Friedrich-Alexander-Universität Erlangen-Nürnberg

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Internship Report

Report Title
Training with Unaligned Dataset: Soft Dynamic Time Warping

submitted by

Quang Hoang Nguyen Vo

submitted

September 2, 2025

Supervisors

Msc. Johannes Zeitler
Prof. Dr. Meinard Müller

AUDIO LABS

International Audio Laboratories Erlangen
A joint institution of the
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and
the Fraunhofer-Institut für Integrierte Schaltungen IIS.

Fraunhofer
IIS

Abstract

The evolution of Deep Neural Networks (DNNs) has shifted the paradigm of music information retrieval (MIR) from heuristic and mathematical models to data-driven approaches, which rely on large amounts of labelled training data. However, it introduces challenges when training with weakly aligned datasets. In this project, we investigate the characteristics of differential dynamic time warping (dDTW) through the soft-DTW (sDTW) algorithm when training with weakly aligned data. The main objective is to integrate soft-DTW as a loss function in the training process of a template-based chord recognition model. The dataset will have its chord label timestamps distorted or removed to simulate weakly or unaligned data. The dDTW loss function will then be used to train the model with the distorted dataset. The results will be compared with those obtained using the original dataset and the Connectionist Temporal Classification (CTC) loss function. Additional tasks may include experimenting and evaluating the performance of sDTW with different stablizing strategies.

Contents

Chapter 1

Introduction

Generally speaking, the introduction chapter should introduce the topic of the thesis and motivate the importance of it. Moreover, the introduction should give an outline of the thesis and point out the contributions of this work.

You can logically group the chapters of the thesis by using so-called parts. An example of how to insert a part-page containing a teaser-image is included in this template. Typically, you will have an introductory chapter that gives a broad overview. Then, the first part of the thesis might start after the introduction.

1.1 Report Organization

You can setup some layout options and elementary data in `config.tex`:

Chapter 2

Soft Dynamic Time Warping Algorithm

Each chapter should start with a small summary discussing the content and the relation of the sections. In this chapter, we elaborate on the theoretical background, foundations and concepts that are being used in the thesis. In particular, we focus on latex construct that are typically used in thesis documents. Section **??** illustrates the usage of math equations. Then, in Section **??**, some examples are given on how to cite literature. In Section 2.3, ...

When using labels, please avoid collisions in the label names.

2.1 Forward Pass

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. These three styles are used in the following paragraph along with labeling, aligning and referencing equations.

A little anecdote tells that Johann Carl Friedrich Gauß was able to compute the sum of numbers in the set $\{i \in \mathbb{N} \mid i \leq N\}$ in a very short amount of time at the age of nine. Apparently, he was able to observe that if N is even, then the numbers in the sum can be reordered and grouped so that

$$\begin{aligned}\sum_{i=1}^{N} i &= 1 + 2 + \ldots + N \\ &= (1 + N) + (2 + N - 1) + \ldots + (\frac{N}{2} + \frac{N}{2} + 1) \qquad (2.1) \\ &= \frac{N \cdot (N+1)}{2} \qquad (2.2)\end{aligned}$$

If N is even, then the reordering Equation (2.1) simply builds $N/2$ summands with value $N + 1$, resulting to Equation (2.2). This Equation also holds for odd N and a similar argument holds in this case.

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Table 2.1. Frequency of Special Characters. Note that this table does not contain any vertical lines which makes the table look more tidy.

2.2 Backward Pass

In a thesis one might want to cite interesting and relevant books [6, 5, 3], scientific papers [4, 2], or websites [1].

2.3 Tables

Because tables cannot, (well, at least not easily) be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper “floating” placement of tables, use the environment *table* to enclose the table’s contents and the table caption. The contents of the table itself must go in the *tabular* environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules, as seen in Table 2.1.

2.4 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper “floating” placement of figures, use the environment *figure* to enclose the figure and its caption. This sample thesis document contains examples of *.pdf* files to be displayable with LaTeX.

