Proposal of the Aesthetic Experience-Oriented Evaluation Framework for Field-recording Sound Retrieval System

Experiments using Acoustic Feature Signatures Based on Multiscale Fractal Dimension

Motohiro Sunouchi
Design Department, Sapporo City University,
Geijutsu-no-mori 1, Minami-ku, Sapporo, Hokkaido
005–0864, Japan
sunouchi@media.scu.ac.jp

Masaharu Yoshioka
Graduate School of Information Science and Technology,
Hokkaido University, Kita 14, Nishi 9, Kita–ku, Sapporo,
Hokkaido, 060–0814, Japan
yoshioka@ist.hokudai.ac.jp

ABSTRACT

Sound designers and musicians often need to retrieve sound materials based on their similarity to aesthetic hearing experiences from sound databases such as Freesound. This study proposes an aesthetic experience-oriented evaluation framework for a field-recording sound retrieval system, using the sound clips extracted from Freesound. Furthermore, we discuss the features of the framework by analyzing the performance of the similarity search system for field-recording sound material using acoustic feature signatures that are based on the multiscale fractal dimension.

CCS CONCEPTS

• **Applied computing**; • **Arts and humanities**; • **Sound and music computing**; • **Information systems**; • **Information retrieval**; • **Specialized information retrieval**; • **Multimedia and multimodal retrieval**; • **Speech / audio search**; • **Evaluation of retrieval results**;

KEYWORDS

aesthetic experience-oriented evaluation framework, field-recording sound analysis, multiscale fractal dimension, content-based retrieval

ACM Reference Format:
Motohiro Sunouchi and Masaharu Yoshioka. 2021. Proposal of the Aesthetic Experience-Oriented Evaluation Framework for Field-recording Sound Retrieval System: Experiments using Acoustic Feature Signatures Based on Multiscale Fractal Dimension. In *2021 3rd International Conference on Image, Video and Signal Processing (IVSP 2021), March 19–21, 2021, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3459212.3459223

1 INTRODUCTION

Over the last decade, online sound sharing services have gained popularity and become accessible to many people. Freesound [1, 2] aims to create a huge collaborative database of creative-commons licensed sounds for musicians and sound lovers. Splice [3] is a cloud-based music creation and collaboration platform with a sound database that provides users with sound materials for music creation. For these online services, one of the most significant functions is to search for sounds that are similar to what each user is looking for.

The Freesound database contains various field-recording sounds that are produced external to recording studios. Field recording sounds include both natural and artificial sounds. These sound recordings are appreciated as an aesthetic experience of sound, and are useful for sound creators, such as sound designers and musicians, to develop new sound works.

In recent years, research on environmental sound recognition (ESR) for understanding a scene and its context has received considerable attention [4]. The workshop challenges on the detection and classification of acoustic scenes and events (DCASE) have demonstrated the performance evaluations of systems for the detection and classification of sound events [5].

2 MOTIVATION

Field-recording sounds have important acoustic features in the time domain with various time scales [6]. When sound creators search for new sound materials in a field-recording sound database, they listen to sounds with an awareness of both specific sound sources (bird, car, insect, etc.) and a phenomenal hearing experience of the entire target sound, including background sounds and noises [7]. Therefore, a similarity search system for field-recording sounds should use acoustic features that can describe the qualities of a timbre with varying time periods.

Recent evaluation frameworks for ESR, such as the DCASE challenge, have focused on tasks that detect pre-defined sound events and that recognize specific sound sources. However, sound creators often need to retrieve sound materials based on their similarity to the aesthetic hearing experience from sound databases such as Freesound. We assume that the tasks of the DCASE challenge cannot evaluate the requirements of sound creators. Therefore, this study proposes an aesthetic experience-oriented evaluation framework for a field-recording sound retrieval system using sound clips extracted from Freesound, which are labeled with a group of tags.

