Soundscape Audio Signal Classification and Segmentation Using Listeners Perception of Background and Foreground Sound *

MILES THOROGOOD, AND JIANYU FAN, AND PHILIPPE PASQUIER
(mthorogo@sfu.ca) (jianyuf@sfu.ca) (pasquier@sfu.ca)

Simon Fraser University, SIAT, Canada

Classification and segmentation are important but time consuming tasks when using soundscape recordings in sound design and research. Background and foreground are criteria when segmenting sound files according to a signal's perceptual attributes. We establish the background and foreground classification task within a musicological and soundscape context, and present a method for the automatic segmentation of soundscape recordings based on this task. We present a soundscape corpus with ground truth data obtained from a human perception study. An analysis of the corpus shows participants have a high level of agreement on the category assigned to background samples (92.5%), foreground samples (80.8%), and background with foreground samples (75.3%). We verify the corpus by training a Support Vector Machines classifier. An analysis of the classifier demonstrates a similarly high degree of certainty for background 96.7%, foreground 80%, and background with foreground 86.7%. Further, we report an experiment evaluating the classifier with different analysis windows sizes, and demonstrate how smaller window sizes affect the performance of the classifier. The classifier is then implemented in a segmentation system. We present the results of an evaluation on three segmentation systems: median filter, k-depth lookahead, and a probabilistic algorithm selecting class association.

0 INTRODUCTION

A soundscape recording (or field recording) is a recording of sounds at a given location at a given time, obtained with one or more fixed or moving microphones. Audio-based creative practices, such as sound design and soundscape composition, and problems presented by audio scene monitoring, require analysis and segmentation of the complex soundscape audio signals to make them relevant. The sounds in a soundscape are background or foreground depending on their salient characteristics, such as proximity, repetition, and spectral attributes. Further, background and foreground sounds often occur simultaneously in a soundscape. Another challenge of working with soundscape recordings is that it is common for recordings to be several hours in length, with current recording systems allowing for days of recording. When working with soundscape recordings, such as audio scene monitoring and sound design, the process of analyzing and extracting regions becomes exceedingly time-consuming.

We address the problem of segmenting and labelling sound files into the background, foreground, and background with foreground classes. In Section 1, we establish the background/foreground classification task within a musicological and production-related context with grounding in the soundscape literature. In Section 2, we discuss related work in the domain of soundscape classification. Next, in Section 3, we outline the creation of a soundscape corpus obtained from an auditory perception experiment. We show results from an experiment evaluating a classifier trained with this corpus. Then, in Section 4, we describe the procedure that tests the hypothesis that shorter analysis windows provide better boundary precision at the cost of classifier performance. Having established the parameters of the analysis window size, we investigate different segmentation algorithms and provide the evaluation of these algorithms in section 5. Finally, in Section 6, we present our conclusions and suggest directions for future work.

*To whom correspondence should be addressed Tel: +1-250-807-9266; Fax: +1-250-807-990; e-mail: mthorogo@sfu.ca

1 BACKGROUND AND FOREGROUND SOUND 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 mix different recordings when generating artificial soundscapes. Any sound can be either background or foreground depending on factors of listening context and attention. For example, the noise of a drop of water in a bathtub is accentuated by the bathrooms 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, noise from a TV is foreground when a show is watched, but becomes background when the viewers attention is turned to a conversation in the kitchen.

Truax [30] 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. Our classifier accounts for context but not attention, primarily because the system does not have the ability to model attention. i.e. the drop of water example will work, but the TV example will not unless the conversation in the kitchen is more prominent in the signal than the TV show.

In regard 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. The background texture of a soundscape is synonymous with ubiquitous sound, specified by Augoyard and Torgue [3] as - "a diffused sound that is 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 as standing out clearly against the background. In soundscape recording, there may be either background sound, foreground sound or a combination of both.

