Chapter 14
Future Perspective

Dan Ellis, Tuomas Virtanen, Mark D. Plumbley, and Bhiksha Raj

Abstract This book has covered the underlying principles and technologies of sound recognition, and described several current application areas. However, the field is still very young; this chapter briefly outlines several emerging areas, particularly relating to the provision of the very large training sets that can be exploited by deep learning approaches. We also forecast some of the technological and application advances we expect in the short-to-medium future.

Keywords Audio content analysis • Sound catalogues • Sound vocabularies • Audio database collection • Audio annotation • Active learning • Weak labels • Applications of sound analysis

14.1 Introduction

The foregoing chapters of this book have provided a comprehensive view of the state of the art in sound scene and event recognition, ranging from perceptual and computational foundations, through core techniques and evaluation, to a series of relatively advanced application areas. However, this is a young field which is

D. Ellis
Google Inc, 111 8th Ave, New York, NY 10027, USA
e-mail: dpwe@google.com

T. Virtanen (✉)
Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
e-mail: tuomas.virtanen@tut.fi

M.D. Plumbley
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford,
Surrey GU2 7XH, UK
e-mail: m.plumbley@surrey.ac.uk

B. Raj
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: bhiksha@cs.cmu.edu

T. Virtanen et al. (eds.), *Computational Analysis of Sound Scenes and Events*,
DOI 10.1007/978-3-319-63450-0_14

developing rapidly. This chapter provides brief descriptions of emerging trends that did not appear in earlier chapters, as well as some brief speculation about the technologies and applications likely to be important in coming years.

Given the dominance of large-scale machine learning in so many fields, in the next section we revisit the question of obtaining the data with which to train such classifiers, including the problem of defining a vocabulary of labels to use, and some approaches to working with imperfectly labeled data. We then go on to briefly outline some applications and approaches we see on the horizon for this field.

14.2 Obtaining Training Data

In many fields, deep learning approaches (as described in Chap. 5) have shown startling abilities to rival human abilities to recognize images, words, etc. In all cases, however, the best performance has relied on a combination of massive computational power with enormous training data. Because the sound scene and event recognition community has a relatively short history, there are no well-established sources for these kinds of "big data" training sets in our field. This deficit is stimulating a variety of current research.

14.2.1 Cataloguing Sounds

A first step to automated content analysis of sound recordings is to compose a catalogue of sound classes to detect in them. The catalogue has two components. The first is the set of *exemplars* or *models* for each of the sound classes. The second is the *labels* we assign to the classes, and by which we refer to them.

In principle the labels could be arbitrary, e.g., random character strings that uniquely identify the sound class. In practice, the output of audio content recognition is generally intended for downstream automated or manual analysis, and it becomes convenient, or even necessary, for the labels to be semantically meaningful words or phrases that a human analyst can interpret. In effect, we must "name" the sounds in linguistic terms. It follows, therefore, that an essential part of building sound catalogues is to "know" how to label sounds in linguistic terms.

Since the sound labels are intended to appeal to the human sense of semantics, the simplest solution is to involve humans directly in the collection and labeling of sounds, an approach we refer to as *manual* sound event vocabulary creation. Eventually, though, in order to obtain a more comprehensive and current catalogue of sound labels, we will require semi- or fully automated techniques that can mine the web and other knowledge sources to create a list of sound labels, a process we refer to as *automatic* vocabulary creation.

14.2.1.1 Manual Sound Event Vocabularies

The most natural way to obtain the labeled data needed for supervised training is to directly solicit labels for snippets of sound from human annotators. Supervised classification generally requires a predefined vocabulary of possible labels, but collecting unconstrained labels (such as the free-text tags on www.freesound.org) can lead to very large vocabularies with many synonyms (distinct terms referring to essentially the same sound event). Thus, it is preferable to establish a predefined, fixed vocabulary and to constrain annotators to only use terms from the vocabulary. This, however, raises several issues when compared to completely free annotation: How to define the vocabulary, and how to make sure the annotators are familiar with all the categories.

Both these problems become more serious as the size and scope of the vocabulary grows. For the Urban Sounds taxonomy, Salamon et al. [38] chose to work with a limited set of 10 sound event classes (from "car horn" to "children playing") that were both common and typical in their urban field recordings. These classes were intended to be representative, so the exact choices were not important, and there was no effort to cover all events. They then went through their raw data marking each instance of each event; raters could be expected to quickly learn the full set of 10 events and identify them without further prompting.