In 2013, we proposed a new acoustic feature signature, namely, the enhanced multiscale fractal dimension (EMFD) signature, and demonstrated the effectiveness of EMFD for a content-based similarity search of field-recording sounds, which were extracted from Freesound [8]. Then, we improved the EMFD signature using the

IVSP 2021, March 19–21, 2021, Singapore, Singapore

ACM ISBN 978-1-4503-8891-7/21/03…$15.00
https://doi.org/10.1145/3459212.3459223

Table 1: Definition of tag categories

Category	Definition
Situation	Tags that represent locations, circumstances, environment, time, and seasons for sound recording
Sound source	Tags that represent specific sound sources that humans can recognize.
Others	Tags that do not have any relation with a feature of sound. Tags that are not categorized according to the situation or sound source category.

kernel density estimation method, which was named the EMFD-KDE signature. Furthermore, we developed another acoustic feature signature based on MFD, namely, the very-long-range MFD (MFD-VL) signature. The MFD-VL signature describes several features of the time-varying envelope for long periods [9].

To verify the effectiveness of our proposed framework, we discuss the features of the framework by analyzing the performance of the similarity search system for field-recording sound material using the EMFD-KDE and MFD-VL acoustic signatures.

3 IDEAS OF AESTHETIC EXPERIENCE-ORIENTED EVALUATION FRAMEWORK

We focused on conditions that sound creators listen to sounds with an awareness of both specific sound sources and a phenomenal hearing experience of the entire target sound, including background sounds and noises when they search for new sound materials in the field-recording sound database. We refer to the former listening attitude as "semantic hearing" and the latter attitude as "aesthetic hearing".

For the evaluation framework, we develop a similarity search system and define the similarity index between the tag group of the search-key sound and that of the retrieved sound to evaluate the performance of the system. The tags of user-generated content, such as sound clips of Freesound, often vary in number and quality for each content. To fix the negative effects of this problem, we define the normalized discounted cumulative gain (nDCG) using the similarity index.

Furthermore, we propose categories of tags and apply them to the framework to evaluate the performance of the similarity search task corresponding to the "aesthetic hearing" and "semantic hearing" of sound creators, respectively. Table 1 shows the definition of tag categories. We assume that the tags in the sound source category relate to the acoustic features for "semantic hearing" and the tags in the situation category can relate to those for "aesthetic hearing". We define the normalized tag frequency that is a measure to evaluate the quality of retrieved sounds using categorized tags. We discuss the results of the similarity search task and the effectiveness of aesthetic experience-oriented evaluation framework through the experiments.

4 EXPERIMENTAL ANALYSIS

4.1 Setup of Similarity Search System and Sound Dataset

To analyze the effectiveness of acoustic feature signatures based on the MFD, we developed a similarity search system using the k-nearest neighbors (k-NN) method. A sound dataset contains 3000 sounds that were collected from Freesound, and tagged with "field-recording," having lengths between 1 and 600s. We chose the top 3000 sounds in descending order of the downloaded numbers by unspecified users that are counted by Freesound's system for each sound. Each sound was converted to a uniform format (1 channel, 44100-Hz sampling frequency, 16-bit depth, and maximal amplitude normalized to -0.1 dB) for normalization before extracting acoustic features, including EMFD-KDE, MFD-VL, and MFCC39. The average length of the sounds is 70.4s. The sound dataset is openly available on the Web [10].

We used tags labeled to each sound in the dataset to evaluate the system performance. We removed the common morphological and inflectional endings from all tags using Porter Stemmer [11] in advance. Moreover, pre-defined stop words, including sound formats, such as “mp3” and “stereo,” and tool makers, such as “sony” and “tascam,” were removed from the tag sets. The tags that contain the term "fieldrecord" were removed from the tag sets because all sounds in the dataset contained them.

