TAPE: AN END-TO-END TIMBRE-AWARE PITCH ESTIMATOR

Nazif Can Tamer[♯*], *Yigitcan Özer*[♭*], *Meinard Müller*[♭], *Xavier Serra*[♯]

[♯] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[♭] International Audio Laboratories Erlangen, Germany

ABSTRACT

Pitch estimation of a target musical source within a multi-source polyphonic signal is of great interest for music performance analysis. One possible approach for extracting the pitch of a target source is to first perform source separation and then estimate the pitch of the separated track. However, as we will show, this typically leads to poor results. As an alternative to this approach, we introduce a timbre-aware pitch estimator (TAPE), which estimates the pitch of a target source in an end-to-end manner without the need for an explicit source separation step. Opposed to existing approaches that assume the predominance of a lead voice, our approach builds upon other cues that only rely on the timbral characteristics. Our results on real violin–piano duets show that, without any pre-processing step, TAPE trained on synthetic mixes outperforms the sequential procedure of source separation and pitch estimation under many settings, even if the target source is not predominant.

Index Terms— Pitch Estimation, Audio Source Separation, Music Performance Analysis, Weakly Supervised Learning

1. INTRODUCTION

Pitch estimation in real-life musical settings is a challenging problem. Most pitch estimation methods restrict the scenario to clean monophonic signals, and use time [1, 2, 3] or frequency [4, 5] domain techniques. Whereas the recent deep learning models, e.g., [6, 7], can handle realistic noisy settings, pitch estimation of a target source within a polyphonic music signal remains to be a challenge [8]. First attempts to solve this problem involve multi-pitch estimation without source assignments [9] and predominant melody extraction, i.e., pitch extraction for the dominant source in a polyphonic setting [9, 10, 11]. Since the dominant source in most musical traditions and genres is the singing voice, research efforts focused in particular on vocal melody extraction [8, 12, 13, 14].

In the literature, one can find two main strategies for vocal melody extraction: the direct approaches [12, 13, 14] and source-separation-based [15] methods. Direct approaches generally employ multi-resolution architectures that model human auditory perception to extract the melody from the polyphonic music [12, 13, 14, 16]. The second strategy for vocal melody extraction involves singing voice separation [8, 15, 17]. Earlier approaches divide the problem into explicit source separation and pitch estimation stages [15]. Recently, joint learning of source separation and pitch estimation has been proposed [8, 17]. In [8], Jansson et al. investigate different strategies for vocal pitch estimation from a mixture of different instruments and indicate that joint learning of source separation and pitch estimation leads to better results for both of these problems.

In this paper, we investigate violin pitch estimation from violin–piano duets. Duets, and especially duets of a monophonic instrument and piano, are the most common scenario in chamber music. Solo instruments, e.g., violin, are commonly played with piano in music exams and auditions. Thus, music performance analysis under the duet setting is of great interest to music education technologies. While predominant melody and vocal melody extraction assume that the main instrument always plays the melody, this assumption does not apply to duets since the piano part also frequently takes over the lead through dialogues and the contrapuntal texture of compositions.

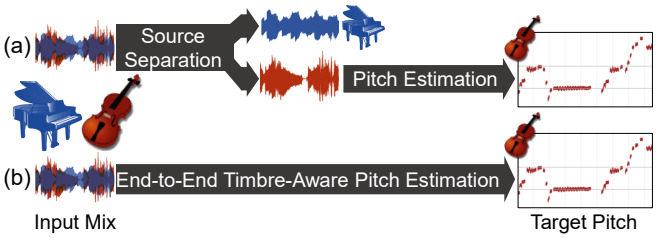

Fig. 1. Comparison of the two methods to estimate the violin pitch in a violin–piano duet. **(a)** Two-stage method with source separation as a preprocessing step, **(b)** End-to-end timbre-aware pitch estimation.

