

REFINING AUDIO-TO-SCORE ALIGNMENT FOR SINGING VOICE TRANSCRIPTION

Miguel P. FERNANDEZ (miguel.perez01@estudiant.upf.edu) (0000-0001-8437-4517)^{1,2},
Holger KIRCHHOFF (0000-0001-8437-4517)¹, **Peter GROSCHKE** (0000-0001-8437-4517)¹, and
Xavier SERRA (0000-0001-8437-4517)²

¹Huawei Research Center, Munich, Germany

²Music Technology group, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Note-level automatic singing transcription, which involves extracting both time boundaries and accurate pitch from a singing voice, remains a significant challenge in Music Information Retrieval (MIR). Despite advancements in deep learning, progress is constrained by the labor-intensive task of annotating datasets, leading to ongoing data scarcity. To tackle this, we introduce a novel audio-to-score alignment algorithm that effectively synchronizes timed events between score and audio. Our method not only delivers precise alignments but also includes a mechanism to evaluate their reliability. Using this approach, we developed a Singing Onset Labels Extracted Automatically (SOLEA) dataset, which facilitated training a model on a diverse range of musical genres and achieved state-of-the-art performance in singing onset estimation for pop music. We plan to make both the alignment algorithm and the SOLEA dataset publicly available for use by other researchers.

1. INTRODUCTION

Automatic Music Transcription (AMT) is a fundamental task in MIR that involves identifying the notes in an audio recording. Each note has three main attributes: onset and offset (which indicate when the note starts and ends) and pitch [1, 2]. A particularly difficult area in AMT is Singing Transcription from Polyphonic Music (STP), which deals with transcribing a single singing voice mixed with multiple instruments. In this challenging scenario, where the voice is blended with other instruments, deep-learning methods have been more successful than their signal processing counterparts [3, 4]. Several of these Deep Learning (DL) methods achieve high pitch extraction metrics, with some relying solely on unlabeled data [5, 6]. However, accurately extracting note boundaries, especially onsets, remains a significant challenge [3, 4]. Training models for precise onset detection requires datasets with accurate alignment between audio and labeled notes. Traditionally, creating annotated datasets for AMT has been a slow manual process that results in small datasets. While these

smaller datasets were enough for evaluating signal processing methods, modern deep-learning approaches need much larger datasets for training. Currently, there are only two public datasets for STP: MIRST500 [7] and N20EMv2 [8], which together provide only 43 hours of music approximately. Such reduced datasets prevents from training large models as they might easily overfit the training data. In contrast, instruments like piano have extensive datasets, such as the 198-hour Maestro dataset [9], made possible by mechanisms that directly link audio performance with piano key presses. Unfortunately, such methods are not applicable to all instruments. Some AMT researchers have turned to synthesized datasets [10], which have been effective in certain scenarios [11]. However, for singing voice, synthesized data alone has not been sufficient due to the difficulty of timing onsets of human vocals [4]. Some AMT researchers have used audio-to-score alignment techniques, but these methods are neither extent from problems [12]; concretely, when the score and audio are automatically retrieved and matched from different sources, difficulties arise in ensuring the score belongs to that specific performance contained in the audio [13, 14].

In this paper, we contribute in two ways. First, we propose a new audio-to-score alignment method. Our approach defines events as specific musical occurrences at precise times, such as an onset at 5.8s, a drum beat at 4.6s, etc. Using an initial alignment as a reference, our method matches events with similar characteristics in both the onset and the score. Based on the amount of matched events, we employ some heuristics to determine if the annotations of certain parts of the audio are reliable enough for training. Our second contribution is a dataset of Singing Onset Labels Extracted Automatically (SOLEA), created by re-aligning the DALI dataset [15] with our proposed algorithm. Our results demonstrate that our proposed alignment method is accurate and effective for creating ready-to-use datasets for AMT.

2. RELATED WORK

In the current literature, a common approach to address data scarcity is leveraging unlabeled data. A widely used technique is the noisy teacher-student method: in [16], the authors used teacher-student training on a model trained with labeled singing voice data to predict pitch and note boundaries for unlabeled data. The notes generated by the

Copyright: © 2025. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

teacher model are used to train a student network. However, pseudo-labeling can lead to biased models [17], as the quality of pseudo-labels depends on the initial training data. Another approach is using weakly-aligned data, where onset timings are not precise. In [3], the authors found that the Connectionist Temporal Classification (CTC) loss [18] can train a singing transcription system with weakly aligned data. Building on this, [4] used advanced singing voice synthesizers to create new training data. These synthesis models have made significant strides [19], producing realistic human-like voices. This synthesis process enables the generation of data that accurately resembles real-world music scenarios. However, the generated data lacks precise note timings, so it must be used with the CTC method proposed in [3]. A limitation of this method is that it teaches the model to predict onsets and pitches in the correct order but not with precise timings. Labeled data with accurate onset labels is still needed to train singing transcription systems.

