Active Learning for Auditory Hierarchy

William Coleman, Charlie Cullen, Ming Yan, Sarah Jane Delany

Active Learning for Auditory Hierarchy*

William Coleman[1,2][0000−0002−7551−882X], Charlie Cullen[1][0000−0002−8435−023X], Ming Yan[2], and Sarah Jane Delany[1][0000−0002−2062−7439]

[1] School of Computer Science, TU Dublin, Kevin Street, D08 X622.
d15126149@mytudublin.ie
[2] Xperi Corporation, Bangor, UK, BT19 7QT

Abstract. Much audio content today is rendered as a static stereo mix: fundamentally a fixed single entity. Object-based audio envisages the delivery of sound content using a collection of individual sound 'objects' controlled by accompanying metadata. This offers potential for audio to be delivered in a dynamic manner providing enhanced audio for consumers. One example of such treatment is the concept of applying varying levels of data compression to sound objects thereby reducing the volume of data to be transmitted in limited bandwidth situations. This application motivates the ability to accurately classify objects in terms of their 'hierarchy'. That is, whether or not an object is a foreground sound, which should be reproduced at full quality if possible, or a background sound, which can be heavily compressed without causing a deterioration in the listening experience. Lack of suitably labelled data is an acknowledged problem in the domain. Active Learning is a method that can greatly reduce the manual effort required to label a large corpus by identifying the most effective instances to train a model to high accuracy levels. This paper compares a number of Active Learning methods to investigate which is most effective in the context of a hierarchical labelling task on an audio dataset. Results show that the number of manual labels required can be reduced to 1.7% of the total dataset while still retaining high prediction accuracy.

Keywords: active learning · auditory hierarchy · machine learning · support vector machine.

1 Introduction

Recent technological advances have driven changes in how media is consumed in home, automotive and mobile contexts. Multi-channel audio home cinema systems have become more prevalent. The consumption of broadcast and gaming content on smart-phone and tablet technology via telecommunications networks is also more common. This has created new possibilities and consequently poses new challenges for audio content delivery such as how media can be optimized for multiple contexts while minimizing file size.

* This work was supported by the Irish Research Council and DTS Licensing Ltd. (now part of Xperi) under project code EBPPG/2016/339.

Object-based audio [9] may offer a solution to this problem by providing audio content at an object level with meta-data which controls how the media is delivered dependant on mode of consumption. In this context, insight into the relative importance of different sounds in the auditory scene will be useful in forming content delivery strategies. In the following the concept of a hierarchy between audio objects in multimedia content which changes over time due to activity in the material, external factors and personal bias will be referred to as *audio object hierarchy*.

How sounds are noted as being of most interest, referred to as Foreground (FG) sounds in the following, is not explicitly understood. Section 2.1 outlines a number of factors hypothesised to have an influence, such as attention [63], prior training [5] and context [57]. It is reasonable to suggest that certain sounds (speech, or alert noises, such as alarms) would likely be consistently categorised as FG. However, detailed knowledge in this respect would be important in the design of any delivery solution for an object-based audio system as every potential influence can be thought of as requiring a weighting appropriate to the degree to which each influences the hierarchy. Accurately mapping such weightings requires a structured study in order to examine each influence in isolation, where possible. In order to do so a dataset of sounds is required, isolated from context in so far as this is possible, to examine the influence of external factors. To our knowledge no such dataset suited to the study of auditory hierarchy currently exists and the lack of labelled datasets of suitable scale is an acknowledged problem in the domain [46]. This paper outlines a method by which large numbers of audio objects can be labelled with minimal human input to high levels of accuracy into FG and Background (BG) categories.

Previous work has outlined evidence of an inherent FG, Neutral (N), BG hierarchy between isolated sounds [10]. Further studies have established that a supervised Machine Learning (ML) approach to auditory hierarchy prediction shows promise [11], in this instance by framing the problem as a binary 'FG' / 'not FG' categorisation task, albeit using a small dataset. State-of-the-art audio ML requires large, labelled datasets [8] in order to achieve high accuracy levels. Datasets are difficult and expensive to compile and label manually, a problem which can be addressed using data augmentation techniques [49]. However, care must be taken that such augmentations do not alter the underlying semantic information of stimuli. It follows that other methods of minimising the manual effort required to compile large datasets are worthy of investigation.

Active Learning (AL) is a supervised ML technique that can be used to minimise the manual effort required to label large datasets [52]. AL strives to identify & label the most informative instances in a dataset, aiming to use a minimum of manual intervention to train a model capable of classifying unseen instances to a high level of accuracy. In this paper, AL is used to apply hierarchical labels to an audio corpus of environmental sounds. A number of selection strategies can be used in AL. Uncertainty Sampling AL (USAL) is a model-based approach which uses an uncertainty measure to identify instances that are most difficult to classify, assuming that these will be most informative for predicting

other instances [52]. Exploration Guided AL (EGAL) [24] is a model-free approach which aims to identify dense clusters of instances that are most diverse from already labelled examples. This operates on the assumption that focusing on cluster centroids most distal from already labelled instances first will allow accurate predictions to be made early in the labelling process.

While popular, USAL is computationally expensive and time consuming, requiring models to be built repeatedly throughout the labelling process. EGAL addresses this problem by basing selection of informative instances solely on dataset features, avoiding the overhead of training models repeatedly. EGAL has been found to be more effective than model-based methods in some applications [41] and to our knowledge has not yet been applied to an audio problem. In addition, a random selection strategy is implemented as a baseline.

Results show that it is possible to dramatically reduce the number of labels required to hierarchically label the audio dataset used in this study. EGAL techniques outperform both USAL and random selection strategies, being able to label to high accuracy using only 1.7% of labelled dataset instances. The next best performing selection method requires 11.7% of labels to achieve the same accuracy level.

The following sections offer an overview of perceptual (Section 2.1) and ML (Section 2.2) research relevant to auditory hierarchy. Section 2.3 introduces AL and outlines the USAL and EGAL selection methods employed in this study. Section 3 offers an overview of methodology covering a subjective labelling exercise (Section 3.1), feature extraction methods (Section 3.2), classifier choice (Section 3.3), a cross validation experiment investigating audio data representation options (Section 3.4) and the structure of the AL experiment implemented in this instance (Section 3.5). Section 4 describes an experiment utilizing a number of AL selection methods and outlines the results observed from these efforts. Finally, Section 5 discusses these findings and suggests future areas for study.

