# A Hybrid Approach to Audio-to-Score Alignment

Ruchit Agrawal [1]    Simon Dixon [1]

## Abstract

Audio-to-score alignment aims at generating an accurate mapping between a performance audio and the score of a given piece. Standard alignment methods are based on Dynamic Time Warping (DTW) and employ handcrafted features. We explore the usage of neural networks as a preprocessing step for DTW-based automatic alignment methods. Experiments on music data from different acoustic conditions demonstrate that this method generates robust alignments whilst being adaptable at the same time.

## 1. Introduction and Motivation

Audio-to-score alignment is the task of finding the optimal mapping between a performance and the score for a given piece of music. Dynamic Time Warping (Sakoe & Chiba, 1978) has been the de facto standard for this task, typically incorporating handcrafted features (Dixon, 2005; Arzt, 2016). Recent advances in Music Information Retrieval have demonstrated the efficacy of Deep Neural Networks (DNNs) to a variety of tasks like music generation (Eck & Schmidhuber, 2002), audio classification (Lee et al., 2009), onset detection (Marolt et al., 2002), music transcription (Marolt, 2001; Hawthorne et al., 2017) as well as music alignment (Dorfer et al., 2018a). The primary advantage of DNNs is that they can learn directly from data in an end-to-end manner, thereby eschewing the need for complex feature engineering. However, DNNs struggle with modelling long-term dependencies (Bengio et al., 1994) in temporal sequences. End-to-end alignment is a challenging task since it incorporates dealing with multiple inputs of different modalities, in addition to handling of very long term dependencies. This paper is an endeavor towards employing neural networks for music alignment. We present a hybrid approach to audio-to-score alignment, which consists of a neural network based preprocessing step as a precursor to

Dynamic Time Warping. This approach involves computing a frame similarity matrix which is then passed on to a DTW algorithm that computes the optimal warping path through this matrix. The advantage of our method is that the preprocessing step is trainable, thereby making our method adaptable to a particular acoustic setting, unlike traditional DTW-based methods which employ handcrafted features.

## 2. Related Work

Early works on feature learning for MIR tasks employ algorithms like Hidden Markov Models (Joder et al., 2013) or deep belief networks (Schmidt et al., 2012). Recently, a number of works have explored feature learning for MIR using deep neural networks (Sigtia & Dixon, 2014; Oramas et al., 2017; Thickstun et al., 2016; Lattner et al., 2018; Arzt & Lattner, 2018; Korzeniowski & Widmer, 2016). Work specifically on learning features for audio-to-score alignment has mainly focused on an evaluation of current feature representations (Joder et al., 2010), learning of the mapping for several common audio representations based on a best-fit criterion (Joder et al., 2011) and learning transposition-invariant features (Arzt & Lattner, 2018) for alignment. (Hamel et al., 2013) propose transfer learning for MIR tasks by learning learn a shared latent representation across related tasks of classification and similarity detection. Weaknesses in standard approaches to choosing similarity thresholds has been explored in (Kinnaird, 2017). (İzmirli & Dannenberg, 2010) propose the idea of learning features for aligning two sequences of music, as opposed to employing a standard chroma-based feature representation. (Nieto & Bello, 2014) present a novel algorithm to capture music segment similarity using two-dimensional Fourier-magnitude coefficients. (Korzeniowski & Widmer, 2016) explore frame-level audio feature learning for chord recognition using artificial neural networks.

## 3. Experiments and Results

The standard feature representation choice for music alignment is a time-chroma representation generated from the log-frequency spectrogram. Since this representation only relies on pitch class information, it ignores variations in timbre and instrumentation, and is not adaptable to different acoustic settings. Using neural networks helps us to

override the manual feature engineering whilst providing the capability to adapt to different settings. Rather than extracting a feature representation from the inputs, we focus on the task of constructing a frame-similarity matrix. This matrix is then passed on to a DTW-based algorithm to generate the alignments. We employ a "Siamese" Convolutional Neural Network (Bromley et al., 1994) for this task. This framework has shown promising results in the field of computer vision for computing image similarity (Zagoruyko & Komodakis, 2015), as well as in the field of natural language processing, for learning sentence similarity (Mueller & Thyagarajan, 2016) and speaker identification (Lukic et al., 2016) amongst others.