Given the ongoing need for labeled data for onset detection, some researchers in AMT have turned to audio-to-score alignment, where the goal is to match the timing between the audio and the score. By scraping audio and scores separated from multiple source in the internet [13, 14, 20] researchers can develop large datasets for music transcription. Most alignment methods follow a similar procedure [12]: features are extracted from the audio at a constant rate, then the score is synthesized and the same features are extracted. Finally, a cost matrix is created to reflect the similarity between the audio and score features. Dynamic Time Warping (DTW) [21] is then applied to the cost matrix to obtain a warping path W that aligns the score time i with the audio time j . The choice of features is crucial; some provide robust global alignments but fail on small local timings and vice versa [22, 23]. To address this, [23] proposes using multiple features, creating and combining various cost matrices before applying DTW. Leveraging this method, [13] combined chroma and onset activations functions to create a violin transcription dataset. The algorithm aligns raw network predictions, not onset events directly. Other alignment algorithms do match events from the score and audio but have drawbacks. For example, [24] relies on bar information, which is not available in all datasets, and [22] can not align events that have multiple matches in the audio or vice versa. In this work, we propose an alignment algorithm that only require a reference alignment to match events. Moreover, we propose to use the alignment produced by the reference iteratively to yield finer alignments between scores and audios.

3. EXPERIMENTS

3.1 Model training

A common DL approach to AMT consists of training a model to make framewise predictions for pitch and onset activations [9, 11]. Such framewise predictions are then combined and transformed into discrete notes. Following [4], we leverage a Wav2Vec2.0 [25] pretrained for speech and we finetune it to obtain 20 logits at a rate of 50Hz:

onset activation, vocal activity, 12 pitch classes, and 6 octaves. We create the target labels, employ the same losses, and create notes from framewise predictions as in [4]. For those frames where there is no active pitch, we mask the losses for pitch class and octave. A batch size of 25 is employed, being each audio sample a chunk of 6 seconds. For audio chunks of less than 6 seconds, we pad both audio and target labels, and mask the loss produced by such padded targets so that they do not contribute to the final loss. The metrics employed are CO_n , CO_nP , CO_nPOff , with *mir_eval* [26] as usually done in STP [3, 4, 7]; these reflect increasing levels of difficulty, as the first one requires to only retrieve the correct onset, the second one also requires the correct pitch, and the latest one also the correct offset. To select the onset and silence threshold in our system we perform a grid search in the evaluation set, as in other works [3, 4]. We perform an evaluation every 325 training steps (approx 13.5 hours of audio), and apply early stopping if the CO_nP did not improve in the validation set for 5 consecutive evaluations. We choose CO_nP rather than CO_nPOff because of two reasons: the proposed dataset provides only the position of the onsets, therefore we do not expect any improvement in offset estimation; secondly, determining the precise time when a note ends is very subjective, moreover, different criteria was used on two of the main STP datasets on when a note should be considered to end [7, 8]. We use Adam [27] with a learning rate of $5e^{-5}$. Models were trained on a V100/32GB using 16 bit mixed precision, leading to training times from 4h to 12h depending on the datasets used for training.

3.2 Aligning events

We build on the DALI dataset [15] to create SOLEA. DALI contains 7,756 popular songs with lyrics and note annotations, spanning 63 music genres and 32 different languages, opening a great opportunity for singing related tasks. However, while it has been quite popular for lyrics transcription [8], its adoption for STP has been lower. Moreover, previous attempts of using it as a dataset for this task failed [7]. Three reasons are likely behind this: first, the dataset is based on crowd-sourced pop song annotations for karaoke scores, which are matched with YouTube videos. As those annotations are not necessarily made by trained musicians, many of the annotated note pitches are wrong, as well as some notes having incorrect durations [7]. Secondly, the procedure employed to align the audio and scores of the DALI dataset resulted in an alignment that although precise enough for the task of lyrics transcription, does not match the strict timings required for STP as we also show later in our experiments. Lastly, is common in large automatically retrieved AMT datasets that, although the audio and score may correspond to the same song, certain sections diverge [14, 15], such as instances where the audio originates from a different performance of the song with variations in elements like the chorus. Our proposed method addresses these challenges by providing an identification method of reliably finding corresponding sections within the audio and score.

