Robust Temporal Point Event Localization through Smoothing and Counting

Julien Schroeter [1] **Kirill Sidorov** [1] **David Marshall** [1]

Abstract

This work addresses the long-standing problem of robustly learning precise temporal point event localization despite only having access to poorly aligned labels for training. To that end, we introduce a novel loss function that relaxes the reliance of the training on the exact position of labels, thus allowing for a softer learning of event localization. We demonstrate state-of-the-art performance against standard benchmarks in challenging experiments.

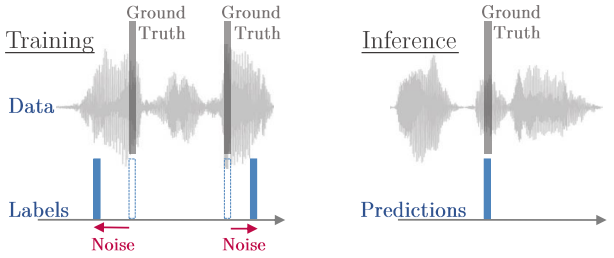


Figure 1. Task illustration. Model training solely relies on noisy labels that differ from the actual ground-truth, while the final inference objective is the *precise* localization of events.

1. Introduction

This work tackles the problem of precise temporal localization of point events (i.e., determining when and which instantaneous events occur) in sequential data (e.g. time series, video, or audio sequences) despite only having access to poorly aligned annotations for training (see Figure 1). This task is characterized by the discrepancy between the noisiness of the training labels and the precision expected of the predictions during inference. Indeed, while models are trained on inaccurate data, they are evaluated on their ability to predict event occurrences as precisely as possible with respect to the actual ground-truth. In such a setting, effective models have to infer event locations more accurately than the labels they relied on for training. This requirement is particularly challenging for most classical approaches that are designed to learn localization by strictly mimicking the provided annotations. Indeed, as the training labels themselves do not accurately reflect the event location, focusing on replicating these unreliable patterns is incompatible with the overall objective of learning the actual ground-truth.

This work introduces a novel model-agnostic loss function that relaxes the reliance of the learning process on the exact temporal location of the annotations. This softer learning approach inherently makes the model more robust to temporally misaligned labels.

[1]Cardiff University, United Kingdom. Correspondence to: Julien Schroeter <SchroeterJ1@cardiff.ac.uk>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

2. Related Works

The literature on temporal noise robustness is limited despite the relevance of this issue. First, Yadati et al. (2018) propose solutions combining noisy and expert labels; however, these methods require a sizable clean subset of annotations, unlike our approach. Second, while Adams and Marlin (2017) achieve increased robustness by augmenting simple classifiers with an explicit probabilistic model of the noise structures, the effectiveness of the approach on more complex temporal models still needs to be demonstrated. Finally, Lea et al. (2017) perform robust temporal action segmentation by introducing an encoder-decoder architecture. However, the coarse temporal encoding comes at the expense of finer-grained temporal information, which is essential for the precise localization of short events (e.g., drum hits). In this paper, rather than a new architecture, we propose a novel and flexible loss function — agnostic to the underlying network — which allows for the robust training of temporal localization networks even in the presence of extensive label misalignment.

Classical Heuristic Our approach is closely linked to the more classical trick of label smoothing or target smearing (e.g., applying a $\tilde{\sigma}^2$-Gaussian filter $\Phi_{\tilde{\sigma}^2}$ to the labels $\mathbf{y}_i$) which has been considered to increase robustness to temporal misalignment of annotations (Hawthorne et al., 2017; Schlüter & Böck, 2014). However, this slight modification of the input data ultimately leads to several issues such as location ambiguity and prediction entanglement (see full discussion in Section 3.1). In contrast, our novel loss function does not suffer from any of these issues, while still achieving a more robust localization learning.

