AUDIO SET: AN ONTOLOGY AND HUMAN-LABELED DATASET FOR AUDIO EVENTS

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen,
Wade Lawrence, R. Channing Moore, Manoj Plakal, Marvin Ritter

Google, Inc., Mountain View, CA, and New York, NY, USA

{jgemmeke,dpwe,freedmand,arenjansen,wadelawrence,channingmoore,plakal,marvinritter}@google.com

ABSTRACT

Audio event recognition, the human-like ability to identify and relate sounds from audio, is a nascent problem in machine perception. Comparable problems such as object detection in images have reaped enormous benefits from comprehensive datasets – principally ImageNet. This paper describes the creation of *Audio Set*, a large-scale dataset of manually-annotated audio events that endeavors to bridge the gap in data availability between image and audio research. Using a carefully structured hierarchical ontology of 632 audio classes guided by the literature and manual curation, we collect data from human labelers to probe the presence of specific audio classes in 10 second segments of YouTube videos. Segments are proposed for labeling using searches based on metadata, context (e.g., links), and content analysis. The result is a dataset of unprecedented breadth and size that will, we hope, substantially stimulate the development of high-performance audio event recognizers.

Index Terms— Audio event detection, sound ontology, audio databases, data collection

1. INTRODUCTION

Within the domain of machine perception, we are interested in artificial sound understanding. We would like to produce an audio event recognizer that can label hundreds or thousands of different sound events in real-world recordings with a time resolution better than one second – just as human listeners can recognize and relate the sounds they hear.

Recently, there have been astonishing results in the analogous problem of image recognition [1, 2, 3]. These make use of the ImageNet dataset, which provides more than 1 million images labeled with 1000 object categories [4]. ImageNet appears to have been a major factor driving these developments, yet nothing of this scale exists for sound sources.

This paper describes the creation of *Audio Set*, a dataset and ontology of audio events that endeavors to provide comprehensive coverage of real-world sounds at ImageNet-like scale. Our intention is to use the dataset to create automatic audio event recognition systems that are comparable to the state-of-the-art in recognizing objects in real-world images.

Audio event recognition has been studied from perceptual and engineering perspectives. Warren & Verbrugge [5] were among the first to connect perceptual properties to acoustic features in their study of bouncing and breaking sounds. Ballas [6] probed the perception of 41 short events obtained from a sound effects collection by relating identification time to acoustic features, measures of familiarity, and environmental prevalence. Gygi, Kidd & Watson [7] used multidimensional scaling to identify acoustic factors such as temporal envelope and pitch measures that predicted similarity ratings among 50 environmental sounds. LeMaitre and Heller [8] proposed a taxonomy of sound events distinguishing objects and actions, and used identification time and priming effects to show that listeners find a "middle range" of abstraction most natural.

Engineering-oriented taxonomies and datasets began with Gaver [9] who used perceptual factors to guide the design of synthetic sound effects conveying different actions and materials (tapping, scraping, etc.). Nakatani & Okuno [10] devised a sound ontology to support real-world computational auditory scene analysis. Burger et al. [11] developed a set of 42 "noisemes" (by analogy with phonemes) to provide a practical basis for fine-grained manual annotation of 5.6 hours of web video soundtrack. Sharing many of the motivations of this paper, Salamon et al. [12] released a dataset of 18.5 hours of urban sound recordings selected from `freesound.org`, labeled at fine temporal resolution with 10 low-level sound categories chosen from their urban sound taxonomy of 101 categories. Most recently, Säger et al. [13] systematically constructed adjective-noun and verb-noun pairs from tags applied to entire `freesound.org` recordings to construct AudioSentiBank, 1,267 hours of audio labeled with 1,123 adjective-noun or verb-noun "sentiment" tags. The labels' time resolution is limited whole clips, which can be up to 15 min long, and there is no guarantee that the word-pairs actually belong together – e.g., "talking bird" yields mostly tracks containing both "talking" and "bird", but very few birds doing the talking.

