Improving Audio Classification by Transitioning from Zero- to Few-Shot

James Taylor[1]*, Wolfgang Mack*[2]

[1]BabbleLabs, Cisco Systems, Inc., Scotland
[2]BabbleLabs, Cisco Systems, Inc., Germany
jamesta2@cisco.com, womack@cisco.com

Abstract

State-of-the-art audio classification often employs a zero-shot approach, which involves comparing audio embeddings with embeddings from text describing the respective audio class. These embeddings are usually generated by neural networks trained through contrastive learning to align audio and text representations. Identifying the optimal text description for an audio class is challenging, particularly when the class comprises a wide variety of sounds. This paper examines few-shot methods designed to improve classification accuracy beyond the zero-shot approach. Specifically, audio embeddings are grouped by class and processed to replace the inherently noisy text embeddings. Our results demonstrate that few-shot classification typically outperforms the zero-shot baseline.

Index Terms: dimensionality reduction, few-shot, audio classification, noise classification, audio embeddings

1. Introduction

Identification and classification of sound events play a vital role in understanding acoustic environments. Competitions like DCASE [1] and accompanying datasets like TUT [2] and CochlScene [3] showcase the importance and interest in acoustic scene classification and event detection. Practitioners working on audio machine learning rely on such tools for crucial downstream tasks, such as understanding urban environments and acoustic surveillance in smart cities [4, 5], fall alerts in care homes [6], monitoring for conservation and ecological research [7], and speech enhancement [8].

Classically, sound event classification is performed using neural networks that map to class activity, which require extensive labeled datasets for each class present in the system [9, 10]. These models, which need retraining whenever a new class is introduced, are inherently limited in adaptability. In dynamic environments where new events frequently emerge, such constraints become increasingly apparent. To address these challenges, few-shot learning approaches have explored how to classify with minimal labeled examples, often leveraging classical machine learning techniques to reduce data dependence and increase generalization. While [11] explores the power of pre-training for a specific domain, [12, 13] demonstrate how classes can be added incrementally and continually. In [14, 15], the authors explore how few-shot approaches perform in multi-label datasets. Recent exploration into lightweight models achieve robust performance with limited labeled data, providing a foundation for more adaptable systems.

Alternatively, inspired by the multi-modality of Contrastive Language-Image Pretraining (CLIP) [16], early attempts to constrastively model text and audio [17, 18] modeled audio and its labels in the same space. Further models such as CLAP showed that by leveraging mass-amounts of audio and natural language descriptions, contrastive models can be successfully used to classify, retrieve, and even caption audio [19, 20, 21, 22]. Bringing the robustness of language modeling to the problem allowed sound event classification to be approached as a zero-shot problem. These models, while successfully modeling audio and their descriptors together, still show room to improve in terms of understanding sound sequences, sound concurrency, and disambiguating foreground and background sounds [23]. In the latest rendition of audio-language models, Pengi framed all problems as text generation problems and has allowed for enhanced captioning, question answering, source-distance discrimination, sentiment/emotion recognition, and music analysis [24]. Meanwhile, Audio-Flamingo shows augmenting large language models to understand audio has added-on the ability to count sound occurrences, describe the ordering of sounds, and assess sound quality [25].

While these models have become increasingly flexible and impressive, they inherently rely on text embeddings. These text embeddings are general purpose, but also introduce variability, sometimes leaving practitioners to try and find the 'best' prompt. In an effort to de-noise these embeddings from other tasks and make them most applicable to such a scenario, we explore light-weight approaches to few-shot, closed-set audio classification. Our approach side-steps the use of these text embeddings by defining classes with a few audio samples instead. In this way, we harness the power of pre-trained audio encoders and remain adaptable to new classes.

Our contribution is a practical and effective extension of modern zero-shot, text-based audio classification tools to a few-shot setting using audio embeddings, showing consistent improvements with minimal data. Another contribution is a rigorous evaluation of the proposed method and baselines. Our empirical results demonstrate that this approach significantly improves on existing methods, offering a more adaptable solution for audio classification in dynamic environments.

