# What is the ground truth?
# Reliability of multi-annotator data for audio tagging

Irene Martín-Morató, Annamaria Mesaros
Computing Sciences, Tampere University
Tampere, FINLAND
Email: {irene.martinmorato, annamaria.mesaros}@tuni.fi

***Abstract*—Crowdsourcing has become a common approach for annotating large amounts of data. It has the advantage of harnessing a large workforce to produce large amounts of data in a short time, but comes with the disadvantage of employing non-expert annotators with different backgrounds. This raises the problem of data reliability, in addition to the general question of how to combine the opinions of multiple annotators in order to estimate the ground truth. This paper presents a study of the annotations and annotators' reliability for audio tagging. We adapt the use of Krippendorf's alpha and multi-annotator competence estimation (MACE) for a multi-labeled data scenario, and present how MACE can be used to estimate a candidate ground truth based on annotations from non-expert users with different levels of expertise and competence.**

## I. Introduction

Annotated audio is a fundamental component in training and evaluation of sound classification. Given the recent advances in environmental sound classification, including sound events classification, tagging, and detection, coupled with the use of deep learning-based solutions, availability of large datasets has become a crucial necessity. While unsupervised learning [1] or automatic methods for predicting labels [2] can provide an alternative to human-annotated data at the training stage, annotated data still plays an important role evaluation [3].

Manual annotation of audio requires repeated listening of the given sample in order to annotate it, and relying on expert annotators makes it a slow process. For this reason, crowdsourcing has emerged as an attractive method for increasing the volume of data [4]–[6]. Its disadvantage is mainly that it relies on non-expert annotators, who may provide incorrect or inconsistent labels. A common processing of such multi-annotator labels in order to create the reference annotation is to aggregate them using a majority vote (consensus), as done for example in the case of OpenMIC 2018 dataset for instrument recognition [6] or proposed for environmental sound classification in [4].

Still, because annotating audio requires both time and resources, having multiple annotators describe each data point remains relatively rare. In the audio domain, the CHIME-Home dataset [7] was obtained using three annotators, and the final annotation was created as a majority vote. Another example is the DCASE 2013 Office Live dataset that was annotated by two persons, with both annotation sets provided with the data; within the challenge, submitted systems were evaluated against each annotator separately, and then the performance was averaged [8]. Multiple expert annotators are more common in medical imaging for automatic diagnostic algorithms. Methods for fusing the expert annotator opinions include different strategies, from simple ones like intersection and union [9], to complex ones that estimate an optimal ground truth using expectation-maximization as done in STAPLE [10] or maximizing the joint agreement between annotators [11]. The method used to estimate the ground truth was found to have a significant effect on the evaluated performance of the system, with STAPLE causing underestimation of performance when only few annotations are available, and consensus overestimating it [12].

In this paper, we tackle an important research problem that has not been yet addressed in the audio domain, namely how to aggregate opinions from non-expert annotators with different backgrounds and levels of expertise in order to create a reliable ground truth for training sound classifiers. We use a subset of the publicly available TAU Urban Acoustic Scenes 2019 dataset [13] that we annotate using sound event tags. We estimate annotators' competence and inter-annotator agreement using established statistical tools, and compare different aggregation procedures for creating the reference annotation. We show that a low agreement does not necessarily reflect a low annotator reliability, instead it partly reflects the difficulty of the annotation task.

The paper is organized as follows: Section II introduces the annotator competence and agreement measures we will use in our analysis, Section III presents the data we use and the annotation process, Section IV presents the analysis of the collected data and further experiments. Finally, Section V presents conclusions and future work.

## II. Annotator and annotation analysis

We propose to adapt and employ a collection of methods that are more familiar to those working in computational linguistics. *Labeling* is the process of assigning a label to an item by an *annotator*. In our study, we deal with *multi-label annotation*, i.e. an item (in our case an audio file) is assigned one or multiple labels from a pre-defined set of labels.

The largest datasets for audio classification typically rely on user-generated material available as web audio, for which labels can be inferred from user-generated data. For example

This paper has received funding from Academy of Finland grant 332063 ”Teaching machines to listen”.

AudioSet [2] consists of automatically labeled and partially verified audio, but has an estimated label error of above 50% for 30% of its classes[1]; FSDnoisy18k [5] was crowdsourced and a subset of it was curated by experts. These are two examples showing the trade-off between accepting noisy labels and the effort necessary for curation. The reliability of the annotation process for audio data and its outcome, rarely analyzed before, is something we aim to do in this study.

