DETECTING SOUND OBJECTS IN AUDIO RECORDINGS

Anurag Kumar, Rita Singh and Bhiksha Raj

Carnegie Mellon University, Pittsburgh USA-15213
{alnu,rsingh,bhiksha}@cs.cmu.edu

ABSTRACT

In this paper we explore the idea of defining *sound objects* and how they may be detected. We try to define sound objects and demonstrate by our experiments the existence of these objects. Most of current works on acoustic event detection focus on detecting a finite set of audio events and the detection of a generic object in sound is not done. The major reason for proposing the idea of sound objects is to work with a generic sound concept instead of working with a small set of acoustic events for detection as is the norm. Our definition tries to conform to notions present in human auditory perception. Our experimental results are promising, and show that the idea of sound objects is worth pursuing and that it could give a new direction to semi-supervised or unsupervised learning of acoustic event detection mechanisms.

Index Terms— Sound Objects, Acoustic Event Detection

1. INTRODUCTION

The key to automated machine understanding of audio data is description of its content. Audio descriptions are generally provided in terms of the distinct, identifiable sound units detected in the recordings. Traditionally, these units have been human-identifiable acoustic events, detectors for which are learned from annotated data. Needless to say, these events must be from a finite vocabulary of events for which it was possible to train detectors – one might anthropomorhpize this to state that the automated system only detects the events it is familiar with.

Contrast this with human (and possibly other animal) listeners. While we definitely do identify sound events that we are familiar with, we are often also able to detect the occurrence of sound events or acoustic "objects" that we have never encountered earlier, based only on how the phenomenon stands out against the background – an ability that greatly enables us to form our own vocabulary of sounds from repetitions of the detected novel phenomena. Cognitively, it is argued by several researchers, there may be an underlying model of "objectness" that human (or animal) listeners subscribe to, and that we are able to detect the occurrence of acoustic phenomena that conform to this notion of objectness, even when we are unfamiliar with the event itself [1, 2, 3].

This observation motivates the work we present in this paper. Instead of working with a small vocabulary of pre-specified acoustic events, we explore a generic concept called sound "objects" which, as we define it, is closely related to human perception of sound. We draw upon psycho-acoustic models of human perception to claim that there are some inherent characteristics to sound objects which are present no matter what the actual object is. Since our goal is computational modeling, we then provide a formal definition of sound objects; however our definition is phenomenological rather than psycho-acoustically motivated – if humans can detect them, they are objects. We then show experimentally that it is indeed possible to automatically detect sound objects as defined, without any reference to the semantics of their content. Although in this work we do not actually explicitly demonstrate it, based on our experimental evidence for detection of sound objects we argue that such detection can hold significant importance for acoustic event detection in semi-supervised and unsupervised paradigms.

Before proceeding, we will first present a very brief survey of some of the recent literature in detecting acoustic events. The literature on the topic is large and, with apologies to authors whose work we may not have cited, our review is only a sampling. We note here that the terms "acoustic event", "sound object", and other permutations of the four words have been used interchangeably in the literature. Consequently, the sound phenomena we call "objects" in this paper may be viewed as "events" by some researchers. We do not attempt to clear the confusion, beyond pointing out that by our convention objects are cognitively distinct units, and while they may form events or be constituents of larger events, they will not themselves be comprised of events.

Mesaros *et al.* [4] use Hidden Markov Models to model each of a set of 60 events. The effects of background noise on detection and temporal localization of events in audio is also studied in this work. Lee *et al.* [5] model events as states of a Markov model. The formulation is intended for weakly-supervised learning scenarios where the exact time-stamps within which events occur in the training data is unknown. A total of 25 events are evaluated. Lu *et al.* [6] utilize an SVM on features derived from segments of the audio, and smooth the labels obtained to reduce false alarms. A particularly effective structure-free approach employs bag-of-words characterizations of the audio, obtained by clustering feature vectors derived from the audio. Pancoast and Akbacak [7], and by Jiang *et al.* [8], report approaches that use these in conjunction with SVM classifiers for event detection in the context of multimedia event detection. They are also used in a slightly different context for copy detection in audio by Liu *et al.* [9]. An interesting method is proposed in Lee *et al.* [10] which models event classes as Gaussians and employs probabilistic latent semantic analysis of Gaussian component histograms on soundtracks of videos to identify types of videos. Zhuang *et al.* [11] use a speech recognition framework based on HMMs for detection of events. They also attempt to identify the right set of features for the purpose in the process. Valenzise *et al.* [12] and Pikrakis *et al.* [13] attempt to detect events relevant to surveillance: Valenzise *et al.* [12] use a GMM-based classifier for gunshot detection, and Pikrakis *et al.* [13] use Bayesian networks for the same purpose. Kumar *et al.* [14] employ a simple Gaussian mixture classifier to detect events.

