Improving Audio Retrieval through Loudness Profile Categorization

Sanjeel Parekh
Technicolor R&D, France
sanjeel.parekh@technicolor.com

Frederic Font and Xavier Serra
Music Technology Group, Universitat Pompeu Fabra
Barcelona, Spain

***Abstract*—The increasing popularity of audio content sharing in online platforms requires the development of techniques to better organize and retrieve this data. In this paper we look at how to improve similarity search through content categorization in the context of Freesound, a popular online sound sharing site. We focus on organization based on morphological description. In particular, we propose to improve search results by incorporating information about query sound's loudness profile. This is performed within a thresholding based framework and can be generalized to structure information about the temporal evolution of other sound attributes. We perform a subjective evaluation to demonstrate the practical relevance of our method.**

***Keywords*-audio similarity search; morphological description; retrieval; freesound**

I. Introduction

There is an exponential increase in the amount of user 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 paper we focus on *Freesound*[1] which is an online database of sounds where audio clips are shared by users under the creative commons licenses [1]. This database, which has now more than four million registered 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 low-level content descriptors, but the descriptors are not tailored to the particular types of content that can be found in it. As a consequence, the search results are often not relevant. This makes conducting structured content search a persisting problem.

Most of the current audio retrieval systems do not take advantage of perceptual criteria or information about sound's temporal evolution, instead they use feature values averaged over a bag of frames approach. As a result we are faced with two primary difficulties:

- 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 have an undefined source or is not adequately described in words, retrieval would be difficult without a perception based advanced search. Some sound engineers or artists also tend to have a template of "how the sound should be ?" in terms of its loudness or pitch profile. We believe filtering sounds based on such criteria would 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.

Thus, to alleviate these limitations there is a need to incorporate information about the primary perceptual attributes of sound, namely loudness, pitch and timbre. Here we propose a generic thresholding-based framework for loudness profile categorization to improve audio retrieval. The rest of the paper is organized as follows: In Sec. II we discuss the previous work followed by a rationale and description for our approach in Sec. III. Subsequently in Sec. IV we present experimental results and conclude in Sec. V.

II. Related Work

A sound can either be described in terms of the previously stated perceptual characteristics or based on its source of generation. Though the source-centric description is important, as discussed earlier, a categorization in terms of perceptual traits (or morphological characteristics) would provide a more generic description for any kind of sound. In this context, we discuss Schaeffer's work [2] on typomorphology which has been utilized to build taxonomies for sound indexing and retrieval [3], [4].

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 [5]. Schaeffer proposes to describe these *objects* using seven morphological components, grouped into three 'criteria'. Fig. 1 concisely presents this. The matter/form (or shape) pair are central to Schaeffer's morphological taxonomy. 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. Interested reader is referred to [6], [3], [4]

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

for details on suggested taxonomies and representations. We use Peeters et al.'s [4] taxonomy with some modifications.

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 1. Schaeffer's morphological criteria [7]

We use [4], [3] as key references. With respect to the previous work, our primary contributions are (i) proposal of a simple, intuitive and flexible approach for loudness profile categorization using thresholding (ii) demonstrating that filtering similarity search results based on query sound's loudness profile improves their relevance.

III. Our Approach

A. Methodology

In this work we concern ourselves with loudness, which, as defined by the American National Standards Institute (ANSI) is *that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud*. A loudness profile can then be described as the temporal evolution of a sound's loudness.

The idea is to threshold a few meaningful parameters to retrieve sounds similar to any kind of loudness profile. Thus, we first extract features relevant for describing a sound's loudness profile. Ideally we would now allow the user to customize the loudness profile parameters for sound retrieval. However, from the viewpoint of evaluation we show that it can be done for the taxonomy shown in Fig. 2.