For the TUT Sound Events 2016 database, Mesaros et al. [30] asked annotators to mark every sound event they perceived with free-text labels specifying a noun (object) and verb (action). These labels were subsequently manually merged, yielding 18 sound event classes with adequate representation (at least 20–30 examples) to be used in classification experiments. The data were limited to two acoustic contexts, "Residential area" (7 event categories) and "Home" (11 event categories), making it possible to obtain reasonable coverage with few classes.

The AudioSet Ontology [15] adopts the ambitious goal of defining a sound event vocabulary that covers all environmental sounds at a uniform level of detail, manually constructed from seeds including WordNet [31] and "Hearst patterns" [16] applied to web text. The resulting hierarchy of around 600 events arranged under 7 top-level categories raises problems of making annotators aware of all the categories. Annotation was primarily achieved via "verification" (presenting annotators with a single or small set of candidate labels which were then marked as present or absent, but without asking whether other sounds were also present); there was some experimentation with an interactive labeling tool that allowed annotators to search for events whose definitions included provided keywords.

14.2.1.2 Automatic Creation of Sound Event Vocabularies

A larger problem is to automatically generate a comprehensive vocabulary of the set of all sound classes that one may expect to encounter in an audio recording. By "listing" sound classes, we mean the listing of their *labels*—words or phrases that represent distinct sounds (regardless of whether we actually have exemplars of these sounds or not). Generating such a list may be used as a precursor to identifying or collecting associated sound samples, e.g., for strongly or weakly supervised learning of classifiers, or more generally as an indicator of the awareness of the existence of these sounds. Such a list will, however, be much longer than would be feasible using the manual procedures mentioned in Sect. 14.2.1.1, and will require automated methods to produce.

The obvious way to compose the vocabulary is to list sounds by name—i.e., the noun or noun phrase canonically designated to represent the sound. Unfortunately, while some sounds do have names, e.g., onomatopoeic terms such as "squeal" and "chirp," etymologically rooted terms such as "music," or explicitly designated names, e.g., "Beethoven's ninth," many more, possibly the majority, are referred to by characteristics of their manner of production. In effect, the labels applied to the sounds are "descriptors" rather than "names."

Consequently, once past the named sounds, sound vocabularies must be extended either by *composing* the sound-descriptor phrases that act as sound labels using appropriate rules or by *mining* web corpora for them, or a combination of the two.

There are, however, several confounding factors. Sound-descriptor terms can refer to the *objects* that produce the sound (e.g., "airplane" or "wind chime"), the environment in which the sound is produced (e.g., "playground"), the mechanism that produces the sound (e.g., "sawing wood") or to more complex characterizations of the entire sound-producing phenomenon (e.g., "metal scraping on concrete" or "children in a playground"). The words composing the labels may themselves have no direct implication of sound, e.g., the term "car idling" evokes a type of sound, but "idling" by itself is not immediately associated with sound (consider "man idling"). Naive construction of sound label lists will thus result in many spurious entries.

Vocabulary-generation mechanisms should therefore take a two-step approach: first generate *candidate* sound labels, and subsequently *filter* them to eliminate spurious candidates. For instance, as mentioned above, the AudioSet Ontology [15] generated candidates by applying modified "Hearst patterns" [16] to identify hyponyms of the word "sound," but this set needed subsequent manual filtering. Säger et al. [37] generated a much larger set of candidate sound terms based on the principle that sound events arise from an object (i.e., a noun) engaged in a particular action (i.e., a verb) or in some specific state (i.e., an adjective). They collected a set of sound-relevant words including 1200 nouns, 40 verbs, and 75 adjectives, and composed all possible adjective–noun pairs (ANPs) and verb–noun pairs (VNPs). This overcomplete set was then pruned by keeping only the pairs that occurred as tag combinations for sound files on www.freesound.org, then further pruning away implausible results by rejecting combinations that relied on a single uploader, or that only occurred in conjunction with the same other pairs. This resulted in over 1000 sound concepts; Table 14.1 shows some of their examples.

