# Test Time Augmentation for Automatic Piano Transcription

Filip Danielsson[0009-0007-9188-3943]

KTH Royal Institute of Technology, fidani@kth.se

## 1 Introduction

Automatic piano transcription (APT) systems have mostly not been trained on augmented data, as it did not increase the performance on validation sets within the same data source [4, 7]. More recent works [2, 6] have experimented with training APT systems on augmented data using minor transformations and noise that avoid modifying the ground truth. These systems have shown an improvement with out-of-domain data suggesting that models trained on a single source such as the Maestro data set [5] can be biased. Using more complex augmentation operations that would modify the ground truth, such as pitch shifting whole semitones or time stretching, has not been explored. The purpose of this work is to explore the potential of these operations for improving APT systems. Firstly, the equivariance of a popular APT system [7] to a specific implementation of pitch shifting and time stretching will be tested. Secondly, test time augmentation (TTA) with pitch shifting and time stretching operations will be used in an attempt to improve the performance of the APT model.

Automatic piano transcription addresses the complex challenge of converting raw audio recordings into precise symbolic music representations. The state-of-the-art consists of deep learning approaches. In this work [7] will be used which uses a regression-based approach that directly predicts onset and offset times. [7] presents a high-resolution piano transcription system that achieves state-of-the-art performance by jointly modeling note timing, velocity, and pedal positions. The proposed architecture processes mel-spectrograms through a series of convolutional and RNN layers to extract both local and long-range musical features. The system was trained and evaluated on the Maestro dataset [5], comprising over 200 hours of virtuosic piano performances recorded under controlled conditions. The dataset provides perfectly aligned pairs of high-quality audio recordings and precise MIDI data captured from Yamaha Disklavier pianos during the International Piano-e-Competition.

## 2 Methodology

Test time augmentation (TTA) for piano transcription operates on the principle of generating multiple variations of an input audio sample to obtain more robust predictions. During inference, the original piano recording undergoes controlled modifications through pitch-shifting and time-stretching transformations. These transformations might include shifting the pitch up and down by small intervals such as one to three semitones, as well as stretching or compressing the temporal dimension by factors around 2-10%. Each transformed version is then independently processed through the transcription model.

The crucial step in TTA involves aligning and combining these varied predictions. For pitch-shifted versions, the resulting piano roll predictions must be inversely shifted to match the original pitch space. For instance, when the input audio is pitched up by one semitone, the corresponding piano roll prediction needs to be shifted down by one semitone to align with the original pitch space. Similarly, time-stretched predictions undergo inverse temporal scaling to match the original time base.

Once all predictions are properly aligned in both pitch and time dimensions, they can be aggregated through averaging across the piano roll representations. This averaging process exploits the tendency of true musical events to maintain consistency across different augmentations, while spurious detections typically exhibit more random behavior and thus get attenuated in the averaging process. For example, a genuine middle C note present in the original recording would likely be detected consistently across all augmented versions, whereas false positives would appear more sporadically and consequently receive lower averaged probabilities. The frame-wise precision, recall, and F1 score is then calculated on the piano rolls and compared to the baseline transcription to measure improvements.

While TTA provides improved transcription accuracy without requiring model retraining, it does introduce additional computational overhead during inference due to the multiple forward passes required for each augmented version. The trade-off between accuracy improvement and computational cost can be adjusted by varying the number and extent of augmentations applied.

Rubberband [1] is an audio processing library that manipulates pitch and time using a phase vocoder approach, which transforms audio into the frequency domain to separately adjust its pitch and duration. Its standout feature is intelligent transient detection that preserves the crispness of percussive sounds while avoiding the unnatural "phasiness" that often occurs in simpler pitch-shifting and time-stretching algorithms. Using rubberband to pitch shift $+$-3 semitones or time-stretch with a multiplier of 0.9 to 1.1 on piano music does not produce major noticeable artifacts.

## 3 Experiments

Each experiment is run on all 177 performances in the Maestro v3 test set consisting of over 20 hours of performances. Precision, Recall, and F1 scores are computed frame-wise with 10ms frames consistent with [7].

### 3.1 Reproducing baseline

The frame-wise performance of [7] was computed on the Maestro test set. We attempt to reporduce the result to quantify how comparable the outcomes might be to related works and what effects minor changes in experiment setup might have. The reproduced results show a slight improvement, this might be due to different methods being used for converting between midi and piano rolls.

| | **Precision** | **Recall** | **F1** |
|---|---|---|---|
| Kong [7] | 88.71 | 90.73 | 89.62 |
| Reproduced | 88.86 | 93.21 | 90.98 |

Table 1: Comparison between the original and reproduced results.