Automatic systems for audio event classification go back to the Muscle Fish content-based sound effects retrieval system [14]. EU Project CHIL conducted an Acoustic Event Detection evaluation [15] over examples of 16 "meeting room acoustic events" comparing results of three systems. A similar task was included in the IEEE-sponsored DCASE 2013 [16] which attracted 7 submissions to detect 16 "office" audio events. F-measures (harmonic mean of precision and recall of detected events) were below 0.2, leaving much room to improve. DCASE 2016 [17], includes an audio event detection task based on 90 minutes of real home and outdoor recordings, carefully annotated with 13 audio event categories spanning "bird singing" to "washing dishes".

Unlike previous work, Audio Set considers all sound events rather than a limited domain. We believe that a large-scale task, in terms of both categories and data, will enable more powerful learning techniques, and hence a step-change in system quality.

2. THE AUDIO SET ONTOLOGY

Given the goal of creating a general-purpose audio event recognizer, we need to define the set of events the system should recognize. This set of classes will allow us to collect labeled data for training and evaluation. In turn, human-labeled data provides a crisp definition of our research goal: to create a system able to predict the human

labelings from audio alone.

Rather than a flat list of audio events, we want to have events structured into an abstraction hierarchy. In training, this indicates classes with nonexclusive relationships; for instance we do not want to our classifiers to attempt to separate "dog sounds" from "bark". During recognition, hierarchical relations allow backing-off to more general descriptions when the classifier encounters ambiguity among several subcategories (e.g., a sound that is recognized ambiguously as "growl", "bark" and "howl" can fall back to a classification of "dog sounds"). Finally, a well-structured hierarchy can aid human labeling by allowing a labeler to quickly and directly find the set of terms that best describe a sound; this was also important during the development of the event set when trying to add categories without overlap and duplication.

This structured set of audio event categories is called the Audio Set Ontology. The specific principles that guided its development were:

- The categories should provide a comprehensive set that can be used to describe the audio events encountered in real-world recordings.
- The category of a particular audio event should correspond to the idea or understanding that immediately comes to the mind of a listener hearing the sound.
- Individual categories should be distinguishable by a "typical" listener. That is, if two categories correspond to sounds that a listener cannot easily or reliably distinguish, the categories should be merged. Sounds that can be distinguished only by an expert (such as particular bird species, or fine distinctions in musical instruments) should not be separated. This is a natural condition to prevent the set becoming unwieldy, although we recognize the possibility of extensions that expand certain nodes with more detailed, expert, distinctions.
- Ideally, individual categories are distinct based on their sound alone, i.e. without relying on accompanying visual information or details of context. Thus, a sound is a "thump", rather than "bare foot stamping on a wooden floor".
- The hierarchical structure allows annotators to identify the best, most specific categories for given audio events as easily as possible. This means that the hierarchy should not be too deep, while the number of children in any node should rarely be more than 10, to facilitate rapid scanning.

To avoid biasing the category set by the orientation of a particular researcher, or the limited diversity of sounds in a particular dataset or task, we sought to start from a neutral, large-scale analysis of web text. We seeded our audio event lexicon using a modified form of the "Hearst patterns" [18] to identify hyponyms of "sound", i.e. terms that frequently occur in constructions of the form "... sounds, such as X and Y ..." or "X, Y, and other sounds ...", etc. Applying these rules over web-scale text yields a very large set of terms, which are then sorted in terms of how well they represent sounds - calculated as a combination of overall frequency of occurrence with how exclusively they are identified as hyponyms of "sound" rather than other terms. This gave us a starting list of over 3000 terms; as part of the matching process, they were also resolved against Freebase/Knowledge Graph entity/machine IDs (MIDs) [19], which we use as stable identifiers.

Starting from the top of the sorted list, we manually assembled a hierarchy to contain these terms in a way that best agreed with our intuitive understanding of the sounds. This step is subjective, but it best captures the role of the hierarchy in supporting human labelers. We stopped when the list seemed to be supplying terms that were all obscure or ill-defined (e.g., "Wilhelm scream", "The Oak Ridge Boys", "Earcon", "Whump"). The resulting structure is not a strict hierarchy as nodes may occur in several places; for instance, "Hiss" appears under "Cat", "Steam", and "Onomatopoeia". There are in total 33 categories that appear more than once.

