NOTE-LEVEL TRANSCRIPTION OF CHORAL MUSIC

Huiran Yu
University of Rochester
hyu56@ur.rochester.edu

Zhiyao Duan
University of Rochester
zhiyao.duan@rochester.edu

ABSTRACT

Choral music is a musical activity with one of the largest participant bases, yet it has drawn little attention from automatic music transcription research. The main reasons we argue are due to the lack of data and technical difficulties arise from diverse acoustic conditions and unique properties of choral singing. To address these challenges, in this paper we propose a Transformer-based framework for note-level transcription of choral music. This framework bypasses the frame-level processing and directly produces a sequence of notes with associated timestamps. We also introduce YouChorale, a novel choral music dataset in a cappella setting curated from the Internet. YouChorale contains 452 real-world recordings in diverse acoustic configurations of choral music from over 100 composers as well as their MIDI scores. Trained on YouChorale, our proposed model achieves state-of-the-art performance in choral music transcription, marking a significant advancement in the field.

1. INTRODUCTION

Choral singing stands as one of the most widely engaged forms of musical expression, uniting voices in harmony across cultures and communities. Despite its profound presence in the musical landscape, choral singing has notably been overlooked in the field of Automatic Music Transcription (AMT), a domain predominantly oriented towards instrumental music [1–3], leaving choral singing with scant attention and few dedicated studies [4, 5]. This oversight not only highlights a gap in AMT research but also underscores the potential for significant advancements in the transcription of choral music, an area waiting for exploration and innovation.

The transcription of choral music introduces unique challenges compared with its instrumental counterparts. One of the main characteristics of choral singing is the soft onset of notes and smooth transitions between notes, resulting in indistinct boundaries and complicating the determination of note onsets. Additionally, the complex acoustic environment enriches choral music performances with reverberation, further complicates transcription efforts. These factors combined present a formidable challenge in accurately capturing the note occurrences in choral music recordings, necessitating novel approaches to AMT that can handle these specific challenges.

Recent methodologies in AMT fall primarily into two categories: Onsets and Frames [1], which estimates frame-level pitch activation informed by note onset predictions, and then combines such results to note estimates; and the use of models like MT3 [2], which conceptualize transcription as a token prediction task. However, both approaches exhibit limitations in addressing the soft onset characteristic of choral singing. Onset and frame detection methods heavily rely on the successful identification of note onsets, a task made difficult by the blurry beginning of vocal notes. Conversely, models like MT3 predict notes as a series of tokens, which can complicate the aggregation of information pertaining to individual notes, thereby obscuring the cohesive representation of choral music.

Another critical hurdle in advancing choral music transcription is the availability of comprehensive and high-quality datasets. Existing resources include the Dagstuhl ChoirSet [6], which offers less than one hour of high-quality recording of two pieces and a set of systematic exercises. The Erkomaishvili Dataset [7] provides around seven hours of recordings, but the sound quality is poor for model training. The Bach Chorale [1] and Barbershop Quartet [2] datasets provide tracked recordings, but the music genre is limited in these datasets. Also, they only involve a small group of singers and a fixed recording environment. This dearth of datasets impedes field progress and highlights the need for more robust and accessible resources for choral music transcription.

In response to these challenges, this paper proposes a novel note-level transcription architecture inspired by advancements in object detection and sound event detection. Instead of predicting the frame-level activation or separated MIDI-like events, this model directly decodes the pitch, onset, and duration from a hidden embedding of each note. To take care of the sequential relationships between the notes, we integrate the Transformer model as the backbone of our network, leveraging its proven efficacy in capturing long-term dependencies. Experiment results show that this model has largely improved the frame-level recall of the transcription output, indicating that the proposed model makes better use of the entire process of note articulation. The proposed model has also shown

© H. Yu and Z. Duan. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** H. Yu and Z. Duan, "Note-Level Transcription of Choral Music", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

[1] https://www.pgmusic.com/bachchorales.htm
[2] http://www.pgmusic.com/barbershopquartet.htm

robustness against the distortion caused by reverberation in the recordings. To address the critical gap in available resources, we have curated a comprehensive dataset for choral music transcription, comprising 496 real-world recordings across a diverse array of acoustic environments and featuring compositions from over 100 composers, accompanied by their corresponding MIDI scores. This dataset not only facilitates the development of our proposed model but also provides a valuable resource for future research in choral music transcription. Through this work, we aim to bridge the existing gap in AMT research, offering novel insights and methodologies that enhance our understanding and capabilities in transcribing choral music.

