INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Insights into end-to-end audio-to-score transcription with real recordings: A case study with saxophone works

Juan C. Martínez-Sevilla, María Alfaro-Contreras, Jose J. Valero-Mas, Jorge Calvo-Zaragoza

University Institute for Computing Research (IUII), University of Alicante, Spain
{jcmartinez, malfaro, jjvalero, jcalvo}@dlsi.ua.es

Abstract

Neural end-to-end Audio-to-Score (A2S) transcription aims to retrieve a score that encodes the music content of an audio recording in a single step. Due to the recentness of this formulation, the existing works have exclusively addressed controlled scenarios with synthetic data that fail to provide conclusions applicable to real-world cases. In response to this gap in the literature, this work introduces a novel assortment of real saxophone recordings—together with their digital scores—and poses several experimental scenarios involving real and synthetic data. The obtained results confirm the adequacy of this A2S framework to deal with real data as well as proving the relevance of leveraging synthetic interpretations to improve the recognition rate in scenarios with real-data scarcity.

Index Terms: Audio-to-score transcription, Real-world data, Deep neural networks

1. Introduction

Attaining structured digital representations from music sources, a process typically known as (music) *transcription*, represents a long-standing problem in the Music Information Retrieval (MIR) area [1, 2]. Automatic Music Transcription (AMT) stands as the research field devoted to devising computational approaches capable of retrieving a high-level symbolic representation of the music content in an audio file [3].

Due to the inherent difficulty of the task, attempts to AMT usually resort to a *note-level* transcription, i.e., that in which the acoustic piece is encoded in terms of the onset, offset, pitch values, and musical instrument of the estimated notes [4]. In this regard, while such representation is deemed useful for a number of tasks, the particular goal of achieving a *score-level* codification—namely, Audio-to-Score (A2S) transcription—has been scarcely addressed in the literature as it entails the additional challenge of inferring non-audible information (e.g., clefs or meter indicators) [5].

Early A2S research works (e.g., [6, 7]) mainly relied on pipeline-based schemes to alleviate the complexity of the task. In those frameworks, each stage estimated individual aspects of the transcription (e.g., notes, key and time signatures, streams, bars, or voices) that were eventually integrated. Nevertheless, the issues related to error propagation between the stages, as well as the adequacy of the A2S methods to specific scenarios based on particular heuristics, have hindered their practical use. In contrast, recent advances in Deep Learning have enabled the development of neural *holistic* or *end-to-end* A2S methods that perform the transcription process in a single step to avoid the aforementioned issues [8, 9, 10].

The previous A2S works have exclusively addressed controlled experimental scenarios, being the sole use of synthetic data one of the most limiting points [11, 12]. Hence, there exists a need to analyze and provide insights when addressing real recordings [13, 14], since their richer expressiveness and varied quality conditions typically entail higher challenges than those cases in which synthetic data is exclusively contemplated.

To our best knowledge, this work constitutes the first to address the commented gap in the neural end-to-end A2S field by comparatively studying the use of recorded and synthetic audio pieces as well as their possible synergistic combination in the context of saxophone interpretations. Our contributions are: (i) the creation of a corpus of real-world saxophone recordings specifically devised for end-to-end A2S transcription; (ii) comprehensive experimentation to quantitatively assess the limitations of learning with real or synthetic data to transcribe recorded pieces in single-step end-to-end A2S pipelines; and (iii) the study of mechanisms to leverage synthetic data to improve the transcription performance when dealing with scenarios of real-data scarcity.

2. Saxophone recordings for A2S

Despite a large number of existing works in the AMT field, there is a lack of benchmark corpora for end-to-end A2S, especially when addressing real data. Hence, this work presents a collection of recorded saxophone performances together with their digital music scores.[1]

Having the saxophone as our main instrument relays in the expressive ability, dynamics (*piano-forte*), variety of effects during sound emission, and diverse musical styles. These characteristics, even though common to the wind family instruments, can be found regularly and accentuated in saxophone recordings. Note that, while the choice of the saxophone may be considered simplistic, as it represents a monophonic instrument, no previous end-to-end A2S work—neither monophonic nor polyphonic—has assessed the capabilities of such formulation with real recordings. In this regard, a thorough experimental study resorting to monophonic saxophone performances represents a valuable contribution to the field, without the question needs to the adequacy of the A2S method (which is still not clear for polyphony).

