Automatic re-labeling of Google AudioSet for improved quality of learned features and pre-training

1st Tomasz Grzywalski, 2nd Dick Botteldooren
Department of Information Technology
Ghent University
Ghent, Belgium
0000-0002-9388-0494 0000-0002-7756-7238

***Abstract*—AudioSet, comprising over 2 million human-labeled sound clips, remains one of the biggest and most versatile publicly available audio events datasets. Deep neural networks trained on this data are able to detect 527 types of sounds organized in a hierarchical (tree-like) structure named ontology. However, these models are also often used as feature extractors or serve as a basis for knowledge transfer to other sound detection and classification tasks. When describing the AudioSet recordings, raters were asked to choose one or more labels from the ontology. Analysis of the dataset reveals that raters were inconsistent and imprecise when dealing with the hierarchy of sounds. For example, some raters selected only the most precise labels while others selected all relevant labels (i.e. all parents of selected child labels). Additionally, a large fraction of sound clips are labeled with general labels without providing any fine-grained labels. These issues harm the quality of features learned by the models trained on AudioSet. As a remedy, we propose two ways in which the dataset can be automatically re-labeled to achieve specific, consistent and complete label definitions on all levels of the ontology tree. Experimental results show significant improvement in the performance of new models trained on features extracted from, or initialized with weights transferred from base models trained with re-labeled AudioSet data. In a more general view, this work highlights the importance of paying attention to the labeling of data as a way to improve model accuracy.**

***Index Terms*—AudioSet, missing labels, feature extraction, transfer learning, pre-training, ontology**

I. Introduction

Two million AudioSet (AS) [1] audio clips were labeled with the use of an ontology of 632 labels organized in a hierarchical structure. The ontology contains 7 root labels: (1) Animal, (2) Channel, environment and background, (3) Human sounds, (4) Music, (5) Natural sounds, (6) Sounds of things, and (7) Source-ambiguous sounds, and has a maximal depth of 5 levels. A small excerpt from the ontology focusing on guitar sounds is presented in Fig. 1. Each AS video clip (including both image and audio) was labeled by human rates whose task was to select one or more labels that best describe the audio content. Out of 632 labels available in the ontology, only 527 were actually used by the raters. The remaining labels were either abstract concepts that served only as intermediate nodes for more concrete child labels or were blacklisted due to being too obscure or difficult to define. The average number of labels per recording in AS is 2.7 and the number of recordings per label varies from 121 to over a million.

Machine learning models trained on AS are able to detect a large variety of sound events, although for some classes of sounds the detection accuracy might be low [2]. Nevertheless, such models have many practical applications, for example in [3] a transformer neural network was utilized to automatically generate AS labels that were used to improve the accuracy of the model performing audio captioning and in [4] a long short-term memory (LSTM) neural network trained on AS was used to identify different types of urban sounds.

Another utility of AS models lies in the quality of learned features. In this case, the trained model, usually a neural network, is fed with an audio segment and an output from a chosen hidden layer is taken as a short, high-level description of that segment. Such embedding can then be used to perform classification, clustering, or nearest neighbor search. In [5] the quality of features learned using AS was investigated and positively evaluated. An example utility of such features includes clusterization and identification of bird calls [6] and estimation of annoyance in urban soundscapes [7].

Last but not least, AS can be used as a pre-training step to solve some downstream sound recognition tasks. In this context the AS allows the model to learn some useful initial features, which speeds up (and often improves the outcome of) training of a machine learning model to solve the downstream (target) task. As an example, in [8] a pre-training on AS was used to improve the detection of abnormal heart sounds in stethoscope recordings, and in [9] such pre-training was utilized to find a better joint audio-video representation of video clips. In [10] the AudioSet was used to pre-train a transformer neural network (in addition to ImageNet pre-training), which improved classification of environmental sounds and recognition of speech commands.

