

A DATASET FOR AUDIO-VISUAL SOUND EVENT DETECTION IN MOVIES

Rajat Hebbal, Digbalay Bose, Krishna Somandepalli, Veena Vijai, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California

ABSTRACT

Audio event detection is a widely studied audio processing task, with applications ranging from self-driving cars to healthcare. In-the-wild datasets such as Audioset have propelled research in this field. However, many efforts typically involve manual annotation and verification, which is expensive to perform at scale. Movies depict various real-life and fictional scenarios which makes them a rich resource for mining a wide-range of audio events. In this work, we present a dataset of audio events called Subtitle-Aligned Movie Sounds (SAM-S). We use publicly-available closed-caption transcripts to automatically mine over 110K audio events from 430 movies. We identify three dimensions to categorize audio events: *sound*, *source*, *quality*, and present the steps involved to produce a final taxonomy of 245 sounds. We discuss the choices involved in generating the taxonomy, and also highlight the human-centered nature of sounds in our dataset. We establish a baseline performance for audio-only sound classification of 34.76% mean average precision, and show that incorporating visual information can further improve the performance by about 5%. Data and code are made available for research at <https://github.com/usc-sail/mica-movie-audio-events>

Index Terms— Audio Event Detection, Movies, Audio Recognition, Audio Visual Dataset

1. INTRODUCTION

Audio events are naturally occurring non-verbal sounds produced by humans/objects. Robust detection of such audio events can reveal information about one’s acoustic environment, their psychological state, and help automate rich transcription of multimedia data. Audio event detection (AED) is used in a wide range of domains, including context-aware smart device applications such as in smartphones [1], smart-speakers [2] and self-driving cars [3, 4], acoustic monitoring for health and well-being applications [5, 6] as well as large-scale multimedia indexing [7, 8]. Recently, the introduction of large-scale “in-the-wild” datasets such as Audioset [7] and VG-GSound [9] has enabled prolific AED research. Neural-network representations learned over Audioset have been used for several audio classification tasks such as emotion recognition [10], gender identification [11] and music classification [10, 12]. However, curating such datasets usually involves manual intervention at multiple stages – during data collection, labeling and taxonomy generation. Furthermore, data collected from YouTube sources are subject to attrition due to videos being taken down or made private.

As *sound effects*, audio events form an integral component of the movie audio stream. Deliberate placement of sounds and background-score in a movie scene helps construct a rich narrative and elicit the intended emotional response from viewers. While a large fraction of sound-effects in movies are naturally produced (human/animal vocalizations, music, etc.), some sounds, known as “Foley sounds” [13] are added in post-production. Foley involves



Fig. 1. Closed caption showing audio events occurring off-screen in a movie

the use of ‘everyday’ objects to create sound effects that imitate naturally occurring audio events in different ambient environments; be it the use of snapping *celery* to imitate the sound of breaking bones, or popping the bottom of *trashcans* to amplify the sound of heartbeats¹. Foley is an effective tool that enables simple and inexpensive reproduction of such sounds. It allows for the possibility of audio events that may not be commonly found in the aforementioned data sources (e.g., vehicles crashing, light footsteps, gunshots). Furthermore, it was shown that Foley sounds are nearly indistinguishable from their naturally produced counterparts [14], which makes it useful for developing AED models.

Closed-captions (CC) are time-aligned transcriptions of character dialogues and sound effects. These captions are mandated for several broadcast media, including movies and TV-shows, in an effort to make media more accessible to the hearing-impaired and non-native speakers. Following the guidelines provided by the Described and Captioned Media Program (DCMP) [15], audio captioners are expected to label audio events that are deemed relevant to the plot of the movie/TV-show. Therefore, existing captions can be used to obtain audio-events from movie data in a precise manner.