2 RELATED WORK

From a listeners perspective, the background and foreground of a soundscape account for the disparity of different sounds in the environment. When analyzing and combining different soundscape recordings in research or practice, segmenting the audio file based on these categories is an important task. The literature on sound design research demonstrates similar approaches of selecting specific sections of recordings from both semantic and salient criteria. For example, Eigenfeldt et al. [10], Janer et al. [17], and Thorogood et al. [29] use a hand-selected procedure of curating a corpus of recordings for generative systems.

The problem of automatic discrimination of background and foreground sound has been approached using environmental sound classification and segmentation systems. Moncrieff et al. discuss the delineation of background and foreground for environment monitoring [22]. Their adaptive model updates what is classified as the background over time, notifying the system of a foreground event when rapid deviations in the signal occur. Slina et al. [7] present another approach to classification, addressing the BF-Classification problem for contextual computing. Slina et al. demonstrate the algorithm using three separate environments (a coffee room, courtyard, and subway) with both background and foreground sound. They report that 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 time alterations of events, which is different from the BF context here that classifies discrete windows of audio features from a signal for class association. A wide range of other approaches model audio signals by testing and ranking various audio features, classifiers, and windowing options. For example, content-based music structure analysis [20], sound identification [5], segmentation and summarization [8], segmentation and classification techniques in surveillance/conference system [18], and audio-adaptive bimodal segmentation [1] have put forward different configurations of audio features, classifiers, and windowing options to model audio signals for specific applications.

Aucouturier et al. [2] present a method of differentiating between environmental sound contexts, such as park and urban. They suggest a classification technique for modelling these environmental contexts. In this technique a Gaussian Mixture Model is 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 [19] demonstrates that this technique does not generalize well across different recordings. Instead, we adapt a solid approach from the music information retrieval literature [32], modelling audio features with a Support Vector Machines classifier. Roma et al. [25] select this method for segmenting soundscape sound files based on Gaver's taxonomy [14] of interacting materials. Their algorithm segments and classifies an audio file into 2-second analysis windows with an overall classification accuracy of 84.56%.

Prior work in audio segmentation has focussed on evaluating segmentation systems by adapting metrics of established music information retrieval methods, such as of precision, recall, accuracy, and F-Measure [9]. Galliano et al. [13] describe these measures for the ESTER evaluation

criteria as the aggregate duration of inserted class events relative to ground truth segment boundaries, and where they occur (recall) and where they are detected (precision). Temko et al. [27] outline a weighted F-Measure, precision, and recall when defining metrics analyzing segmentation systems for the application of acoustic events. Ramona and Gel [24] report the F-Measure when comparing implementations of SVM classification to segmenting music, speech, and mixed signals. Wichern et al. [33] report the mean average accuracy for an HMM segmentation system regarding performance, ranging between 0.125 and 0.567 depending on different test conditions.

Our technique of classification and segmentation models background and foreground sound, a set of perceptually motivated classes used by sound designers and researchers. We include an audio feature selection step in our technique, and evaluate our approach with an experiment on the classifiers performance using progressively smaller analysis windows. We evaluat the different segmentation systems using precision, recall, and F-Measure. The contributions presented in this paper include: establishing the background/foreground classification task within a musicological and production-related context; creating a background/foreground labelled soundscape corpus using human participants; describing and presenting an experiment testing the effect of analysis window size on boundary precision; finally, we establish and evaluate different segmentation algorithms for fragmenting the background/foreground parts of soundscape audio signals.

3 BF-CLASSIFER

Our BF-Classifier models the soundscape categories background, foreground, and background with foreground sound. We extract audio feature vectors from the BF labeled corpus, which is used to train a Support Vector Machines classifier (SVM). In adopting this supervised machine learning approach, we first create a corpus of training data from a perceptual study.

3.1 Corpus

We create the soundscape recording corpus from the World Soundscape Project Tape Library database [31] (WSPTL). The WSPTL contains five unique collections of soundscape recordings, with a total of 2545 individual sound files amounting to over 223 hours of carefully selected recordings. The collections gathered between 1972 and 2010 are comprised of recordings from across Canada and Europe. The researchers use a Nagra IV-S field recorder and a pair of AKG condenser microphones. Collections have since been digitized at 44.1kHz 16bit and stored online at Simon Fraser University.