A common approach to estimate the pitch of an instrumental source in a duet is using source separation as a preprocessing step, as depicted in Figure 1a. Instead of applying an explicit source separation step, we propose an end-to-end timbre-aware pitch estimator (see Figure 1b). Our proposed model shares similarities with the multi-instrument transcription task [18, 19], which aims at inferring the instrument labels alongside transcription. Similar to the multi-instrument transcription, we aim to identify the instrument with its pitch contour. However, we only focus on a single instrument and strive for higher precision in the frequency domain.

As shown here with the violin–piano duets as an example, this scenario can be extended to other instruments' duets with piano and allow the analysis of musical performances in multi-instrument settings. Our main contributions in this paper are as follows:

- We introduce the Timbre-Aware Pitch Estimator (TAPE[1]), which works directly on the polyphonic mix waveform and can estimate the pitch of a target source (violin),
- We propose a novel synthetic audio mixing strategy that enables the training of timbre-aware pitch estimators using only single-instrument datasets in a curriculum,
- We benchmark different music source separation (MSS) methods for the task of downstream violin pitch estimation,
- As our main result, we show that the proposed TAPE model significantly outperforms SOTA pitch estimators, even if they receive source-separated audio as input.

*Equal contribution

[1] https://github.com/MTG/tape

2. TIMBRE-AWARE PITCH ESTIMATOR

The proposed two-stream architecture in Figure 2 works on raw audio sampled at 16 kHz and closely resembles the previous multi-resolution architectures in the vocal melody extraction literature [12, 13, 14]. We train this model on single-instrument datasets using a novel synthetic mixing paradigm that ensures timbre-awareness.

2.1. Two-Streams Model Architecture

As shown in Figure 2, TAPE comprises two convolutional neural networks with equal channel capacity and a transformer module that enables the information flow between them. Following the two-stream modeling of the human auditory system [16], we use the six convolutional layers of CREPE [6] as the two feature-extracting streams and connect them through a two-layer transformer [20]. Our use of convolutional and attention-based architectures is inspired by the findings from Dai et al. [21]. A detailed description of the four TAPE modules in Figure 2 is as follows:

Main Stream is the first six layers of CREPE [6] pitch estimator without any modification. This is the main stream responsible for pitch estimation, which receives 1024 waveform samples as input.

Attendant Stream is structurally identical to the *main stream*, but with a larger receptive field through dilations and strides in the convolution. Human accuracy in pitch tracking increases with sample duration [22], and *attendant stream* serves similar to a temporal smoothing step to eliminate the need for Viterbi post-processing.

In duets, it is possible that the other instrument dominates the audio for a time instant and occludes the target pitch. Thus, the large window size of the *attendant stream* helps in focusing on the target instrument in such conditions. By default, the *attendant stream* window size is 16384-samples, i.e., 1024 ms. We also report TAPE's performance on different *attendant stream* window sizes.

Transformer that we use here is a simple two-stage encoder-decoder module that enables the information flow from the *attendant stream* to the *main stream* through attention. We used two layers for both the encoder and decoder of the transformer with sinusoidal positional embeddings. We refer to Vaswani et al. [20] for details.

Fully-Connected (FC) is the final fully-connected layer to obtain the output pitch activations. The transformer decoder output with a shape of 512×4 is reshaped into 2048×1 and converted into 480 activations as shown in Figure 2. Instead of the final 360 activation bins previously used in deep learning-based pitch estimators [6], we use 480 bins that match the violin range and provide a higher frequency precision: Our model predicts in the pitch range E_3–E_8, with 12.5 cents between bin centers.

2.2. Timbre Awareness through Synthetic Audio Mixing

One main novelty of TAPE is its new training strategy based on synthetic audio mixing. To our knowledge, synthetic mixing strategies have not been adopted for pitch estimation or melody extraction.

While we create random mixes from single-source violin and piano datasets similar to [23], what differentiates our methodology is the different mixing parameters that we use to control and schedule the signal-to-noise ratio (SNR) during training. The main dataset we use for the training is the recently-introduced *Violin Etudes* [24], which is a large-scale violin performance dataset collected from teacher performances of the pedagogical violin repertoire. *Violin Etudes* comprises automatically extracted f_0 labels, and the resynthesized audio matching with the extracted f_0 values. To ensure the timbre awareness of the violin pitch tracker in violin-piano duets, we apply an artificial mixing strategy using the *Maestro v3* dataset [25], which is one of the most extensive open-source piano datasets. We use an $80-20\%$ train–validation split on both of the datasets and 16384-sample audio waveforms with a sampling rate of 16 kHz.

