds, for all the classes (except impulsive), we consider threshold over a single feature.

1. Impulsive: ED40 $\leq \delta$ or ED10 $\leq \gamma$, where $\delta = 0.25$ and $\gamma = 0.3$

2. Stable: ED80 $\geq 1 - \delta$, where $\delta = 0.3$
3. Increasing : RD1 $\geq 1 - \delta$, where $\delta = 0.3$
4. Decreasing : RD2 $\geq 1 - \delta$, where $\delta = 0.3$
5. Delta: $|RD1 - 0.5| \leq \delta$, where $\delta = 0.1$
6. Others: According to the definitions above, the others class has two components others-increasing: $0.6 < RD1 < 0.7$ and others-decreasing $0.3 < RD1 < 0.4$

We also make the observation that for a thresholding approach we would first need to separate the impulsive and stable class from the set of single events. This is also a result of the features we extract. For instance, after a straight line approximation the features of a stable sound might be very similar to that of a sound from the delta class. This also explains the need for the effective duration features we mentioned in the previous section.



(A) Stable (B) Increasing (C) Decreasing (D) Delta

Figure 3.5: Loudness Profile : The first row displays the ideal case templates for each of the categories, however (as shown in the second row) the real case would almost always have an "increasing-decreasing" profile. The dashed lines in each of the graphs denote the thresholds we must determine

3.2 Experimental Results

3.2.1 Dataset

For this set of experiments we use the **FS-SFX** dataset which was created by downloading content from Freesound using the 'fx' tag as a filter. It was also ensured that all the sounds were less than 10s in duration. It contains the high quality ogg preview, content descriptors and metadata for a total of 5248 sounds. All the content based experiments have been carried out using this dataset. A subset of this data (238 sounds) was manually annotated according to the loudness profile in order to perform a first check on several classification tasks. The details of this reduced dataset, referred to as **SFX-Reduced** are given in Table 3.1.

Class	Number of sounds
Complex	57
Impulsive	53
Stable	28
Increasing	30
Decreasing	36
Delta	34

Table 3.1: SFX-Resuced Dataset Details

3.2.2 Objective Evaluation

1. **Complex/Single Event Detection - SFX-Reduced**

   Through this experiment we evaluate the accuracy of our methodology for complex/single event classification. The parameters were tuned through trial and error over the SFX-Reduced dataset. It was important to have a small number of false positives i.e. to prevent single events from being labeled as complex.

   Critical parameters: Lowpass filter cut-off= 2Hz, mean threshold = 0.5

Results

		Predicted	
		Complex	Single
Actual	Complex	45	12
	Single	28	153

Table 3.2: Confusion Matrix: Complex/Single Event Classification

Class	Number of sounds
Complex	1944
Single	3304

Table 3.3: Complex/single event classificaton for FS-SFX dataset

From Table 3.2 we observe that our system gives reasonably good accuracy by classifying correctly 78.9% (45/57) of complex and 84.5% (153/181) of single events.

The number of false positives depended upon the lowpass filter cut-off. A high cut-off meant that though all the complex sounds were segregated, many single events were also mistakenly discarded. Thus, optimal performance depended on both, the mean threshold parameter of the *Onsets* function and the low pass filter cut-off.

2. **Loudness Profile Classification - SFX-Reduced**

Here we provide a comparison and an objective evaluation of our thresholding based approach. First we present the results of both, the thresholding approach and WEKA's decision tree-based rule learning algorithm PART on the SFX-Reduced dataset. PART algorithm's performance was evaluated with 10-fold cross validation

Parameters: The thresholds were set as described in Sec. 3.1.2. WEKA implementation of the PART algorithm was used where the default parameter settings gave the best accuracy.

Results

		Predicted							
		Imp	Stb	Inc	Dec	Delta	Oth-Inc	Oth-Dec	Cmp
Actual	Imp	**32**	0	6	4	1	2	0	8
	Stb	2	**11**	6	0	0	0	0	9
	Inc	4	0	**21**	0	0	1	0	4
	Dec	2	0	3	**19**	5	0	5	5
	Delta	9	0	6	1	8	2	6	2

Table 3.4: Confusion Matrix: Loudness Profile Classification

		Predicted				
		Imp	Stb	Inc	Dec	Delta
Actual	Imp	**40**	1	2	5	5
	Stb	1	**24**	1	0	2
	Inc	3	0	**22**	0	5
	Dec	4	0	1	**27**	4
	Delta	2	3	7	5	**17**