Our method aligns a sequence of events in a reference

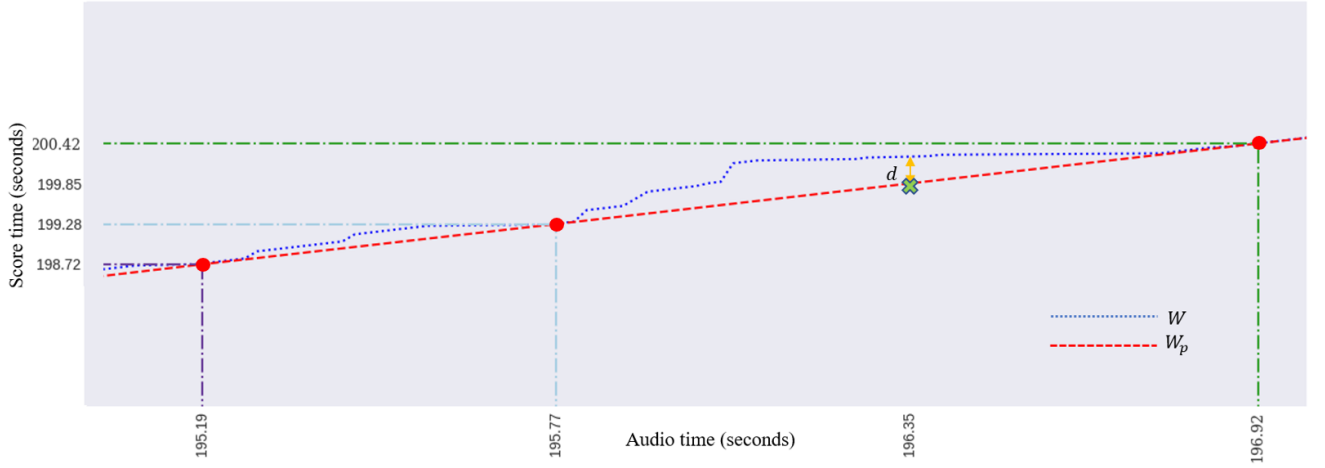


Figure 1: The figure shows a reference alignment W and the timestamps for 4 events in the audio, and 4 in the score. For each combination of events in score and audio, we calculate a distance $d = |s_i^t - W(a_j^t)|$ reflecting how their alignment diverge from W . In this example, 3 out of 4 onset events in the score were matched to audio onsets. These matched pairs are represented with red circles. The third event in both the audio and the score were left unmatched. If matched, they would have resulted in a new point of alignment between the audio and the score, marked as a green cross. However, the distance from such green cross to W is large, indicating a disagreement between the potential alignment provided by the events and W . We therefore consider that such pair of events can not be reliably associated. The matched events result in a new warping path W_p . When aligning iteratively, these unmatched events can be associated when using W_p as reference.

| Dataset | # Songs | # Chunks | # Onsets | Duration |
|--------------|---------|----------|----------|----------|
| SOLEA* | 4,763 | 23.4k | 280,812 | 39.1h |
| SOLEA | 5,249 | 25.8k | 303,238 | 42.5h |
| N20EMv2 [8] | 157 | — | 38,857 | 8.4h |
| HSD [28] | 55 | — | 20,399 | 3.8h |
| MIRST500 [7] | 500 | — | 162,438 | 34.4h |
| DALI [15] | 7,756 | — | > 1.6M | 488.1h |

Table 1: Overview of different datasets for singing voice transcription

score with events automatically extracted from the audio. We associate events based on a reference alignment W providing a global correspondence between the audio and the score. Such alignment might be manually annotated or automatically extracted with any of the methods in Section 2. The events from the audio and the score are aligned while minimizing the discrepancy between the aligned events, and the reference W . An example of our algorithm can be seen in Figure 1. Let an event be defined as $e^{t,l}$ where t is the timestamp of the event and l is the label associated with it (e.g., an onset, beat, downbeat, etc). We subsequently define the events in the score as $S = \{s_1^{t,l}, s_2^{t,l}, \dots, s_{|S|}^{t,l}\}$, and $A = \{a_1^{t,l}, a_2^{t,l}, \dots, a_{|A|}^{t,l}\}$ as the events in the audio. The first step in our algorithm is to construct a “matching” matrix $B^{|S| \times |A|}$ that reflects which events from S can be matched to events from A . This matrix contains 1’s where an event $s_i^{t,l}$ can be associated with an $a_j^{t,l}$, as detailed in Equation 1.

$$B(i, j) \begin{cases} 0, & \text{if } s_i^l \neq a_j^l \\ 0, & \text{if } |s_i^t - W(a_j^t)| > \varepsilon \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

Where $i < |S|$ and $j < |A|$; s_i^t and a_j^t are the timestamps at which the events occur in the score and the audio, while s_i^l and a_j^l are their associated labels. The threshold ε controls for how much the alignment between the events is allowed to differ from the alignment provided by W . If the difference is bigger than ε , we consider that such pair of events can not be matched. In summary, Equation 1 reflects that events with different labels can not be associated, and neither those whose alignment differ notably from the reference W , with the hyper-parameter ε allowing to control how strict should be the agreement between the aligned events and W .