4.2 Feature Signature Extraction

A fractal dimension is an index value that represents the characteristics of a fractal by quantifying its complexity in detail as a ratio of the change to the change in scale. Acoustic features based on fractal dimensions have been proposed and utilized in various practical applications. Maragos *et al.* [12] proposed the short-time fractal dimension of speech signals as an acoustic feature, using it for speech segmentation and sound classification. Zlatintsi and Maragos [13] proposed a MFD profile as a short-time descriptor, and concluded that the MFD profile can discriminate several aspects among different musical instruments.

In our previous work [8], we proposed an EMFD signature that can describe both the frequency-domain and time-domain features of field-recording sounds. The EMFD signature is a feature vector, that consists of time-varying MFD values. Then, we extended EMFD using the kernel density estimation method (EMFD-KDE), which results in increased stability and robustness against small fluctuations in the parameters of sound sources. Furthermore, we proposed another acoustic feature signature based on MFD, namely MFD-VL. The MFD-VL signature describes several features of the time-varying envelope for long periods [9].

Table 2 shows the acoustic feature sets that were used in the experimental evaluation. We used a feature vector of MFCC39 as a feature set of the baseline, which is a popular acoustic feature that is used for analyzing environmental sounds. MFCC39, which represents the first- and second-order derivatives of MFCC13, was computed using the SPTK toolkit [14] with a fixed-width analysis

Table 2: List of acoustic feature sets for the evaluations.

Acoustic Feature Sets	L1	L2
1 MFCC39 [baseline]	39	24
2 MFCC39 + MFD-VL (x5.5)	49	27
3 MFCC39 + EMFD-KDE (x1)	551	114
4 MFCC39 + EMFD-KDE (x1) + MFD-VL (x0.8)	561	114

window of 50-ms length. The feature sets of MFCC39 consist of the mean values of its coefficients of the analysis window. Column L1 in Table 2 shows the total number of features in the concatenated feature sets, and column L2 shows the number of features in the feature sets after dimensionality reduction through principal component analysis (PCA). To extract eigenvectors for dimensionality reduction, PCA was applied to the feature vectors of the 600 most frequently downloaded sounds in the dataset. The "prcomp" function of R language was used for the PCA. The L2s of feature sets 1, and 2 were determined such that each of their cumulative contribution ratios was 99%. The L2s of feature sets 3, and 4 were set to 114. Suffix "($\times\gamma$)" of each feature vector denotes weighting coefficient γ. Each value of the feature vectors is multiplied by γ when its feature vector is concatenated with other feature(s). Through experimental evaluation, weighting coefficient γ for each feature vector was appropriately chosen to obtain the best result.

4.3 Evaluation of Similarity Search Performance Using Similarity Index and nDCG

To evaluate the results of the field-recording sounds returned by the similarity search system, we defined the similarity index SI, as in Eq. (1), which represents the similarity between the tag set of the search-key sound $\mathrm{tags}_{\mathrm{key}}$ and that of the retrieved sound $tags_s$. This index is known as the Jaccard similarity coefficient, which measures the similarity between finite sample sets.

$$SI = \frac{\mathrm{card}\left(tags_{key} \cap tags_s\right)}{\mathrm{card}\left(tags_{key} \cup tags_s\right)} \tag{1}$$

For each of the 3000 sounds in the dataset, SIs between a search-key sound and each retrieved sound in the search-result list were computed. Then, we computed the discounted cumulative gain (DCG) [15] to measure the ranking quality of the search-result list. Let $SI_{S,\,i}$ be the SI between search-key sound S and the i-th retrieved sound in the search-result list. We defined $DCG_{S,\,k}$ accumulated at a particular rank position k, as shown in Eq. (2).

$$DCG_{S,\,k} = \sum\nolimits_{i=1}^{k} \frac{SI_{S,\,i}}{\log 2\,(i+1)} \tag{2}$$