We select 200 4-second samples from the WSPTL. Independent listeners confirmed 4 seconds is a sufficient length for identifying the context of the sound. Additionally, the corpus is compact so participants finish the study with minimum listening fatigue. Further, the short samples to preserve their class homogeneity for the machine learning. The types of sounds cover the following six soundscape categories defined by Schafer [26].

- Natural sounds: bird, chicken, rain, sea shore;
- Human sounds: laugh, whisper, shouts, talk, cough;
- Sounds and society: party, concert, grocery store;
- Mechanical sounds: engine, cars, air conditioner;
- Quiet and silence: wild space, silent forest;
- Sounds as indicators: clock, doorbell, siren.

Samples in the corpus range from indoor and outdoor settings, both with and without music in the soundscape. The expert commentary accompanying recordings demarcates foreground and background regions, and we subjectively select from these regions based on consistent texture and dynamics. No normalization is applied to the original recordings or the extracted regions. The audio is mixed down to mono. Thereby, stereo information is lost 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.

The study group consists of 31 participants from the student body at Simon Fraser University, Canada. Before the study, an example of each of the categories, background, foreground, and background with foreground is played, and a short textual description of the classes presented. Participants are asked to use headphones when listening to samples. Samples are played using an HTML5 audio player object. Depending on the browser software, the audio format for the study is either MP3 at 196 kps or Vorbis [12] at an equivalent bitrate. Participants have no time limit and can listen to recordings repeatedly.

Each participant receives the 200 samples in a randomized order. They then select a category from a set of radio buttons after listening to a sample (Figure 1). Participants confirm a choice by pressing a button to hear to the next segment. Upon completion of the study, participants classification results are uploaded onto a database for analysis.

The accumulated study results are used to find the most agreed upon category for each of the corpus samples using a simple max operation. We added the 30 results with the agreement for each category for the final corpus, and disposed of the remaining samples. Figure 2 shows the mean lines for the participant agreement of class association for the selected samples[1].

A quantitative analysis of responses against the final corpus show that a participant agrees on 92.5% (SD=3.6%) of the background samples class association, 80.8% (SD=9.5%) of the foreground samples class association, and 75.3% (SD= 11.3%) of the background with foreground samples class association. The minimum agreement for a single recording categorized as background is 87% while the highest agreement is 100%. Further, the lower quartile and upper quartile, 90.3% and

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

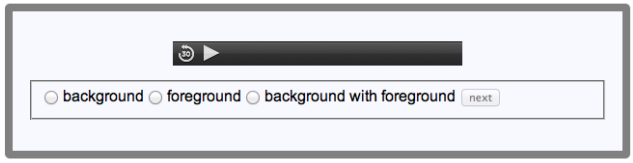


Fig. 1. The graphical interface presented to study participants. Responses are entered by the participant using the radio buttons corresponding to background, foreground, and background with foreground. The response is logged when the participant requests the next recording.

96.7% respectively, demonstrate that most people share the opinion on 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 class is 64.5%, the highest agreement is 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 as 61.2%; the highest agreement is 96.7%, with a lower quartile of 64.5% and an upper quartile of 87%.

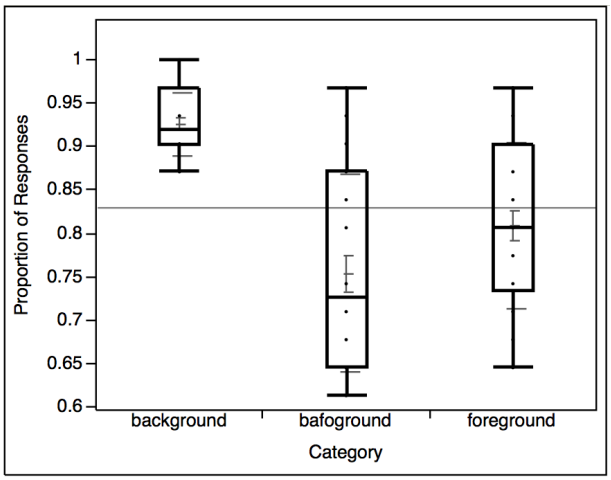


Fig. 2. Box plots and mean lines for the agreement of labels for the corpus of background, foreground, and background with foreground recordings. The light grey line represents the overall mean agreement for the three classes.