Notably, although all of these and other relevant literature approach the problem of acoustic event detection from many perspectives, in all cases the attempt has been to develop detectors for a finite (and inexhaustive) set of *known* events : the notion of just detecting objects without reference to their underlying category, *i.e.* without

specific *a priori* information about them has been absent.

The rest of the paper is organized as follows: In Section 2 we discuss the idea of a sound object, placing it in the context of studies on human cognition. In Section 3 we present our human-centric definition of sound objects. In Section 4 describes our approach for automatic detection of sound objects. In Section 5 we present our experimental set up and results and in Section 6 we discuss our results, and present our conclusions.

2. WHAT IS A SOUND OBJECT

At the basis of our work is the definition of a sound object. We begin by noting that the concept of an object is itself hard to articulate, and philosophers though time, from Leucippus and Plato to Kant [15] and Russel [16] have struggled to define it. Cognitive scientists too have found it difficult to arrive at a precise definition. From a purely cognitive point-of-view, objects are defined as the bases of experience [2] – a definition that the Merriam Webster dictionary agrees with in defining an object as "something material that may be perceived by the senses". In effect, by these definitions, objects are perceptual entities that are fundamentally a function of the sensory processes that perceive them. Even so, a concrete definition remains elusive, with definitions largely being in the nature of a composition of parts or percepts, *e.g.* [17, 18].

And yet, in the physical world, the concept of an object is something all of us are familiar with, in spite of inherent ambiguities (*e. g.* a door is handle an object, but then so is a door that it is a part of). To contextomize American jurist Potter Stewart, "we know one when we see one". The word "see" is particularly relevant here – although it is agreed that the concept of "objecthood" extends to perceptions derived from all senses, the majority of the discussions and descriptions in the literature have centered on *visual* objects.

Not surprisingly then, researchers in *computer* vision too have attempted to emulate this human facility of detecting (or hallucinating) objects in their visual field. In the context of computer vision, this translates to detecting objects in digital images, based only on their "objectness". For computational purposes, however, they have avoided hierarchical-grouping-based definition of objects such as in [18], and work from the simpler saliency based description given in Alexe *et al.* [19]: an image object is defined as something which holds at least one of the following three characteristics (a) it has a well defined closed boundary in space (b) it has an appearance which is different from the surrounding (c) stands out as salient or is unique within the image. A significant literature has since sprung up that builds on the concept of objectness to detect objects in images.

In this paper we are, however, interested in detecting *sound* objects. In particular, we are interested in detecting them from the signal without using spatial cues, since such cues will not be available in typical digital audio recordings. Pierre Schaefer's work [20] on sound objects is one of the early works on theory of sound objects and our definition of sound objects closely follows his idea of it.

The concept of a "sound object" follows the same rationale as an "image" object. Sound objects are distinct acoustic percepts that we distinguish from the background as possibly representing a cogent phenomenon or source that stands apart from the background. Unfortunately, they too suffer from the Potter-Stewart syndrome: we know them when we encounter them, but they are hard to define.

Since objects are essentially cognitive constructs, let us then refer to researchers in auditory cognition to obtain a definition. From a cognition perspective, sound objects – if they exist – are sensory entities built upon auditory stimuli; hence, while we refer to "sound" objects in deference to the fact that these objects are definitive units that we expect to detect in digital recordings of sound, the literature in cognition refers to "auditory" objects. For want of a convincing argument to the contrary, we will assume that sound objects are the same as auditory objects.

Even among auditory neuroscientists we encounter debate and diversity of definition. Bregman [21] rejects the very notion of an auditory object, claiming instead that auditory "streams" are the fundamental units of the auditory world. More commonly, auditory objects are associated with sources [2]. The flaw in this association is that a source may produce more than one type of sound; other phenomena that a human may identify as a sound object (*e.g.* a snippet of music) may have many sources. The peripheral auditory system constructs two-dimensional characterizations of sound akin to "images" [22, 23]. Based on this researchers such as Kubovy and Van Valkemberg [1] propose a visual analogy that resembles the propositions in Alexe *et al.* [19]: auditory objects are defined as salient phenomena that lie within clear boundaries in two-dimensional characterizations such as spectrograms. This analogy however breaks down quickly; by this rule individual formants in a recording would stand out as objects. In fact, although we have used the visual analogy to hypothesize sound objects, it is questionable how much the analogy can be stretched. The fundamental difference is that visual objects are formed from *presence* of physical objects, while sound objects result from their *actions*. Lemaitre and Heller [24] studied a taxonomy of everyday sound and concluded that sounds produced through "specific actions" are easier to identify compared to general ones, a characteristic that may be at variance with visual objects.