Table 3.5: PART Algorithm Confusion Matrix: Loudness Profile Classification

- Please note that the sounds were manually chosen (based on perception) to belong to one of the five classes namely impulsive, stable, increasing, decreasing and delta. However, the idea was to test the thresholding system's performance as a whole. Hence, the confusion matrix in Table 3.4 is not square and includes the others (oth-inc and oth-dec) and the complex classes for predictions. As we see, the system does mis-classify 28/181 single events into the complex category. This is not the case for PART algorithm where all the sounds were

Class	Number of sounds
Impulsive	1308
Stable	175
Increasing	508
Decreasing	449
Delta	544
Others-increasing	142
Others-decreasing	178

Table 3.6: Thresholding approach classification for single events over the whole FS-SFX dataset

treated as single events and were to be classified into one of the five categories only. (Table 3.5).

- Though the PART algorithm gives us a classification accuracy of 71.8% it requires *19* rules to do so. Whereas the thresholding approach gives us comparable performance with just *5 rules*. Moreover, with our approach, we have the flexibility to let the user decide on certain meaningful parameters such as relative duration etc.
- For both the approaches, a sound belonging to other categories has been mis-classified into the impulsive class. It is particularly evident for the delta class in Table 3.4. This implies that, for the thresholding approach, using only the effective duration descriptors for the impulsive class is not sufficient. For borderline cases this proves to be detrimental.
- The inclusion of the others class i.e. Oth-Inc and Oth-Dec is positive. Its requirement is especially illustrative in the spread of system predictions for delta class sounds over all the other classes, for both the algorithms.

3.2.3 Subjective Evaluation

For this subjective evaluation our goal was to analyze the utility of our framework for the use case of similarity search. In particular, we compare the performance of the current Freesound similarity search with a modified version of it. For the modified system the similarity search results are obtained after filtering current system's results according to the query sound's loudness profile category.

Experiment Design

Each candidate was presented with retrieval results from two systems for eight query sounds. Each query sound was followed by top 5 results from the two systems presented in separate columns, labeled I and II. For each sound, the system presented in each column was randomized. The task was to carefully go through each query sound along with its results. The candidate was then asked to indicate his/her preference for system in column I or II based on the similarity of its retrieval results to the query sound. They were also provided with a 'No Preference' option, in case they did not find any of the systems to be better than the other. They could listen to each sound multiple times. A screenshot of the online survey is given in Fig. 3.6.

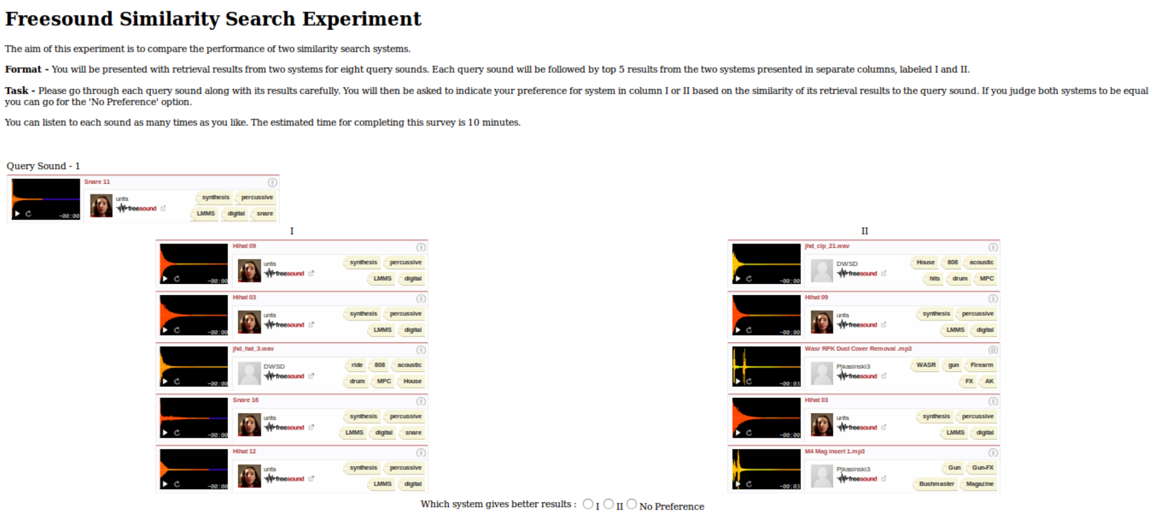
For each user the 8 query sounds were chosen from a pool of 91 sounds selected from the SFX-Reduced dataset. These were sounds which were correctly classified into the five categories namely impulsive, stable, increasing, decreasing and delta by our thresholding based system (refer to Table 3.4). For the modified system, the filters used for refining similarity search results are presented in Table 3.7.