3.2 Audio Feature Selection

A recursive feature elimination and selection step automatically selects audio features from a larger set, as defined in [23, 4], extracted from all 4-second samples in the labeled soundscape corpus using the YAAFE software [21]. We resampled the audio from 44100Hz AIF format to 22500 Hz and applied a Hamming window of 1024 samples with 512 samples overlapping. The mean and standard deviation of features is calculated and logged. This windowing configuration and subsequent analysis step result in a high descriptive power for representing the texture and overall dynamics of the sound. Since we achieve good results with this method, we did not explore other window configurations.

We apply a dimension reduction method for features. We split the corpus into a training set for selecting features and a validation set for evaluating the classifier. 20% of the corpus is allocated to the training set, with the remaining 80% allocated to validating the classifier. Features are recursively eliminated using the training set and an SVM technique [15] implemented in the WEKA software [16]. We select the top 10th percentile of ranked audio features for our experiment. Table 3.2 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. We think it significant that perceptual features that model the human auditory system perform better than those that do not in this classification task, where the perception of the human listener is an important consideration. As such, properties of the selected features, such as loudness response curves, apply well to soundscape-related classification tasks.

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

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

3.3 Support Vector Machines

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 algorithm with a linear kernel suited to smaller feature vectors and training set [6].

3.4 Evaluation of the Classifier

We trained the classifier with features and labels from the corpus training set and evaluated with the corpus validation set. We perform 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 3.4. The classifier achieves an overall true positive rate of 87.77%. An inter-rater reliability analysis using the kappa statistic de-

termines the consistency of the classification. In this case, the kappa statistic of 0.8167 shows a strong reliability of the classification results over the 10-fold validations.

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

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

In Table 3.4, the true positive rate for background classification (96.7%) shows most of the samples identified as background were labelled as such. The BF-Classifier correctly classified a majority of the foreground (80%) and background with foreground (86.7%) samples correctly, showing a similarly high true positive rate for these classes.

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

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

4 DIMINISHING ANALYSIS WINDOWS

The corpus evaluation described in Section 3.4. is based on the mean and standard deviation of features over a 4-second length window since it is the size of the sounds humans were using for the classification task of the corpus. It is practical for the BF-Classifier to delineate more precisely the segment boundary using smaller window lengths. Hence, we conduct an experiment to evaluate the classifier on smaller analysis windows.

In this experiment, we evaluate the classifier on 2-second, 1-second, 500-millisecond, 250-millisecond, and 125-millisecond analysis windows to ascertain if performance degrades with diminishing window lengths. First, we generate a ground truth corpus of BF labeled segments for setting a benchmark of the classifier performance and generalizing the classifier under the test conditions. Labels are automatically applied to samples in the corpus using the trained BF-Classifier described in Section 3.3. We generate the ground truth corpus for this experiment from recordings in the commercially available Sound Ideas XSeries sound effects database[2]. Those recordings are professionally curated with a similar range of foci to the WSPTL corpus described in Section 3.1. The BF-Classifier was used to segment a subset of the files from the database.

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

Next, the BF-Classifier classifies each labeled segment with the different length analysis windows and we log the results. We analyze this data using established music information retrieval methods of precision, recall, and F-Measure [9]. Figure 3. shows the precision, recall, and F-Measure of the BF-Classifier on analysis windows of 4 seconds, 2 seconds, 1 second, 500 milliseconds, 250 milliseconds, and 125 milliseconds. An F-Measure of 0.0 demonstrates the poorest performance while an F-Measure of 1.0 means 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 remains 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 exhibits 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, which is not surprising since the unique combination of background and foreground sounds can cause the moment to moment classification errors for this class. Foreground classification is reasonably stable after an initial decrease in performance (F: 1.0, 0.72, 0.72, 0.64, 0.48).