Yet others researchers, *e.g.* [3] propose that auditory objects are formed through grouping of time-frequency components based on perceptual expectations. Unfortunately this definition is simultaneously over generative and over inclusive – grouping can group parts of what a human would identify as an object separately; at the same time a background, which too is not an object, too would be grouped. Moreover, segregation and grouping of time-frequency components may occur at many scales, not all of which conform to our notion of objects. Other scientists have proposed *detection-based* definitions, via models of the cognitive processes that find acoustic objects: Shamma [23] suggests that objects are identified from coincident spike discharges in separated auditory nerves, Husain *et al.* [25] propose that objects may be detected through a hierarchical 3-stage process in the brain, Griffiths and Warren [2] suggest a cascade of processes, and Adams and Janata [26] propose a template based model for auditory object detection.

Clearly then, although the literature provides many hints, it does not provide a definitive answer to what a sound object may be, and how it can be characterized in terms that will allow us to build a computational model. Some of the principles proposed in the literature are, however, useful. Griffiths *et al.* [2] aver that auditory objects must be temporally restricted. Clearly sound objects possess saliency, as suggested by Kubovy and Vanvalkemberg [1]. They must stand out against the background. We can also clearly specify what sound objects are *not* – just as visual objects are distinctly not textures, sound objects too are not sound textures. Sound textures are analogous to visual textures and are produced by superposition of multiple and rapidly occurring acoustic events [27, 28]. Examples are sounds produced by naturally occurring events such as rainstorms, insect swarms, fires, galloping horse etc. And yet, although they are composed from events, textures end up possessing temporal homogeneity, which we claim objects need not have.

Beyond these guidelines, we must eventually state a computationally addressable definition of sound objects ourselves.

3. DEFINING SOUND OBJECTS

Possibly the clearest way to define sound objects is one alluded to earlier: based on human judgment. Simply stated, our definition will be – if a human can detect it, it is an object.

This leads us then, to an empirical characterization based on human annotated data. We will ask humans to detect and annotate sound events heard in a collection of audio recordings. Since the annotations are done by humans, we can claim that these are guided by the human cognitive notion of sound objects. We will then evaluate whether these annotations can be used to learn a sound-object detector.

Following this, a collection of audio recording was given to human annotators, who were asked to detect and annotate the start and end times of sound events heard in the recordings. For purposes we explain shortly, the annotators were also asked to assign labels to the detected events. The labels, which were selected by the annotators themselves, included sounds such as "birds", 'crowd cheer", "drums" etc. However, the labels were post hoc and not imposed. Some level of inconsistency is to be expected. The act of assigning labels imposes cognitive constraints; some events will be missed. Moreover the basic notion of being able to identify *unfamiliar* objects may not be satisfied. Nevertheless, we can hypothesize that the annotation provides an empirical characterization of both – what *is* an object, through the objects that are tagged, and what is *not* an object, through the regions of the audio that are not tagged as being objects. The question of whether a generic concept of sound objects exists will now simply be answered as follows: is it possible to learn a detector from the data that can detect objects of categories/labels that are not included in the training data? Without leaving the reader in suspense, the answer, as we will show later is "yes".

Post-hoc, we note from the annotated audio and find the following operative definitions to hold which follow the lines of Alexe *et al.* [19]. We note that the definition of sound objects is complicated by various factors such as that different audio events can overlap, and change in audio content can be attributed several factors such as amplitude, rhythm etc. Yet, similar issues such as occlusion, perspective etc. are also faced in computer vision; these are lateral to the definition of objects themselves. Sound object are generally noted to have the first two of the following properties and may or may not satisfy the third property. (a) they have clear onset time (b) they are acoustically salient – they posses characteristics that distinctly separate them from what listeners may perceive as background, noise or silence, and (c) the offset time is clear. The relaxation on offset time is because we find that the onset of a detected object is always noticeable to listeners, whereas the offset time may not be distinctly noted and has poor inter-annotator agreement.

