

# Source Separation-based Data Augmentation for Improved Joint Beat and Downbeat Tracking

Ching-Yu Chiu

Graduate Program of Multimedia Systems and Intelligent Computing  
National Cheng Kung University and Academia Sinica, Taiwan  
sunnycyc@citi.sinica.edu.tw

Joann Ching

Research Center for IT Innovation  
Academia Sinica, Taiwan  
joann8512@citi.sinica.edu.tw

Wen-Yi Hsiao

Yating Music Team  
Taiwan AI Labs, Taiwan  
wayne391@ailabs.tw

Yu-Hua Chen

Yating Music Team  
Taiwan AI Labs, Taiwan  
cloudstrife60138@gmail.com

Alvin Wen-Yu Su

Dept. CSIE  
National Cheng Kung University  
alvinsu@mail.ncku.edu.tw

Yi-Hsuan Yang

Research Center for IT Innovation  
Academia Sinica, Taiwan  
yang@citi.sinica.edu.tw

**Abstract**—Due to advances in deep learning, the performance of automatic beat and downbeat tracking in musical audio signals has seen great improvement in recent years. In training such deep learning based models, data augmentation has been found an important technique. However, existing data augmentation methods for this task mainly target at balancing the distribution of the training data with respect to their tempo. In this paper, we investigate another approach for data augmentation, to account for the composition of the training data in terms of the percussive and non-percussive sound sources. Specifically, we propose to employ a blind drum separation model to segregate the drum and non-drum sounds from each training audio signal, filtering out training signals that are drumless, and then use the obtained drum and non-drum stems to augment the training data. We report experiments on four completely unseen test sets, validating the effectiveness of the proposed method, and accordingly the importance of drum sound composition in the training data for beat and downbeat tracking.

**Index Terms**—Beat tracking, downbeat tracking, source separation, data augmentation

## I. INTRODUCTION

Beats and downbeats are usually referred to as a sequence of time instants that human would tap their feet to [1], [2]. Beats and downbeats define the metrical structure of a musical piece, and enable numerous higher level tasks such as structure segmentation, automated DJ mixing, and score alignment [2]–[4]. Moreover, they are essential foundations for machines to understand and analyze music. As humans are able to track musical beats without difficulty in various musical genres regardless of the existence of drum sounds in an audio signal, people expect machines to have similar skills. Thanks to cumulative efforts in the research community, recent years witness great success of deep learning-based supervised models for *joint beat and downbeat tracking*—i.e., a single model that estimates beats and downbeats at the same time—for music with steady and strong beats, such as pop, rock, and dance music [5]. We also focus on the joint beat/downbeat tracking task in this paper.

Existing deep learning-based methods for beat/downbeat tracking are mainly supervised models that require labeled

training data, namely audio with annotated labels of beat and downbeat timing. These approaches usually consist of a neural network module and a post-processing dynamic model. The neural network module takes low-level features such as chroma, spectral flux, or first-order derivatives of magnitude spectrograms [2], [6]–[8] as input, and outputs an activation function that indicates the most likely beats and downbeats time candidates. The post-processing model, be it a dynamic Bayesian network (DBN), hidden Markov model (HMM) or conditional random field (CRF), makes binary predictions from the activation function [2], [8], [9].

Challenges for beat/downbeat tracking can be viewed from the inherent signal characteristics for different type of music, such as strong tempo variation, various tempo distribution, rhythmic variation or syncopation for purposes of creating metrical tension, and the rather blurred note onsets and transitions caused by non-percussive instruments [10]–[13]. From a machine learning perspective, the difficulty may also be related to limitations of model capacity, insufficient labeled data for training, or potential imbalance of training data. For example, we note that most available training datasets for beat/downbeat tracking are pop/rock songs [5]. Moreover, while deep learning models perform well, the performance of a deep learning model for beat/downbeat tracking seems to be sensitive to the composition of training data. We found the training models with the same model architecture, comparable training data size, yet different data constitution (e.g., tempo distribution, percussive/non-percussive sound source percentage) can lead to significant performance difference on different test sets.

Although not often mentioned together with beat tracking, music source separation has been another important topic in music information research. Given a monaural input mixture (i.e., an audio recording of multiple instruments sounding together), a source separation model generates separated stems of different sound sources that composed the mixture [14]–[18]. Noticing that the total duration and metrical structure of the stems are basically the same, we conjecture that it may be useful to apply drum/non-drum separation as a means of

data augmentation, to increase the amount of drum or non-drum data in our training set. It is therefore our goal in this paper to investigate variants of such augmentation data and their usefulness in improving the performance of joint beat and downbeat tracking. To the best of our knowledge, such a study has never been presented before.