Among the models trained on AS, the most notable are the PANNs [11]. They have been successfully employed in all three aforementioned scenarios in various areas of digital audio processing. However, their performance in sound event detection is far from perfect, reaching only 0.439 mean average precision (mAP, [12]) on the official AS validation set. One

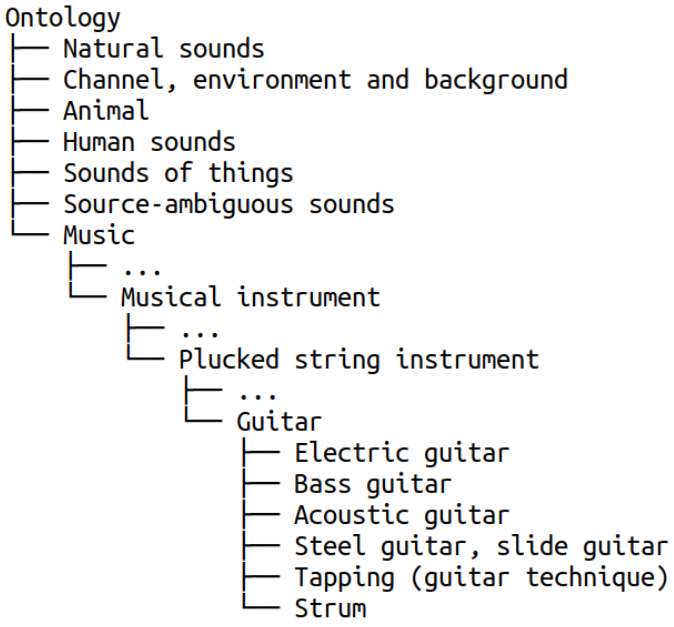

Fig. 1. Fragment of the ontology used by AudioSet focusing on guitar sounds. The ontology tree includes seven root labels and has a maximal depth of 5 levels.

of the main factors that limit the models' performance is the prevalent issue of missing labels, which is the main focus of this work.

The issue of missing labels in AS was previously reported in [13] and [14]. In the latter, the authors proposed to exclude the most critical missing labels from training, with the missing labels being identified through an iterative teacher-student training process. The solution is, however, computationally demanding and is not guaranteed to fix all the issues. In fact, this method works under the assumption the teacher model is good enough to be able to identify missing labels. As we show in the next chapter, in some cases the proportion of missing labels is so high that this requirement cannot be met. In contrast, we propose automatic re-labeling of the whole AS using logical rules that are derived directly from the hierarchy of labels in the ontology. In the re-labeled data, most of the issues found in the original (noisy) labels are avoided. The usefulness of the proposed solution is tested in a scenario where the re-labeled data is used for pre-training or training of an embedding extractor, which is another contribution of this paper.

II. Issues with AudioSet labels

According to our investigation, the main source of missing labels in AS stems from the different approaches taken by the raters to selecting more general labels related to the identified fine-grained labels. For instance, when an electric guitar sound was identified in the recording, some raters selected only the "Electric guitar" label while others selected all patent labels as well, i.e. "Guitar", "Plucked string instrument", "Musical instrument" and "Music". There were also those who selected, for example, only "Electric guitar" and "Guitar" while leaving the remaining parent labels unselected.

When looking at the whole database we can see that, when a child label was selected, only in 52% of cases the parent label was added as well (not counting cases when the parent label was abstract or blacklisted). However, it should be also noted that these proportions vary a lot between different sections of the ontology tree. For example, among 12024 recordings labeled as "Electric guitar", 10243 (85%) were also labeled as "Guitar". Among 51597 recordings labeled as "Guitar", 40266 (78%) were also labeled as "Plucked string instrument". In contrast, among 3392 recordings labeled as "Rain", only 24 (0.7%) were labeled as "Water" and among 2603 recordings labeled as "Duck", only 5 (0.2%) were labeled as "Fowl".

This inconsistency presents a major challenge when training a machine learning model, because in this case very similar recordings are presented to the model with conflicting training targets. Indeed, it is unclear if the "Plucked string instrument" label is relevant for all plucked string instrument sounds or only in the cases when a concrete type of plucked string instrument cannot be determined or can be determined, but is not present in the ontology. Without a clear label definition, the model cannot fully converge leading to learning of sub-optimal features.

Another source of ambiguity in the labels stems from the inability of the raters to identify some fine-grained labels. Please note that raters were relying heavily on the video content when making their assessment. Visibility of the sound source must have had a major impact on their decision and quite often a more general label was selected without specifying any children if the sound source was not clearly visible. For example, among 51597 recordings labeled as "Guitar", only 30863 (60%) were labeled with any of the child labels. Surely among these 40% of recordings, for which the concrete guitar type was not specified, there were some electric, acoustic, and bass guitars but they were difficult to identify. This also poses a major challenge for the training as all these 20734 recordings are used as negative examples of the electric, bass, and acoustic guitar classes when surely some of them actually contain such guitar sounds.

