USING THE SYNC TOOLBOX FOR AN EXPERIMENT ON HIGH-RESOLUTION MUSIC ALIGNMENT

Yigitcan Özer, Michael Krause, Meinard Müller
International Audio Laboratories Erlangen, Germany
{yigitcan.oezer, michael.krause, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Music synchronization aims to automatically align multiple music representations such as audio recordings, MIDI files, and sheet music. For this task, we have recently published the Sync Toolbox [1], an open-source Python package for efficient, robust, and accurate music synchronization. This work combines spectral flux used as onset features with conventional chroma features to increase the alignment accuracy. We conduct some experiments within the Sync Toolbox framework to show that our approach preserves the accuracy compared with another high-resolution approach while being computationally simpler.

1. INTRODUCTION

In the field of music information retrieval (MIR), alignment techniques are central for several applications such as score following, content-based retrieval, automatic accompaniment, or performance analysis. Besides these applications, music synchronization has great potential for simplifying data augmentation, data annotation, and model evaluation. For example, music synchronization can be used to obtain additional training data for deep learning approaches semi-automatically.

The Sync Toolbox [1] is a recent open-source Python package that provides all the components of a music synchronization pipeline to produce state-of-the-art alignment results regarding efficiency and accuracy. It builds a complete system that allows users to reproduce research results from the literature and provides well-documented functions for all required basic building blocks for feature extraction, alignment, and evaluation. Its algorithmic core is based on the high-resolution multiscale dynamic time warping (DTW) approach from [2]. High-resolution chroma onset features from [3] are used per default on the finest layer of the alignment algorithm. These 12-dimensional chroma onset features are called *decaying locally adaptive normalized chroma onset features (DLNCO)* and have achieved state-of-the-art results, particularly in the synchronization of piano music.

Spectral flux (SF) is an alternative to DLNCO features for capturing onset information [4]. This work shows that using SF in combination with chroma features on the finest layer of the multiscale DTW pipeline results in a similar accuracy and robustness compared to the original DLNCO-based approach. As a main benefit, using one-dimensional SF discards the need for the elaborate computation of 12-dimensional DLNCO features. Moreover, we show that both feature combinations that comprise chroma and onset information outperform the alignment approach using only chroma features.

For the reproducibility of our experiments, we provide Jupyter notebooks and the Python implementation of the SF in our publicly available GitHub repository [1] .

2. SPECTRAL FLUX

In the following, we recall the computation of SF as described in [5]. A first-order differentiator is applied on the log-compressed magnitude spectrum of a music recording to compute the SF. Half-wave rectification follows the differentiation to keep only positive differences between subsequent frames. We subtract a local average function to enhance the peak structure and apply an additional temporal decay as a postprocessing step as in [3].

3. EXPERIMENTS

In this section, we evaluate audio alignments obtained using chroma and SF features on the Schubert Winterreise Dataset (SWD) [6] and compare them to alignments obtained using chroma features and the combination of chroma features with DLNCO features.

SWD comprises several representations of the song cycle *Winterreise* D911 (Op. 89), which consists of 24 songs composed for solo voice with piano accompaniment. For our experiments, we focus on the music recordings by the baritones Gerhard Hüsch and Randall Scarlata and the corresponding measure annotations. These two versions are publicly available, which allows reproducing all our experiments based on open-source code and open-source data.

As the synchronization result, DTW yields a *warping path* that indicates the alignment between two feature sequences. Our experiments use the resulting warping path