### *A. Annotator competence estimation*

When a large pool of annotators that annotate partially the same data is available, the competence of these annotators can be estimated using MACE - Multi-Annotator Competence Estimation [14]. The method allows identification of trustworthy annotators and prediction of correct underlying labels, by using an unsupervised model that learns from redundant annotations.

The model considers that annotator $j$ produces label $A_{ij}$ on instance $i$. The annotated label depends on the true label $T_i$, and whether annotator $j$ is spamming (selecting the answer at random). Annotation behavior is modeled by binary variable $S_{ij}$ drawn from a Bernoulli distribution with parameter $(1-\theta_j)$. The behavior assumes that when an annotator is not spamming on instance $i$ ($S_{ij}=0$), the annotation $A_{ij}$ corresponds to the true label. When the annotator is spamming, $S_{ij}=1$, $A_{ij}$ is sampled from a multinomial distribution with parameter vector $\xi_j$. The annotations $A_{ij}$ are observed, the true labels $T_i$ and the spamming indicators $S_{ij}$ are unobserved. The model parameter $\theta_j$ specifies the probability of trustworthiness for annotator $j$, while $\xi_j$ determines the spamming behavior of annotator $j$.

The model parameters are estimated using the expectation maximization algorithm, to maximize the probability of the observed data [14]:

$$P(\mathbf{A};\theta,\xi)=\sum_{T,S}\left[\prod_{i=1}^{N}P(T_i)\prod_{j=1}^{M}P(S_{ij};\theta_j)P(A_{ij}|S_{ij},T_i;\xi_i)\right] \tag{1}$$

where $\mathbf{A}$ is the matrix of annotations, $\mathbf{S}$ is the matrix of competence indicators, and $\mathbf{T}$ is the vector of true labels. The method was shown to produce predicted labels very accurately in comparison with ground truth data on a few tasks. At the same time, the model's $\theta_j$ was shown to correlate strongly with annotator proficiency [14].

We use MACE to study the behavior of our annotators and to predict different sets of aggregated labels. It is important to note that MACE does not discard annotators, but weighs their opinion based on their competence, which results in a different procedure than majority voting which trusts and weighs all annotators equally.

### *B. Inter-annotator agreement*

Many measures that assess inter-annotator agreement are developed for only two annotators. In addition, simple measures like percentage of agreement or correlation suffer from various biases related to chance agreement and statistical independence of annotators and annotated data [15]. We select for our analysis Krippendorff's alpha as a general agreement metric that is able to cope with more than two annotators per item and with missing data (overlap in annotated items only among few of the annotators).

Krippendorff's alpha is defined as:

$$\alpha = 1-\frac{D_o}{D_e} \tag{2}$$

where $D_o$ is the observed disagreement and $D_e$ is the expected disagreement. In the case of multiple annotators $m$, multiple nominal categories, and missing values, the formulation uses *nominal* $\alpha$ defined as [16, pp.230-231]:

$$_{nominal}\alpha = 1-(n-1)\frac{n-\sum_c o_{cc}}{n^2-\sum_c n_c^2} \tag{3}$$

where:

$$o_{ck}=\sum\nolimits_u \frac{\text{Number of } c\text{-}k \text{ pairs in } u}{m_u-1} \tag{4}$$

is the number of observed coincidences of two categories $c$ and $k$ assigned to the same item by two different annotators, and $m_u$ is the number of values assigned to item $u$ (number of annotators that labeled this item). Observed coincidences are calculated based on a coincidence matrix that considers the annotators interchangeable, therefore pairing the contingencies in both directions: if $x_{ck}$ is the number of times a particular observer uses $c$ while the other uses $k$, then the number of coincidences is $o_{ck}=x_{ck}+x_{kc}$.

Krippendorff's alpha is reported in [4] for annotation of audio, but no discussion on its values for the annotated data is provided. Krippendorff's alpha has been also used to measure inter-annotator agreement for images [17] and video annotations [18].

## III. Multi-annotator data collection

The dataset used in our experiments is a subset of TAU Urban Acoustic Scenes 2019 [13], consisting of audio from three acoustic scenes (airport, public square, and park). The audio files are 10 seconds long, and some of them are consecutive segments of one long recording from a single location. Each audio file was annotated by five different annotators, following a single-pass multi-label annotation procedure [4], in which the annotator selected for one audio file a number of labels presented as a list. Candidate labels were *birds singing, dog barking, adults talking, children voices, traffic noise, music, footsteps, siren, announcement speech* and *announcement jingle*. According to the annotation procedure, a positive label is explicitly provided by the annotator, while a negative label is implicit, by not being selected.