2 Related Work

2.1 Auditory Scene Analysis

Research in object-based broadcasting [9] and auditory object categorization [62] has underlined a growing interest in how such concepts can be applied to modern media consumption. The concept of a variable compression codec is but one such possibility, addressed in this paper by outlining how a dataset can be formulated on which a model can be trained to predict auditory hierarchy. Auditory Scene Analysis (ASA) involves a constant activity of sound categorization which Bregman [7] outlines as both a conscious (schematic or *top-down*) and unconscious (primitive or *bottom-up*) process of soundscape perception. Guastavino [18] has noted converging evidence from both behavioral and neurophysiological domains that provides support for the notion that amalgamation of these processes is integrated, rather than serial. In the context of auditory hierarchy ASA can therefore be considered as a constant analysis of the surrounding sound scene,

subject to varying levels of influence from a number of external factors. These have been noted to include physical properties of sounds [45], the level of attention granted by listeners [63], volume level [44], proximity [31], sound event context and listening mode [57], level of anticipation [26], prior training [5], experience [35] and even other senses (sight [17], smell [64] and touch [54]). This process involves continual identification of interesting sounds which may then be consciously analysed for semantic information or further meaning, or not, as deemed necessary.

This is therefore not a trivial problem to approach, as any fully-featured model predicting FG elements would be expected to incorporate input from many, continually changing, factors, each requiring careful examination both to evaluate the relative weight each carries with respect to auditory hierarchy and to assess how they interact over time. Furthermore, the process is subjective, each subject having a different perspective on which sounds are important and which are not, either explicitly or implicitly. Considering this, the most effective way to examine the effect of each factor is to first form a dataset with stimuli isolated from external influence in so far as this is possible. This paper therefore investigates audio object hierarchy prediction as it pertains to sounds isolated from context in so far as this is practical. Future work will involve investigation of other factors identified as having an influence on hierarchical categorization.

Predicting auditory hierarchy for modern media applications involves an investigation of individual subjective judgement of sound, specifically with regard to which sounds are most important when. As such, this should be seen as distinct from studies utilising ITU-R standards such as BS.1116-3 [28] and BS.1534-3 (MuSHRA) [29] which focus on the minutiae of variations in Basic Audio Quality (BAQ) [28] between experimental stimuli when evaluating output from different loudspeakers [53] or the qualities of ambisonic microphones [4], for instance. Our focus is on subjective perception of macro sound categorisation on a hierarchical level, rather than on micro differences between stimuli. This study focusses on stimuli suitable for use in game audio, visual streaming media and broadcasting content. This represents a broad palette of environmental sounds deemed most appropriate in terms of the envisaged end use of a hierarchical model. Framing the experimental and stimuli requirements in this manner allows us to prioritise accumulating volume of labels via an online environment over maintaining strict laboratory assessment conditions as required by the stricter standards.

2.2 Machine Learning for Auditory Hierarchy

There is a considerable extant audio ML literature [59] and a rich recent history in the application of such knowledge to the areas of acoustic scene classification [21], music information retrieval [60], various so-called 'hearable' technologies such as Google Home [16] and Amazon Echo [1] and many others [36, 61, 34]. In particular, Deep Learning (DL) algorithms such as Convolutional Neural Networks and Recurrent Neural Networks have gained a reputation as being good predictors for a wide variety of tasks and are considered state-of-the-art [8]

in many audio domains such as speech recognition [23] and environmental sound categorization [48]. However, while these algorithms are capable of high accuracy they also require large amounts of data in order to achieve such scores [51]. Indeed, the lack of large, labelled datasets for experimental purposes is an acknowledged problem in the field [3, 46].

It follows that securing a sufficient volume of suitable auditory stimuli is of primary importance in order to train a model that will accurately predict on unseen instances. This can be addressed by crowd sourcing labels for auditory stimuli and by using ML data augmentation techniques [46]. In the audio domain these include temporal and pitch variations, random cropping, dynamic range compression and the introduction of background noise [49]. However, it is still a difficult task to scale to large datasets using these methods and each are subject to limitations. Label quality is a concern with crowd sourced labels and care must be taken that data augmentation techniques do not change the underlying semantic content of the stimuli.

Auditory hierarchy has been investigated in a number of studies to date but there are methodological differences that preclude the use of these datasets for this study. For example, Lewis et al. [32] examine subject rating of approximately 256 sounds on an 'object like' versus 'scene like' axis for a selection of mechanical and environmental sounds. Thorogood et al. [55] use 200 soundscape recordings of 4 seconds in length derived from the World Soundscape Project Tape Library database [58] and categorize them in BG, FG and 'FG with BG' categories. These datasets are not of a suitable scale for our purposes, however. Salamon et al. [50] apply subjective labelling to 8,732 BG and FG urban sounds and validate label accuracy with experimental testing, but the sounds used are confined to urban settings and are not isolated from context. The authors are unaware of any large, publicly available database of sounds with hierarchical labels suitable for a study of multiple influences on auditory hierarchy.

A number of environmental sound databases are publicly accessible for research purposes (a useful summary is available [22]) and from these the Dataset for Environmental Sound Classification (ESC) [43] has been selected as it provides a large number of potential stimuli ($>$ 250,000 in total) which can be parsed for instances that contain isolated sounds suitable for hierarchical labelling. Further details on the stimuli selected for this experiment are given in Section 3.1.

2.3 Active Learning

AL is a supervised ML technique originally designed to build classifiers with minimal manual labelling effort which can be used to label large datasets [52]. As outlined in Figure 1, when a prediction model cannot confidently predict class membership the informativeness of those instances can be assessed using a selection technique. Those deemed most informative are presented to a human oracle for labelling and used to improve the prediction model. The AL process is applied iteratively and more instances are presented for labelling until the performance of a model trained on labelled instances reaches a predetermined

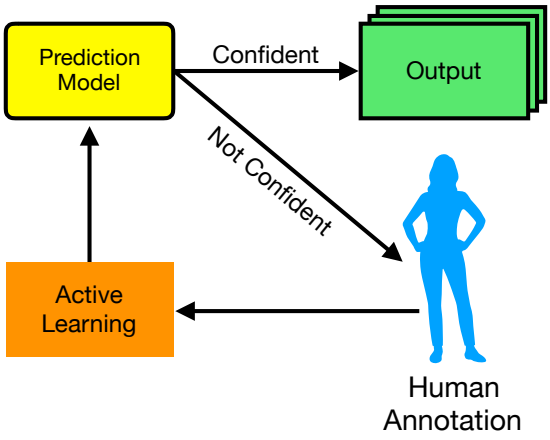

Fig. 1. An outline of the Active Learning process. The confidence of a prediction model (USAL) is one method of selecting instances for labelling. A feature space exploration technique (EGAL) will also be explored.

level or there are no more instances to label. The objective is to use minimal manual effort to label the entire dataset to a high level of accuracy.