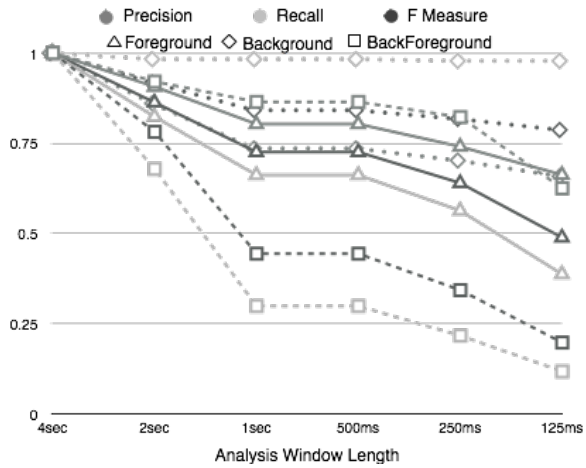


Fig. 3. Precision (grey), recall (light grey), and F-Measure (dark grey) of the BF-Classifier on analysis windows of 4 seconds, 2 seconds, 1 second, 500 milliseconds, 250 milliseconds, and 125 milliseconds. Foreground (triangle), background (diamond), and background with foreground (square).

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

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

5 SEGMENTATION

In this section, we propose a segmentation algorithm for complimenting the BF classifier. Three different approaches to segmentation are evaluated: a median filter, k-depth lookahead, and a technique maximizing posterior probability for BF-classes. The aim of the segmentation algorithm here is to group BF-classified windows perceived as belonging to the same class. We describe the corpus used for experiments. Next, the segmentation algorithms are defined. Finally, we evaluate the segmentation algorithms and report on the results of the experiment.

5.1 Corpus

The following evaluation experiments are carried out on a ground truth set of BF-labelled samples. The ground truth set contains 600 BF samples. To remove any bias toward the classifier on the sample length, the segment boundaries vary in duration between 3 and 6 seconds. Further, to keep the number of transition between classes fair, samples are arranged by BF-label permutations that are concatenated in such a way as no BF-label repeats, i.e. the transition from one sample to the next is always a different BF-label.

To obtain the ground truth set, we generate a corpus of BF-labelled regions with a 6-second duration using the method outlined in Section 3.4. Ground truth samples are chosen at random and truncated between 50% and 100% of the original duration based upon a uniform random sampling. The total duration of segment boundaries labelled with foreground is approximately 890 seconds, background 890 seconds, and background with foreground is 900 seconds. Experiments evaluating the segmentation algorithms are carried out using this corpus and the BF-Classifier described in section 3.3, with audio feature statistics on 250ms analysis windows.

5.2 Median Filter

A simple median filter smooths the SVM confidence intervals on classified analysis windows [24]. The window size has been empirically tuned to 7 frames, which corresponds to a 1.75s window. Next, the class for maximizing the posterior probability is selected for analysis windows. Finally, adjacent analysis windows with the same BF-label are concatenated.

5.3 k-depth Lookahead

The k-depth segmentation system operates by looking ahead for a BF-label and backtracking to reclassify, and as such, relabels segments when it encounters the initial class. The lookahead length k is parameterized for concatenating segments to different depths. For example (see Figure 4.), $k = 3$ will conduct a label equality test with the analysis windows three positions ahead, and decrement the position until the result of the test returns true, or the starting position is reached. A small k value will result in grouping segments with the same label for sounds with relatively short duration but at the cost of losing the coherence of more sparsely distributed sounds. A k value of 1 (one) has the same effect as concatenating adjacent segments with the same class label. Larger k values will result in grouping intermittent sounds spread over larger intervals, but with a greater likelihood of grouping sounds not belonging together. The value of k has been empirically tuned to look ahead seven frames, corresponding to 1.75 seconds.