The paper is structured as follows. In Section 2, we introduce a signal model and embedding-based zero-shot audio classification. The proposed method is presented in Section 3. Section 4 contains information about the hyperparameters and the data we used. Finally, we evaluate the performance in Section 5 followed by a brief conclusion in Section 6.

2. Fundamentals

We consider a dataset $\mathcal{X} = \{x_n \mid n = 1, \ldots, N\}$, consisting of N discrete audio signals $x_n \in \mathbb{R}^{T_n}$, where $T_n \in \mathbb{N}$ specifies the number of samples in each signal. Correspondingly, we define a set of labels $\mathcal{Y} = \{y_n \mid n = 1, \ldots, N\}$, with $y_n \in \mathcal{C}$

representing the class label of the n-th audio recording, and $\mathcal{C}$ is the set of possible classes. To handle the variability in audio lengths and to extract meaningful representations, we employ an embedding-extractor model $\mathcal{F}_A \colon \mathbb{R}^{T_n} \to \mathbb{R}^L$. This model maps each audio recording x_n to a fixed-dimensional embedding $e_n \in \mathbb{R}^L$ of length $L \in \mathbb{N}$, defined as

$$e_n = \mathcal{F}_A(x_n), \tag{1}$$

resulting in the set of embeddings $\mathcal{E}_\mathcal{A} = \{e_n \mid n = 1, \ldots, N\}$. Similarly, we define a set of text embeddings E_T using an embedding-extractor $\mathcal{F}_T$, which processes text akin to $\mathcal{F}_A$. The training of $\mathcal{F}_A$ and $\mathcal{F}_T$ involves reducing the distance between paired audio-text embeddings in a contrastive manner. During inference, each audio class is described textually and transformed via $\mathcal{F}_T$ to yield a reference embedding $e_c \in \mathbb{R}^L$ for class c. Audio samples are classified by transforming them using $\mathcal{F}_A$ and mapping them to the closest text embedding in a zero-shot manner [22]. The challenge lies in selecting text that yields the 'optimal' embedding for classifying a specific audio class, due to the complexity and variability of language. Different texts may describe the same audio class in various ways, making it difficult to determine which description will best facilitate classification.

3. Few-Shot Audio Classification Methods

In many practical scenarios, a small number of audio samples for all classes already exists or can be gathered with minimal effort. Extracting audio embeddings from such a development set and constructing e_c from them in a few-shot manner instead of text leads to an improved classification, as we can see in Section 5. Few-shot classification is advantageous over zero-shot classification here, as e_c is directly obtained in a way that optimizes the assignment of embeddings to classes. In the following, we present different few-shot methods to compute e_c. We define the set of audio embeddings per class c used to compute e_c as $\mathcal{E}^c$.

- **Averaged Embeddings**: In this method, we compute the average of the embeddings for each class c using

$$e_c = \sum_{e_i \in \mathcal{E}^c} w_i \cdot e_i, \tag{2}$$

  where w_i denotes the weight associated with embedding e_i. Two different weighting schemes are employed. The first scheme assigns $w_i = \frac{1}{|\mathcal{E}^c|}$, where $|\mathcal{E}^c|$ is the number of embeddings in the set, effectively making (2) the arithmetic mean. The second scheme assigns weights as $w_i = \frac{1}{|\mathcal{E}^c| \cdot \|e_i\|_2}$, where $\|e_i\|_2$ denotes the l_2-norm of e_i, resulting in Equation (2) representing the average of normalized embeddings. This normalization aligns with the properties of cosine similarity, where the direction of vectors, rather than their magnitude, is critical. This approach is consistent with the training methodology of CLAP, where cosine similarity was used.

  Subsequently, an audio embedding e_n is assigned to a class c based on a minimum distance criterion d given by

$$c = \arg\min_{c \in \mathcal{C}} d(e_c, e_n). \tag{3}$$

  We employ two distance metrics for this classification: mean squared error (MSE) and cosine similarity (COS), the latter of which is consistent with the training process of CLAP.