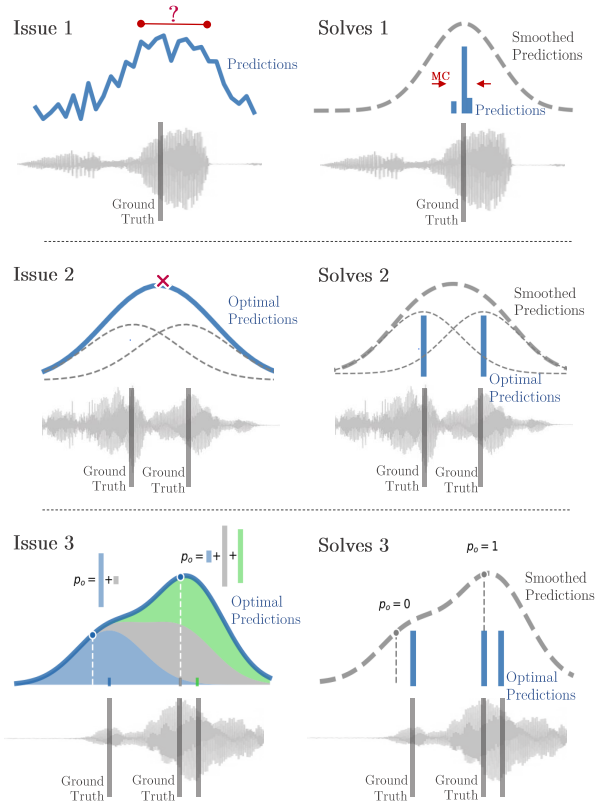


Figure 2. Our approach solves the inherent drawbacks of the classical trick of smoothing the labels only. **Issue 1**: ambiguous predictions of event locations require the use of additional heuristics. **Issue 2**: close events cannot be easily disentangled. **Issue 3**: awareness of past and future event occurrences is required to make optimal predictions (e.g., left tail estimation for causal models).

3. The *S*oftLoc Model

3.1. Drawbacks of Labels Smoothing

Label smoothing (e.g., applying a Gaussian filter to the point label) is a common and state-of-the-art methodology in 2D image point detection applications where spatial uncertainty must be dealt with, e.g., human pose estimation (Tompson et al., 2014). This slight modification of the labels converts the original point prediction problem into a distribution prediction problem, as the smoothing transforms the point labels into distributions. The models are then trained to predict these distributions, which eventually have to be transformed back to point predictions using hand-crafted peak picking heuristics. However, despite its intuitive nature, this solution presents inherent drawbacks when applied to temporal point localization (see Figure 2):

(Issue 1) As the model is designed to yield heatmap predictions that are spread out over several timesteps, additional heuristics (e.g. peak picking (Böck et al., 2013)) are required to obtain precise point predictions. The learning of point localization is thus not done in an end-to-end fashion (see Figure A.1 for illustration).

(Issue 2) Even advanced peak picking struggles to *disentangle* close events. For instance, a single maximum might emerge in the middle of two events, thus significantly harming the temporal resolution of the final predictions.

Even in a noise-free setting, by transforming the point targets into distributions, the optimal solution with respect to the training loss (i.e., heatmap prediction) does not match the goal of the pipeline (i.e., precise point prediction):

$$\begin{aligned}\hat{y}_i^{opt}(t) &= \sum_{\tau=0}^{T_i} y_i(\tau)\Phi_{\tilde{\sigma}^2}(\tau - t) \\ &= \underbrace{\sum_{\tau \le t-1} y_i(\tau)\Phi_{\tilde{\sigma}^2}(\tau - t)}_{\text{A}} + y_i(t)\Phi_{\tilde{\sigma}^2}(0) \\ &\quad + \underbrace{\sum_{\tau \ge t+1} y_i(\tau)\Phi_{\tilde{\sigma}^2}(\tau - t)}_{\text{B}} .\end{aligned} \tag{1}$$

(Issue 3) The optimal prediction at any given time does not only depend on previous event occurrences (Eq. 1 A), but also on all closely upcoming events (Eq. 1 B). This implies that correctly detecting an event is not enough; the context — before and after — also has to be estimated accurately. For instance, the optimal prediction value for a given event occurrence is different whether it stands alone or it is directly followed by other event occurrences. This cross-influence from other timesteps is especially problematic for causal models (i.e., models that make predictions at time t only with data up to time t). Indeed, these models have little or even no ability to integrate information from future timesteps. Thus, for example, requiring them to estimate the left tail of the label distribution might compel them to learn irrelevant features preceding the actual event occurrence, leading to poor generalization.

