BF-Classifier: Background/Foreground Classification and Segmentation of Soundscape Recordings

Miles Thorogood
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
mthorogo@sfu.ca

Jianyu Fan
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
jianyuf@sfu.ca

Philippe Pasquier
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
pasquier@sfu.ca

ABSTRACT

Segmentation and classification is an important but time consuming part of the process of using soundscape recordings in sound design and research. Background and foreground are general classes referring to a signal's perceptual attributes, and used as a criteria by sound designers when segmenting sound files. We establish the background / foreground classification task within a musicological and production-related context, and present a method for automatic segmentation of soundscape recordings based on this task. We created a soundscape corpus with ground truth data obtained from a human perception study. An analysis of the corpus showed an average agreement of each class - background 92.5%, foreground 80.8%, and background with foreground 75.3%. We then used the corpus to train a machine learning technique using a Support Vector Machines classifier. An analysis of the classifier demonstrated similar results to the average human performance (background 96.7%, foreground 80%, and background with foreground 86.7%). We then report an experiment evaluating the classifier with different analysis windows sizes, which demonstrates how smaller window sizes result in a diminishing performance of the classifier.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Methodologies and techniques

1. INTRODUCTION

Audio based creative practices, such as sound design and soundscape composition, use recordings to create expressive works. A soundscape recording (or field recording) is a recording of sounds at a given locale at a given time, obtained with one or more fixed or moving microphones. These recordings are traditionally used in sound design production using manual information retrieval and segmentation processes. One challenge in working with soundscape recordings is the huge amount of information they contain. It is not uncommon for recordings to range from 5 minutes to over half an hour in length. When working with many recordings, the process of analyzing and extracting regions becomes exceedingly time-consuming.

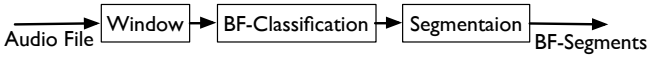

Figure 1: Our BF-Classification approach first divides an audio file into discrete windows before classification and final concatenation into segments.

In this paper, we address the problem of cutting up and labelling sound-files for sound design production and generative systems. Specifically, we describe a technique for automatic classification and segmentation of soundscape recordings. Our technique utilizes a soundscape recording dataset obtained from a human listening study, and an adaptation of well-established methods for classification and feature extraction, all enhanced with feature-selection step.

Our contributions are: 1) The establishment of the background / foreground classification task within a musicological and production-related context; 2) The creation of a soundscape corpus with ground truth data obtained from a human perception study; 3) The adaptation and implementation of a machine learning method, which includes an audio feature selection step; 4) Lastly, we show results from an experiment that evaluates this classification approach when using different analysis window lengths.

The remainder of this paper is organized as follows. In Section 2, we define the classification categories with grounding in the soundscape literature. In Section 3, we discuss the related work in the domain of soundscape classification. Section 4, outlines the method and evaluation of the classifier used in our research, including the acquisition of a soundscape corpus used to train the classifier, and the feature selection method. In Section 5, we show the results of the classifier in an experiment applying different analysis window sizes. Finally, in Section 6, we present our conclusions and suggest directions for future work.

AM '15 October 7-9, 2015, Thessaloniki, Greece.

2. SOUNDSCAPE BACKGROUND AND FOREGROUND CATEGORIES

In this section we define the categories background, foreground, and background with foreground. Background and foreground are general classes referring to a signal's perceptual attributes. These categories are important for sound designers who will mix different recordings in generating artificial soundscapes. A sound can be either background or foreground depending on factors of listening context and attention. For example, Schafer [20] outlines a taxonomy of sound types:

1. Natural sounds: e.g. birds, insects, rain;
2. Human sounds: e.g. laugh, whisper;
3. Sounds and society: e.g. party, church bells, concert;
4. Mechanical sounds: e.g. airplane, machines, cars;
5. Quiet and silence: e.g. dark night, wild space; and
6. Sounds as indicators: e.g. clock, doorbell, siren.

With any of these types of sounds, the external listening context and listener attention influence background and foreground classification. For example, the sound of a drop of water in the bathtub is accentuated by the bathroom environment, whereas it becomes a part of the background texture when in the ocean. A listeners' attention is the second factor in perceiving a sound as background or foreground. For example, the sound from the TV is foreground when a show is being watched, but becomes background when the viewers attention is turned to a conversation in the kitchen.

