A Machine Learning Approach to Hierarchical Categorisation of Auditory Objects

William Coleman
Technological University Dublin, d15126149@mytudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

See next page for additional authors

Recommended Citation

Coleman, W. et al. (2010) A Machine Learning Approach to Hierarchical Categorisation of Auditory Objects,*The Journal of the Audio Engineering Society*, Vol. 1, No. 1, 2010. DOI: 10.17743/jaes.2020.0001

Funder: Irish Research Council

Authors

William Coleman, Sarah Jane Delany, Charlie Cullen, and Ming Yan Dr.

A Machine Learning Approach to Hierarchical Categorisation of Auditory Objects

William Coleman[1*], *AES Student Member*, **Sarah Jane Delany**[1], **Ming Yan**[3], **AND Charlie Cullen**[12]
(william.coleman4@mydit.ie) (sarahjane.delany@dit.ie) (ming.yan@xperi.com) (charlie.cullen@dit.it)

[1]*Technological University Dublin, Ireland* [2]*University of the West of Scotland,* [3]*DTS Inc. now part of Xperi*

With the advent of new audio delivery technologies comes opportunities and challenges for content creators and providers. The proliferation of consumption modes (stereo headphones, home cinema systems, 'hearables'), media formats (mp3, CD, video and audio streaming) and content types (gaming, music, drama & current affairs broadcasting) has given rise to a complicated landscape where content must often be adapted for multiple end use scenarios. The concept of object-based audio envisages content delivery not via a fixed mix but as a series of auditory objects which can then be controlled either by consumers or by content creators & providers via accompanying metadata. Such a separation of audio assets facilitates the concept of Variable Asset Compression (VAC) where the most important elements from a perceptual standpoint are prioritised before others. In order to implement such a system however, insight is first required into what objects are most important and secondly, how this importance changes over time. This paper investigates the first of these questions, the hierarchical classification of isolated auditory objects, using machine learning techniques. We present results which suggest audio object hierarchies can be successfully modelled and outline considerations for future research.

0 INTRODUCTION

Recent technological developments have created new modes of audio consumption. Increased mobile network capacities have made possible the streaming of high definition video content while 'on the move'. Smart home speakers (becoming known as 'hearables'), such as the Google Home [1], connect a mono speaker to a voice search capacity which allows the control of music streaming services and numerous other functions. New home entertainment technology, such as sound bars [2] and multi-speaker home cinema systems [3] have become more prevalent. This proliferation of consumption paradigms brings challenges and opportunities for audio delivery infrastructure. Where previously a stereo audio mix was the standard for the majority of scenarios, the plethora of possibilities now available to consumers creates an impetus towards optimising audio delivery for multiple cases.

Object-based audio is an active area of research [4] which conceives of audio content being delivered as a collection of individual audio objects controlled by metadata. Such flexibility gives rise to numerous possibilities for content creation and delivery. For example, audio professionals can be given direct control of how their content is delivered across multiple formats, accommodating stereo, mono or multi-speaker audio presentation in one download with no great increase in media file size. The BBC has experimented with end-user command of audio elements with broadcasts offering mix control to viewers [5]. Additionally, breaking audio content into discrete objects offers the possibility of intra-object variable compression which could be utilised to adapt audio file sizes in constrained bandwidth situations. For this scenario it follows that insight into the relative Hierarchy of Importance (HoI) between individual sound objects in an auditory scene is critical as it can be used to implement a Variable Asset Compression (VAC) schema which maps how audio object importance changes over time.

We approach this problem by first establishing context for the task as an extension of Auditory Scene Analysis (ASA), Semantic Audio, and Machine Learning (ML) research in Section 1. In Section 2 we review an experiment investigating subjective hierarchical ratings of isolated sounds [6] which we shall use as the basis for the current study, the methodology for which is described in Section 3. We will outline results of the experiment in Section 4 and in Section 5 we will discuss these in the context of planned future work.

[*] To whom correspondence should be addressed Email: william.coleman4@mydit.ie

1 RESEARCH CONTEXT

Bregman [7] has described ASA as the process by which auditory scenes are parsed into individual sounds which we are referring to as auditory *objects*. This is a complex task because sounds are interleaved and overlap in both temporal and frequency domains, and the human auditory system only has access to an amalgam of all sounds that are presented to the ear at any one moment. Bregman describes how the human auditory system addresses this using processes of sequential and simultaneous grouping where perception is governed by primitive low-level and schematic high-level structures that parse the sound scene for individual objects.