Query Sound Category	Filter
Impulsive	Impulsive
Stable	Stable
Increasing	Increasing + Others Increasing
Decreasing	Decreasing + Others Decreasing
Delta	Others Decreasing + Delta + Others Increasing

Table 3.7: Modified system: Filters for refining similarity search results

Results

Nine candidates participated in this online experiment. We obtained a total of 72 judgements (8 per candidate). Out of these, 25% were 'No preference' judgments. If we discard these cases, we see from Table 3.8 that 72.2% of the



Freesound Similarity Search Experiment

The aim of this experiment is to compare the performance of two similarity search systems.

Format - You will be presented with retrieval results from two systems for eight query sounds. Each query sound will be followed by top 5 results from the two systems presented in separate columns, labeled I and II.

Task - Please go through each query sound along with its results carefully. You will then be asked to indicate your preference for system in column I or II based on the similarity of its retrieval results to the query sound. If you judge both systems to be equal you can go for the 'No Preference' option.

You can listen to each sound as many times as you like. The estimated time for completing this survey is 10 minutes.

Query Sound - 1

Which system gives better results : ○ I ○ II ○ No Preference

Figure 3.6: Online Experiment Interface

	Preference		
Participants	MOD System	FS System	No Preference
1	4	2	2
2	4	2	2
3	5	1	2
4	5	2	1
5	3	1	4
6	3	3	2
7	5	1	2
8	4	2	2
9	6	1	1
Total	39	15	18

Table 3.8: Online Experiment Participant Results - Here each column under the Preference heading represents, for each participant, the number of responses in favor of Modified System (MOD System), Freesound System (FS System) and 'No Preference' respectively

		Preference		
Class	Number of sounds	MOD System	FS System	No Preference
Impulsive	25	17	3	5
Increasing	17	9	5	3
Decreasing	17	7	2	8
Stable	8	4	3	1
Delta	5	2	2	1
Total	72	39	15	18

Table 3.9: Online Experiment Profile-Specific Results - Here each column under the Preference heading represents, for each class, the number of responses in favor of Modified System (MOD System), Freesound System (FS System) and 'No Preference' respectively

judgements were in favor of the modified system. To further validate the performance, we observe in Table 3.8 that for all the candidates, the number of responses in favor of the modified system were always greater than or equal to those in the favor of current Freesound system (note that equality held only in one case). This gives us strong preliminary evidence to claim that the modified version is an improvement over the current Freesound system.

It is also interesting to see Table 3.9 which is an analysis of user preference for systems with respect to the loudness profile of the query sound. The table should be read as follows: for instance, 17 of the 72 responses were for sounds belonging to the increasing class, where 9 responses were in favor of the modified system. We see that the modified system performs particularly well for impulsive and increasing class. However, that is not the case for stable and delta class. Two primary reasons for this could be: (i) We do not take into account the timbral similarity while filtering for loudness profile (ii) The classes are not sufficiently well defined.

3.3 Conclusion and discussion

In this chapter we have described our framework for content categorization on the basis of loudness profiles. Through the proposed thresholding-based approach we pre-define a set of classes in accordance with the previous work carried out in this area [Peeters and Deruty, 2010, Ricard, 2004b]. An objective evaluation of our scheme shows that we achieve performance comparable to machine learning techniques. Moreover, we have the added advantage of control over meaningful parameters such as relative duration and slopes to assist in classification or filtering for retrieval. The online experiment carried out for the use case of similarity search provides evidence for successful application of our technique. As the results indicate, our system was clearly preferred over the current similarity search facility in Freesound. Due to time constraints the evaluation could only be carried out over a small number of participants. To the best of our knowledge this was one of the first such subjective evaluations and also the first study to be carried out on a real-world dataset. This framework is not restricted to SFX and can be generalized to any kind of audio.