Typically, in addition to a *figure* environment, we use use PowerPoint that allows us to easily align and position subfigures as well as labels, as done in Figure 2.1. By the way, the curves shown in Figure 2.1 have been generated with MATLAB. Please find the generating sample script in `figures/figures.m`.

Including pixel graphics like in Figure 2.2 is also possible. As mentioned before, pixel graphics have to be converted to the allowed formats beforehand.

2.5 Theorem-like Constructs

Other common constructs that may occur in your thesis are the forms for logical constructs like theorems, axioms, corollaries and proofs. See the following example:

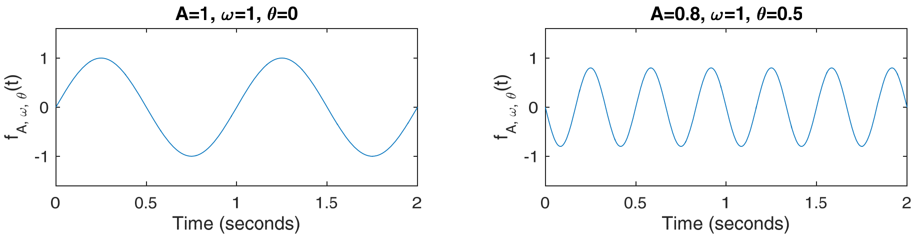

Figure 2.1. Sine functions with different amplitudes A, frequencies ω and phases θ can be calculated as $f_{A,\omega,\theta}(t) = A\sin(2\pi(\omega t - \theta))$.

Figure 2.2. Pixel graphics can also be included, but in low resolution it looks terrible and in high resolution is takes ages to load.

Theorem 2.1 *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a). \tag{2.3}$$

The following is a definition:

Definition 2.1 *If z is irrational, then by e^z we mean the unique number which has logarithm z:*

$$\log e^z = z. \tag{2.4}$$

Chapter 3

Experimental Setup

3.1 Data Preparation

The dataset will have its chord label timestamps distorted or removed to simulate weakly or unaligned data. The dDTW loss function will then be used to train the model with the distorted dataset. The results will be compared with those obtained using the original dataset and the Connectionist Temporal Classification (CTC) loss function. Additional tasks may include experimenting and evaluating the performance of sDTW with different stablizing strategies.

3.2 Model Architecture

The model architecture will be based on a standard sequence-to-sequence framework, with an encoder-decoder structure. The encoder will process the input audio features, while the decoder will generate the corresponding chord labels. Attention mechanisms may be employed to improve alignment between the input and output sequences.

3.3 Training Procedure

The training procedure will involve optimizing the model parameters using the dDTW loss function on the distorted dataset. The model will be evaluated on both the distorted and original datasets to assess its robustness to label noise. Hyperparameter tuning and regularization techniques may be applied to improve generalization.

3.4 Results

3.5 Discussion

In this section, we will discuss the implications of the results obtained from the experiments. We will analyze the performance of the proposed method in comparison to existing approaches and highlight its strengths and weaknesses. Additionally, we will explore potential avenues for future research and improvements.

Chapter 4

Conclusions

4.1 Limitation

4.2 Future Work

Draw the conclusions in the big picture of the thesis! Then, indicate future work.

Bibliography

[1] *Music free to download, print out, perform and distribute.* `http://www.mutopiaproject.org`, Retrieved 12.05.2009.

[2] B. Edler, S. Disch, S. Bayer, R. Geiger, and G. Fuchs, *A time-warped mdct approach to speech transform coding*, in Proceedings of the Audio Engineering Society Conference (AES), May 2009.

[3] E. A. P. Habets, *Single- and multi-microphone speechdereverberation using spectral enhancement*, PhD thesis, 2007.

[4] J. Herre and L. Terentiv, *Parametric coding of audio objects: Technology, performance and opportunities*, in Proceedings of the Audio Engineering Society Conference (AES), Ilmenau, Germany, 2011.

[5] M. Müller, *Information Retrieval for Music and Motion*, Springer Verlag, 2007.

[6] M. Müller, *Fundamentals of Music Processing – Audio, Analysis, Algorithms, Applications*, Springer Verlag, 2015.