To compare the DCG values for different search-key sounds, the cumulative gain at each rank position k should be normalized. We computed the SIs between the two sounds of all possible combinations of the 3000 sounds. For each of the 3000 sounds, we produced perfect search-result lists that were sorted in descending order of SI values. When the search-key sound S is used as a search query and its perfect search-result list with SIs values is known, $DCG_{S,k}^{perfect}$, which accumulated at a particular rank position k, is computed by Eq. (2). Then, let SS_{key} be the sound dataset that contains search-key sounds. The normalized DCG (nDCG) at a particular rank position k was obtained using Eq. (3), which represents the average performance of the similarity search system.

$$nDCG@k = \frac{1}{\left|SS_{key}\right|} \sum\nolimits_{S}^{SS_{key}} \frac{DCG_{S,k}}{DCG_{S,k}^{perfect}} \tag{3}$$

Tables 3 and 4 show the scores of nDCG@1 and nDCG@3, respectively. We chose the top eight most popular tags in the dataset, and used them to create subsets of search-key sounds. These tables show the nDCG scores obtained by each subset of search-key sounds for each feature set. Suffix "(x)" in the column heading represents the number of search-key sounds labeled with each tag. Bold numbers are the best nDCG scores for each subset. The numbers with ++ and + show that the difference from the baseline is statistically significant using a t-test ($p < 0.01$ and $p < 0.05$ respectively).

The nDCG scores for the 3000 sounds show that both the EMFD-KDE and MFD-VL signatures improve the performance of the similarity search task and describe some acoustic features that are not described by MFCC39. The nDCG scores for feature set 4 showed the best similarity search performance. Hence, it was confirmed that the EMFD-KDE and MFD-VL signatures have different descriptions compared with MFCC39 for field-recording sounds.

4.4 Analysis of Additional Matched Tags Using MFD-VL and EMFD-KDE

To reveal the descriptiveness of the EMFD-KDE and MFD-VL signatures in detail, we analyzed the occurrence rates of the matched tags for each category between the tag set of the search-key sound and that of the retrieved sounds. Let $tags_{key}$ be the tag set of the search-key sound, and $tags_i$ be the set of the i-th retrieved sound. The tag multiset $TAGS_{k,ft}^{matched}$ denotes the summation of matched tag sets between the search-key sound and top-k retrieved sounds using the feature set ft, as in Eq. (4). Let SS_{key} be a sound dataset containing search-key sounds. Then, we defined the tag multiset $TAGS_{k,ft}^{diff}$, as in Eq. (5), which contains the summation of additional matched tags using the feature set ft, and compared with them using the baseline feature set (MFCC39) for each query sound in dataset SS_{key}.

$$TAGS_{k,ft}^{matched} = \sum\nolimits_{i=1}^{k} \left\{tags_{key} \cap tags_i\right\} \tag{4}$$

$$TAGS_{k,ft}^{diff} = \sum\nolimits^{SS_{key}} \left\{TAGS_{k,ft}^{matched} \setminus TAGS_{k,baseline}^{matched}\right\} \tag{5}$$

Table 3: nDCG@1 scores. bird*=(bird, birdsong), ambien*=(ambienc, ambianc, ambient), env-res=(environmentalsoundresearch), rain*=(rain, rainfall)

Feature Sets		situation				sound source			
	all(3000)	natur(585)	ambien*(498)	env-res(187)	citi(180)	bird*(476)	water(251)	car(163)	rain*(149)
1 MFCC39 (baseline)	0.293	0.385	0.305	0.344	0.277	0.352	0.338	0.286	0.412
2 MFCC39 + MFD-VL (x5.5)	0.298+	0.393	0.310	0.337	0.284	0.356	0.329	0.299	0.420
3 MFCC39 + EMFD-KDE (x1)	0.316++	0.404+	0.323+	**0.391++**	**0.301**	0.368	**0.356**	0.308	**0.446**
4 MFCC39 + EMFD-KDE (x1) + MFD-VL (x0.8)	**0.321++**	**0.415++**	**0.324**	**0.391+**	0.292	**0.376+**	0.351	**0.330+**	0.430

