NELS - Never-Ending Learner of Sounds

Benjamin Elizalde *†, **Rohan Badlani** *‡, **Ankit Shah** *◇, **Anurag Kumar**◇, **Bhiksha Raj**◇*
◇LTI, †ECE, Carnegie Mellon University, Pittsburgh, PA
‡Department of Computer Science, BITS Pilani, India
bmartin1,aps1,alnu@andrew.cmu.edu, rohan.badlani@gmail.com

Abstract

Sounds are essential to how humans perceive and interact with the world. These sounds are captured in recordings and shared on the Internet on a minute-by-minute basis. These recordings, which are predominantly videos, constitute the largest archive of sounds we've ever seen. However, most of these recordings have undescribed content making necessary methods for automatic audio content analysis, indexing and retrieval. These methods have to address multiple challenges, such as the relation between sounds and language, numerous and diverse sound classes, and large-scale evaluation. We propose a system that continuously learns from the web relations between sounds and language, improves sound recognition models over time and evaluates its learning competency in the large-scale without references. We introduce the Never-Ending Learner of Sounds (NELS), a project for continuously learning of sounds and their associated knowledge, available on line in **nels.cs.cmu.edu**.

1 Introduction and Related Work

The ability to automatically recognize sounds is essential in a large number of applications such as identifying emergencies in elderly care and patients in hospitals (choking, falling down) [1, 2], where cameras are unwanted due to privacy concerns [3], allowing self-driving cars to respond safely to warning sounds and emergency vehicles (ambulance siren) [4], improving airport and house surveillance, where any number of unusual phenomena have acoustic signatures (alarm, footsteps, glass breaking) [5], expanding our interaction with digital assistants through non-verbal communication (clapping, laughing), and analyzing and retrieving video by its content, perhaps the most explored application [6, 7, 8, 9, 10]. By year 2021, every second a million minutes of videos will be uploaded to the Internet; this will constitute 82% of all consumer traffic [2]. The ability to recognize sounds in all these recordings is key to organizing, understanding, and exploiting the rapidly growing audio and multimedia data.

In recent years, sound recognition research has focused on curated data and guidelines and although successful and necessary, the literature has under explored the challenges of web audio. Curated audio recordings [11, 12, 4, 13] have carefully collected and recorded sound classes with limited set of samples, have manually annotated descriptions of their content, and have defined task-oriented sets of sound classes. In contrast, web audio is recorded in an unstructured manner with an every-day growing number of samples, has descriptions that are insufficient, unavailable or wrong, and has an unbounded number of sound classes. Hence, we should not only test how state of the art sound recognition performs in the web context, but also explore new paradigms to continuously learn from the web audio.

---

*All authors contributed equally.

[2]https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

A never-ending learning system learns many types of knowledge from years of diverse sources, using previously learned knowledge to improve subsequent learning and with sufficient self-reflection to avoid learning stagnation, as pointed in [14]. This contrasts with non continuous systems, which learn from a finite curated source, their learning is limited to the scope of the source and the optimization objective and do not improve learning over time. The never-ending paradigm has been employed in ongoing projects: Never-Ending Language Learner [14] and Never-Ending Image Learner [15]. This paradigm can address challenges in sound learning and to this day it is unexplored. For example, learn associations between sounds and language (metadata, ontologies, descriptive terms); increase acoustic vocabularies and improve robustness of sound recognizers; and evaluate the subjectivity of sound recognition in the absence of prior knowledge of the source or generation process.

We introduce the Never-Ending Learner of Sounds (NELS), a project for large-scale continuous learning of sounds and their associated knowledge. Examples of associated knowledge are semantics related to objects, events, actions, places [16], cities [17, 18] or qualities [19, 20]. Our contribution begins with a working framework that serves for audio content indexing, for evaluation of our sound recognition on web audio and for searching the indexed sounds. Since its inception in 2016, NELS has reported in several research publications discussed in Section 3 and has won the 2017 Gandhian Young Technological Innovation (GYTI) award in India.