The presented assortment comprises 1026 recordings of real interpretations from two different types of saxophone—tenor and alto—along with their corresponding scores in Kern format [15]. This notation format is one of the most frequently used representations in computational music analysis due to its features, including a simple vocabulary, an easy-to-parse file structure—convenient for end-to-end A2S applications—and its compatibility with dedicated music software [16, 17] that can automatically convert it to other music encodings.

[1] Accessible at https://grfia.dlsi.ua.es/audio-to-score

10.21437/Interspeech.2023-88

The compositions in the set, which span for approximately 3 hours of total duration, comprise examples of melodies, exercises, scales, and a small number of music incipits extracted from [18]. Table 1 provides an additional description of the main features of this assortment.

Table 1: *Data description in terms of the average duration, mean number of symbol annotations per score, and pitch range for the contemplated saxophone types.*

Saxophone	Average duration (s)	Mean symbols per score	Pitch range (transposed to C)
Tenor	10.8 ± 2.8	27.0 ± 8.6	$A\flat_2$ - G_5
Alto			$D\flat_3$ - C_6

Regarding the collection process, all pieces were recorded in a home studio by musicians proficiently trained in the instrument. Different tempi, styles, and rhythm metrics were considered to increase the variability in the data, being a metronome used to avoid considerable tempo deviations. Note that, the transposing nature of the saxophone—i.e., music notation is not written at concert pitch—prevents its direct use in A2S since a given note token does not represent the same pitch for all variants of this family of instruments. Hence, all scores were processed to unify the reference pitches to the C note: scores in the alto saxophone (tuned in $E\flat$) were applied an ascendant minor third whereas the tenor saxophone annotations (tuned in $B\flat$) were transposed in a descendant major second.

Finally, the scores include additional annotations related to the interpretation—i.e., *altissimo*, *bend*, *breath*, *fall*, *false fingering*, *glissando*, *growl*, *trill*, and *vibrato*—as well as metadata detailing the piece type (scale, melody, exercise, etc.), the instrument model, the mouthpiece, the reed, and the performer profile. Such descriptions are expected to enable the use of this assortment in other MIR works beyond A2S transcription as, for instance, expressive performance recognition.

3. Methodology

3.1. Learning framework

Based on other works addressing A2S transcription [8], we consider a Convolutional Recurrent Neural Network (CRNN) scheme. This architecture comprises a block of *convolutional* layers, which learn the adequate set of features, followed by a group of *recurrent* stages, which model the temporal dependencies of the feature-learning block, and a final fully-connected network with a *softmax* activation that retrieves the posteriogram to be decoded. The Connectionist Temporal Classification (CTC) training procedure [19] is contemplated to achieve an end-to-end scheme as it allows training the network using unsegmented sequential data.

Formally, let $\mathcal{T} \subset \mathcal{X} \times \Sigma^*$ be a set of data where sample $x_i \in \mathcal{X}$ of acoustic recordings is related to symbol sequence $\mathbf{z}_i = \left(z_{i1}, z_{i2}, \ldots, z_{i|\mathbf{z}_i|}\right) \in \Sigma^*$, where Σ represents the symbol vocabulary used for encoding the music score. Note that the use of CTC to model the transcription task as an end-to-end *sequence labeling* framework requires the inclusion of an additional "*blank*" symbol in the Σ vocabulary, i.e., $\Sigma' = \Sigma \cup \{blank\}$.

At prediction, for a given datum $x_i \in \mathcal{X}$, the model outputs a posteriogram $p_i \in \mathbb{R}^{|\Sigma'| \times K}$, where K represents the number of frames given by the recurrent stage. Eventually, the predicted sequence $\hat{\mathbf{z}}_i$ is commonly obtained resorting to a *greedy* policy that retrieves the most probable symbol per frame in p_i and a subsequent mapping function that merges consecutive repeated symbols and removes the *blank* labels.

3.2. Train data scenarios

With the aim of quantitatively assessing the transcription performance of real music works when using real or synthetic train data, we pose two different scenarios that essentially differ in the nature of such train data: (i) a first one, denoted as $\mathcal{T}_r$ in the rest of the work, that uses the real saxophone recordings presented in Section 2; and (ii) a second case, denoted as $\mathcal{T}_s$, that contemplates sound synthesis procedures to generate an artificial version of the real recordings based on the annotated score.