Truax [22] outlines how listening is a dynamic process of different listening modes. Listening modes can treat any sound as either background or foreground depending on the level of attention being paid at any given moment. However, background listening tends to favour background sound, just as foreground listening tends to favour foreground sounds.

We present a method of segmenting soundscape recordings to address background and foreground sound perception. For simplicity, we call this the BF-Classification problem, and our solution the BF-Classifier. However, our classifier accounts for context but not attention i.e. the drop of water example will work, but the TV example will not.

In regards to listening context, background sounds either seem to come from farther away than foreground sounds, or are continuous enough to belong to the aggregate of all sounds that make up the background texture of a soundscape. This is synonymous with a ubiquitous sound, specified by Augoyard and Torgue[3] as - "a sound that is diffuse, omnidirectional, constant, and prone to sound absorption and reflection factors having an overall effect on the quality of the sound". Urban drones and the hum of insects are two examples of background sound. Conversely, Foreground sounds are typically heard standing out clearly against the background. At any moment in soundscape recording, there may be either background sound, foreground sound or a combination of both.

3. RELATED WORK

The problem of discriminating the background from the foreground has been approached using environmental sound classification and segmentation techniques. For example, Moncrieff et al. discuss the delineation of background and foreground for environment monitoring [17]. Their adaptive model updates what is classified as background over time notifying the system of a foreground event when rapid deviations in the signal occur. Slina et al. [6] present another approach to classification addressing the BF-Classification problem for contextual computing. In their research sound from three separate environments (coffee room, courtyard, and subway) with both background and foreground sound are used to demonstrate their algorithm. They report the detection accuracy of background sound varies between 82.5% and 92.1%, and foreground 63.5% and 75.9% depending on the environmental context.

For the most part, these approaches rely on the monitoring of the time alterations of the occurred sound events, which is different from the BF context here. A wide range of other approaches have been used for modelling audio signals by testing and ranking of different audio features, classifiers, and windowing options. For example, content-based music structure analysis [15], audio identification [4], segmentation and summarization [7], segmentation and classification techniques in surveillance /conference system [13], audio-adaptive bimodal segmentation [1] have put forward different configurations of audio features, classifier, and windowing option to model audio signals for specific applications.

Our BF-Classifier approaches background and foreground classification by analyzing discrete analysis windows from a corpus labelled from a perceptual study. A similar approach to the research here is presented by Aucouturier et al. [2], who suggested a classification technique for modelling different environmental contexts. This technique involves a Gaussian Mixture Model trained with the long-term statistical distribution of Mel Frequency Cepstral Coefficients - accounting for long durations of audio data, and thus presents an attractive model for soundscape classification that often has sounds that evolve over time. However, recent scrutiny of the approach [14] demonstrates this technique does not generalize well across different recordings. Instead, we adapt a solid approach from the music information retrieval literature [24] that models audio features with a Support Vector Machines classifier. Roma et al. [19] select this method for segmenting soundscape sound files. Their segmentation algorithm splits an audio file into 2-second analysis windows for classification according to classes from Gaver's taxonomy [10] of interacting materials. They report an overall classification accuracy of 84.56%.

Our approach adapts the standard segmentation with classification technique to a set of perceptually motivated classes for sound designers and generative systems. We include an audio feature selection step, and evaluate our approach with an experiment on the classifiers performance using progressively smaller analysis windows.

4. BF CLASSIFICATION

The BF-Classifier in our research models the soundscape categories background, foreground, and background with foreground sound. Audio feature vectors were extracted from BF labeled corpus and used to train a Support Vector Machines classifier (SVM). In adopting this supervised machine learning approach, we first created a corpus of training data from a perceptual study.

4.1 Corpus

We created a corpus of soundscape recording samples from the World Soundscape Project Tape Library database [23] (WSPTL). The WSPTL contains five unique collections of soundscape recordings, with a total of 2545 individual sound files amounting to over 223 hours of high quality carefully selected recordings. The collections gathered between 1972 and 2010 comprise of recordings from across Canada and Europe. Recording equipment included a Nagra IV-S field recorder and a pair of AKG condenser microphones. Collections were originally on analog tape and have since been digitized and held online by Simon Fraser University.