As our focus is on data augmentation, for the model architecture, we simply adopt the pipeline of cascading a long short-term memory (LSTM) network with an HMM-based dynamic network proposed by Madmom [2], [19], which represents the state-of-the-art for joint beat and downbeat tracking across various genres. We evaluate the performance of our models on four completely unseen datasets in our experiments, namely ASAP [20], Rock [21], HJDB [11], and RWC Royalty-Free [22]. We plan to open source our code at <https://github.com/SunnyCYC/aug4beat>.

## II. BACKGROUND

As the available labeled data for training supervised deep learning models for beat tracking is limited, research has been done to enhance the generalizability of models, or to compensate for the relatively insufficient properties in the training set. Giorgi *et al.* [23], for example, proposed a deterministic time-warping operation to help their model learn rhythmic patterns independently of tempo. Böck and Davies [5] devised a novel multi-task approach to leverage shared connections in musical structure, and to simultaneously estimate tempo, beat, and downbeat. They also proposed an augmentation method based on changing parameters of the short-time Fourier transform to expose their model to a wider range of tempi. Zapata and Gomez [3] proposed an audio voice suppression technique and a simple low-pass filter to improve beat tracking.

Efforts have been made to tackle beat tracking by treating percussive/non-percussive features separately. Goto [24] developed a method to judge if the input audio signal contains drum sounds, and then used different ways to track the beats for music with or without drum sounds. Gkiokas *et al.* [25] presented an tempo estimation and beat tracking algorithm which utilized source separation to extract features of percussive/harmonic components separately. Our work is different from these prior arts in that we use source separation as a means to create augmented data for training a supervised neural network model.

We implement a beat/downbeat tracking model on our own following the LSTM+HMM architecture described in the Madmom library [2], [19]. The network consists of three fully-connected bidirectional recurrent layers with 25 LSTM units each. After the LSTM layers, a softmax classification layer with three units produces three activation functions (i.e., curves) corresponding to the probability of a frame being ‘a beat but no a downbeat,’ ‘a downbeat’ or ‘a non-beat’ position. The output activation functions are then processed by an HMM to produce the final binary beat/downbeat predictions. In other words, the model predicts beat and downbeat jointly. In this paper, the bar length setting required by HMM is set as three or four beats, following the default setting of Madmom.

TABLE I  
THE DATASETS EMPLOYED FOR MODEL TRAINING (TOP) AND TESTING (BOTTOM) IN OUR EXPERIMENTS; ‘#’ DENOTES THE NUMBER OF MUSICAL PIECES. THE LAST TWO COLUMNS INDICATE THE PERCENTAGE OF THE SELECTED DRUM STEMS FOR THE DATASET. MOST OF THE 10 DATASETS LISTED HERE HAVE THE DRUM STEMS AFTER SOURCE SEPARATION, EXCEPT FOR RWC CLASSICAL [22] AND ASAP [20], WHICH HAS NO PRESENCE OF DRUM SOUNDS AT ALL.

Dataset	#	Total duration	drum presence rate	
			ABSM	OSFQ
RWC Classical [22]	54	5h 19m	0.0%	0.0%
RWC Jazz Music [22]	50	3h 42m	29.8%	64.1%
RWC Music Genre [26]	100	7h 20m	39.3%	62.6%
Ballroom [27], [28]	685	5h 57m	59.1%	82.2%
Hainsworth [29]	222	3h 19m	65.0%	75.1%
GTZAN [30], [31]	999	8h 20m	71.2%	79.0%
Carnatic [32]	176	16h 38m	72.8%	84.2%
Beatles [33]	180	8h 09m	75.8%	94.0%
RWC Popular [22]	100	6h 47m	89.5%	90.8%
Robbie Williams [34]	65	4h 31m	93.5%	98.5%
ASAP [20]	520	48h 07m	0.0%	0.0%
Rock [21]	200	12h 53m	83.2%	97.1%
HJDB [11]	235	3h 19m	99.2%	99.2%
RWC Royalty-Free [22]	15	27m	99.9%	99.9%