In the context of Freesound, an important outcome of this investigation is that such a categorization can be used as a method for automatically generating content-based metadata. For instance, a sound could be labeled as 'stable' and a simple text search could be used for retrieval. Note that we now have the ability to retrieve audio based on perceptual attributes from the text search itself. This might be useful for some systems. However, there are a few limitations of our current approach. We do not take into account the slope features. As a result, we lose some information about the attack or decay leading to confusions primarily within impulsive, decreasing and delta classes. Also, the classification experiment indicates the need to better define the impulsive class.

With some modifications our approach can be extended to include profiles for other sound attributes. We briefly discuss one such extension to pitch profiles which also provides us with first steps towards future work. The pitch profiles were defined as the temporal evolution of fundamental frequency and classified into five classes: stable, increasing, decreasing, delta and inverted delta. A similar modeling approach was followed and class definitions were based on sign of the slope features S1 and S2. Preliminary experiments threw light on the following associated issues:

- Due to the introduction of the inverted delta class it is not possible for us to fix the center knot for B-Spline modeling at t_{max}. Hence, we set it to $t = 0.5$ for all the sounds. This reduces the accuracy of our system
- The FS-SFX dataset contained very few clear examples for each of the aforementioned pitch profile classes. Moreover most of the dataset is made up of complex pitch profiles and noisy sounds. This made conducting evaluation difficult.

Though the idea of profile classification into broad categories finds use in applications such as similarity search, it might seem too restrictive in cases where the user wants to be very specific. It is possible to incorporate such a requirement through our approach by considering the slope and relative duration parameters. Also, the thresholds in our approach can be looked at as *presets*, which can be modified according to the user's need.

In the next chapter we shift our focus towards metadata categorization through latent dirichlet allocation.

Chapter 4

METADATA CATEGORIZATION USING LATENT DIRICHLET ALLOCATION

Having dealt with content in the previous chapter, in this part of the thesis we cater to the objective of organizing and utilizing tag information. We use a popular model known as Latent Dirichlet Allocation [Blei et al., 2003] for extracting and categorizing tag information for each sound in our database in terms of "topics". These topics are learnt in an unsupervised manner from the tags attached with each sound. This chapter is dedicated to understanding the relevant technical details and the intuition behind this technique. Subsequently, in order to establish the usefulness of this approach we run three evaluation experiments. The first two are aimed at demonstrating the ability of this method to automatically group together "similar" words i.e. words/tags coming from the same "topic". The last one illustrates the use of this topic based representation in similarity search.

4.1 Defining the Task

1. **Task** - Given the tags for each sound in our dataset we wish to discover the underlying topics and thus obtain a representation of each sound in terms of the learnt categories.

2. **Input (Text) Representation** - We choose the vector space representation also known as the bag-of-words approach where for a dictionary of N words, each document (in this case sound) is represented as an N-

dimensional binary feature vector, $W = \{w_i\}_{i=1}^{N}$ where $w_i = 1$ if the sound contains the i^{th} word in the dictionary and 0 otherwise.

3. **Output (Topic) Representation** The model would output based on the chosen number of topics, K a feature vector for each sound that captures the proportion in which any particular topic is present.

4.2 LDA Details

Formally, the latent dirichlet allocation (LDA) is a generative probabilistic graphical framework for modeling the process of the generation of a corpus. The graphical model is shown in Fig. 4.1. The generative process for each document in the dataset can be described as follows:

1. Choose $\theta \sim Dirichlet(\alpha)$ - drawing a topic distribution from a uniform dirichlet distribution with parameter α
2. For each word w:
   - Draw a topic $z_n \sim$ Multinomial(θ).
   - Choose a word w_n from $p(w_n \mid z_n, \beta)$, a probability conditioned on the topic, z_n. Here β is a word-topic probability matrix $p(w|z)$

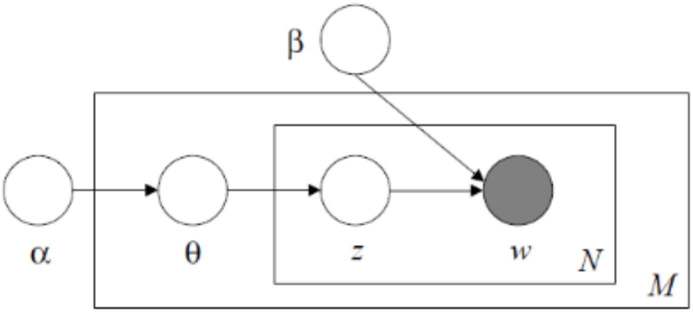


Figure 4.1: Latent Dirichlet Allocation - A generative probabilistic model

The central computational problem is that of computing the posterior i.e. the conditional distribution of the latent variables given the documents. It must be noted that exact inference is intractable. Various approximate methods like gibbs sampling, variational inference and variational EM exist in order to perform inference and parameter estimation for LDA. The reader is referred to [Blei et al., 2003] for a detailed discussion on inference and parameter estimation techniques.