We refined this initial structure by comparing it with existing audio event lists or taxonomies, including [9, 20, 11, 12, 21, 8, 17]. Our hope was that the initial list would subsume these earlier efforts; in fact, numerous gaps were exposed, although eventually we reached a point where nearly every class from other sets was covered. Some classes from other sets were not incorporated because they were too specific, or otherwise did not meet our criteria of being readily identifiable. For instance, the urban sounds taxonomy of [12] includes "Car radio" as a source of recorded music, which makes sense when labeling city street ambience recordings but is too specialized or context-dependent for our set (although we do have "Radio" for sounds that are clearly being produced by a radio). The 165 verified-distinct sound clips used in [21] include detailed examples such "Trumpet jazz solo" and "Walking on leaves"; in our scheme, these would be simply "Trumpet" and "Walk, footsteps", respectively.

We then began using this set to label recordings, as well as trying to find examples for every category (as described in section 2.1 below). Feedback from the larger group of people involved led to further modifications to the category set.

Our final list contains 632 audio event categories, arranged in a hierarchy with a maximum depth of 6 levels; an example from among the eight level-6 nodes is "Sounds of things" → "Vehicle" → "Motor vehicle" → "Emergency vehicle" → "Siren" → "Ambulance (siren)" (a category that was appropriated from the Urban Sound taxonomy [12]). Figure 1 shows the 50 first- and second-level nodes in this hierarchy.

2.1. Ontology Data Release

The ontology is released[1] as a JSON file containing the following fields for each category:

- **ID**: The Knowledge Graph Machine ID (MID) best matching the sound or source, used as the primary identifier for the class. In Knowledge Graph, many MIDs refer to specific objects (e.g., `/m/0gy1t2s`, "Bicycle bell" or `/m/02y_763`, "Sliding door"); when used in the ontology, they are understood to mean "the sound readily associated with the specified object". For instance, in the case of "Sliding door", the sound made by someone crashing into a closed door would be emanating from the door; however, it would not be the characterisic short burst of bearing noise that suggests "Sliding door", so it should instead be labeled "Smash" or "Thump".
- **Display name**: A brief one or two word name for the sound, sometimes with a small number of comma-separated alternatives (e.g., "Burst, pop"), and sometimes with parenthesized disambiguation (e.g. "Fill (with liquid)"). Display names are intended to be unambiguous even when shown without their ancestor nodes.
- **Description**: A longer description, typically one or two sentences, used to provide more explanation of the meaning and limits of the class. In many cases these are based on

[1] `g.co/audioset`

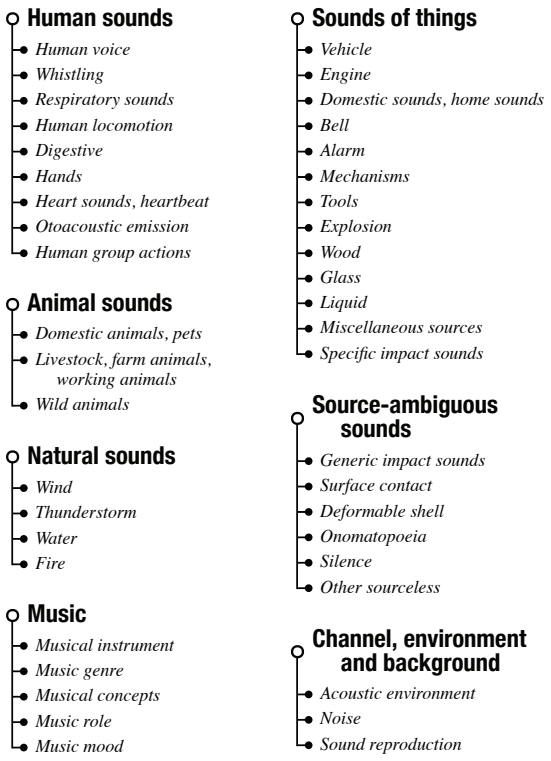

Fig. 1: The top two layers of the Audio Set ontology.