## III. PROPOSED METHOD

### A. Source Separation-based Data Augmentation

We propose to employ a source separation model to isolate out the drum sounds and non-drum sounds for the musical pieces in the training data for beat/downbeat tracking. In doing so, we employ Spleeter [16], the current state-of-the-art model for musical source separation that has publicly available model checkpoints. Spleeter can generate four separated stems, each stem corresponding to the sounds of the vocals, drums, bass, and others, respectively. We simply sum all but the drum stem as the *non-drum* stem. If we call the original pieces in the training set as *mixes*, we would get the same number of drum stems and non-drum stems after source separation. For those pieces that correspond to the same song, we assume that they share the same metrical structure. Accordingly, we can assume that the labels of beat and downbeat timing apply equally well to that corresponding drum stem and non-drum stem.

### B. Drum Stem Selection

Since some of our training data may not contain any drum sounds originally, the drum stems derived from them would be nearly silent and could be harmful if adopted as training data. We devise two drum selection criteria to tackle this. **ABSM**: Exclude a drum stem if the mean value of its absolute magnitude in the time domain is less than 0.01 (while the maximum absolute magnitude is 1), an empirically set threshold. **OSFQ**: Exclude a drum stem if the mean value of its absolute magnitude is less than 0.001 (i.e., looser than ABSM), or if it has less than one prominent onset per second, as detected by `librosa.onset.onset_detect` [35]—i.e., when its onset patterns is sparse or irregular. See Figure 1 for an example of bass drum excluded by OSFQ.

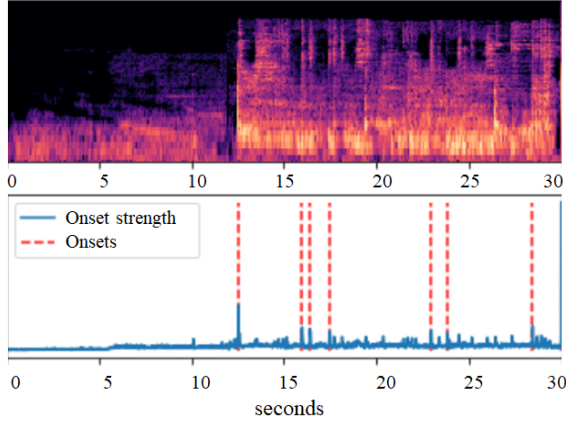


Fig. 1. An example of a separated drum stem that would pass the ABSM filter as it has recognizable drum sounds, yet would not pass the OSFQ filter, for it lacks regular beat structure as can be seen from the result of onset detection in the bottom figure. The top figure shows its spectrogram.

After drum selection, we have in total four types of data: *Mix*, *non-drum*, *only-drum* (ABSM) and *only-drum* (OSFQ), for all the pieces in our training set. To manipulate the percentage of drum sounds in our training set as data augmentation, we can use different combinations of them. For example, we can use *Mix* with *non-drum/only-drum* to reduce/increase the drum sound percentage of our training data, while both usage can result in comparable total amount of training data. Please see Table I for the datasets employed in training our models.

#### IV. EXPERIMENT

We implement models with five data type combinations that are enabled by source separation, as listed in Table II. The goal is to see any of these can outperform the baseline case of using the *Mix* alone for model training, as it represents the standard approach as of now. For these models, we randomly split the top 10 datasets listed in Table I into 80%, 10%, 10% for training, validation and test. During validation, the parameters of the post-processing HMM model are optimized. The proposed data augmentation applies to the training sets only. To alleviate the random effect, we repeat the training of all models for five times and report the averaged evaluation results. Due to space limit, we only report the results on the four unseen test sets on the bottom of Table I. The result on the testing split of the top 10 datasets is consistent with that on the unseen test sets.

We consider the model trained with the original mixes as the *baseline* in our experiments. We call it “Madmom-like,” for it uses the same 3-layer bidirectional LSTM+HMM architecture as the RNNDowBeatProcessor of Madmom [2], [19].

##### A. Result on the Four Unseen Test Sets

Table III shows the F-measure (F1) results of the proposed methods. The following observations can be made. Except for case of the ASAP test set, which is composed of expressive classical piano performance, the augmented models obtain

TABLE II  
THE DATA TYPE COMBINATIONS TESTED IN OUR EXPERIMENTS, ALONG WITH THE SHORT NAMES OF THE CORRESPONDING MODELS. EACH MODEL IS TRAINED FROM SCRATCH USING THE TICKED DATA TYPES OF THE TRAINING SPLIT OF THE 10 DATASETS LISTED IN TABLE I.