5.4 Maximizing Posterior Probability

The Maximizing Posterior Probability (MPP) algorithm computes the function maximizing the posterior probability of a class based on symbolic level BF-labelled windows. The algorithm iterates over the length of the sound file computing the class probability from a subset of the windows (see Figure 5.). In the case of ties, the initial analysis label is given precedence. The size of the subset is empirically tuned to a length of 4, corresponding to 1 second.

5.5 Metrics

We use an evaluation scheme with three standard metrics - precision, recall, and F-Measure, to evaluate the segmentation algorithms on their ability to retrieve items for a given BF-class, and the relevance of those items retrieved [13]. These measures take into account the aggregate duration of inserted class events in relation to ground truth segment boundaries. Recall r denotes how many of the items detected are relevant while precision p informs us how many relevant items are detected. The F-measure F is the harmonic mean of these two measures, calculated as $F = 2\frac{pr}{p+r}$.

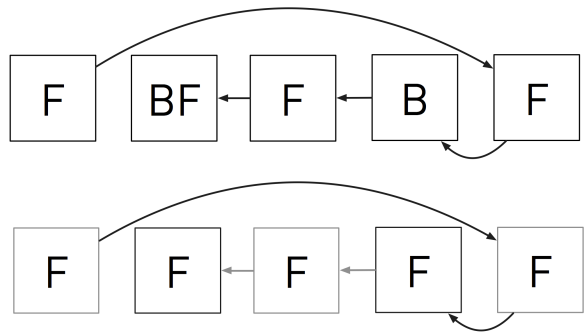


Fig. 4. k-depth segmentation system jumping $k = 3$ then searching $k - 1$ and backtracking to relabel windows.

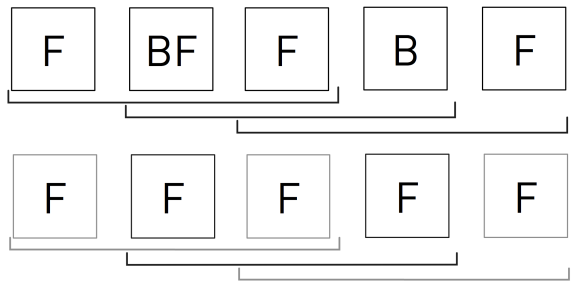


Fig. 5. MPP segmentation system with span=3. In this case, all windows are relabelled as foreground.

5.6 Results

We compute the precision, recall, and F-Measure to compare the performance of the segmentation systems, median filter, k-depth lookahead, and MPP (see Table 4). The k-depth lookahead technique achieved a higher mean average precision (0.812) than the MPP (0.802), and median filter (0.687). In regard to recall, the MPP technique shows a mean average performance (0.828) slightly higher than the k-depth lookahead algorithm (0.824). Both these results are superior to those achieved by the median filter (0.755). Further, the k-depth lookahead algorithm achieves an F-Measure (0.813), and then the MPP (0.799), and median filter (0.678), respectively.

These results suggest that the k-depth lookahead technique achieves better performance segmenting the corpus than the other two approaches. However, it should be noted that this performance approximates the measures of the classifier given the analysis window of 250ms, which evidently does not improve the overall precision and recall of classification.

Table 4. Mean average precision, recall, and F-Measure for median filter, k-depth look-ahead, and max posterior segmentation systems.

Segmentation type	precision	recall	F-Measure
median	0.687	0.755	0.678
k-depth	0.812	0.824	0.813
MPP	0.802	0.828	0.799

6 CONCLUSIONS

The BF-Classifier classifies fixed-length analysis windows across the length of the audio file, providing a quick means of indicating where a difference in classification occurs. We demonstrate the BF-Classifier with a 250ms rectangular non-overlapping analysis window by implementing a segmentation system to segment an audio file. We note the trade-off between boundary resolution and classification accuracy when using different sized analysis windows. The results of the BF-Classifier and further segmentation approach highlight that an analysis window of this size will obtain a high degree of performance in delineating background segments from those with foreground sounds.