We train our Siamese CNN model to determine if two patches, one from the audio and one from the synthesized MIDI "match" or not. A similar approach has been used by (İzmirli & Dannenberg, 2010), however they use a Multi-Layer Perceptron (MLP) framework to compute if two frames are the same or not. In addition to using an enhanced framework which is optimal for this task, our work differs from them in that we also compute similarity labels (non-binary) and use this distance matrix further for alignment. We explain the preprocessing steps below:

In order to keep the modality constant, we first convert the MIDI files to audio using FluidSynth (Henningsson & Team, 2011). We then transform the frame-level audio patches to image spectrograms using librosa (McFee et al., 2015), a Python library for audio and music analysis. We conduct experiments using both the Short-Time Fourier transform (STFT) as well as the Constant-Q transform (CQT) transformations of the raw audios. We briefly explain our choice of loss function below:

The objective of our Siamese CNN model is not to classify the inputs, but to differentiate between them. Hence, a contrastive loss function is much better suited to this task than a standard classification loss function like cross entropy. The contrastive loss function is computed as follows:

$$(1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

where $D_w$ is the Euclidean Distance between the outputs of the two Siamese twin networks. More formally, $D_w$ can be expressed as follows:

$$\sqrt{\{G_W(X_1) - G_W(X_2)\}^2}$$

where $G_w$ is the output of each of the twin networks and $X_1$ and $X_2$ are the two inputs.

We train the model on the MAPS database (Emiya et al., 2010), where we have MIDI-aligned audio for a range of acoustic settings. We only select the subset containing the recordings played using a real piano, and discard the ones which are software-synthesized. We compute the similarity matrix using two mechanisms:

- Using binary labels: For this we employ the output labels of our Siamese CNN. 0's imply similar pairs, 1's imply dissimilar pairs.

- Using distances: For this we calculate $D_w$. The distance directly corresponds to the dissimilarity between the two inputs. Higher the value of $D_w$, higher the dissimilarity.

We then generate an alignment path through this matrix using fast-DTW (Salvador & Chan, 2007), through a readily available DTW implementation in Python [1]. We test the performance of our model on a subset of the Mazurka dataset (Sapp, 2007) which contains recordings from various acoustic settings. The results obtained using both methods are given in Table 1.

*Table 1.* Alignment accuracy (in %)

| Type of matrix | STFT | CQT |
|----------------|------|------|
| Binary | 76.3 | 78.6 |
| Distance | 78.1 | 81.4 |

Our results suggest that this method is a promising approach to alignment, especially in non-standard acoustic conditions; since the pre-processing Siamese Network is trainable on such data, unlike the manually handcrafted features used by standard DTW-based algorithms.

## 4. Conclusion and Future Work

We demonstrated that our hybrid approach to audio-to-score alignment is capable of generating robust alignments across various acoustic conditions. The advantage of our method is that it is adaptable to a particular acoustic setting without requiring a large amount of labeled training data. In the future, we would like to conduct an exhaustive evaluation of this approach on musically relevant parameters and analyze its limitations. We would also like to work on learning the features as well as the alignments in a completely end-to-end manner.

[1]https://pypi.org/project/fastdtw/

# References

Arzt, A. *Flexible and robust music tracking*. PhD thesis, Ph. D. thesis, Universität Linz, Linz, 2016.

Arzt, A. and Lattner, S. Audio-to-score alignment using transposition-invariant features. *arXiv preprint arXiv:1807.07278*, 2018.

Arzt, A., Widmer, G., and Dixon, S. Adaptive distance normalization for real-time music tracking. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2689–2693. IEEE, 2012.

Bell, S. and Bala, K. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 34(4):98, 2015.

Bengio, Y., Simard, P., Frasconi, P., et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pp. 737–744, 1994.

Carabias-Orti, J. J., Rodríguez-Serrano, F. J., Vera-Candeas, P., Ruiz-Reyes, N., and Cañadas-Quesada, F. J. An audio to score alignment framework using spectral factorization and dynamic time warping. In *ISMIR*, pp. 742–748, 2015.

Dixon, S. An on-line time warping algorithm for tracking musical performances. In *IJCAI*, pp. 1727–1728, 2005.

Dorfer, M., Arzt, A., and Widmer, G. Towards score following in sheet music images. *arXiv preprint arXiv:1612.05050*, 2016.

Dorfer, M., Arzt, A., and Widmer, G. Learning audio-sheet music correspondences for score identification and offline alignment. *arXiv preprint arXiv:1707.09887*, 2017.