Considerable sensory research exists regarding soundscapes [8, 9, 10], sound categorisation [11, 12, 13] sound taxonomies [14, 15, 16] and how attentional, contextual and other processes affect our perception of the environment [17, 18, 19]. The recent multi-stable *Yanny/Laurel* percept [20] is a current example of such phenomena. However, the authors are unaware of any research using subjectively derived hierarchical ratings as labels in a ML task using objective measures to predict sound importance. Lewis et al. [21] provide a sound hierarchy rating on an *object-like* versus *scene-like* axis for a selection of mechanical and environmental sounds. Thorogood et al. [22] use a selection of soundscape recordings derived from the World Soundscape Project Tape Library database [23] and categorise them in Background (BG), Foreground (FG) and 'FG with BG' categories. These sounds were selected with the intention of allowing the listener to identify sound context. Salamon et. al [14] perform subjective labelling of BG and FG urban sounds and validate their accuracy with experimental testing, but the sounds used are confined to urban contexts and are not isolated from context.

ML is an active area of research both generally [24] and in audio terms [25]. There is a rich recent history in the application of such knowledge to a number of auditory research areas which provide significant context for the hierarchical categorisation task proposed. Considerable progress in speech recognition [26, 27, 28], music information retrieval [29, 30, 31] and environmental sound classification [32, 33, 34] tasks, including the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [35, 36], provide background to a variety of different sound classification tasks and suggest ML techniques as a fruitful path for development of a VAC schema.

The following will outline research focussed on subjective perception of macro sound categorisation on a hierarchical level, as opposed to sound quality differences between stimuli that occur on a micro level. Hierarchical categorisation of audio objects is a variation on the environmental sound classification problem for content such as computer games, drama, entertainment and current affairs broadcasting. This is a process of deriving meaning from sounds, a subset of the field of Semantic Audio, the study of the 'abstraction and processing of information relating to audio signals' [37]. As such, it involves an investigation of individual subjective judgement of sound hierarchy, which is distinct from studies focussed on variations in Basic Audio Quality (BAQ) between experimental stimuli, which typically involve assessment of audio equipment, such as loudspeakers [38] or compression codecs [39]. In this context, providing a basis for subsequent investigation of effects such as context, expectation and individual training across the broadest spread of categories requires an investigation into the existence of an inherent HoI between isolated sounds. The authors have previously outlined a perceptual study [6], summarised in Section 2, which suggests the existence of such a structure by quantifying human subjective hierarchical ratings of sounds. The next step is to derive labels from these data for use in a ML classification exercise which establishes the feasibility of predicting the hierarchy of a sound set using purely objective measurements.

2 SUBJECTIVE HIERARCHICAL SOUND RATINGS

In order to maximise participants, the experiment outlined in [6] was deployed in an online environment providing detailed instructions as to its use and a training phase for test environment familiarisation purposes. Subjects were asked to complete the test using headphones in a quiet environment, were required to submit basic demographic information and then rate 40 sounds in a BG—Neutral (N)—FG evaluation task. A total of 112 complete test results were collected from 36 women and 76 men. The majority (65%) of respondents were 25—44 years of age. For study purposes, FG and BG were defined as follows:

A FG sound: One you are likely to think prominent and give greater attention.

A BG sound: One you are likely to think less important and give less attention.

Informed consent was obtained for all participants following guidelines approved by the Dublin Institute of Technology Research Ethics Committee. Figure 1 shows the stimulus presentation and scale rating portion of the test environment.

It was decided to use stimuli analogous to visual streaming content as this is the envisaged end-use of object-based audio in media consumption scenarios and thus is ecologically valid. Stimuli were sourced from the ESC-50 [40] sound set and presentation was randomised so as to minimise order effects. The Environmental Sound Classification 50 classes dataset (ESC50) files are provided in the .ogg format. To maximise browser compatibility the files were converted to .mp3s in Audacity [41] at 320kbps, the highest possible bitrate. All files have a sample rate of 44.1kHz. The ESC-50 dataset has been compiled for use in computational audio scene analysis contexts for training and testing automatic classification of sounds. Dataset recordings are of approximately 5 seconds duration and are organised into 5 broad classes:

- Animals

- Natural soundscapes and water sounds
- Human, non-speech sounds
- Interior/domestic sounds
- Exterior/urban sounds

Subject responses were collated and a frequency table compiled (summarised in Table 1) which shows counts of BG, N and FG selections for each sound. We consider the hierarchical spectrum as a ranked continuum indicating level of sound importance. Thus, the median sound rating is used for categorisation purposes as this is generally accepted as the appropriate measure of centre for ordinal data. The numerical coding for categorisation was used to calculate standard deviation and mean score for each sound which gives an indication of consensus between participants as to sound category. Ranking using these measures results in the order used in Table 1.