Additional details of the data are in the experimental section.

4. DETECTING SOUND OBJECTS

Studies in computer vision have derived detectors based on saliency models, and characterizations obtained from the prescribed definition of objects, to build object detectors. Given the rather more empirical definition we have come up with, we will therefore treat the problem of sound-object detection one of binary classification, and attempt to learn models based on objects identified by human annotators. Note, once again, that we are not concerned with a particular class of sound events; instead we are concerned with the generic concept of sound objects. We would like to confirm that this description leads to the ability to detect sound objects within a computational paradigm. Thus we will separate the data we have by label, such that the labels that are present in the training data are not present in the test data and vice versa.

We employ a rather simple classification paradigm: the principle being that it sound objects are indeed distinctive, even a simple classifier must do a reasonable job of detecting them. We model the characteristics of sound objects using Gaussian Mixture Models(GMM). We train a background GMM of M components from a large collection of audio recording containing both sound objects and non-sound objects. We then generate two different features using this background GMM [14]. The first set of features represent the distribution of the data over the *components* of background GMM. The second set of features represent the actual distribution of data over the entire GMM. Since we aim to capture generic sound-object characteristics, we keep the number of Gaussian components small, to avoid learning highly discriminative clusters obtained using a large number of Gaussians, which might capture characteristics of individual events.

4.1. Probabilistic Feature Vectors

We repesent all audio as sequences of mel-frequency cepstral coefficient (MFCC) vectors, since they are perceptually motivated. The MFCCs for any recording are represented by *d* dimensional vector $\vec{x_t}$ where *t* goes from 1 to T. T is the total number of mel-frequency cepstral coefficients vectors for the given recording. For illustration, we will view the Gaussians in the GMMs as clusters. Then for each component i of the background GMM the probability that the data belonging to the i^{th} cluster is computed as:

$$P(i) = \sum_{i=1}^{T} p(\vec{x_t}|\lambda_i) \tag{1}$$

where λ_i collectively represents the mean and covariance parameters for the i^{th} Gaussian component of the background Gaussian Mixture Model. Since we are not looking for fixed length audio, instance normalization for varying lengths for different audio is done as

$$P(i) = \frac{1}{T} \sum_{i=1}^{T} p(\vec{x_t}|\lambda_i) \tag{2}$$

M dimensional feature vector $\vec{F}$ is obtained from this where each element of $\vec{F}$ is equal to $P(i)$. This feature vector $\vec{F}$ captures the probabilistic distribution of all the mel-frequency cepstral vectors of the recording over the Gaussian components of the background GMM. It is similar to the word-frequency histogram representation used in the bag-of-words representations, but is much more robust [14] due to soft assignment and the more detailed characterization in GMMs.

4.2. GMM-MAP features

To obtain a more effective and robust feature representation of sound objects, we obtain GMMs for each audio recording by adapting the background GMM. The means of the background GMM are adapted to each training recording using the maximum-a-posteriori (MAP) criterion as described in [29]. This is done as follows for i^{th} component of the mixture

$$Pr(i|\vec{x_t}) = \frac{w_i p(\vec{x_t}|\lambda_i)}{\sum_{j=1}^{M} p(\vec{x_t}|\lambda_j)} \tag{3}$$

$$n_i = \sum_{t=1}^{T} Pr(i|\vec{x_t}) \tag{4}$$

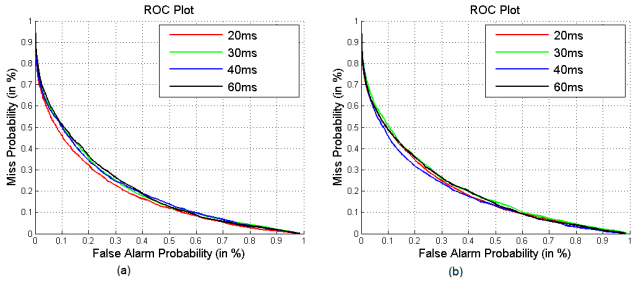

Fig. 1. ROC plots for sound objects detection with different window sizes

$$E_i(\vec{x}) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\vec{x_t})\vec{x_t} \tag{5}$$

w_i is the weight of i_{th} Gaussian component.
Finally the updated means are

$$\hat{\vec{\mu_i}} = \frac{n_i}{n_i + r} E_i(\vec{x}) + \frac{r}{n_i + r}\vec{\mu_i} \tag{6}$$

where $\vec{\mu_i}$ is the mean vector of i^{th} component of the background GMM and r is a relevance factor. The means of all components are then appended to form a new vector of $M \times d$ dimensions. This vector is a representative of the distribution of vectors in the audio. The above two features namely $\vec{F}$ and GMM-MAP features are unsupervised methods of characterizing acoustic events in audio.