Two selection strategies are investigated in the following. The first of these is the most commonly used method, USAL, a model-based approach which uses uncertainty in model prediction as a metric to select instances for labelling. It has used in a variety of audio applications including environmental sound classification [20], bird sound categorization [47] and speech emotion recognition [65]. The hypothesis behind USAL holds that the instances about which the classifier is most confident will provide the least useful information and that the instances most difficult to categorize will be more informative, allowing greater accuracy from fewer manually applied labels. It therefore selects these instances for labelling first. Uncertainty can be identified in different ways. The least confident method ranks classification confidence based on the best prediction and takes the lowest ranking instance for labelling. An entropy measure can also be used to assess the average information content of an instance. The margin method is employed in this case, which ranks instances by their proximity to a classifier decision boundary, presenting those closest for labelling as they are the instances most difficult to categorize. This model-based approach is computationally expensive as it requires a model to be built at each iteration which can be time consuming and may not be practical for some applications.

The EGAL selection strategy addresses this shortcoming by eschewing use of a model and identifying useful instances for classification purposes in relation to their location in the feature space relative to neighbouring instances and proximity to already labelled instances. This can be expected to reduce the computational overhead and time required to label a corpus of instances compared

to a model-based technique such as USAL as there is no need to retrain a model for each iteration. EGAL has been used in text classification applications [24] but to our knowledge this is the first application of this technique to an audio problem. The algorithm seeks to identify instances in clusters that are furthest from labelled instances on the assumption that dense clusters which are diverse from labelled instances will be most informative for classification purposes. This is implemented by first calculating a *density* value per instance, defined as the sum of similarities between the instance and all other instances within a certain radius. Here, the inverse of Euclidean distance is used for this measure. Secondly, a *diversity* value is calculated by measuring the distance between labelled and unlabelled instances.

3 Methodology

Audio stimuli and label collection methods used in a subjective labelling exercise are described in Section 3.1. Feature extraction and data preparation are covered in Section 3.2 and classifier choice is outlined in Section 3.3. A cross validation experiment investigating feature representations and classifiers is outlined in Section 3.4. The Python language was used for implementation using the associated Scikit-learn [42], SciPy [30] and Pandas [38] libraries.

3.1 Dataset

The ESC datasets [43] have been compiled from the Freesound website (*freesound.org*) for use in computational audio scene analysis contexts for training and testing automatic classification of sounds. They have been selected in this instance as they provide a large bank (>250,000) of potential stimuli with associated sound class metadata. Dataset recordings are of approximately 5 seconds duration and are provided in stereo at a sample rate of 44.1kHz. In excess of 20,000 sounds were reviewed by the authors for suitability of use in this study with care taken to exclude sounds which evinced more than one sound event in order to provide a corpus of stimuli isolated from context in so far as this is possible. This resulted in the selection of 10,166 sounds as suitable for inclusion as they did not evince more than one audio 'object'.

A random selection of these sounds were used in a subjective labelling exercise carried out by participants from Xperi/DTS Inc. and from researchers in the School of Media, Technological University Dublin. The labelling environment was deployed using a website because this facilitates access for a physically distributed cohort of participants. As discussed in Section 2.1 auditory hierarchy categorisation concerns perception on a macro rather than a micro level so participants were asked to label sounds using headphones in a quiet environment accepting a variance in acoustic rendering in favor of maximising participant numbers. Presentation order was randomised using random orders were sourced from *random.org*, a source for true random sequences cited in a number of peer-reviewed publications [19] in order to ensure there was no imbalance in sound

class representation for each labelling session. In all, 3,002 sounds were labelled a minimum of 3 times on a FG, Neutral, BG scale by 149 participants (73% male, 7% 18-24, 49% 25-44). An average of 83.42 sounds were rated per participant with each given the opportunity to rate 100 stimuli. The average time taken to complete the rating process excluding outliers >1 hour in duration was 19 minutes 54 seconds. The numerical coding for each category (BG - 1, N - 2, FG - 3) was used to generate mean and standard deviation scores for each sound. For illustrative purposes the sounds are organized into 12 broad classes as outlined in Table 1 which reproduces the average rating score and standard deviation per class.

Table 1. A summary of instance count, average score and standard deviation (σ) per class for all 3,002 sounds rated. The highest occurrences are reproduced in **bold**, lowest are underlined.

Class	No.	Average Score	σ
Nature	**523**	1.655	0.578
Ambience	507	1.477	0.504
Animal	408	2.121	0.569
Urban	370	1.382	0.437
Machine	285	1.941	0.585
Human	266	2.131	0.461
Other	226	2.325	0.564
Domestic	145	2.307	0.527
Travel	115	1.285	0.356
Actions	67	2.269	0.573
Alarms	55	**2.535**	0.41
Bells	35	2.41	**0.715**

This table shows that sounds such as 'Alarms' are likely to be labelled as FG. Sounds categorised as 'Travel' are most likely to be labelled BG reflecting the interior public transport hum ambience present in many of these sounds. Standard deviation per sound class varies between 0.41 and 0.715. The variance in average rating is outlined in the boxplots reproduced in Figure 2 and this gives an indication of the variance in the data which in this instance shows the degree of subject consensus on BG - Neutral - FG sounds.

For illustrative purposes the sounds were organised into three average rating score bands. There are 1,156 instances with an average rating of 1.5 or under, designated BG sounds. There are 608 sounds with an average rating of >2.5 and these are designated FG sounds. The remaining 1,238 sounds have an average rating >1.5 & <2.5 and are referred to as neutral sounds. The width of each box plot is proportionate to the number of instances summarized in each rating band.