The structure of this paper is as follows: Section 2 covers the related works of this study; Section 3 describes the transcription architecture we proposed for the choral singing task; Section 4 introduces the YouChorale dataset, the experimental settings and the results; finally, Section 5 concludes the paper.

2. RELATED WORK

Automatic Music Transcription (AMT) has been a largely investigated task in Music Information Retrieval (MIR), and people have proposed various methods to address this problem. Onsets and Frames [1] represents the start of a group of methods that uses Convolutional Neural Networks (CNN) to extract the onset activation and frame activation in the spectrogram based on which a final note prediction output is aggregated through post-processing. Many other methods have inherited this idea, and several methods have been proposed for piano [8] and multi-instrument transcription [3, 9]. To fully use the activation detection and produce holistic transcription results, Yan et al. [10] proposed a neural semi-CRF-based method that predicts the best interval combinations of the frame-level estimations.

Another choice is to use sequence-to-sequence models that transcribe tokens describing different aspects of the notes, such as note-on and note-off events, velocity, and time stamps [11]. MT3 [2] expanded this method to multi-instrument transcription, and Simon et al. [12] further augmented the training data of such model by mixing monophonic recordings. There also exist methods that use generative diffusion models [13] to perform transcription. However, the performance of this method is still not comparable with other works.

For choral music transcription, Schramm et al. [4] proposed a spectrogram factorization method to transcribe a cappella performances. McLeod et al. [5] proposed using extended probabilistic latent component analysis and music language model to improve the performance further. There is also literature on score transcription of choral music [14], but they focus on producing the music score instead of the precise physical timing of each note in the recordings.

We can view automatic music transcription as a special form of sound event detection, which aims to identify the note entities in the audio recordings. The strong timing correlations between the notes drive us to detection methods with sequential modeling abilities. Carion et al. [15] proposed an end-to-end object detection architecture with Transformers, which uses the Transformer encoder and decoder to attend to the input image to detect sound events and their corresponding bounding boxes. Such an idea is also adapted in sound event detection, represented by works from Kong et al. [16].

3. METHOD

We demonstrate the architecture of the model in Figure 1. The input mel-spectrogram first goes through a prefiltering CNN network. After adding the positional encoding, two multi-head attention and feed-forward encoder layers further aggregate information in the spectrogram. The processed spectrogram then goes into the transformer decoder to auto-regressively generate an array of note embeddings. Finally, we employ three Multi-Layer Perceptron (MLP) modules as the feed-forward network to predict the MIDI pitch, onset time, and duration from the embedding of each note.

Inspired by the Transformer-based object detection methods and sound event detection methods, we regard the onset and offset as the "bounding box" of each note. Like pitch, they are the note's built-in attributes. Since Transformer models are well-known for their capability of learning long-term dependencies, we here let the encoder and decoder layers fully take care of the aggregation of information of each note to achieve end-to-end music transcription.

The model's input is a batch of segmented spectrograms with the shape of (B, L, M), where B is the batch size, L is the length of the segment, and M is the number of frequency bins. The model's output is three parallel arrays of pitch, onset time, and duration, with the shapes of (B, N, K), $(B, N, 1)$, $(B, N, 1)$, respectively. N is the length of the transcribed note sequence, and K is the number of possible pitch entries. To ensure the one-dimensional note sequence is unique for a polyphonic score, we serialize the notes first in chronological order from earliest to latest and second in pitch order from highest to lowest.

3.1 CNN Preprocessing

We use two layers of the 1D-Convolutional Neural Network (CNN) to preprocess the input mel-spectrogram. The network has a kernel size of 9 and is activated with the ReLU function, creating a receptive field of around 300 ms at each output frame. After the CNN, the shape of the output is (B, L, C), where C is the size of the hidden dimension. Then, we add the output with positional encoding and feed it into the encoder layers.