In addition, based on recent work in the related Automatic Speech Recognition field [20], we study the possibility of leveraging synthetic data to improve A2S transcription when dealing with cases of real-data scarcity. For that, we propose the use of a combined train set $\mathcal{T}_c = \{x \in_R \mathcal{T}_r\}^{\theta \cdot |\mathcal{T}_r|} \cup \{x \in_R \mathcal{T}_s\}^{(1-\theta) \cdot |\mathcal{T}_s|}$, where $\theta \in [0, 1]$—parameter used for modeling different experimental scenarios—denotes the percentage of real recordings in the assortment.

Note that, while the $\mathcal{T}_r$ and $\mathcal{T}_s$ cases consider the standard strategy of using all available data to train the model, two different policies are considered for the $\mathcal{T}_c$ combined one: (i) a *Mixed* case in which the set $\mathcal{T}_c$ is directly used for training the model; and (ii) a *Fine-tuning* approach that restricts to the synthetic part of the assortment to train the model so that, after convergence, the available set of real data is used for fine-tuning.

4. Experimentation

4.1. Data preprocessing and representation

As commented, this work constitutes the first to gather insights related to the use of recorded and synthetic audio pieces in A2S transcription. Thus, while the quality and expressiveness of the real interpretations may be deemed as fixed attending to the skills of the musician and the room conditions of the recording session, synthetic elements severely differ based on the particular generation procedure contemplated.

Having said that, we consider two different audio synthesis methods for saxophone timbre to avoid possible biases and obtain more general conclusions: (i) the sample-based FluidSynth software with a regular-quality SoundFont bank; and (ii) the MIDI Differentiable Digital Signal Processing (MIDI-DDSP) neural synthesis model [21] due to its capability of producing remarkably realistic interpretations.

The input and output representations are based on those used in recent works [10, 22]. First, all pieces—recorded and synthesized—are resampled to a rate of 44 100Hz. A Short-Time Fourier Transform representation with log-spaced bins is then obtained considering $A_4 = 440$Hz as the reference pitch with 48 bins per octave, a 4096-sample window (92.88ms), and an 882-sample hop size (20ms).

We consider a subword-based encoding for the output codification, exemplified in Fig. 1, in which each note token is disentangled into its duration, pitch, and alteration components, remaining the rest of the score-level elements—clefs, key signatures, bar lines, measures, and tempo indications—unaltered. This results in a cardinality of $|\Sigma| = 75$ and $|\Sigma| = 73$ symbols for the tenor and alto saxophones, respectively.

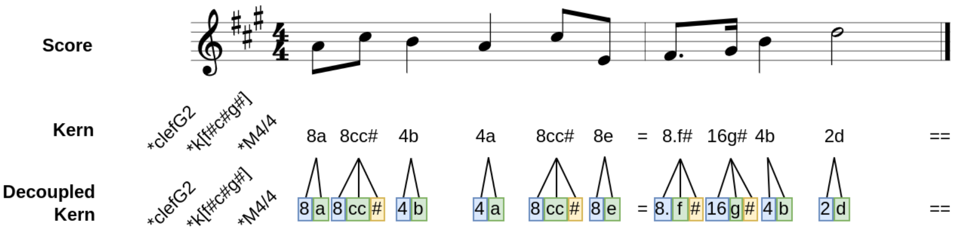
Score
Kern
*clefG2
*k[f#c#g#]
*M4/4
8a 8cc# 4b 4a 8cc# 8e = 8.f# 16g# 4b 2d ==
Decoupled Kern
*clefG2
*k[f#c#g#]
*M4/4
8 a 8 cc # 4 b 4 a 8 cc # 8 e = 8. f # 16 g # 4 b 2 d ==

Figure 1: *Graphical example of the score encoding. The Kern row depicts the symbolic equivalent of the Score one, whereas the Decoupled Kern row shows the format used by our transcription model. Note that the latter representation disentangles note tokens into their rhythm (in blue), pitch (in green), and accidental (in yellow) components, remaining the rest of the symbols unaltered.*

Finally, a 5-fold cross-validation scheme is contemplated in which, for each fold, the corpus is divided into three partitions at a file level with sizes 70%, 10%, and 20% for the train, validation, and test sets, respectively.

4.2. Evaluation metrics

We consider two complementary figures of merit typically used in the A2S field [9]: (i) the *Symbol Error Rate* (SER), a non-musical metric, that is computed as the average number of elementary editing operations (insertions, deletions, or substitutions) required to convert prediction $\hat{\mathbf{z}}_i$ into reference $\mathbf{z}_i$, normalized by the length of the latter; and (ii) the *MV2H* metric, specifically devised for A2S and introduced in [23], that summarizes, in a single value, the performance of the scheme in terms of its multi-pitch detection, voice separation, metrical alignment, note value detection, and harmonic analysis capabilities. For simplicity, while originally devised for polyphonic data, no components are adapted or discarded in the *MV2H* metric to the presented case of monophonic saxophone pieces.