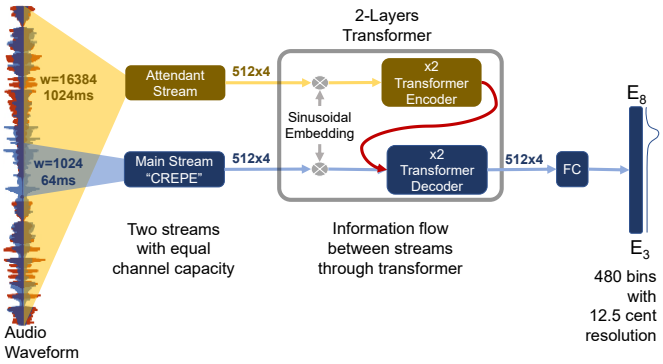

Fig. 2. TAPE architecture. Input mix audio waveform is analyzed with two convolutional streams and a transformer that provides the information flow between them. The output is the violin pitch activations encompassing the pitch range E_3–E_8 with 480 bins.

Training Curriculum starts with training on clean violin tracks until the Raw Pitch Accuracy (RPA50) reaches 90% on the clean validation data. Later, we start creating artificial mixes from random violin and piano patches. In order to control the relative loudness of the violin and piano sounds during training, we fixed the root mean square (RMS) amplitude of the violin signal and varied the RMS amplitude of the piano signal according to SNR in dB as parameter:

$$\mathrm{SNR\ (dB)} = 20 \log_{10} \frac{\mathrm{RMS}_{\mathrm{violin}}}{\mathrm{RMS}_{\mathrm{piano}}} \tag{1}$$

To enhance the robustness of our model against different recording conditions, we use a different SNR value for each sample in a batch when mixing the violin and piano patches. Concisely, using a batch size of 256, we generate 256 linearly-spaced SNR values between $\mathrm{SNR}_{\mathrm{min}}$ and $\mathrm{SNR}_{\mathrm{max}}$, and apply them to the corresponding samples in the batch. We set $\mathrm{SNR}_{\mathrm{max}}$ to 60 dB throughout the training, which corresponds to the simplest *monophonic* scenario. On the other hand, $\mathrm{SNR}_{\mathrm{min}}$ is gradually decreased from 60 dB to -30 dB, where the violin signal is barely audible. After this level, the training continues in the linearly-spaced SNR range from -30 dB to 60 dB.

Aside from the above-mentioned method, we also experimented with other curriculum strategies, such as constantly increasing the SNR or random SNR per batch. However, we did not report the results for the latter because they led to unstable training outcomes.

Other Training Details: We adopt the Binary Cross Entropy (BCE) loss function as in [6] and minimize the BCE loss between the prediction and the target vector that is smoothed by a Gaussian kernel with a standard deviation of 15 cents. We train using Adam optimizer with learning rate 10^{-5} and finish the training after one epoch on the 45 million pitch samples of the *Violin Etudes*.

Note that we use different datasets for training and testing. We train our network with artificial mixes, following [23] since random mixes lead to acceptable source separation results when the training datasets are big enough. Our test dataset comprises real violin–piano duets, which we will explain in detail in the next section.

