CHORALE MUSIC TRANSCRIPTION WITH SOFT-DTW TRAINING LOSS

Huiran Yu
University of Rochester
hyu56@ur.rochester.edu

ABSTRACT

Chorale music transcription has been an infrequent topic in the field of automatic music transcription due to the homogeneity of the sound sources, the smooth attack of the notes, and especially the lack of detailly labeled training data. Meanwhile, the soft-DTW method can align two sequences of different lengths with gradients to enable backpropagation and can be used when time-aligned data are unavailable. Therefore, soft-DTW training loss can help alleviate the problem of data shortage. In this work, we explore soft-DTW loss in the chorale music transcription task and conclude that it performs better than DTW-aligned target training and is suitable for end-to-end music transcription model training when people do not have a time-aligned dataset.

1. INTRODUCTION

Automatic Music Transcription (AMT) has long been a popular research topic in Music Information Retrieval (MIR), as it is the basis of many other symbolic analysis downstream tasks, such as motif analysis and structure analysis. Meanwhile, a high-performance AMT system can also be used to create large datasets, which plays an important role in supporting the trend of large model training.

While most of the works in AMT focus on the transcription of instrumental music, such as piano and symphonic performance, the transcription of choral music is still an under-explored area. Compared with instrumental pieces, the homogeneity of sound sources and the vagueness of the note onsets in chorale music performances add difficulty to the transcription task. Also, the quantity of existing datasets can hardly support the training of large models. They are typically less than one hour, and some datasets only contain 20 minutes of recordings [1]. Meanwhile, there are plenty of recordings on YouTube and human-transcribed MIDI scores on the archive websites, but the only information that links them together is the song title and the composer's name. They could contain the same music content, but typically, the time stamps of the notes are totally off. One method that could solve the alignment problem is Dynamic Time Warping (DTW); however, it is not differentiable and can not be used in end-to-end model training.

This situation urges us to develop methods to train neural transcription models using large amounts of unaligned data. In 2017, M. Cuturi et al. proposed soft-DTW [2], which introduces gradient into the Dynamic Time Warping (DTW) procedure. Then, training with unaligned data becomes feasible. This paper will explore whether soft-DTW training loss can improve the model's performance in the chorale transcription task. The following sections are constructed as follows: section 2 describes the soft-DTW method and model architecture; Section 3 discusses the experiment, including the dataset, experiment settings, and results; and finally, section 4 concludes the paper.

2. METHOD

I used two main parts of the method in this work: soft-DTW [2] and the architecture proposed by [3]. Soft-DTW provides a convenient way to compute the loss between unaligned audio and midi annotations, and the model implements the transcriber.

2.1 Soft-DTW

First proposed in [2], the soft-DTW method is a standard of differentiably matching two sequences with different lengths.

Given two sequences $\mathbf{x} \in \mathbb{R}^{k,m}$, $\mathbf{y} \in \mathbb{R}^{k,n}$, k is the dimension of each elements in the sequence, m is the length of $\mathbf{x}$, n is the length of $\mathbf{y}$. The first step to do time warping between two sequences is to calculate their distance matrix $\Delta(x, y) := [\delta(x_i, y_j)]_{ij}$. People usually use distance functions such as cosine-similarity or binary cross-entropy as function δ. In this work, we model the transcription task as a binary classification problem at each time frame in each note entry. Therefore, we use binary cross-entropy as δ.

Let's consider the DTW [4] alignment and Global Alignment kernel (GAK) [5] method to find the alignment path. If we denote the alignment matrix as A, then the alignment score can be expressed as $\langle A, \Delta \rangle$, the inner product of the two matrices. The score can also be expressed by:

$$\text{DTW}(\mathbf{x}, \mathbf{y}) = \min_{A} \langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle \tag{1}$$

$$k_{GA}^{\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{A} e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma} \tag{2}$$

When $\gamma = 0$, the GAK case degenerates into the normal DTW case.