The contributions of our work are three-fold:

1. We use simple and scalable methods to automatically extract audio captions from movies and categorize an audio event along the dimensions of *sound*, *source* and *quality*.
2. We propose a flat taxonomy of sounds, decoupled from their sources. Unlike previous AED taxonomies, this enables us to group together acoustically similar sounds from different sources.
3. We leverage visual-cues using early multimodal-fusion of audio and video features in a transformer setup to establish baseline audio event detection performance on our dataset.

The rest of the paper is organized as follows: Section 2 discusses existing resources and methodology. In Sec 3, we outline the taxonomy generation process using subtitle tags and compare with audioset taxonomy. In Sec. 4, we describe the methods used to develop baseline AED models on our dataset.

¹blog.storyblocks.com/inspiration/foley-sfx-everyday-household-objects

2. RELATED WORK

Data Resources: Audioset [7], one of the first large-scale AED datasets, includes over 2-million clips with weakly-tagged audio-event classes. The audioset ontology, is the most comprehensive taxonomy of audio-events, comprising 527 different audio-events in a hierarchical structure based on the source of an audio-event. Human-raters were used to label the Audioset data at clip-level.

In order to reduce the manual effort involved in labeling, VGGSound [9] dataset was proposed, which used a scalable pipeline of mining *visually-grounded* audio events from YouTube. Existing machine learning models were used to automatically verify presence of visual signature, and to reject possible false-positive audio classes during data curation. However, a shortcoming is that such methods still do not guarantee occurrence of a tagged sound-event, allowing for some label-noise in exchange for reduced manual effort. Furthermore, it is often the case that an audio-event is heard but not shown on screen, a scenario that is not covered by the VGGSound dataset.

FSD50K [16] consists of over 50K audio events collected in-the-wild, which are annotated across 200 audioset classes on the freesound platform [17]. Apart from these, several smaller-scale AED datasets exist such as Mivia [18], DESED [19], UrbanSound8k [20]. These are typically targeted toward a specific subset of sounds such as indoor, outdoor and rare audio events.

Table 1. Details of different audio event detection datasets. Here, SL refers to whether the audio event labels are precise or are weakly-labeled, PA refers to whether the dataset is publicly available or not; FSD - Freesound platform.

| Dataset | Domain | Clips | Classes | SL | PA | Annotation |
|-------------------|-----------|-------|---------|----|----------------|----------------|
| Mivia [18] | Synthetic | 6K | 3 | ✓ | ✓ | Manual |
| UrbanSound8k [20] | FSD | 8.7K | 10 | ✓ | ✓ | Manual |
| DESED [19] | FSD | 12K | 10 | ✓ | ✓ | Manual |
| FSD50K [16] | FSD | 50K | 200 | ✗ | ✓ | Manual |
| Audioset [7] | Youtube | 2.1M | 527 | ✗ | ✗ ² | Manual |
| VGGSound [9] | Youtube | 200K | 309 | ✗ | ✗ | Automatic |
| SAM-S | Movies | 110K | 191 | ✓ | ✓ | Semi-automatic |

AED: Until recently, convolutional (CNN) models have been used widely for AED. A light-weight version of VGG-16, called *VGGish* [21] was the first benchmark model for AED on Audioset. Several commonly used CNN architectures were compared on Audioset [10], and it was shown that light-weight CNNs can obtain comparable performance to their larger counterparts. More recently, transformer architectures such as VATT [22] and AST [23] have shown state-of-the-art (SoA) performance for audio-only AED. VATT used a self-supervised contrastive loss to pretrain general multimodal representations, which were then finetuned for AED. AST incorporated state-of-the-art vision transformers [24] in AED using spectrogram features.

There have been a few audio-visual methods proposed for audio event detection and localization. Mid-level attention based fusion was used on audio-visual streams with CNN backbones [25]. An optimal multimodal fusion mechanism, called *gradient-blending* [26], was proposed to address variable overfitting rates across modalities. More recently, attention-bottlenecks in multimodal transformer architectures have been proposed [27], showing SoA AV-performance on Audioset using early-fusion. Cross-modal attention mechanisms have been used for audio-visual localization of event sources, and weakly-supervised detection [28, 29, 30].