Table 4: nDCG@3 scores

Feature Sets		situation				sound source			
	all(3000)	natur(585)	ambien*(498)	env-res(187)	citi(180)	bird*(476)	water(251)	car(163)	rain*(149)
1 MFCC39 (baseline)	0.257	0.350	0.268	0.280	0.237	0.316	0.277	0.264	0.366
2 MFCC39 + MFD-VL (x5.5)	0.261++	0.354	0.273	0.279	0.241	0.322+	0.279	0.272	0.372
3 MFCC39 + EMFD-KDE (x1)	0.277++	0.372++	0.283+	**0.317++**	**0.258+**	0.338++	0.298+	0.285+	0.387
4 MFCC39 + EMFD-KDE (x1) + MFD-VL (x0.8)	**0.280++**	**0.375++**	**0.285++**	0.313++	0.248	**0.339++**	**0.300+**	**0.294++**	**0.390+**

We obtained the list of matched tags with their occurrence numbers from $TAGS^{diff}_{3,ft}$ for feature sets 2 and 3, using all 3000 sounds as SS_{key}. Let DS be a sound dataset containing 3000 sounds. To normalize each tag's occurrence number, we use the inverse tag frequency itf_{tagX}, as defined in Eq. (6). In this experiment, the total number of sounds in dataset $|DS|$ was constantly 3000. Then, the normalized tag frequency at ranking position k ($ntf@k_{tagX}$) for $tagX$ was obtained using Eq. (7).

$$itf_{tagX} = \log \frac{|DS|}{|\{sounds \in DS|\text{sounds labeled with tagX in DS}\}|} \tag{6}$$

$$ntf@k_{tagX} = \left|\left\{TAGS^{diff}_{k,ft}|tagX\right\}\right| \cdot itf_{tagX} \tag{7}$$

Table 5 lists the additional matched tags with the top 20 ntf@3 for each feature set grouped by the categories. The normalized tag-frequency rate of categoryA $ntfr_{k,\ categoryA}$ denotes the ratio of the summation of $ntf@k_{tagX}$ for the tags in categoryA to that of tags in all categories, as in Eq. 8). We use this $ntfr_{k,\ categoryA}$ value to evaluate the descriptiveness of each feature set for each category. Similarly, we calculated the lists of the additional matched tags with the top 20 of ntf@3 using the search-key sounds labeled with "natur", "bird or birdsong", and "water", which are the top 3 most popular tags in the sound dataset; the results of which are shown in Tables 6, 7, and 8, respectively. In Tables 5-8, for all cases, we confirmed that the MFD-VL and EMFD-KDE signatures can describe the acoustic features related to the tags in both the "situation" and "sound source" categories. In cases 2, 3, and 4, the additional matched tag with the highest score of ntf@3 for each feature set is the same as that used to produce the subset of search-key sounds for each case.

$$ntfr_{k,\ categoryA} = \frac{\sum^{tags\ in\ categoryA} ntf@k_{tagX}}{\sum^{tags\ in\ all\ categories} ntf@k_{tagX}} \tag{8}$$

4.5 Evaluation of Feature Sets Using ntfr

Figure 1 shows the $ntfr_3$ values for each feature set in the stacked bar chart that are grouped by the cases 1-4. The $ntfr_3$ values of the situation category, when using EMFD-KDE (feature set 3 and 4), are greater than that when using MFD-VL (feature set 2) for each case. In particular, in cases 1, 2, and 3, the $ntfr_3$ values of the situation category when using EMFD-KDE are relatively high. The difference of rates in each case between feature set 3 and 4 are much smaller than that between feature set 2 and 3. As a consequence of these facts, we confirmed that the descriptiveness of the EMFD-KDE related to the situation category tends to be higher than that of the MFD-VL and the impact of the EMFD-KDE on the rates of the categories is larger than that of MFD-VL.

