REFORMULATING SOFT DYNAMIC TIME WARPING: INSIGHTS INTO TARGET ARTIFACTS AND PREDICTION QUALITY

Johannes Zeitler and Meinard Müller
International Audio Laboratories Erlangen, Germany
{johannes.zeitler, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Training deep neural networks for music information retrieval (MIR) often relies on strongly aligned data, where each frame has a precisely annotated target label. To reduce this dependency, soft dynamic time warping (SDTW) enables training with weakly aligned data by replacing hard decisions with weighted sums, allowing for gradient-based learning while aligning feature sequences to shorter, often binary, target sequences. However, SDTW introduces gradient artifacts that can cause blurring and degrade predictions, impacting the learning process. In this work, we analyze the sources and effects of these artifacts and propose a reformulation of SDTW that expresses its gradient in terms of an equivalent strongly aligned target representation. This reformulation provides an intuitive interpretation of learned representations and insights into the impact of SDTW hyperparameters on the prediction quality. Using multi-pitch estimation as a case study, we systematically investigate these modified targets and demonstrate their potential for improving training stability, interpretability, and alignment quality in MIR tasks.

1. INTRODUCTION AND RELATED WORK

Many state-of-the-art methods for classification and regression rely on training deep neural networks (DNNs) using large amounts of labeled data. While accurately labeled training data is widely available in fields such as image recognition, real-world time series data is rarely strongly aligned, meaning there is no precise frame-by-frame correspondence between the input signal and its label. This lack of alignment is mainly due to the high cost of manual annotation. In music information retrieval (MIR), examples of such strong targets primarily include Disklavier recordings [1, 2] and synthetic data [3]. In contrast, weakly aligned labels, which provide a global correspondence to the input but lack precise frame-level alignment, are easier to obtain. For example, in musical instrument transcription, the start and end times of segments can be annotated in a music recording and its corresponding symbolic score. This results in a “weak target” representation, where all notes appear in the correct order but their precise onset times and durations remain uncertain.

To train DNNs with weak targets, an alignment step between network predictions and the weak targets is essential. One common approach, as applied in [4, 5], is to perform an offline alignment between the predicted features and weak targets using classical dynamic time warping (DTW) [6]. The aligned labels are then treated as strong targets for training, and this alignment can be iteratively refined after each training step in an expectation-maximization-like process. A second approach incorporates the alignment step directly into the computation of the training loss, ensuring that predictions and weak targets are aligned implicitly. A well-known example is the connectionist temporal classification (CTC) loss function, widely used in automatic speech recognition [7] and also adopted in multi-pitch estimation (MPE) [8]. Another method extends DTW by introducing a differentiable minimum function [9–11], leading to the soft dynamic time warping (SDTW) algorithm. SDTW has been applied to tasks such as multi-pitch and pitch class estimation [12, 13]. Unlike CTC, SDTW is not limited to a finite target alphabet and avoids the combinatorial explosion that arises when targets are represented by high-dimensional multi-hot vectors, as in MPE.

Despite its advantages, SDTW-based training exhibits instabilities, with performance variations influenced by data representation [12], training strategies [13], and hyperparameter choices such as step weights [14]. Previous experiments have observed issues like alignment collapse and diagonalization [13], but to date, there has been no straightforward way to analyze the SDTW training process. One key challenge is that the network parameter updates resulting from the SDTW loss are difficult to interpret due to the algorithm’s mathematical complexity. In contrast, when training with strongly aligned targets using element-wise loss functions such as mean squared error (MSE) or binary cross-entropy (BCE), the optimization process is more transparent, as the network simply minimizes the distance between predictions and strong targets.

In this work, we address the following fundamental questions: How does SDTW training differ from standard training with strongly aligned targets? What does a network actually learn when trained with an SDTW loss? To provide answers, as a key contribution of this paper we reformulate the SDTW gradient into an equivalent represen-

Attribution: J. Zeitler and M. Müller, “Reformulating Soft Dynamic Time Warping: Insights into Target Artifacts and Prediction Quality”, in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