Dorfer, M., Hajič Jr, J., Arzt, A., Frostel, H., and Widmer, G. Learning audio–sheet music correspondences for cross-modal retrieval and piece identification. *Transactions of the International Society for Music Information Retrieval*, 1(1), 2018a.

Dorfer, M., Henkel, F., and Widmer, G. Learning to listen, read, and follow: Score following as a reinforcement learning game. *arXiv preprint arXiv:1807.06391*, 2018b.

Eck, D. and Schmidhuber, J. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103, 2002.

Emiya, V., Bertin, N., David, B., and Badeau, R. Maps-a piano database for multipitch estimation and automatic transcription of music. 2010.

Hamel, P., Davies, M., Yoshii, K., and Goto, M. Transfer learning in mir: Sharing learned latent representations for music audio classification and similarity. 2013.

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

Henningsson, D. and Team, F. D. Fluidsynth real-time and thread safety challenges. In *Proceedings of the 9th International Linux Audio Conference, Maynooth University, Ireland*, pp. 123–128, 2011.

Hu, N., Dannenberg, R. B., and Tzanetakis, G. Polyphonic audio matching and alignment for music retrieval. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 185–188. IEEE, 2003.

İzmirli, Ö. and Dannenberg, R. B. Understanding features and distance functions for music sequence alignment. In *ISMIR*, pp. 411–416. Citeseer, 2010.

Joder, C., Essid, S., and Richard, G. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 409–412. IEEE, 2010.

Joder, C., Essid, S., and Richard, G. Optimizing the mapping from a symbolic to an audio representation for music-to-score alignment. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 121–124. IEEE, 2011.

Joder, C., Essid, S., and Richard, G. Learning optimal features for polyphonic audio-to-score alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2118–2128, 2013.

Kinnaird, K. M. Examining musical meaning in similarity thresholds. In *ISMIR*, pp. 635–641, 2017.

Korzeniowski, F. and Widmer, G. Feature learning for chord recognition: The deep chroma extractor. *arXiv preprint arXiv:1612.05065*, 2016.

Lattner, S., Grachten, M., and Widmer, G. Learning transposition-invariant interval features from symbolic music and audio. *arXiv preprint arXiv:1806.08236*, 2018.

Lee, H., Pham, P., Largman, Y., and Ng, A. Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pp. 1096–1104, 2009.

Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. Speaker identification and clustering using convolutional neural

networks. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6. IEEE, 2016.

Mandel, M. I., Poliner, G. E., and Ellis, D. P. Support vector machine active learning for music retrieval. *Multimedia systems*, 12(1):3–13, 2006.

Marolt, M. Sonic: Transcription of polyphonic piano music with neural networks. In *Workshop on Current Research Directions in Computer Music*, pp. 217–224, 2001.

Marolt, M., Kavcic, A., and Privosnik, M. Neural networks for note onset detection in piano music. In *Proceedings of the 2002 International Computer Music Conference*, 2002.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

Mueller, J. and Thyagarajan, A. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Muller, M., Ewert, S., and Kreuzer, S. Making chroma features more robust to timbre changes. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1877–1880. IEEE, 2009.

Nieto, O. and Bello, J. P. Music segment similarity using 2d-fourier magnitude coefficients. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 664–668. IEEE, 2014.

Oramas, S., Nieto, O., Barbieri, F., and Serra, X. Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint arXiv:1707.04916*, 2017.

Pampalk, E., Dixon, S., and Widmer, G. On the evaluation of perceptual similarity measures for music. In *of: Proceedings of the sixth international conference on digital audio effects (DAFx-03)*, pp. 7–12, 2003.

Pons, J., Nieto, O., Prockup, M., Schmidt, E., Ehmann, A., and Serra, X. End-to-end learning for music audio tagging at scale. *arXiv preprint arXiv:1711.02520*, 2017.

Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1): 43–49, 1978.

Salvador, S. and Chan, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

Sapp, C. S. Comparative analysis of multiple musical performances. In *ISMIR*, pp. 497–500, 2007.

Schmidt, E. M., Scott, J. J., and Kim, Y. E. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *ISMIR*, pp. 325–330. Citeseer, 2012.

Sigtia, S. and Dixon, S. Improved music feature learning with deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6959–6963. IEEE, 2014.

Thickstun, J., Harchaoui, Z., and Kakade, S. Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*, 2016.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Zagoruyko, S. and Komodakis, N. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4353–4361, 2015.