Table 5: List of additional matched tags with the top 20 ntf@3 obtained using all 3000 sounds as queries env_research = environmental sounds research

Case1: all (3000)							
Feature set 2: MFCC39+MFD-VL				**Feature set 3: MFCC39+EMFD-KDE**			
categories	ntfr$_3$	tags	ntf@3	categories	ntfr$_3$	tags	ntf@3
situation	31.0%	natur	68.7	situation	49.7%	natur	224.0
		spring	36.4			citi	132.2
		citi	28.1			ambianc	110.4
		ambienc	22.7			crowd	98.9
		ambient	22.4			ambienc	98.3
		atmospher	20.9			atmospher	80.5
sound source	69.0%	bird	73.5			sea	80.5
		water	42.2			street	70.3
		car	37.9			ambient	69.9
		machin	36.6			env_research	69.4
		rain	30.1	sound source	50.3%	bird	216.6
		fire	29.9			wave	145.0
		train	27.8			water	134.0
		wave	27.6			rain	105.3
		thunder	25.4			car	96.1
		wind	25.3			wind	79.1
		thunderstorm	23.1			thunder	69.9
		metal	22.9			footstep	68.0
		voic	21.3			audienc	66.3
		footstep	20.4			peopl	66.0

Table 6: List of additional matched tags with the top 20 ntf@3 obtained using sounds labeled with "natur" tag as queries

Case2: natur (585)							
Feature set 2: MFCC39+MFD-VL				**Feature set 3: MFCC39+EMFD-KDE**			
categories	ntfr$_3$	tags	ntf@3	categories	ntfr$_3$	tags	ntf@3
situation	47.2%	natur	68.7	situation	59.4%	natur	224.0
		spring	23.1			ambianc	56.6
		park	13.0			forest	53.9
		southspain	9.9			spring	49.6
		night	7.4			southspain	49.5
		summer	7.1			summer	35.4
		weather	6.7			weather	33.5
		citi	5.6			ambient	30.8
sound source	52.8%	bird	30.1			soundscap	30.4
		insect	16.7			atmospher	29.8
		water	14.9			storm	29.0
		wave	13.8	sound source	40.6%	bird	130.0
		stream	12.8			water	42.2
		thunder	12.7			wave	41.4
		thunderstorm	11.5			thunder	38.1
		cricket	10.8			insect	36.8
		birdsong	9.7			rain	36.1
		wind	9.5			birdsong	35.5
		rain	9.0			stream	34.2
		traffic	6.8			wind	31.6

Table 7: List of additional matched tags with the top 20 ntf@3 obtained using sounds labeled with "bird" or "birdsong" tag as queries

Case3: bird, birdsong (476)							
Feature set 2: MFCC39+MFD-VL				**Feature set 3: MFCC39+EMFD-KDE**			
categories	ntfr$_3$	tags	ntf@3	categories	ntfr$_3$	tags	ntf@3
situation	54.0%	natur	36.0	situation	58.0%	natur	111.2
		spring	29.8			forest	53.9
		southspain	13.2			ambianc	43.1
		park	13.0			spring	43.0
		countrysid	9.1			southspain	29.7
		summer	7.1			atmospher	26.8
		forest	6.7			ambienc	22.7
		atmospher	6.0			park	21.6
		ambient	5.6			ambient	19.6
		ambienc	5.0			soundscap	19.0
		streetnois	3.3			outdoor	16.4
		citi	2.8			summer	14.2
		ambianc	2.7			env_research	13.9
sound source	44.4%	bird	73.5			donana	13.8
		birdsong*	19.4	sound source	40.5%	bird	216.6
		insect	10.0			birdsong*	48.4
		crow	4.6			insect	23.4
		hors	4.4			wind	12.6
		cricket	3.6			rain	12.0
others	1.6%	flickr	4.1	others	1.5%	xy	11.6

Table 8: List of additional matched tags with the top 20 ntf@3 obtained using sounds labeled with "water" tag as queries