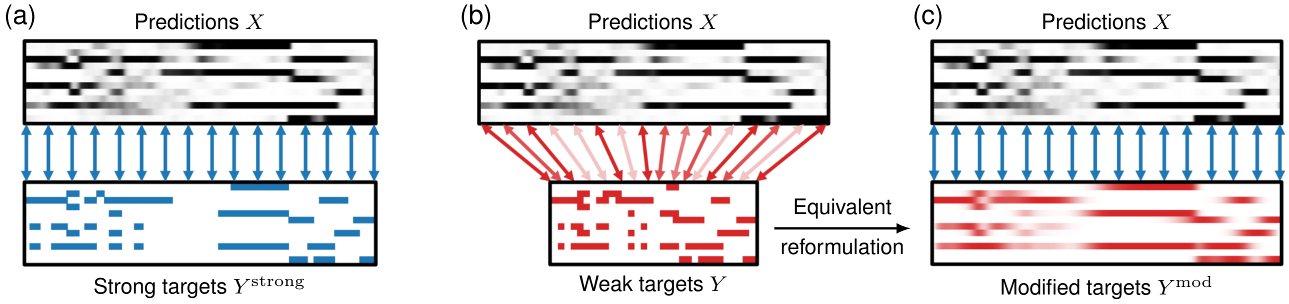

Figure 1: Overview of different training and alignment strategies. **(a)** Strong targets Y^{strong} with direct frame-wise correspondence to the predictions X. **(b)** Weak targets Y that require alignment to the predictions. **(c)** Reformulation of weakly aligned targets into modified targets Y^{mod}, ensuring frame-wise correspondence with the predictions.

tation derived from element-wise MSE and BCE losses. This reformulation introduces interpretable *modified targets* that, when used as strong targets in an element-wise loss function, produce identical network updates to those obtained under SDTW with weak targets. In other words, the DNN learns to minimize the distance (e.g., MSE or BCE) to these modified targets. By inspecting the modified targets, we gain insights into what the DNN actually learns, allowing us to analyze the impact of SDTW hyperparameters and training strategies on network performance. Furthermore, these modified targets can be visualized or sonified at early training stages to provide qualitative assessments of the learning process. Distance measures to strongly aligned reference targets can also be computed, facilitating quantitative evaluations.

The remainder of this paper is structured as follows. Section 2 introduces the problem and methodology. Section 3 provides background on the SDTW algorithm. In Section 4.1, we reformulate the gradient of the SDTW loss into the canonical form of element-wise MSE and BCE losses, yielding the so-called modified targets. Section 4.2 discusses the properties of these modified targets. Sections 5.1 and 5.2 outline our experimental framework and demonstrate the impact of different training configurations using SDTW. Finally, we discuss our findings and their implications in Section 5.3 and conclude in Section 6.

2. PROBLEM FORMULATION

We consider the task of training a DNN that takes an input sequence and predicts features $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ with $\mathbf{x}_n \in \mathbb{R}^D$, and frame index $n \in \{1, 2, \ldots, N\}$. For example, in the case of MPE, the input sequence can be a spectral representation of an audio recording and X is a sequence of estimated pitch vectors. Ideally, to train such a network, we have access to a strongly aligned target sequence $Y^{\text{strong}} = \left(\mathbf{y}_1^{\text{strong}}, \ldots, \mathbf{y}_N^{\text{strong}}\right)$ with $\mathbf{y}_n^{\text{strong}} \in \mathbb{R}^D$ which temporally corresponds to X on the frame level, as visualized in Figure 1a. In the example of MPE, the target features are typically encoded as binary multi-hot vectors indicating the presence of certain pitches. In the case of strong targets, we can use an element-wise loss function to train the network, such as MSE

$$c_{\text{MSE}}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \tag{1}$$

or BCE

$$c_{\text{BCE}}(\mathbf{x}, \mathbf{y}) = -\mathbf{y}^\top \log \mathbf{x} - (1 - \mathbf{y})^\top \log(1 - \mathbf{x}), \tag{2}$$

where the logarithm of a vector is defined element-wise.

In practice, strongly aligned targets are rarely available in MIR. Instead, weakly labeled targets, which share only a global correspondence with the input data, are more readily obtainable. For instance, in MPE, weak targets $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ with $\mathbf{y}_m \in \mathbb{R}^D$ and $m \in \{1, 2, \ldots, M\}$ can be derived from note events in a musical score. While Y and Y^{strong} contain the same set of feature vectors, they differ in the number of repetitions of each vector. To train DNNs on weakly aligned data, the SDTW loss function is used to align predictions X with weak targets Y during loss computation (see Figure 1b and Section 3). Although SDTW-based training has shown promising results, it is highly sensitive to training strategies and hyperparameter choices [12–14]. Understanding these stability issues requires deeper insights into what the network actually learns. However, unlike training with strong targets and element-wise loss functions, interpreting SDTW-based training remains a challenge due to its complex alignment process.