Similar to other research [10] a greater consensus is noted among subjects as regards sounds considered most FG or most BG; There is less variance in

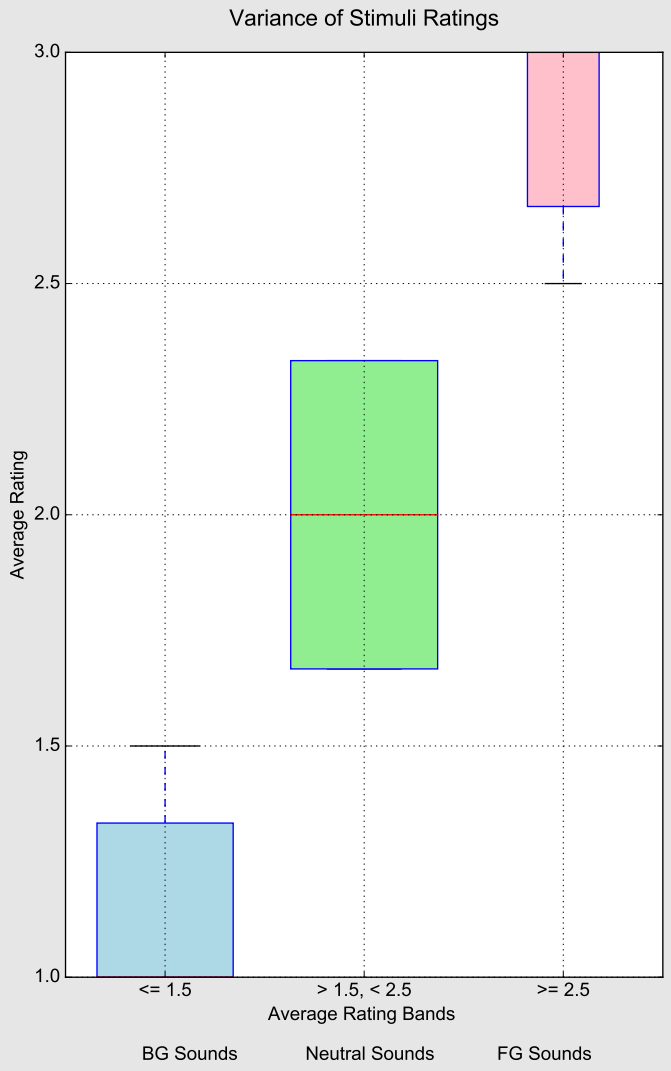

Fig. 2. Boxplots outlining the variance in average sound ratings grouped in broad bands. Note that the minimum average score for BG sounds is 1, hence there is no quartile or minimum whisker below this value. Similarly, the maximum average score for FG sounds is 3, hence this band has no quartile or maximum whisker above this value. Also, the width of each boxplot is proportional to the number of instances summarised in each band.

the ratings for these bands than those sounds considered Neutral. Interquartile range for both FG and BG bands is approximately 0.33 of a rating score. Neutral sounds on the other hand exhibit greater variance in rating scores compared to BG and FG sounds, interquartile range here being twice that of BG and FG sounds, 0.67 of a rating score. The high degree of variance for some sounds, indicating a lack of consensus between subjects as to correct sound class, is to be expected with a subjective labelling task and the dataset evinces disagreement between raters as to the correct hierarchical category for many instances. The proposed application of a variable compression codec suggests a priority of identifying FG sounds so for the purposes of the following experiments it was decided to address the data as a binary classification problem. Accordingly, all sounds achieving an average score ≥ 2.5 (608 instances, 20.25%) are categorized as 'FG'. All others (2,394 instances, 79.7%) are categorized as 'nonFG' sounds.

3.2 Feature Extraction

The Python LibROSA [37] package was used to extract three different feature vectors for each audio stimulus. Mel Frequency Cepstral Coeficients (MFCC) and Log Power Mel Spectrogram (LPMS) representations were extracted based on their popularity in audio machine learning applications [46] and a Chromagram representation was also extracted based on its usefulness in previous experiments by the authors [11]. All files were first downsampled to 16kHz to account for the variable recording quality of sounds sourced from Freesound, such as the ESC datasets. A Hann window of the form outlined in Equation 1, (where n = sample number, M = the number of points in the output window) is used to extract audio data.

$$w(n) = 0.5 - 0.5cos\left(\frac{2\pi n}{M-1}\right),\ 0 \leq n \leq M-1 \tag{1}$$

In line with similar experiments [49, 14] a window size of 128 ms (2048 samples at 16 kHz) and stride of 32 ms was used to extract 12 frequency bands of Chromagram, 13 bands of zero-order MFCC feature vectors and 40 bands of LPMS features. A representation of 128 bands of LPMS data was also experimented with but no performance improvement was observed although training time increased markedly due to the greater volume of data involved. From these zero-order, delta, double delta and fifth-order delta representations were extracted as delta features were prominent in previous experiments by the authors investigating hierarchical categorization [11]. All data is scaled and bands from each data matrix are flattened and organized into 12 data subsets, 4 each for the MFCC, Chroma and LPMS data, a summary of which is presented in Table 2.

3.3 Classifier

A Support Vector Machine (SVM) classifier is used as it has been used extensively on audio ML applications [55, 56, 6]. SVMs aim to find the optimal hyperplane which separates instances by maximising the margin of distance from

Table 2. A summary of feature representation data vectors and their dimensions. For each representation (MFCC, Chroma & LPMS) zero-order and 1st, 2nd and 5th order delta vectors are computed, resulting in a total of 12 initial representations.

Type	Dimensions	Flattened Dimensions
MFCC	3002 x 13 x 156	3002 x 2,028
Chroma	3002 x 12 x 156	3002 x 1,872
LPMS	3002 x 40 x 156	3002 x 6,240

hyperplane to data point [12]. A number of different kernels can be used with a SVM. Here, three are investigated: the Radial Basis Function (RBF), Polynomial and Linear kernels.

3.4 Cross Validation Experiment

In a preliminary experiment optimal feature representation for distinguishing between FG and nonFG sounds and which SVM kernel works best on this data is investigated. A SVM with three different kernels (RBF, polynomial and linear) is applied using default parameters outlined in Table 3. Class weights are adjusted to penalise mistakes inversely proportional to the number of instances in each class to adjust for the class imbalance in the dataset. Results showed that extracted delta representations give no improvement on zero-order representations and so these were discarded. Average Class Accuracy (ACA) and single class accuracy scores per kernel and representation are provided in Table 4.

Table 3. Default parameters used per kernel in the initial classification exercise. The 'scale' value for the gamma parameter uses $1/(no.features * variance)$ as value of gamma.

Kernel	Parameters
Radial Basis Function	C=1, gamma='scale'
Linear	C=1
Polynomial	C=1, degree=3, gamma='scale'

In addition to MFCC, Chroma and LPMS zero-order representations one further representation, a concatenation of these three, was investigated. This is labelled the 'All' representation in Table 4. Marginally better performance was observed using the 'All' representation at the cost of a significant increase in time taken to run the analysis due to its larger size. The MFCC and LPMS representations performed similarly to the 'All' representation, while the Chroma representation was notably poorer. It was decided to proceed with the LPMS representation as it performed slightly better than the MFCC and takes significantly less time to train than the 'All' representation, while achieving scores only slightly lower. With regard to kernel choice, RBF and polynomial kernels were observed to perform more strongly than the linear kernel. The overall difference

Table 4. Average Class Accuracy (ACA), and Class Accuracy scores for FG and nonFG classes per kernel and feature representation. The 'All' representation is a concatenation of the other 3 representations.