3 HIERARCHICAL CATEGORISATION USING OBJECTIVE MEASURES

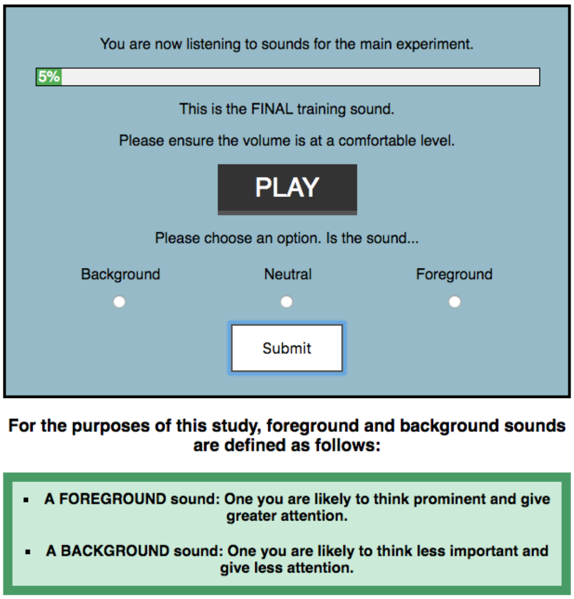
In approaching the problem of modelling auditory hierarchy from objective measures it was decided to prioritise identifying FG sounds. Any real-world implementation of a VAC would in theory require all important sounds be correctly identified and have tolerance for accepting some misclassified objects. For this reason it was decided to frame the task as a binary classification problem using the target labels of ‘FG’ and ‘notFG’, the latter of which is simply the set of all sounds identified as ‘N’ and ‘BG’ according to the median rating derived in [6].

3.1 Feature Extraction

Objective measures of the sound stimuli were generated in Matlab [42] using the ‘Matlab Audio Analysis Library’ [43] as detailed in [44]. A Hamming window of the form outlined in Equation 1 (where n = sample number, N = the number of samples in the window, window length $L = N + 1$) was implemented with a size and step length of 0.05 and 0.025 secs (50% overlap) respectively. This resulted in an initial Analytics Base Table (ABT) of 35 features, which are described in Table 2.

$$w(n) = 0.54 - 0.46\, cos(2\pi \frac{n}{N}),\ 0 \leq n \leq N \qquad (1)$$

Standard statistical summaries (mean, median, standard deviation, standard deviation by mean ratio, maximum, minimum, mean non-zero, and median non-zero) were ap-


You are now listening to sounds for the main experiment.

This is the FINAL training sound.

Please ensure the volume is at a comfortable level.

Please choose an option. Is the sound...

Background ○ Neutral ○ Foreground ○

For the purposes of this study, foreground and background sounds are defined as follows:

- **A FOREGROUND sound: One you are likely to think prominent and give greater attention.**
- **A BACKGROUND sound: One you are likely to think less important and give less attention.**

Fig. 1. The online test environment.

Table 1. Summary results ordered by mean sound rating from top to bottom. Sounds ranked *More Background* are towards the top while those *More Foreground* are to the bottom.

Sound	BG	N	FG	Category
Birds	95	12	5	BG
Keyboard_Tapping	81	25	6	BG
Clock_Tick	79	25	8	BG
Fire	76	31	5	BG
Crickets	81	16	15	BG
Water_Drops	73	23	16	BG
Wind	69	28	15	BG
Engine	69	23	20	BG
Helicopter	68	22	22	BG
Train	62	19	31	BG
Washing_Machine	61	20	31	BG
Rain	55	28	29	N
Drink_Sipping	51	31	30	N
Hen	50	32	30	N
Can_Open	53	25	34	N
Pouring_Water	50	26	36	N
Coughing	43	38	31	N
Snoring	46	28	38	N
Crow	45	27	40	N
Brushing_Teeth	42	33	37	N
Handsaw	36	40	36	N
Fireworks	37	30	45	N
Clapping	35	31	46	N
Pig	31	35	46	N
Church_Bells	34	26	52	N
Dog	28	36	48	N
Cow	35	20	57	FG
Door_Wood_Creak	31	25	56	FG
Insects	28	27	57	FG
Thunderstorm	30	22	60	FG
Rooster	24	25	63	FG
Cat	24	18	70	FG
Laughing	17	30	65	FG
Breathing	19	22	71	FG
Chainsaw	12	16	84	FG
Siren	11	12	89	FG
Baby_Crying	6	7	99	FG
Door_Knock	3	10	99	FG
Glass Breaking	2	11	99	FG
Clock_Alarm	1	7	104	FG

Table 2. A description of features extracted as objective measures of the sound stimuli.