We selected 200 4-second samples from the WSPTL covering the six soundscape categories defined by Schafer [20]. We found 4-seconds sufficient length for identifying the context of the sound, which was confirmed by independent listeners. Further, we wanted the corpus to be compact so participants could finish the study with minimum listening fatigue. Additionally, we wanted samples short to preserve their class homogeneity for the machine leaning.

Samples range from indoor and outdoor settings, both with and without music in the soundscape. Expert commentary accompanying recordings demarcates foreground and background regions, and we subjectively selected from these regions based on consistent texture and dynamics. No normalization was applied to the original recordings or the extracted regions. Audio was mixed down to mono; thereby losing stereo information in favour of a higher degree of generality of the system for recordings not obtained with similar high precision equipment, or for those recorded in mono.

There was a total of 31 participants in the study group from the student body at Simon Fraser University, Canada. Before the study, an example for each of the categories, background, foreground, and background with foreground was played, and a short textual description of the classes presented. Participants were asked to use headphones when listening to samples. Samples were played using an HTML5 audio player object. Depending on the browser software, the audio format for the study was either MP3 at 196 kps or Vorbis at an equivalent bit rate. Participants had no time limit and could listen to recordings repeatedly.

Each participant received the 200 samples in a randomized order. They then selected a category from a set of radio buttons after listening to a sample (Figure 2). Participants confirmed each choice when pressing a button to listen to the next segment. On completing the study, the participants classification results were uploaded into a database for analysis.

Results of the study were accumulated to find the most

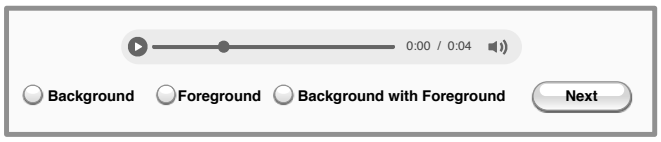

Figure 2: The graphical interface presented to study participants. Responses are entered by the participant using the radio buttons corresponding to background, foreground, and background with foreground. The response is logged when the participant requests the next recording.

agreed upon category for each of the samples using a simple majority. The 30 results with the highest majority for each category were added to the final corpus, and remaining samples disposed. Figure 3 shows the mean lines for the agreement of the final corpus[1].

A quantitative analysis of responses against the final corpus showed the average participant agreed on the categories: background 92.5% (SD=3.6%); foreground 80.8% (SD=9.5%), and; background with foreground 75.3% (SD= 11.3%). The minimum agreement for a single recording categorized as background was 87% while the highest agreement was 100%. Further, the lower quartile and upper quartile, 90.3% and 96.7% respectively, demonstrate that most people share the opinion of which sounds from the corpus belong to the background category. The category foreground shows a less strong consensus. The minimum agreement for a recording of this category was 64.5%, the highest agreement was 96.7%, with the lower quartile and upper quartile 73.3% and 90.3% respectively. Similarly, the category background with foreground shows the minimum agreement for a recording was 61.2%; the highest agreement was 96.7%, with a lower quartile of 64.5% and an upper quartile of 87%.

4.2 Audio Features

Audio features are automatically chosen from a recursive feature elimination and selection step. A large set of audio features [18] was extracted from all 4-second samples in the labeled soundscape corpus using the YAAFE software [16]. Audio was resampled from 44100Hz AIF format to 22500 Hz. The analysis step is 512 samples with a Hamming window of 1024 samples. The mean and standard deviation of features was calculated and logged. This windowing configuration and subsequent analysis step results in a high descriptive power for representing the texture of the sound and overall dynamics of the audio signal. Since we achieved good results with this method we did not explore other window configurations.

We applied a method of dimension reduction for features. We split the corpus into a test and validation set of the corpus. The test set was used for selecting features and contained 20% of the corpus while the validation set kept the remaining 80%. A method of recursive feature elimination

[1]Corpus and dataset accessed April 2015 `http://www.sfu.ca/~mthorogo/bfcorpus/`.