Then, we generalize the definition of the $\min$ operator:

$$\min{}^{\gamma}\{a_1, \ldots, a_n\} = \begin{cases} \min_{1 \le i \le n} a_i, & \gamma = 0 \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0 \end{cases} \tag{3}$$

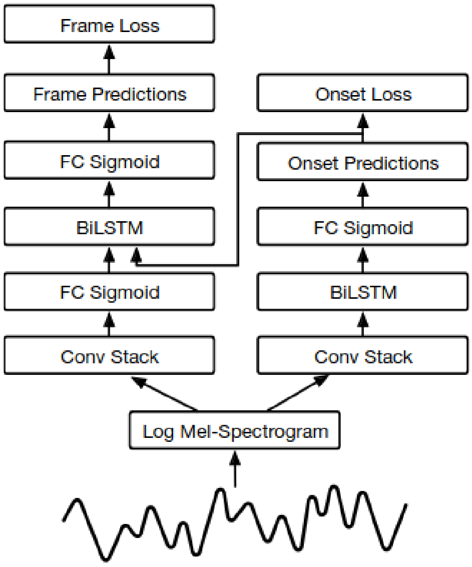

Figure 1. Onsets and Frames architecture from [3]

With this operator, we can link the DTW and GAK together to define γ-soft-DTW:

$$\mathrm{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \min{}^\gamma\{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle\} \tag{4}$$

Now let's consider the derivative of the $\mathrm{dtw}_\gamma(\mathbf{x},\mathbf{y})$. For simplicity, here we only discuss the case when $\gamma > 0$. According to the chain rule,

$$\nabla_{\mathbf{x}}\mathrm{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \left(\frac{\partial\Delta(\mathbf{x},\mathbf{y})}{\partial\mathbf{x}}\right)^\intercal \mathbf{E}_\gamma[A] \tag{5}$$

, where

$$\mathbf{E}_\gamma[A] = \frac{1}{k^\gamma_{GA}(\mathbf{x},\mathbf{y})}\sum_A e^{-\langle A,\Delta(\mathbf{x},\mathbf{y})/\gamma\rangle}A \tag{6}$$

Since all the parameters are known during computation, we can perform the forward and backward calculations during training. Section 2.3 of [2] shows the detailed backward calculation algorithm.

2.2 Model Architecture

For the transcription model, I used the model from [3], which takes mel-spectrogram as input and detects the onset, offset, and activation in each frame. The "activation" block and the "onset" block are shown in Figure 1. The convolution stacks after the mel-spectrogram input act as an acoustic model, and the bi-directional LSTM layer computes the time-dependent information. In real implementation, there is also an "offset" block to calculate the end time of the notes with the same structure as the "onset" blocks, and the onset and offset predictions will all be sent to the activated bi-directional LSTM of the "activation" block to calculate the final note-activation matrix per frame.

The loss $\mathcal{L}$ is computed with three components, L_{onset}, L_{offset} and L_{frame}. The ground truths are the multi-hot activation matrix $M \in \mathbb{R}^{N,T}$, where N is the number of MIDI notes, and T is the number of the frames in the ground truth. The prediction output of each block is matched with the corresponding ground truth with soft-DTW, respectively, and finally, we sum all the losses together to get the final loss $\mathcal{L}$.

3. EXPERIMENT

3.1 Dataset

We used the BachChorale [6] dataset to do training and evaluation. It contains 54 chorale music pieces composed by J.S. Bach, each with a performance recording and the corresponding time-aligned MIDI file. After data cleaning, we divided the dataset into 37 training songs, five evaluation songs, and five test songs.

3.2 Baseline Model

B. Maman et al. proposed a DTW-based method to achieve unaligned supervision in training music transcription models [7]. They pre-trained the model with synthesized MIDI data to provide it with a standard pitch recognition ability. Then, the model was trained in a two-stage manner: first, to create the time-aligned training target, they aligned the ground truth with the transcription from the model; then, they used the aligned training target to train the model with normal binary cross-entropy loss and back-propagation. This is a get-away method from the problem of DTW's non-differentiable property, and we will compare its performance with the proposed soft-DTW trained model.