4.3. Neural model configuration

The CRNN scheme is based on that used in recent works [1, 2]: two convolutional layers that apply 8 filters of size 2×10 and 5×8, respectively, considering a Leaky ReLU activation with a negative slope of $\alpha = 0.2$ and max-pooling stages of size and striding factors of 2×2 and 1×2. These feature maps are fed into two Bidirectional Long Short-Time Memory layers with 256 hidden units each, and a dropout value of $d = 50\%$ followed by a fully-connected network with $|\Sigma'|$ units.

The models were trained with a batch size of 16 elements considering the ADAM optimizer with a fixed learning rate of 10^{-3}. We iterate for 300 epochs, keeping the weights that minimize the SER metric in the validation partition. Finally, all experiments were run using the Python language (v. 3.8.13) with the PyTorch framework (v. 1.13.0) on a single NVIDIA A100 card with 40GB of video memory.

4.4. Results

Table 2 presents the average test results obtained with the proposed experimental scheme in terms of the SER and MV2H metrics. Note that all results correspond to the case of using real test data except for the *All synthetic* column in which both train and test constitute synthesized performances.

Attending to the figures obtained, similar recognition rates are observed in the scenarios in which both the train and the test data come from the same distribution. More precisely, the *Real* column—the case of recorded data—depicts average SER and MV2H values of 23% and 55%, respectively, being the corresponding scores in the *All synthetic* column—data entirely synthesized—around 20% and 70%, for the same metrics. This proves the validity of the contemplated end-to-end A2S formulation for real data, given that the results obtained do not remarkably degrade from those using synthetic data for both the train and test partitions (as in all previous works).

Focusing on the case of real test data (*Single train scenario* column), the use of the $\mathcal{T}_s$ synthetic data assortment to train the recognition model (columns *FluidSynth* and *MIDI-DDSP*, depending on the synthesis method) entails a steep performance decrease compared to that obtained when considering the $\mathcal{T}_r$ set of real data (column *Real*). More in detail, the SER metric loosely degrades from a value of 23% for both types of saxophone to error figures of 55% and 40%, depending on the synthesis approach. Similarly, the MV2H roughly decreases from a score of 55% to a value of 13%, independently of the nature of the synthetic data and saxophone type. Such a fact evinces the limitations of these sound synthesis methods to train recognition systems aimed at transcribing real recordings, most likely due to the lack of realism in certain aspects (e.g., musical expressiveness, recording artifacts, or timing deviations).

In addition, while both synthesis approaches prove to be insufficient for the posed real data transcription task, it is observed that the *MIDI-DDSP* case systematically yields more competitive results than the *FluidSynth* one. This suggests that more adequate and realistic methods can lead to performance improvements since the former method currently represents a state-of-the-art neural synthesis approach known by the realism of the results, at least when compared to other techniques (e.g., the *FluidSynth* one). In this regard, the rest of the experiments in the work only contemplate the *MIDI-DDSP* synthesis method.

Regarding the $\mathcal{T}_c$ case of combining real and synthetic data to target real recordings (column *Combined train scenario)*, the blending values of $\theta \in \{10\%, 20\%\}$ have been contemplated to simulate a shortage of real train data. For reference purposes, the *No synth* case provides the figures obtained when exclusively considering the subset of real data.

The results obtained in the commented scenario prove the validity of leveraging synthetic data to palliate real data scarcity in A2S transcription. The recognition rates in the $\theta = 10\%$ case depict certain improvements from the sole use of real recordings to the combined train data of up to 10% and 20% for the SER and MV2H metrics, respectively, depending on the particular saxophone and train policy. This point suggests a synergistic relationship between real and synthetic data since, even with a small number of recordings, the method outperforms the case of exclusively contemplating real pieces. Note that these observations match those in the $\theta = 20\%$ case with the only difference