Model	Data			
	mix	non-drum	onlyDrum ABSM	OSFQ
Mix (baseline)	✓			
Mix+no-drum	✓	✓		
Mix+only-drum_ABSM	✓		✓	
Mix+only-drum_OSFQ	✓			✓
3combABSM	✓	✓	✓	
3combOSFQ	✓	✓		✓

different levels of improvement in the other three test sets. We speculate that this is due to the fact that the dataset characteristics (e.g., local tempo variation, rhythmic pattern, note onset/offset patterns) of ASAP is quite different from most of our training sets. As a result, even though the *non-drum* stems are adopted in *Mix+no-drum* to increase the percentage of non-drum audio signals in the training set, the difference between the *non-drum* of the train sets and ASAP still deteriorate the performance.

For results of HJDB—a drum-heavy dataset comprising Hardcore, Jungle and Drum&Bass music excerpts [11], models augmented by *onlyDrum* stems gains clear improvement (e.g., at least +5.5% relative improvement in downbeat), while model augmented by *non-drum* stem gets relatively limited improvement. Such an observation implies the noticeable influence of trainset sound source composition on beat/downbeat tracking performance on different test sets. We also notice the unexpected downbeat improvement of *Mix+no-drum* on RWC Royalty-Free. One reason for this lies within the difference between HJDB and RWC Royalty-Free. While songs in HJDB are dominated by drum sound all the time, songs in RWC Royalty-Free may be without drum sound for several measures.

The performance improvement on the Rock dataset is less obvious than that on either HJDB or RWC Royalty-Free. This may be due to the fact that the majority of our training data (as shown in Table I) is composed of rock/pop music—the baseline model may have learned most of the beat/downbeat related patterns required for tracking Rock music. The downbeat improvements are consistently larger than the beat ones, which may suggest downbeat gains more advantage from the proposed data augmentation.

In summary, it is clear that the test sets with high drum presence rate can benefit from the extra stems derived from source separation and drum selection. And the exact best combinations/composition of training data depends on the characteristics of target test set.

##### B. Case Study: Analysis of Improvement on HJDB

To have an idea of the underlying causes for improvements on HJDB, we compare the model activation functions,

TABLE III

THE RESULTING F1 SCORES (AVERAGED OVER FIVE RUNS) IN BEAT AND DOWNBEAT TRACKING IN FOUR DIFFERENT TEST SETS UNSEEN AT TRAINING TIME. EACH MODEL IS TRAINED FROM SCRATCH USING THE TICKED DATA TYPES, SHOWN IN TABLE II, OF THE TRAINING SPLIT OF THE 10 DATASETS LISTED IN TABLE I. IN EACH COLUMN, WE HIGHLIGHT THE RESULT THAT OUTPERFORMS BASELINE IN BOLD, AND THE BEST RESULT WITH UNDERLINE.

Model	ASAP		Rock		HJDB		RWC Royalty-Free	
	beat F1	downbeat F1	beat F1	downbeat F1	beat F1	downbeat F1	beat F1	downbeat F1
Madmom [2], [19]-like baseline	0.591	0.406	0.874	0.788	0.882	0.679	0.833	0.810
Mix+no-drum	0.586	0.398	0.869	<b>0.796</b>	<b>0.900</b>	<b>0.685</b>	<b>0.861</b>	<b>0.853</b>
Mix+only-drum_ABSM	0.589	0.402	<b>0.882</b>	<b>0.800</b>	<b>0.896</b>	<b>0.735</b>	<b>0.865</b>	<b>0.852</b>
Mix+only-drum_OSFQ	0.590	0.401	0.874	<b>0.790</b>	<b>0.894</b>	<b>0.771</b>	<b>0.836</b>	<b>0.826</b>
3combABSM	0.583	0.394	0.870	<b>0.794</b>	<b>0.886</b>	<b>0.758</b>	<b>0.837</b>	<b>0.812</b>
3combOSFQ	0.581	0.399	0.868	0.783	<u>0.902</u>	<u>0.757</u>	<b>0.847</b>	<b>0.843</b>