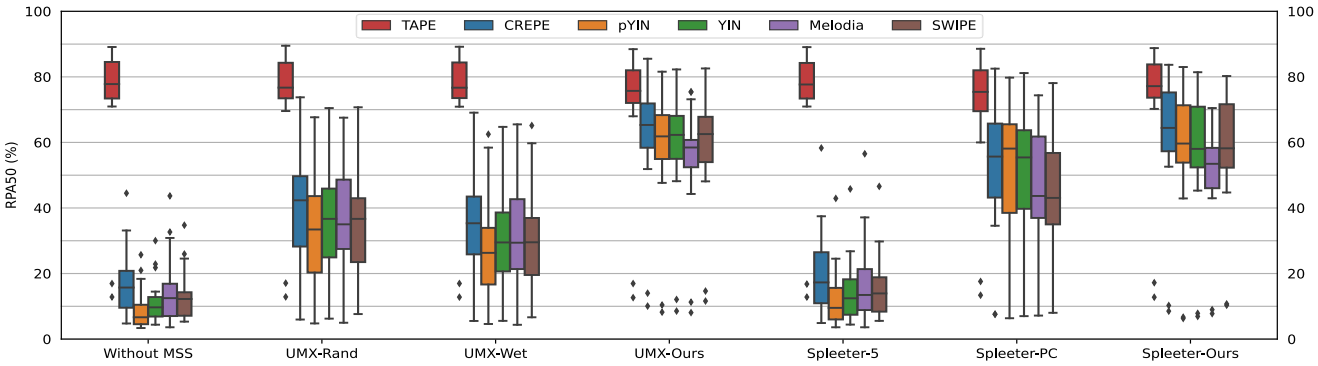

Fig. 3. Boxplots summarizing the Raw Pitch Accuracy (RPA50, %) under different input conditions on MusicNet violin–piano duets. Experiments without MSS compared with six source separators (UMX-Rand, UMX-Wet, UMX-Ours, Spleeter-5, Spleeter-PC, and Spleeter-Ours).

3. EXPERIMENTS

We test TAPE on *MusicNet* [26] and *MedleyDB* [27] datasets, which are disjoint to the training datasets, and compare its performance with monophonic pitch trackers preceded by MSS algorithms. We also investigate different MSS models and their impact on the subsequent pitch estimation task. For evaluation, we use the `mir_eval`[2] implementation of Raw Pitch Accuracy (RPA), computed with two thresholds: RPA50, the conventional threshold with 50-cent tolerance, and RPA5 fine-grained pitch accuracy metric, with 5-cent tolerance, for intonation analysis.

MusicNet [26] is a large-scale music transcription dataset with automatically-generated MIDI alignments. We provide our test results on all the violin–piano duets from this dataset, i.e., 22 tracks comprising Beethoven violin–piano sonatas of a duration of 180 minutes in total. To compute RPA50 from MIDI note numbers, we convert aligned violin MIDI events to frame-level pitch values with 2 ms between frames, and report the results on the monophonic passages played by the violin. RPA50 benchmarks reported for the MusicNet are bounded by two types of inherent errors in the labels: the alignment and the performer's intonation errors. The authors of *MusicNet* estimate the alignment errors to be around $8-10\%$.

MedleyDB [27] is one of the most commonly-used benchmark datasets in MIR with melody annotations. We report our experiments on a violin–piano subset previously identified as leakage-free by Chiu et al. [23]. Since we need violin pitch annotations, we created an even smaller subset from the segments where a violin and a piano are active, and the violin has a melody annotation. This small subset corresponds to only 4 minutes. However, it enables testing the robustness against different microphone placements thanks to the availability of both violin and piano stems. Furthermore, since the melody annotations are semi-automatically generated and corrected, the annotations are reliable for studying RPA5, which is fine-grained pitch accuracy and therefore is crucial in intonation analysis.

We use the following implementations for the baseline pitch estimators: CREPE [6] from its official repository[3], PitchMelodia [10] from Essentia[4], and pYIN [3], YIN [2] and SWIPE [5] from libf0[5]. All the pitch trackers, except for CREPE, allow setting min and max frequencies; we used frequency values that correspond to E_3 (min) E_8 (max) to match with the violin range for a fair evaluation.