In this work, we provide deeper insights into SDTW-based training by reformulating the SDTW gradient into an equivalent representation using strongly aligned modified targets $Y^{\text{mod}} = \left(\mathbf{y}_1^{\text{mod}}, \ldots, \mathbf{y}_N^{\text{mod}}\right)$. Training with these modified targets and a standard element-wise loss function results in the same network updates as training with SDTW directly. Importantly, this reformulation does not alter the SDTW training process but instead serves as a tool for better interpreting how the model learns. By analyzing these modified targets (Figure 1c), we gain a clearer understanding of the features the DNN actually learns, making SDTW training as interpretable as training with strongly aligned targets.

3. SOFT DYNAMIC TIME WARPING

We aim to compute and minimize the soft alignment cost between the sequences X and Y. Without loss of generality, let X represent a sequence of DNN predictions and Y the corresponding weak targets. To quantify the alignment cost between elements of X and Y, we employ a local cost function $c : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ such as MSE or BCE. The resulting element-wise costs are stored in the local cost matrix $\mathbf{C} \in \mathbb{R}^{N \times M}$ defined as:

$$\mathbf{C}\,(n,m) = c(\mathbf{x}_n, \mathbf{y}_m)\,. \tag{3}$$

To ensure differentiability, we use a smooth approximation of the minimum function, defined as:

$$\min^\gamma(\mathcal{S}) = -\gamma \log \sum_{s \in \mathcal{S}} \exp\{-s/\gamma\}\,, \tag{4}$$

where $\gamma > 0$ is a temperature hyperparameter, and $\mathcal{S}$ is a list of real numbers [9]. The parameter γ controls the degree of smoothness, with the function converging to the hard minimum as $\gamma \to 0$. Using this formulation, we define the SDTW forward recursion to compute the accumulated cost matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$ as:

$$\begin{aligned}\mathbf{D}(n,m) = \min^\gamma(\{&w_\mathrm{h}\mathbf{C}(n,m) + \mathbf{D}(n-1,m),\\ &w_\mathrm{v}\mathbf{C}(n,m) + \mathbf{D}(n,m-1),\\ &w_\mathrm{d}\mathbf{C}(n,m) + \mathbf{D}(n-1,m-1)\})\,,\end{aligned} \tag{5}$$

where w_h, w_v, and w_d are step weights controlling the cost contribution of horizontal, vertical, and diagonal steps in the alignment process [9, 14]. The original SDTW formulation from [9] is recovered for $w_\mathrm{h} = w_\mathrm{v} = w_\mathrm{d} = 1$. The overall alignment cost is given by the final element of the accumulated cost matrix:

$$\mathrm{SDTW}(\mathbf{C}) = \mathbf{D}(N,M)\,. \tag{6}$$

The gradient $\mathbf{H} \in \mathbb{R}^{N \times M}$ of the SDTW cost w.r.t. the cost matrix:

$$\mathbf{H}(n,m) := \frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{C}(n,m)} \tag{7}$$

can be computed efficiently by a second recursion in reverse order. We refer to [14] for the technical details of the gradient computation. When using uniform step weights, i.e., $w_\mathrm{h} = w_\mathrm{v} = w_\mathrm{d} = 1$, the elements $\mathbf{H}(n,m) \in [0,1]$ can be interpreted as a form of pseudo-probability, indicating the degree to which the sequence elements $\mathbf{x}_n$ and $\mathbf{y}_m$ are aligned. $\mathbf{H}$ is therefore also called the "soft alignment matrix" (see [14] for a discussion of SDTW alignments). For $\gamma \to 0$, SDTW converges to the classical "hard" DTW algorithm, yielding a binary alignment matrix with $\mathbf{H}(n,m) \in \{0,1\}$. The gradient of the SDTW cost w.r.t. the network outputs $\mathbf{x}_n$ is obtained by applying the chain rule:

$$\frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{x}_n} = \sum_{m=1}^{M} \mathbf{H}\,(n,m) \cdot \frac{\partial\, c(\mathbf{x}_n, \mathbf{y}_m)}{\partial\,\mathbf{x}_n}\,. \tag{8}$$

This gradient is typically computed using the automatic differentiation modules available in modern deep learning frameworks.