beat/downbeat estimations/annotations and corresponding input feature, and show one representative example from HJDB in Figure 2. From the top three feature plots, we can observe the following features of this song: the relative sparseness of non-drum feature reflects the drum sound dominance. The number of beats in a time span of three seconds, and the several drum onsets in between beats indicate this is a fast tempo song with frequent use of syncopation. These are features distinguishing HJDB from the other datasets we employed. From the beat/downbeat estimations (blue/black vertical lines), we can see that the drum stem-augmented models (i.e. *3combOSFQ* and *Mix+only-drum\_OSFQ*) lead to predictions that are closer to the ground-truth (red vertical lines) than the baseline. However, such an improvement can not be achieved without the help of HMM. From the several ‘false-positive’ activation peaks between beats produced by the three models, we can see the syncopation clearly ‘misleads’ the three models. As the underlying mechanism of HMM decoding is to find the global optimal path, HMM has the chance to ‘correct’ the local activation error as long as the model also produces peaks high enough at the right beat positions. From the three red squares indicating the critical beat positions, we can see that *Mix+only-drum\_OSFQ* is able to generate peak at the right positions, while *3combOSFQ* fails for the second square but benefits from HMM. And the failure of *baseline* to generate clear peaks at the three positions not only prevents it from the help of HMM but also deteriorates HMM’s judgement on downbeat (highlighted by green square). Looking back to the drum feature again, it seems reasonable that the augmented models may have better chance than *baseline* to learn the idea of syncopation. While the truth beat/downbeat positions may fall in nearly silent positions (e.g., between drum onsets) in drum stems, such cases are less frequent in the original audio mixture due to the presence of other non-drum sounds. In summary, this example demonstrates the underlying reasons for the performance improvement derived by the proposed methods, and reveals that the HMM plays some role in taking advantage of the improved network activation prediction.

## V. CONCLUSION

In this paper, we have presented a source separation-based data augmentation technique that utilizes the drum/non-drum

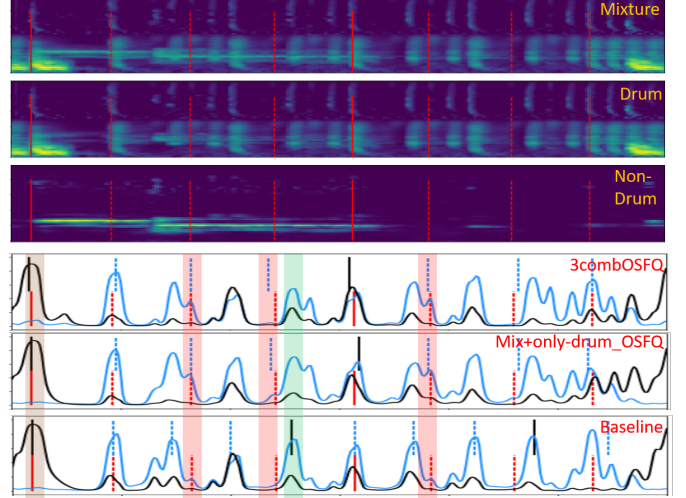


Fig. 2. Estimation (vertical lines) and activation (curves) of beat (blue) and downbeat (black) on a clip of a song (cut from 17–21 seconds) from HJDB [11], named “Sweet Vibrations.” The input features (magnitude spectrogram and its first order derivative) calculated from mixture, drum stem, and non-drum stem are also partially shown for comparison. Red vertical lines indicate the ground-truth annotation of beats (dashed) and downbeats (solid). In this example, the beat/downbeat estimation (finalized by HMM) of the augmented models, *3combOSFQ* and *Mix+only-drum\_OSFQ*, are comparable and are closer to ground-truth, while the *baseline* model gradually deviates from the ground-truth annotations. Brown square on the left hand side highlights the correct estimation of downbeats and ensures the latter errors are not propagated from previous unshown part of the song. Red/green squares highlight the critical positions of beats/downbeats that cause the failure of *baseline*. The difficulty of this song mainly comes from the frequent syncopation (i.e., the drum onsets between beats), which could cause false-positive activation peaks between beats. Although all the three models exhibit such false-positive activation peaks, the augmented models somehow learn the idea of syncopation and are able to generate activation peaks at the nearly silent beat positions, enabling the HMM to ‘correct’ the final results.

stems separated out by source separation to increase the presence of drum-only and non-drum data in the training set. We have also reported experiments showing that the performance of joint beat/downbeat tracking can depend on the sound source composition of the training set, and that the proposed data augmentation method leads to performance improvement across different test sets. For future work, we are interested in building a beat/downbeat tracking model that incorporates source separation as part of the model, and in conducting more experiments on classical music pieces.