The presence of strong label misalignment further worsens all these issues as increased noise commonly warrants increased smoothing, dispersing the label (and consequently the prediction) mass even more.

3.2. Soft Localization Learning Loss

Smoothing Many of the drawbacks arising from the asymmetric nature of the one-sided smoothing can however be alleviated by filtering not only the labels (i.e., $\Phi_{S_M} * \mathbf{y}_i(\cdot)$), but also the predictions (i.e., $\Phi_{S_M} * \hat{\mathbf{y}}_i(\cdot)$) with a softness parameter s_M. The comparison of these two smoothed processes yields a relaxed loss function for the soft learning of the location that deals on its own with the temporal uncertainty of the labels. Indeed, in such a setting, the model is given point labels and directly infers *point predictions* in an end-to-end fashion without having to resort to heatmaps nor distributions (see Figure A.1 for illustration);

it is only the loss function that views these point labels and point predictions as smoothed processes. In discrete time settings, the loss can be written as:

$$\mathscr{L}_{\text{SLL}}(\theta) = \sum_i \mathcal{L}\big(\Phi_{\mathcal{S}_M} * \hat{\mathbf{y}}_{i,\theta}(\cdot), \Phi_{\mathcal{S}_M} * \mathbf{y}_i(\cdot)\big), \quad (2)$$

with $\Phi_{\mathcal{S}_M}$ a $\mathcal{S}_M$-Gaussian filter. For all experiments in Section 4, $\mathcal{L}$ is set to the average local mean-squared error. The learning is characterized as *soft* since the loss is not strictly constraining in terms of precision or mass concentration. Indeed, the mass of each event can be both scattered over numerous timesteps and slightly shifted temporally without any abrupt increase in loss. Thus, the model's reliance on exact label locations is relaxed.

Properties Symmetrically smoothing *both* the labels and predictions directly solves several of the issues highlighted in the previous section (see Figure 2). Indeed, in a noise-free setting, the optimal predictions with respect to $\mathscr{L}_{\text{SLL}}$ are the original annotations themselves, i.e.,

$$\hat{y}_i^{opt}(t) = y_i(t). \quad (3)$$

Thus, the training objective and the task objective (i.e., precise point predictions) are aligned.

(Solves 3) Since the optimal behavior (with respect to $\mathscr{L}_{\text{SLL}}$) is to predict events at the exact moment they occur (Eq. 3), there is no cross-influence across the different timesteps anymore. Thus, smoothing both the labels and predictions does not only simplify the localization learning (compared to Eq. 1), but also allows causal models to deal with uncertainty.

(Partially Solves 1 and 2) As points (rather than smoothed heatmaps) are inferred, the prediction mass of a particular event is not necessarily dispersed over time. For instance, in noise-free settings, the point targets themselves are the solution to the optimization problem (Eq. 3). However, $\mathscr{L}_{\text{SLL}}$ is not a hard constraint against splitting a detection into multiple lower-likelihood point predictions.

3.3. $\mathcal{S}$oftLoc Loss

Counting Both the potential dispersion of the prediction mass (i.e. prediction of a single event split into a series of lower-likelihood triggers) and its direct consequences on localization performance still need to be addressed. To that end, we propose to leverage the properties of the event-counting loss function defined in (Schroeter et al., 2019):

$$\mathscr{L}_{\text{MC}}(\theta) = -\sum_i \log\left(\sum_{A\in F}\prod_{l\in A}\hat{y}_{i,\theta}(l)\prod_{j\in A^c}(1-\hat{y}_{i,\theta}(j))\right), \quad (4)$$

where F is the set of all subsets of $\{1,...,T_i\}$ of size $\sum_k \mathbf{y}_i(k)$. Indeed, the loss exhibits an implicit strong mass convergence property, which concentrates the scattered prediction mass toward well-defined single points in time. More precisely, this loss function highly penalizes scattered low-likelihood predictions to the benefit of more sparse high-likelihood predictions. For instance, given a sequence with exactly one event occurrence, a unique prediction with probability $p = 0.9$ would induce a loss of $-\log(0.9)$, while two predictions of $p = 0.45$ would yield a much higher contribution of $-\log(0.495)$, regardless of the temporal position of the predictions.