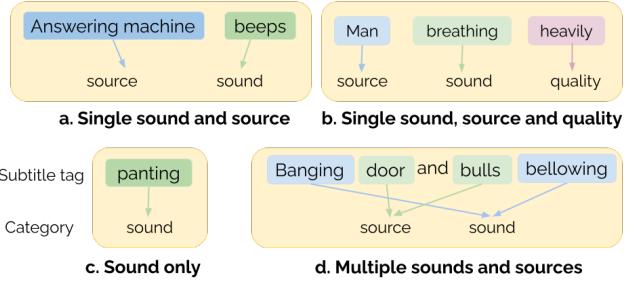


Fig. 2. Annotation examples for sound, source and quality categorization in movie audio events

In the context of multimedia, there have been limited works that have analysed audio events. Gunshot and explosion classification was studied based on dictionary learning from MFCC features [31]. Audio event change detection was explored via clustering methods in a set of 8 movies [32].

3. SUBTITLE ALIGNED MOVIE SOUNDS (SAM-S)

The SAM-S dataset we introduce in this work comprises 430 top-grossing Hollywood films from the years 2014 to 2018. In order to mine audio events, subtitle files for each of the movies were obtained automatically³. Closed-captioned subtitles extracted in this manner contain time-aligned character dialogues and plot-relevant event tags. It is important to note here that closed-captions are not exhaustive in labeling all audio events that occur in a movie, i.e., the tagging process has low-recall. However, the tagged captions are accurate in terms of the labeled sound, i.e., high-precision. This precision-recall trade-off means that while we lose potentially useful data, we ensure minimal additional human effort for annotation and cleanup, and a large-enough set of sounds to develop and evaluate AED models.

These tags are typically enclosed in braces. We automatically extract these tags and the associated time-stamp from the subtitle file. In total, we obtain just over 116K subtitle tags, of which 20,817 are unique. These subtitle tags are descriptive in nature, and include information about the sound, source and quality of the audio event occurring (see Figure 2). While information about sound is always present, often the source and quality of the audio event are not tagged. In fact, following DCMP guidelines, captioners are expected to label the source of the sound with the exception of the instances where the source is clearly visible on-screen. Out of the 21K unique tags, 1.5K are unigrams - usually indicating only the sound, 11K are bigrams - which include the sound and source, and the rest are n-grams, $n \geq 3$, which could refer to the quality of the audio event or multiple simultaneously occurring events. Due to the presence of source-ambiguous audio events in our data, we chose to adopt a flat taxonomy as opposed to a hierarchical one as in [7], which we discuss in more detail in Sec. 3.2.

3.1. Taxonomy generation

The following steps outline the procedure to label, refine and condense our final taxonomy: **Categorization:** We conducted an anno-

²Youtube policy: <https://www.youtube.com/static?gl=US&template=terms>

³<https://github.com/ruediger/VobSub2SRT>

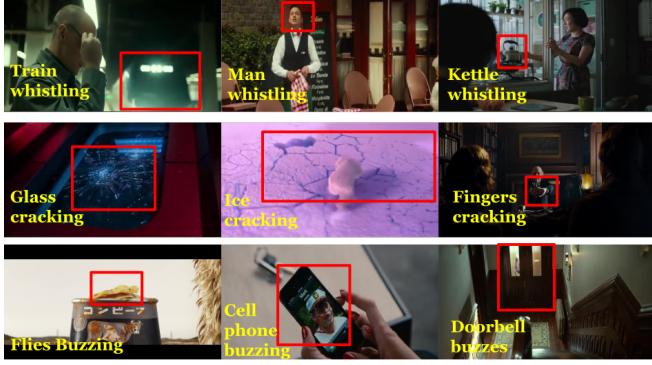


Fig. 3. Examples of sounds originating from multiple sources in movies: (Top) "Whistling" sound from different sources in SAM-S, (Middle) "Cracking", (Bottom) "Buzzing"