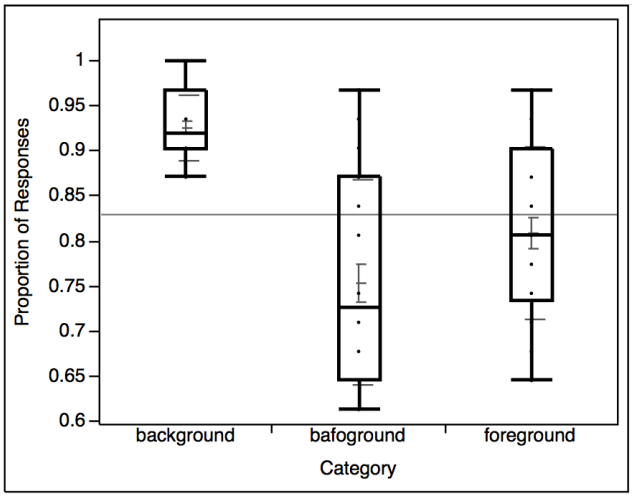

Figure 3: Box plots and mean lines for the agreement of labels for the corpus of background, foreground, and background with foreground recordings.

with a SVM classifier [11] using the WEKA software [12] evaluated the features. We selected the top 10th percentile of ranked audio features for our experiment.

Table 1 shows the reduced set of descriptors. The audio feature set contains spectral and perceptual audio descriptors, including the means and standard deviations of Mel Frequency Cepstral Coefficients, total loudness, perceptual spread, and spectral flux. Perceptual descriptors such as these model the human auditory system, which is desirable from the point of view of soundscape studies, where the perception of the human listener is an important consideration.

Audio Features
MFCC mean (coef. 8,11,15,28,36)
MFCC std dev (coef. 1,2,5,6,18,20,32,34)
Total Loudness mean & std dev
Perceptual Spread mean
Spectral Flux std dev

Table 1: The set of audio features output from the analysis of the soundscape corpus test set.

4.3 SVM Classifier

A Support Vector Machines (SVM) classifier is a binary non-probabilistic linear classifier that learns the optimal separating hyperplane of the data with the maximum margin. Non-linear decision boundaries, as is common with complex environmental sound, can be represented linearly in a higher dimension space than the input space with a kernel function. Additionally, the SVM can be extended for multi-class problems such as our BF classification problem using the one-versus-the-rest approach. We use the C-support vector classification (C-SVC) algorithm with a linear kernel suited to smaller feature vectors and training sets [5].

4.4 Evaluating the BF-Classifier

The classifier was trained with features and labels from the corpus test set, and evaluated with the corpus validation set. We performed an evaluation of the BF-Classifier using a 10-fold cross-validation strategy on the corpus validation set. This method randomly partitions the validation set into $k = 10$ equally sized sub-samples before iteratively testing the remaining sub-samples against each k-partition. The results summary is shown in Table 2. The classifier achieved an overall true positive rate of 87.77%. An inter-rater reliability analysis using the kappa statistic was performed to determine consistency of the classification. The kappa statistic of 0.8167 shows a strong reliability of the classification results over the 10-fold validations.

True Positive	87.77%
False Positve	12.22 %
Kappa statistic	.8167

Table 2: Average true positive and false positive classification of SVM classifier.

In Table 3, the true positive rate for background classification (96.7%) shows most samples identified background were labelled as such. Foreground (80%) and background with foreground (86.7%) classes show similarly high true positive rate demonstrating the BF-Classifier correctly classified a majority of these samples correctly.

Class	True positive rate
B	96.7%
F	80%
BF	86.7%

Table 3: Detailed accuracy by class of SVM classifier for the categories background (B), foreground (F), and background with foreground (BF).

5. DIMINISHING ANALYSIS WINDOWS

The corpus evaluation described in Section 4.4. is based on the mean and standard deviation of features over a 4-second length window. It is desirable for BF-Classifier to delineate precisely the segment boundary to a varying degree of window lengths. Hence, we conducted an experiment to evaluate the classifier on smaller analysis windows.

In this experiment, we evaluated the classifier on 2-second, 1-second, 500-millisecond, 250-millisecond, and 125-millisecond analysis windows to ascertain if performance degrades with diminishing analysis windows. First, we generated a larger ground truth corpus of BF labeled segments for generalizing the classifier performance under the test conditions. Labels were automatically applied to samples in the corpus using the trained BF-Classifier described in Section 4.3. We generated the ground truth corpus for this experiment from recordings in the commercially available Sound Ideas XSeries sound effects database[2]. Those recordings have been

[2]Sound Ideas website accessed April 17, 2015 `www.sound-ideas.com`

professionally curated with a similar range of foci to the WSPTL corpus described in Section 4.1. The BF-Classifier was used to segment a subset of the files from the database.