3.2 Encoder and Decoder Layers

We inherit the encoder and decoder design in the original Transformer [17], which includes multi-head attention

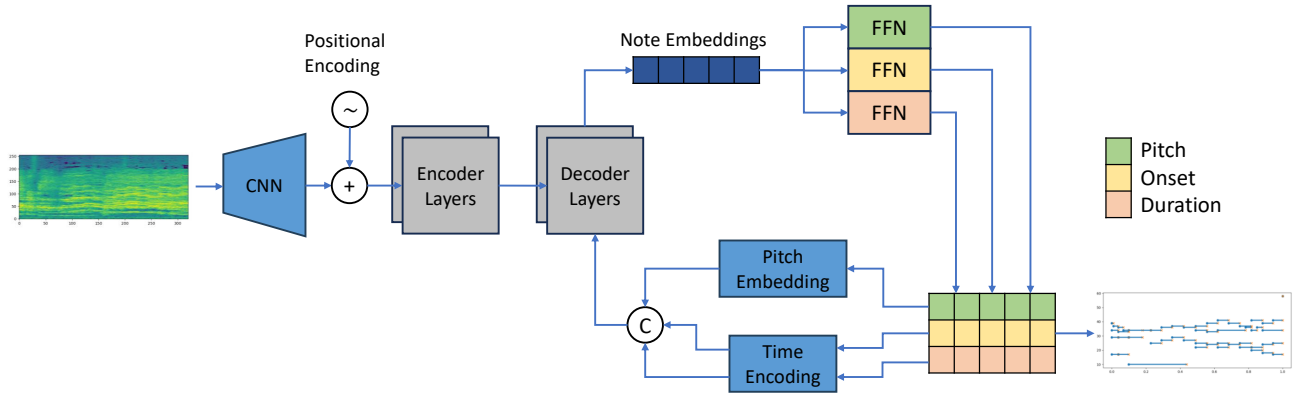

Figure 1. The overall architecture of the transcription model.

blocks and feed-forward layers. The encoder is conducting self-attention with the CNN-processed spectrogram; the decoder also attends to the output of the encoder after a self-attention layer. In our model, we use two layers of encoder layers and decoder layers.

During inference, the decoder performs auto-regressive decoding of the final note sequence. In model implementation, we normalize the time within one segment to $[0, 1]$ and calculate the onset time t_o and duration t_d accordingly to reduce the difficulties in training.

3.3 Positional Encoding

After the mel-spectrogram goes through the CNN filter banks, it will be added to a positional encoding to let the encoder layers learn the sequential relationship between the frames. For the positional encoding before the encoder layers, we adopt the original design in Transformer [17]. Given the frame from the processed spectrogram at pos out of the L possible positions and denote the dimensionality of the Transformer as d_{model}, we define the positional encoding PE at the $2i$ and $2i+1$ dimension as

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}), \quad (1)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}). \quad (2)$$

When we encode the continuous onset time and duration as the input of the decoder layers, we would like to align the decoder time encoding with the encoder's positional encoding. An onset time at t_o will have the time embedding TE identical to the PE of the corresponding frame position:

$$TE_{(t_o,i)} = PE_{(t_o \times L,i)}. \quad (3)$$

Similarly, the duration of the note t_d is encoded with the same equation, replacing t_o with t_d in Equation (3). In this way, we can align the time in the spectrogram and the onset prediction of the model, which will help the model better find the relationship between the frames in the spectrogram and the time in the final transcription result. After we get the pitch embeddings and time encodings of the previously generated notes, we concatenate them together and send them into the decoder layers.

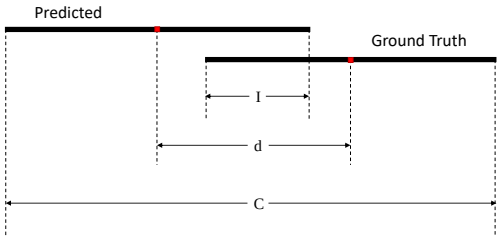

Figure 2. A demonstration of DIoU calculation.