Kumar et al. [26] took the pattern-based approach used by CMU's never-ending language learner, "NELL" [32], to generate sound labels from the ClueWeb corpus [11]. They noted that sound-descriptor phrases can often be disambiguated based on whether they can be prefixed by the words "*sound of*" without changing their meaning. Consequently, by matching the template "sound(s) of <Y>" where *Y* is any

Table 14.1 Examples of detected "sound" ANPs and VNPs

Howling dog	Splashing water
Heavy rain	Howling wolf
Crackling footsteps	Echoing phone
Heavy metal	Extreme noise
Gurgling water	Breaking snow

Table 14.2 Patterns for discovering sound concepts in text

Pattern		Example concept
P1	(DT) VBG NN(S)	Honking cars
P3	(DT) NN(S) VBG	Dogs barking
P5	(DT) NN NN(S)	String quartet
P6	(DT) JJ NN(S)	Classical music

VBG is the part of speech tag for verbs in the gerund form, *NN* for nouns, *DT* for determiners, and *JJ* for adjectives

phrase of up to four words to identify candidate phrases, followed by the application of a rule-based classifier to eliminate noisy candidates, they obtained a list of over 100,000 sound labels. Table 14.2 shows some of their grammatical patterns along with representative matches. Further, by applying a classifier to features extracted from a dependency path between a manually listed set of acoustic *scenes* and the discovered sound labels, they were also able to discover ontological relations, for instance, that *forests* may be associated with the sounds of "birds singing," "breaking twigs," "cooing," and "falling water," and that *churches* are associated with "children laughing," "church bells," "singing," and "applause."

The solutions described so far only consider *contiguous* word sequences as candidates for sound labels. More generally, sound-describing phrases may also be *extracted* from longer noun or verb phrases in which not all constituents relate to the sound. For instance, "a cat runs past a dog mewling" has the constituent "cat mewling." In preliminary experiments Pillai and Qazi [35] found that candidates may be formed by parsing sentences into their components and evaluating combinations of various constituents of the sentence. Subsequent classification of formed candidates using a support-vector machine applied to vector-space representations of the phrases derived from a neural network returned lists of sound labels that were judged to be more than 80% accurate through manual inspection.

The lists obtained by these techniques can be further refined by considering frequency of occurrence across different webpages, their co-occurrence or affinity with one another, and the *contexts* they occur in. For instance, Pillai and Qazi [35] found that phrases derived from sound-related Wikipedia pages had a much higher likelihood of being valid candidates than those obtained from the larger web. Eventually, however, the true test is whether these phrases can indeed be associated with audio recordings. Thus a key feature that may be tested is to determine the frequency with which the phrases co-occur with sound files, particularly in contexts where phrases co-occurring with sounds may be expected (e.g., on sites such as soundforge.org or YouTube, or on Wikipedia pages about sounds).

14.2.2 Opportunistic Data Collection

When deep learning approaches were being developed for image recognition, the need for large numbers of labeled examples was addressed by mining the many billions of online digital images for examples whose captions suggested they contained the desired object. Even if captioned images make up only a tiny fraction of the billion-plus photos uploaded every day, there is still a very good chance of finding labeled examples for any common subject, and very often those examples will be clear, well-composed pictures.

Sound events are different. There is no widespread culture of uploading brief recordings of specific sounds; the closest is www.freesound.org, which has only a few hundred thousand sound files total. However, videos, while perhaps a thousand times less numerous than photos, are still available in ample volumes, including a proportion with associated text and other metadata. Users are perhaps more likely to describe the objects in their videos instead of the sounds, but those sounds are often associated with specific objects.

Thus, by a series of assumptions (some strong, some more reliable), we can take soundtrack snippets from videos whose metadata makes us believe they could contain the particular sound event for which we are collecting examples. Our assumptions might not work out, in which case we have a snippet labeled positive for containing sound X which in fact does not; we call this the “noisy labeling” problem. Even if the sound event does occur, there may be a lot of uncertainty about *when* it occurs; for instance, a 2 min video whose title is “AWESOME GLASS SMASH” may contain only a single glass breaking sound lasting under a second; this uncertainty around event timing we call the “weak labeling” problem.