Wikipedia or WordNet descriptions (with appropriate citation URIs), adapted to emphasize their specific use for audio events.

- **Examples**: As an alternative to the textual descriptions, we also collected at least one example of the sound (excepting "Abstract" classes, described below). At present, all examples are provided as URLs indicating short excerpts from public YouTube videos. The exercise of finding concrete examples for each class was helpful in flushing out indistinct or ambiguous categories from earlier versions of the ontology.
- **Children**: The hierarchy is encoded by including within each node the MIDs of all the immediate children of that category.
- **Restrictions**: Of the 632 categories, 56 are "blacklisted", meaning they are not exposed to labelers because they have turned out to be obscure (e.g., "Alto saxophone") or confusing (e.g., "Sounds of things"). Another 22 nodes are marked "Abstract" (e.g., "Onomatopoeia"), meaning that they exist purely as intermediate nodes to help structure the ontology, and are not expected to ever be used directly as labels. These flags appear in the **Restrictions** field.

Figure 2 shows the complete record for a single example category, "Bird vocalization, bird call, bird song".

3. AUDIO SET DATASET

The Audio Set YouTube Corpus consists of labeled YouTube segments, structured as a CSV file[2] comprising YouTube identifiers, start time, end time and one or more labels. Dataset segments are

[2] g.co/audioset

```
{
id: '/m/020bb7',
name: 'Bird vocalization, bird call, bird song',
description: 'Bird sounds that are melodious to
the human ear.',
citation_uri: 'http://en.wikipedia.org/wiki/Bird_
vocalization',
examples: 'youtu.be/vRg6EQm8pBw?start=30&end=40',
children: '/m/07pggtf,/m/07pggtn,/m/07sx8x_',
restrictions: ''
}
```

Fig. 2: The full data record for one sound category from the Audio Set Ontology. The children correspond to "Tweet", "Chirp", and "Squawk".

all 10 seconds long (except when that exceeds the length of the underlying video). Each dataset segment carries one or more ontology class labels.

3.1. Human rating

Human raters were presented with a 10-second segments including both the video and audio components, but did not have access to the title or other meta-information of the underlying YouTube video. If the duration of the video was less than 10 seconds, the entire video was used for rating. We experimented both with shorter segments and audio-only presentation, but raters found these conditions far more difficult, possibly due to the fine-grained nature of our audio event ontology.

For each segment, raters were asked to independently rate the presence of one or more labels. The possible ratings were "present", "not present" and "unsure". Each segment was rated by three raters and a majority vote was required to record an overall rating. For speed, a segment's third rating was not collected if the first two ratings agreed for all labels.

The raters were unanimous in 76.2% of votes. The "unsure" rating was rare, representing only 0.5% of responses, so 2:1 majority votes account for 23.6% of the decisions. Categories that achieved the highest rater agreement include "Christmas music", "Accordion" and "Babbling" (> 0.92); while some categories with low agreement include "Basketball bounce", "Boiling" and "Bicycle" (< 0.17).

Spot checking revealed a small number of labeling errors which were mostly attributed to: 1) confusing labels, 2) human error, and 3) difference in detection of faint/non-salient audio events. As an additional check we analyzed correlations between "present" labels and words in the video's metadata. This exposed some commonly-misinterpreted labels, which were then removed from the ontology. Due to the scale of the data and since majority agreement was very high, no other other corrective actions were taken.

3.2. Selecting segments for rating

In order to obtain a sufficient number of positive ratings from a moderate volume of labeling effort, it is important to send for rating only those that have a good chance of containing particular audio events.