There can still be undesired cases where an event in the audio is matched to multiple events in the score, and vice versa. E.g. an onset extractor predicts two onsets in a region where only one score onset exists, revealing a case of a false onset prediction. Therefore, after obtaining the matrix indicating the feasibility of two events being related, we process it to maximize the number of matches between the audio and the score with the following conditions: no audio event is associated with more than one score event and vice-versa, and all the matched events occur in increasing order. The matched audio-score events will be used to construct the new warping path W_p as observed in Figure 1, using linear interpolation in-between them.

This new W_p can then serve as a refined reference alignment, enabling the possibility of iterative alignment. That is, W_p can be used as the reference for further alignment, potentially matching additional events where W_p proves to be more accurate than the initial alignment W . An example of this process is illustrated in Figure 1.

| Training set | SOLEA? | MIRST500 | | | N20EMv2 | | | HSD | | |
|----------------------------|----------|-----------------|-------------------|----------------------|-----------------|-------------------|----------------------|-----------------|-------------------|----------------------|
| | | CO _n | CO _n P | CO _n POff | CO _n | CO _n P | CO _n POff | CO _n | CO _n P | CO _n POff |
| DALI [15] | NA | 74.8 | — | — | 51.7 | — | — | 76.8 | — | — |
| SOLEA | NA | 81.1 | — | — | 62.0 | — | — | 78.7 | — | — |
| MIRST500 | X | 81.5 | 75.4 | 59.8 | 65.7 | 56.0 | 34.0 | 78.6 | 72.5 | 44.8 |
| | ✓ | 82.4 | 77.6 | 60.4 | 66.4 | 49.6 | 31.3 | 79.3 | 73.1 | 47.3 |
| MIRST500 + N20EMv2 | X | 81.9 | 76.8 | 60.5 | 84.3 | 69.1 | 59.0 | 78.4 | 72.6 | 44.0 |
| | ✓ | 82.3 | 77.5 | 60.9 | 83.8 | 67.8 | 57.6 | 79.9 | 73.5 | 44.8 |
| N20EMv2 | X | 77.0 | 64.6 | 47.2 | 84.1 | 66.6 | 54.1 | 75.8 | 63.2 | 33.7 |
| | ✓ | 79.3 | 61.4 | 46.9 | 82.0 | 64.3 | 51.3 | 76.8 | 52.0 | 28.6 |
| CTC+CE [3] | NA | 79.6 | 74.3 | 57.4 | 66.2* | 54.9* | 17.8* | 74.5* | 67.8* | 42.5* |
| T3MS [29] | NA | 80.6 | 77.1 | 61.0 | — | — | — | 78.2 | 73.4 | 51.4 |
| Wav2Vec _{N20} [8] | NA | 78.0 | 70.0 | 52.4 | 93.6 | 79.6 | 74.1 | — | — | — |

Table 2: Results of our experiments along some baselines. We show the effects that adding SOLEA to the training has when evaluating on different datasets. The last three rows indicate some baselines for comparison. Whenever possible we used the officially available trained models to obtain results for the other datasets. Best results are highlighted in bold.

3.3 The SOLEA dataset

We apply our described method to obtain reliable singing voice onsets annotations from the DALI dataset. First, we train a system as detailed in Subsection 3.1 on the MIRST500 and N20EMv2 datasets, with this we obtain a reference alignment W as in [13]. We then obtain onset events from the DALI audio, and follow the alignment method described in Subsection 3.2 to synchronize the audio onsets with the score onsets. A perfect alignment for the entirety of the audio and score cannot always be ensured, and it's possible that annotations are accurate only for certain segments of the audio. Thus, SOLEA is not constituted by entire songs, but we rather use our method to extract segments of at least 2.5 seconds in which all audio onsets can be matched with onsets in the score (with $\varepsilon = 75ms$). These aligned and filtered audio chunks form the SOLEA dataset. We use this dataset, together with MIRST500 and N20EMv2, to train a model for singing voice transcription. A summary of the alignment results along some other relevant datasets is presented in Table 1. An asterisk (*) indicates that the iterative process, described at the end of Subsection 3.2, was not applied. Including this iterative process results in a significantly larger number of aligned audio and score chunks. This implies that the W_p yielded by our algorithm reflects a higher agreement between the newly obtained alignment and the events present in both audio and score. The other datasets included in the table are MIRST500, HSD, and N20EMv2. MIRST500 is a widely used dataset for STP, featuring approximately 34 hours of professionally produced music sourced from YouTube. Notice that MIRST500 provides a training/test split without any official validation split. In our work we reserve the last 30 songs from the original training set to be used for validation. HSD [28] is also sourced from YouTube from commercial music equally to MIRST500, it has however, a smaller size and therefore we use it only for testing. We process the dataset as it was proposed in [29] to obtain the correct labels. Lastly, we added N20EMv2 to our dataset list, which differs from the

previous datasets in two key ways: it was recorded by the dataset authors instead of relying on YouTube links, and its singers are predominantly amateur.