Feature	Description
Zero Crossing Rate	The number of times the signal changes value, negative to positive and vice versa, divided by frame length.
Energy	Sometimes referred to as the *power* of a signal, calculated as the sum of the squares of signal values normalised by the respective frame length.
Entropy of Energy	A measure of the abrupt changes in the energy of an audio signal, which can be thought of as an indication of signal predictability.
Spectral Centroid	An indicator of timbre. Higher values equate to brighter sounds.
Spectral Spread	A measure of how the sound spectrum is distributed about the spectral centroid. Higher values result from spectra not tightly grouped about the centroid, exhibiting more variety.
Spectral Entropy	Similar to energy entropy, but in the frequency domain. A measure of abrupt changes.
Spectral Flux	The degree of change in the frequency domain between two analysis frames.
Spectral Rolloff	Generally used to indicate the frequency below which 90% of the magnitude distribution of the spectrum is focussed.
MFCCs	Mel Frequency Cepstral Coefficients capture timbre detail of a signal efficiently. The frequency bands used to split the signal are not linear but distributed according to the mel-scale which is modelled on the human auditory system. In this instance, 13 bands are extracted.
Harmonic Ratio	The maximum value of the normalised autocorrelation function (the correlation of an analysis frame with itself at a defined time lag, in this instance, one analysis frame).
Fundamental Frequency	An estimate of the frequency equivalent of the length of the fundamental period of the signal.
Chroma Vector	A 12-element representation of spectral energy where the bins represent the 12 equal-tempered pitch classes of western music (semitone spacing).

plied to each feature resulting in an initial vector of 280 features per sound. In addition to these global summaries, delta and double delta measures for the original 35 features were calculated to capture detail of local variation in the stimuli. These were derived from the frame level data and summarised using mean, median, standard deviation, standard deviation by mean ratio, maximum & minimum values resulting in a further 420 features. This resulted in a final ABT of dimensions 40 sounds detailed by 700 features.

3.2 Algorithm Choice

Numerous ML algorithms, Random Forest (RF) [45, 46], k Nearest Neighbours (KNN) [32, 47], Naive Bayes (NB) [48], logistic regression [49], and Support Vector Machine (SVM) [50, 22] have been successfully applied to audio problems. Additionally, Convolutional Neural Network (CNN)s and Recurrent Neural Network (RNN)s feature strongly throughout the environmental sound categorisation literature [51, 52] and in successful solutions to the DCASE 2017 [53] environmental sound classification challenge [54, 55]. This list is not intended to be exhaustive, but gives an indication of the broad range of options open to the researcher for automated audio classification. It was decided to use SVM and RF in this instance.

The relatively small size of the available dataset is a factor in algorithm choice, as there are noted strengths and weaknesses for the different ML methods. For instance, as pointed out in [56], SVMs tend to outperform other algorithms on small datasets. Also, Deep Neural Networks require large amounts of data to outperform SVMs, which are noted to perform well using up to 10,000 instances, but deteriorate in performance thereafter [57]. This suggests that better results will be achieved with the current dataset using algorithms known to perform well with relatively small datasets, such as SVMs, which aim to find the optimal hyperplane which separates instances by maximising the margin of distance from hyperplane to data point [58].

A RF is an ensemble of decision trees used extensively in ML classification problems [57]. Where a single decision tree can overfit the training data, an ensemble of trees is less prone to this problem as the tendency to overfit in single trees can be averaged out throughout the ensemble. RF are often used to provide insight into relative feature importance to assist in the process of dimension reduction.

3.3 Dimension Reduction

There are a number of potential evaluation procedures for ML features which include filter-based, wrapper and Principle Component Analysis (PCA) approaches. A comprehensive overview is beyond the scope of this paper, see [59] for a review.

A *wrapper* approach was applied in this instance because the relatively small dataset size meant that the computational load, which can be excessive [60], was not prohibitive. The *wrapper* technique uses a prediction algorithm, the *'wrapper'*, to reduce the dimensionality of a dataset while incorporating interacting effects among features by searching the feature set for subsets that perform best [61]. This is achieved either via a process of *forward sequential selection*, where the search starts with a single feature and iteratively adds more, or *backward sequential selection*, where the search starts with the full feature set and iteratively eliminates single features from each subsequent trial.

Two rounds of Recursive Feature Elimination (RFE), a backward sequential feature selection procedure, were applied in this instance to reduce the dimensionality of the initial dataset. Firstly, 5 subsets were generated using a RF *wrapper*, as it was noted that this resulted in variations in which features were selected and the total number of features chosen. Each of these initial subsets was then subjected to a further round of dimension reduction using a *wrapper* based on the final prediction algorithm, either RF or SVM. This produced smaller data subsets of sizes ranging from 2 to c. 200 features.

3.4 Model Training and Validation

5-fold cross-validation was implemented on the dataset to measure performance and a 4-fold cross-validation was used on the training set to select features and to fix parameters using a grid search. Baseline models were first run for later comparison. The data was normalised in the case of SVMs as required [57]. After the dimension reduction process, detailed in Section 3.3, another parameter search was performed to isolate the best parameter and subset choice. Once identified, the best performing models at this stage were then tested on the held out fold of unseen data to provide a robust assessment of model performance.