3.4 Training Objectives

We optimize the loss of pitch estimation and timing estimation. For pitch estimation, we use the cross entropy loss $\mathcal{L}_p$. For time estimation, we first apply the L1 loss:

$$\mathcal{L}_{time} = \sum_{i=1}^{N} ||\hat{t}_o^{(i)} - t_o^{(i)}||_1 + \sum_{i=1}^{N} ||\hat{t}_d^{(i)} - t_d^{(i)}||_1, \quad (4)$$

where t_o is the ground-truth onset time, $\hat{t}_o$ is the predicted onset time; t_d is the ground-truth duration, $\hat{t}_d$ is the predicted duration.

We also adapt the DIoU (Distance-IoU, Intersection over Union) loss [18] from object detection to 1D scenario, as in Figure 2:

$$\mathcal{L}_{DIoU} = 1 - IoU + \frac{d^2}{C^2}. \quad (5)$$

Here, IoU is the ratio between the intersection and the union of the predicted time span and the ground-truth time span; d is the distance between the center of the prediction and the ground truth time span to add more penalties to the far away predictions; C is the length of the minimum bounding box that can cover both prediction and ground truth. Note that when there is an overlap between prediction and ground truth, union$= C$; when there is no overlapping between them, $IoU = 0$, we define union as the summation of the length of the two segments.

In the experiments, we trained two models with $\mathcal{L}_{time}$ and $\mathcal{L}_{DIoU}$ respectively and tested their performances.

	Singers in Each Part		Reverberation Time			Number of Parts						
	≤3	>3	long	medium	short	2∼3	4	5	6	7	8	≥9
Train	89	303	57	283	52	17	218	54	46	5	39	13
Validation	9	21	3	25	2	0	11	6	1	1	9	1
Test	10	20	5	23	2	0	11	9	4	0	5	1
Total	108	344	65	331	56	17	240	69	51	6	53	15

Table 1. Statistics of the YouChorale dataset.

4. EXPERIMENT

In this section, we describe experiments that evaluate our models against baselines.

4.1 YouChorale Dataset

In an effort to address the scarcity of resources for choral music transcription, we curated a dataset, YouChorale, from YouTube and a variety of MIDI archive sources [3] [4] [5], focusing exclusively on a cappella choral singing. With a total length of 22 hours 25 minutes, the YouChorale dataset contains 452 recordings of 261 compositions from 118 composers, representing a wide range of historical periods, styles, and complexities inherent to choral music. We have made the dataset publicly available at `https://github.com/ella-granger/YouChorale`, enriching a comprehensive resource of choral music for further exploration and development in the field of Automatic Music Transcription (AMT).

We split the dataset into train, validation and test set by the ratio of 392:30:30. To ensure the intonation of the evaluation recordings, we only selected performances by well-known choirs into the validation and the test sets. The detailed statistics of the dataset are shown in Table 1. The metric "singers in each part" indicates whether the performance is from a small a cappella group (less than or equals to three singers per part) or a larger ensemble (more than three singers per part). "Reverberation time" is an indication of the acoustic environment and how the signal is blurred or distorted. "Number of parts" indicates the complexity of the piece. Most of the pieces contain four to six parts, for example SATB or SSATTB, but there are also extreme cases where over nine parts appear in one composition.

We are also providing an aligned version of MIDI file along with the recordings. The alignment is achieved through the following steps: First, we adjust the key signature of the MIDI files to match the recording. Next, we render the waveform of the MIDI notation and align the Constant-Q Transform [19] feature of the synthesized audio and the performance recording by the soft-DTW algorithm [20]. Finally, we smooth the alignment curve to remove abrupt tempo changes in the aligned MIDI.

[3] www.learnchoralmusic.co.uk
[4] gasilvis.net
[5] http://www.maennerchor-sg.ch/midi/

4.2 BachChorale Dataset

The accurately labeled BachChorale dataset also serves as a benchmark for evaluating the transcription performance of our model and the baselines. This two-volume dataset contains 53 four-part choral compositions by J.S. Bach, with a total length of two hours. Note that during the collection of YouChorale, all the Bach pieces we found are accompanied by organ or orchestra. Therefore, we excluded Bach pieces to keep the dataset in a cappella settings, which also means that using BachChorale as a test dataset does not introduce label leakage.