We used a variety of methods to identify videos that were likely to be relevant to particular labels: About half of the audio events corresponded to labels already predicted by an internal video-level automatic annotation system, and we used the videos bearing the label to provide segments for rating. The labels from the internal video labeling system are automatically generated and thus in-

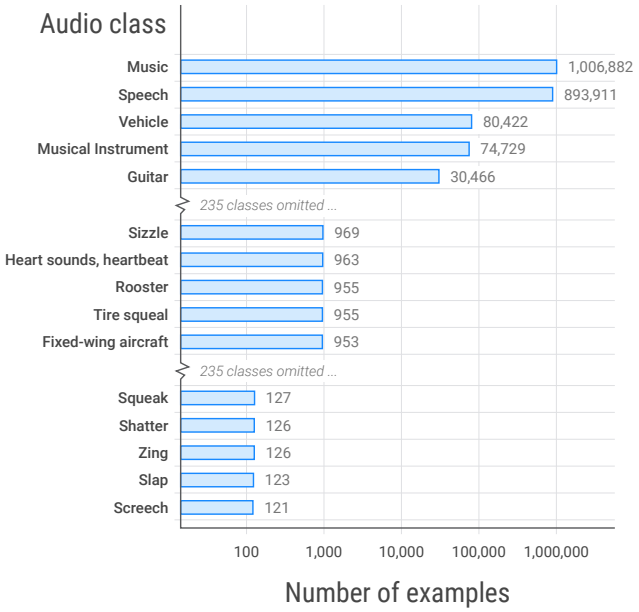

Fig. 3: Histogram of label counts over the entire 1,789,621 segment dataset.

evitably imperfect. They are, however, based on multiple complementary sources of information, such as metadata, anchor text of incoming links, comments, and user engagement signals, and their reliability is greatest for popular videos which have more of these attributes. We therefore select only videos with at least 1,000 views.

In addition, videos were selected by searching titles and other metadata for relevant keywords. We used a ranking-based approach where videos were scored by matching not only against the display name of the audio event, but also, with decreasing weight, its parents in the ontology. Including the term "sound" as a root node in these searches improved the precision of results. Temporal selection of segments with the returned videos was arbitrary, typically taking segments with a start time at 30 sec to avoid any introduction or channel branding at the start of the video. For some of the data, temporal selection was based on content analysis such as nearest-neighbour search between the manually-verified clips clips and the label-specific candidate videos.

Using all these techniques, there were some categories for which we were unable to find enough positive examples to fully populate the dataset, but this proportion is now very small (and shrinking).

In general we found that segments nominated by the automatic video annotation system performed the best (49% of segments rated present, versus 41% for the metadata-based approach), with the caveat that not all audio classes are included. For audio classes not included in the automatic annotations, 36% of the metadata-based segments were rated present. The effectiveness of the content-similarity approaches varied widely depending on the quality of the example and the candidate set of videos.

Regardless of the mechanism by which a segment was nominated, we gathered ratings for all labels associated with the segment by any method. Additionally, we always collected ratings for "Speech" and "Music" (excepting a few experimental runs).

3.3. Dataset construction and characteristics

The released dataset constitutes a subset of the collected material: Only "present" ratings are represented in the release. Rating was conducted with an effort to maximize balance across audio event labels. However, since segments can be labeled with multiple audio events (including the always-rated "Speech" and "Music"), certain labels appear much more frequently. A second objective was to avoid drawing more than one segment from any given video, to avoid correlation of examples.

These objectives were achieved by iteratively adding segments for the least-represented class (for which further examples are available). Out of the set of candidate segments for this audio event class, preference is given to segments bearing the greatest number of labels. We also provide maximally-balanced train and test subsets (from disjoint videos), chosen to provide at least 50 positive examples (in both subsets) for as many classes as possible. These sets were constructed by first collecting examples for the rarest classes, then moving on to less-rare classes and adding more segments only where they had not already passed the threshold of 50. Even so, very common labels such as "Music" ended up with more than 5000 labels.