2 NELS Framework

In its current form, NELS continuously (24/7) creates a content-based index for web audio. It crawls YouTube videos and analyzes their text and audio content for over 600 sound classes. The sound recognizers were trained from a variety of sources, including web audio itself. NELS also evaluates the quality of the recognition through human feedback. The audio content is indexed combining the crawled metadata, sound recognition predictions and human feedback. As of October 20, 2017, we have crawled over 300 hours of audio corresponding to 4 million video segments of 2.3 seconds. The indexed audio content is available for search and retrieval using our engine in the website.
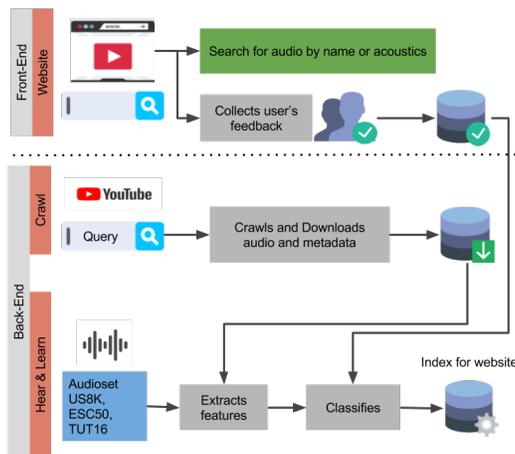

Figure 1: The framework serves as a continuously audio content indexer, a sound recognition evaluation and a search engine for the indexed audio.

2.1 Crawl

In this module, a search query is used to crawl audio and metadata from YouTube videos. The query corresponds to 605 sound event labels from four different datasets. The video metadata is extracted using the Pafy API[3] and corresponds to 12 attributes, such as title, URL and description. The crawled metadata is use to index the audio content.

Although NELS will eventually feed from different sound web archives, we selected YouTube as our first source due to its diversity of sounds and the available metadata associated to them. In contrast to audio-only recordings, collecting audio from videos poses several challenges. YouTube contains massive amount of videos and a proper formulation of the search query is necessary to filter videos with higher chances of containing the desired sound event. Typing a query composed by a noun

---

[3]https://pypi.python.org/pypi/pafy

such as *air conditioner* will not necessarily fetch a video containing such sound event because the associated metadata often corresponds to the visual content; contrary to audio-only archives such as *freesounds.org*. Therefore, we modified the query to be a combination of keywords: "<sound event label> sound", for example,"air conditioner sound". Although the results empirically improved, the sound event was not always found to be occurring and even if it was, sometimes it was present within a short duration, overlapping with other sounds and with low volume. We discarded videos longer than ten minutes and shorter than two seconds because they were either likely to contain unrelated sounds or were too short to be processed.

2.2 Hear & Learn

In this module, we used 605 sound events from four annotated datasets to train classifiers and run them on the crawled YouTube video segments, which are unlabeled. The class predictions are also used to index the audio content.

The framework is being developed so that given a set of guidelines, new datasets could be added seamlessly. NELS should be able to take advantage of existing curated annotations, however dealing with mismatch conditions. The current four datasets are: ESC50, US8k, TUT16 (Task-3) and AudioSet. *ESC-50* [21] has 50 classes from five broad categories: animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds and exterior sounds. The dataset consists of 2,000 audio segments with an average duration of 5 seconds each. *The US8K or UrbanSounds8K* [13] has 10 classes like *gun shot, jackhammer, children playing*. The dataset consists of 8,732 audio segments with an average duration of 3.5 seconds each. *TUT 2016 (Task-3)* [22] has 18 classes like *car passing by, bird singing, door banging* from two major sound contexts namely home context and residential area. The dataset consists of 954 audio segments with an average duration of 5 seconds each. *AudioSet* [23] has 527 classes and 2.1 million audio segments with an average duration of 10 seconds each.

The audio from both the datasets and crawled video segments are preprocessed and classified. NELS is meant to be classifier agnostic. We currently follow Convolutional Neural Networks (CNNs) classifier setup described in [21]. Recordings are segmented into 2.3 seconds and converted into 16-bit encoding, mono-channel, and 44.1 kHz sampling rate WAV files as in [21]. Then, we extracted features comprising of log-scaled mel-spectrograms with 60 mel-bands with a window size of 1024 (23 ms) and hop size is 512. Lastly, the features are used to train multi-class classifiers using CNNs for each of the datasets.