4.3 Training Settings

For the training data, we downsampled the audio to 16 kHz, and extracted the mel-spectrogram with $N_{FFT} = 2048$, hop length = 256, and number of frequency bins $M = 256$. The length of each segment $L = 320$, which corresponds to 5.12 seconds of audio. The hidden dimension of the model $C = 256$. During training, we set the batch size = 8, and use an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The warmup step is set to 12000. We use teacher forcing during the training phase, which provides the ground truth notes as the context and lets the model predict only the next note. We have released our code at `https://github.com/ella-granger/NoteTranscription`.

4.4 Results

We choose Schramm et al. [4], Onsets and Frames [1] and MT3 [2] as our baselines. For Schramm et al. [4], we list their reported frame-level multi-pitch estimation result which was also evaluated on the BachChorale dataset. For Onsets and Frames, we train a new model with the YouChorale training set from scratch; for MT3, we use the provided multi-instrument checkpoint. For the Onsets and Frames, MT3, and the proposed model, we evaluate the transcription result after they produce the final MIDI notes output.

We evaluated the frame-level activation detection and the note onset detection with a tolerance of 50 ms and 100 ms, respectively. The results on the BachChorale dataset are shown in Table 2, and the results on the YouChorale test set are shown in Table 3. We can see that compared with Onsets and Frames or MT3, our proposed model has a more balanced performance on precision and recall at the frame level, and produces the highest f1 score among the

Model	Frame			Onset (50ms)			Onset (100ms)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Schramm et al. [4]	0.713	**0.709**	**0.710**	-	-	-	-	-	-
Onsets and Frames [1]	**0.832**	0.440	0.571	**0.411**	0.130	**0.196**	**0.730**	0.231	**0.348**
MT3 [2]	0.645	0.411	0.502	0.117	**0.249**	0.157	0.201	**0.426**	0.269
Proposed-$\mathcal{L}_{time}$	0.663	0.616	0.639	0.162	0.225	0.185	0.263	0.368	0.301
Proposed-$\mathcal{L}_{DIoU}$	0.611	0.639	0.624	0.189	0.182	0.183	0.284	0.274	0.275

Table 2. Model performances on BachChorale dataset.

Model	Frame			Onset (50ms)			Onset (100ms)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Onsets and Frames [1]	**0.806**	0.326	0.428	**0.450**	0.178	**0.242**	**0.688**	0.248	**0.344**
MT3 [2]	0.590	0.243	0.344	0.117	0.148	0.127	0.200	0.255	0.217
Proposed-$\mathcal{L}_{time}$	0.670	0.596	0.631	0.181	**0.221**	0.192	0.284	**0.339**	0.299
Proposed-$\mathcal{L}_{DIoU}$	0.630	**0.658**	**0.644**	0.210	0.209	0.203	0.309	0.304	0.297

Table 3. Model performances on YouChorale test set.

Model	Frame				Onset (50ms)			
	Precision	Recall	F1	Δ F1	Precision	Recall	F1	Δ F1
Onsets and Frames [1]	**0.604**	0.313	0.406	-0.165 (-28.9%)	**0.126**	0.094	0.107	-0.089 (-45.4%)
MT3 [2]	0.553	0.398	0.463	**-0.039 (-7.8%)**	0.022	0.040	0.028	-0.129 (-82.2%)
Proposed-$\mathcal{L}_{time}$	0.518	**0.518**	**0.518**	-0.121 (-18.9%)	0.089	**0.180**	**0.114**	**-0.069 (-37.7%)**

Table 4. Model performance under reverb distortion on BachChorale dataset.

deep learning methods. Although the Onsets and Frames model still reaches a higher precision value on the onset time of the note, the significantly higher recall of our model at the frame level indicates that it places greater emphasis on the entire process of note articulation, not just the onset and offset of the notes, which achieves our goal with holistic note transcription. The deep-learning methods still have some room for improvement towards Schramm et al. [4] on the BachChorale dataset, however, since the dataset only have one singer for each part, Schramm et al. might have some advantage as it was trained on solo singing.