The annotation platform was a simple web-based interface that presented users with audio files to annotate one by one. Annotators registered using their email, which permitted the annotation process to be paused and resumed later on next login. A set of instructions and examples of annotation were provided at the beginning. Annotators were instructed to work

[1] See https://research.google.com/audioset/dataset/index.html for an explanation of the quality assessment. Information accessed January 2021

TABLE I
ANNOTATION MATRIX EXAMPLE WITH EXPLICIT/IMPLICIT ANNOTATIONS PRODUCED BY $m$ ANNOTATORS

| items (file, label) | 1 | 2 | 3 | 4 | .. | m |
|---|---|---|---|---|---|---|
| airport-Paris-0, footsteps | 1 | 1 | 0 | 1 | ... | - |
| airport-Paris-0, adults-talking | 0 | 0 | 0 | 1 | ... | - |
| airport-Paris-0, dog-barking | - | - | - | - | ... | 1 |
| ... | ... | ... | ... | ... | ... | |
| airport-Helsinki-4, footsteps | 1 | 1 | 1 | - | ... | - |

in small batches and to use good quality headphones. It was allowed to listen to each audio file multiple times before selecting one or more of the candidate labels. A total of 133 annotators, students taking an audio signal processing course, were randomly assigned a maximum of 131 files to annotate. The total number of files annotated is 3930. Annotators were assigned into 30 groups, aiming that each group will provide annotations to the same set of files.

Because neither MACE nor the inter-annotator agreement metrics are defined for multi-labeled items, we represent the annotations as a set of binary *yes/no* labels per file, with explicit/implicit presence as explained before. In consequence, each (file, label) pair is considered an independently annotated item, equivalent to a multiple-pass binary annotation [4]. However, because the annotation process did not request producing the labels themselves, we consider that this assumption has sufficient grounds. Complete annotations are represented as a matrix containing the answers of all annotators, illustrated in Table I. Each row refers to a (file, label) item, and each column represents the answer of one annotator in the format $[0, 1, -]$, marking the presence (1, explicit) or absence (0, implicit) of this label within the audio file; ”$-$” indicates that this file was not assigned to this specific annotator. The resulting matrix contains a total of 39300 items.

## IV. DATA ANALYSIS

We first consider the aggregation of multiple annotations. The simplest one, union, assigns a label to a file if at least one of the annotators has considered it active. The most commonly used aggregation method, majority vote, assigns a label to a file if most annotators have considered it active. The statistics of the resulting classes are presented in Table II for the individual classes (first two columns), while Fig. 1 shows the resulting number of labels per file. The resulting annotations are largely unbalanced, with the most common label *adults talking* being assigned to 3168 files, and least common *announcement jingle* to 116 files. Majority voting reduces their frequency in the resulting annotation to 2401 and 8, respectively.

### A. Predicting ground truth with MACE

As explained, MACE predicts the true labels by estimating the annotators' competences and their effect on the true labels within the same model. This creates a weighing procedure on the different opinions which is dependent on how trusted the respective annotator is. In addition, the produced estimation for ground truth can be constrained using a threshold $n$, with MACE@n containing the $n\%$ of predicted labels for which the method is most confident [14].

TABLE II
STATISTICS OF CLASS LABELS RESULTING FROM DIFFERENT METHODS OF COMBINING THE MULTIPLE ANNOTATIONS

| class labels | union | maj. vote | MACE | MACE@90 |
|---|---|---|---|---|
| adults talking | 3168 | 2401 | 2983 | 2831 |
| footsteps | 2560 | 859 | 1969 | 1583 |
| traffic noise | 2418 | 680 | 1713 | 1178 |
| children voices | 1467 | 513 | 1046 | 821 |
| birds singing | 1332 | 672 | 1035 | 855 |
| music | 306 | 106 | 212 | 174 |
| ann. speech | 273 | 73 | 148 | 108 |
| dog barking | 177 | 42 | 108 | 79 |
| siren | 177 | 38 | 99 | 61 |
| ann. jingle | 116 | 8 | 38 | 16 |

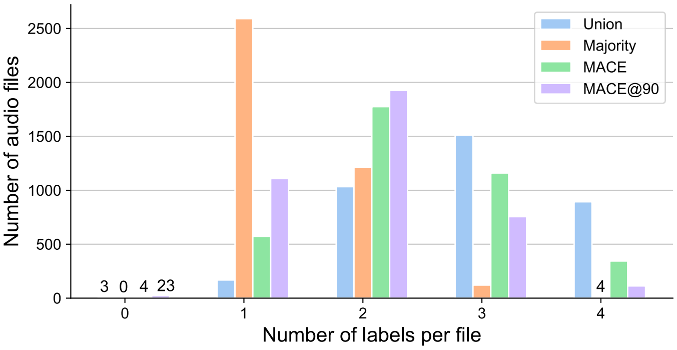

Fig. 1. Average number of labels per file produced by different approaches to estimate the true labels: union, majority vote, MACE, and MACE@90