Table 2: *Results in terms of the SER (%) and MV2H (%) metrics when considering real recordings ($\mathcal{T}_r$), synthetic corpora ($\mathcal{T}_s$), or their combination ($\mathcal{T}_c$) as train data, denoting parameter θ the overall percentage of real recordings in the latter scenario. All cases are evaluated on real recordings except for the All synthetic column in which both train and test constitute synthesized performances. Symbols ↑ and ↓ depict whether the metrics are positively or negatively valued, respectively. Note that the $\mathcal{T}_c$ combined train scenario only contemplates the MIDI-DDSP synthesis method.*

Metric	Saxophone type	Single train scenario ($\mathcal{T}$)			Combined train scenario ($\mathcal{T}_c$)			All synthetic (train & test)	
		FluidSynth	MIDI-DDSP	Real	Case	$\theta = 10\%$	$\theta = 20\%$	FluidSynth	MIDI-DDSP
↓ SER	Tenor	55.7	41.6	23.9	No synth	41.3	34.1	21.4	19.8
					Fine-tuning	39.5	32.4		
					Mixed	31.6	28.9		
	Alto	52.3	39.2	22.9	No synth	39.3	34.8	18.8	19.8
					Fine-tuning	34.1	32.8		
					Mixed	29.4	28.7		
↑ MV2H	Tenor	13.1	13.4	51.9	No synth	14.8	25.3	72.7	68.5
					Fine-tuning	24.6	33.3		
					Mixed	33.3	40.6		
	Alto	12.6	14.8	58.0	No synth	15.5	24.4	73.8	65.1
					Fine-tuning	30.5	33.3		
					Mixed	38.2	39.7		

of reporting narrower improvement margins due to the better reference recognition rate of the *No synth* case as a consequence of the larger amount of real recordings.

In relation to the particular train policies for this $\mathcal{T}_c$ combined scenario, the *Mixed* method systematically achieves better recognition rates than the *Fine-tuning* one, independently of the θ case, evaluation metric, or saxophone type. Such a difference in the performance suggests that the former strategy allows the model to better exploit the nuances provided by the subset of real data than the two-stage approach of the latter policy.

Finally, Fig. 2 shows the transcription results of a piece when training the model with either the entire set of real recordings ($\mathcal{T}_r$) or with the combined assortment ($\mathcal{T}_c$) of synthetic elements with $\theta = 10\%$ of real data. As expected, the set $\mathcal{T}_r$ provides a more accurate transcription (despite overestimating bar lines, it almost tracks all notes) than $\mathcal{T}_c$ (which misses several notes and produces an *enharmony* error).[2] However, as aforementioned, the competitive results obtained in the combined case reinforce the previous insights of a possible synergistic relationship between real and synthetic data in this particular training approach.

5. Conclusions

This work tackles a gap in the neural end-to-end Audio-to-Score (A2S) field. We studied and compared different scenarios of real and synthetic data to provide general insights about the transcription performance in those cases. For that, we contribute with a novel collection of real-world saxophone recordings with their corresponding score-level annotations—which represents the first assortment of this type specifically devised for A2S transcription—and pose different experimental scenarios comprising real, synthetic, and blended collections of data.

The obtained results validate the capabilities of the A2S framework to transcribe real data as the recognition rates match those observed in the works exclusively relying on synthetic corpora. Moreover, the use of synthetic train data proves to be useful for addressing scenarios with real-data scarcity since, with the adequate policy, it considerably improves the overall transcription performance.

Future work seeks to expand the presented assortment by contemplating other instruments and music textures, such as polyphony or ensemble music, to provide other challenging scenarios from which to extract additional insights. Furthermore, in light of the results obtained, another promising venue is to examine adequate data augmentation pipelines to increase the robustness of the transcription schemes and, hence, narrow the synthetic-to-real recognition gap.

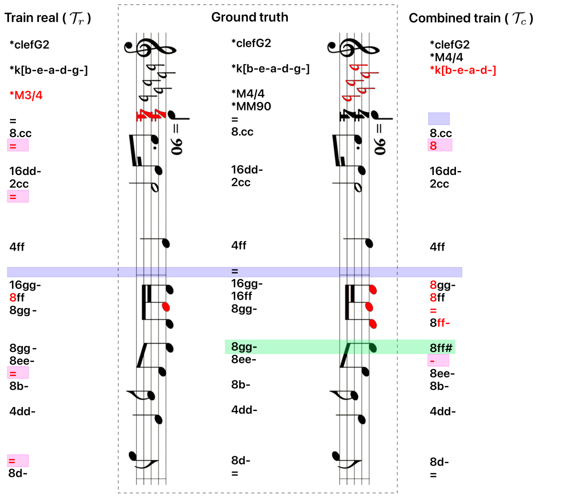

Figure 2: *Transcription examples when training the model with the $\mathcal{T}_r$ complete set of real recordings (left) or the $\mathcal{T}_c$ combined assortment of synthetic elements with $\theta = 10\%$ of real data (right). The ground-truth transcription in Kern (center) is provided, being also rendered as two independent scores for each recognition case in which the individual errors are highlighted.*