We would also like to compare the performance of the two loss function $\mathcal{L}_{time}$ and $\mathcal{L}_{DIoU}$. From the results we can see that the $\mathcal{L}_{time}$ trained model tends to have high precision and low recall at frame level and low precision and high recall on note onsets, while the $\mathcal{L}_{DIoU}$ trained model has the opposite behavior. It indicates that the $\mathcal{L}_{time}$ trained model usually extracts shorter fragments of the notes and the $\mathcal{L}_{DIoU}$ trained model longer full notes. This is due to the property of the two loss functions: The L_1 based time loss function focuses more on the absolute distance between the boundary of the predicted notes and the ground-truth notes, while the L_{DIoU} based loss function puts more emphasis on the overall intersection of the prediction and ground truth, and will have the boundaries not as precise as the L_1 loss.

4.5 Performance Under Reverb Distortion

In real-world choral music performances, reverberation is an unignorable part of acoustic effects. For example, concert halls create reverb with a long reberveration time, which introduces distortions into the spectrogram. To evaluate the resilience of our model against common distortions encountered in live settings, we apply an artificial reverb [6] to our test set to simulate the complex acoustic environmental characteristic of real-world choral performances. The performance of each model is shown in Table 4. The findings indicate that our proposed model still holds a relatively high performance, and the proposed model together with Onset and Frames trained on the YouChorale dataset, retains some ability to predict onset timing while the MT3 model nearly failed to predict any reasonable onset of the notes. This resilience highlights the importance of incorporating diverse, real-world data in training AMT models, ensuring their applicability and effectiveness in practical, everyday transcription scenarios.

Through cautious dataset curation and strategic model design, the experiments have shown our proposed model's capabilities in the realm of choral music transcription. By directly addressing the nuanced challenges of this genre,

[6] https://ccrma.stanford.edu/ jos/pasp/Freeverb.html. The roomsize is set to one.

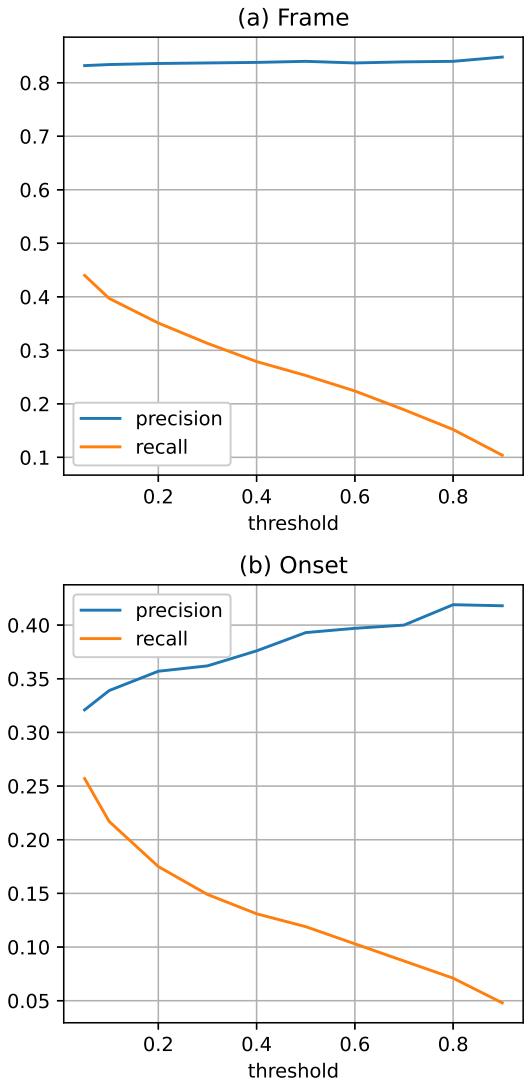

Figure 3. The precision/recall v.s. threshold curves of frame-level and note-level (onset) transcription from Onsets and Frames.

from soft note onsets to complex acoustic environments, we not only advance the state of AMT but also pave the way for future innovations in the transcription of polyphonic vocal music.

4.6 Limitations of Onsets and Frames Model

If we take a closer look at the result in Table 5, we may find that for Onsets and Frames model, there is a big gap between the frame-level precision and recall. After extracting the frame activation before post-processing and calculating its objectives, we get the result in Table 5. We can see that although post-processing improves prediction precision, it discards a large amount of true-positive frame activations.

