PSEUDO-LABEL TRANSFER FROM FRAME-LEVEL TO NOTE-LEVEL IN A TEACHER-STUDENT FRAMEWORK FOR SINGING TRANSCRIPTION FROM POLYPHONIC MUSIC

Sangeun Kum[1]   *Jongpil Lee*[1]   *Keunhyoung Luke Kim*[1]   *Taehyoung Kim*[1]   *Juhan Nam*[1,2]

[1]Neutune Research, Seoul, South Korea
[2]Graduate School of Culture Technology, KAIST, Daejeon, South Korea

ABSTRACT

Lack of large-scale note-level labeled data is the major obstacle to singing transcription from polyphonic music. We address the issue by using pseudo labels from vocal pitch estimation models given unlabeled data. The proposed method first converts the frame-level pseudo labels to note-level through pitch and rhythm quantization steps. Then, it further improves the label quality through self-training in a teacher-student framework. To validate the method, we conduct various experiment settings by investigating two vocal pitch estimation models as pseudo-label generators, two setups of teacher-student frameworks, and the number of iterations in self-training. The results show that the proposed method can effectively leverage large-scale unlabeled audio data and self-training with the noisy student model helps to improve performance. Finally, we show that the model trained with only unlabeled data has comparable performance to previous works and the model trained with additional labeled data achieves higher accuracy than the model trained with only labeled data.

Index Terms— singing transcription from polyphonic music, pseudo label, teacher-student framework, music information retrieval

1. INTRODUCTION

The goal of singing transcription from polyphonic music (STP) is to transcribe the monophonic vocal in polyphonic music into musical notes denoted by onset, offset, and the score pitch. STP includes several sub-tasks such as singing voice detection, pitch estimation, note-level segmentation, and onset/offset detection [1]. It is a challenging task due to the high variability of singing voice in terms of timbre, expressions or formant modulation. Furthermore, multiple instrument sources in polyphonic music make the task even harder. Therefore, many of the singing transcription methods have focused on monophonic vocal sources [2, 1] or have employed pre-trained source separation models as pre-processing to extract vocal sources from polyphonic music [3, 4].

Although we can leverage the power of a deep neural network in supervised settings for STP, the major obstacle is the lack of large-scale note-level label data. While there are several datasets released for singing transcription, they cover only monophonic singing voice [5, 6] or have low quality as the labeling was automated without manual refinement [4, 7]. The lack of note-level label data can be addressed by additional singing voices that sung the song [8] but it requires more resources and has an issue of data distribution. A recent work released a dataset of human-annotated note-level labels for 500 Chinese songs [4]. While this is highly beneficial for STP

research, such manual note-level labeling is time-exhausting and expensive.

We address the lack of labeled data by using pseudo labels from vocal pitch estimation models. Vocal pitch estimation or melody extraction is a sub-task of STP that estimates the frame-level instantaneous pitch of vocal source music [9]. Recently, deep neural network models for the task have been extensively studied, showing superior performances [10, 11, 12, 13, 14]. Using the vocal pitch estimation models, we first convert the frame-level pitch contours to note-level piano roll as pseudo labels. We then use the the note-level pseudo labels for self-training of a STP model in a teacher-student framework. Along with the noisy student model that employs various data augmentation techniques in a teacher-student framework [15], we show that the newly trained neural network model consistently improves the performance for STP. We validate the proposed method in various experiment setups using the Cmedia and MIR-ST500 datasets. Finally, we show that the model trained with only unlabeled data has reasonable performances compared to previous works and, the model trained with additional labeled data, achieves higher accuracy than the model trained with only labeled data.

2. METHOD

The proposed method is composed of two stages. The first stage extracts F0 pitch contours from vocal pitch estimation models and converts them to note-level pseudo labels. The second stage uses the note-level pseudo labels to train a new neural network for STP in a teacher-student framework. The details of each stage are explained below.