Full $\mathcal{S}$oftLoc Model Incorporating this mass convergence loss as a regularizer to our soft localization learning loss $\mathscr{L}_{\text{SLL}}$ allows the model to directly achieve unique precise impulse-like localization (i.e. a single high likelihood trigger per event), without weakening its noise robustness properties. Thus, this eliminates prediction temporal ambiguity and disentanglement issues, as only a single point prediction is outputted per event occurrence **(Solves 1 & 2)**. We define the $\mathcal{S}$oftLoc loss as follow:

$$\mathscr{L}_{\mathcal{S}\text{oftLoc}}(\theta) = \big(1-\alpha_\tau\big)\mathscr{L}_{\text{SLL}}(\theta) + \alpha_\tau\mathscr{L}_{\text{MC}}(\theta). \quad (5)$$

Overall, when trained with this novel loss function, models simultaneously softly learn to mimic the location annotations, while converging the scattered mass toward single impulse-like predictions. In this equation, α_τ regulates the predominance of the mass convergence against the soft location learning (for training iteration τ).

End-to-end Learning of Localization One of the key factors of the predominance of the deep learning models over classical ones relies on their ability to solve problems in an end-to-end fashion, without the need to resort to partial optimization or hand-crafted heuristics. In contrast to more classical approaches (see Section 3.1), our proposed method is an end-to-end solution to the problem of temporal localization in the presence of misaligned labels (see Figure A.1 for illustration) and thus can then be expected to better serve the task at hand.

4. Experiments

In order to demonstrate the effectiveness and flexibility of our approach, a broad range of challenging experiments are conducted (video action detection and music event detection). *Implementation details*[1].

4.1. Golf Swing Sequencing in Video

In this section, we replicate the golf swing event detection experiment from (McNally et al., 2019) using either the original cross-entropy loss (CE), the one-sided smoothing (oSS), or our proposed loss ($\mathscr{L}_{\mathcal{S}\text{oftLoc}}$) for training (leaving everything else the same). The task consists in the precise detection (within a one frame tolerance) of eight different golf swing events in video extracts (e.g., address and impact). To assess robustness to noisy annotations, rounded

[1] https://github.com/SchroeterJulien/2020-ICML-UDL-Workshop-Robust-Temporal-Point-Event-Localization.git

Table 1. Golf Swing Action Detection. Performance comparison with respect to various label misalignment levels $\lfloor \mathcal{N}(0, \sigma^2) \rceil$. The cross-validated (4-folds) mean accuracy is reported.

Loss	$\sigma = 0$	1	2	3	4
CE	68.1	60.4	51.6	43.1	36.9
oSS	**69.1**	66.2	60.6	54.7	50.7
$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	67.2	**68.0**	**65.6**	**58.6**	**54.2**

normally distributed misalignments (i.e., $\epsilon_m \sim \lfloor \mathcal{N}(0, \sigma^2) \rceil$) are applied to the event timestamps of the training samples, while the test labels are kept intact for unbiased inference.

Results Table 1 (a) confirms the intuitive understanding that the cross-entropy (CE) is not well suited to effectively deal with label misalignment. Indeed, we observe here that attempting to strictly mimic unreliable annotations leads to poor generalization performance. The results further reveal that even just one of the issues presented in Section 3.1 (e.g., here prediction ambiguity) can negatively impact the prediction accuracy, as shown by the significant performance gap between our approach ($\mathscr{L}_{\mathcal{S}\text{oftLoc}}$) and the one-sided smoothing (oSS) in noisy settings. Indeed, while our approach yields sharp predictions, the oSS predictions are highly ambiguous as illustrated in Figure 3. In strict settings with reduced error tolerance, this lack of preciseness can certainly lead to suboptimal performance. (Even more clear-cut predictions can be obtained for $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ by training longer than the predefined 10k iterations, which would allow for a full convergence of the mass convergence loss.)

4.2. Piano Onset Experiment

Piano transcription and more specifically piano onset detection is a difficult problem, as it requires precise and simultaneous detection of hits from 88 different polyphonic channels. In this section, we reproduce the experiment from Hawthorne et al. (2017) using the MAPS database (Emiya et al., 2010). (Only onsets are considered for the comparison.) Once again, to evaluate the robustness, the training labels are artificially perturbed according to a normal distribution ($\epsilon_m \sim \mathcal{N}(0, \sigma^2)$).