As mentioned in 3.2, many notes in DALI were annotated with incorrect lengths. A drawback of our method, is that although it can align precisely on the onset positions, it can not ensure the alignment at other points i.e. offset positions. Whenever sampling from SOLEA or DALI we train only the onset detector of our model, to avoid possible noisy data during training. Consequently, in scenarios where only these datasets are used for training, we select the checkpoint for evaluation based on the best CO_n as only the onset detector is trained.

4. RESULTS

The experimental results presented in Table 2 illustrate the impact of training on SOLEA alone as well as in combination with other datasets. For comparison, we also add the results of training on DALI with its given alignment. The results demonstrate that training on SOLEA yields better onset detection than training in DALI [15], with our proposed dataset performing on par with, or even outperforming, manually annotated datasets in some scenarios. Specifically, either using SOLEA alone or along any of the other datasets notably improves the results for both HSD and MIRST500. An exception occurs with N20EMv2: here, training on either MIRST500 or SOLEA results in a poor performance, and only when using N20EMv2's own training set do the results improve. Conversely, when only using N20EMv2's training set, adding SOLEA notably improves the results obtained for MIRST500 and HSD. This points that N20EMv2 represents very different kinds of dataset, and that SOLEA is more similar to MIRST500 and HSD than to N20EMv2. CO_nP consistently yields lower values than the less restrictive CO_n . For MIRST500 and HSD, the gap between these metrics is typically around 5-6 points, but the difference becomes more pronounced when evaluating N20EMv2.