3.5 Model Evaluation

The final step in the modelling process is measuring the performance of the methods chosen, for which there are a number of popular metrics. The applicability of these varies for different use cases. Given the priority of correctly classifying FG sounds outlined in earlier in Section 3, it was decided to use FG class accuracy (also referred to as Recall) and accuracy as measures of model success. Class accuracy indicates correct predictions of FG sounds only. Accuracy, on the other hand, indicates how many ‘FG’ & ‘notFG’ predictions are correct.

Scores from baseline and optimised models from each of the 5 cross-validation folds implemented in the experiment were compared using the Kruskal-Wallace H-test, a non-parametric statistical test for comparing two or more independent examples which can be applied to data samples of 5 or more observations. A significance level of $p < 0.05$ was adopted in this instance to indicate a statistically significant difference between model scores [62].

4 RESULTS

Table 3 summarises the results of ML modelling. The baseline class accuracy scores are poor, 30% of FG sounds captured by RF and 50% by SVM. However, accuracy scores are more promising with RF successfully categorising 60.8% of sounds and SVM scoring 67.7%. Taken together, these results suggest that HoI may plausibly be modelled using machine learning techniques, though considerable improvement in categorisation success rates will likely be required for any real-world implementation.

Table 3. Summary results for baseline (BL) & optimised (OP) models. CA is the FG class accuracy rate (or FG recall rate). AC is model accuracy rate for both FG and ‘notFG’ classes.

Metric	RF-BL	RF-OP	SVM-BL	SVM-OP
CA	30 %	73.3 %	50 %	93.3 %
AC	60.8 %	80.3 %	67.7 %	88.1 %

The parameter tuning and dimension reduction process described in the foregoing were implemented in an attempt to improve these baseline scores to levels comparable with other studies. If successful, this would strengthen the case for utilisation of ML in the domain. In the case of RF, class accuracy improves from 30% to 73.3%, and accuracy from 60.8% to 80.3%. When comparing the fold scores using the Kruskal-Wallace test, the difference between class accuracy baseline and optimised models is statistically significant at the 95% level. Accuracy scores are not statistically significant, but only marginally so (p = 0.057). SVM class accuracy improves from 50% to 93.3%, and accuracy from 67.7% to 88.1%. Both of these results are statistically significant. While it is yet to be determined if these success rates would be effective in the implementation of a VAC codec, the SVM class accuracy score of 93.3% is encouraging, given the stated priority of classifying FG sounds. Furthermore, the optimised model scores are comparable to similar studies [22, 46] which indicate that experimentation with feature generation approaches may lead to further improvements. Finally, when comparing optimised RF with SVM fold scores, while we report better metrics for SVM models in Table 3, the difference between optimised learning algorithms was not statistically significant in this instance.

5 DISCUSSION & FUTURE WORK

The aim of the current study was to establish if modelling HoI from objective measures of the sounds is feasible and it can be regarded as successful in this respect. The process has revealed two primary issues that need to be addressed in the development of VAC functionality for real-world application.

Firstly, the amount of labelled data available is a significant issue to address before further ML analysis. The dataset of 40 sounds derived previously [6] is useful for initial modelling attempts to assess the application of ML techniques to the domain. However, given the performance of Deep Learning (DL) algorithms in the environmental sound classification literature it should be regarded as likely that a superior performing model can be derived once a suitable dataset is compiled. We also note that Mesaros et. al [63] recommend quantity over quality of data for sound classification applications, while accepting poor quality data invalidates findings. Application of DL techniques to this domain will require a labelled dataset of significantly greater size than used in the foregoing. This could potentially be compiled by combining subjective ratings with Active Learning [64] techniques.

Secondly, further investigation is required on the hypothesised impact of how attentional, contextual and other processes, as outlined in Section 1, affect our perception of auditory hierarchies. Sound context, for example, may prove a more important indicator of importance than visual accompaniment, suggesting that a weighted schema could be derived experimentally which would model how different factors affect hierarchical categorisation and auditory scene perception. Once complete, such a schema would inform the functioning of a VAC codec meaning that auditory objects could be compressed in terms of their importance to sound scene perception. Thus, audio content could be flexibly delivered to consumers taking cognisance of the mode of consumption and the capacity of the delivery mechanism involved.

6 ACKNOWLEDGMENT

This work was supported by the Irish Research Council and DTS Licensing Ltd. under project code EBPPG/2016/339.

7 REFERENCES

[1] "Google Home - Smart Speaker & Home Assistant - Google Store," URL `https://store.google.com/product/google{_}home`.

[2] J. Seo, J.-H. Yoo, T. Park, T. Lee, M. Yoo, G. Jang, J.-H. Won, Y. Choi, "Soundbar System with Embedded Multichannel Digital Amplifier SoC," presented at the *Proceedings of the 138th Convention of the Audio Engineering Society (2015)* (2015 May).

[3] E.-J. Völker, "Home Cinema Surround Sound-Acoustics and Neighborhood," presented at the *Proceedings of the 100th Convention of the Audio Engineering Society (1996)* (1996 May).