There are two additional advantages of using LDA. The topic representation works as a dimensionality reduction method where we can obtain a "higher" level representation. If looked at from the perspective of matrix decomposition algorithms, it is a kind of PCA analysis [Blei, 2012]. Moreover, we also discover the probability with which each word is associated with any particular topic. The model can be used for both: text-based retrieval of sounds and similarity search based on tags. In the next section we put this technique to use for determining the topic structure and also perform similarity search based on topic feature vectors. For all the experiments python package LDA 1.0.2 implementation is used.

4.3 Experimental Results

4.3.1 Datasets

Here, in addition to the FS-SFX dataset, we also use the **FS-CLS** dataset. FS-CLS contains 120 sounds in total, equally distributed among the following five classes: Soundscapes, Instrument Samples, Speech/Voice, SFX, Music [Dimitriou, 2014]. The human similarity judgements for each of the sounds were also obtained in order to evaluate LDA for similarity search.

4.3.2 Experiments

The following experiments are designed to analyze two goals concerned with tag categorization into topics:

- LDA's ability to discover 'meaningful' topics over Freesound datasets (Experiment 1 and 2)
- Evaluating application of topic-vector representation of each sound to similarity search (Experiment 3)

1. **Topic Extraction - FS-CLS**

   The aim of this experiment is to illustrate the ability of LDA to discover the underlying topic structure. It has been carried out as a sanity check for use with the kind of data we possess. Since we already know the five classes which form the FS-CLS dataset namely Soundscapes, Instrument Samples, Speech/Voice, SFX, Music, we wish to discover them automatically (only using tags).

Parameters: Number of topics, $K = 5$, Iterations $= 1000$

Results

Table 4.1 shows the tags attached with each of the topics extracted by LDA. Clearly, the 5 "topics" present in FS-CLS dataset can be extracted. One must note that LDA requires the number of topics as an input parameter. Thus, had K been set to more than five for this dataset, the model would have further decomposed and similarly, consolidated for a value of less than five.

Topics	Tags
Soundscapes	field-recording, soundscape, ambience, atmos, seoul, wind, korea
Instrument Samples	multisample, people, acoustic, drum, guitar, children, break
Speech/Voice	voice, speech, male, noise, man, talk, deep
SFX	ambient, industrial, percussion, fx, hit, dark, metal
Music	loop, beat, drum, melody, processed, synth, pad

Table 4.1: LDA topic-word relations for 5 topics

2. **Topic Analysis - FS-SFX**

Similar to the topic extraction task for the FS-CLS dataset, here we aim to learn the categories from the sound effects dataset. Since we know that the sounds belong to the class of sound effects, for each sound we discard tags such as sfx, effects, sound-effects etc in order to avoid their influence on the topic-word matrix because of their frequent occurrence.

Parameters: The results are shown for 20 Topics and 1000 iterations

Results

LDA topic-word relations for FS-SFX - Revised

Topic 0: noise, glitch, digital, slices, cuts, step, blip

Topic 1: glitch, circuit, bent, bending, electronic, korg, experimental

Topic 2: acoustic, instrument, mbira, thumb-piano, kalimba, african, audio-triggered-synth

Topic 3: synth, analog, synthesizer, bleep, filter, soundesign, bass

Topic 4: digital, noise, crash, reverb, click, downlifter, roomworks

Topic 5: drum, percussion, hit, industrial, field-recording, ambient, zynaddsubfx

