Sound event detection with soft labels: a new perspective on evaluation

Manu Harju, Irene Martín-Morató, Toni Heittola, Annamaria Mesaros
Signal Processing Research Centre, *Tampere University*, Tampere, Finland
{manu.harju, irene.martinmorato, toni.heittola, annamaria.mesaros}@tuni.fi

***Abstract*—Sound event detection has been an essential task in the DCASE Challenge since the beginning, with various alterations over the years. The 2023 Challenge presented for the first time a sound event detection task for which the reference labels representing sound class activity were provided as real numbers on the interval from zero to one, in addition to binary labels. In this paper we provide an overview of the sound event detection with soft labels task in DCASE 2023 Challenge, and re-evaluate the challenge submissions using a soft metric. The use of a soft metric allows computing precision, recall and F-score directly using the soft labels, and thus avoids the optimization step for binarizing both the reference and predictions using a threshold. We analyze the behavior of the soft metric on a large number of systems, and show that for the softly labeled reference data, the results obtained with the soft metrics represent very well the system's ability to follow the data, and is a good proxy for entropy-based measures.**

***Index Terms*—sound event detection, soft labels**

I. Introduction

Sound event detection (SED) is defined as the task of providing the event class and the event time localization given that multiple events can be present in an audio recording. SED has been a fundamental task of the DCASE Challenge. Throughout the years, the task has seen various modifications and different data, and evolved from a basic setup with training data available as strongly-labeled (label and temporal information available for each event instance) [1], [2] to being trained with weakly-labeled real data (only label, no temporal information) and finally to unlabeled data [3], [4]. A major reason driving these changes was the scarcity of strongly-labeled data, with labeling procedures often complex, time-consuming and expensive. Efforts in labeling real-life data with strong labels have produced small datasets [4], [5]; the largest dataset with strong labels is currently AudioSet [6], containing annotations for approximately 117k clips with a length of 10 seconds.

One way to solve the annotation problem is to distribute the workload by using crowdsourcing platforms like Amazon Mechanical Turk. However, on such platforms it is beneficial to keep the individual annotation tasks as simple as possible, and strong labeling is not simple. A method for obtaining temporally strong labels was recently proposed in [7] through collecting multiple temporally weak labels for short, overlapping excerpts of the audio with a small stride. The multiple opinions were aggregated by weighted averaging [8], resulting in temporally strong annotation with a time resolution determined by the stride, and sound event activity expressed as a number between 0 and 1, i.e. a *soft label*. Soft labels can also be obtained for example by using mixup or label smoothing data augmentation procedures. Furthermore, soft labels can represent the uncertainties in the data, similar to the outcome obtained in [7].

In real life applications, hard labels are often more useful than soft labels, as it is beneficial to get a clear indication when an event is occurring. However, the model outputs are typically from some continuous scale, and setting up the system usually includes choosing a threshold value for binarizing the outputs into positive and negative predictions. The operating point determined by this threshold should be selected according to the tolerance to different error types [9], with the best operating point ideally chosen based on some development data. Recently, a metric for jointly computing system performance on an evaluation set for all possible thresholds was proposed [10], in an effort to create a threshold-independent metric. The metric also allows selecting the threshold which best fulfills the requirements of a given application. However, when the reference annotation is softly labeled, this needs to be binarized too into hard labels, in order to use the standard evaluation metrics. Most commonly, hard labels are produced as a majority vote (threshold of 0.5); Morato et. al. [8] recently showed that using the mid-range of the soft labels also produces a reasonable set of hard labels.

In this work, we propose the use of a soft definition for evaluating SED systems directly against soft labels, hence avoiding the problem of choosing a threshold value for both the reference annotation and the system predictions, and the noise introduced by the quantization. We recently introduced a novel definition for soft precision and recall using fuzzy set theory [11], and showed that, for a small number of models, the ranking based on soft metrics agrees well with the ranking based on KL-divergence, and confidence intervals for the soft metrics are smaller than for the hard metrics. We build upon that work through a comprehensive investigation on a large number of different systems in different task setups.