[4] M. Armstrong, M. Brooks, A. Churnside, M. Evans, F. Melchior, M. Shotton, "Object-based Broadcasting – Curation, Responsiveness and User Experience," (2014), URL `http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP285.pdf`.

[5] T. Churnside, "Object-Based Broadcasting," (2013), URL `http://www.bbc.co.uk/rd/blog/2013-05-object-based-approach-to-broadcasting`.

[6] W. Coleman, C. Cullen, M. Yan, "Categorisation of Isolated Sounds on a Background - Neutral - Foreground Scale," presented at the *Proceedings of the 144th Convention of the Audio Engineering Society*, pp. 1–9 (2018), doi:https://doi.org/10.13140/2.1.1598.6882.

[7] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organisation of Sound* (The MIT Press, Cambridge, MA) (1990).

[8] W. J. Davies, M. D. Adams, N. S. Bruce, R. Cain, A. Carlyle, P. Cusack, D. A. Hall, K. I. Hume, A. Irwin, P. Jennings, M. Marselle, C. J. Plack, J. Poxon, "Perception of Soundscapes: An Interdisciplinary Approach," *Applied Acoustics*, vol. 74, no. 2, pp. 224–231 (2013), doi:https://doi.org/10.1016/j.apacoust.2012.05.010.

[9] M. Raimbault, D. Dubois, "Urban Soundscapes: Experiences and Knowledge," *Cities*, vol. 22, no. 5, pp. 339–350 (2005), doi:https://doi.org/10.1016/j.cities.2005.05.003.

[10] C. Guastavino, "The Ideal Urban Soundscape: Investigating the Sound Quality of French Cities," *Acta Acustica united with Acustica*, vol. 92, no. 2006, pp. 945–951 (2006).

[11] J. Woodcock, W. J. Davies, T. J. Cox, A. Member, F. Melchior, "Categorization of Broadcast Audio Objects in Complex Auditory Scenes," *Journal of the Audio Engineering Society*, vol. 64, no. 6 (2016), doi:https://doi.org/10.17743/jaes.2016.0007.

[12] O. Rummukainen, J. Radun, T. Virtanen, V. Pulkki, M. M. Murray, "Categorization of Natural Dynamic Audiovisual Scenes," *PLoS ONE*, vol. 9, no. 5, p. 14 (2014), doi:https://doi.org/10.1371/.

[13] C. Guastavino, "Categorization of Environmental Sounds," *Canadian Journal of Experimental Psychology*, vol. 61, no. 1, pp. 54–63 (2007), doi:https://doi.org/10.1037/cjep2007006.

[14] J. Salamon, C. Jacoby, J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," presented at the *Proceedings of the ACM International Conference on Multimedia - MM '14*, pp. 1041–1044 (2014), doi:https://doi.org/10.1145/2647868.2655045.

[15] A. L. Brown, J. Kang, T. Gjestland, "Towards Standardization in Soundscape Preference Assessment," *Applied Acoustics*, vol. 72, no. 6, pp. 387–392 (2011 May), doi:https://doi.org/10.1016/j.apacoust.2011.01.001.

[16] P. Lindborg, "A Taxonomy of Sound Sources in Restaurants," *Applied Acoustics*, vol. 110, pp. 297–310 (2016), doi:https://doi.org/10.1016/j.apacoust.2016.03.032.

[17] J. Woodcock, W. J. Davies, T. J. Cox, "A Cognitive Framework for the Categorisation of Auditory Objects in Urban Soundscapes," *Applied Acoustics*, vol. 121, no. 2017, pp. 56–64 (2017), doi:https://doi.org/10.1016/j.apacoust.2017.01.027.

[18] J. Sussman-Fort, E. Sussman, "The Effect of Stimulus Context on the Buildup to Stream Segregation," *Frontiers in Neuroscience*, vol. 8, no. 8 APR, pp. 1–8 (2014), doi:https://doi.org/10.3389/fnins.2014.00093.

[19] B. Gygi, G. R. Kidd, C. S. Watson, "Similarity and Categorization of Environmental Sounds," *Perception & Psychophysics*, vol. 69, no. 6, pp. 839–855 (2007).

[20] D. Pressnitzer, J. Graves, C. Chambers, V. de Gardelle, P. Egré, "Auditory Perception: Laurel and Yanny Together at Last," *Current Biology*, vol. 28, no. 13, pp. R739 – R741 (2018), doi:https://doi.org/10.1016/j.cub.2018.06.002.

[21] J. W. Lewis, W. J. Talkington, K. C. Tallaksen, C. a. Frum, "Auditory Object Salience: Human Cortical Processing of Non-Biological Action Sounds and their Acoustic Signal Attributes," *Frontiers in Systems Neuroscience*, vol. 6, no. May, pp. 1–15 (2012), doi:https://doi.org/10.3389/fnsys.2012.00027.