[2] `https://craffel.github.io/mir_eval`
[3] `https://github.com/marl/crepe`
[4] `https://essentia.upf.edu`
[5] `https://github.com/groupmm/libf0`

3.1. Music Source Separation (MSS) Baselines

Monophonic pitch estimators, by design, require single source stems as the input. However, our application focuses on the pitch estimation of violin in violin–piano duets. To have the relevant baselines, we use four different pre-trained MSS models as a preprocessing step for pitch estimation. Furthermore, we train two well-known MSS models on the same datasets as TAPE for a fair comparison. To this end, we adopt spectral-based MSS models, which learn to approximate the magnitude spectrogram of a target source and reconstruct the separated audio signals through soft masking or multichannel Wiener filtering [28].

As a starting point, we run the violin pitch estimators using the violin–piano duets *without MSS*, i.e., without an explicit source separation step. As our first MSS baselines, we choose the pre-trained violin–piano MSS models by Chiu et al. [23] based on the BLSTM-based OpenUnMix [29]. We denote the pre-trained model trained with random mixes as *UMX-Rand* and the one trained using additional pink noise for data augmentation as *UMX-Wet*. For further details regarding the training strategies of the models, we refer to [23].

As a second MSS baseline, we consider the 5-stem Spleeter model [30] (*Spleeter-5*), which addresses the separation of piano, vocals, bass, drums, and other for popular music recordings. Combining the resulting non-piano magnitude spectrograms, we use the pre-trained model as a binary source separator, as in [23].

Third, we use the Spleeter-based pre-trained model by Özer and Müller [31] (*Spleeter-PC*), which focuses on splitting piano concerto recordings into piano as the lead instrument, and orchestra as the accompaniment. We regard the resulting separated orchestra as solo violin, which is feasible regarding the high inclusion of strings in the orchestral works used for the training of Spleeter-PC.

Fourth, we train UMX- and Spleeter-based models using the same datasets as TAPE, i.e., artificial random mixes from Violin Etudes [24] and Maestro v3 [25]. We denote these models as *UMX-Ours* and *Spleeter-Ours*.

Figure 3 illustrates the violin pitch estimation performances under different input conditions. First, we observe that using MSS as a preprocessing step enhances the performance of the monophonic pitch estimators as expected, whereas TAPE exhibits its best performance without an explicit MSS step. Among the MSS models, Spleeter-Ours yields the best MSS performance for the downstream task of violin pitch estimation. We also see that the deep-learning-based CREPE performs better than the traditional pitch estimators under most settings. Yet, among all the pitch estimators, TAPE performs the best under each setting and preserves its robustness despite the artifacts arising from non-optimal MSS results.

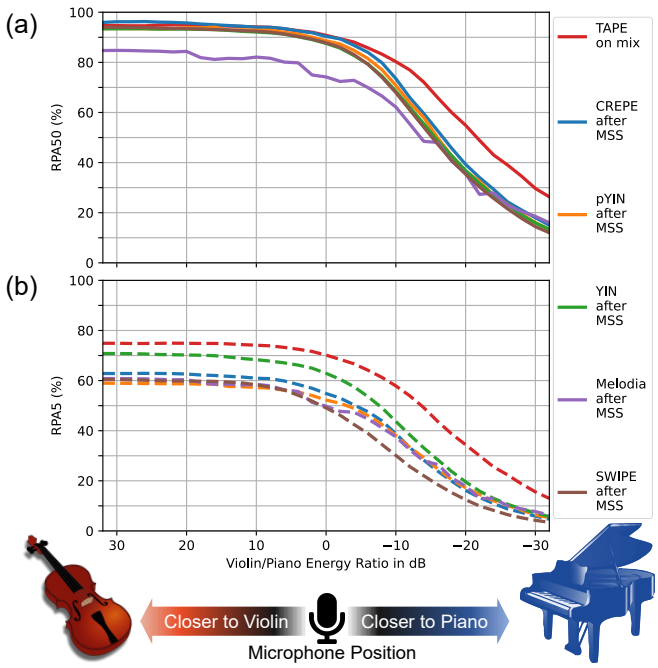

Fig. 4. Plots comparing **(a)** RPA50 and **(b)** RPA5 results on MedleyDB violin–piano duets. SNR values in the horizontal axis are obtained by mixing violin–piano stems with different gains to study microphone positioning. CREPE, pYIN, YIN, Melodia, and SWIPE are preceded by the best-performing MSS model Spleeter-Ours, whereas TAPE receives the violin–piano mixes directly.