III. Proposed modification of AudioSet labels

We argue that the task of training a machine learning model using AS can be made much easier and provide better results if we redefine the AS labels in a way that ensures concrete class definitions, consistency of labeling and minimizes the potential for conflicting training targets.

A. Parent-expanded labels (PEL)

The easiest and most straightforward solution to the first problem (inconsistency in adding parent labels) is to automatically add all missing parent labels. This should be done recursively up to the ontology tree root label(s), starting form each original label. An opposite action (removing of all parent labels) would be incorrect because there might be multiple sources of sounds that would justify the existence of parent

label even if a child label is also present, e.g. a duet of acoustic guitar ("Acoustic guitar" label) and a banjo ("Guitar" label).

In the remainder of this work, we will refer to the modified labeling where all missing parents are automatically added as parent-expanded labeling (PEL) and the original AS labeling will be abbreviated as OL.

B. Parent-expanded and children-masked labels (PE-CML)

As a further extension of the previous idea, we propose a method to cope with potential errors in the AS labels related to the possibility of the rater selecting only a more general label when more concrete labels were available in the ontology. To remove this kind of potential error, we need to assume that whenever a parent label is selected from ontology but none of its children are, then one or more of its children might actually be a correct label. Since we don't know which ones are correct, we need to act as if the applicability of all child labels is unknown.

This is realized by adding a new category of labels - masked labels. A masked label is a label that is excluded from training by which we mean that it does not participate in loss calculation and gradient propagation for a given recording. As such, we define a third category of labels which we name PE-CML (parent-expanded and children-masked labeling). In this case, we automatically add all missing parents to the set of labels (PE) and additionally mask all child labels related to the original labels if none of the child labels related to the original label was selected by raters (CM). In the latter, child labels should be understood as all labels below the original label on the ontology tree, down to leaf label(s).

C. Impact of re-labeling on dataset structure

Please note that the alternative labelings significantly change the structure of the labels in the dataset. In comparison to OL, the PEL is expected to include much more labels which intuitively might have a positive impact on the training. On the other hand, in the PEL and PE-CML the imbalance between the number of examples per label significantly increases because parent labels become much more prevalent in the dataset while child labels (especially leaves) don't, which can further reduce impact of the underrepresented classes on the training.

IV. Experiment design

The proposed modifications of AS labels are difficult to be evaluated directly. This is because the modifications influence the whole dataset - the training as well as the testing subsets. Results obtained on OL, PEL and PE-CML testing subsets cannot be compared directly because in each case the structure of labels is different. When using, for example, OL testing set, it is expected that the best results will be achieved with OL training set because it has a matching (albeit inconsistent, therefore sub-optimal) label structure.

To solve this issue we propose a two-step evaluation scenario that simulates a situation which is very common in machine learning practice. In the first step, we train the "base" model. In the second step, the base model is used (a) as a feature (embedding) extractor, or (b) as a source of pre-trained weights to train the "target" model. In this scenario, we train different base models using different AS labeling options and for each such base model, we train a target model, which is trained and tested the same way in all cases. A consistent improvement in the performance of the target model achieved with a particular type of labeling being used during training of the base model would be a proof that this type of labeling is superior to others.

Our simulated two-step scenario is realized by splitting the AS dataset based on ontology. In particular, we split the AS into "target" and "base" datasets by extracting one ontology root label and all of its sub-labels and their recordings as the "target" dataset and leaving the rest as the "base" dataset. In case a recording has labels belonging to both the base and the target ontologies it is assigned to the target set and its base ontology labels are discarded.

We use alternative labelings (OL, PEL, and PE-CML) when training base models using base recordings to perform classification within the base ontology. Next, for each type of base model, we use target recordings to conduct transfer learning or train a new model based on features extracted from the base model to perform classification within the target ontology. The split between base and target ontologies is repeated multiple times for different root labels which allows us to test our hypothesis for different domains of sounds and different proportions between base and target dataset sizes. The transfer learning experiments also include an option in which the new model is trained from scratch (no knowledge transfer).

In order to increase the number of test cases and to strengthen our conclusions, the target training for each split is also performed for all three considered types of labeling. The diagram depicting the proposed experiment design is provided in Fig. 2.

V. Experiment realization

A. Dataset

An excerpt from AS was prepared by downloading all available audio recordings from the "evaluation" and "balanced train" parts of the dataset, which was further complemented by a selection of recordings from the "unbalanced train" part of the dataset. This third part of the dataset was used only to get more examples of the rare labels. Eventually, the excerpt included 102578 recordings with the least represented class ("Gargling") being represented by 107 examples.