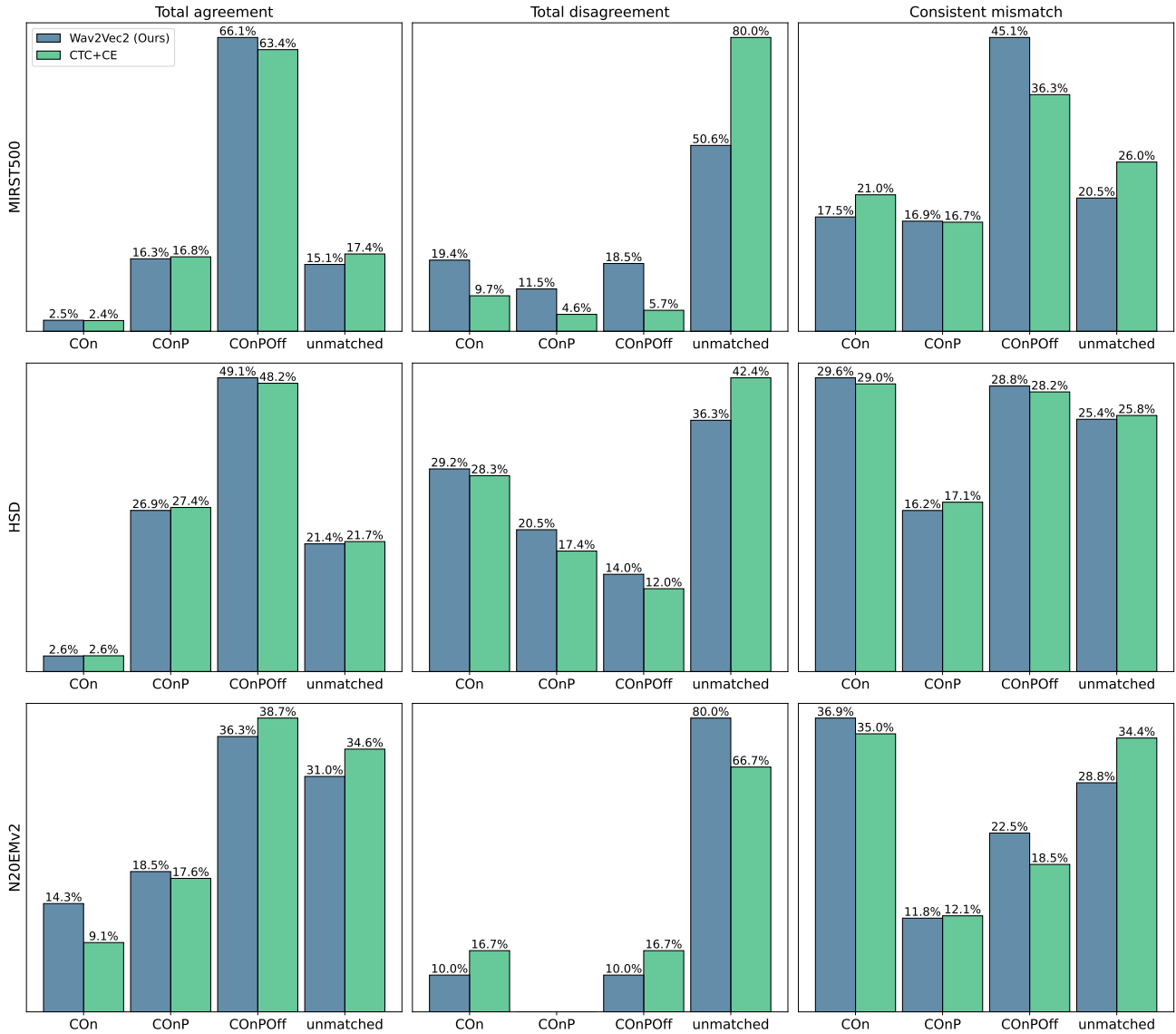


Figure 2: Effect of agreement cases on transcription performance with two different transcription systems: our trained Wav2Vec2 model, and the CNN from [3]. The figure shows how different agreement types (total agreement, total disagreement, consistent mismatch) correlate with note transcription accuracy. Notes in the total agreement category are transcribed most accurately, while total disagreement cases are often left unmatched. In N20EMv2, most consistent mismatch cases result in onset-only matches.

4.1 Analyzing the Onset-Pitch gap

It is not trivial why the gap between *COn* and *COnP* is notably larger in N20EMv2, and it could stem from the performance of the transcription systems, or reflect a change in data domain i.e. MIR500 and HSD are mostly composed from professional singers as opposed to N20EMv2. We hypothesize that amateurs can potentially produce unclear pitches that might lead to a lower performance of our models, when such models are trained only with data from professional musicians. Inspired by [30], we developed an automatic approach to assess such “pitch clarity” by leveraging f_0 extractors: we use three different pitch extractors (CREPE [31], PESTO [6], and PYIN [6]) to estimate framewise f_0 from vocals isolated with Demucs [32]. For each note, defined by its reference start and end times, we

quantize the extracted f_0 values to the nearest semitone and take the mode of the note’s time boundaries as the “automatic” pitch annotation. Therefore obtaining 3 automatically extracted pitches in addition to the reference one given for each dataset. We define 3 cases that we consider are most relevant based on the agreement between the different pitches obtained:

- **Total Agreement:** All automatic annotators agree with the reference.
- **Total Disagreement:** All automatic annotators disagree with each other and the reference.
- **Consistent Mismatch:** Total agreement within automatic annotators, but they disagree with the reference.

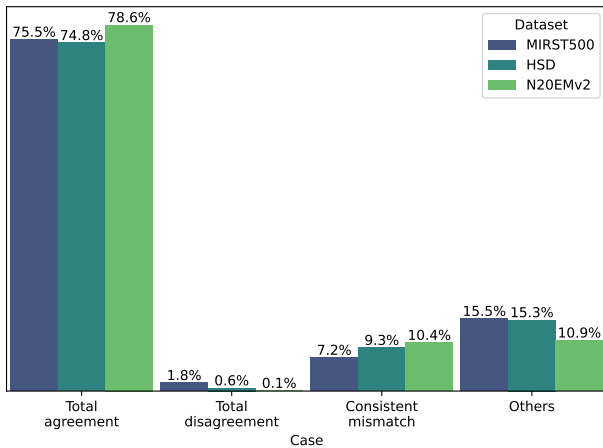


Figure 3: Agreement distribution across datasets: total agreement, disagreement, and consistent mismatch.

Figure 3 displays the distribution of these cases across the datasets. We observe that total agreement is the most common scenario. Total disagreement is rare across the three datasets. Cases with a consistent mismatch (although far from the majority) conform a sensible amount of cases.

Figure 2 further explores how these agreement correlate with transcription performance—that is, whether the reference notes are successfully matched in the transcription or left unmatched. We compare results from two transcription systems: Wav2Vec2 and the CTC+CE model from [3]. For a fair comparison between the two system, we use checkpoints trained solely on MIRST500, as the original model in [3] was also trained exclusively on that dataset.