- **linear discriminant analysis and mutual information**: We train a linear discriminant analysis (LDA) model on the audio embeddings $\{\mathcal{E}^c \forall c \in C\}$ classes to obtain a classification system. During inference, the trained LDA model is used to estimate c from e_n. In Section 5, we observe a reduced performance of LDA compared to simple averaging. We assume the curse of dimensionality [26] to be a reason for the reduced performance. Hence, we additionally employ dimensionality reduction [27] of e_n before LDA using the mutual information (MI) approach [28, 29].

4. Experimental Setup

4.1. Data

The performance evaluation is based on three datasets: BBL (internal), ESC-50, and FSD50K. These datasets provide a diverse range of sound events and evaluation scenarios. **BBL** consists of 3600 audio signals divided equally among 24 humanly annotated sound classes. The audio originates from AudioSet [30]. The classes cover a wide range of indoor and outdoor sounds, such as cars, wind, vacuum cleaners, as well as human non-speech sounds like baby crying or laughter. The 150 signals per class are split into non-overlapping sets of 100 for development and 50 for evaluation. **ESC-50** [31] contains 2000 audio signals, each 5 seconds long, distributed equally among 50 sound event classes. The classes are organized into five broad topics: animals, natural sounds, human sounds, interior/domestic, and exterior/urban noises. Each audio sample is assigned a single label. The dataset uses a 5-fold cross-validation split for evaluation. **FSD50K** [32] includes more than 50,000 audio signals distributed across 200 sound classes. The audio is organized hierarchically based on the AudioSet ontology, leading to multi-label annotations where a single recording may belong to multiple related categories. The classes cover a wide array of everyday sounds, such as musical instruments, animals, household sounds, and crowd chatter. Recordings vary in length and quality, reflecting real-world variability. The dataset is provided with a development and evaluation split, and performance is typically reported using mean average precision (mAP).

4.2. Few-Shot Methods

To obtain $\mathcal{E}^c$ for the few-shot (FS) methods, we choose $|\mathcal{E}^c|$ training samples from the dev set. In the case of FSD50K, we specifically choose those with the least class overlap with other files. We then preprocess audio using the pretrained audio encoder of CLAP (Version 2023) for each dataset separately. To obtain e_c, the embeddings are subsequently either averaged per class (AVG), used to train an LDA system, or dimensionality reduced using mutual information (MI) plus subsequently AVG or LDA. We refer to the different FS methods as $\text{FS}_\bullet^{|E^c|}$ where $\bullet \in \{\text{AVG}, \text{LDA}, \text{MI+AVG}, \text{MI+LDA}\}$ represents the respective method. When using MI, we vary the number of features selected by MI a K ratio of $|\mathcal{C}|$ with $K \in \{1/2, 1, 2, 4, 8, 16, 32\}$ such that the embedding size after MI is $|\mathcal{C}| \cdot K$. We decided to vary K in this way in order to create a fair comparison between the datasets, as we believe $|\mathcal{C}|$ and the number of features needed to differentiate between classes are closely linked.

5. Performance Evaluation

We compared different zero-shot and FS methods to each other using their classification accuracy (BBL and ESC-50) or the

Table 1: *Top shows zero-shot baselines, bottom shows FS methods. For ESC-50 and FSD50K, results are obtained from the respective papers. For BBL, results are obtained using open-source implementations. For methods that use feature selection via MI, we use the best K for a given method/dataset pair (see Figure 3).*

Model (ZS)	**BBL (dev/eval, acc)**			**ESC-50 (5-fold, acc)**	**FSD50K (dev/eval, mAP)**		
CLAP22	0.602			0.826	0.302		
PENGI	0.547			0.920	0.468		
CLAP23	0.623			0.948 (FT: **0.983**)	0.485		
Model (FS)	$\lvert\mathcal{E}^c\rvert$=*10*	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*50*	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*10*	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*50*
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{AVG}}$	0.678	0.700	**0.716**	0.970	0.540	0.561	0.579
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{LDA}}$	0.615	0.623	0.632	0.968	0.440	0.487	0.526
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{MI+AVG}}$	0.681	0.698	**0.716**	**0.971**	0.559	0.576	**0.586**
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{MI+LDA}}$	0.563	0.612	0.650	0.969	0.509	0.551	0.559