The resulting statistics of the estimated ground truth in terms of number of produced labels and number of labels per file are presented in Table II and Fig. 1 for comparison with the union and majority vote. With MACE, the number of labels estimated for the ground truth is significantly higher than using majority vote for all categories, indicating that for some cases a minority of annotators is reliable enough to justify the label. Even when eliminating the least confident 10% of predictions (MACE@90), the number of resulting labels is higher than with the majority vote, showing that this method has the potential to overcome the problem of missing labels caused by an insufficient number of votes, which can cause label noise for learning [19].

### B. Annotator competence analysis

The estimated competence of our annotators, obtained using MACE, is illustrated in Fig. 2. We observe that there are a number of annotators that are highly trustable (64 over 0.8), while a small number of them have much lower estimated competence. Even though the annotators do seem mostly reliable, the agreement on the labels is not very high, with Krippendorff's alpha for the entire dataset being 0.696.

We hypothesize that inter-annotator disagreement can come from two sources. One is the annotator competence: an annotator who does not pay attention to the task and completes it at random will not have high agreement with an annotator who is very diligent about the task. A second source of disagreement

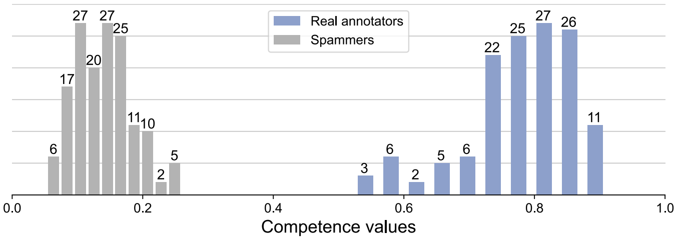

Fig. 2. Annotator competence estimated using MACE

is the annotator's personal experience and perception. Experiments in cognitive psychology have shown that life experience brings much subjectivity in categorization [20]. The audio data in our experiment is recorded in the wild, with no control over the sound sources present, their prominence in the scene and their overlaps, which makes it rather difficult to annotate and allows personal interpretation. In addition, some studies have shown that visual stimuli help with audio annotation [21], but our experiment did not provide any visual information.

In absence of the gold standard (which would allow us to estimate the upper bound for agreement and evaluate the estimated ground truth), we simulate the lower bound. We simulate a group of spammer annotators for the task, that provide *yes/no* indicators per file for the set of 10 labels. We create 150 random annotators, with each being randomly assigned a number of 130 files from the set of 3930 available. We then analyze their output in terms of labels statistics, majority vote, MACE competence, and inter-annotator agreement metrics.

The estimated competence of these spammers, presented in Fig.2, shows that even though the real annotators disagree on the labels, they are in fact diligent and not answering at random. The distribution of labels per file for the random annotators is much more uniform, and inter-annotator agreement (Krippendorff's alpha) is practically 0. This suggests perception differences as being the main cause of disagreement of annotators, but because we cannot separate the effects of the two in the data, we cannot draw a definite conclusion.

### C. Inter-annotator agreement and improving data reliability

Krippendorff's alpha was calculated for the overall data, and separately for each class. The results are presented in Table III. We observe a wide variation in the class-wise agreement, with the highest agreement on the more rare *dog barking* class. On the other hand, the more frequent classes *footsteps* and *traffic noise* have similar frequency in our data but very different agreement values. Their different acoustic characteristics also indicate perception as a reason for disagreement, as explained in the previous paragraph.

A straightforward way to improve the data reliability when we have knowledge about annotators competence is to eliminate annotations produced by the least trusted annotators, in order to obtain a set of annotations which is produced by the most reliable ones. Of course, gradual elimination of annotators will result in a reduced set of annotations available, until the extreme case of having only a single annotator. Table III presents the calculated alpha for the case of only using annotators with estimated competence of at least 0.6 (124 annotators) and at least 0.8 (64 annotators).