Hershey et al. [17] use YouTube metadata to assign labels to videos; these labels are both weak (each label is assumed relevant to the entire video, whereas in fact it may only relate to specific time ranges within the soundtrack) and noisy (the label inference may have assigned a label that is not relevant at all). Their 3000 labels (“song,” “motorcycle,” etc.) are oriented towards YouTube searching behavior and thus may not necessarily relate to sound events, but their results—obtaining a mean Average Precision of up to 0.2, where random guessing would give something less than 0.01—show that, overall, the problems of weak and noisy labels are less devastating than might be expected.

14.2.3 Active Learning

Since human annotators are the ultimate authority for labeling sound examples, there is a strong incentive to maximize the value obtained for the expense of human labeling by ensuring they are shown the most important examples. This is the idea behind *active learning*, which includes a human annotator within a machine learning “loop,” so that each new label provides maximum value to the automatic system [8]. Thus, instead of gathering labels for a large number of examples that simply confirm the confident predictions of a classifier trained on existing labels, only examples that the current system is most likely to misclassify (i.e., those on the boundaries between decision regions) are labeled. As the new labels improve the classifier, this boundary will shift, and the examples selected for labeling will change. Another aspect of this approach is to reduce annotator load by asking for less-precise judgments when the system can automatically refine them given some initial guidance; this philosophy has been effectively applied in image object segmentation, where annotators simply clicked a single point within an object, leaving the system to infer the most likely bounding box of the clicked object [34].

14.2.4 Using Unsupervised Data

Obtaining labeled sound event data is difficult or at least expensive, but unlabeled audio is plentiful, favoring any method able to exploit it. In [19], Jansen et al. process a million YouTube soundtracks (about 5 years of audio) using an online clustering system to produce millions of clusters. Although labels are not used in creating the clusters, they show that the resulting clusters are correlated with labels for labeled items, meaning that the unlabeled data can be used to help "regularize" a classifier trained on a smaller amount of labeled data.

Unlabeled data can also be used as a source of candidates for annotation [45]. Since many sound classes are rare, simply annotating randomly selected sound excerpts will have a very inefficient yield. Instead, given a few positive examples, excerpts with high acoustic similarity (by some measure) can be prioritized for annotation. An acoustic similarity measure that better approximates human similarity judgments (such as the embedding layer of a trained classifier) will give a correspondingly more useful prioritization.

14.2.4.1 Training with Weak Labels

While unlabeled data may be used to organize audio, eventually labeled data are needed to train classifiers for sound events. Ideally, these data would comprise isolated or cleanly segmented recordings of the target sound events. As mentioned earlier, such "strongly" labeled data are hard to come by, since the effort required to produce them is considerable.

It is much easier and cheaper to obtain weaker labels that merely indicate whether a particular sound event is *present* within a recording, without specifying additional details, such as the precise location of the event or even the number of times it occurs. Such labels may be obtained through manual annotation, e.g., the Google AudioSet corpus [15] which provides weak labels for 10-s snippets of audio. Alternatively, the labels may be inferred from the metadata or text attached to a recording, from analysis of any accompanying video, etc.

The focus now shifts to how best to train sound event classifiers with such weakly labeled data. How do we train classifiers to similarly tag (weakly label) other similarly sized snippets? At a finer level, can we use the weak labels to actually infer the *location* of the target sounds within the training data itself? Can we develop detectors to find and localize instances of the target events in novel *test* data?

These questions are analogous to a well-studied problem in the machine-learning literature: *multiple-instance learning* (MIL) [4]. Within the MIL paradigm, data instances are assumed to be grouped into *bags*. It is assumed that only bag-level labels are available, which indicate whether a bag contains representatives of a target class or not. The MIL tasks are now to (a) learn to best classify other *bags* (to determine what classes are present within them), and (b) to learn to classify *individual instances*, including those in the training data bags themselves. A number of algorithms have been proposed for MIL including methods based on boosting [2], random forests [27], support-vector machines [1], and neural networks [47]. MIL has been successfully applied to a variety of tasks including image recognition [29], text categorization [23], drug activity prediction [46], and bioinformatics [5].

In the sound-classification framework, the analogue of a bag of instances is a weakly labeled recording. The recording can in turn can be split into short temporal or time-frequency segments, e.g., by uniformly segmenting it into fixed-length sections, for instance, half a second or one second long. These would comprise the individual *instances* in the bags. MIL techniques can now be directly applied. Classifying the individual instances (segments) as belonging or not belonging to the target sound event will naturally also localize the event to within the granularity of the segments.