The OL in this excerpt included an average of 2.5 labels per recording. PEL included 5.9 labels per recording and in the PE-CML an average of 28.7 labels per recording were masked.

B. Splitting

The aforementioned split between base and target datasets was repeated four times for the following root labels: (1) Human sounds, (2) Sounds of things, (3) Source ambiguous sounds and (4) Natural sounds. Proportions between base and target datasets' sizes are summarized in Table I. Each

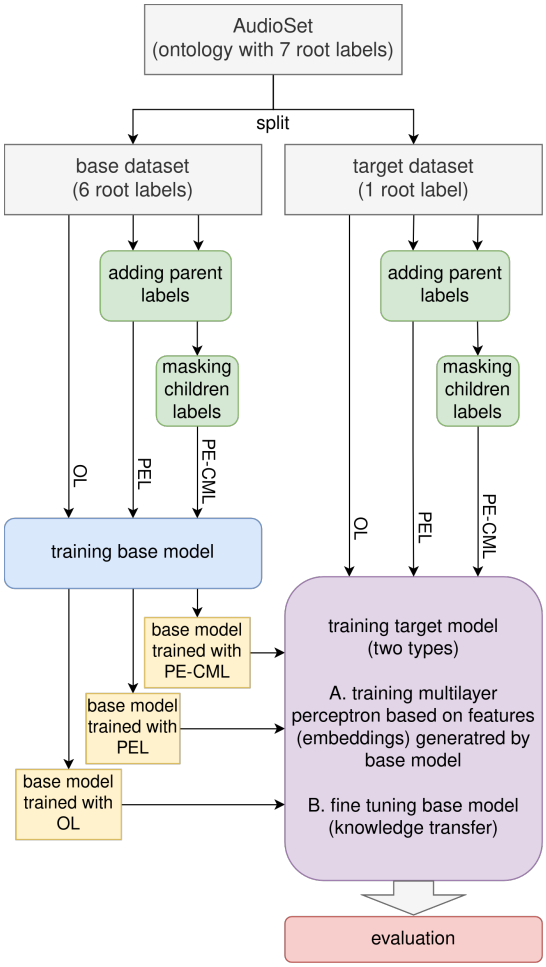

Fig. 2. Diagram depicting the experiment design (single split between base and target datasets). During the realization of the experiment, four splits were performed.

TABLE I
AUDIOSET SPLITS - NUMBER OF TARGET RECORDINGS AND TARGET LABELS FROM OL. THE NUMBER IN THE BRACKET INDICATES A PORTION OF THE AS INCLUDED IN THE TARGET DATASET (THE REMAINDER BECAME THE BASE DATASET).

Target dataset root label	Target recordings	Target labels
Human sounds	35644 (35%)	72 (14%)
Sounds of things	27329 (27%)	176 (33%)
Source ambiguous sounds	14256 (14%)	55 (10%)
Natural sounds	3168 (3%)	18 (3%)

dataset was further divided into 80% training recordings, 10% validation recordings, and 10% testing recordings.

C. Neural network

Base trainings were conducted using a deep convolutional neural network that accepts on input a vector of 80k raw audio samples (10 seconds times 8k samples per second). The network consists of 9 convolutional layers followed by two fully-connected layers and an output layer with the number of neurons equal to the number of labels in the training set. All layers except the last layer utilize ELU (Exponential Linear Unit) non-linearity and batch normalization. The last layer uses sigmoid activation. The network includes a total of 1.8M trainable parameters. A summary of the architecture of the neural network is provided in Table II.

TABLE II
ARCHITECTURE OF THE NEURAL NETWORK USED IN THE STUDY

Layer	Filters	Kernel size	Strides	Output length
Conv 1D	64	17	3	26662
Conv 1D	64	16	3	8883
Conv 1D	96	15	3	2957
Conv 1D	96	14	3	982
Conv 1D	128	13	3	324
Conv 1D	128	12	3	105
Conv 1D	192	9	3	33
Conv 1D	192	9	3	9
Conv 1D	256	9	1	1
Dense	256	-	-	-
Dense	128	-	-	-
Dense	N	-	-	-

When used for feature extraction, the base model's output from the next-to-last layer was taken without non-linearity and batch normalization (128 features). The target models that were trained on these features consisted of one fully-connected layer with 256 neurons, ELU non-linearity, and batch normalization and an output dense layer with sigmoid non-linearity. Deeper fully-connected models were also considered in this experiment but they did not provide any improvement over the two-layer network.