Case4: water (251)							
Feature set 2: MFCC39+MFD-VL				**Feature set 3: MFCC39+EMFD-KDE**			
categories	ntfr$_3$	tags	ntf@3	categories	ntfr$_3$	tags	ntf@3
situation	27.1%	natur	11.4	situation	32.8%	natur	32.7
		citi	8.4			citi	25.3
		countrysid	4.6			ocean	24.8
		wet	4.6			sea	23.0
		winter	4.2			beach	15.4
		sea	3.8			ambienc	12.6
		forest	3.4			atmospher	11.9
		atmospher	3.0			soundscap	11.4
sound source	65.0%	water	42.2			ambient	11.2
		wave	17.3			environ	9.2
		stream	12.8	sound source	65.5%	water	134.0
		bird	7.5			wave	79.4
		rain	6.0			stream	51.2
		human	4.3			drip	18.3
		river	4.2			rain	18.1
		footstep	3.4			river	17.0
		birdsong	3.2			bird	15.1
		wind	3.2			birdsong	12.9
others	8.0%	soundeffect	8.4			splash	9.0
		unprocess	4.4	others	1.7%	folei	9.0

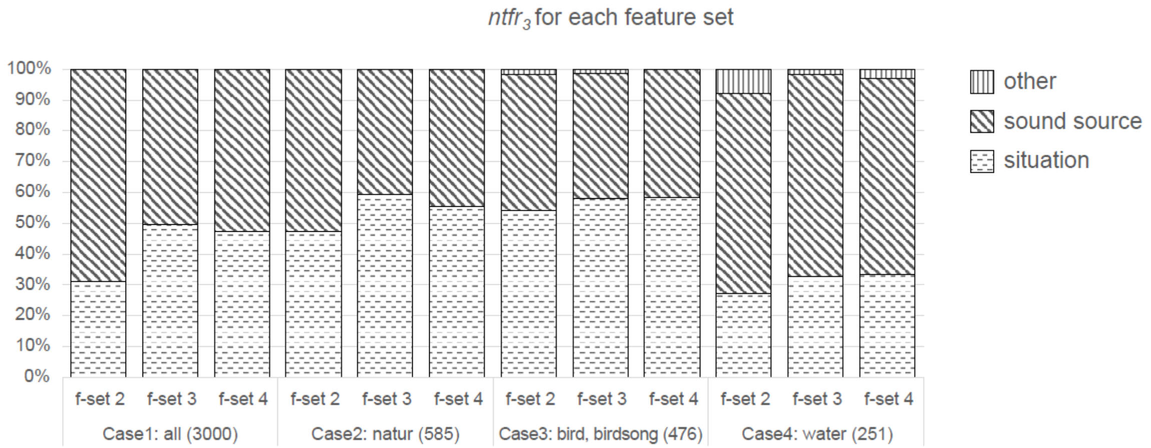

Figure 1: The normalized tag-frequency rates at ranking position 3 $ntfr_3$ for each feature set. They are grouped by the cases 1-4 that use the different search-key sound sets.

5 CONCLUSIONS

Our final goal is to realize a similarity search system in which sound creators can search for new sound materials. We focused on the conditions that sound creators listen to sounds with an awareness of both specific sound sources and a phenomenal hearing experience of the entire target sound that includes background sounds and noises when they search for new sound materials in the database. We proposed the aesthetic experience-oriented evaluation framework to evaluate the performance of similarity search system corresponding to the "aesthetic hearing" and "semantic hearing" of sound creators, respectively.

In Subsection 4.3, we confirmed that the MFD-VL and EMFD-KDE signatures have different descriptions compared with MFCC39 for field-recording sounds. In Subsection 4.4 and 4.5, we confirmed that the MFD-VL and EMFD-KDE signatures can describe the acoustic features related to the tags in both the situation and sound source categories. The descriptiveness of the EMFD-KDE related to the situation category tends to be higher than that of the MFD-VL. Furthermore, we confirmed that our developed feature signatures can describe the acoustic features related to the situation category with relatively high efficiency in the case where we use the search-key sounds labeled with "natur", "bird", and "birdsong" for the similarity search task.