tation task using Mechanical Turk in order to categorize a given subtitle tag into ‘source’, ‘sound’ and ‘quality’ classes. A few sample examples of the annotation task is provided in Figure 2. Annotators were explicitly asked to only sample from the words in the tag, and not introduce/interpolate from context (f.e, a subtitle tag of “flickering”, would only have a sound of “flickering” and no source). Three annotations were used for each tag, and majority voting was used for each category. All ties were resolved by an author. For this task, we chose the set of subtitle tags that occur at least 5-times in the dataset. This set of 2161 tags covers 80% (~91K) of the audio-visual events that occur in the dataset.

Lemmatization: For each of the categories, the set of annotations obtained were lemmatized using an open-source NLP-toolkit - spaCy⁴. The lemmatization process was manually verified and errors were corrected by an author. Following this, a total of 254 sources, 254 sounds and 115 qualities form the initial taxonomy of our dataset. **Automatic tagging:** Next, we created a dictionary mapping the original words in the subtitle tags to the transformed version for each of the categories. Using this dictionary, we attempt to automatically label the ~25K tags which were left out of the manual annotation process due to low frequency. In cases where both sound and source were detected, an additional check was added to ensure that the sound-source combination was seen in the manual annotation-scheme. New combinations were disregarded as labeling error and such samples were not used. Any audio event without a sound tagged was discarded. We were able to automatically tag around 10K more sounds and 5K sources in this manner.

Label set refinement: We do a final manual pass of the unique sound and source tags and combine classes that were not taken care of by the lemmatization, e.g., “laugh” and “laughter”, “explode” and “explosion”, and “thunder” and “thunderclap”. We use a named entity recognizer⁵ to detect names of persons and merge into a single source class. The resultant dataset consists of 95,452 samples covering 101,311 sounds from 245 classes, 21,460 sources from 183 classes and 7212 qualities from 93 classes.

3.2. Acoustic and Semantic grouping of sounds

Several audio event “sounds” in our dataset can be associated with multiple distinct sources (See Fig. 3). For example, the sound

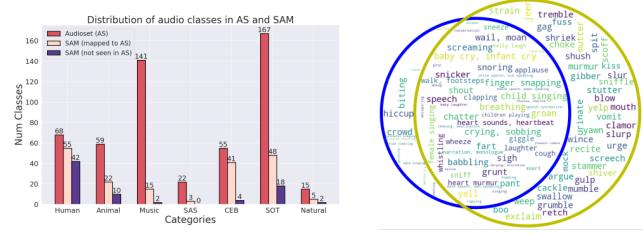


Fig. 4. a) Number of classes in each category in the AudioSet taxonomy, b) Venn diagram for human-centered audio events in AudioSet (Blue) and SAM-S (Yellow)

of buzzing is associated with different sources - flies/insects, cell-phone, doorbell and alarm, in our dataset. The sounds of a cell-phone buzzing and an alarm buzzing can be considered to be acoustically similar, so can the sounds of flies and bees buzzing. However, the buzz sound of a fly/insect has distinct acoustic signature related to its frequency spectrum and timing characteristics that distinguishes it from the cell-phone or alarm buzzes. In a semantic sense, these are all generally referred to as buzzing. Hence, for modeling purposes, one of two options can be considered: 1) Retain source-specific sounds as individual classes. 2) Merge acoustically and semantically similar sounds

If such sounds are considered as a single class, we reduce the total number of classes and obtain more representative samples per class, while at the same time increase the acoustic variability within a single sound class.

As an example, the audioset (AS) taxonomy [7] is hierarchical, with the different branches of the hierarchy being organized by the source. Here, the sound “buzz” is seen in 5 different audioset hierarchies, under ‘alarm’, ‘telephone’, ‘fly’, ‘bees’ and ‘onomatopoeia’. Most modeling techniques developed on the Audioset data adopt option 2, by flattening out the hierarchy and considering each sound as a single class.