Across all datasets, notes in the total agreement category are most frequently transcribed with correct pitch and offset, while total disagreement cases are typically left unmatched. In the case of consistent mismatches, we observe a notable difference between N20EMv2 and the other datasets. For MIRST500, the systems successfully match most notes, including the correct offset, a trend that also holds for HSD, though to a lesser extent. However, for N20EMv2, the majority of matched notes are only aligned at the onset level, failing to capture the correct pitch.

This discrepancy is partly due to the systems’ inability to detect notes accurately in these scenarios, which also occurs for MIRST500 and HSD. However, considering the larger *CON-CONP* gap for N20EMv2 in both models, and the consistent pattern of errors in the “consistent mismatch” cases, it suggests that these notes, where only the onset is detected correctly but the pitch is missed, requires further investigation. We do not claim that these pitch annotations are necessarily incorrect, but rather that certain scenarios, such as when the singer is off-pitch yet the annotator correctly identifies the intended pitch based on the musical context provided by the accompaniment, may be worth exploring from a perceptual perspective. This type of analysis is beyond the scope of this paper.

5. CONCLUSIONS

In this paper, we present a novel audio-to-score alignment method that synchronizes events between an audio signal and its corresponding score. Starting with a reference alignment, our algorithm identifies matching events in both domains and uses these as anchors to generate a refined warping path. Rather than replacing or competing with existing alignment techniques [13, 22], our approach leverages any available method to produce a reference alignment and then refines it around precise event positions. We also introduce heuristics that assess the reliability of each aligned section based on the number of matched events, a feature particularly useful for scores and audio obtained through automated means, such as web scraping. To demonstrate our method, we re-aligned the DALI dataset using note onset annotations and trained a singing voice transcription model on the resulting data. Models trained on this dataset achieved state-of-the-art results, with improved *CON* metrics and consequently led to better *CONP* performance in most cases. Additionally, we propose a methodology to examine cases where the annotated pitch may be unclear—particularly in instances of “consistent mismatch” as detailed in Section 4.1. Our method has the potential to impact research beyond automatic music transcription, benefiting any task that requires precise audio-score alignment, such as beat tracking. Accordingly, future extensions could adapt our approach to align other types of events, including beats and downbeats. Both the alignment code and instructions to download SOLEA are available on github¹.

Acknowledgments

This research has resulted from a collaboration between the Universitat Pompeu Fabra and the Huawei Munich Research Center.

6. REFERENCES

- [1] Salamon, Justin, “Melody Extraction from Polyphonic Signals,” Ph.D. dissertation, Universitat Pompeu Fabra, 2013.
- [2] Emilio Molina and Ana M. Barbancho and Lorenzo J. Tardón and Isabel Barbancho, “Evaluation Framework for Automatic Singing Transcription,” in *Proceedings of the 15th ISMIR*, Taipei, Taiwan, 2014.
- [3] Wang, Jun-You and Jang, Jyh-Shing Roger, “Training a Singing Transcription Model Using Connectionist Temporal Classification Loss and Cross-Entropy Loss,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [4] Qiu, Yao and Zhang, Jinchao and Shan, Yong and Zhou, Jie, “Enhancing Note-Level Singing Transcription Model with Unlabeled and Weakly Labeled Data,” in *ICASSP*, 2024.