Topic 6: space, dub, reggae, lofi, rasta, siren, jungle

Topic 7: synth, bass, techno, electronic, loop, electro, dance

Topic 8: voice, horror, vocal, sound, experimental, noise, human

Topic 9: ambient, experimental, dark, drone, deep, reverb, pad

Topic 10: dub, rave, electronic, house, dub-step, club, techno

Topic 11: sci-fi, processed, computer, game, noise, robot, synthetic

Topic 12: sound, sample, synthesiser, 4x4-records, snare, analogique, j-lee

Topic 13: game, drop, click, plastic, short, fall, water

Topic 14: house, acid, ping, quality, synth, electronic, door

Topic 15: drum, drums, percussion, acoustic, sound, snare, perc

Topic 16: hit, hardstyle, stab, electro, house, metal, metallic

Topic 17: kick, virus, ti, beat, distortion, long, sample

Topic 18: digital, sound-design, electric, sci-fi, sounddesign, strange, noise

Topic 19: space, alien, sci-fi, futuristic, noise, scifi, gun

- We notice in the box above that the highlighted (in yellow) topics have been well learnt. In other words, the tags clustered under each topic are indeed closely related.
- Certain words occur in multiple topics. Though this can certainly be the case, for us this occurs quite frequently for a few of them. A

possible reason could be that some words are used in multiple contexts for the sound effects class.

3. **Tag-based Similarity Search - FS-CLS**

We consider the problem of performing similarity search using only the topic representation for sounds in the FS-CLS dataset.

Methodology

For this task we make a comparison between topic and tf-idf representations. From each representation, a cosine similarity matrix is computed, referred to as the candidate similarity matrix. For LDA we also compute the hellinger distance similarity matrix (eqn. 4.1). To check the 'goodness' of these candidate similarity matrices we see how well they conincide with a reference similarity matrix obtained from human judgments. The reference similarity matrix is a result of an online experiment conducted in [Dimitriou, 2014]. In this experiment users were asked to rate similarity between pairs of sounds on a scale of 0 to 10. Approximately 11500 comparisons were obtained [Dimitriou, 2014]. The ratings received for each pair were averaged to construct the reference similarity matrix.

Logan measure [Berenzweig et al., 2004] is used as the evaluation metric for comparing candidate and reference similarity matrices. Computation of this metric is the following two-step process:

- First a Top-N agreement score, s_i is computed. For this, rows in both matrices are sorted in decreasing order. Top-N values in the i^{th} row of each matrix now represent the retrieval results for the i^{th} query sound. Thus, for each row, the top N 'hits' from the sorted reference similarity matrix are exponentially weighed with a factor α_{ref}^{r-1} depending upon their rank r after sorting. Similarly, for each sound at rank r in the reference matrix corresponding rank k_r is determined in the sorted candidate matrix and weighed with another factor, $\alpha_{can}^{k_r-1}$. These values are combined according to eqn 4.1. α_{ref} and α_{can} are experimental constants that represent the sensitivity to ordering in reference and candidate similarity matrix respectively.
- The Top-N agreement scores are then normalized with the maximum and the mean is computed to give the final metric .

$$s_i = \sum_{r=1}^{N} (\alpha_{ref})^{r-1} (\alpha_{can})^{k_r - 1} \tag{4.1}$$

To summarize, we compare the topic based representation with the standard PCA vector obtained from TF-IDF representation. For LDA we test with the cosine and hellinger similarity measures. The latter is specific to finding similarity between probability distributions (eqn.4.2).

$$\text{document-similarity}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\theta}_{d,k} - \sqrt{\theta}_{f,k} \right)^2 \tag{4.2}$$

Parameters: Feature Vector Dimension, K = [5, 10], N=5, $\alpha_{ref} = 0.5^{1/3}$, $\alpha_{can} = (\alpha_{ref})^2$, LDA Iterations=1500

Results

The results are presented in Table 4.2

- The results are presented for the top 5 hits and feature vector dimensions of 5 and 10. It is evident that LDA based methods outperform the standard tf-idf approach. The primary advantage of LDA is that even for very low dimension feature vectors we can expect to get a better performance than conventional methods.
- For LDA based methods we observe a slight decrease in the logan metric for K=10. A possible reason for this could be that decomposition into 5 topics is better than that into 10.

Methods	K=5	K=10
LDA-Cosine	**0.146**	**0.143**
LDA-Hellinger	0.141	0.136
tf-idf-Cosine	0.117	0.129

Table 4.2: Tag-based Similarity Search evaluation for different methods based on Logan measure

4.4 Conclusion and Discussion

In this chapter we have dealt with the rationale and intuition behind latent dirichlet allocation. After an explanation of the basic technical details we show its applicability to Freesound dataset through three evaluation experiments. In our similarity search experiment we provide evidence for its superiority over conventional tf-idf approach. LDA proves to be a powerful tool for obtaining automatic categorization and a low dimensional representation of tag information associated with each audio file. The representation in terms of topics also helps us reduce the noise present in tags.