Kernel	Measure	MFCC	Chroma	LPMS	All
RBF	ACA	72.2%	65.7%	73.9%	74.3%
	FG	67.3%	53.6%	67.1%	69.7%
	nonFG	77.2%	77.8%	80.7%	78.9%
Linear	ACA	63.9%	53.1%	63.1%	60.3%
	FG	45.9%	31.9%	38.5%	35.9%
	nonFG	81.9%	74.3%	87.8%	84.8%
Polynomial	ACA	72.4%	63.0%	73.4%	74.4%
	FG	66.6%	61.0%	69.1%	70.1%
	nonFG	78.3%	65.0%	77.7%	78.7%

between RBF and polynomial was marginal so the RBF kernel was selected as it is more commonly used [40].

Parameters were fitted for the RBF kernel to the LPMS representation using a grid search and a separate validation set of 20% of the dataset within a 5-fold cross validation. The ACA achieved was 76.9% which provides a point of comparison with ACA scores achieved using minimal labelled instances during AL labelling.

3.5 Active Learning Process

As 3,002 instances were pre-labelled as described in Section 3.1, a simulated labelling exercise was conducted to assess AL for auditory hierarchy, extracting a stratified, randomly selected hold-out test set of 501 instances to measure performance. The remaining 2,501 instances form the pool of 'unlabelled' examples. Due to the random nature of the hold-out test set and 'unlabelled' pool three random splits are formed in total to counteract the chance of a single iteration providing a misleading result. The results reported are therefore averages over 3 iterations.

To select the first set of instances agglomerative clustering was employed on the 'unlabelled' pool to form 5 distinct clusters and then select an initial batch of 10 instances as this has been shown to be an effective way to initiate AL [25, 24]. During labelling runs Average Class Accuracy (ACA) was used to measure performance due to the imbalanced class distribution. The initial instances are labelled, a model was trained on them and an ACA score calculated on the hold-out test set. The selection method was then used to pick the next batch of 10 instances from the 'unlabelled' pool, these were labelled, added to the other labelled instances and a new accuracy score calculated on the hold-out test set. This process continued until there were no more instances left to be 'labelled'. The ACA values were used to plot a learning curve which was then used to compare methods both visually and with an Area Under the Learning Curve

(AULC) value. As a baseline a random selection strategy was also implemented which does not seek to intelligently select instances for labelling.

4 Results

An overview of USAL and EGAL selection methods has previously been offered in Section 2.3. These methods are now applied to the selected representation to identify which is most effective in terms of identifying the minimal number of instances for manual labelling that allow a model to classify to high accuracy levels. In total five selection methods are investigated:

- USAL, which uses a SVM to identify the instances closest to the classification decision boundary.
- Diversity EGAL, which uses the diversity measure from EGAL to select instances that are most diverse from already labelled instances.
- Density EGAL, which uses the density measure from EGAL to select cluster centroids from the most densely populated areas of the feature space.
- Hybrid EGAL, which combines density and diversity EGAL measures to select cluster centroids that are most diverse from already labelled instances.
- Random selection, selects instances randomly. 3 random selection runs are executed to account for randomness.

Figure 3 shows results of labelling runs from 10 to eventually 2,501 'labelled' instances. It includes a shaded area that denotes the maximum and minimum values achieved by random selection for each batch which demonstrates large variance.

The EGAL runs are noticeably strongest early in the training runs, all quickly achieving scores in excess of 70% accuracy. USAL does not match this performance and indeed is surprisingly, given the effectiveness of the method in other domains [52], less effective than Random selection apart from the earliest section of the run under 70 labels. There is considerable variance between the maximum and minimum scores from the random selection method showing that this is not a reliable method for selection in this application. Figure 4 focusses on the early portion of the labelling run which tracks scores achieved between 0 - 500 labels.

This highlights the success of diversity EGAL, which achieves 74% ACA using only 50 labels. The other EGAL variants are fractionally behind this early result, but perform similarly up to approximately 120 labels with the performance of density EGAL being notably strong beyond this point. The random selection strategy does not improve on the accuracy level of diversity EGAL at 50 labels until it is provided 350 labels. USAL requires 1,410 labels to achieve the same. Table 5 offers a summary of ACA and AULC scores at different points from each labelling run.

Given the ACA achieved on the whole dataset is 76.9% across 5 stratified folds, the score of 74% from 50 labels is a strong result, meaning that AL in this instance can achieve 96.1% of total possible model accuracy using only 1.7% of

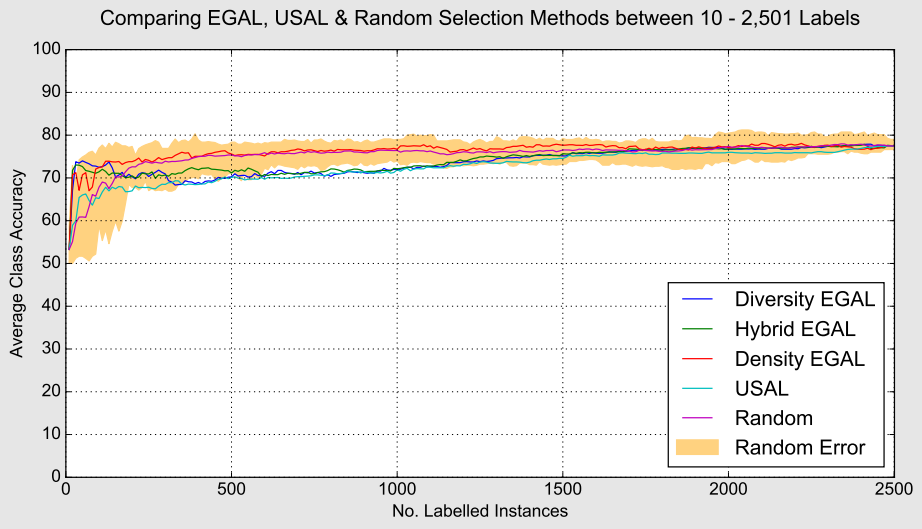

Fig. 3. Comparison of Active Learning selection methods displaying ACA scores (Balanced Accuracy) achieved from 10 - 2,501 labels. Each line denotes the overall average score for each method per batch. The shaded area denotes the variance observed from the random selection method.