Since the Onsets and Frames model will not transcribe any new note until it finds a new onset, the model's capability of correctly predicting the onset significantly affects the overall performance. Figure 3 shows the precision and recall curve of frame and onset prediction with respect to the onset decision threshold. The curves are unbalanced, and the reported result in Table 2 is at the threshold value of 0.05, which means we almost extract all the possible onsets as long as there is a trace amount of activation. All the evidence shows that the limitation of putting too much attention to onsets becomes especially pronounced in the context of choral music, where soft onsets and smooth transitions are prevalent. Instead, we should leverage the information contained in the frame activation and view each note as a whole, and let the model decide where to locate the notes, which is the design principle of our proposed method.

Model	Precision	Recall	F1
Onsets and Frames [1]	**0.851**	0.267	0.400
O&F (frame activation)	0.801	**0.632**	**0.704**

Table 5. Comparison between the final transcription result and frame-level activation of Onsets and Frames on Bach-Chorale dataset.

5. CONCLUSIONS

We proposed a novel transcription model architecture for choral music, which conducts holistic note transcription, addressing the soft onset and complex acoustic environment issues. We also introduced a newly curated a cappella dataset for the development of automatic music transcription. Tested on the BachChorale dataset, our model has shown competent performance on the choral music transcription task, particularly in its robustness against reverb. By addressing the noted limitations of existing models and contributing a valuable dataset to the research community, our work paves the way for future innovations in AMT, enhancing the accessibility and understanding of choral music through technology.

The next step of this work would be to distinguish and stream different parts in choral singing, and explore the potential of model architecture to general music transcription tasks.

6. ACKNOWLEDGEMENT

This work is supported in part by National Science Foundation (NSF) grants 1846184 and 2222129 and synergistic activities funded by NSF grant DGE-1922591.

7. REFERENCES

[1] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and Frames: Dual-Objective Piano Transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2018, pp. 50–57.

[2] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-Task Multitrack Music Tran-

scription," in *International Conference on Learning Representations*. ICLR, 2021.

[3] B. Maman and A. H. Bermano, "Unaligned Supervision for Automatic Music Transcription in The Wild," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, 17–23 Jul 2022, pp. 14 918–14 934.

[4] R. Schramm and E. Benetos, "Automatic Transcription of a Cappella recordings from Multiple Singers," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[5] A. McLeod, R. Schramm, M. Steedman, and E. Benetos, "Automatic transcription of polyphonic vocal music," *Applied Sciences*, vol. 7, no. 12, p. 1285, 2017.

[6] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, "Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing," *Transactions of the International Society for Music Information Retrieval*, Jul 2020.

[7] S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, and M. Müller, "Erkomaishvili Dataset: A Curated Corpus of Traditional Georgian Vocal Music for Computational Musicology," *Transactions of the International Society for Music Information Retrieval*, Apr 2020.

[8] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[9] K. W. Cheuk, D. Herremans, and L. Su, "Reconvat: A semi-supervised automatic music transcription framework for low-resource real-world data," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3918–3926.

[10] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the Frame-Level: Event-Based Piano Transcription With Neural Semi-CRFs," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 20 583–20 595.

[11] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-To-Sequence Piano Transcription With Transformers," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 246–253.

[12] I. Simon, J. Gardner, C. Hawthorne, E. Manilow, and J. Engel, "Scaling Polyphonic Transcription with Mixtures of Monophonic Transcriptions," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 44–51.

[13] K. W. Cheuk, R. Sawata, T. Uesaka, N. Murata, N. Takahashi, S. Takahashi, D. Herremans, and Y. Mitsufuji, "Diffroll: Diffusion-Based Generative Music Transcription with Unsupervised Pretraining Capability," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "A Holistic Approach to Polyphonic Music Transcription With Neural Networks," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, 2019, pp. 731–737.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[16] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound Event Detection of Weakly Labelled Data With CNN-Transformer and Automatic Threshold Optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 993–13 000, Apr. 2020.

[19] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[20] M. Cuturi and M. Blondel, "Soft-DTW: A Differentiable Loss Function for Time-Series," in *International conference on machine learning*. PMLR, 2017, pp. 894–903.