4.3. Classification

We use a random forest classifier in our experiments [30]. It is method of ensemble learning which builds a "forest" of decision trees, each node of which is computed using only a random subset of input variables. The final decision is performed through voting. In a study by [31] it has proven be among the best classification strategies .

5. EXPERIMENTS AND RESULTS

We use the TRECVID MED 2011 database in our experiments. A set of 31 events spanning over 410 audio files from the TRECVID database is used as our set of sound objects. The labels assigned are varied, such as *crowd cheer*, *drums*, etc. and represent a variety of types of acoustic saliency. The complete list of all 31 events is not given here due to space constraints. Each of these 410 audio files were manually annotated; the part of audio data belonging to the set of 31 events were instances of sound objects. We divide these sound objects into two groups, one with 16 randomly chosen events and another with the remaining 15. Two-way jackknife experiments were performed, using one set for training and another for test in each case to get a performance measure over the entire data. Negative instances for sound objects are obtained from those audio files in which annotators have not found any of these events. Again we would like to emphasize here that there are no positive training instances for the set of test events and hence our experiment is generic.

Mel Frequency Cepstral Coefficients (MFCCs) are used as the raw features for the audio data. 20 dimensional MFCC feature vectors are extracted with different window sizes. The window sizes considered are 20ms, 30ms, 40ms and 60ms. The window in each case is moved by 10ms resulting in 50%, 66.66%, 75% and 83.33% overlap respectively. Experiments are performed using each of these window sizes separately. In another set of experiments we use delta MFFCs along with MFCCs resulting in a total of 40 dimensional raw features.
The background GMM is learned from a completely different set of audio data which includes both sound objects and background. The total amount of data for GMM training is slightly more than 10 hours. 4 different GMMs with component size M=64 are learned corresponding to 4 different window sizes used for MFCCs extraction. The same setup is used for the case when delta coefficients are also used. The value of 'r' in GMM-MAP adaptation is fixed to be 0.5. The number of trees in random forest is set to be 500.
Fig 1(a) shows the miss probability vs. false alarm probability for different window sizes without using delta coefficients of MFCCs. Fig 1(b) shows the same values when we use delta coefficients along with MFCCs in our experiments. The area under the curve (AUC) and Equal Error Rate (EER) are two characteristics normally used as single metrics for evaluating ROC plots. Since we are plotting error curves the area under the curve should be as low as possible. The EER too must be also be as small as possible. Table 1 shows these values for different cases.

Window (ms)	MFCCs		MFCCs+Delta	
	AUC	EER	AUC	EER
20	0.1827	0.259	0.1986	0.271
30	0.2014	0.273	0.2099	0.281
40	0.2007	0.272	0.1887	0.264
60	0.2042	0.283	0.2032	0.283

Table 1. Area Under Curve (AUC) and EER values for ROC curves in Fig 1

6. DISCUSSIONS AND CONCLUSIONS

Through very simple characterization techniques we are able to achieve a reasonable performance in detection of sound objects. Although using different window sizes in the current setup does not show any remarkable change in the performance we do believe that since we are dealing with a general concept viz. sound objects we need to check with different window sizes to see which is best suited. This might actually be visible in different characterization techniques. The best performance among all is achieved on window of 20ms. This might be attributed to the fact that small window will spread even very short lasting sound objects over multiple frames and thus help in detecting some signature characteristics. The most important conclusion drawn from this work is that we can consider the generic concept of sound objects as a complement to target-class-driven acoustic event detection. Even though there was no instance of test acoustic events in training, we are still able to get a reasonable performance using the very simple strategy shown in this paper, validating this idea. This might, in turn enable us to build up vocabularies of sound objects or events for audio annotation – a process that may in fact mimic how we as humans ourselves gain our vocabularies. Semantics may be obtained through the similarity between detected objects and their association with information from other modalities. This could also be used in hierarchical analyses, or as a bootstrap for obtaining human labels, to reduce annotation costs.

7. REFERENCES

[1] M. Kubovy and D. Van Valkenburg, "Auditory and visual objects," *Cognition*, vol. 80, pp. 97–126, 2001.