It must be noted that the number of topics must be given as input to LDA. This parameter should be fine tuned through repeated experimentation to achieve the right amount of granularity in terms of topics for a given dataset and problem. The analysis carried out in this thesis must be extended to larger datasets.

Chapter 5

SUMMARY AND FUTURE WORK

5.1 Summary and Contributions

Main contributions of this thesis can be summarized as follows:

1. **Content Categorization**

   - We propose a framework for taxonomical organization and thresholding-based classification of loudness profiles.
   - Conducted a subjective evaluation (online experiment) that shows our system's superior performance over Freesound's current similarity search

2. **Metadata Categorization**

   - Proposed the use of Latent Dirichlet Allocation (LDA), a popular topic model for tag information representation in the context of Freesound.
   - Evaluated use of topic-based feature vectors for audio similarity search. This gives us promising results, in terms of both, the performance measure and dimensionality reduction.

In addition, we provide a comprehensive overview of literature on morphological description taxonomies and content-based audio retrieval systems.

Through this thesis we have contributed to the field of audio retrieval, particularly in the context of Freesound. The proposed thresholding approach for content categorization was successfully built upon the B-spline approximation of profiles introduced in [Peeters and Deruty, 2010]. The results of our experiments not only show that the method is comparable to machine learning techniques but emphasize its flexibility with regard to incorporation of the "others" class. The subjective evaluation strengthens evidence for the method's ability to improve Freesound similarity search performance. Our preliminary work (described in Sec. 3.3) shows the extendability of this method to pitch profiles. To the best of our knowledge this study is the first attempt at morphological description of a real-world dataset consisting of more than 5000 sounds.

On the front of metadata categorization, contrary to the conventional tag/tf-idf representation based feature vectors, we propose to use topic models. Their effectiveness in providing meaningful, high level, low dimensional representation is illustrated.

5.2 Future Work

We conclude this thesis by presenting future research avenues. For both, content and metadata categorization we have identified several shortcomings of our approach. Moreover, our work also opens up several very interesting research paths. Following is a list of suggested future work for immediately extending and improving the work carried out in this thesis:

- Develop a similar framework for timbre profile inclusion and also progress with the pitch profile characterization. Inclusion of features for better description of classes such as impulsive is essential.

- Incorporating the slope/relative duration descriptors into a web-based advanced content search facility. Possibly, giving the user control over parameters such as the modulation rate, extent and relative duration. A unique and intuitive way of doing this could be to ask the user to draw the profile. Relevant parameters can then be extracted from the drawing.

- In the context of Freesound, developing a framework for iterative sound description would be an important contribution

- Since the LDA based model has shown positive results, next step would be to integrate these features with the content based features so as to enhance the current similarity search.

- LDA is a generic model. Evaluating and building several other problem-specific topic models with larger datasets would prove to be more beneficial

- Topics can be used as a visualization tool for assisting in retrieval tasks. This could enrich the user experience.

Bibliography

[Berenzweig et al., 2004] Berenzweig, A., Logan, B., Ellis, D. P., and Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Bogdanov et al., 2013] Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *ISMIR*, pages 493–498.

[Cano and Koppenberger, 2004] Cano, P. and Koppenberger, M. (2004). Automatic sound annotation. In *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, pages 391–400. IEEE.

[Cano et al., 2005] Cano, P., Koppenberger, M., Groux, S. L., Ricard, J., Wack, N., and Herrera, P. (2005). Nearest-neighbor automatic sound annotation with a wordnet taxonomy. *Journal of Intelligent Information Systems*, 24(2):99–111.

[Cano et al., 2004] Cano, P., Koppenberger, M., Herrera, P., Celma, O., and Tarasov, V. (2004). Sound effects taxonomy management in production environments. In *Proc. AES 25th Int. Conf.*

[Chaney and Blei, 2012] Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *ICWSM*.

[Chion, 1983] Chion, M. (1983). Guide to sound objects. pierre schaeffer and musical research. *Trans. John Dack and Christine North), http://www. ears. dmu. ac. uk*.

[De Cheveigné and Kawahara, 2002] De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

[Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

[Dimitriou, 2014] Dimitriou, C. A. (2014). Query-by-example model creation methodology in freesound. *Master's Thesis, Universitat Pompeu Fabra*.