We applied the following method of refining the corpus. Firstly, adjacent segments with the same label were concatenated. Next, we extracted a 4-second span centred on the mid-point of regions longer than two segments (i.e > 8 seconds). Lastly, the extracted segments were run through the BF-Classifier for verification with the initial classification. Samples violating the original classification were rejected. One remaining segment from each analyzed file was chosen at random resulting in 142 foreground, 407 background, and 171 background with foreground samples in the corpus[3].

Next, each labeled segment with the different length analysis windows was classified and the results logged. This data was analyzed using established music information retrieval methods of precision, recall, and F-Measure [8]. Figure 4. shows the precision, recall, and F-Measure of the BF-Classifier on analysis windows of 4 second, 2 second, 1 second, 500 milliseconds, 250 milliseconds, and 125 milliseconds. An F-Measure of 0.0 demonstrates poorest performance, while an F-Measure of 1.0 means a perfect retrieval. Although we expect a 4-second window to achieve perfect recall, we include it here as an indication of the change in classification performance with smaller analysis windows.

The BF-Classifier performance remained high for all analysis windows for background, with only a moderate rate of decline (F: 1.0, 0.91, 0.84, 0.84, 0.8, 0.78). Background with foreground classification exhibited by far the greatest performance losses (F: 1.0, 0.78, 0.44, 0.44, 0.34, 0.19). That rapid decline corresponds to smaller analysis windows, and is not surprising since the unique combination of background and foreground sounds can cause moment to moment classification errors for this class. Foreground classification was reasonably stable after an initial decrease in performance (F: 1.0, 0.72, 0.72, 0.64, 0.48).

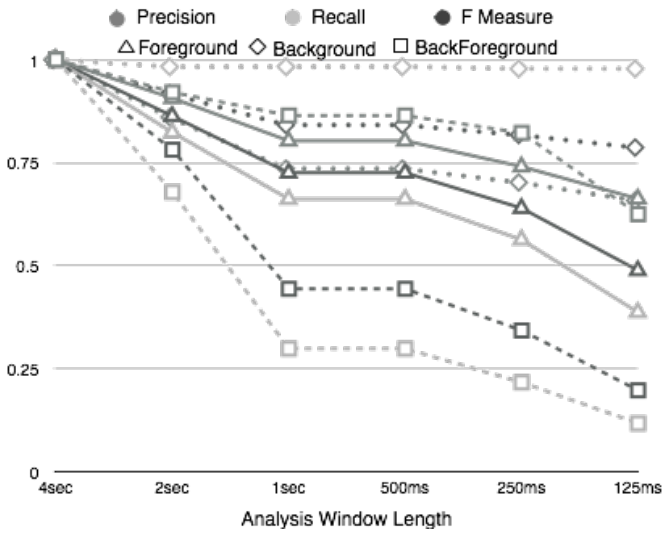

Figure 4: Graph showing the precision (grey), recall (light grey), and F-Measure (dark grey) of the BF-Classifier on analysis windows of 4 second, 2 second, 1 second, 500 milliseconds, 250 milliseconds, and 125 milliseconds. The results for each set of experiments is given for foreground (triangle), background (diamond), and background with foreground (square).

6. CONCLUSIONS AND FUTURE WORK

The BF-Classifier here classifies fixed-length analysis windows across the length of the audio file, providing a quick means of indicating where a difference in classification occurs. For example, we could use a 250ms rectangular non-overlapping analysis window with a BF-Classifier to segment an audio file; keeping in mind the trade-off between boundary resolution and classification accuracy when using different sized analysis windows. However, the results of the BF-Classifier here demonstrate that an analysis window of this size will obtain a high degree of performance in delineating background segments from those with foreground.

The BF-Classifier can automatically iterate over the length of the audio file while classifying and labelling segments with BF-classes with a much greater speed than if done by hand [4]. Further, we have described the creation of a soundscape recording corpus generated from results of a perceptual study with human participants. Further, we conducted an evaluation of the corpus showing it can be modelled using machine learning techniques with performance closely correlated to the average human classification. Next, we used well-established MIR techniques to observe the effect of how different window lengths affect our classification approach.