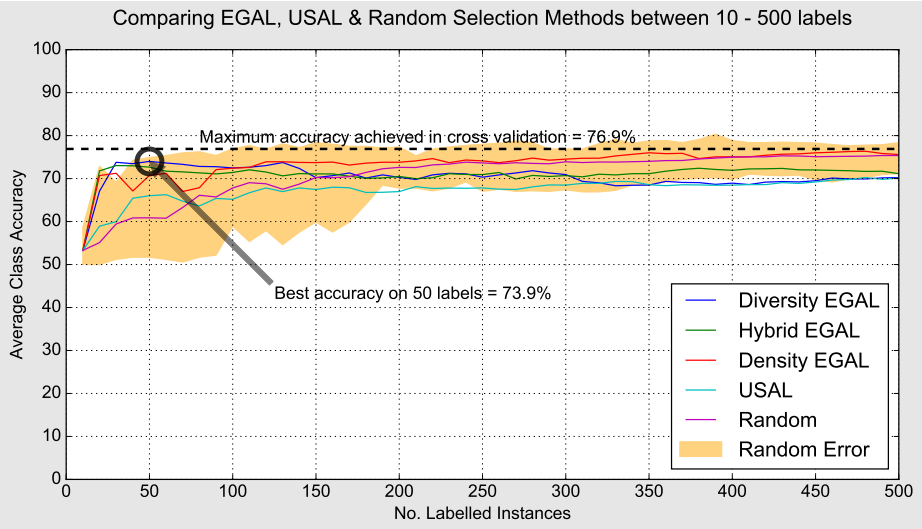

Fig. 4. Comparison of selection methods for early stage (between 0 and 500 labels) Active Learning runs. Each line denotes the overall average score for each method per batch. The shaded area denotes the variance observed from the random selection method.

Table 5. A summary of model accuracy and AULC scores for points in the labelling run per AL method. Using Diversity EGAL it is possible to achieve high classification accuracy (74%) relative to the model maximum using minimal manual labelling (50 labels) on this dataset.

Method	**50**	**100**	**200**	**500**	**2501**
	Average Class Accuracy Scores				
Diversity EGAL	74.0%	72.5%	70.3%	70.2%	77.5%
Hybrid EGAL	72.7%	71.4%	70.5%	71.1%	77.5%
Density EGAL	70.7%	72.4%	73.8%	75.6%	77.5%
USAL	66.0%	65.2%	67.0%	70.1%	77.5%
Random	60.9%	67.8%	72.6%	75.4%	77.5%
	AULC Scores				
Diversity EGAL	20.4	57.1	128.9	338.4	1838.2
Hybrid EGAL	20.8	56.7	127.7	341.6	1845.2
Density EGAL	20.2	54.9	128.2	353.3	1900.8
USAL	17.8	50.4	117.5	323.1	1808.6
Random	17.2	48.6	117.9	340.4	1877.5

labels. As noted, using random selection 350 labels, or 11.7% of the total, is required to improve on this accuracy level.

For statistical analysis the Friedman test was used to compare more than two samples with the Wilcoxon signed-rank test used as a post-hoc test between pairs of samples. In the case of the Wilcoxon test a Bonferroni correction was applied to the significance level in order to reduce the Type I error rate (identifying a significant effect where there is none) [15]. This resulted in a revised significance level of 0.005 for the post-hoc Wilcoxon tests as 10 comparisons were made. Additionally for the Wilcoxon test runs which have 20 measurements are compared as comparisons below this point are not recommended due to sample size [30]. The Friedman and Wilcoxon are non-parametric tests that look for differences between related samples and are noted to be a safer option than using parametric tests as they do not assume normal distributions or homogeneity of variance [13].

A Friedman test on the AL ACA values up to 200 labels provided is significant at the 95% level (p = 8.03E-10). The Wilcoxon tests reveal that the differences between EGAL variants are not significant to the revised significance level. However, the differences between EGAL and USAL, and between EGAL and Random selection methods are significant to the revised significance level. This indicates that EGAL is superior to both USAL and Random selection at selecting instances on which a classifier can be built to achieve high accuracy levels with minimal labelling. These results also suggest that there is little difference between the EGAL variants in this instance, as the results of the Wilcoxon comparisons between EGAL runs are not significant.

5 Discussion

This research has explored a series of Active Learning approaches to an auditory hierarchy labelling problem. It has been found that in this instance it is possible to classify to 96.1% of maximum model accuracy by labelling only 1.7% of dataset instances using the EGAL selection method. Using a random selection strategy it is necessary to select 350 instances (11.7% of the total) to surpass this accuracy level, however, as noted the large variance observed in scores from using the random selection strategy makes this an unreliable method in this instance. These results suggest EGAL is an effective method to minimise the manual effort required to label audio instances with hierarchical labels.

DL techniques are acknowledged as state-of-the-art in the audio classification domain [8] but are limited in terms of application to specific problems by the existence of suitable, large, appropriately labelled datasets. In a real-world scenario where potentially millions of labelled instances are required for DL applications the performance of EGAL in this instance suggests a potential for significant savings on manual labelling effort in both time and money terms for many different audio ML problems based on subjective human perception and evaluation of environmental sounds.

This is particularly interesting given the significance accorded to the emergence of datasets of this scale in other domains. For instance, the existence of ImageNet [27], consisting of over 14 million labelled images, is considered an important factor in the success of computer vision techniques and the influence of the DL methods applied to them [46]. While a number of large audio datasets are available [2, 33, 39] they are not universally appropriate for all audio ML problems, particularly those where subjective judgement is required. Therefore, the ability to quickly and efficiently take existing sound corpora and label them for a bespoke categorization tasks has the potential to facilitate the study of many more specific questions than would be the case if datasets were restricted to those consisting of manually labelled instances. Auditory hierarchy applied to the concept of a variable compression codec is one example of such a task.

6 Future Work

Our intention is to use these methods to label a large corpus and build a DL classifier to improve classification accuracy on hierarchically labelled audio data, ultimately validating machine labelled instances with human subjects. This extended corpus will also be suitable for use in deeper investigations on the functioning of auditory hierarchy which has been noted in Section 2.1 to be influenced by a series of factors such as sound context, the experience level of the subject and the physical characteristics of the sound itself. Having in-depth knowledge of the functioning of auditory hierarchy has application to media file delivery strategies, auto-mixing applications and object-based audio broadcasting scenarios.

Further combinations of AL methods with Self Learning elements, where labels are assigned to instances based on predictions from a model, or the concept

of Co-Training, where labels are derived via a combination of prediction and selection methods on different feature representations may also lead to improvements.