2.1. Pseudo Labels: Frame-level to Note-level

Figure 1 (a) illustrates the conversion steps from frame-level pitch contours to note-level piano rolls. The pitch contours are obtained from a vocal pitch estimation model given unlabeled audio data (polyphonic music with vocals). The first step is pitch quantization which rounds the continuous pitch to semi-tone steps. In many pitch estimation models based on neural networks, the output is represented as the softmax function over quantized pitch values where the neighboring pitches are much smaller than one semi-tone [12, 14]. We take the pitch with the highest confidence and quantize it to the nearest MIDI note number. The second step is rhythm quantization. This processing "snaps" the fragments of the quantized pitch lines to beat-based units. It is carried out by smoothing the quantized pitch with a series of three median filters. In order to make the filtered outputs beat-based units, we set the sizes of

(a) Conversion from pitch contours to piano rolls

(b) The proposed teacher-student framework.
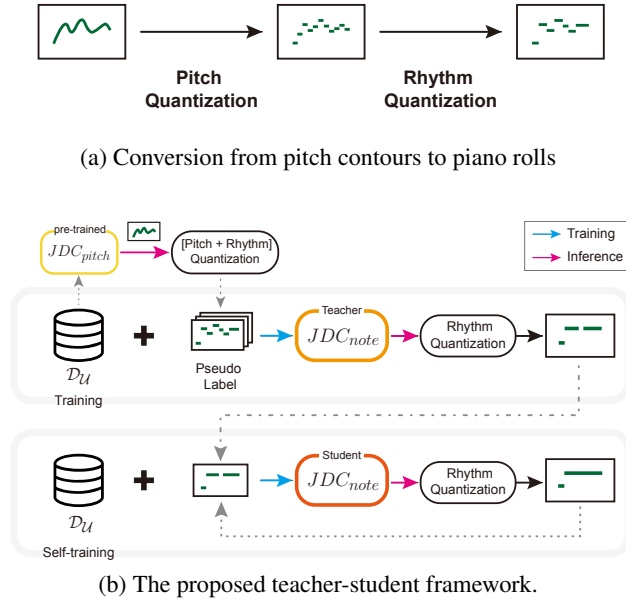
Fig. 1. Illustrations of pseudo label transfer from frame-level to note-level in the teacher-student framework. $\mathcal{D}_{\mathcal{U}}$ is a large-scale unlabeled audio data, JDC_{pitch} is a pretrained vocal pitch estimation model, and JDC_{note} is a STP model trained with note-level pseudo labels.

the median filters to 1/32, 1/16, and 1/12 beat, respectively, given a tempo. We estimated the tempo for each audio track using the function in librosa [16]. We observed that the three cascaded filters progressively improved the label quality. In addition, we removed small fragments that were too short to be considered as singing notes. In the experiment, we set the threshold to 1/16 beat. Finally, we added a simple rule to minimize octave errors.

2.2. Self-Training in the Teacher-Student Framework

Figure 1 illustrates the proposed teacher-student framework. We first build a new neural network model for STP and train it using the note-level pseudo labels obtained from the first stage. We then iterate the training using the same configuration of neural network in the teacher-student framework. This self-training in a teacher-student framework has been studied in diverse domains such as image classification [17, 18], speech recognition [19], and audio classification [20]. In particular, random data augmentation is a key element to improve the model performance. In our method, we primarily used the noisy student model where the student network, takes unlabeled data that is perturbed by the random audio augmentation while the teacher network takes the original input to generate the pseudo-labels [15]. For random audio augmentation, multiple audio effects are randomly applied such as overdrive, reverb, audio equalizer, and audio filter to manipulate the audio. In addition, between the teacher network and the student network, we add the rhythm quantization as a post-processing to enhance the transcription performance. For applying the median filtering and the following rules in the rhythm quantization, we use one-hot vectors (i.e., hard label) as pseudo labels. In the mean time, we compare the noisy student
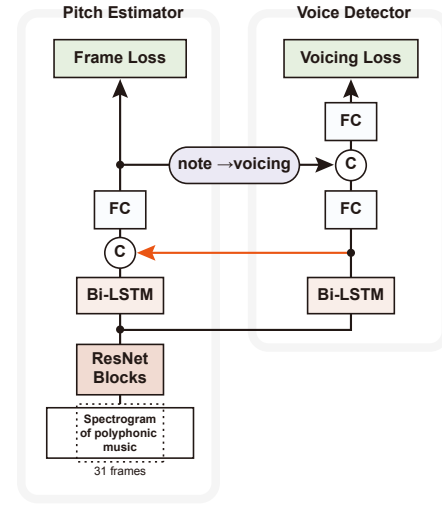
Fig. 2. The model architecture for JDC_{note}. "C" indicates feature concatenation.

model to the basic teacher-student model which does not use the random audio augmentation in the experiment.