The resulting dataset includes 1,789,621 segments (4,971 hours), comprising at least 100 instances for 485 audio event categories. The remaining categories are either excluded (blacklisted / abstract as described in section 2.1), or difficult to find using our current approaches. We will continue to develop methods for proposing segments for rating, and aim eventually to cover all non-excluded classes. The unbalanced train set contains 1,771,873 segments and the evaluation set contains 17,748. Because single segments can have multiple labels (on average 2.7 labels per segment), the overall count of labels is not uniform, and is distributed as shown in Fig. 3. "Music" is particularly common, present in 56% of the segments.

4. BENCHMARK

To give a sense of the performance possible with this data, we have trained a simple baseline system. Using the embedding layer representation of a deep-network classifier trained on a large set of generic video topic labels [22], we used the training portion of the Audio Set YouTube Corpus to train a shallow fully-connected neural network classifier for the 485 categories in the released segments. We evaluated on the test partition by applying the classifier to 1 sec frames taken from each segment, averaging the scores, then for each category ranking all segments by their scores. This system gave a balanced mean Average Precision across the 485 categories of 0.314, and an average AUC of 0.959 (corresponding to a d-prime class separation of 2.452). The category with the best AP was "Music" with AP / AUC / d-prime of 0.896 / 0.951 / 2.34 (reflecting its high prior); the worst AP was for "Rattle" with 0.020 / 0.796 / 1.168.

5. CONCLUSION

We have introduced the *Audio Set* dataset of generic audio events, comprising an ontology of 632 audio event categories and a collection of 1,789,621 labeled 10 sec excerpts from YouTube videos. The ontology is hierarchically structured with the goal of covering all acoustic distinctions made by a 'typical' listener. We are releasing this data to accelerate research in the area of acoustic event detection, just as ImageNet has driven research in image understanding. In the future, we hope to be able to make larger and improved releases, ideally including contributions from the wider community.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[5] William H Warren and Robert R Verbrugge, "Auditory perception of breaking and bouncing events: a case study in ecological acoustics," *Journal of Experimental Psychology: Human perception and performance*, vol. 10, no. 5, pp. 704, 1984.

[6] James A Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of experimental psychology: human perception and performance*, vol. 19, no. 2, pp. 250, 1993.

[7] Brian Gygi, Gary R Kidd, and Charles S Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[8] Guillaume Lemaitre and Laurie M. Heller, "Evidence for a basic level in a taxonomy of everyday action sounds," *Experimental Brain Research*, vol. 226, pp. 253–264, 2013.

[9] William W Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[10] Tohomiro Nakatani and Hiroshi G Okuno, "Sound ontology for computational auditory scene analysis," in *Proceeding for the 1998 conference of the American Association for Artificial Intelligence*, 1998.

[11] Susanne Burger, Qin Jin, Peter F Schulam, and Florian Metze, "Noisemes: Manual annotation of environmental noise in audio streams," Tech. Rep. CMU-LTI-12-07, Carnegie Mellon University, 2012.

[12] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[13] Sebastian Sager, Damian Borth, Benjamin Elizalde, Christian Schulze, Bhiksha Raj, Ian Lane, and Andreas Dengel, "Audiosentibank: Large-scale semantic ontology of acoustic concepts for audio content analysis," *arXiv preprint arXiv:1607.03766*, 2016.

[14] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[15] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo, "Clear evaluation of acoustic event detection and classification systems," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 311–322.

[16] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[17] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016, `http://www.cs.tut.fi/sgn/arg/dcase2016/`.

[18] Marti A Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992, pp. 539–545.

[19] Amit Singhal, "Introducing the knowledge graph: things, not strings," 2012, Official Google blog, `https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`.

[20] George A Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[21] Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott, "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition," *Neuron*, vol. 88, no. 6, pp. 1281–1296, 2015.

[22] Shawn Hershey, Sourish Chaudhury, Daniel P. W. Ellis, Jort Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Rif A. Saurous, Brian Seybold, Malcolm Slaney, and Ron Weiss, "CNN architectures for large-scale audio classification," in *IEEE ICASSP 2017*, New Orleans, 2017.