4. GRADIENT REFORMULATION INTO MODIFIED TARGETS

While the gradient is well-defined and efficient computation methods exist, it remains unclear which features the DNN actually learns when trained with an SDTW loss. In contrast, when training with strongly aligned targets using standard element-wise loss functions such as MSE or BCE, the optimization process is more transparent. For instance, given the gradients:

$$\frac{\partial\, c_\mathrm{MSE}(\mathbf{x},\mathbf{y})}{\partial\,\mathbf{x}} = \mathbf{x} - \mathbf{y} \tag{9}$$

$$\frac{\partial\, c_\mathrm{BCE}(\mathbf{x},\mathbf{y})}{\partial\,\mathbf{x}} = -\frac{\mathbf{y}}{\mathbf{x}} + \frac{1-\mathbf{y}}{1-\mathbf{x}}\,, \tag{10}$$

it is evident that the network parameters are adjusted to bring the predictions $\mathbf{x}$ closer to the targets $\mathbf{y}$. However, in the case of SDTW, the alignment process introduces additional complexity, making it less intuitive to interpret what the network is optimizing towards.

4.1 Derivation for MSE and BCE Cost

Next, we reformulate the SDTW gradient from (8) into the standard element-wise loss functions defined in (9) and (10). We then demonstrate the equivalence of the modified target representations for MSE and BCE, offering a unified perspective on SDTW-based training.

4.1.1 MSE as local cost

For $c = c_\mathrm{MSE}$, the gradient of the SDTW cost w.r.t. the DNN predictions is given by:

$$\frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{x}_n} = \sum_{m=1}^{M} \mathbf{H}\,(n,m) \cdot (\mathbf{x}_n - \mathbf{y}_m)\;, \tag{11}$$

which follows from (8) and (9). Our goal is to reformulate this expression into the standard form of (9). We introduce the row sum of the gradient matrix:

$$h(n) := \sum_{m=1}^{M} \mathbf{H}(n,m) \in \mathbb{R} \tag{12}$$

and define the row-normalized gradient matrix as:

$$\tilde{\mathbf{H}}(n,m) := \mathbf{H}(n,m)/h(n) \in \mathbb{R}^{N \times M}\,. \tag{13}$$

Using these definitions, we can rewrite the gradient from (11) as:

$$\begin{aligned}\frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{x}_n} &= h(n)\cdot\mathbf{x}_n - \sum_{m=1}^{M}\mathbf{H}\,(n,m)\,\mathbf{y}_m\\ &= h(n)\cdot\left(\mathbf{x}_n - \sum_{m=1}^{M}\tilde{\mathbf{H}}\,(n,m)\,\mathbf{y}_m\right)\\ &= h(n)\cdot\left(\mathbf{x}_n - \mathbf{y}_n^\mathrm{mod}\right)\;,\end{aligned} \tag{14}$$

where we define the modified targets:

$$\mathbf{y}_n^\mathrm{mod} := \sum_{m=1}^{M} \tilde{\mathbf{H}}\,(n,m) \cdot \mathbf{y}_m \in \mathbb{R}^D\,. \tag{15}$$

Thus, the network parameters are updated such that the predictions $\mathbf{x}_n$ move closer to the modified targets $\mathbf{y}_n^{\mathrm{mod}}$. The update magnitude is determined by $h(n)$.

4.1.2 *BCE as local cost*

A similar reformulation applies to the SDTW gradient when using BCE as the local cost function. Starting with the gradient:

$$\frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{x}_n} = \sum_{m=1}^{M} \mathbf{H}\,(n,m)\cdot\left(-\frac{\mathbf{y}_m}{\mathbf{x}_n}+\frac{1-\mathbf{y}_m}{1-\mathbf{x}_n}\right), \tag{16}$$

we rewrite this expression in the form of (10) as follows:

$$\begin{aligned}\frac{\partial\,\mathrm{SDTW}(\mathbf{C})}{\partial\,\mathbf{x}_n} &= -\frac{\sum_m \mathbf{H}\,(n,m)\,\mathbf{y}_m}{\mathbf{x}_n} \\ &\quad+\frac{\sum_m \mathbf{H}\,(n,m)-\sum_m \mathbf{H}\,(n,m)\,\mathbf{y}_m}{1-\mathbf{x}_n} \\ &= h(n)\cdot\Bigg(-\frac{\sum_m \tilde{\mathbf{H}}\,(n,m)\,\mathbf{y}_m}{\mathbf{x}_n} \\ &\quad+\frac{1-\sum_m \tilde{\mathbf{H}}\,(n,m)\,\mathbf{y}_m}{1-\mathbf{x}_n}\Bigg) \\ &= h(n)\cdot\left(-\frac{\mathbf{y}_n^{\mathrm{mod}}}{\mathbf{x}_n}+\frac{1-\mathbf{y}_n^{\mathrm{mod}}}{1-\mathbf{x}_n}\right).\end{aligned} \tag{17}$$

Notably, this reformulation uses the same modified targets $\mathbf{y}_n^{\mathrm{mod}}$ and weighting factor $h(n)$ as in the MSE case. These modified targets can be visualized or sonified alongside the original signal, providing an intuitive way to analyze the SDTW training process. Importantly, this approach does not require knowledge of the strong reference targets, making it highly suitable for analyzing training behavior in real-world scenarios where only weakly aligned data is available.