¹ <https://github.com/migperfer/solea>

- [5] Gfeller, Beat and Frank, Christian and Roblek, Dominik and Sharifi, Matt and Tagliasacchi, Marco and Velimirović, Mihajlo, “SPICE: Self-Supervised Pitch Estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.
- [6] Riou, Alain and Lattner, Stefan and Hadjeres, Gaëtan and Peeters, Geoffroy, “PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective,” in *Proceedings of the 24th ISMIR*, 2023.
- [7] Wang, Jun-You and Jang, Jyh-Shing Roger, “On the Preparation and Validation of a Large-Scale Dataset of Singing Transcription,” in *ICASSP*, 2021.
- [8] Gu, Xiangming and Ou, Longshen and Zeng, Wei and Zhang, Jianan and Wong, Nicholas and Wang, Ye, “Automatic Lyric Transcription and Automatic Music Transcription from Multimodal Singing,” *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024.
- [9] Hawthorne, Curtis and Stasyuk, Andriy and Roberts, Adam and Simon, Ian and Cheng-Zhi and Huang, Anna and Dieleman, Sander and Erich, Elsen and Engel, Jesse and Eck, Douglas, “Enabling Factorized Piano Music Modeling and Generation with the MAE-STRO Dataset,” in *Proceedings of the 7th International Conference on Learning Representations*, 2019.
- [10] Manilow, Ethan and Wichern, Gordon and Seetharaman, Prem and Le Roux, Jonathan, “Cutting Music Source Separation Some Slack: A Dataset to Study the Impact of Training Data Quality and Quantity,” in *WASPAA*. IEEE, 2019.
- [11] Bittner, Rachel M. and Bosch, Juan José and Rubinstein, David and Meseguer-Brocal, Gabriel and Ewert, Sebastian, “A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation,” in *ICASSP*, Singapore, 2022.
- [12] Alia Morsi and Xavier Serra, “Bottlenecks and Solutions for Audio to Score Alignment Research,” *Proceedings of the 23rd ISMIR*, 2022.
- [13] Nazif Can Tamer and Yigitcan Özer and Meinard Müller and Xavier Serra, “High-Resolution Violin Transcription using Weak Labels,” in *Proceedings of the 24th ISMIR*, Milan, Italy, 2023.
- [14] Maman, Ben and Bermanno, Amit H, “Unaligned Supervision for Automatic Music Transcription in The Wild,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. PMLR, 17–23 Jul 2022.
- [15] Meseguer-Brocal, Gabriel and Cohen-Hadria, Alice and Peeters, Geoffroy, “Creating DALI, a Large Dataset of Synchronized Audio, Lyrics, and Notes,” *Transactions of the International Society for Music Information Retrieval*, Jun 2020.
- [16] Kum, Sangeun and Lee, Jongpil and Kim, Keunhyoung Luke and Kim, Taehyoung and Nam, Juhan, “Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music,” in *ICASSP*, 2022.
- [17] Eric Arazo and Diego Ortego and Paul Albert and Noel E O’Connor and Kevin McGuinness, “Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning,” in *IJCNN*. IEEE, 2020.
- [18] Graves, Alex and Fernández, Santiago and Gomez, Faustino and Schmidhuber, Jürgen, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 2006.
- [19] Yamamoto, Ryuichi and Yoneyama, Reo and Toda, Tomoki, “NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit,” *arXiv preprint arXiv:2210.15987*, 2022.
- [20] Nazif Can, Tamer and Ramoneda, Pedro and Serra, Xavier, “Violin Etudes: A Comprehensive Dataset for f0 Estimation and Performance Analysis,” *Proceedings of the 23rd ISMIR*, 2022.
- [21] Sakoe, H. and Chiba, S., “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, no. 1, 1978.
- [22] Bernhard Niedermayer and Gerhard Widmer, “A Multi-pass Algorithm for Accurate Audio-to-Score Alignment,” in *Proceedings of the 11th ISMIR, Utrecht, Netherlands, August 9-13, 2010*. ISMIR, 2010.
- [23] Yigitcan Özer and Matěj Ištvanek and Vlora Arifi-Müller and Meinard Müller, “Using Activation Functions for Improving Measure-Level Audio Synchronization,” in *Proceedings of ISMIR*, Bengaluru, India, 2022.
- [24] Peter, Silvan David and Cancino-Chacón, Carlos Eduardo and Foscarin, Francesco and McLeod, Andrew Philip and Henkel, Florian and Karystinaios, Emmanouil and Widmer, Gerhard, “Automatic Note-Level Score-to-Performance Alignments in the ASAP Dataset,” *Transactions of the International Society for Music Information Retrieval*, Jun 2023.
- [25] Baeovski, Alexei and Zhou, Henry and Mohamed, Abdelrahman and Auli, Michael, “wav2vec 2.0: a framework for self-supervised learning of speech representations,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

- [26] Colin Raffel and Brian McFee and Eric J. Humphrey and Justin Salamon and Oriol Nieto and Dawen Liang and Daniel P. W. Ellis, “mir_eval: A Transparent Implementation of Common MIR Metrics,” in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014.
- [27] Kingma, Diederick P and Ba, Jimmy, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2015.
- [28] X. Fu, X. Yuan, and J. Hu, “Hsd: A hierarchical singing annotation dataset,” in *2022 IEEE International Symposium on Multimedia (ISM)*, 2022, pp. 245–246.
- [29] L. Kim, S. Jeon, W. Heo, and J. Park, “Note-level singing melody transcription for time-aligned musical score generation,” *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1–15, 2025.
- [30] Rosenzweig, Sebastian and Scherbaum, Frank and Müller, Meinard, “Reliability Assessment of Singing Voice F0-Estimates Using Multiple Algorithms,” in *Proceedings of ICASSP*, 2021.
- [31] Kim, Jong Wook and Salamon, Justin and Li, Peter and Bello, Juan Pablo, “Crepe: A Convolutional Representation for Pitch Estimation,” in *ICASSP*, 2018.
- [32] Rouard, Simon and Massa, Francisco and Défossez, Alexandre, “Hybrid Transformers for Music Source Separation,” in *ICASSP*, 2023.