Benchmarks Three additional classical benchmarks based on the state-of-the-art model (on clean data) proposed by (Hawthorne et al., 2017) are considered: first, the original model itself which is highly representative of models aiming for optimal performance with little regard for annotation noise (ORIGINAL); second, a version with extended onset length (i.e., target smearing) (EXTENDED); finally, a version trained with the soft bootstrapping loss proposed by (Reed et al., 2014) instead of the cross-entropy for increased robustness.

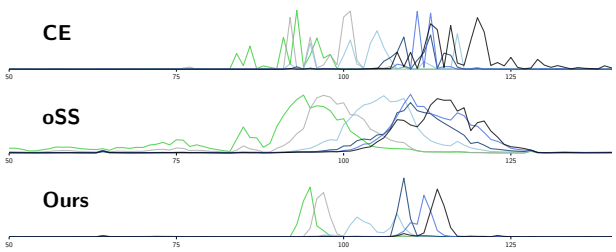


Figure 3. Out-of-Sample Golf Swing Action Predictions. **Ours**: sharp predictions, **oSS**: ambiguous predictions, **CE**: multiple peaks. (Test sequence: 0, split: 1, noise level: $\sigma = 3$ frames.)

Architecture, Training, and Evaluation See B.2.

Results As summarized in Table 2, our proposed $\mathcal{S}$oftLoc approach displays strong robustness against label misalignment; in contrast to all benchmarks, the performance appears almost invariant to the noise level. For instance, at $\sigma = 150$ms, only 26% of training labels lie within the 50ms tolerance (see Figure B.2 in Appendix B.1 for illustration); in such a context, the score achieved by our $\mathcal{S}$oftLoc model (i.e., $\sim 75\%$) is unattainable for classical approaches, which do not take label uncertainty into account and attempt to strictly fit the noisy annotations. While standard tricks, such as label smoothing (oSS) or label smearing (EXTENDED) slightly improve noise robustness, their effectiveness is limited. The results also reveal that, as the noise level increases, the addition of the mass convergence regularizer $\mathscr{L}_{\text{MC}}$ to $\mathscr{L}_{\text{SLL}}$ is key to achieve strong robustness. Finally, a fixed parameter set is used throughout this experiment, which explains the small performance gap between our approach and (Hawthorne et al., 2017) for the noise-free case. This could easily be remedied by adapting the loss settings (e.g., $\alpha_\tau = 1$, $\mathcal{S}_M^2 \to 0$ms and $\mathcal{L}(\cdot) = -\log(1 - |\cdot|)$ as our loss is a strict generalization of the standard cross-entropy.

5. Conclusion

In this work, we introduced a novel loss function that allows for the training of precise temporal localization models even in the presence of poorly aligned annotations. While a softer learning of event localization is already made possible through classical heuristics (e.g., label smoothing), we showed that these approaches inherently suffer from multiple drawbacks (e.g., entanglement and ambiguity of predictions). We demonstrated the effectiveness of our approach in a number of challenging tasks.

Table 2. Piano Onset Detection. Performance comparison with respect to label misalignment distribution $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$.

Loss	$\sigma = 0$ms	50ms	100ms	150ms	200ms
Haw. (ORIGINAL)	**82.1**	38.5	2.0	0.5	0.2
Haw. (EXTENDED)	77.7	68.0	30.7	9.2	3.9
Haw. (BOOTSTRAP)	79.1	74.2	32.5	15.4	6.9
oSS	73.1	70.5	59.2	41.3	28.0
$\mathscr{L}_{\text{SLL}}$	76.1	76.0	75.1	66.9	46.9
$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	76.0	**76.3**	**75.9**	**74.0**	**73.7**

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

References

Adams, R. and Marlin, B. Learning Time Series Detection Models from Temporally Imprecise Labels. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 157–165. PMLR, 2017.

Böck, S., Schlüter, J., and Widmer, G. Enhanced peak picking for onset detection with recurrent neural networks. In *Proceedings of the 6th International Workshop on Machine Learning and Music (MML)*, pp. 15–18, 2013.