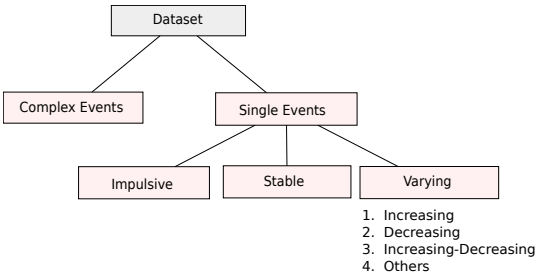

Figure 2. Content categorization scheme based on loudness profiles

B. Modeling Temporal Evolution - Feature Extraction

We modify the modeling strategy used in [4] to incorporate our requirements.

1) **Loudness Computation** - Spectrum is computed and outer-mid ear filtering is performed [8]. Next, the energy in each bark band, denoted by $E(z,t)$ is obtained. The loudness is then computed as

$$l(t) = \sum_z l'(z,t) \text{ where } l'(z,t) = E(z,t)^{0.23} \quad (1)$$

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 filter cut-off was set at 2 Hz. The time axis is normalized for all the sounds.

2) **Complex/Single Event Classification** - Clearly, the profile description would apply only to single sound events i.e. sounds with one loudness envelope. Since the dataset we use also contains complex events, we automatically separate complex and single events using the loudness curve. First, an onset detection function is constructed from the derivative of the loudness profile and then, peaks of this function are detected using a running mean threshold. Any sound with more than one peak is classified as complex. Hereafter, we only consider the loudness curves for single events.
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. Now we have a straight line approximation for the filtered, thresholded log-scale loudness curve.
4) **Extracted Features** - As shown in Fig. 3, we extract the following slope and relative duration features from this representation. We denote the time instances corresponding to *start*, *maximum* and *end* of the profile with t_s, t_M and t_e respectively:
   - RD1 - Relative duration - $t_M - t_s$
   - RD2 - Relative duration - $t_e - t_M$ or 1-RD1
   - S1 - Slope of the approximation from where it begins (t_s) to the maximum (t_M)
   - S2 - Slope of the approximation between maximum (t_M) and the end point (t_e)

   We 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.

C. Profile Categorization

The template of a short/long sound or increasing/decreasing profile can be defined intuitively. We demonstrate here how that intuition can be translated into parameter thresholds for broad categorization into classes shown in Fig. 2. As indicated, for single events the loudness curve could belong to one of the following categories: *impulsive* if it has either a sharp attack or is of a very short duration; *stable*

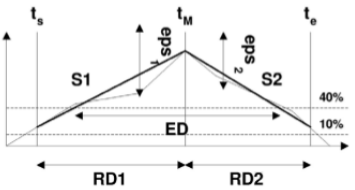

Figure 3. Loudness Profile Descriptors [4]

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

Table I
SFX-REDUCED DATASET DETAILS

if the loudness of the sound does not vary much; It is of *increasing* (or decreasing) category if the loudness increases (or decreases) for significant portion of the sound's duration. and of *delta* class if it is perceived of increasing-decreasing loudness. The *others* class would contain sounds which lie in the "confusion" areas

1) Setting Thresholds: Consider for instance the increasing class, ideally the sound would increase in loudness for all its duration, however, in reality, the 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.

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 (Oth Inc.): $0.6 < RD1 < 0.7$ and others-decreasing (Oth Dec.): $0.3 < RD1 < 0.4$

Observe that 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 previously.

IV. EXPERIMENTAL RESULTS

A. Dataset

First the **FS-SFX** dataset, containing 5248 sounds 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. For the experiments discussed in this paper we also use **SFX-Reduced** dataset (238 sounds), a subset of FS-SFX manually annotated according to the loudness profile. The details are provided in Table I.

B. Sanity Test

We perform this as a sanity check for our manually thresholded categories. Classification results are presented on the SFX-Reduced dataset. We use Essentia [9] for extracting onsets and loudness curve. We make the following observations from Table II:

1) Note that the sounds were manually annotated to belong to one of the five classes in Table I. Hence, the confusion matrix is not square and includes the others and the complex classes for predictions. The system mis-classifies 28/181 single events into complex category. We emphasize that we are only aiming for broad categorization and satisfactorily achieve that with very less complexity.
2) A sound belonging to other categories has been misclassified into the impulsive class. It is particularly evident for the delta class. This implies that, using only the effective duration descriptors for the impulsive class is not sufficient.