[22] M. Thorogood, J. Fan, P. Pasquier, "Soundscape Audio Signal Classification and Segmentation Using Listener's Perception of Background and Foreground Sound," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 484–492 (2016), doi:https://doi.org/10.17743/jaes.2016.0021.

[23] B. Truax, "World Soundscape Project Tape Library," (2015), URL `http://www.sfu.ca/sonic-studio/srs/index2.html`.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York), 1st ed. (2006), doi: https://doi.org/10.1117/1.2819119.

[25] T. Virtanen, M. D. Plumbley, D. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events* (Springer International Publishing) (2018), doi:https://doi.org/10.1007/978-3-319-63450-0.

[26] N. Dave, "Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition," *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. Vi, pp. 1–5 (2013).

[27] H. Hermansky, N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589 (1994).

[28] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," (1989), doi:https://doi.org/10.1109/5.18626.

[29] Y.-H. Yang, H. H. Chen, "Machine Recognition of Music Emotion: A Review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 1–30 (2012), doi:https://doi.org/10.1145/2168752.2168754.

[30] T. Virtanen, M. D. Plumbley, D. P. W. Ellis, "Introduction to Sound Scene and Event Analysis," in T. Virtanen, M. D. Plumbley, D. P. W. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events*, chap. 1, pp. 3–12 (Springer International Publishing), 1st ed. (2018).

[31] C. Joder, S. Essid, G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 174–186 (2009), doi:https://doi.org/10.1109/TASL.2008.2007613.

[32] J.-C. Wang, J.-F. Wang, K. W. He, C.-S. Hsu, "Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-level Descriptor," presented at the *International Joint Conference on Neural Networks, 2006. IJCNN'06.*, pp. 1731–1735 (2006).

[33] V. Bountourakis, L. Vrysis, G. Papanikolaou, "Machine Learning Algorithms for Environmental Sound Recognition," presented at the *Proceedings of the Audio Mostly 2015 on Interaction With Sound - (AM15)*, pp. 1–7 (2015), doi:https://doi.org/10.1145/2814895.2814905.

[34] H. B. Sailor, D. M. Agrawal, H. A. Patil, "Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification," *Proceedings of Interspeech 2017*, pp. 3107–3111 (2017).

[35] "DCASE 2017 Challenge," URL `http://www.cs.tut.fi/sgn/arg/dcase2017/index`.

[36] "DCASE 2018 Challenge," URL `http://dcase.community/challenge2018/index`.

[37] F. Rumsey, "Semantic Audio," *J. Audio Eng. Soc*, vol. 62, no. 4, pp. 281–285 (2014).

[38] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *Journal of the Audio Engineering Society*, vol. 33, no. 1/2, pp. 2–32 (1985).

[39] U. Wüstenhagen, B. Feiten, J. Kroll, A. Raake, M. Wältermann, "Evaluation of Super-Wideband Speech and Audio Codecs," presented at the *Proceedings of the 129th Convention of the Audio Engineering Society (2010)*, p. Paper 8205 (2010).

[40] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," (2015), doi:https://doi.org/10.1145/2733373.2806390.

[41] "Audacity ® — Free, open source, cross-platform audio software for multi-track recording and editing." URL `https://www.audacityteam.org/`.

[42] "What is MATLAB?" URL `https://uk.mathworks.com/discovery/what-is-matlab.html`.

[43] T. Giannakopoulos, "Matlab Audio Analysis Library," (2014), URL `https://uk.mathworks.com/matlabcentral/fileexchange/45831-matlab-audio-analysis-library`.

[44] T. Giannakopoulos, A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach* (Elsevier Academic Press, Oxford, UK) (2014), doi:https://doi.org/10.1016/C2012-0-03524-7.

[45] L. Yang, F. Su, "Auditory Context Classification using Random Forests," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, pp. 2349–2352 (2012).

[46] M. Malfante, J. I. Mars, M. Dalla Mura, C. Gervaise, "Automatic Fish Classification," *Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2834–2846 (2018), doi:https://doi.org/10.1121/1.5036628.

[47] M. Esfahanian, H. Zhuang, N. Erdol, "Sparse Representation for Classification of Dolphin Whistles by Type," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. EL1–EL7 (2014), doi: https://doi.org/10.1121/1.4881320.

[48] R. Cai, L. Lu, A. Hanjalic, H. J. Zhang, L. H. Cai, "A Flexible Framework for Key Audio Effects Detection and Auditory Context Inference," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1026–1038 (2006), doi:https://doi.org/10.1109/TSA.2005.857575.

[49] P. Ruvolo, I. Fasel, J. R. Movellan, "A Learning Approach to Hierarchical Feature Selection and Aggregation for Audio Classification," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1535–1542 (2010).