## REFERENCES

- [1] S. Böck, F. Krebs, and G. Widmer, "A multi-model approach to beat tracking considering heterogeneous music styles," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 603–608, 2014.
- [2] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 255–261, 2016.
- [3] J. R. Zapata and E. Gomez, "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 51–55, 2013.
- [4] H. Grohganz, M. Clausen, and M. Müller, "Estimating musical time information from performed MIDI files," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 35–40, 2014.
- [5] S. Böck and M. E. P. Davies, "Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation," *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2020.
- [6] S. Durand, J. P. Bello, B. David, and G. Richard, "Downbeat tracking with multiple features and deep neural networks," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 409–413, 2015.
- [7] F. Krebs, S. Böck, M. Dorfer, and G. Widmer, "Downbeat tracking using beat-synchronous features and recurrent neural networks," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 129–135, 2016.
- [8] M. Fuentes, B. McFee, H. C. Crayencour, S. Essid, and J. P. Bello, "Analysis of common design choices in deep learning systems for downbeat tracking," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 106–112, 2018.
- [9] S. Durand, J. P. Bello, B. David, and G. Richard, "Robust downbeat tracking using an ensemble of convolutional networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 72–85, 2017.
- [10] G. Sioros, M. Miron, M. E. P. Davies, F. Gouyon, and G. Madison, "Syncopation creates the sensation of groove in synthesized music examples," *Frontiers in Psychology*, vol. 5, pp. 1036, 2014.
- [11] J. A. Hockman, M. E. P. Davies, and I. Fujinaga, "One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 169–174, 2012.
- [12] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? A case study on Chopin Mazurkas," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 649–654, 2010.
- [13] B. McFee and D. P. Ellis, "Better beat tracking through robust onset aggregation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2154–2158, 2014.
- [14] S. Uhlich et al., "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 261–265.
- [15] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.
- [16] R. Hennequin, A. Khilif, F. Voituret, and M. Moussallam, "Spleeter: A fast and state-of-the-art music source separation tool with pre-trained models," *J. Open Source Softw.*, 2020.
- [17] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, pp. 1667, 2019.
- [18] C.-Y. Chiu, W.-Y. Hsiao, Y.-C. Yeh, Y.-H. Yang, and A. W.-Y. Su, "Mixing-specific data augmentation techniques for improved blind violin/piano source separation," *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, 2020.
- [19] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "Madmom: A new python audio and music signal processing library," *Proc. ACM Multimed. Conf.*, pp. 1174–1178, 2016.
- [20] F. Foscarin, A. McLeod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: a dataset of aligned scores and performances for piano transcription," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2020, pp. 534–541.
- [21] T. de Clercq and D. Temperley, "A corpus analysis of rock harmony," *Popular Music*, vol. 30, no. 1, pp. 47–70, 2011.
- [22] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, Classical, and Jazz music databases," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2002, pp. 287–288.
- [23] B. D. Giorgi, M. Mauch, and M. Levy, "Downbeat tracking with tempo-invariant convolutional neural networks," *Proc. Int. Soc. Music Inf. Retr. Conf.*, pp. 216–222, 2020.
- [24] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.
- [25] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stajylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 421–424, 2012.
- [26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music genre database and musical instrument sound database," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2003, pp. 229–230.
- [27] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1832–1844, 2006.
- [28] F. Krebs, S. Böck, and G. Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2013.
- [29] M. Macleod and S. Hainsworth, "Particle filtering applied to musical tempo tracking," *EURASIP Journal on Advances in Signal Processing*, no. 927847, 2004.
- [30] U. Marchand and G. Peeters, "Swing ratio estimation," in *Proc. Digital Audio Effects*, Trondheim, Norway, 2015.
- [31] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [32] A. Srinivasamurthy and X. Serra, "A supervised approach to hierarchical metrical cycle tracking from audio music recordings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 5217–5221.
- [33] M. E. P. Davies, N. Degara Quintela, and M. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," in *Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.
- [34] B. D. Giorgi, M. Zanoni, A. Sarti, and S. Tubaro, "Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony," in *Proc. Int. Workshop on Multidimensional Systems*, 2013, pp. 1–6.
- [35] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," *Proc. Python in Science Conf.*, pp. 18–24, 2015.