The contributions of this paper are as follows: (1) we re-evaluate the submissions of DCASE 2023 Challenge Task 4B: Sound event detection with soft labels using the proposed soft F-score, examining the differences in the metrics and resulting rankings, to understand the behavior and best use of the soft

This work was supported by Academy of Finland grant 332063 "Teaching machines to listen".

metrics; (2) we re-evaluate the submissions of Task 4A: Sound Event Detection with Weak Labels and Synthetic Soundscapes, to confirm that the proposed metrics work consistently with the existing metrics in case of hard reference labels; (3) we highlight the benefits of using a soft metric when the reference annotation is soft, and provide further recommendations for evaluation and ranking.

The paper is organized as follows: Section II presents the task setup with soft labels, data and baseline system; Section II-B introduces the metric definitions for the softly-labeled data; Section III presents the re-evaluation of the sound event detection systems using the soft metric, and discusses the differences between the official ranking metrics and the soft F-score in Section III-C. Finally, Section IV presents conclusions and future work.

II. Sound Event Detection with Soft Labels

In 2023 DCASE Challenge, the Sound Event Detection task had for the first time a subtask that focused on training systems using softly labeled data. Subtask A had the familiar setup, with training data consisting of real data with temporally weak labels, unlabeled real data, and synthetic data with temporally strong labels in binary format. In comparison, Subtask B training data consisted of a single dataset with temporally strong soft labels having clear temporal boundaries, and non-binary activity values for the sound event instances. An illustration of soft labels per 1s-segment and (binarized) corresponding hard labels is presented in Fig. 1.

The main goal of the subtask was to study if the information in soft labels can be leveraged in training of SED systems. In addition to the provided softly labeled dataset, the participants were allowed to use external data and embeddings, but the task rules required that the soft labels are used in training of the systems. Furthermore, it was allowed to also use hard labels in the training, and the provided dataset also included hard labels obtained using a threshold of 0.5.

A. Dataset, baseline, and challenge setup

The challenge used the MAESTRO Real dataset [7] which consists of approximately 3-minute long clips of audio recorded in real-life acoustic scenes. The annotation was done by crowdsourcing temporally weak labels for 10-second audio segments of the long recordings, overlapping with a 1-second stride to provide a dense annotation of the entire length of the clip. The obtained annotations were processed using MACE [12] to estimate annotators' competence, which was then used as a weight factor in the aggregation of multiple opinions [8]. The resulting value is regarded as a soft label, with each 1-second audio clip being processed separately based on the individual opinions available; hence, the temporal resolution of the resulting soft labels is one second. The development dataset contains 49 files (190 minutes in total), while the evaluation data consists of 26 files (97 minutes), with 17 event classes. The development set also includes a five-fold cross-validation setup.

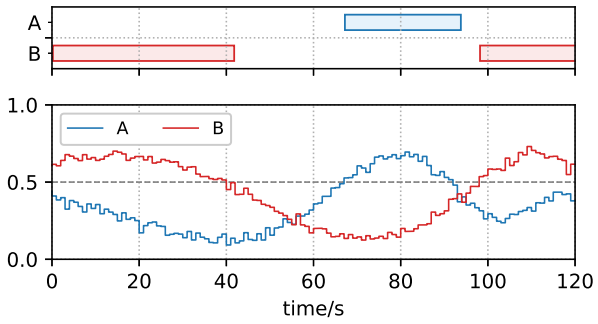

Fig. 1. Illustration of soft labels and the corresponding hard labels obtained by using a 0.5 threshold value.

The challenge provided a standard CNN baseline system comprising three convolutional blocks, each consisting of a 2D convolution with 128 filters and 3×3 kernels, batch normalization, ReLU, dropout and max pooling in feature dimension. The CNN part is followed by a single bidirectional GRU layer of 32 units. Outputs were predicted for all the 17 labels present in the data, and the system was trained with a regression setup using MSE loss [8].

For evaluation, the reference data was binarized using a 0.5 threshold. However, six classes have no soft labels above the threshold, or the amount of such segments is very small; for this reason, the official challenge evaluation and ranking was done using only the other 11 event classes. The submissions were evaluated and ranked according to optimal threshold (OT) macro F-scores [10], which finds the optimal class-wise threshold values for binarizing the predictions, such that the per-class F-score is maximized, resulting in a best possible overall F-score value.