3.2. Effect of Microphone Position

In real-life recordings, obtaining a perfect sound balance is a challenging task. Thus, any performance analyzer that works on audio mixtures should strive for invariance to different microphone placements. In this work, we study the robustness to diverse microphone positioning by mixing violin–piano stems from MedleyDB with varying gains. After calculating the RMS energy of violin and piano, we mix them at different ratios that correspond to the SNR values seen in Figure 4. The results using the conventional RPA50 metric indicate that, when the violin is predominant, the two-stage procedure of MSS and pitch estimation performs as well as TAPE on raw input mixture. However, when the piano is louder than the violin, TAPE significantly outperforms the two-stage methods. Furthermore, results using RPA5 demonstrate that the TAPE estimates are more precise across all mixing ratios and microphone positions.

3.3. Results

Table 1 provides the comparison of TAPE with the state-of-the-art pitch estimators in the literature. Here, we show the MedleyDB results on the perfect microphone placement scenario, i.e., equal violin–piano energy corresponding to $\mathrm{SNR} = 0$ (see Figure 4). For a fair evaluation, we use the monophonic pitch estimators preceded by the best-performing MSS model (Spleeter-Ours) to provide a separated violin signal. The results show that the TAPE on raw mix waveform outperforms the two-stage methods across all evaluation metrics. TAPE yields an RPA50 result of 75.9%, whereas the second best pitch estimator, CREPE after Spleeter-Ours, results in a significantly lower RPA50 of 64.9% for the subset from MusicNet. For MedleyDB, all the methods achieve higher pitch estimation results. For example, TAPE yields an RPA50 of 90.8% and an RPA5 of 70.2%, whereas the second best model CREPE after MSS leads to an RPA50 of 90.4% and an RPA5 of 54.8%. Note that, as shown by the RPA5 values, TAPE yields more precise pitch estimates than the other methods. Thus, it is more suitable for intonation analysis.

	MusicNet	MedleyDB	
	RPA50	RPA50	RPA5
TAPE on mix	**75.9**	**90.8**	**70.2**
CREPE after MSS	64.9	90.4	54.8
pYIN after MSS	62.1	88.6	52.2
YIN after MSS	60.7	87.6	62.9
Melodia after MSS	53.2	74.2	49.8
SWIPE after MSS	60.0	87.8	49.1

Table 1. Violin Raw Pitch Accuracy (RPA, %) results on violin–piano duets from MusicNet (180 min) and MedleyDB (4 min). CREPE, pYIN, YIN, Melodia, and SWIPE are preceded by the best-performing MSS model Spleeter-Ours whereas TAPE receives the violin–piano audio directly as the input.

Attendant Stream	MusicNet	MedleyDB	
window size (ms)	RPA50	RPA50	RPA5
64	73.8	87.7	67.8
128	74.1	88.1	68.2
256	75.4	89.6	69.0
512	75.6	90.8	69.9
1024	75.9	90.8	70.2
2048	**78.6**	**92.4**	**71.3**

Table 2. Effect of TAPE *attendant stream* window size on Raw Pitch Accuracy, in %, for violin pitch estimation in violin–piano duets.

In Table 2, we study the effect of TAPE's attendant stream window size on violin pitch estimation in duets. With the same model weights, enlarging the attendant stream window to 2048 ms through dilations at inference time yields an RPA50 of 78.6% on MusicNet, and an RPA50 of 92.4% and RPA5 of 71.3% on MedleyDB. Results indicate that the performance of the pitch detection can be substantially improved with larger attendant stream windows.

4. CONCLUSION

In this work, we explored pitch estimation of target sources in multi-instrument music, in particular, violin pitch estimation for violin–piano duets. We introduced a violin timbre-aware pitch estimator (TAPE) which was trained on single-instrument datasets using a new synthetic mixing strategy. We showed that our proposed TAPE outperformed the conventional pipeline of source separation and pitch estimation in real-life violin–piano duets. We believe that the timbre-aware pitch estimation is an important step towards end-to-end music performance analysis for accompanied instruments.