2.3 Website

The module is on line in **nels.cs.cmu.edu** and currently serves for two purposes. First, to evaluate sound recognition using human feedback, which we include as part of the audio indexing. Second, to provide a search engine of the indexed audio content. The goal is to eventually be able to search for audio based on descriptive (subjective) terms, onomatopoeias and acoustic content [24].

The website provides a search field to capture a term (text-query) for sound searching. The term is mapped to the closest of our sound classes. The mapping uses the tool *word2vec* [25] and a precomputed vocabulary of 400 thousand words called *Glove* [26]. Word2Vec computes vector representations of the vocabulary words and the text-query. Then, computes cosine similarity between the text-query, the precomputed vocabulary and our list of sounds. Given that our list of sound classes does not necessarily match with all the words of the precomputed vocabulary, we only consider words within a similarity threshold of 0.15, else no results will be retrieved. Additionally, we provide another feature on the website, the user can paste a YouTube video link on a second text field and NELS will yield the dominant sound.

Each displayed video segment resulting from the text-based search can be evaluated by the user with two button-options, *Correct* or *Incorrect*. That is, whether the human claims that the system's predicted class was present within the segment or not. Examples can be seen in Figure 2.

3 Discussion

In this section we discuss three sound learning challenges where NELS was involved.

Relation between sounds and language. Language can describe audio content, be used to search for sounds, and help to define sound vocabularies [27]. However, the relation between sounds and language is still an inchoate topic. To better understand the usage of language for our indexing, we carried two studies.

Figure 2: Examples of indexed video segments using NELS. First, shows an example where for a *can-opening* sound, the title and images are clearly related. Second, shows an example where a *siren wailing* sound and title are related, but not the visual sound source, which is a child rather than an electronic device. Third, shows an example of *pig-oink* sound, which matches with the visuals, but not with the title and text metadata. Fourth, shows the thumbnail of a video that was indexed, but eventually deleted by the user.

First, sound recognition results [11, 22] evidenced how although two sound events: *quiet street* and *busy street* defined audio from streets, the qualifier implied differences in the acoustic content. These kind of nuances can be described with Adjective-Noun Pairs (ANPs) and Verb-Noun Pairs (VNPS) [19]. We collected one thousand pair-labels derived from different audio ontologies. The audio recordings containing the sound events and their pair-labels were crawled from the collective archive *freesounds.org*. We concluded that despite of the subjectivity of the labels, there is a degree of consistency between sound events and both types of pairs.

Second, in [20] we wanted to identify text phrases which contain a notion of audibility and can be termed as a sound events. We 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. Hence, by matching the combination "sound(s) of <Y>" where Y is any phrase of up to four words to identify candidate phrases, followed by the application of a rule-based classifier to eliminate noisy candidates, we obtained a list of over 100,000 sound labels. 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, we were also able to discover ontological relations. For example, forests may be associated with the sounds of *birds singing*, *breaking twigs*, *cooing* and *falling water*.

Continuous semi-supervised learning of sounds.   NELS should take advantage of existing curated sound datasets and non curated web audio to improve its learning. Previously, semi-supervised self-training approaches have been used to improve sound event recognizers [28, 29]. In [29], we used an earlier version of NELS. Its classifiers were trained and tested using the US8K dataset consisting of about 8,000 labeled samples for 10 sounds. For re-training, we used 200,000 unlabeled YouTube video segments. Similarly to the first paper, but with mismatched data, we achieved about 1.4% overall precision improvement. Regardless of the improvement, we quickly reached a learning plateau. This could be due to mismatch conditions between training and self-training audio. The initial class bias introduced by the hand-crafted dataset. The use of ambiguous YouTube audio to self-train sound classes. Hence, to learn from the daily growing source of web audio, further exploration is needed.

Evaluation of the learning quality.   NELS indexes audio content 24/7, but these segments are unlabeled or have weak or wrong descriptions. Therefore, it is a priority to find methods for automatic evaluation of quality in the large-scale. A solution is to include human intervention [30]. Hence, our website allows collection of human feedback to asses consistency and correctness of sound event indexing. Nevertheless, human feedback may be slow or costly, hence it is important to combine it with other methods that estimate performance in the large-scale.