Emiya, V., Badeau, R., and David, B. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2010.

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., and Eck, D. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*, 2017.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. Temporal convolutional networks for action segmentation and detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 156–165. IEEE, 2017.

McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., and Wong, A. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., and Ellis, D. P. `mir_eval`: A transparent implementation of common MIR metrics. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, pp. 367–372, 2014.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Schlüter, J. and Böck, S. Improved musical onset detection with convolutional neural networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983. IEEE, 2014.

Schroeter, J., Sidorov, K., and Marshall, D. Weakly-supervised temporal localization via occurrence count learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 5649–5659, 2019.

Stevens, S. S., Volkmann, J., and Newman, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1799–1807, 2014.

Yadati, K., Larson, M., Liem, C. C., and Hanjalic, A. Detecting socially significant music events using temporally noisy labels. *IEEE Transactions on Multimedia*, 20(9):2526–2540, 2018.

A. End-to-End Learning of Localization

Figure A.1 presents the clear contrast in terms of localization learning between our proposed approach and classical ones. Indeed, training with our novel loss function allows for an end-to-end learning of localization without relying on additional heuristics to obtain clear-cut point predictions in time.

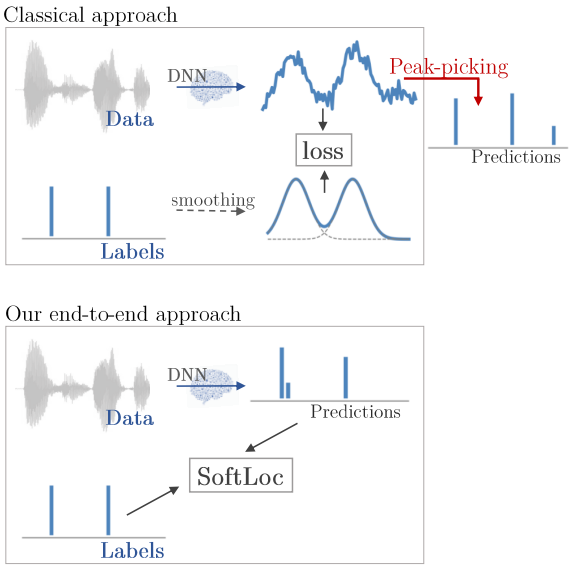
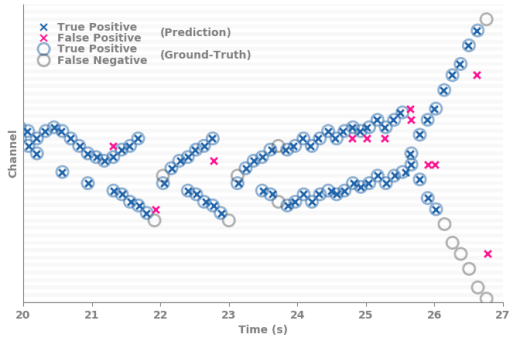
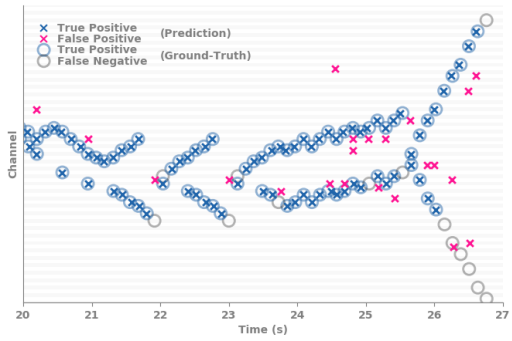


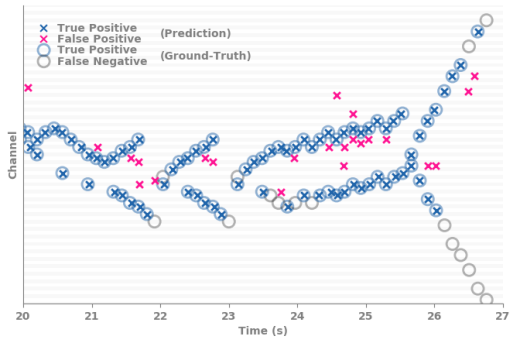
Figure A.1. Modeling novelty. By smoothing *both* the labels and predictions, our model directly infers point predictions rather than distributions. Among other things, this modification allows for an end-to-end learning of localization.