B. Evaluation using soft metrics

To evaluate the system outputs directly against the soft labels we use an extension of precision, recall and F-score, as introduced in [11]. The method is based on interpreting predictions and references as fuzzy sets of the 1-second segments, where the soft labels of each segment are interpreted as the membership values of the segment belonging to the sets of predictions or references. We use the standard set theoretical definitions for precision, recall and F-score. In particular, here we use the standard fuzzy set intersection, namely the minimum of the membership values.

More formally, for the i'th segment let $\hat{y}_i$ and y_i be the predicted value and the reference label value, respectively. Assume that y_i and $\hat{y}_i$ are bounded on the unit interval. We use the following definitions for the soft precision, recall, and F_1-score:

$$\text{Precision} = \frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i \hat{y}_i}, \quad (1)$$

$$\text{Recall} = \frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i y_i}, \quad (2)$$

$$\text{F}_1 - \text{score} = 2\frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i (\hat{y}_i + y_i)}. \quad (3)$$

The formulation allows calculating the scores for non-binary values, and coincides with the well-established definitions in case of binary labels. Extensive details on the metric behavior are provided in [11].

We use this metric to re-evaluate the DCASE 2023 Challenge Task 4 B submissions. We evaluate the soft predictions directly against the reference soft labels, and compare the results with the scores obtained by binarizing the reference data and using the optimal threshold metrics. We also re-evaluate the submissions in Task 4 A, where the reference is provided directly as hard labels.

III. Re-evaluation of Challenge Entries

A. Soft evaluation for soft reference labels

The soft labels task received 22 submissions from 7 teams; some systems relied on pretrained models or precomputed embeddings, most commonly BEATs [13] and AST [14]. None of the participants used any other data than the provided dataset, but data augmentation techniques like mixup [15], SpecAugment [16] and oversampling proved to be useful for improving model performance.

We evaluated all submissions using the soft F-score defined in Section II-B, to compare them with the official task evaluation setup, aiming to introduce a better way to measure performance of SED. We also estimated 95% confidence intervals for both variants of F-score using the jackknife [17] procedure, by performing leave-one-out using one-second segments, as they are the evaluation unit sample. The results are presented in Table I. For brevity we only list the scores for the best system of each team, with the teams ordered according to their official challenge rank. Team names in the table correspond to the format on the challenge website. The rankings for all the systems in both evaluations are plotted in Fig. 2, and 95% CI for the top ranked teams are presented in Fig. 3. Fig. 2 shows that the best systems of the OT evaluation are placed around midrange in the soft evaluation, and a number of systems get a small increase in the soft ranking, placing slightly above the diagonal in the figure. Fig. 3 shows that the confidence intervals for the OT F-score overlap, whereas for the soft F-score they are better separated and much smaller.

A closer look at the performance indicates that the biggest changes in the rankings were caused by changes in the precision/recall balance. In the OT evaluation only the shape of the system output matters, as the evaluation method is finding the best threshold. However, in the soft metric evaluation the system outputs actually need to follow closely the reference labels in order to get good scores, and this can result in decreased rank in soft evaluation.

B. Soft evaluation for hard reference labels

Task A received submissions from 15 teams, and the participants were allowed to submit 8 systems per team. Out of the 82 submitted systems, 43 were ensembles, whereas 39 submissions were single model systems. We re-evaluated the submissions on the 1051 ten-second audio clips of the evaluation set, containing data for ten event classes, calculating

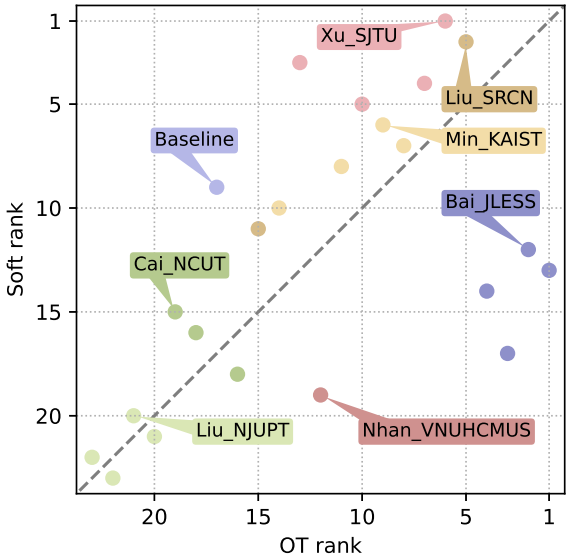

Fig. 2. Ranking of the systems in both evaluations. Color indicates systems of one team. Systems placed above the diagonal are ranked higher in the soft metric evaluation, whereas those under the diagonal are getting a lower rank.