When used for knowledge transfer, the base model had its output layer replaced with a new layer with the number of neurons equal to the number of target labels.

D. Training

In all three cases (base model training, fine-tuning (knowledge transfer), and training of a new model based on extracted features) an Adam optimizer [15] was used with a learning rate of 0.001 and a learning rate decay of 0.987 per epoch. Trainings were conducted in batches of 24 recordings for 100 epochs. In all cases, the Mean Squared Error loss function was used with additional masking in the case of PE-CML. After each epoch mean average precision (mAP) was checked on the validation data and the model weights with the highest mAP were eventually tested on the test set which yielded the final result.

VI. RESULTS

Results for base models are presented in Table III. The performance of base models trained with PEL is significantly higher compared to those trained with OL, and PE-CML labeling provides further improvement. However, as explained in the "Experiment design" section, these results shouldn't be compared directly between columns of Table III because in each case the testing sets used different label structure. Instead,

TABLE III
PERFORMANCE (MAP) OF BASE MODELS IN THE CLASSIFICATION OF SOUNDS WITHIN THE BASE DOMAIN OF SOUNDS.

	OL	**PEL**	**PE-CML**
Base for Human sounds	0.361	0.497	0.562
Base for Sounds of things	0.469	0.595	0.646
Base for Source ambiguous sounds	0.408	0.513	0.575
Base for Natural sounds	0.410	0.504	0.557

the increased performance indicates that the re-labeling of AS data made the task easier for the neural model which hints on the improved convergence and the quality of learned features.

The results of the target experiments are summarized in Table IV. They show a clear advantage of using PEL and PE-CML during the training of base models, regardless of whether the base models are used for fine-tuning or as feature extractors.

In the case of feature extraction, training of base models with PEL resulted in an improved model performance in the target domain in all 12 cases. What's important - the improvement is present even if the target training is performed with original labels (OL), which suggests that the improvement should translate well to many different sound recognition tasks. Training of base models with PE-CML shows improvement over OL in 10 of 12 cases, however, the PE-CML shows the highest average mAP - 0.553 as compared to 0.548 achieved with PEL and 0.533 shown by OL.

In the case of knowledge transfer, the PEL base has shown only small improvement over OL base with mean mAP equal to 0.541 in the former case and 0.536 in the latter, and PEL winning in 7 categories of 12. However, the PE-CML shows improvement over the OL base in all 12 categories with mean mAP reaching 0.553. For comparison, the average performance of the model trained without knowledge transfer (initialized with random weights) was 0.499.

As one might expect, the advantage of using alternative labeling becomes more prominent in cases when the target set is smaller. For example, in the case of Human sounds (35644 training recordings) the improvement from using alternative AS labelings is under 0.01 mAP for both feature extraction and transfer learning while for Natural sounds (3168 training recordings) this improvement reaches 0.06 mAP.

VII. CONCLUSIONS

In this work two major issues with the original AudioSet labels were identified and for each an automatic re-labeling procedure was proposed that alleviates the problem. The exhaustive testing included splitting of the AudioSet into different subsets based on ontology and performing base training and two types of target training (new models trained on extracted features and fine-tinning) within different domains of sounds. Results reveal that both types of target training benefit from base training being performed using the corrected data, with an average mAP improvement of 0.02 and a maximum improvement of 0.06 mAP in cases where the target training set is small. This gives strong evidence that the proposed modified labels allow the neural network to learn more informative features, better suited as the basis for the next generation of models, especially when few target training examples are available.

Given these promising results, it might be worth to re-train the state-of-the-art audio event classification models, like PANNs, using the re-labeled AudioSet for the benefit of all derived machine learning solutions.

REFERENCES

[1] 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.

[2] "AudioSet — research.google.com," https://research.google.com/audioset/dataset/index.html, [Accessed 07-11-2023].

[3] Thodoris Kouzelis, Grigoris Bastas, Athanasios Katsamanis, and Alexandros Potamianos, "Efficient audio captioning transformer with patchout and text guidance," https://doi.org/10.48550/arXiv.2304.02916, 2023, [v1, accessed 07-11-2023].