4.2 Properties of the Modified Targets and Magnitude Decay

In this section, we analyze the theoretical properties of the modified targets and illustrate them using a synthetic example (Figure 2). Specifically, we investigate how the SDTW loss influences blurring and magnitude variations when predicting short and long features. To demonstrate these effects, we construct a sequence Y consisting of six binary feature vectors $\mathbf{y} \in \{0,1\}^{12}$, three of which are all-zero (see Figure 2a). Next, we construct a sequence X (as seen in Figure 2b), by repeating the non-zero elements of Y two, one, and five times, respectively, and repeating the all-zero elements three times each. Note that in the case of classical DTW with hard alignments, in our example X and Y could be aligned with zero cost. We are interested in the alignment of these sequences under SDTW loss and thus compute the forward and backward passes with uniform step weights $w_{\mathrm{h}} = w_{\mathrm{v}} = w_{\mathrm{d}} = 1$ and $\gamma = 1$ as described in [14]. We then derive the alignment matrix $\tilde{\mathbf{H}}$ and compute the modified targets Y^{mod}, displaying the results in Figure 2c and Figure 2d, respectively. Ideally, as we chose X to be an unfolded version of Y without additional noise, Y^{mod} should closely resemble X.

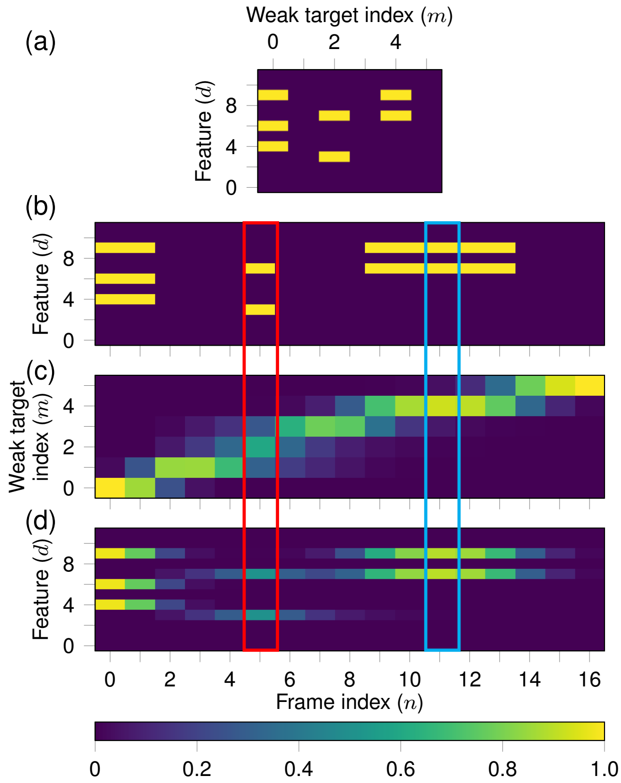

Figure 2: Illustration of magnitude decay in modified targets. **(a)** Weakly aligned target sequence Y. **(b)** Predicted sequence X, which corresponds to a perfectly aligned and unfolded version of Y. **(c)** Alignment matrix $\tilde{\mathbf{H}}^{\top}$. **(d)** Modified targets Y^{mod}, showing significant variations in magnitude. A short feature instance experiencing magnitude decay is highlighted in red, while a long feature instance with near-complete magnitude preservation is highlighted in blue.

Due to the relatively high softmin temperature $\gamma = 1$, we observe significant temporal blurring in both the modified targets and the alignment matrix. By definition, the row sums of the alignment matrix $\tilde{\mathbf{H}}$ are normalized, i.e., $\sum_{m=1}^{M} \tilde{\mathbf{H}}(n,m) = 1$ for all $n \in \{1,\ldots,N\}$. This normalization ensures that each predicted frame is assigned a convex combination of weak target frames. However, when targets are aligned for only a short duration (e.g., staccato notes), the temporal blurring from neighboring frames overlaps with the actual target, leading to a significant reduction in its magnitude after normalization by $h(n)$. This effect is illustrated in Figure 2, where the frame marked in red shows how short events are particularly affected. Conversely, for targets aligned over a longer duration (e.g., sustained notes), the influence of temporal blurring is primarily limited to the onset and offset regions, leaving the central frames mostly unaffected. This phenomenon is visible in the section marked in blue in Figure 2, where the corresponding frame in Y^{mod} retains a relatively high magnitude. Consequently, with a higher softmin temperature γ, SDTW introduces a pronounced magnitude imbalance:

short events tend to have reduced magnitudes, while the central frames of long events maintain higher magnitudes.