The earliest reported application of MIL to audio analysis was by Mandel and Ellis [28], who applied it to music. Musical labels are generally applied to artists, albums, or individual tracks. However, a label may not apply to the entirety of an album or a track, e.g., a track tagged as “saxophone” may contain segments that have no sax in them at all. Mandel and Ellis attempted to apply MIL to obtain finer-grained tags from the high-level labels, i.e., to tag individual segments (at 10-s granularity) within the music, and reported being able to do so with reasonable accuracy.

Briggs et al. [6] applied a variant of MIL known as *multiple-instance multiple label* (MIML) learning to identify bird sounds in recordings. In a typical natural recording of bird sounds many different birds can be heard. The labels on training data generally only identify all the birds heard in them, but do not (and often cannot) isolate the individual birds. The authors demonstrated that the MIML solution provides better accuracies than other methods at identifying all birds in a test recording. Their solution was, however, restricted to performing bag-level classification; they did not attempt to isolate individual bird calls in either the test or the training data.

Kumar and Raj [24, 25] reported one of the earliest applications of MIL to the problem of generic sound event detection, and proposed a variety of solutions based on different classifier formalisms including support-vector machines and neural networks. The primary task addressed in their work was that of learning to *detect* and localize sound events from weakly labeled training data. As a byproduct, their solutions also obtained temporal localization of the sound events within the training data itself. They were able to achieve both classification performance and temporal localization of some sound event classes comparable to that achieved with *strongly* labeled training data over a small vocabulary of sound events.

Xu et al. [44] and Kong et al. [22] proposed an alternative to the MIL approach, which treats the problem of learning to classify from weak labels as one of learning to pay *attention* to the right training instances in each bag. Their solution uses a combination of two neural networks, one which determines the importance of each instance in a bag, expressed as a weight assigned to the instance, and a second which attempts to classify it. The bag-level classifier output is a weighted combination obtained by summing the instance-wise multiplication of the outputs of both neural networks over all the instances in the bag. Both networks are jointly trained to minimize bag-level error. They showed that the resulting classifier is not only able to achieve highly accurate recording-level classification of audio, but also able to accurately localize events within both training and test recordings.

All of these proposed solutions have limited scope; their efficacy has only been demonstrated on small datasets and vocabularies. More recently, DCASE has issued a large-scale challenge on learning to classify sound events from weakly labeled data [13] that greatly increases the size, if not the vocabulary of the datasets. The challenge is expected to generate increased interest in the problem of learning from weakly labeled data.

Other outstanding problems include that the weak labels are often noisy. For instance, the AudioSet corpus reports that many of the weak labels in their dataset are inaccurate, with the accuracy of the labels falling below 50% for some categories, even with human annotators. Correia et al. [10] propose MIL solutions for cases where a confidence in the annotation may be established; more generally, however, the problem of training from noisy weak labels remains a challenge. A secondary, associated problem lies with the annotation of *negative* bags. Weak labels generally only indicate the *presence* of target sounds in a recording. The *absence* of sound events in a recording is rarely, if ever, annotated. Thus, the bags used as negative exemplars are only *assumed* to be negative, and are not guaranteed to be so. MIL solutions for noisy labels generally focus on noisy positive labels; the issue of noise in negative labels has been less considered. These are among the issues that must be resolved for truly scalable solutions for training with weakly labeled data.

14.2.4.2 Exploiting Visual Information

Thanks to the proliferation of smart phones, there are now billions of people carrying devices able to make recordings of their everyday experiences in both audio and video modalities; the hundreds of hours of video uploaded every minute to YouTube and similar services present both an important application domain and a rich source of training material for automatic sound scene and event recognition systems. But the fact that these environmental audio recordings also include simultaneous visual information is an opportunity not to be ignored. In particular, given the difficulties in constructing accurate labels for audio recordings, can we glean useful labels from the video channel?

As mentioned above, image recognition systems are already very powerful, so a natural idea is to use existing image classifiers to provide the labels for training sound classifiers. One problem is that the labels provided by the image classifier—the objects visible in the scene, or some global label for the scene depicted—are at best only related to the sound events we would want to detect. At worst, they can be unrelated, either because the sound sources are not in the field of view or perhaps because the video has had an unrelated soundtrack dubbed on. However, the potential for enormous training sets can counterbalance these potential weaknesses in the labels.

