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 3, 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.

Chapter 2

Soft Dynamic Time Warping Algorithm

2.1 Forward Pass

2.2 Backward Pass

2.3 Soft Alignment

Chapter 3

Experimental Setup

In this section, we present the objectives of our experiment, the dataset used for training alongside with the network architecture and the training process.

3.1 Dataset

For the experiment, we use the Beatles dataset retrieved from Isophonics [**?**], consisting of four audio recordings with respective annotations. Due to the simplicity of the current chord recognition network, we need to simplify the chord labels, so that the annotations only have major and minor chords. We split the dataset into training, validation, and test sets. For test set, a short segment of Let It Be is used, while the rest of the dataset is split into 3:1 ratio for training and validation.

We choose a sequence length of 150 samples for training and validating the model, creating 43 and 12 segments for training and validation, respectively. In case of soft alignment, we remove the adjacent repetitions in the sequence, effectively reducing its length by around 85% on average (see Figure **??**). However, this method introduce a problem where batching target sequences of different lengths is not possible. To address this issue, after reduction, we pad the sequences repeating each frame uniformly until they reach a desired length, or "soft length". After some experiments, we found that a soft length of 16 covers all of possible reduced sequences while keeping the reduction ratio high.

3.2 Model Architecture

Given the aim of this experiment is to evaluate the performance of the proposed SDTW loss function, the network architecture plays a minor role and are kept simple. Therefore, we used a simple chord recognition network (dChord) that based on the template-based chord recognition algorithm. This network consists of a single layer that acts as the chord template to predict a 24-dimensional chord label activation vector, corresponding to 12 chromas with their respective major or minor variant. Combined with log-compression and feature normalization layers, the network has a total of 25 trainable parameters. Table 3.1 illustrate the components of the

Layer	Input Dimension	Output Dimension	Parameters
Log-compression	(12, T)	(12, T)	0
Normalization	(12, T)	(12, T)	0
dChord	(12, T)	(24, T)	25
softmax	(24, T)	(24, T)	0

Table 3.1. Architecture of the chord recognition network. T is the number of time frames.

architecture with their respective input and output dimensions.

During training, we use Adam optimizer with a learning rate of 0.01.

3.3 Results and Discussion

Chapter 4

Conclusions

Bibliography