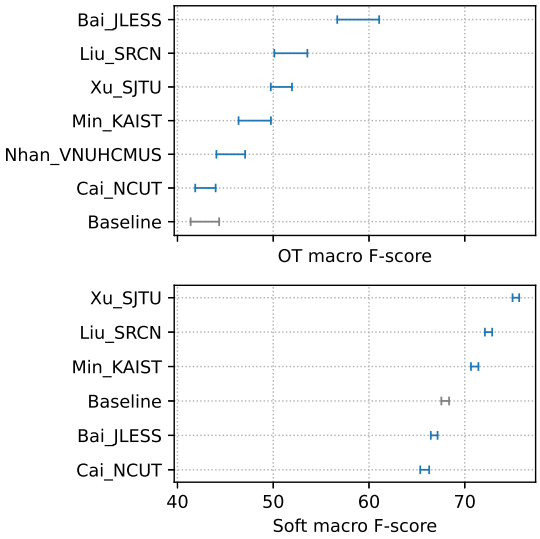

Fig. 3. 95% confidence intervals for the best performing systems of the top ranked teams in both evaluations.

the soft metrics at the resolution provided by each submission. Confidence intervals were computed using jackknife and leaving out one-second segments.

The results of the re-evaluation are well aligned with the ones from the soft labels. Major changes in the ranks are caused due to the precision/recall imbalance, whereas the overall ranking is mainly preserved. However, in this task the differences between the best performing systems' scores are comparatively small, thus the changes in the ranks are more drastic. This can be seen in Figure 4, where the single model system ranks are plotted for the official rank and a soft F1-score based rank. The ensembles are left out for clarity.

TABLE I

TEAM RANKING ACCORDING TO THE OPTIMAL THRESHOLD (OT) HARD MACRO F-SCORE ALONG WITH THE SOFT SCORES AND RANKS. F-SCORES MARKED WITH * HAVE NO OVERLAP IN THE CONFIDENCE INTERVALS WITH THE NEXT TEAM. OT P AND OT R ARE PRECISION AND RECALL IN THE OT EVALUATION. SOFT P AND SOFT R DENOTE SOFT PRECISION AND RECALL CALCULATED ACCORDING TO EQ. (1).

OT rank	Team	OT F_M	OT P	OT R	Soft rank	Soft F_M	Soft P	Soft R
1	Bai_JLESS	58.9 ± 2.2*	63.2 ± 3.6	57.0 ± 2.4	5	66.8 ± 0.4*	83.3 ± 1.2	58.7 ± 0.3
2	Liu_SRCN	51.8 ± 1.7	46.4 ± 1.7	64.6 ± 7.1	2	72.5 ± 0.4*	72.9 ± 0.5	72.4 ± 0.5
3	Xu_SJTU	50.9 ± 1.1	44.3 ± 1.0	80.0 ± 14.0	1	**75.3 ± 0.3***	76.4 ± 0.5	74.6 ± 0.4
4	Min_KAIST	48.1 ± 1.7	46.1 ± 2.0	58.0 ± 4.7	3	71.0 ± 0.4*	71.3 ± 0.5	71.2 ± 0.4
5	Nhan_VNUHCMUS	45.6 ± 1.5*	42.4 ± 1.5	53.0 ± 4.6	7	32.7 ± 0.1*	40.3 ± 0.4	32.3 ± 0.2
6	Cai_NCUT	42.9 ± 1.1	39.1 ± 1.0	68.9 ± 11.8	6	65.8 ± 0.5*	64.4 ± 0.6	68.6 ± 0.6
7	Baseline	42.9 ± 1.5*	37.2 ± 1.4	73.5 ± 12.6	4	68.0 ± 0.4*	63.8 ± 0.5	75.3 ± 0.6
8	Liu_NJUPT	23.7 ± 1.1	20.0 ± 7.1	61.2 ± 6.5	8	30.8 ± 0.4	42.3 ± 4.6	32.6 ± 0.6