[Font et al., 2013] Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *ACM International Conference on Multimedia (MM'13)*, pages 411–412, Barcelona, Spain. ACM, ACM.

[Font et al., 2014] Font, F., Serrà, J., and Serra, X. (2014). Audio clip classification using social tags and the effect of tag expansion. In *AES 53rd International Conference on Semantic Audio*, London.

[Foote, 1997] Foote, J. T. (1997). Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pages 138–147. International Society for Optics and Photonics.

[Gaver, 1993] Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29.

[Gu et al., 2005] Gu, J., Lu, L., Cai, R., Zhang, H.-J., and Yang, J. (2005). Dominant feature vectors based audio similarity measure. In *Advances in Multimedia Information Processing-PCM 2004*, pages 890–897. Springer.

[Hartmann, 1996] Hartmann, W. M. (1996). Pitch, periodicity, and auditory organization. *The Journal of the Acoustical Society of America*, 100(6):3491–3502.

[Helen and Lahti, 2006] Helen, M. and Lahti, T. (2006). Query by example methods for audio signals. In *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, pages 302–305.

[Helén and Virtanen, 2007a] Helén, M. and Virtanen, T. (2007a). Query by example of audio signals using euclidean distance between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–225. IEEE.

[Helén and Virtanen, 2007b] Helén, M. and Virtanen, T. (2007b). A similarity measure for audio query by example based on perceptual coding and compression. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*.

[Herrera-Boyer et al., 2003] Herrera-Boyer, P., Peeters, G., and Dubnov, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21.

[Hoffman et al., 2009] Hoffman, M. D., Blei, D. M., and Cook, P. R. (2009). Easy as cba: A simple probabilistic model for tagging music. In *ISMIR*, volume 9, pages 369–374.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Kabal, 2002] Kabal, P. (2002). An examination and interpretation of itu-r bs. 1387: Perceptual evaluation of audio quality. *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, pages 1–89.

[Klapuri, 2005] Klapuri, A. P. (2005). A perceptually motivated multiple-f0 estimation method. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 291–294. IEEE.

[Lee and Seung, 2001] Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

[McAdams et al., 1995] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192.

[Mesaros et al., 2013] Mesaros, A., Heittola, T., and Palomaki, K. (2013). Query-by-example retrieval of sound events using an integrated similarity measure of content and label. In *WIAMIS'13*, pages 1–4.

[Moore and Glasberg, 1996] Moore, B. C. and Glasberg, B. R. (1996). A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345.

[Peeters and Deruty, 2010] Peeters, G. and Deruty, E. (2010). Sound indexing using morphological description. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):675–687.

[Peeters et al., 2011] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916.

[Ricard, 2004a] Ricard, J. (2004a). Towards computational morphological description of sound. *DEA pre-thesis research work, Universitat Pompeu Fabra, Barcelona*.

[Ricard, 2004b] Ricard, J. (2004b). Towards computational morphological description of sound. Master's thesis.

[Ricard and Herrera, 2004] Ricard, J. and Herrera, P. (2004). Morphological sound description: Computational model and usability evaluation. In *Audio Engineering Society Convention 116*. Audio Engineering Society.

[Schaeffer, 1966] Schaeffer, P. (1966). Traité des objets musicaux.

[Schedl et al., 2014] Schedl, M., Gómez, E., and Urbano, J. (2014). *Music Information Retrieval: Recent Developments and Applications*. now Publishers.

[Shamma, 2003] Shamma, S. (2003). Encoding sound timbre in the auditory system. *IETE Journal of research*, 49(2-3):145–156.

[Slaney, 2002] Slaney, M. (2002). Mixtures of probability experts for audio retrieval and indexing. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 345–348 vol.1.

[Sordo, 2012] Sordo, M. (2012). *Semantic Annotation of Music Collections: A Computational Approach*. PhD thesis, Universitat Pompeu Fabra, Barcelona (Spain).

[Spevak and Favreau, 2002] Spevak, C. and Favreau, E. (2002). Soundspotter-a prototype system for content-based audio retrieval. In *Proceedings of the 5th International Conference on Digital Audio Effects*.

[Timoney et al., 2004] Timoney, J., Lysaght, T., Schoenwiesner, M., and MacManus, L. (2004). Implementing loudness models in matlab.

[Turnbull et al., 2008] Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):467–476.

[Virtanen and Helen, 2007] Virtanen, T. and Helen, M. (2007). Probabilistic model based similarity measures for audio query-by-example. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 82–85.