In our taxonomy, we make a practical choice of not following the audioset method due to two reasons: 1. Keeping source-ambiguous sounds separate significantly reduces the number of samples available to train/evaluate machine learning models, 2. Following DCMP guidelines, we do not always have information about the source of an audio-event.

3.3. Overlap with Audioset

We are interested in understanding the distribution of sound events in movies and how they compare with existing datasets. In order to do this, we distribute each of the sound classes in our dataset into two groups, a) shared sounds, b) movie-specific sounds. Shared sounds refer to the classes which exist in both our taxonomy as well as the Audioset taxonomy while movie-specific sounds are those that exist in our taxonomy alone.

For each of the sounds in SAM-S, we matched one or more corresponding classes in Audioset taxonomy [7], in order to analyze label coverage. Pairwise cosine similarity scores were extracted between sentence transformer embeddings (MiniLM-L6-v2) [33] of the two taxonomies. For each sound, top-5 audioset class matches were then manually verified. Out of 245 sound classes, we found one or more direct matches in Audioset for 170 classes. The remaining classes were manually mapped, if found relevant, to an equivalent AS class. Sounds originating from multiple possible sources were mapped to each of the relevant categories. This resulted in a total

⁴<https://spacy.io/api/lemmatizer>

⁵<https://spacy.io/usage/linguistic-features#named-entities>

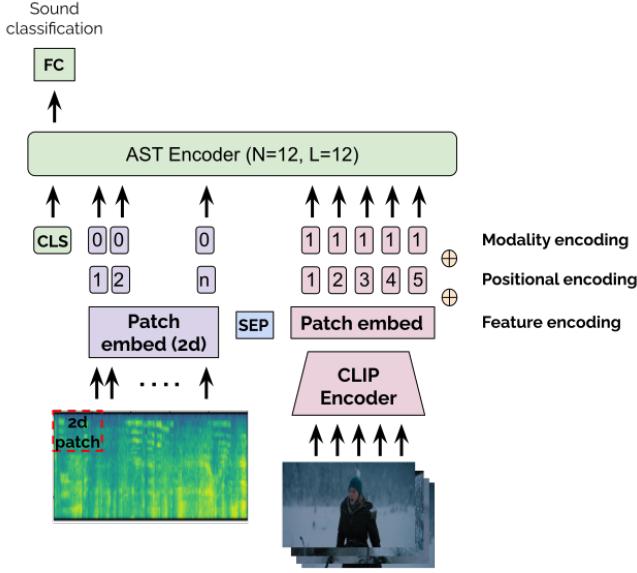


Fig. 5. AST-MM: Multimodal Transformer architecture for audio event detection in movies.

of 189 shared sound classes. It is interesting to note that although some of these classes exist in the audioset taxonomy, they have no representative data samples, for example; screech, blare, yawn and boooing.

Finally the remaining ‘movie-specific sounds’ are manually grouped into the high-level categories. The distribution of the ‘shared sounds’ and ‘movie-specific sounds’ are shown in Fig. 4a. We can see that highest coverage is obtained for human sounds and source-ambiguous sounds. In the case of music, instances in SAM-S typically do not specify music instruments/genres, hence we do not observe the level of detail as in Audioset.

In Fig. 4b, we can see that SAM-S includes most of the human sound classes found in Audioset. However, it also contains many sounds not found in Audioset; including single-person sounds such as yawn, sniffle, strain and scoff, as well as crowd-sounds such as murmur, clamor and argue. Therefore, SAM-S can also be used to augment Audioset with more fine grained human-centric sounds.

4. EXPERIMENTS

In order to transfer knowledge from large-scale models, we create 10s segments by adding context (~5s) on either side of the audio-event. We create a train-validation-test (80-10-10) split based on the movies, i.e., we use 344 movies for training, 43 for validation and 43 for final evaluation. For development and evaluation purposes, we restrict to sound classes that explain at least 0.1% of the entire data, which results in 120 sound classes.