5. EVALUATION

In this section, we examine modified targets in a DNN training scenario. To ensure clarity and intuitive accessibility, we select MPE as a case study—a straightforward yet relevant task well-suited for providing insights into the model behavior. MPE is particularly appropriate due to its diverse weakly labeled datasets, consisting of score–audio pairs, and its broad range of target durations, from short staccato notes to sustained tones. Additionally, it provides an intuitive framework for visualization and analysis. Rather than aiming to advance the state of the art in MPE or report performance benchmarks, our objective is to demonstrate how early-stage inspection of modified targets can yield valuable insights into the learning process and serve as a reliable predictor of final model performance. Concluding this section, we integrate theoretical findings from the previous section with empirical insights from our MPE experiment to formulate practical recommendations for training state-of-the-art MIR models with SDTW. Sonifications for all presented examples and further links to Pytorch implementations are available at our website. [1]

5.1 Experimental Setup

As an example architecture for MPE, we adopt a single convolutional stack from the Python implementation of the *Onsets and Frames* model [15]. The stack processes a Mel spectrogram as input and consists of three convolutional layers with batch normalization, max pooling, and dropout, followed by two fully connected layers with sigmoid activation. In total, the model comprises approximately 4.3 million trainable parameters. We choose the *Onsets and Frames* model as our basis due to its widespread use in the literature and its proven effectiveness for transcription tasks. However, we simplify the architecture by using only a single stack, reducing interdependencies between multiple stacks present in the original model. This modification not only decreases the model size but also enhances interpretability by removing recurrent neural networks from the pipeline.

We pre-train the model on strongly aligned data from the MAESTRO [1] dataset for 100000 steps with audio of 20 s length and a batch size of 8, BCE loss, Adam optimizer [16] with an initial learning rate of $6 \cdot 10^{-4}$ and a reduction of the learning rate by a factor of 0.98 every 10000 steps, and gradient clipping.

We fine-tune the model using weakly aligned data from the Beethoven Piano Sonata Dataset (BPSD) [17]. For this, we automatically generate training samples by pairing multi-pitch labels with corresponding audio segments, each spanning 8 measures. If a segment exceeds 20 seconds in duration, it is excluded from training due to hardware memory constraints. The segments are grouped into batches of size 8. For optimization, we use the weighted SDTW loss [14, 18] with BCE as the local cost function. The model is trained for 5000 steps using the Adam optimizer [16] with a learning rate of 10^{-3}. We evaluate performance on a test set from the BPSD with versions that were not included in the training. The pretrained model achieves an F-measure of 0.60 on the test set.

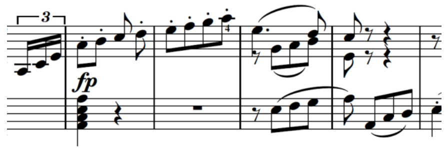

Figure 3: Musical score for running example.

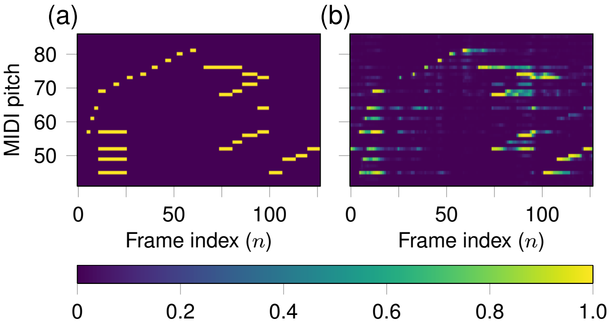

Figure 4: Pianoroll representation of the running example. **(a)** Strongly aligned reference targets. **(b)** Predictions of pretrained model.

The weak targets Y are derived by removing all repetitions from the strongly aligned target sequence Y^{strong}, which is provided in the BPSD. For additional details on the network architecture, we refer to [15], and for a comprehensive explanation of the weighted SDTW loss along with a Python implementation, we refer to [14].