Having the BF-Classifier automatically iterate over the length of the audio file while classifying and labelling segments with BF-classes happens at a much greater speed than if done by hand [4]. To obtain these results we describe the creation of a soundscape recording corpus generated from the results of a perceptual study with human participants. Then, 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. Finally, we explored the problem of connecting fragmented sounds to address the issue of grouping audio regions of sounds with longer temporal evolution with three segmentation systems: median filter, k-depth lookahead, and maximizing posterior probability.

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 [28, 11] the feasibility of modelling properties of a soundscape, such as affective representations of pleasantness and eventfulness. The perception-based classification and segmentation of soundscape recordings will be tremendously useful for sound designers in research and creative practice. As part of our larger research goals, we will be applying these techniques to computer-assisted tools for sound designers and researchers.

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

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 Truax for the guidance and knowledge he imparted in contributing to this research.

8 REFERENCES

[1] Eren Akdemir and Tolga Ciloglu. Bimodal automatic speech segmentation based on audio and visual information fusion. *Speech Communication*, 53(6):889 – 902, 2011. http://dx.doi.org/10.1016/j.specom.2011.03.001.

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

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

[4] B.Mathieu. YAAFE Features [online]. 2010. Available at `http://yaafe.sourceforge.net/features.html`; visited on May 8th 2016.

[5] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41(3):271–284, 2005. http://dx.doi.org/10.1007/s11265-005-4151-3.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems Technology*, 2(3):27:1–27:27, 2011. http://dx.doi.org/10.1145/1961189.1961199.

[7] Selina Chu, S. Narayanan, and C.-C.J. Kuo. A semi-supervised learning approach to online audio background

detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1629–1632, 2009. http://dx.doi.org/10.1109/ICASSP.2009.4959912.

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

[9] Stephen Downie, Andreas. Ehmann, Mert Bay, and Cameron 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. http://dx.doi.org/10.1007/978-3-642-11674-2_5.

[10] Arne Eigenfeldt and Philippe Pasquier. Negotiated content: Generative soundscape composition by autonomous musical agents in coming together: Freesound. In *Proceedings of the Second International Conference on Computational Creativity*, pages 27–32, 2011.

[11] Jianyu Fan, Miles Thorogood, and Philippe Pasquier. Impress: Automatic recognition of eventfulness and pleasantness of soundscape. In *Proceedings of the 10th Audio Mostly*, Thessaloniki, Greece, 2015. http://dx.doi.org/10.1145/2814895.2814927.

[12] Xiph.Org Foundation. Vorbis I Specification [online]. 2015. Available at `http://xiph.org/vorbis/doc/Vorbis_I_spec.html`; visited on May 8th 2016.

[13] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586, 2009.

[14] William W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5:1–29, 1993. http://dx.doi.org/10.1207/s15326969eco0501_1.

[15] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002. http://dx.doi.org/10.1023/A:1012487302797.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. http://dx.doi.org/10.1145/1656274.1656278.

[17] Jordi Janer, Gerard Roma, and Stefan Kersten. Authoring augmented soundscapes with user-contributed content. In *ISMAR Workshop on Authoring Solutions for Augmented Reality*, 2011.

[18] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alen Waibel. Strategies for automatic segmentation of audio data. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1423–1426, 2000. http://dx.doi.org/10.1109/ICASSP.2000.861862.

[19] Mathieu Lagrange, Grégoire Lafay, Boris Defreville, and Jean-Julien Aucouturier. The bag-of-frames approach: a not so sufficient model for urban soundscapes. *arXiv preprint arXiv:1412.4052*, 2014. http://dx.doi.org/10.1121/1.4935350.

[20] Namunu C. Maddage, Changsheng Xu, Mohan S. Kankanhalli, and Xi 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. http://dx.doi.org/10.1145/1027527.1027549.

[21] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël 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.

[22] Simon Moncrieff, Svetha Venkatesh, and Geoff West. Online audio background determination for complex audio environments. *ACM Transactions on Multimedia Computing and Communications Applications*, 3, 2007. http://dx.doi.org/10.1145/1230812.1230814.

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

[24] Mathieu Ramona and Gel Richard. Comparison of different strategies for a svm-based audio segmentation. In *Signal Processing Conference, 2009 17th European*, pages 20–24. IEEE, 2009.