Model (ZS)	**BBL (dev/eval, acc)**		**ESC-50 (5-fold, acc)**	**FSD50K (dev/eval, mAP)**	
CLAP22	0.602		0.826	0.302	
PENGI	0.547		0.920	0.468	
CLAP23	0.623		0.948 (FT: **0.983**)	0.485	
Model (FS)	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*50*	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*20*	$\lvert\mathcal{E}^c\rvert$=*50*
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{AVG}}$	0.700	**0.716**	0.970	0.561	0.579
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{LDA}}$	0.623	0.632	0.968	0.487	0.526
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{MI+AVG}}$	0.698	**0.716**	**0.971**	0.576	**0.586**
$\text{FS}^{\lvert\mathcal{E}^c\rvert}_{\text{MI+LDA}}$	0.612	0.650	0.969	0.551	0.559

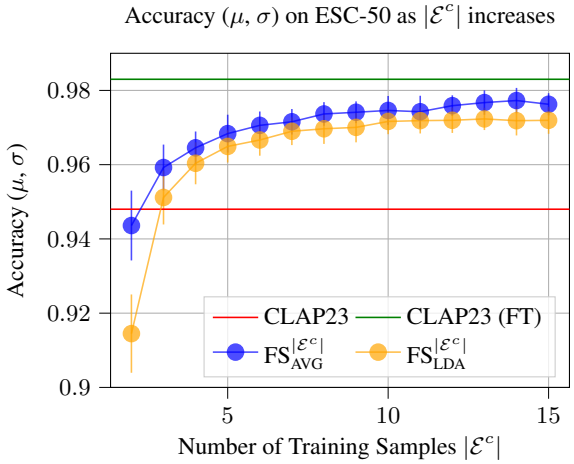

Figure 1: *Mean and standard deviation of $FS^{|\mathcal{E}^c|}_{AVG}$ and $FS^{|\mathcal{E}^c|}_{LDA}$ using between 2 and 15 training samples, using 30 runs for each $|\mathcal{E}^c|$ value. CLAP23 (FT) is the fine-tuned CLAP model for that specific data set.*

mAP (FSD50K). A summary of the results is given in Table **??**. The FS methods used COS as the distance metric and non-normalized embeddings. Using the MSE instead of COS as the distance metric strongly deteriorated the results. Normalizing the embeddings showed limited influence on the results.

5.1. Zero-Shot vs. Few-Shot Methods

A comparison of zero-shot and FS methods is given in Table **??**. For the zero-shot methods, CLAP 23 performs best for all three datasets. For the FS methods, we see increased performance for an increasing $|\mathcal{E}^c|$. Comparing $\text{FS}^{10}_{\text{AVG}}$ to CLAP 23 the accuracy in BBL increases from 0.623 to 0.678. For $\text{FS}^{50}_{\text{AVG}}$, the increase is even higher to 0.716. Consequently, a small number of reference samples per class can be used to improve the performance compared to zero-shot baselines. We use zero-shot baselines as a reference point to assess how our system improves when transitioning from a zero-shot to a few-shot system. For $\text{FS}^{|\mathcal{E}^c|}_{\text{AVG}}$, the performance is always better than the zero-shot baselines. Interestingly, for $\text{FS}^{|\mathcal{E}^c|}_{\text{LDA}}$ the performance only improves for higher $|\mathcal{E}^c|$. For $|\mathcal{E}^c| = 10$, the performance is even reduced compared to the zero-shot baselines.