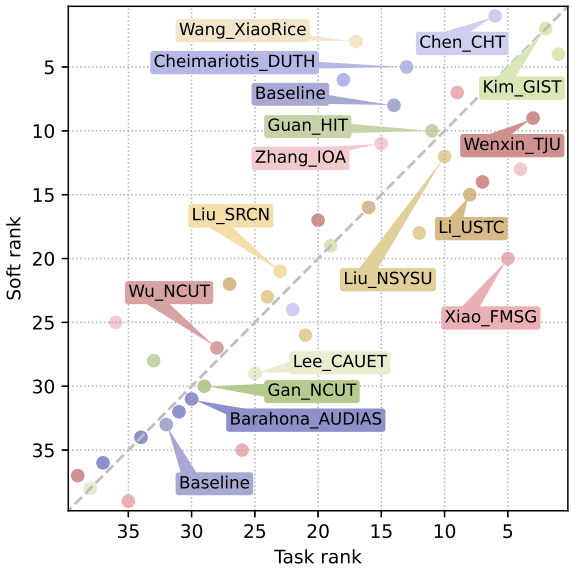

Fig. 4. Ranking of Subtask A single model systems with official and soft evaluation. Color indicates systems of one team. Systems placed above the diagonal are ranked higher in the soft metric evaluation, whereas those under the diagonal are getting a lower rank.

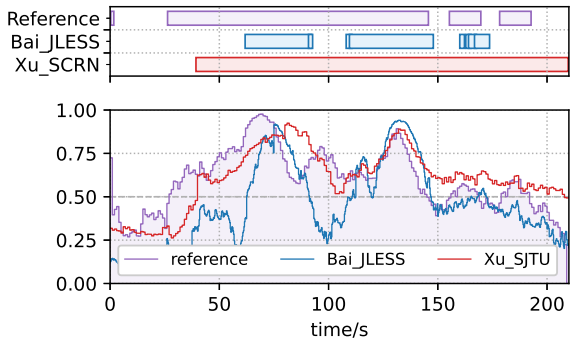

Fig. 5. Predictions of the best systems in each ranking for reference labels of sound class "car" in a single audio file. The hard labels in the upper panel are obtained by using the found optimal thresholds for the predictions and 0.5 threshold for the reference data.

C. Discussion

While the OT evaluation used in Task B tries to find the best F-scores, it needs to balance between precision and recall. The optimization problem is not symmetric: for any system the threshold can be always made lower to improve recall, whereas the maximum value of precision depends on the actual system outputs. For most systems, the OT evaluation chose a threshold maximizing recall; this makes it the more unstable component in the leave-one-out evaluation, noticeable in Table I as large confidence intervals for OT R. Furthermore, the binary evaluation methods (OT included) require hard reference labels, and the common assumption is that a 0.5 threshold (*i.e.,* majority vote) is the correct value for all the data. However, experiments using soft labels showed that it could be justified to use class-specific values, e.g. the mid-range of the class-wise soft labels [8], as otherwise the soft labels may not exceed the 0.5 threshold for some classes.

With no consensus on a most suitable way to aggregate and binarize information at the annotation stage, it is not at all obvious what is the correct (or optimal) threshold value; these issues simply carry over to the evaluation stage and inadvertently affect the measured performance. While thresholding of predictions can be avoided in the evaluation through use of ROC or log-loss based metrics, the reference is typically binary. We illustrate one such example in Fig. 5, comprising hard and soft predictions and reference labels for the class "car" in one audio clip. At the end of the clip the reference fluctuates around the threshold value, resulting in two different event instances in the hard labels, even though the soft values are not so different.

We calculate the Spearman rank correlations [18] between different ranking methods for all the systems in subtask B, to measure the agreement of different evaluation methods; The numbers are shown in Table II. MSE and KL-divergence (KLD) based rankings are computed directly using the soft labels. The ranking obtained by using the soft F-score correlates very strongly with the KLD and MSE rankings (0.90 and 0.87, respectively). However, the rankings computed against the binarized reference data are less correlated with the KLD and MSE rankings (0.76 and 0.75, respectively), indicating that the soft F-score is better suited for measuring the system's ability to follow the reference labels.

Ideally, we want to evaluate how closely the reference and predictions match, but MSE and KLD lack a clear interpretation as metric. On the other hand, the proposed soft F-score has an intuitive interpretation, its value is between 0 (only errors)

TABLE II
SPEARMAN RANK CORRELATIONS FOR RANKING METHODS IN TASK B.