4.1. Baseline Audio Models

We conduct audio-only baseline experiments using two SoA models. The first is a Resnet-18 model [9] using 512-dim log-spectrograms pre-trained on VGGSound dataset. We also fine tune a transformer-based AED model - AST [23], which has shown state-of-the-art performance for audio event detection on Audioset. 128-dim log-mel

Table 2. Uni- and multi- modal results on SAM-C

| Model | Modality | mAP | mAUC | d-prime |
|------------|----------|-------------|--------------|-------------|
| VGGSound | A | 14.1 | 87 | 1.59 |
| AST | A | 34.76 | 95.02 | 2.33 |
| AST-MM (S) | AV | 35.67 | 95.05 | 2.33 |
| AST-MM (B) | AV | 35.82 | 95.11 | 2.34 |
| AST-MM (L) | AV | 36.3 | 95.25 | 2.36 |

spectrograms are used as features to AST. For augmentation, we use mixup [34] with probability 0.5 and sample the mixup-lambda from a Beta-distribution with parameters $\alpha=\beta=10$. We also use SpecAugment [35] with a time-frequency mask of 192x48.

For our experiments, we use a batch size of 20, initial learning rate of 1e-5 and a multi-step learning rate scheduler at epoch 5 and 25 with decay of 0.85 similar to as in [23]. We train each model for 30 epochs. Since our classes are multi-label and we use mix-up, we use binary cross-entropy loss. As evaluation metrics, we use mean average precision (mAP), area under curve (mAUC), and d-prime.

4.2. AST-MM

As visual features, we use the output of CLIP-encoder (ViT-B/32) [36]. CLIP was trained in a contrastive manner using language-image pairs and has been widely used in a number of image recognition tasks. We also chose CLIP because of its ability to generalize well to unseen objects and scenes, which is often the case in movies. We extract 512-dim CLIP features at 1fps, and pad/crop the resulting features to a sequence length of 12.

We use position embeddings and modality-specific segment embeddings to encode multiple modalities in a transformer setup as in previous work[37]. We pass CLIP features through a linear layer to match-dimensions of the audio patch embeddings (768-dim). We experiment with three different position embeddings: 1. Fixed sinusoidal position embeddings (S) [38], 2. Learnable embeddings initialized with pretrained BERT position embeddings (B) [39], 3. Learnable, randomly initialized.

We also use a separator token to distinguish audio and visual sequences as we find it helps empirically. The final input representation to the encoder is obtained by adding the patch, position and modality embeddings for each of the sequences.

4.3. Results

From Table 2, we see that the audio-only model trained on Audioset clearly outperforms the one trained on VGGSound. Apart from the size of the datasets and model architecture, a reason for this could be that the constraint on audio-visual correspondence limits the range of sound classes seen in VGGSound, hence affecting its transferability to other domains. The multimodal AST-MM model shows a 5% relative improvement over the audio-only model.

5. CONCLUSION

In this paper, we release a dataset curated for audio-event detection in movies. We describe a scalable method for generating a flat-taxonomy for audio events, and compare it with existing taxonomy popularly used for audio event detection. We designed an annotation scheme to categorize the sound and source of an audio event, solely from subtitle tags. We employ state of the art machine learning models to establish baseline AED performance on the SAM-S corpus.

We incorporate visual information using CLIP-encoder features in a early-fusion manner to further improve AED multimodally.