To investigate the influence of $|\mathcal{E}^c|$ further, we plot the mean and standard deviation of the accuracy for multiple runs using ESC-50 in Figure 1. For small $|\mathcal{E}^c|$, the standard deviation of the accuracy over multiple runs is higher than for a large $|\mathcal{E}^c|$. This is expected as more audio embeddings reduce the influence of individual files, leading to a more reliable e_c. Increasing $|\mathcal{E}^c|$ leads to an increase in accuracy for all FS methods. The mean-performance of $\text{FS}^{|\mathcal{E}^c|}_{\text{LDA}}$ is always lower than of $\text{FS}^{|\mathcal{E}^c|}_{\text{AVG}}$. We assume that AVG is preferable over LDA here, as the audio embeddings are high-dimensional, and the curse of dimensionality reduces the performance of LDA. Compared to CLAP 23, less than five embeddings per class are required to improve the results. Compared to a CLAP model that is fine-tuned on ESC-50, the FS methods perform slightly worse. Note that fine-tuning comes with higher computational and tempo-

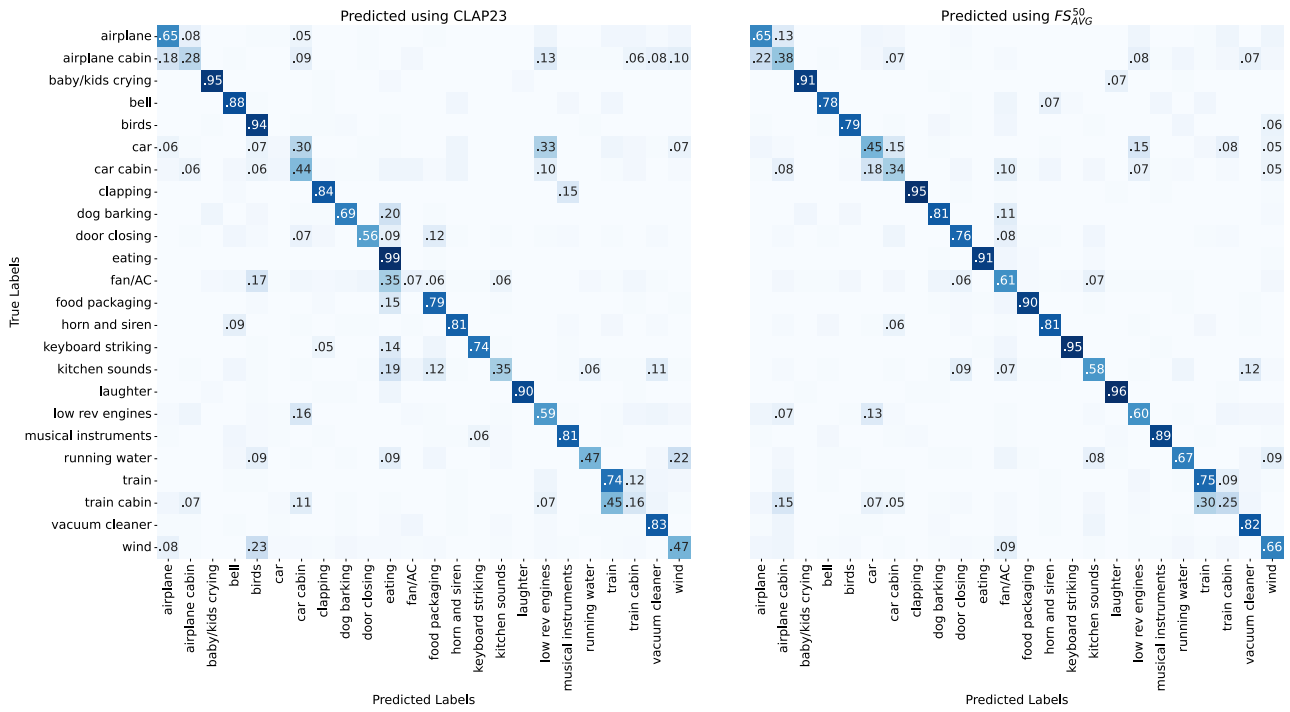

Figure 2: *Confusion matrix for BBL. On the left, we see the comparison between the true label and using base CLAP23. On the right, we see a much stronger diagonal trend and an improvement in most classes when using* $FS^{|\mathcal{E}^c|}_{AVG}$