[50] H. Jiang, J. Bai, S. Zhang, B. Xu, "SVM-based audio scene classification," presented at the *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05*, vol. Oct 30 - N, pp. 131–136 (2005), doi: https://doi.org/10.1109/NLPKE.2005.1598721.

[51] R. N. Tak, D. Agrawal, H. Patil, "Novel Phase Encoded Mel Filterbank Energies for Environmental Sound

Classification," presented at the *International Conference on Pattern Recognition and Machine Intelligence 2017*, pp. 317–325 (2017).

[52] G. Parascandolo, H. Huttunen, T. Virtanen, "Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings," presented at the *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6440–6444 (2016 Mar.), doi: https://doi.org/10.1109/ICASSP.2016.7472917.

[53] "DCASE2017," (2017), URL `http://www.cs.tut.fi/sgn/arg/dcase2017/index`.

[54] S. Mun, S. Park, D. Han, H. Ko, "Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane," Tech. rep., DCASE2017 Challenge, Munich, Germany; 16th November (2017 Sep.).

[55] Y. Han, J. Park, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," Tech. rep., DCASE2017 Challenge, Munich, Germany; 16th November (2017 Sep.).

[56] S. Krstulovic, "Audio Event Recognition in the Smart Home," in T. Virtanen, M. D. Plumbley, D. P. W. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events*, chap. 12, pp. 335–371 (Springer International Publishing), 1st ed. (2018).

[57] A. C. Muller, S. Guido, *Introduction to Machine Learning with Python* (O'Reilly Media, Sebastopol, United States) (2017), doi:https://doi.org/10.1017/CBO9781107415324.004.

[58] P. Cunningham, M. Cord, S. J. Delany, "Supervised Learning," in M. Cord, P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia*, pp. 21–49 (Springer Berlin Heidelberg, Berlin, Heidelberg) (2008), doi:https://doi.org/10.1007/978-3-540-75171-7_2.

[59] T. Özseven, "A Novel Feature Selection Method for Speech Emotion Recognition," *Applied Acoustics*, vol. 146, pp. 320–326 (2019 Mar.), doi:https://doi.org/10.1016/J.APACOUST.2018.11.028.

[60] P. Cunningham, "Dimension Reduction," in M. Cord, P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia.*, chap. 4, pp. 1–24 (Springer, Berlin, Heidelberg, Dublin) (2008).

[61] R. Kohavi, G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324 (1997 Dec.), doi:https://doi.org/10.1016/S0004-3702(97)00043-X.

[62] G. W. Corder, D. I. Foreman, *Nonparametric Statistics for Non-Statisticians: A Step-by-step Approach* (John Wiley & Sons, Inc., Hoboken, NJ, USA) (2009 May), doi:https://doi.org/10.1002/9781118165881.

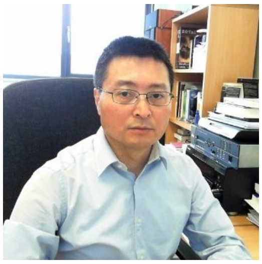
[63] A. Mesaros, T. Heittola, D. Ellis, "Datasets and Evaluation," in T. Virtanen, M. D. Plumbley, D. P. W. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events*, chap. 6, pp. 147–179 (Springer International Publishing), 1st ed. (2018).

[64] S. Tong, E. Chang, "Support Vector Machine Active Learning for Image Retrieval," presented at the *Proceedings of the 9th ACM international Conference on Multimedia*, pp. 107–118 (2001).

THE AUTHORS

William Coleman

Sarah Jane Delany


Ming Yan

Charlie Cullen

William Coleman is an Irish Research Council funded Ph.D candidate at Technological University Dublin. He received his Masters in Music and Media Technology from Trinity College Dublin in 2015. His research interests cover both scientific and artistic concerns, including the psychoacoustics of sound hierarchies and machine hearing around which his doctoral research is based and the application of technology to novel compositional systems, such as visual music. He is a published academic and has presented & performed extensively in commercial, academic and creative contexts.

●

Professor Sarah Jane Delany is Head of Postgraduate Studies and Research in the School of Computer Science at Technological University Dublin. Research interests include machine learning, active learning, text data analysis and handling concept drift in online learning.

●

Dr. Ming Yan is currently a Fellow, Audio Processing at DTS/Xperi. He has 25 years industrial experience in the field of audio codec and processing technologies. He is also responsible for building and enhancing DTS R&D and has managed several collaborative projects with partners from UK/Ireland/EU. Research interests include immersive audio technologies, audio compression, upmixing and processing, and machine learning.

Dr. Charlie Cullen is Head of the Institute of Creative Technologies and Applied Computing (ICTAC) as a Reader in Creative Computing in the University of the West of Scotland, where he combines lecturing and research into various aspects of digital media technologies. Dr. Cullen has completed many research projects, and has published widely in many areas related to creative technologies.