This line of thinking was neatly developed by Aytar et al. [3]. They trained an audio classifier to predict the classification of the corresponding visual frame using existing pre-trained visual object and scene classifiers; the internal representation of this classifier was then used to train a simple SVM classifier for audio scene and event recognition tasks, substantially outperforming the best published results which did not have the benefit of the large audio-visual training set they were able to exploit.

14.2.5 Evaluation Tasks

The astonishing progress in image classification bears a substantial debt to the existence of the ImageNet [36] dataset and the associated evaluations. ImageNet provided at least 1000 positive example images for 1000 object categories, giving enough data to support the training of high-performance, deep network classifiers, and a broad enough range of object categories to give a passable attempt at general-purpose recognition.

ImageNet was the inspiration for AudioSet, a collection of manually labeled 10 s excerpts from the soundtracks of YouTube videos, providing at least 100 examples for over 500 sound event categories. While still much smaller than ImageNet, it at least attempts to provide comprehensive coverage of sounds rather than being limited to the small, specialized subsets of sound events that have been used in evaluations to date such as CLEAR [41] and DCASE [30, 40]. A standard evaluation based around AudioSet may similarly emerge as the common standard to push forward sound event detection.

14.3 Future Perspectives

14.3.1 Applications

Audio classification promises powerful applications in "embedded" intelligent devices that can benefit from adapting to unpredictable environments, from smart phones to self-driving cars. One significant recent development in this category is the smart home assistant, pioneered Amazon's Echo [43]. This kind of hands-free smart assistant naturally relies on sound input for control; currently, this is exclusively via speech commands, but it is natural for it to use other information available in the acoustic channel, including the kind of home surveillance applications presented in Chap. 12, and raising all the privacy issues discussed there.

Another promising application area is personal hearing devices. Hearing aids that adapt automatically to their environments have been under development since at least 2005 [7], but the past few years have seen the emergence of intelligent "hearables," presented as augmented earphones that promise features such as automatic removal of unwanted noise while passing through important or desired sounds [12]. Given their extreme constraints of size and power consumption, today's devices merely suggest the kinds of functionality that will become possible as technology improves.

14.3.2 Approaches

The current generation of acoustic recognizers, as typified by the DCASE evaluations, focuses on an explicit set of output categories—either scenes or specific sound events. Despite recent efforts to develop a complete "ontology" of sound events [15], this approach seems doomed since there is an unlimited variety of sounds and subcategories within sounds that might be distinguished. One trend in fields including text and vision analysis is to work with an "embedding space," a moderately sized feature space (e.g., 128 dimensions) where each object or event is mapped to a point such that semantically similar objects are close together [18]. Such a representation is intrinsically continuous, supporting arbitrarily fine distinctions between similar objects. Classification is not required, but if desired it can be accomplished by a simple quantization of the space. The embedding space is conveniently obtained as the activation of an intermediate layer in a neural net, trained by any method ranging from classical supervised training on explicitly labeled examples through to "triplet loss" approaches that require only same/different labels for pairs of examples, leaving implicit the underlying classes [39].

A common problem in trained classifiers is mismatch between training and test data: When tested on data that is systematically different from the examples used for training, performance may be arbitrarily degraded. This kind of mismatch can include things that human listeners subconsciously ignore, for instance, the difference between the same sound source recorded in differently sized rooms (i.e., room acoustics), or mixtures with different background noises. To achieve the goal of human-level robustness at recognizing sound events, we will either need to collect training sets that span all relevant combinations of sources and environments (which becomes exponentially expensive) or devise alternative approaches to achieving this kind of generalization. Work in speech recognition has attempted to identify acoustic features that are relatively invariant to acoustic variations [20], although the alternative solution of collecting speech in very many acoustic conditions has ultimately proven more successful.

The idea of "transfer learning" [33] is aimed at situations where there is substantial out-of-domain training data that can nonetheless contribute to a task. Embedding space representations can be used for this kind of transfer: an embedding trained on one set of sound events—in which data straddles a wide range of recording conditions—can provide an embedding providing some invariance to recording conditions; if the embedding preserves enough source-relevant information to discriminate classes in a new task, then a classifier trained on the embedding representation of a small set of in-domain examples may result in a classifier that "inherits" the invariance from the larger dataset.