5.2 Analyzing Modified Target Representations

We now analyze the modified targets for both the pretrained model and the final predictions after fine-tuning with the SDTW loss over 5000 training steps. Our goal is to examine how the predictions of the fine-tuned model align with the modified targets obtained from the pretrained model. For this analysis, we use an excerpt from the first movement of Beethoven's second piano sonata (Op. 2 No. 2), as shown in Figure 3. The corresponding reference targets (strongly aligned) and the pretrained model's predictions for a performance by Alfred Brendel (1996) are illustrated in Figure 4a and Figure 4b, respectively. This excerpt features a transition from a staccato passage to a legato section with sustained notes, making it a representative test case for illustrating how SDTW handles variations in note duration.

For our experiments, we use step weights of $w_{\text{h}} = 0.1$, $w_{\text{v}} = 1$, and $w_{\text{d}} = 1$, reducing the weight of horizontal steps (target repetition) to enhance robustness against prediction outliers [14]. We vary the softmin temperature $\gamma \in \{0.1, 1.0, 10.0\}$ and present the results for the modified training targets of the pretrained model and the predictions of the fine-tuned model in Figure 5.

[1] `https://audiolabs-erlangen.de/resources/MIR/2025_ZeitlerM_softDTW_modTargets_ISMIR`

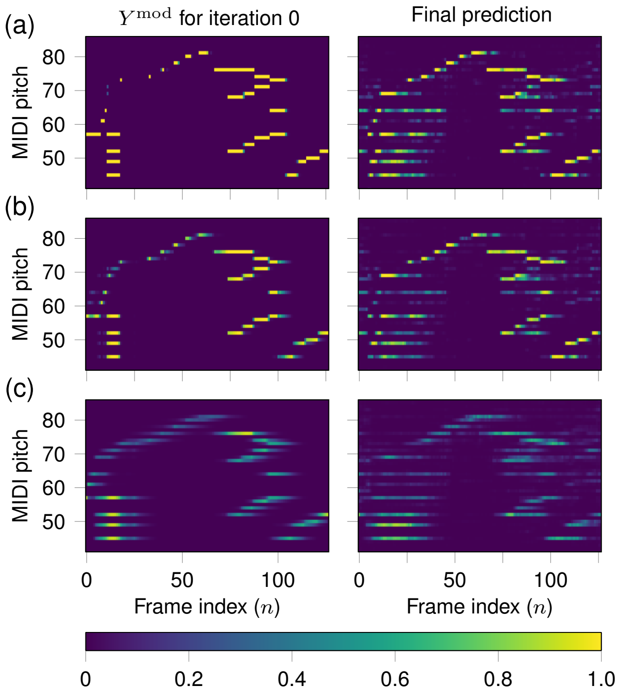

Figure 5: Visualization of modified targets (left) and model predictions (right) for different SDTW configurations. **(a)** $\gamma = 0.1$. **(b)** $\gamma = 1$. **(c)** $\gamma = 10$.

For $\gamma = 0.1$ (Figure 5a), the modified targets align closely with the reference targets from Figure 4, showing only slight blurring. This leads to final predictions that capture all notes with relatively high and consistent magnitude across detected events, resulting in a test F-measure of 0.67. For $\gamma = 1.0$ (Figure 5b), the SDTW loss causes slight blurring at the note onsets and offsets of the modified targets, accompanied by a reduced magnitude for notes in the staccato movement. As the predictions can never get better than what is given by the training targets, also the predicted notes of the fine-tuned model reveal stronger blurring and slightly lower magnitudes than for $\gamma = 0.1$. The test F-measure slightly reduces to 0.64. With $\gamma = 10.0$ (Figure 5c), note events in the modified targets become even more blurred. For all but the longest notes, magnitudes drop considerably, falling below 0.5 in the staccato movement. The fine-tuned model's predictions closely follow this pattern, showing a pronounced magnitude reduction for most notes, with a test F-measure of only 0.36. Notably, the detected staccato notes fall below 0.5, which is problematic for post-processing tasks that often discard events under this threshold.

5.3 Practical Implications of SDTW Reformulation

In this section, we outline some practical implications of the proposed reformulation of SDTW when training DNNs from scratch. Previous studies [13, 19] have observed that training a DNN from a poor initialization requires a relatively high softmin temperature parameter γ. On the one hand, when $\gamma \to 0$, SDTW alignments often degenerate, leading to a collapse in model training. On the other hand, for sufficiently high γ, the network is exposed to a weighted combination of multiple alignments, which facilitates successful training initialization.