In future work, we will explore the problem of connecting fragmented sounds to address the problem of grouping audio regions of sounds with longer temporal evolution. For example, a limited lookahead-search, or median filter could be applied to link sounds spread across longer regions by concatenating adjacent and more isolated segments with the same label.

Moreover, soundscape classification continues to provide many challenges. Not in the least is the subjective interpretation of soundscape, demonstrated by the disparity between participants classifications of soundscape samples. We have shown in other work [21, 9] the feasibility of modelling properties of soundscape, such as affective representations of pleasantness and eventfulness. Perception-based classification and segmentation of soundscape recordings will be tremendously useful for sound designers in research and creative practice. As part of out larger research goals, we will be applying these techniques to computer-assisted tools for sound designers, and generative systems.

7. ACKNOWLEDGMENTS

We would like to acknowledge the National Science and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council of Canada for their ongoing financial support. Thank you to Professor Barry

[3]Corpus and dataset accessed April 2015 `http://www.sfu.ca/~mthorogo/bfcorpus/`.

[4]A demonstration of the BF-Classifier can be accessed at `http://www.audiometaphor.ca/bfclassifer`

Truax for his guidance and knowledge he imparted, contributing to this research. Finally, we would like to thank the reviewers, who through their thoughtful comments have been assisting with this publication.

8. REFERENCES

[1] E. Akdemir and T. Ciloglu. Bimodal automatic speech segmentation based on audio and visual information fusion. *Speech Communication*, 53(6):889 – 902, 2011.

[2] J.-J. Aucouturier and B. Defreville. Sounds like a park: A computational technique to recognize soundscapes holistically, without source identification. *19th International Congress on Acoustics*, 2007.

[3] J. Augoyard and H. Torgue. *Sonic Experience: A Guide to Everyday Sounds*. McGill-Queen's University Press, 2006.

[4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41(3):271–284, 2005.

[5] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems Technology*, 2(3):27:1–27:27, 2011.

[6] S. Chu, S. Narayanan, and C.-C. Kuo. A semi-supervised learning approach to online audio background detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009*, pages 1629–1632, 2009.

[7] M. L. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proceedings of the International Symposium on Music Information Retrieval*, 2002.

[8] S. Downie, A. Ehmann, M. Bay, and C. Jones. The music information retrieval evaluation exchange: Some observations and insights. In *Advances in Music Information Retrieval*, volume 274, pages 93–115. Springer Berlin Heidelberg, 2010.

[9] J. Fan, M. Thorogood, and P. Pasquier. Impress: Automatic recognition of eventfulness and pleasantness of soundscape. In *Proceedings of the 10th Audio Mostly*, Thessaloniki, Greece, 2015.

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

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[13] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *In Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1423–1426, 2000.

[14] M. Lagrange, G. Lafay, B. Defreville, and J.-J. Aucouturier. The bag-of-frames approach: a not so sufficient model for urban soundscapes. *arXiv preprint arXiv:1412.4052*, 2014.

[15] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 112–119, New York, NY, USA, 2004.

[16] B. Mathieu, S. Essid, T. Fillon, J.Prado, and G.Richard. YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software. In *Proceedings of the 2010 International Society for Music Information Retrieval Conference*, 2010.

[17] S. Moncrieff, S. Venkatesh, and G. West. Online audio background determination for complex audio environments. *ACM Transactions on Multimedia Computing and Communications Applications*, 3, 2007.

[18] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, 2004.

[19] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra. Ecological acoustics perspective for content-based retrieval of environmental sounds. *EURASIP Journal of Audio Speech Music Process*, 7:1–11, 2010.

[20] R. Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1977.

[21] M. Thorogood and P. Pasquier. Impress: A machine learning approach to soundscape affect classification for a music performance environment. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 256–260, Daejeon, Republic of Korea, May 27-30 2013.

[22] B. Truax. *Acoustic Communication: Second Edition*. Ablex Publishing, 2001.

[23] B. Truax. World Soundscape Project - Tape Library, 2015. Available online at `http://www.sfu.ca/sonic-studio/srs/index2.html`; visited on January 12th 2015.

[24] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.