		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 II
CONFUSION MATRIX: LOUDNESS PROFILE CLASSIFICATION

C. Subjective Evaluation

We analyze the utility of our framework for the use case of similarity search. This is a very useful application of morphological description, which, to the best of our knowledge is not deployed by any online audio sharing platform. In particular, we compare the performance of the current Freesound similarity search with the proposed modified version of it.

Current implementation in freesound performs a kNN search over PCA features extracted using the Essentia framework. The euclidean similarity measure is used. Features consist of statistics computed over various low-level

features. [2] Thus the information embedded in temporal variations is lost. For the modified system the results are obtained after filtering current system's results according to the query sound's loudness profile category.

Query Sound Category	Filter
Impulsive	Impulsive
Stable	Stable
Increasing	Inc. + Oth Inc.
Decreasing	Dec. + Oth Dec.
Delta	Oth Dec. + Delta + Oth Inc.

Table III
MODIFIED SYSTEM: FILTERS FOR REFINING SEARCH RESULTS

1) Experiment Design: We ran an online survey where each user was asked to rate retrieval results from two systems for eight query sounds. Each query sound was followed by top 5 results (computed over FS-SFX dataset) 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 user 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. The users were also provided with a 'No Preference' option, in case they did not find any of the systems to be better than the other. For each user 8 query sounds were chosen from a pool of 91 sounds selected from SFX-Reduced dataset. These were sounds which were correctly classified into the five categories by our system (refer to Table II). For the modified system, the filters used for refining similarity search results are presented in Table III.

2) Results and discussion: 13 candidates participated in this online experiment. We obtained a total of 104 judgements (8/candidate). Out of these, 25.9% (27/104) were 'No preference'. Discarding these, we see from Fig. 4 that 74.02% (57/77) of the judgements were in favor of the modified system. To further validate the performance, we see that for all the candidates, the number of responses in favor of the modified system were always greater than or equal to those in favor of current Freesound system (equality held only in one case). This gives us strong evidence to claim that the modified version is an improvement over the current Freesound system.

V. CONCLUSION

We have successfully demonstrated the use of high-level loudness descriptors for generic audio similarity search. More generally we observe that simple search strategies can be improved significantly by incorporating domain-specific constraints. We must highlight that we wanted to achieve broad categorization and not accurate classification. Hence, intuitively choosing soft bounds over meaningful parameters suffices for our purposes. Moreover, the thresholds are only presets which can be adjusted according to a user's preference.

[2]File in freesound code repository specifying similarity search setting - https://github.com/MTG/freesound/blob/master/similarity/presets/lowlevel.yaml

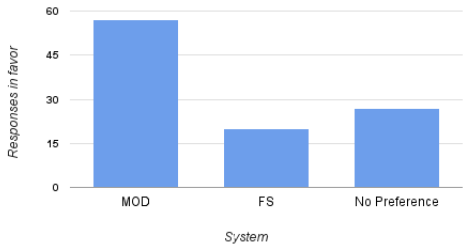

Figure 4. Shows no. of responses in favor of Modified system (MOD), Freesound system (FS) and 'No Preference' respectively

Immediate applications would include: (i) Extension of this framework to other sound attributes (ii) Advanced search facility in freesound (iii) Automatic content-based labeling. Overall, this work is relevant for sound databases where information can be organized based on perceptual criteria for better content-based retrieval and metadata generation.

REFERENCES

[1] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, 2013, pp. 411–412.

[2] P. Schaeffer, "Traité des objets musicaux," 1966.

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

[4] G. Peeters and E. Deruty, "Sound indexing using morphological description," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 675–687, 2010.

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

[6] J. Ricard and P. Herrera, "Using morphological description for generic sound retrieval." in *ISMIR*, 2003.

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

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

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