Given that a high γ is necessary to initialize DNN training with weak targets, we now examine its implications in the context of a common transcription scenario such as *Onsets and Frames* [15]. Since transcription models aim to predict discrete events (e.g., symbolic note information), a common post-processing step involves applying a detection threshold to the raw network output, treating events above the threshold as active. One direct consequence of the temporal blurring induced by high γ is that predictions fade in and out before and after the actual event. If the magnitude within these fading regions lies above the detection threshold, the thresholded predictions extend before and after the actual event, effectively widening the detected temporal span. In time-sensitive tasks such as onset estimation, this can lead to an undesirable temporal shift in post-processed predictions.

A second observed effect, both theoretically derived and empirically demonstrated, is a decay in magnitude for short events. In onset estimation, where events are inherently short, this decay can be particularly problematic. If the magnitude of short events falls below the detection threshold, these events may be lost entirely after post-processing. This issue is especially critical in models like *Onsets and Frames*, where note activation is conditioned on a preceding onset [15].

Based on these conceptual findings and in line with previous work [13], we offer the following recommendations for using SDTW in DNN training scenarios with poor initialization: **1. High γ for initialization:** Begin training with a relatively high softmin temperature γ to ensure stable convergence. During this phase, post-processed performance metrics (e.g., note accuracy in transcription) may be unreliable due to temporal shifting and magnitude decay. **2. Target inspection:** Despite the shortcomings of metrics after post-processing, the modified targets can be inspected at any point to verify whether the DNN is learning meaningful patterns. **3. Gradual γ reduction:** After the initialization phase, progressively lower γ until the modified targets exhibit less temporal blurring and a balanced magnitude for both short and long events. **4. Resume training:** With refined SDTW parameters, continue training to allow predictions to converge toward the improved modified targets.

6. CONCLUSION

In this paper, we introduced a reformulation of the SDTW gradient into interpretable modified targets, which yield identical network parameter updates when used with standard element-wise loss functions. Through theoretical analysis and a controlled experiment, we demonstrated that temporal blurring and magnitude decay are inherently part of training with SDTW, even though it is not visible in the underlying weak targets. By making the training process more transparent, our approach provides researchers and practitioners with deeper insights into SDTW-based learning and offers an intuitive, practical method for analyzing weakly supervised training strategies.

7. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 500643750 (MU 2686/15-1) and Grant No. 521420645 (MU 2686/17-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

8. REFERENCES

[1] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019.

[2] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (SMD)," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, USA, 2011.

[3] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[4] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proceedings of the International Conference on Machine Learning (ICML)*, Baltimore, Maryland, USA, 2022, pp. 14 918–14 934.

[5] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, pp. 1051–1055.

[6] M. Müller, *Fundamentals of Music Processing – Using Python and Jupyter Notebooks*, 2nd ed. Springer Verlag, 2021.

[7] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA, 2006, pp. 369–376.

[8] C. Weiß and G. Peeters, "Learning multi-pitch estimation from weakly aligned score-audio pairs using a multi-label CTC loss," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2021, pp. 121–125.

[9] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, NSW, Australia, 2017, pp. 894–903.

[10] A. Mensch and M. Blondel, "Differentiable dynamic programming for structured prediction and attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholmsmässan, Stockholm, Sweden, 2018, pp. 3459–3468.

[11] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 1801–1810.

[12] M. Krause, C. Weiß, and M. Müller, "Soft dynamic time warping for multi-pitch estimation and beyond," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[13] J. Zeitler, S. Deniffel, M. Krause, and M. Müller, "Stabilizing training with soft dynamic time warping: A case study for pitch class estimation with weakly aligned targets," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milano, Italy, 2023, pp. 433–439.

[14] J. Zeitler, M. Krause, and M. Müller, "Soft dynamic time warping with variable step weights," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, pp. 356–360.

[15] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference, (ISMIR)*, Paris, France, 2018, pp. 50–57.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015.

[17] J. Zeitler, C. Weiß, V. Arifi-Müller, and M. Müller, "BPSD: A coherent multi-version dataset for analyzing the first movements of Beethoven's piano sonatas." *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 2024.

[18] M. Maghoumi, E. M. Taranta, and J. LaViola, "DeepNAG: Deep non-adversarial gesture generation," in *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, College Station, Texas, USA, 2021, pp. 213–223.

[19] M. Krause, S. Strahl, and M. Müller, "Weakly supervised multi-pitch estimation using cross-version alignment," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milano, Italy, 2023.