3.3 Training Settings

We use the soft-DTW loss with gamma starting at 3 and decreases $1e^{-4}$ every step. According to [8], decreasing gamma in soft-DTW can help stabilize the alignment path when we just start the training. The target MIDI sequence in the soft-DTW scenario is a standard MIDI with no tempo change or *fermata* notations, therefore the model and the soft-DTW need to find the true path by themselves.

The two models are trained with batch size of 4, learning rate of 0.0001.

3.4 Experiment Results

Table 1 shows the experiment result. We can see that the soft-DTW method outperforms the DTW-aligned model in frame activation precision and achieves a comparable, if not better, performance in onset detection. In figure 2 and 3, we can see that the precision of frame prediction is governed by the proposed method, and the onset detection performance of the proposed method is slightly better in most of the thresholds. The frame activation recall of the proposed method is lower than the DTW-aligned method, within an acceptable range. Generally, the numerical results show that soft-DTW significantly improves the precision of frame activation detection and is capable of acting as the training criteria with unaligned datasets in chorale transcription tasks.

Figure 4 shows an example of transcription. From top to bottom are the MIDI ground truth notation, DTW-aligned

Method	Frame Activation			Onset Activation (60ms)		
	Prec.	Recall	f1	Prec.	Recall	f1
DTW-aligned [7]	0.717	0.825	0.767	0.632	0.507	0.562
Soft-DTW(proposed)	0.589	0.844	0.694	0.626	0.491	0.550

Table 1. Precision, recall, and f1 score of the baseline and the proposed method. The decision threshold is set to 0.5, and the accepting window of the onset is set to 60ms.

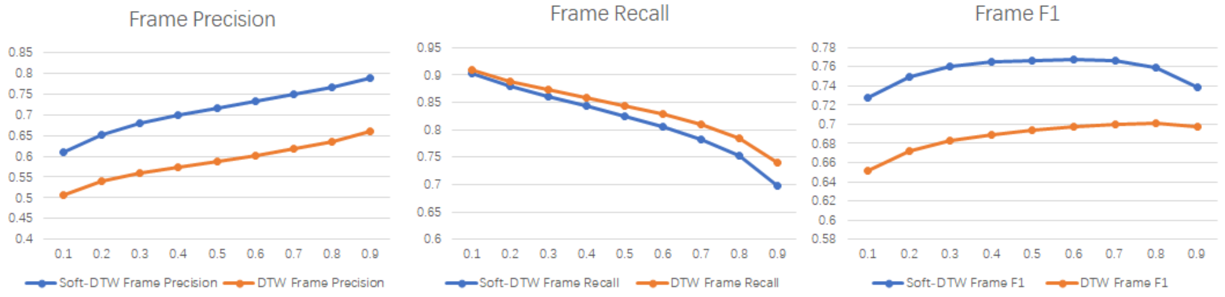

Figure 2. Frame activation precision, recall, and f1 score with different decision thresholds.

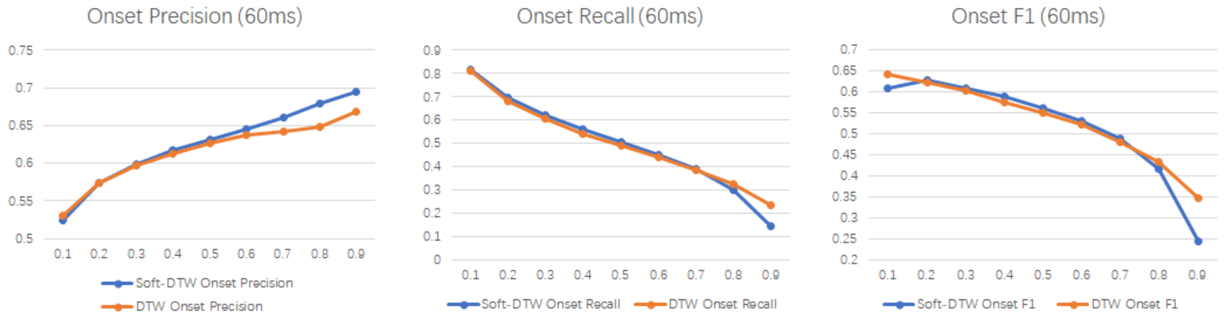

Figure 3. Onset activation precision, recall, and f1 score with different decision thresholds.