(a) Noise-free training data ($\sigma = 0$ms)

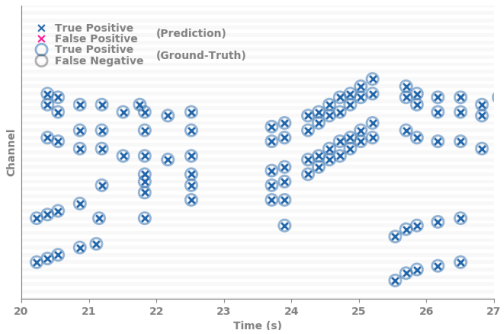


(b) Very noisy training data ($\sigma = 100$ms)

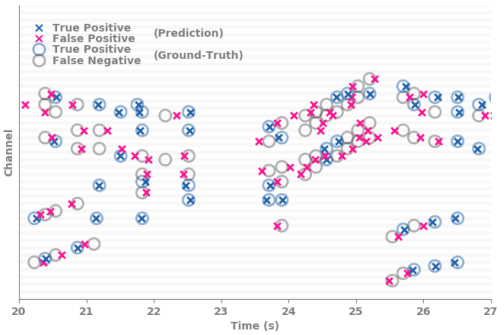


(c) Extremely noisy training data ($\sigma = 200$ms)

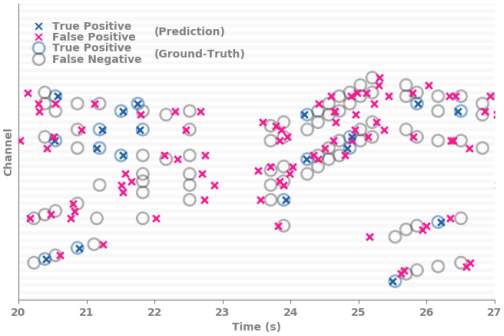
Figure B.1. Out-of-sample predictions of our $\mathcal{S}$oftLoc model trained on data subject to various levels of noise, ranging from (a) the noise-free to (d) the extremely noisy setting. *(Schubert – Piano Sonata in A minor, D 784, Opus 143, 3. Mov)*



(a) Noise-free training data ($\sigma = 0$ms)



(b) Very noisy training data ($\sigma = 100$ms)



(c) Extremely noisy training data ($\sigma = 200$ms)

Figure B.2. In-sample performance of the noisy training labels themselves (*as predictions*) when compared to the clean ground-truth. *(Liszt – Hungarian Rhapsody No. 10)*

B. Piano Onset Detection

B.1. Noisy Labels and Ground-Truth Discrepancy

To further illustrate the complexity of the localization task when annotations are subject to misalignment, we consider the training labels as predictions and then compare them to the clean ground-truth. Figure B.2 displays an example of the quality of the training labels. Obviously, in the noise-free setting (i.e., $\sigma = 0\text{ms}$), the localization is spotless as the training labels and the ground-truths are identical. However, as the noise level increases, the proportion of labels that stay within the 50ms tolerance window decreases significantly. More precisely, the performance (i.e., F_1-score) of the labels themselves is 68.2%, 39.8% and 23.7% for σ equal to 50ms, 100ms and 200ms respectively.

This contrast with the performance of our approach, which appears almost invariant to the noise level (see Figure **??**).

B.2. Implementation Details

The network is comprised of six convolutional layers (representation learning) followed by a 128-unit LSTM (temporal dependencies learning) and two fully-connected layers (prediction mapping). The network is trained using mel-spectrograms (Stevens et al., 1937) and their first derivatives stacked together as model input, while data augmentation in the form of sample rate variations is applied for increased robustness and performance. The models are evaluated on the *noise-free* test set using the *mir_eval* library (Raffel et al., 2014) with a 50ms tolerance (Hawthorne et al., 2017). ($s_M = 100\text{ms}$, $\alpha_\tau = \max(\min(\frac{\tau - 10^5}{10^5}, .9), .2)$.)