TABLE III
KRIPPENDORFF'S ALPHA FOR SELECTED SUBSETS OF ANNOTATIONS

| class-wise | all annot | competence> 0.6 | competence> 0.8 |
|---:|:---:|:---:|:---:|
| adults talking | 0.676 | 0.690 | 0.717 |
| footsteps | 0.271 | 0.284 | 0.236 |
| traffic noise | 0.590 | 0.607 | 0.635 |
| children voices | 0.712 | 0.714 | 0.729 |
| birds singing | 0.613 | 0.619 | 0.657 |
| music | 0.606 | 0.615 | 0.679 |
| ann. speech | 0.485 | 0.501 | 0.548 |
| dog barking | 0.713 | 0.730 | 0.764 |
| siren | 0.550 | 0.569 | 0.624 |
| ann. jingle | 0.404 | 0.430 | 0.525 |
| **overall** | **0.696** | **0.708** | **0.745** |

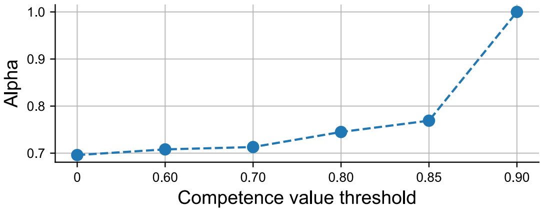

Fig. 3. Alpha values when gradually removing annotators with competence values under a given threshold, until only one is left

All agreement values increase when using the more reliable annotators, except for *footsteps*, while the overall agreement increases significantly when using the top annotators. Figure 3 shows the evolution of $\alpha$ when annotators under a given competence are gradually eliminated, with the final case being a single annotator. As a comparison, we note that standards adopted in social sciences consider a 0.8 agreement reasonable, and consider values between 0.667 and 0.8 only for drawing tentative conclusions. However, Krippendorff argues that the acceptable level of agreement must be chosen depending on the costs of drawing invalid conclusions [16, p. 241]. Therefore we can state that the employed methods allow creating reference annotations that can be trusted for training and evaluation of acoustic models, compared to noisy data[2].

## V. CONCLUSIONS AND FUTURE WORK

This paper presented a study of annotator and annotations reliability for crowdsourced audio tags. We showed that the aggregation of raw multi-annotator labels using annotator competence estimation produces a plausible and trustable ground truth, with gradually improving levels of agreement in the data. However, in our experiment we cannot evaluate the correctness of the estimated ground truth. For this reason, we plan to repeat this experiment in controlled conditions, using generated synthetic data for which ground truth is produced at the same time with the audio. We will try to mimic as closely as possible the classes and acoustic characteristics of the data used in the presented experiment.

[2] The MATS (Multi-Annotator Tagged Soundscapes) dataset is available at https://doi.org/10.5281/zenodo.4774959

# References

[1] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised contrastive learning of sound event representations," *arXiv preprint arXiv:2011.07616*, 2020.

[2] J. F. Gemmeke, D. P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[3] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Springer, 2018, ch. 6, pp. 147–179.

[4] M. Cartwright, G. Dove, A. E. Méndez Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–11.

[5] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 21–25.

[6] E. Humphrey, S. Durand, and B. McFee, "OpenMIC-2018: An open data-set for multiple instrument recognition." in *ISMIR*, 2018.

[7] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "CHiME-Home: A dataset for sound source recognition in a domestic environment," in *Proc. of the 9th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[9] T. Kauppi, J.-K. Kamarainen, L. Lensu, V. Kalesnykiene, I. Sorri, H. Kälviäinen, H. Uusitalo, and J. Pietilä, "Fusion of multiple expert annotations and overall score selection for medical image diagnosis," in *Image Analysis*, A.-B. Salberg, J. Y. Hardeberg, and R. Jenssen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 760–769.

[10] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[11] J.-K. Kamarainen, L. Lensu, and T. Kauppi, "Combining multiple image segmentations by maximizing expert agreement," in *Machine Learning in Medical Imaging*, F. Wang, D. Shen, P. Yan, and K. Suzuki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 193–200.

[12] T. A. Lampert, A. Stumpf, and P. Gançarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2557–2572, 2016.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, Nov 2019.

[14] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1120–1130.

[15] K. Krippendorff, "Agreement and information in the reliability of coding," *Communication Methods and Measures*, vol. 5, no. 2, pp. 93–112, 2011.

[16] ——, *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, 2004.

[17] J. Nassar, V. Pavon-Harr, M. Bosch, and I. McCulloh, "Assessing data quality of annotations with Krippendorff alpha for applications in computer vision," *arXiv preprint arXiv:1912.10107*, 2019.

[18] S. Park, G. Mohammadi, R. Artstein, and L.-P. Morency, "Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface," in *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, ser. CrowdMM '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 29–34.

[19] E. Fonseca, S. Hershey, M. Plakal, D. P. W. Ellis, A. Jansen, and R. C. 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.

[20] C. Guastavino, "Everyday sound categorization," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Springer, 2018, ch. 7, pp. 183–213.

[21] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.