[1] https://github.com/meinardmueller/synctoolbox

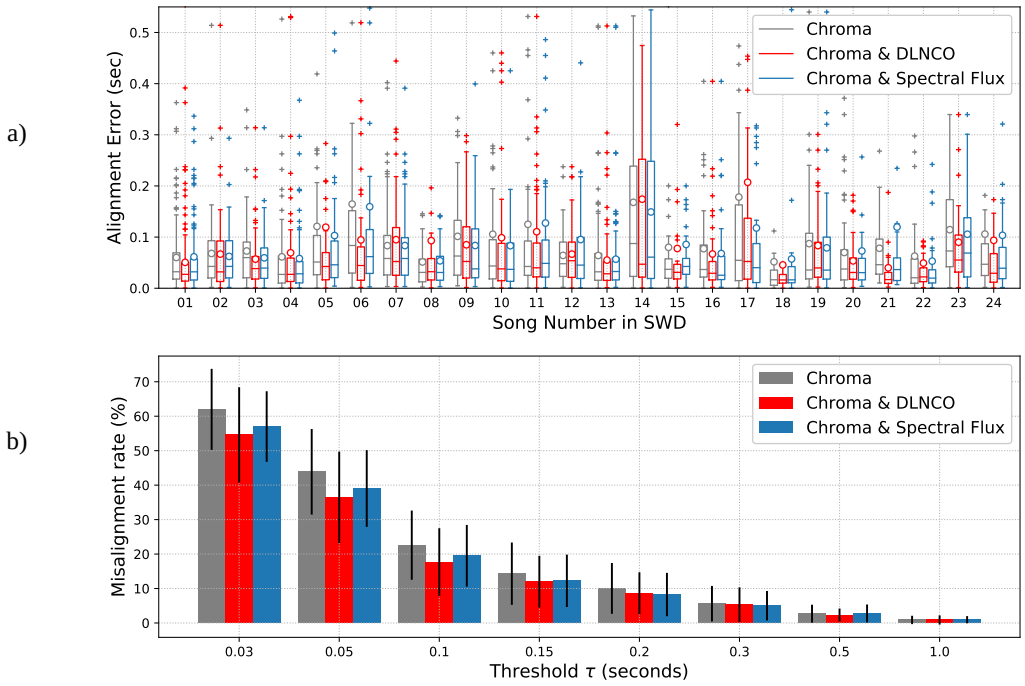

Figure 1: (a) Pairwise alignment error ϵ_P on a song-specific level, based on ground-truth measure annotations. The boxes are created from the first quartile (25th percentile) to the third quartile (75th percentile), and the horizontal line within the boxes indicates the median. The lower and upper whiskers respectively depict the minimum and maximum of the alignment errors. The symbol ○ indicates the mean alignment error, and + marks the outliers. Outliers above 0.55 seconds are not shown. (b) Mean and standard deviation over misalignment rates for all songs in the dataset, given a threshold τ.

to transfer the measure positions annotated for the first recording to the second. Since SWD has the annotations for all the music recordings, this allows us to evaluate the synchronization accuracy on the measure level.

For the evaluation, we utilize the pairwise alignment error ϵ_P from [7]. Given two versions of the same music piece with the time-continuous axes $[0, T_1]$ and $[0, T_2]$, the monotonous alignment can be modeled as a function

$$\mathcal{A} : [0, T_1] \to [0, T_2].$$

The pairwise alignment error ϵ_P for a given alignment of two recording is specified as the mean over the values

$$\epsilon_P(g_1) := |\mathcal{A}(g_1) - g_2|,$$

where $(g_1, g_2) \in [0, T_1] \times [0, T_2]$ indicates the ground-truth pairs of measure annotations.

Figure 1a shows the mean alignment error per song in the SWD song cycle for the alignment methods using different feature combinations. Using both chroma and onset features yields a better synchronization accuracy. Furthermore, the synchronization performance of the recordings, in which the singing voice is dominant, e.g., song No. 6, No. 14, and No. 23, are worse. Besides, the song No. 17 has a repetitive pattern in the accompaniment, which impedes the alignment with chroma-based features, while SF in combination with chroma features improves the synchronization accuracy.

In addition to the pair-wise alignment error, one may also consider the misalignment rate, which identifies the percentage of measure positions in an alignment with an error above a given threshold τ. Figure 1b illustrates that the alignment using chroma features combined with DLNCO or SF reveals higher synchronization accuracy for each threshold. At the same time, the misalignment rates are similar in both scenarios.

In conclusion, our experiments show that the combination of chroma features with DLNCO or SF as onset features improves the alignment results in a similar fashion. The advantage of SF lies in its simpler computation compared to 12-dimensional DLNCO features. As a good common practice [8], the source code as well as the data (audio, annotations) are publicly available and provided within a notebook linked to the Sync Toolbox. This allows for reproducing the entire experiment.

4. ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG MU 2686/7-2, MU 2686/11-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

5. REFERENCES

[1] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization," *Journal of Open Source Software (JOSS)*, vol. 6, no. 64, pp. 3434:1–4, 2021.

[2] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 569–573.

[3] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 1869–1872.

[4] P. Grosche, M. Müller, and S. Ewert, "Combination of onset-features with applications to high-resolution music synchronization," in *Proceedings of the International Conference on Acoustics (NAG/DAGA)*, 2009, pp. 357–360.

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

[6] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. Grohganz, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 25:1–18, 2021.

[7] T. Prätzlich and M. Müller, "Triple-based analysis of music alignments without the need of ground-truth annotations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, March 2016, pp. 266–270.

[8] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 128–137, 2019.