### 3.2 Equivariance of APT to pitch-shifing and time-stretching

If the performance of [7] is equivariant to the pitch-shifting and time-stretching operations provided by rubberband the same F1 score would be expected between the augmented transcription and its corresponding ground truth. As seen in Figure 1, a pitch shift of +-1 semitones or time stretching with a multiplier of 0.98/1.02 resulted in a decrease in the F1 score of at least 2%. Increasing the intensity of the augmentation further reduces the transcription performance. There are no directly audible artifacts in the audio produced by rubberband at these levels, suggesting a bias towards certain qualities used in the Maestro data set.

The Maestro dataset is recorded on Yamaha Disklaviers [5]. Previous works [2] tested [7] on a studio dataset that was a re-recording of the Maestro dataset using Yamaha Disklavier playback in a different studio environment. In that test, the note onset F1 score decreased by 1%. This suggests that even a change in the piano model or recording environment can affect transcription accuracy. If a similar drop in performance occurred due to the change in studio or instrument, it follows that further alterations, such as pitch-shifting or time-stretching using Rubberband, which modify both the spectral content and transients of the signal, would likely cause a more significant decrease in performance. These findings imply that the model is sensitive to subtle variations in the environment and that such manipulations of the signal, even without noticeable audible artifacts, can disrupt its transcription capabilities.

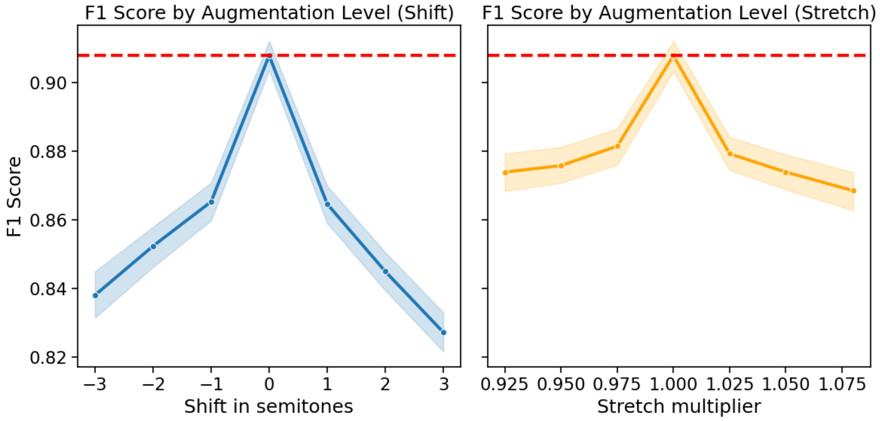

Fig. 1: Transcription performance vs. augmentation intensity

### 3.3 Test time augmentation for automatic piano transcription

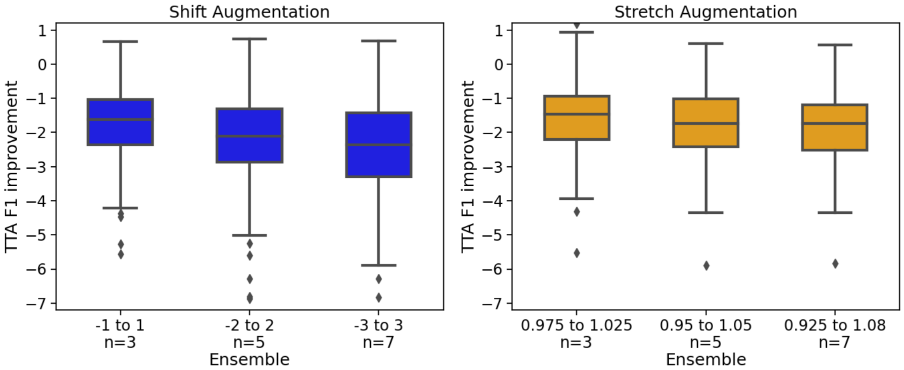

Fig. 2: Aggregated transcription performance vs. Ensemble size

Each file in the Maestro test set was augmented 6 times with a spread ±3 semitones for pitch shifting and 6 times with multipliers between 0.925 and 1.08 for time stretching. The original unaugmented audio together with the augmented files are then then transcribed using [7]. The resulting midi files are converted to piano rolls and grouped into ensembles of size 3, 5 and 7, centered around the unaugmented version. The piano rolls in these ensembles are then combined by taking the frame-wise mode (identical to rounded average for binary values). The F1 score of the combined piano roll is then compared to the F1 score of the original unaugmented file by itself. The spread of the change in F1 score for different ensamble sizes across the whole test set are shown in Figure 2.