5. ACKNOWLEDGMENTS

This research is funded by the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI), and by the German Research Foundation (DFG MU 2686/10-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

6. REFERENCES

[1] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[2] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music.," *Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 659–663.

[4] Robert C Maher and James W Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[5] Arturo Camacho and John G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[6] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 161–165.

[7] Xiaoheng Sun, Xia Liang, Qiqi He, Bilei Zhu, and Zejun Ma, "Gio: A timbre-informed approach for pitch tracking in highly noisy environments," in *Proc. ICMR*, 2022, pp. 480–488.

[8] Andreas Jansson, Rachel M. Bittner, Sebastian Ewert, and Tillman Weyde, "Joint singing voice separation and F0 estimation with deep U-net architectures," in *Proc. EUSIPCO*, A Coruña, Spain, 2019.

[9] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P. Bello, "Deep salience representations for F0 tracking in polyphonic music," in *Proc. ISMIR*, Suzhou, China, 2017, pp. 63–70.

[10] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[11] Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[12] Ping Gao, Cheng-You You, and Tai-Shih Chi, "A multi-scale fully convolutional network for singing melody extraction," in *Proc. APSIPA ASC*. IEEE, 2019, pp. 1288–1293.

[13] Hsin Chou, Ming-Tso Chen, and Tai-Shih Chi, "A hybrid neural network based on the duplex model of pitch perception for singing melody extraction," in *Prpc. ICASSP*. IEEE, 2018, pp. 381–385.

[14] Shuai Yu, Xiaoheng Sun, Yi Yu, and Wei Li, "Frequency-temporal attention network for singing melody extraction," in *Proc. ICASSP*. IEEE, 2021, pp. 251–255.

[15] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1482–1491, 2012.

[16] Taishih Chi, Powen Ru, and Shihab A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[17] Tomoyasu Nakano, Kazuyoshi Yoshii, Yiming Wu, Ryo Nishikimi, Kin Wah Edward Lin, and Masataka Goto, "Joint singing pitch estimation and voice separation based on a neural harmonic structure renderer," in *Proc. WASPAA*. IEEE, 2019, pp. 160–164.

[18] Yu-Te Wu, Berlin Chen, and Li Su, "Multi-instrument automatic music transcription with self-attention-based instance segmentation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 2796–2809, 2020.

[19] Joshua P Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel, "Mt3: Multi-task multitrack music transcription," in *Proc. ICLR*, 2022.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[21] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[22] Alain de Cheveigné, "Pitch perception," *Oxf. Handb. Audit. Sci. Hear*, vol. 3, pp. 71, 2010.

[23] Ching-Yu Chiu, Wen-Yi Hsiao, Yin-Cheng Yeh, Yi-Hsuan Yang, and Alvin Wen-Yu Su, "Mixing-specific data augmentation techniques for improved blind violin/piano source separation," in *Proc. MMSP*, 2020, pp. 1–6.

[24] Nazif Can Tamer, Pedro Ramoneda, and Xavier Serra, "Violin etudes: a comprehensive dataset for f0 estimation and performance analysis," in *Proc. ISMIR*, Bengaluru, India, 2022.

[25] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse H. Engel, and Douglas Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. ICLR*, New Orleans, Louisiana, USA, 2019.

[26] John Thickstun, Zaïd Harchaoui, and Sham M. Kakade, "Learning features of music from scratch," in *Proc. ICLR*, Toulon, France, 2017.

[27] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. ISMIR*, Taipei, Taiwan, 2014, pp. 155–160.

[28] Antoine Liutkus and Roland Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. ICASSP*, Brisbane, Australia, April 2015, pp. 266–270.

[29] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-Unmix – A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, 2019.

[30] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *JOSS*, vol. 5, no. 50, pp. 2154, 2020, Deezer Research.

[31] Yigitcan Özer and Meinard Müller, "Source separation of piano concertos with test-time adaptation," in *Proc. ISMIR*, Bengaluru, India, 2022.