References

1. Amazon Echo (2nd generation) — Alexa Speaker, *https://www.amazon.com/all-new-amazon-echo-speaker-with-wifi-alexa-dark-charcoal/dp/B06XCM9LJ4* [Accessed: 2018-08-27]
2. AudioSet, *https://research.google.com/audioset/ontology/index.html* [Accessed: 2018-08-27]
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning Sound Representations from Unlabeled Video. In: 30th Conference on Neural Information Processing Systems (NIPS 2016). pp. 892–900. Barcelona, Spain; December 5 - 10 (2016)
4. Bates, E., Gorzel, M., Ferguson, L., O'Dwyer, H., Boland, F.M.: Comparing Ambisonic Microphones: Part 1. In: 2016 AES International Conference on Sound Field Control. pp. 1–10. No. 6-3, AES, Guildford, UK (2016)
5. Bigand, E., Poulin-Charronnat, B.: Are We "Experienced Listeners"? A Review of the Musical Capacities that do not Depend on Formal Musical Training. Cognition **100**(2006), 100–130 (2006). https://doi.org/10.1016/j.cognition.2005.11.007
6. Bountourakis, V., Vrysis, L., Papanikolaou, G.: Machine Learning Algorithms for Environmental Sound Recognition. In: Proceedings of the Audio Mostly 2015 on Interaction With Sound - (AM15). pp. 1–7. Thessaloniki, Greece; October 07-09 (2015). https://doi.org/10.1145/2814895.2814905
7. Bregman, A.S.: Auditory Scene Analysis: The Perceptual Organisation of Sound. The MIT Press, Cambridge, MA (1990)
8. Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T., Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T.: Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **25**(6), 1291–1303 (2017)
9. Churnside, T.: Object-Based Broadcasting (2013), *http://www.bbc.co.uk/rd/blog/2013-05-object-based-approach-to-broadcasting* [Accessed: 2017-10-27]
10. Coleman, W., Cullen, C., Yan, M.: Categorisation of Isolated Sounds on a Background - Neutral - Foreground Scale. In: Proceedings of the 144th Convention of the Audio Engineering Society. pp. 1–9. Milan, Italy; May 23-26 (2018)
11. Coleman, W., Cullen, C., Yan, M., Delany, S.J.: A Machine Learning Approach to Hierarchical Categorisation of Auditory Objects. Journal Audio Eng. Soc. **68**(1/2), 48–56 (2020)
12. Cunningham, P., Cord, M., Delany, S.J.: Supervised Learning. In: Cord, M., Cunningham, P. (eds.) Machine Learning Techniques for Multimedia, pp. 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-75171-7_2
13. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research **7**, 1–30 (2006)
14. Espi, M., Fujimoto, M., Kinoshita, K., Nakatani, T.: Exploiting Spectro-temporal Locality in Deep Learning Based Acoustic Event Detection. Eurasip Journal on Audio, Speech, and Music Processing (2015). https://doi.org/10.1186/s13636-015-0069-2

15. Field, A.: Discovering Statistics Using SPSS, vol. 58. SAGE Publications, London, UK, 3rd edn. (2009). https://doi.org/10.1234/12345678
16. Google Home - Smart Speaker & Home Assistant - Google Store, *https://store.google.com/product/google_home* [Accessed: 2018-08-27]
17. Gruters, K.G., Murphy, D.L.K., Smith, D.W., Shera, C.A., Groh, J.M.: The Eardrum Moves when the Eyes Move: A Multisensory Effect on the Mechanics of Hearing. bioRxiv **156570** (2017). https://doi.org/10.1101/156570
18. Guastavino, C.: Everyday Sound Categorization. In: Virtanen, T., Plumbley, M.D., Ellis, D. (eds.) Computational Analysis of Sound Scenes and Events, pp. 183–213. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-63450-0_7
19. Haahr, M., Haahr, S.: Random.org (2018), *https://www.random.org/media/* [Accessed: 2018-01-04]
20. Han, W., Coutinho, E., Ruan, H., Li, H., Schuller, B., Yu, X., Zhu, X.: Semi-supervised Active Learning for Sound Classification in Hybrid Learning Environments. PloS one **11**(9) (2016)
21. Han, Y., Park, J.: Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification. Tech. rep., DCASE2017 Challenge, Munich, Germany; 16th November (Sep 2017)
22. Heittola, T.: Datasets - Toni Heittola, *https://www.cs.tut.fi/ heittolt/datasets* [Accessed: 2019-08-28]
23. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine **29**(6), 82–97 (Nov 2012). https://doi.org/10.1109/MSP.2012.2205597
24. Hu, R., Delany, S.J., Mac Namee, B.: EGAL: Exploration Guided Active Learning for TCBR. In: Proceedings of ICCBR. pp. 156–170. Alessandria, Italy; 19-22 July (2010). https://doi.org/10.1007/978-3-642-14274-1_13
25. Hu, R., Mac Namee, B., Delany, S.J.: Off to a Good Start: Using Clustering to Select the Initial Training set in Active Learning. In: Twenty-Third International FLAIRS Conference. Florida; 19-21 May (2010). https://doi.org/10.21427/D7Q89W
26. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. The MIT Press, Cambridge, MA, USA (2006)
27. ImageNet, *http://www.image-net.org/* [Accessed: 2019-09-23]
28. International Telecommunication Union: ITU-R BS.1116-3, Methods for the Subjective Assessment of Small Impairments in Audio Systems. ITU-R Recommendation **1116**(3) (2015)
29. International Telecommunication Union: ITU-R BS.1534-3, Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems. ITU-R Recommendation **1534-3** (2015)
30. Jones, E., Oliphant, T., Peterson, P., Others: SciPy: Open source scientific tools for Python (2001), http://www.scipy.org/
31. Lavandier, C., Defréville, B.: The Contribution of Sound Source Characteristics in the Assessment of Urban Soundscapes. Acta Acustica united with Acustica **92**, 912–921 (2006)
32. Lewis, J.W., Talkington, W.J., Tallaksen, K.C., Frum, C.a.: Auditory Object Salience: Human Cortical Processing of Non-Biological Action Sounds and their Acoustic Signal Attributes. Frontiers in Systems Neuroscience **6**(May), 1–15 (2012). https://doi.org/10.3389/fnsys.2012.00027