[25] Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera, and Xavier Serra. Ecological acoustics perspective for content-based retrieval of environmental sounds. *EURASIP Journal of Audio Speech Music Process*, 7:1–11, 2010. http://dx.doi.org/10.1155/2010/960863.

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

[27] Andrey Temko, Climent Nadeu, Dušan Macho, Robert Malkin, Christian Zieger, and Maurizio Omologo. *Computers in the Human Interaction Loop*, chapter Acoustic Event Detection and Classification, pages 61–73. Springer London, London, 2009. http://dx.doi.org/10.1007/978-1-84882-054-8_7.

[28] Miles Thorogood and Philippe 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.

[29] Miles Thorogood, Philippe Pasquier, and Arne Eigenfeldt. Audio metaphor: Audio information retrieval for soundscape composition. In *Proceedings of the 6th Sound and Music Computing Conference*, pages 372–378, 2012.

[30] Barry Truax. *Acoustic Communication: Second Edition*. Ablex Publishing, 2001.

[31] Barry Truax. World Soundscape Project - Tape Library [online]. 2015. Available online at `http://www.`

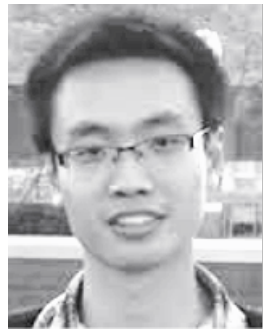
sfu.ca/sonic-studio/srs/index2.html; visited on January 12th 2015.

[32] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. http://dx.doi.org/10.1109/TSA.2002.800560.

[33] Gordon Wichern, Jiachen Xue, Harvey Thornburg, Brandon Mechtley, and Andreas Spanias. Segmentation, indexing, and retrieval for environmental and natural sounds. *Transactions on Audio, Speech and Language Processing.*, 18(3):688–707, 2010. http://dx.doi.org/10.1109/TASL.2010.2041384.

THE AUTHORS

Miles Thorogood



Jianyu Fan

Philippe Pasquier

Miles Thorogood is a PhD candidate at Simon Fraser University, and lecturer of computer science and digital media at the University of British Columbia, Okanagan. He received his Master of Applied Arts in signal modelling and human factors in 2010 from Emily Carr University of Art and Design. His doctoral dissertation deals with computationally creative systems to assist in human creative tasks. Since 2006, he has worked on audio-based research at institutes including the Australian National University, the Australian Commonwealth Scientific and Industry Research Organization, and the Intersection Digital Studio in Vancouver. The application of his academic research has been applied in industry and is regularly featured in public displays, such as the Vancouver Olympics in 2010, and City of Vancouver public works. His research interests include soundscape signal analysis, music information retrieval, modelling sound design process, and interdisciplinary research.

●

Jianyu Fan is currently a Ph.D. student at School of Interactive Arts and Technology, Simon Fraser University. He holds a Bachelors degree from Beihang University and a Masters degree from Dartmouth College. He is interested in the area of artificial intelligence, conducting research aimed at endowing machines with creative autonomous behavior. In particular, he focuses on music information retrieval systems, and generative systems for music and videos. As a scientist, he has presented his papers in various international scientific venues. As an artist, he has played the piano for over 19 years. His artworks have been presented at conferences and art festivals.

●

Philippe Pasquier is Associate Professor in Interactive Arts and Technology and an adjunct professor in Cognitive Science at Simon Fraser University. He is both a scientist and a multi-disciplinary artist. His contributions range from theoretical research in artificial intelligence, multi-agent systems and machine learning to applied artistic research and practice in digital art, computer music, and generative art. Philippe is the Chair and investigator of the AAAI international workshops on Musical Metacreation (MUME), the MUME concerts series, the international workshops on Movement and Computation (MOCO), and he was director of the Vancouver edition of the International Symposium on Electronic Arts (ISEA2015). He has co-authored over 120 peer-reviewed contributions, and presented in forums ranging from the most rigorous scientific venues to the most subjective artistic ones.