Figure 2 shows that, in general, TTA reduces transcription performance across different ensemble sizes and augmentation techniques. Increasing the ensemble size by adding augmentations of higher intensity lowers the performance even further. The change in the F1 score is positive only in certain outlier cases, 29 out of 1062 ensembles improved the resulting score.

TTA is most effective when a model exhibits low bias and moderate variance. Low bias ensures that the model has adequately captured the underlying data distribution, allowing it to generalize well to the augmented inputs. Moderate variance is essential for the model to be responsive to meaningful variations introduced by augmentations while avoiding excessive sensitivity to noise or spurious changes. The significantly lowered performance across the whole test set for single augmentations in section 3.2 shows that [7] is biased against pitch-shifted and time-stretched audio samples. The decrease in performance after TTA with various ensembles suggests that this bias is to too large and the variance too small for a stable improvement with TTA over multiple samples.

### 3.4 Error Analysis

As seen in table 2, the majority of the errors created by TTA stem from a decrease in recall and an increase in false negatives. Increasing the ensemble size with shifting reduces precision, recall, and F1 close to linearly. On the other hand, an increase in the spread for stretching increases precision, dampening the decrease in F1 with a larger ensemble size.

| **Method** | **Ensemble** | ***Δ*P** | ***Δ*R** | ***Δ*F1** |
|---|---|---|---|---|
| Shift | -1 to 1 | -0.55 | -3.03 | -1.77 |
| | -2 to 2 | -0.78 | -3.59 | -2.17 |
| | -3 to 3 | -0.93 | -3.88 | -2.39 |
| Stretch | 0.975 to 1.025 | -0.83 | -2.43 | -1.61 |
| | 0.95 to 1.05 | -0.75 | -2.92 | -1.80 |
| | 0.925 to 1.08 | -0.65 | -3.22 | -1.90 |

Table 2: Change in precision, recall, and F1 score for Shift and Stretch methods across different ensembles.

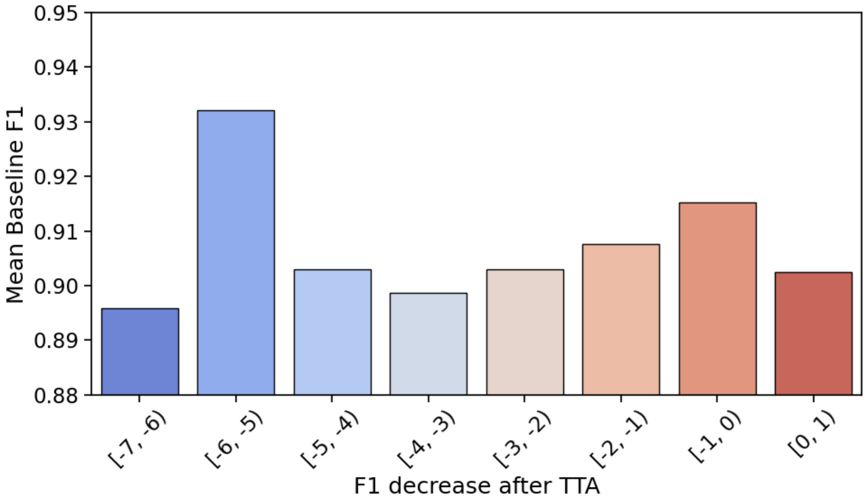

Fig. 3: Mean baseline F1 in post-TTA improvement bins.

The change in performance does not seem to be dependent on the baseline F1 score. In other words, TTA impacts the transcription performance of pieces the same regardless of their transcription difficulty. The correlation between the baseline F1 score and the change in aggregate F1 score is 0.0847. Figure 3 visualizes the mean baseline F1 score in bins of different post-TTA F1 changes. The positive correlation can be seen as the value of each bin increases slightly on average.

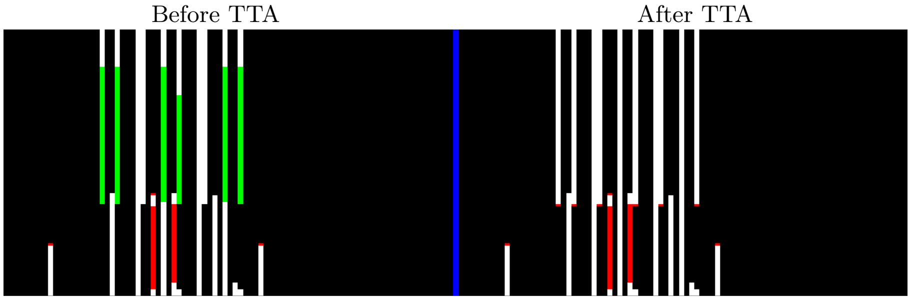

Fig. 4: Improvements, while rare, are often the result of extending multiple note durations or finding the correct pedal-release point.