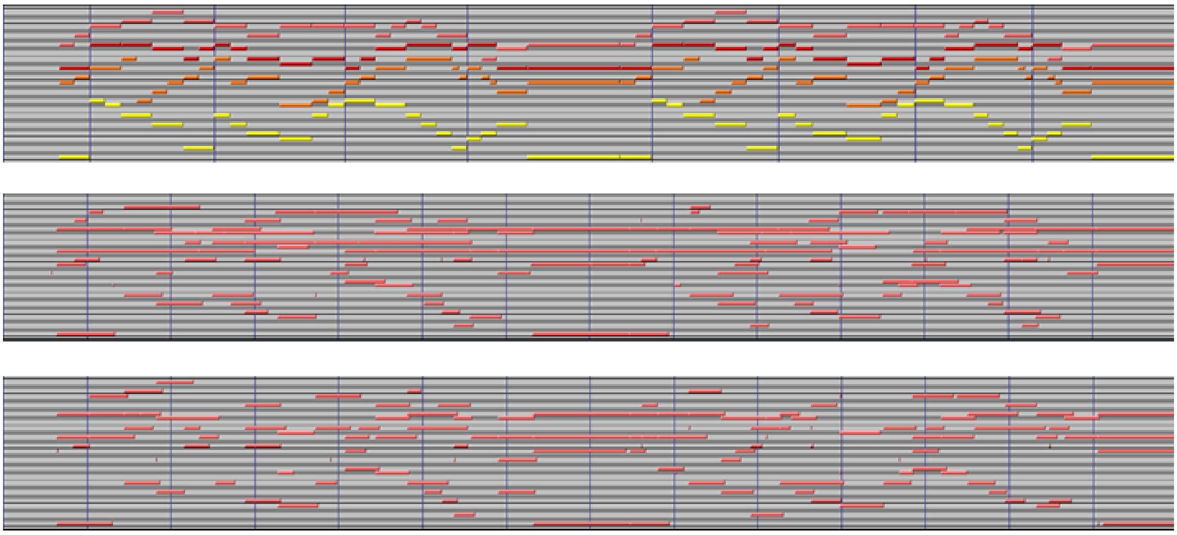

Figure 4. An example of transcription result. From top to bottom: MIDI ground truth notation, DTW-aligned model result, soft-DTW model result.

model result, soft-DTW model result. We can see the soft-DTW method reduces the false positives from the DTW-aligned method. It also reduces incorrect onset detections.

4. CONCLUSION AND FUTURE WORK

This paper shows that when people cannot find a time-aligned dataset to train transcription models, soft-DTW training loss can help them train a relatively good model with the unaligned dataset. To further prove the ability of the soft-DTW method to make use of large data, we need to collect more performance recordings and the corresponding notations and train the model with a larger corpus to see whether the performance will improve.

5. REFERENCES

[1] J. Narang, V. De La Vega, X. Lizarraga, O. Mayor, H. Parra, J. Janer, and X. Serra, "Choralsynth: Synthetic dataset of choral singing," *arXiv preprint arXiv:2311.08350*, 2023.

[2] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *International conference on machine learning*. PMLR, 2017, pp. 894–903.

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

[4] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[5] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–413.

[6] "Bach chorale dataset," https://www.pgmusic.com/bachchorales.htm.

[7] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *International Conference on Machine Learning*. PMLR, 2022, pp. 14 918–14 934.

[8] 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," *arXiv preprint arXiv:2308.05429*, 2023.