Current classifiers achieve robustness to background noise primarily through training on noisy examples, so essentially it is the combined properties of target event and interference that are being recognized. However, as discussed in Chap. 3, human perception appears to analyze complex scenes into distinct representations of individually perceived sources. Such a source separation or "Computational Auditory Scene Analysis" [9, 42] approach is conceptually appealing: an independent process able to divide a complex mixture into multiple, noise-free source sounds (along the lines of the matrix factorization techniques described in Sect. 8.3.3.2) would make the job of a subsequent event recognizer much easier. In practice, however, it is unlikely that such ideal source separation can be achieved without incorporating prior knowledge about source characteristics, so some kind of combined source separation and recognition process (reminiscent of the joint estimation of multiple sources in [14] and [21]) may turn out to be the most successful approach to source separation, and ultimately to robust sound event detection as well.

Although we have considered the classification of sound scenes and sound events as distinct tasks, they are of course related: a sound scene is essentially defined as a particular combination of sound events. Ideally, these two tasks can be unified, with scene classification emerging as a judgment over the set of detected events, although this approach is, for the moment, unlikely to rival global classification applied to the raw features from the scene.

14.4 Summary

We can safely expect high-accuracy automatic sound scene and event recognition in the near future, and it will lead to new and valuable applications in interactive systems and archive management. Sound provides critical information for us as inhabitants of the real world, and our automatic systems must and will take advantage of that information at our behest. The chapters in this book have provided detail on the current state of the art in the technology and applications of environmental sound recognition, and we look forward to the exciting developments that will unfold in the coming years.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 577–584 (2003)
2. Auer, P., Ortner, R.: A boosting approach to multiple instance learning. In: European Conference on Machine Learning, pp. 63–74. Springer, Berlin (2004)
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems, pp. 892–900 (2016)
4. Babenko, B.: Multiple instance learning: algorithms and applications. Technical Report, Department of Computer Science and Engineering, University of California, San Diego (2008)
5. Bandyopadhyay, S., Ghosh, D., Mitra, R., Zhao, Z.: MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci. Rep. **5**, 8004 (2015)
6. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc. Am. **131**(6), 4640–4650 (2012)
7. Büchler, M., Allegro, S., Launer, S., Dillier, N.: Sound classification in hearing aids inspired by auditory scene analysis. EURASIP J. Adv. Signal Process. **2005**(18), 387845 (2005)
8. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. J. Artif. Intell. Res. **4**(1), 129–145 (1996)
9. Cooke, M., Ellis, D.P.: The auditory organization of speech and other sources in listeners and computational models. Speech Commun. **35**(3), 141–177 (2001)
10. Correia, J., Trancoso, I., Raj, B.: Adaptation of SVM for MIL for inferring the polarity of movies and movie reviews. In: Spoken Language Technology Workshop (SLT), 2016 IEEE, pp. 258–264. IEEE, New York (2016)
11. Dalvi, B., Callan, J., Cohen, W.W.: Entity list completion using set expansion techniques. In: Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010). NIST, Gaithersburg MD (2011)
12. Doppler Labs: HearOne wireless smart earbuds (2017). http://hereplus.me
13. Elizalde, B., Raj, B., Vincent, E.: Large-scale weakly supervised sound event detection for smart cars (2017). http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection
14. Frey, B.J., Deng, L., Acero, A., Kristjansson, T.T.: ALGONQUIN: iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In: INTERSPEECH, pp. 901–904 (2001)
15. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: an ontology and human-labeled dataset for audio events. In: IEEE ICASSP 2017, New Orleans (2017). https://research.google.com/pubs/pub45857.html

16. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, vol. 2, pp. 539–545. Association for Computational Linguistics, Stroudsburg, PA (1992)
17. Hershey, S., Chaudhury, S., Ellis, D.P.W., Gemmeke, J., Jansen, A., Moore, R.C., Plakal, M., Sauros, R.A., Seybold, B., Slaney, M., Weiss, R.: CNN architectures for large-scale audio classification. In: IEEE ICASSP 2017, New Orleans (2017). https://research.google.com/pubs/pub45611.html
18. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
19. Jansen, A., Gemmeke, J.F., Ellis, D.P.W., Liu, X., Lawrence, W., Freedman, D.: Large-scale audio event discovery in one million youtube videos. In: IEEE ICASSP 2017, New Orleans (2017)
20. Kingsbury, B.E., Morgan, N., Greenberg, S.: Robust speech recognition using the modulation spectrogram. Speech Commun. **25**(1), 117–132 (1998)
21. Klapuri, A.: Multiple fundamental frequency estimation by summing harmonic amplitudes. In: ISMIR, pp. 216–221 (2006)
22. Kong, Q., Xu, Y., Wang, W., Plumbley, M.D.: A joint detection-classification model for audio tagging of weakly labelled data. CoRR abs/1610.01797 (2016). http://arxiv.org/abs/1610.01797
23. Kotzias, D., Denil, M., De Freitas, N., Smyth, P.: From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 597–606. ACM, New York (2015)
24. Kumar, A., Raj, B.: Audio event detection using weakly labeled data. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 1038–1047. ACM, New York (2016)
25. Kumar, A., Raj, B.: Weakly supervised scalable audio content analysis. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, New York (2016)
26. Kumar, A., Raj, B., Nakashole, N.: Discovering sound concepts and acoustic relations in text. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New York (2017)
27. Leistner, C., Saffari, A., Bischof, H.: Miforests: multiple-instance learning with randomized trees. In: Computer Vision–ECCV 2010, pp. 29–42 (2010)
28. Mandel, M.I., Ellis, D.P.: Multiple-instance learning for music information retrieval. In: ISMIR, pp. 577–582 (2008)
29. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: ICML, vol. 98, pp. 341–349 (1998)
30. Mesaros, A., Heittola, T., Virtanen, T.: Tut database for acoustic scene classification and sound event detection. In: Signal Processing Conference (EUSIPCO), 2016 24th European, pp. 1128–1132. IEEE, New York (2016). http://www.cs.tut.fi/~mesaros/pubs/mesaros_eusipco2016-dcase.pdf
31. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
32. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., Welling, J.: Never-ending learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15) (2015)
33. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
34. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Training object class detectors with click supervision. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii (2017). ArXiv preprint arXiv:1704.06189
35. Pillai, R., Qazi, U.W.: Acoustic analysis of text (aat): Extracting sound out of words. QSIURP Research Report, Carnegie Mellon University Qatar (2016)

36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
37. Sager, S., Borth, D., Elizalde, B., Schulze, C., Raj, B., Lane, I., Dengel, A.: AudioSentiBank: large-scale semantic ontology of acoustic concepts for audio content analysis. arXiv preprint (arXiv:1607.03766) (2016)
38. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1041–1044. ACM, New York (2014). https://serv.cusp.nyu.edu/projects/urbansounddataset/salamon_urbansound_acmmm14.pdf
39. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
40. Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M.D.: Detection and classification of acoustic scenes and events. IEEE Trans. Multimedia **17**(10), 1733–1746 (2015)
41. Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M.: Clear evaluation of acoustic event detection and classification systems. In: International Evaluation Workshop on Classification of Events, Activities and Relationships, pp. 311–322. Springer, New York (2006)
42. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley/IEEE Press, New York (2006)
43. Wikipedia: Amazon Echo (2017). https://en.wikipedia.org/wiki/Amazon_Echo
44. Xu, Y., Kong, Q., Huang, Q., Wang, W., Plumbley, M.D.: Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. CoRR abs/1703.06052 (2017). http://arxiv.org/abs/1703.06052
45. Zhao, S., Heittola, T., Virtanen, T.: Active learning for sound event classification by clustering unlabeled data. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (2017)
46. Zhao, Z., Fu, G., Liu, S., Elokely, K.M., Doerksen, R.J., Chen, Y., Wilkins, D.E.: Drug activity prediction using multiple-instance learning via joint instance and feature selection. BMC Bioinf. **14**(14), S16 (2013)
47. Zhou, Z.H., Zhang, M.L.: Neural networks for multi-instance learning. In: Proceedings of the International Conference on Intelligent Information Technology, Beijing, pp. 455–459 (2002)