**Color legend:** false negative false positive true positive true negative

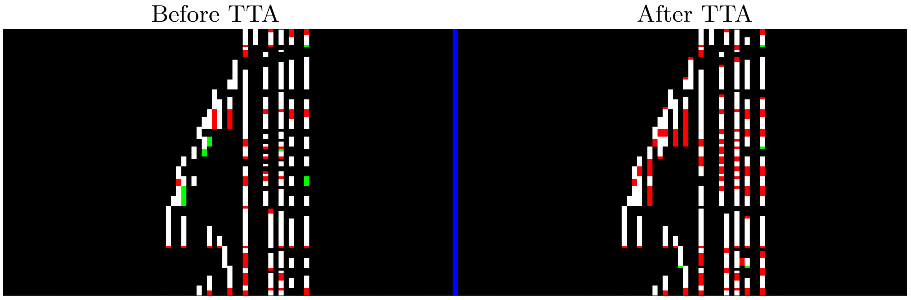

Fig. 5: Some false positives are created from short repeating notes being connected.

The errors or improvements produced by TTA are varied. Certain patterns from pieces and ensembles are shown for pieces with best (Figure 4) and worst (Figure 5) changes in F1 scores. The colors in the figures symbolize the following: green=false negative, red=false positive, white=true positive, black=true negative.

## 4 Future work

It is possible that using TTA with time stretching and pitch shifting from rubberband would yield much better results and could hence be used more broadly if the APT model was less biased. Testing to what degree bias is holding improvements back could be done by either using a more robust model such as [2] or lowering the baselines by testing on out-of-distribution data such as the MAPS dataset [3].

Other methods for shifting and stretching should be compared to understand to what degree the augmentation method impacts the resulting performance. For example, librosa's phase vocoder [8] implementation using the STFT, which processes overlapping frames in the frequency domain but can introduce artifacts in transient-rich piano recordings. In contrast, élastique [9] is an alternative to rubberband which combines time and frequency domain processing with specialized transient preservation, potentially offering higher quality transformations for piano audio, these differences could affect the model's ability to detect note events accurately.

## 5 Conclusion

This work explored the application of test-time augmentation (TTA) using pitch-shifting and time-stretching transformations to improve automatic piano transcription (APT). Through extensive evaluation on the Maestro v3 dataset, we found that TTA, as implemented in this work, failed to enhance overall transcription performance and, in most cases, led to performance degradation. This outcome can be attributed to the bias of the APT model against augmented inputs, resulting in limited adaptability to the variations introduced by TTA.

Despite the negative results, this work highlights the importance of addressing model bias when incorporating data augmentations. Future research could explore the potential of TTA with more robust transcription models or investigate its effectiveness on out-of-distribution datasets, such as MAPS, and compare alternative audio transformation techniques. While TTA did not yield the desired improvements in this case, the method shows promise in scenarios where models demonstrate better generalization to augmented inputs.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Breakfast Quay: Rubber Band Library: Technical Information (2025), https://breakfastquay.com/rubberband/technical.html, accessed: 2025-02-03
2. Edwards, D., Dixon, S., Benetos, E., Maezawa, A., Kusaka, Y.: A data-driven analysis of robust automatic piano transcription (2024), https://arxiv.org/abs/2402.01424
3. Emiya, V., Bertin, N., David, B., Badeau, R.: Maps - a piano database for multipitch estimation and automatic transcription of music (07 2010)
4. Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., Eck, D.: Onsets and frames: Dual-objective piano transcription (2018), https://arxiv.org/abs/1710.11153
5. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=r1lYRjC9F7
6. Kim, Y., Lerch, A.: Towards robust transcription: Exploring noise injection strategies for training data augmentation (2024), https://arxiv.org/abs/2410.14122
7. Kong, Q., Li, B., Song, X., Wan, Y., Wang, Y.: High-resolution piano transcription with pedals by regressing onsets and offsets times. CoRR **abs/2010.01815** (2020), https://arxiv.org/abs/2010.01815
8. Librosa Developers: Librosa: phase_vocoder Function. https://librosa.org/doc/main/generated/librosa.phase_vocoder.html (2025), accessed: 2025-02-03
9. zplane.development: élastique Pro V3.x SDK Documentation (2021), https://licensing.zplane.de/uploads/SDK/ELASTIQUE-PRO/V3/manual/elastique_pro_v3_sdk_documentation.pdf, accessed: 2025-03-25