[2] Timothdy D. Griffith and Jason D. Warren, "What is an auditory object," *Nature Reviews Neuroscience*, vol. 7:2, pp. 252–256, 2003.

[3] B. J. Dyson and C. Alain, "Representation of concurrent acoustic objects in primary auditory cortex," *Journal of the Acoustic Society of America*, vol. 115(1), pp. 280–288, 2004.

[4] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *18th European Signal Processing Conference*, 2010, pp. 1267–1271.

[5] K. Lee, D. PW Ellis, and Alexander C Loui, "Detecting local semantic concepts in environmental sounds using markov model based clustering," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE Intl. Conf. on*. IEEE, 2010, pp. 2278–2281.

[6] Li Lu, Fengpei Ge, Qingwei Zhao, and Yonghong Yan, "A svm-based audio event detection system," in *Electrical and Control Engineering (ICECE), 2010 Intl. Conf. on*. IEEE, 2010, pp. 292–295.

[7] Stephanie Pancoast and Murat Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. of Interspeech*, 2012.

[8] Y Jiang, X Zeng, Guangnan Ye, S Bhattacharya, D Ellis, M Shah, and S Chang, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVid Workshop*, 2010, vol. 2, p. 6.

[9] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proceedings of the ACM Intl Conf. on Image and Video Retrieval*. ACM, 2010, pp. 89–96.

[10] Keansub Lee and Daniel PW Ellis, "Audio-based semantic concept classification for consumer video," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 18, no. 6, pp. 1406–1416, 2010.

[11] Xiaodan Zhuang, Xi Zhou, Thomas S Huang, and Mark Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE Intl. Conf. on*. IEEE, 2008, pp. 17–20.

[12] Gerosa Valenzise, L Gerosa, M Tagliasacchi, F Antonacci, and A Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conf. on*. IEEE, 2007, pp. 21–26.

[13] A. Pikrakis, T Giannakopoulos, and S Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE Intl. Conf. on*. IEEE, 2008, pp. 21–24.

[14] Anurag Kumar, Rajesh M Hegde, Rita Singh, and Bhiksha Raj, "Event detection in short duration audio using gaussian mixture model and random forest classifier," in *21st European Signal Processing Conference 2013 (EUSIPCO 2013)*, Marrakech, Morocco, Sept. 2013.

[15] I. Kant, *Critique of pure reason*, Macmillan, 2003.

[16] B. Russell, *A history of western philosophy*, Simon and Schuster, 1945.

[17] A. Treisman, "Properties, parts and objects," *Handbook of perception and human performance, Cognitive processes and performance*, vol. 2, pp. 35–1:35–70, 1986.

[18] Jacob Feldman, "What is a visual object," *Trends in Cognitive Science*, vol. 5, pp. 887–892, 2004.

[19] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "What is an object?," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 73–80.

[20] Michel Chion, "Guide to sound objects. pierre schaeffer and musical research," *Trans. John Dack and Christine North), http://www. ears. dmu. ac. uk*, 1983.

[21] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[22] R.D. Patterson, M.H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform," *Journal of the acoustic society of america*, vol. 98, pp. 1890–1894, 1995.

[23] Shihab Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Science*, vol. 5, pp. 340–348, 2001.

[24] Guillaume Lemaitre and Laurie M Heller, "Evidence for a basic level in a taxonomy of everyday action sounds," *Experimental Brain Research*, pp. 1–12, 2013.

[25] F. T. Husain, M. A. Tagamets, S. J. Fromm, A. R. Braun, and B. Horwitz, "Relating neuronal dynamics for auditory object processing to neuroimaging activity: a computational modeling and an fmri study," *Neuroimage*, vol. 21, pp. 1701–1720, 2004.

[26] R. B. Adams and Janata P., "A comparison of neural circuits underlying auditory and visual object categorization," *Neuroimage*, vol. 16, pp. 361–377, 2002.

[27] Josh H McDermott, Andrew J Oxenham, and Eero P Simoncelli, "Sound texture synthesis via filter statistics," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*. IEEE, 2009, pp. 297–300.

[28] Josh H McDermott and Eero P Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[29] F. Bimbot, Jean Bonastre, C Fredouille, G. Gravier, Ivan Magrin, S. Meignier, T. Merlin, Ortega Javier, Dijana Petrovska, and D A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.

[30] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] Rich Caruana and Alexandru Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.