Through the experiments using the framework, we could evaluate the difference in the descriptiveness between the MFD-VL and EMFD-KDE signatures using the $ntfr$ value of category based on the ntf. We assume that the acoustic features related to the situation category strongly affect the "aesthetic hearing" when sound creators search for new sound materials. We demonstrated that the evaluation of the sound similarity measured using only the tags in the sound source category is not enough. For the tasks based on the requirements of sound creators, we should evaluate the sound similarity measured using the tags in the situation category in addition to those in the sound source category. However, further studies are required to verify the validity of the proposed categories of tags. In addition, we should verify the integrity of dataset, because not all sounds are labeled with tags in the situation category.

In conclusion, this study has demonstrated that the aesthetic experience-oriented evaluation framework is useful for understanding the similarity search system for sound creators. Further studies are needed to develop ideal methods for applying this framework and these acoustic feature signatures to machine-learning systems and other applications.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 18K11378.

REFERENCES

[1] V. Akkermans, F. Font, J. Funollet, B. de Jong, G. Roma, S. Togias, and X. Serra, “FREESOUND 2.0: An Improved Platformfor Sharing Audio Clips,” 12th Int. Soc. Music Inf. Retr. Conf., (2011).

[2] Music Technology Group of Universitat Pompeu Fabra, “The Freesound Project.,” https://www.freesound.org/

[3] Distributed Creation Inc., "Splice - Royalty-Free Sounds & Rent-to-Own Plugins," Retrieved Feb 4, 2021 from https://splice.com/

[4] S. Chachada and C. C. J. Kuo, “Environmental sound recognition: A survey,” APSIPA Trans. Signal Inf. Process., vol. 3, (2014).

[5] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and Classification of Acoustic Scenes and Events,” IEEE Trans. Multimed., vol. 17, no. 10, pp. 1733–1746, (2015).

[6] S. Chu, S. Narayanan, and C.-C. Kuo, “Environmental Sound Recognition With Time-Frequency Audio Features,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 17, no. 6, pp. 1142–1158, (2009).

[7] M. Solomos, “A Phenomenological Experience of Sound: Notes on Francisco López,” Contemp. Music Rev., vol. 38, no. 1–2, pp. 94–106, 2019.

[8] M. Sunouchi, and Y. Tanaka, “Similarity Search of Freesound Environmental Sound Based on Their Enhanced Multiscale Fractal Dimension,” Sound Music Comput. Conf. 2013, SMC 2013, pp. 715–721, (2013).

[9] M. Sunouchi, and M. Yoshioka. "Diversity-Robust Acoustic Feature Signatures Based on Multiscale Fractal Dimension for Similarity Search of Environmental Sounds." arXiv preprint arXiv:2102.02964 (2021).

[10] "Sound Dataset extracted from Freesound," Online available. https://labs.43d.jp/fs3000_dataset/fs3000_dataset.tar.bz2 (17GB)

[11] M. Porter, “An algorithm for suffix stripping,” Progr. Electron. Libr. Inf. Syst., vol. 14, no. 3, pp. 130–137, (1980).

[12] P. Maragos and A. Potamianos, “Fractal dimensions of speech sounds: computation and application to automatic speech recognition.” J. Acoust. Soc. Am., vol. 105, no. 3, pp. 1925–1932, (1999).

[13] A. Zlatintsi and P. Maragos, “Multiscale Fractal Analysis of Musical Instrument Signals With Application to Recognition,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 21, no. 4, pp. 737–748, (2013).

[14] SPTK working group, “Speech Signal Processing Toolkit (SPTK),” http://sp-tk.sourceforge.net/, (Retrieved 2021-2-4).

[15] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, and W. Chen, “A Theoretical Analysis of NDCG Type Ranking Measures,” Proc. 26th Annu. Conf. Learn. Theory, pp. 1–30, 2013.