[4] Deepank Verma, Arnab Jana, and Krithi Ramamritham, "Classification and mapping of sound sources in local urban streets through AudioSet data and bayesian optimized neural networks," *Noise Mapping*, vol. 6, no. 1, pp. 52–71, Jan. 2019.

[5] Florian Schmid, Khaled Koutini, and Gerhard Widmer, "Low-complexity audio embedding extractors," in *2023 31st European Signal Processing Conference (EUSIPCO)*. Sept. 2023, IEEE.

[6] Irina Tolkova, Brian Chu, Marcel Hedman, Stefan Kahl, and Holger Klinck, "Parsing birdsong with deep audio embeddings," https://doi.org/10.48550/arXiv.2108.09203, 2021, [v1, accessed 07-11-2023].

[7] Yuanbo Hou, Siyang Song, Cheng Luo, Andrew Mitchell, Qiaoqiao Ren, Weicheng Xie, Jian Kang, Wenwu Wang, and Dick Botteldooren, "Joint prediction of audio event and annoyance rating in an urban soundscape by hierarchical graph representation learning," in *INTERSPEECH 2023*. Aug. 2023, ISCA.

[8] Tomoya Koike, Kun Qian, Qiuqiang Kong, Mark D. Plumbley, Björn W. Schuller, and Yoshiharu Yamamoto, "Audio for audio is better? an investigation on transfer learning models for heart sound classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 74–77.

[9] Elvis Nunez, Yanzi Jin, Mohammad Rastegari, Sachin Mehta, and Maxwell Horton, "Diffusion models as masked audio-video learners," https://doi.org/10.48550/arXiv.2310.03937, 2023, [v1, accessed 07-11-2023].

[10] Yuan Gong, Yu-An Chung, and James Glass, "Ast: Audio spectrogram transformer," in *Interspeech 2021*. Aug. 2021, ISCA.

[11] 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.

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Sept. 2009.

[13] Irene Martin-Morato and Annamaria Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *2021 29th European Signal Processing Conference (EUSIPCO)*. Aug. 2021, IEEE.

[14] Eduardo Fonseca, Shawn Hershey, Manoj Plakal, Daniel P. W. Ellis, Aren Jansen, and R. Channing Moore, "Addressing missing labels in large-scale sound event recognition using a teacher-student framework with loss masking," *IEEE Signal Processing Letters*, vol. 27, pp. 1235–1239, 2020.

[15] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," https://doi.org/10.48550/arXiv.1412.6980, 2014, [v9, accessed 07-11-2023].

TABLE IV

PERFORMANCE OF TARGET MODELS PERFORMING SOUND CLASSIFICATION WITHIN GIVEN DOMAINS OF SOUNDS, TRAINED WITH ALTERNATIVE LABELS (COLUMNS) DEPENDING ON THE LABELING USED DURING TRAINING OF THE BASE MODELS (ROWS). ALL NUMBERS REPRESENT MEAN AVERAGE PRECISION (MAP). EACH COLUMN OF THIS TABLE SHOULD BE CONSIDERED AS A SEPARATE TEST CASE.

	Human sounds			**Sounds of things**			**Source ambiguous sounds**			**Natural sounds**		
target labels:	OL	PEL	PE-CML	OL	PEL	PE-CML	OL	PEL	PE-CML	O	PEL	PE-CML
Training multilayer perceptron based on extracted features (embeddings)												
base trained with OL	0.790	0.805	0.812	0.195	0.304	0.323	0.461	0.461	0.465	0.468	0.656	0.653
base trained with PEL	**0.795**	**0.809**	0.817	**0.207**	**0.314**	**0.333**	0.487	0.492	**0.494**	0.487	0.669	0.670
base trained with PE-CML	0.793	0.808	**0.820**	0.199	0.303	0.319	**0.492**	**0.497**	0.493	**0.529**	**0.690**	**0.691**
Fine tuning (knowledge transfer) or training whole model from scratch												
from scratch	0.782	0.800	**0.811**	0.139	0.310	0.343	0.385	0.396	0.440	0.411	0.596	0.579
base trained with OL	0.770	0.796	0.808	0.186	0.311	0.338	0.464	0.471	0.462	0.476	0.674	0.678
base trained with PEL	0.774	0.793	0.810	0.187	0.333	**0.358**	0.477	0.468	0.477	0.472	0.665	0.674
base trained with PE-CML	**0.783**	**0.806**	0.810	**0.191**	**0.334**	0.357	**0.483**	**0.476**	**0.486**	**0.514**	**0.693**	**0.698**