6. REFERENCES

- [1] “Discover built-in sound classification in soundanalysis,” .
- [2] “Identifying sounds in audio streams,” .
- [3] Mahesh Kumar Nandwana and Taufiq Hasan, “Towards smart-cars that can listen: Abnormal acoustic event detection on the road..,” in *INTER-SPEECH*, 2016, pp. 2968–2971.
- [4] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, “Dcase 2017 challenge setup: Tasks, datasets and baseline system,” in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.
- [5] Arindam Jati, Amrutha Nadarajan, Raghveer Peri, Karel Mundnich, Tiantian Feng, Benjamin Girault, and Shrikanth Narayanan, “Temporal dynamics of workplace acoustic scenes: Egocentric analysis and prediction,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 756–769, 2021.
- [6] Stefan Goetze, Jens Schroder, Stephan Gerlach, Danilo Hollosi, Jens-E Appell, and Frank Wallhoff, “Acoustic monitoring and localization for social care,” *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.
- [7] Jort F Gemmeke, Daniel PW 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [8] Benjamin Ma, Timothy Greer, Dillon Knox, and Shrikanth Narayanan, “A computational lens into how music characterizes genre in film,” *Plos one*, vol. 16, no. 4, pp. e0249957, 2021.
- [9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.
- [10] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [11] Rajat Hebbar, Krishna Somandepalli, and Shrikanth S Narayanan, “Improving gender identification in movie audio using cross-domain data.,” in *Interspeech*, 2018, pp. 282–286.
- [12] Jaime Ramírez and M. Julia Flores, “Machine learning for music genre: multifaceted review and experimentation with audioset,” *Journal of Intelligent Information Systems*, pp. 1 – 31, 2019.
- [13] Claire Nozaic, *An introduction to audio post-production for film*, Ph.D. thesis, Stellenbosch: Stellenbosch University, 2006.
- [14] Amalia De Götzen, Erik Sikström, Francesco Grani, and Stefania Serafin, “Real, foley or synthetic? an evaluation of everyday walking sounds,” *Proceedings of SMC*, 2013.
- [15] “Captioning sound effects in tv and movies – 3play media,” <https://www.3playmedia.com/blog/captioning-sound-effects-in-tv-and-movies/>.
- [16] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, “Fsd50k: an open dataset of human-labeled sound events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.
- [17] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, “Freesound datasets: a platform for the creation of open audio datasets,” in *Proceedings of the 18th ISMIR Conference*; p. 486-93., 2017.
- [18] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento, “Reliable detection of audio events in highly noisy environments,” *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.
- [19] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 253–257.
- [20] 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*, 2014, pp. 1041–1044.
- [21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “Cnn architectures for large-scale audio classification,” in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
- [22] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, “Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [23] Yuan Gong, Yu-An Chung, and James Glass, “Ast: Audio spectrogram transformer,” *arXiv preprint arXiv:2104.01778*, 2021.
- [24] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 10347–10357.
- [25] Haytham M Fayek and Anurag Kumar, “Large scale audiovisual learning of sounds with weakly labeled data,” *arXiv preprint arXiv:2006.01595*, 2020.
- [26] Weiyao Wang, Du Tran, and Matt Feiszli, “What makes training multimodal classification networks hard?,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12695–12705.
- [27] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, “Attention bottlenecks for multimodal fusion,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [28] Mathilde Brousmiche, Stéphane Dupont, and Jean Rout, “Intra and inter-modality interactions for audio-visual event detection,” in *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*, 2020, pp. 5–11.
- [29] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang, “Dual attention matching for audio-visual event localization,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6292–6300.
- [30] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan, “Cross-modal attention network for temporal inconsistent audio-visual event localization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 279–286.
- [31] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros, “Audio event detection in movies using multiple audio words and contextual bayesian networks,” *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 17–22, 2013.
- [32] Ji-chen Yang, Lei-an Liu, Qing-wei Qin, and Min Zhang, “Audio event change detection and clustering in movies,” *Journal of Multimedia*, vol. 8, no. 2, pp. 113, 2013.
- [33] Nils Reimers and Iryna Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 11 2019, Association for Computational Linguistics.
- [34] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, “Learning from between-class examples for deep sound recognition,” *arXiv preprint arXiv:1711.10282*, 2017.
- [35] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barrett Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.

- [37] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine, “Supervised multimodal bitransformers for classifying images and text,” *arXiv preprint arXiv:1909.02950*, 2019.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.