In [31], we used an earlier version of NELS with a recognizer trained on 78 sound events from three different datasets. After, we crawled audio from YouTube videos using the sound event labels from the datasets as the search query. The query was a combination of keywords: "<*sound event label*> sound", for example,"air conditioner sound". Then, we evaluated the highest-40 recognized segments per sound class based on both types of references (ground truth), human feedback and search query. The search query is a summary of the video's metadata describing the video, but it was interesting to know to what extent it holds at the video's segment level. Results showed how the performance trends of using both types of references are similar and relatively close with less than an absolute 10% difference of precision. This trend suggests that the query could be used as a lower-bound of human inspection. In other words, it could serve as a preliminary reference to evaluate sound recognition. Further exploration on this and other associated metadata and multimedia cues could be used as alternative measurements.

4

References

[1] Patrice Guyot, Julien Pinquier, Xavier Valero, and Francesc Alias, "Two-step detection of water sound events for the diagnostic and monitoring of dementia," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013.

[2] BM Rocha, L Mendes, I Chouvarda, P Carvalho, and RP Paiva, "Detection of cough and adventitious respiratory sounds in audio recordings by internal sound analysis," in *Precision Medicine Powered by pHealth and Connected Health*, pp. 51–55. Springer, 2018.

[3] Patrice Caire, Assaad Moawad, Vasilis Efthymiou, Antonis Bikakis, and Yves Le Traon, "Privacy challenges in ambient intelligence systems," *Journal of Ambient Intelligence and Smart Environments*, vol. 8, no. 6, pp. 619–644, 2016.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[5] Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli, "Audio based event detection for multimedia surveillance," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 5.

[6] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching.," in *TRECVID*, 2010, vol. 2, pp. 3–2.

[7] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander Hauptmann, "Double fusion for multimedia event detection," *Advances in Multimedia Modeling*, pp. 173–185, 2012.

[8] Hui Cheng, Jingen Liu, Saad Ali, Omar Javed, Qian Yu, Amir Tamrakar, Ajay Divakaran, Harpreet S Sawhney, R Manmatha, James Allan, et al., "Sri-sarnoff aurora system at trecvid 2012: Multimedia event detection and recounting," in *Proceedings of TRECVID*, 2012.

[9] Peter Schäuble, *Multimedia information retrieval: content-based information retrieval from large text and audio databases*, vol. 397, Springer Science & Business Media, 2012.

[10] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[11] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," in *2013 IEEE WASPAA*. IEEE, 2013, pp. 1–4.

[12] Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Mark D. Plumbley, Peter Foster, Emmanouil Benetos, and Mathieu Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, Tampere University of Technology. Department of Signal Processing, 2016.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22st ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014.

[14] Tom M Mitchell, William W Cohen, Estevam R Hruschka Jr, Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, et al., "Never ending learning.," in *AAAI*, 2015, pp. 2302–2310.

[15] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta, "Neil: Extracting visual knowledge from web data," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[16] Richard F Lyon, "Machine hearing: An emerging field [exploratory dsp]," *Ieee signal processing magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[17] Anurag Kumar and Bhiksha Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[18] Benjamin Elizalde, Guan-Lin Chao, Ming Zeng, and Ian Lane, "City-identification of flickr videos using semantic acoustic features," in *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 303–306.

[19] 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.

[20] Anurag Kumar, Bhiksha Raj, and Ndapandula Nakashole, "Discovering sound concepts and acoustic relations in text," *arXiv preprint arXiv:1609.07384*, 2016.

[21] Karol J Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[22] 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.

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[24] 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.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[27] Dan Ellis, Tuomas Virtanen, Mark D. Plumbley, and Bhiksha Raj, *Future Perspective*, Springer International Publishing, 2018.

[28] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Björn Schuller, Xiaojie Yu, and Xuan Zhu, "Semi-supervised active learning for sound classification in hybrid learning environments," *PloS one*, vol. 11, no. 9, pp. e0162075, 2016.

[29] Ankit Shah, Rohan Badlani, Anurag Kumar, Benjamin Elizalde, and Bhiksha Raj, "An approach for self-training audio event detectors using web data," *arXiv preprint arXiv:1609.06026*, 2016.

[30] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA*, 2017.

[31] Rohan Badlani, Ankit Shah, Benjamin Elizalde, Anurag Kumar, and Bhiksha Raj, "Framework for evaluation of sound event detection in web videos," in *In submission to IEEE ICASSP 2018*, 2018.