6. Acknowledgments

This paper is part of the project I+D+i PID2020-118447RA-I00 (MultiScore), funded by MCIN/AEI/10.13039/501100011033. The second author is supported by grant FPU19/04957 from the Spanish Ministerio de Universidades.

[2] Enharmony refers to different note names but same audible pitch (e.g., $B\flat$ and $A\#$).

7. References

[1] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Late multimodal fusion for image and audio music transcription," *Expert Systems with Applications*, vol. 216, p. 119491, 2023.

[2] C. de la Fuente, J. J. Valero-Mas, F. J. Castellanos, and J. Calvo-Zaragoza, "Multimodal image and audio music transcription," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 77–84, 2022.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[4] L. Liu and E. Benetos, "From audio to music notation," in *Handbook of Artificial Intelligence for Music*. Springer, 2021, pp. 693–714.

[5] K. Shibata, E. Nakamura, and K. Yoshii, "Non-local musical statistics as guides for audio-to-score piano transcription," *Information Sciences*, vol. 566, pp. 262–280, 2021.

[6] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing Human Piano Performances into Music Notation," in *17th Int. Society for Music Information Retrieval Conf.*, New York, USA, 2016, pp. 758–764.

[7] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, "Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 101–105.

[8] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "Data representations for audio-to-score monophonic music transcription," *Expert Systems with Applications*, vol. 162, p. 113769, 2020.

[9] L. Liu, V. Morfi, and E. Benetos, "Joint Multi-Pitch Detection and Score Transcription for Polyphonic Piano Music," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, pp. 281–285.

[10] V. Arroyo, J. J. Valero-Mas, J. Calvo-Zaragoza, and A. Pertusa, "Neural audio-to-score music transcription for unconstrained polyphony using compact output representations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Singapore, Singapore, 2022, pp. 4603–4607.

[11] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "An end-to-end framework for audio-to-score music transcription on monophonic excerpts," in *19th Int. Society for Music Information Retrieval Conf.*, Paris, France, 2018, pp. 34–41.

[12] R. Nishikimi, E. Nakamura, M. Goto, and K. Yoshii, "Audio-to-score singing transcription based on a CRNN-HSMM hybrid model," *APSIPA Transactions on Signal and Information Processing*, vol. 10, p. e7, 2021.

[13] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "A holistic approach to polyphonic music transcription with neural networks," in *20th Int. Society for Music Information Retrieval Conf.*, Delft, The Netherlands, 2019, pp. 731–737.

[14] Y. Hiramatsu, E. Nakamura, and K. Yoshii, "Joint estimation of note values and voices for audio-to-score piano transcription," in *22nd Int. Society for Music Information Retrieval Conf.*, 2021, pp. 278–284.

[15] C. S. Sapp, "Online Database of Scores in the Humdrum File Format," in *6th Int. Society for Music Information Retrieval Conf.*, London, UK, 2005, pp. 664–665.

[16] ——, "Verovio Humdrum Viewer," in *Music Encoding Conference*, Tours, France, 2017.

[17] L. Pugin, R. Zitellini, and P. Roland, "Verovio - A library for Engraving MEI Music Notation into SVG," in *15th Int. Society for Music Information Retrieval Conf.*, jan 2014.

[18] J. Calvo-Zaragoza and D. Rizo, "End-to-end neural optical music recognition of monophonic scores," *Applied Sciences*, vol. 8, no. 4, 2018.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *23rd Int. Conf. on Machine Learning*, 2006, pp. 369–376.

[20] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, pp. 5674–5678.

[21] Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.-Z. A. Huang, and J. Engel, "MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling," in *Int. Conf. on Learning Representations*, 2022.

[22] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *17th Int. Society for Music Information Retrieval Conf.*, New York City, United States, 2016, pp. 475–481.

[23] A. McLeod and M. Steedman, "Evaluating Automatic Polyphonic Music Transcription," in *19th Int. Society for Music Information Retrieval Conf.*, Paris, France, 2018, pp. 42–49.