33. Linguistic Data Consortium, *https://catalog.ldc.upenn.edu/* [Accessed: 2019-09-23]
34. Malfante, M., Mars, J.I., Dalla Mura, M., Gervaise, C.: Automatic Fish Classification. Journal of the Acoustical Society of America **143**(5), 2834–2846 (2018). https://doi.org/10.1121/1.5036628
35. McAdams, S.: Recognition of Sound Sources and Events. In: McAdams, S., Bigand, E. (eds.) Thinking in Sound: The Cognitive Psychology of Human Audition, chap. 6, pp. 146–198. Clarendon Press, Oxford, UK (1993)
36. McFee, B.: Statistical Methods for Scene and Event Classification. In: Virtanen, T., Plumbley, M.D., Ellis, D.P.W. (eds.) Computational Analysis of Sound Scenes and Events, chap. 5, pp. 103–146. Springer International Publishing, 1 edn. (2018)
37. Mcfee, B., Raffel, C., Liang, D., Ellis, D.P.W., Mcvicar, M., Battenberg, E., Nieto, O.: librosa: Audio and Music Signal Analysis in Python. In: Proceedings of the 14$^{\text{th}}$ Python in Science Conference (SciPy 2015). Austin, USA; July 6-12 (2015)
38. Mckinney, W.: Data Structures for Statistical Computing in Python. In: PROC. OF THE 9$^{\text{th}}$ PYTHON IN SCIENCE CONF. (SCIPY 2010). p. 51. Austin, USA; June 28 - July 3 (2010)
39. Million Song Dataset, *http://millionsongdataset.com/* [Accessed: 2019-09-23]
40. Nisbet, R., Miner, G., Yale, K., Nisbet, R., Miner, G., Yale, K.: Advanced Algorithms for Data Mining. In: Handbook of Statistical Analysis and Data Mining Applications, pp. 149–167. Academic Press (Jan 2018). https://doi.org/10.1016/B978-0-12-416632-5.00008-6
41. O'Neill, J., Delany, S.J., Macnamee, B.: Model-free and Model-based Active Learning for Regression. In: Angelov, P., Gegov, A., C., J., Shen, Q. (eds.) Advances in Intelligent Systems and Computing, vol. 513, pp. 375–386. Springer Verlag (2017). https://doi.org/10.1007/978-3-319-46562-3_24
42. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(Oct), 2825–2830 (2011)
43. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification (2015). https://doi.org/10.1145/2733373.2806390
44. Pollack, I., Pickett, J.: Cocktail Party Effect. Journal of the Acoustical Society of America **29**(11), 1262–1262 (1957)
45. Pressnitzer, D., Graves, J., Chambers, C., de Gardelle, V., Egré, P.: Auditory Perception: Laurel and Yanny Together at Last. Current Biology **28**(13), R739 – R741 (2018). https://doi.org/10.1016/j.cub.2018.06.002
46. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.y., Sainath, T.: Deep Learning for Audio Signal Processing. Journal of Selected Topics of Signal Processing **13**(2), 206–219 (2019). https://doi.org/10.1109/JSTSP.2019.2908700
47. Qian, K., Zhang, Z., Baird, A., Schuller, B.: Active Learning for Bird Sound Classification via a Kernel-based Extreme Learning Machine. The Journal of the Acoustical Society of America **142**(4), 1796–1804 (2017). https://doi.org/10.1121/1.5004570
48. Sailor, H.B., Agrawal, D.M., Patil, H.A.: Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification. Proceedings of Interspeech 2017 pp. 3107–3111 (2017)
49. Salamon, J., Bello, J.P.: Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters **24**(3), 279–283 (2017)

50. Salamon, J., Jacoby, C., Bello, J.P.: A Dataset and Taxonomy for Urban Sound Research. In: Proceedings of the ACM International Conference on Multimedia - MM '14. pp. 1041–1044. Orlando, Florida, USA; November 3-7 (2014). https://doi.org/10.1145/2647868.2655045
51. Schröder, J., Moritz, N., Anemüller, J., Goetze, S., Kollmeier, B.: Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016. IEEE/ACM Transactions on Audio Speech and Language Processing (2017). https://doi.org/10.1109/TASLP.2017.2690569
52. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012)
53. Soren, B.: Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data. In: Proceedings of the 8th International Conference of the Audio Engineering Society. May, Washington D.C., USA (1990)
54. Steffens, J., Steele, D., Guastavino, C.: Situational and Person-related Factors Influencing Momentary and Retrospective Soundscape Evaluations in Day-to-day Life. The Journal of the Acoustical Society of America **141**(3), 1414–1425 (2017). https://doi.org/10.1121/1.4976627
55. Thorogood, M., Fan, J., Pasquier, P.: Soundscape Audio Signal Classification and Segmentation Using Listener's Perception of Background and Foreground Sound. Journal of the Audio Engineering Society **64**(7/8), 484–492 (2016). https://doi.org/10.17743/jaes.2016.0021
56. Torija, A.J., Ruiz, D.P., Ramos-Ridao, Á.F.: A Tool for Urban Soundscape Evaluation Applying Support Vector Machines for Developing a Soundscape Classification Model. Science of the Total Environment **482-483**(1), 440–451 (2014). https://doi.org/10.1016/j.scitotenv.2013.07.108
57. Truax, B.: Acoustic Communication. Ablex Publishing Corporation, Norwood, NJ, USA, 1st edn. (1984)
58. Truax, B.: World Soundscape Project Tape Library (2015), *http://www.sfu.ca/sonic-studio/srs/index2.html* [Accessed: 2017-03-07]
59. Virtanen, T., Plumbley, M.D., Ellis, D. (eds.): Computational Analysis of Sound Scenes and Events. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-63450-0
60. Virtanen, T., Plumbley, M.D., Ellis, D.P.W.: Introduction to Sound Scene and Event Analysis. In: Virtanen, T., Plumbley, M.D., Ellis, D.P.W. (eds.) Computational Analysis of Sound Scenes and Events, chap. 1, pp. 3–12. Springer International Publishing, 1 edn. (2018)
61. Wang, D., Chen, J.: Supervised Speech Separation Based on Deep Learning: An Overview. Computing Research Repository (CoRR) **abs/1708.0** (2017)
62. Woodcock, J., Davies, W.J., Cox, T.J.: A Cognitive Framework for the Categorisation of Auditory Objects in Urban Soundscapes. Applied Acoustics **121**(2017), 56–64 (2017). https://doi.org/10.1016/j.apacoust.2017.01.027
63. Woods, K.J.P., McDermott, J.H.: Attentive Tracking of Sound Sources. Current Biology **25**(17), 2238–2246 (2015)
64. Yong Jeon, J., Jik Lee, P., Young Hong, J., Cabrera, D.: Non-auditory Factors Affecting Urban Soundscape Evaluation. The Journal of the Acoustical Society of America **130**(6), 3761–3770 (2011). https://doi.org/10.1121/1.3652902
65. Zhang, Z., Schuller, B.: Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition. In: 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012). pp. 362–365. Portland, OR, USA; September 9 - 13 (2012)