	MSE	KLD	Soft F_M	OT F_M
MSE	1.00			
KLD	0.96	1.00		
Soft F_M	0.87	**0.90**	1.00	
OT F_M	0.75	0.76	0.56	1.00

TABLE III
SPEARMAN RANK CORRELATIONS FOR RANKING METHODS IN TASK A.

	Soft F_M	Score	PSDS1	PSDS2
Soft F_M	1.00			
Score	**0.73**	1.00		
PSDS1	0.71	0.99	1.00	
PSDS2	0.26	0.43	0.33	1.00

and 1 (no errors), and is already familiar to those working with classification, recognition and retrieval, which makes it an excellent candidate for the task. Its independence from any threshold, both at the reference and at the system prediction end, is an advantage compared to the hard metrics, where part of the information is always lost through thresholding.

Finally, the soft F-score can also be used with binary reference labels. When all metrics, including KL-divergence, are calculated using soft predictions and hard reference, the rank correlations are 0.88 for KLD and soft F-score, and 0.92 for KLD and OT F-score. While the OT method yields a slightly higher number, both methods show highly correlated rankings to KLD, confirming that the soft F-score represents well the performance also when the reference is given as hard labels. The soft F-score reduces to the standard F-score if all the data in the evaluation process is in binary format.

We also examine the overall behavior of the soft F-score on the systems submitted to subtask A, in which the ground truth is available as strong hard labels, by calculating the Spearman rank correlation between the metrics used in the official ranking; "score" is a combination of PSDS1 and PDSD2 [10]. As shown in Table III, the soft F-score is highly correlated with the ranking score and with PSDS1, which measures the ability of the systems to react fast. Furthermore, for the single model systems the rank correlation between the soft macro F-score and the task ranking metric increases to 0.85. At the same time, we observe that PSDS1 and PSDS 2 are quite weakly correlated, and PSDS1 is very highly correlated with the official ranking score, effectively driving the ranking of the systems. Nevertheless, it shows that the soft F-score agrees with a well-established SED metric.

IV. CONCLUSIONS

This work presented the re-evaluation of an extensive set of sound event detection systems using a metric tailored for soft labels, but which works equally well for hard labels. The soft F-score provides a method to evaluate system outputs directly against soft reference labels, leaving out the selection of the reference label threshold values. In comparison with the hard metrics, the soft F-score yields a ranking that is more correlated with the MSE and KLD rankings, making it a preferable method for measuring the systems' ability to track the reference labels. This is valid with hard labels too, as the soft F-score is correlated to the temporally strict polyphonic detection score. Finally, the confidence intervals for soft F-score are shorter than for the OT metric and contain less overlap, conveniently indicating statistically significant differences without additional measurements.

REFERENCES

[1] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M.D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[2] 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 TASLP*, vol. 27, no. 6, pp. 992–1006, 2019.

[3] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 200–204.

[4] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "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 Workshop (DCASE2019)*, New York City, United States, October 2019.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conf. (EUSIPCO)*, 2016, pp. 1128–1132.

[6] S. Hershey, D. Ellis, E. Fonseca, A. Jansen, C. Liu, C.R. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[7] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[8] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Training sound event detection with soft labels from crowdsourced annotations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[9] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M.D. Plumbley, and D. Ellis, Eds. 2018, pp. 335–371, Springer Int. Publishing.

[10] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022.

[11] M. Harju and A. Mesaros, "Evaluating classification systems against soft labels with fuzzy precision and recall," in *Proc. of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 46–50.

[12] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy, "Learning whom to trust with MACE," in *Proc. of the 2013 NAACL: Human Language Technologies*, Atlanta, Georgia, June 2013, pp. 1120–1130, Association for Computational Linguistics.

[13] S. Chen, Y. Wu, c. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *40th Int. Conf. on Machine Learning (ICML) 2023*, Honolulu, Hawaii, 2023.

[14] Yuan Gong, Yu-An Chung, and James Glass, "AST: Audio spectrogram transformer," in *Interspeech*, Brno, Czech Republic, 2021, pp. 571–575.

[15] H. Zhang, M. Cissé, Y.N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th Int. Conf. on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conf. Track Proceedings*, 2018.

[16] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, and Q.V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019, pp. 2613–2617.

[17] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.

[18] C. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, pp. 88–103, 1904.