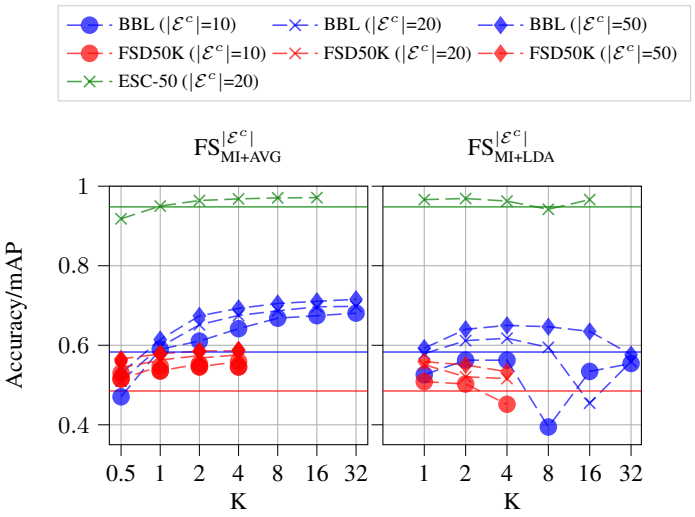

Figure 3: *Accuracy as we increase* K. *Solid horizontal lines indicate the baseline, CLAP23. The left plot shows continual improvements as we use MI as a feature selector before averaging, while we see a more complex relationship when using MI as an intermediary step before LDA.*

ral costs compared to the investigated FS methods and might deteriorate performance on unseen classes.

To assess the effect of zero-shot versus FS methods on a class-by-class comparison, we visualize the confusion matrix for BBL using CLAP 23[1] and $\text{FS}^{|\mathcal{E}^c|}_{\text{AVG}}$ in Figure 2. Overall, accuracy improves, with $\text{FS}^{|\mathcal{E}^c|}_{\text{AVG}}$ outperforming CLAP 23 by approximately 10 percentage points. Notable gains occur in classes that CLAP 23 misclassifies (e.g., 'car', 'fan/AC') or over-predicts (e.g., 'eating'), though some, like 'baby/kids crying' or 'eating', see slight accuracy drops. Audio-based embeddings help distinguish classes that textual descriptions confuse, such as 'car' vs. 'car cabin', improving 'car' accuracy from nearly 0% to 45%. However, class overlap remains challenging, particularly for paired categories like 'airplane' and 'airplane cabin'.

5.2. Few-Shot: Dimensionality Reduction using Mutual Information

We introduce MI as a preprocessing step of the embeddings to retain only the most relevant features, aiming to improve classification performance while mitigating the curse of dimensionality, particularly for LDA. A comparison of $\text{FS}^{|\mathcal{E}^c|}_{\text{MI+AVG}}$ and $\text{FS}^{|\mathcal{E}^c|}_{\text{AVG}}$ in Table **??** shows minor positive effect of MI. For $\text{FS}^{|\mathcal{E}^c|}_{\text{MI+LDA}}$, a positive effect of MI can be seen for $|\mathcal{E}^c| = 50$.

To investigate the number of required features, we plot the accuracy/mAP over K in Figure 3. For MI+AVG, increasing K improves the performance till it slightly surpasses using all 1024 features. For the largest K, around 700 to 800 features are selected. This suggests that most features are relevant to distinguish between audio classes. For MI+LDA, increasing K does not automatically lead to increased performance. LDA has difficulties handling high-dimensional embeddings (curse of di-

[1]Some label adjustments were needed to align with CLAP's mappings, e.g., 'low rev engines' → 'idling tractor' and 'baby/kids crying' → 'baby crying', highlighting the limitations of text embeddings.

mensionality), such that the performance decreases. These observations are in line with the overall reduced performance of LDA compared to a simple averaging approach. As most features are important, and LDA struggles with high-dimensional data, AVG performs better and more stable.

6. Conclusion

In this study, we improved audio classification by transitioning from zero-shot to few-shot methods, addressing the limitations of noisy text embeddings. Our few-shot approach, utilizing a small number of audio samples, consistently outperformed zero-shot classifiers on different datasets by 2 to 10 % points in accuracy. This success is due to the more reliable class representations formed by direct audio embeddings, enhancing the system's robustness. Future work could explore extending this approach to more complex tasks, such as hierarchical classification or scenarios involving overlapping and concurrent sound events, further advancing the capabilities of few-shot audio classification.

7. References

[1] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound Event Detection in the DCASE 2017 Challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 992–1006, 2019.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Eur. Signal Process. Conf.*, 2016, pp. 1128–1132.

[3] I. Y. Jeong and J. Park, "CochlScene: Acquisition of acoustic scene data using crowdsourcing," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC*, 2022, pp. 17–21.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *MM - Proc. ACM Conf. Multimed.*, ser. MM '14, 2014, pp. 1041–1044.

[5] J. Holub and J. Svatos, "AI-Based Acoustic Surveillance System for Smart Cities," in *8th IEEE Int. Forum Res. Technol. Soc. Ind. Innov. RTSI 2024 - Proceeding*, 2024, pp. 295–299.

[6] NHS Transformation Directorate, "Acoustic monitoring integrated with electronic care planning," https://transform.england.nhs.uk/ai-lab/explore-all-resources/understand-ai/acoustic-monitoring-integrated-electronic-care-planning/, 2020.

[7] E. Browning, R. Gibb, P. Glover-Kapfer, K. E. Jones, G. Billington, Z. Burivalova, D. Clink, J. De Ridder, J. Halls, T. Hastings, D. Jacoby, A. Kalan, A. Kershenbaum, S. Linke, S. Lucas, R. Machado, P. Owens, C. Sutter, P. Trethowan, R. Whytock, and P. Wrege, "Passive acoustic monitoring in ecology and conservation," WWF-UK, Tech. Rep. 2, 2017.

[8] A. Sabra, C. Wronka, M. Mao, and S. Hijazi, "SECP: A Speech Enhancement-Based Curation Pipeline for Scalable Acquisition of Clean Speech," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2024, pp. 11 981–11 985.

[9] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, 2015, pp. 1–6.

[10] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 3, pp. 540–552, 2015.

[11] J. W. Jung, H. J. Shim, J. H. Kim, and H. J. Yu, "DCasenet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2021, pp. 621–625.

[12] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2021, pp. 321–325.

[13] R. Singh, P. Nema, and V. K. Kurmi, "Towards Robust Few-shot Class Incremental Learning in Audio Classification using Contrastive Representation," in *Interspeech*, 2024, pp. 5023–5027.

[14] K. H. Cheng, S. Y. Chou, and Y. H. Yang, "Multi-label Few-shot Learning for Sound Event Recognition," in *IEEE 21st Int. Work. Multimed. Signal Process. MMSP*, 2019, pp. 1–5.

[15] J. Liang, H. Phan, and E. Benetos, "Learning From Taxonomy: Multi-Label Few-Shot Classification for Everyday Sound Recognition," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2024, pp. 771–775.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8748–8763, 2021.

[17] H. H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2Clip: Learning Robust Audio Representations From Clip," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2022, pp. 4563–4567.

[18] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending Clip To Image, Text and Audio," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2022, pp. 976–980.

[19] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2023.

[20] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation," in *Interspeech*, 2024, pp. 57–61.

[21] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2023, pp. 1–5.

[22] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision For General-Purpose Audio Representations," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2024, pp. 336–340.

[23] H. H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-Text Models Do Not Yet Leverage Natural Language," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2023, pp. 1–5.

[24] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An Audio Language Model for Audio Tasks," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.

[25] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities," in *Proc. Mach. Learn. Res.*, vol. 235, 2024, pp. 25 125–25 148.

[26] N. P. Faísca, K. I. Kouramas, B. Rustem, and E. N. Pistikopoulos, *Dynamic Programming*, ser. Rand Corporation research study, 2014, vol. 1-7.

[27] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, no. C, 2021.

[28] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[29] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, pp. 175–186, 2014.

[30] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2017, pp. 776–780.

[31] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *MM 2015 - Proc. 2015 ACM Multimed. Conf.*, ser. MM '15, 2015, pp. 1015–1018.

[32] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, 2022.