2.3. Model Architecture

The model architecture for STP is based on the joint detection and classification (JDC) model, which which is originally proposed for vocal melody extraction [14]. The JDC model is a convolutional recurrent neural network (CRNN) with two outputs where one detects the presence of a singing voice and the other estimates the pitch. We modified the JDC model in two folds. First, we set the softmax output of the pitch estimator to have the semi-tone resolution. Second, we added a connection from the output of Bi-LSTM in the voice detector to the pitch estimator as seen in Figure 2. The STP model is denoted as JDC_{note}. In our preliminary study, we observed that the modified JDC model performs better than the original one in STP (the average COnPOff is 2.25% higher on the MIR-ST500 and Cmedia test sets). In the experiments, we also use the original JDC model as a pretrained vocal pitch estimator. The model is denoted as JDC_{pitch} to distinguish it from JDC_{note}.

3. EXPERIMENTS

3.1. Datasets

We used various labeled or unlabeled datasets that contain singing voice for training or testing the STP model.

3.1.1. Training data

We used a large-scale unlabeled datasets ($\mathcal{D}_{\mathcal{U}}$) to generate pseudo labels in the proposed method. They include two public datasets (DSD100 [21] and Free Music Archive (FMA) [22]) and an in-house dataset. As to FMA, we used fma_large, a subset of FMAs with up to 106,574 tracks. Since it contains numerous non-vocal tracks covering a variety of genres and sounds, we mitigated the data imbalance between vocal and non-vocal examples by selecting only vocal tracks as in [15]. The in-house dataset contains 2000 K-pop songs crawled from YouTube. We also used a labeled dataset

	Initial Pseudo Labels		JDC_{note}	
Repurposed Models	Demucs + CREPE	JDC_{pitch}	Demucs + CREPE	JDC_{pitch}
COnPOff	22.43	25.44	24.71	28.97
COnP	45.01	48.48	48.64	53.32
COn	57.65	61.94	62.32	64.74

Table 1. Comparison of pitch estimation models for their initial note-level pseudo labels (Initial Pseudo Labels) and the predicted results from JDC_{note} trained with the initial pseudo labels. They were all evaluated on Cmedia.

	Cmedia		MIR-ST500	
Models	TS	NS	TS	NS
COnPOff	28.97	29.62	22.12	22.62
COnP	53.32	54.55	40.01	40.70
COn	64.74	65.61	56.90	57.87

Table 2. Comparison of the basic teacher-student model (TS) and noisy student (NS) model.

for STP for the ablation study. MIR-ST500 [4] consists of 500 Chinese pop songs and human-labeled notes that correspond to the vocal melody. Among them, we used the official split of 400 songs as a training set.

3.1.2. Test data

We evaluated the models mainly with two public test sets, Cmedia and MIR-ST500 (the test split). The Cmedia dataset provides note-level labels and YouTube links to download audio files. It has been used in the Music Information Retrieval Evaluation eXchange (MIREX 2020) [1]. It contains 100 Chinese pop songs. We used it only for testing in this experiment to compare performance with other models. For MIR-ST500, we used the official split of 100 songs as a test set.

3.2. Training Details

The model configuration of the CRNN architecture in JDC_{note} is almost the same as JDC_{pitch} in [14]. We trained the JDC_{note} using the Adam optimizer for 100 epochs with a learning rate of 0.003, batch size of 64 on 1 GPU. We used a learning rate schedule that reduces the learning rate by 0.7 times if validation accuracy did not increase within 9 epochs. The model and the training procedures were implemented using Keras with TensorFlow 2.3. The source code of the proposed method and the models are available at `https://github.com/keums/icassp2022-vocal-transcription`.

4. ABLATION STUDY

We conducted an ablation study to verify the proposed method for STP. In the following experiments, we evaluated STP performance using the metrics proposed in [6]: COn (Correct Onset), COnP (Correct Onset and Pitch), and COnPOff (Correct Onset, Pitch and

Offset). We set the onset threshold to 50 msec, and the offset threshold to 50 msec or (0.2×note duration), whichever is higher, following the previous works. We computed the average F1-score for the metrics using *mir_eval* [23].

4.1. Comparison of Pitch Estimation Models

The goal of this experiment is to evaluate the efficacy of the repurposed neural network models that predict vocal pitch contours as a note-level pseudo-label generator in our method. While we primarily used the pretrained JDC_{pitch} [14][2] for vocal pitch estimation in the experiments, we also compared it to a combination of Demucs [24] and CREPE [12] because they are widely used pretrained models for sound source separation and monophonic pitch estimation, respectively. Specifically, we separated out vocal audio from the Demucs outputs and fed them to CREPE to estimate frame-level pitch. We converted the frame-level pitch contours to note-level pseudo labels based on the method in Section 2.1. We first evaluated the accuracy of the note-level pseudo labels and show the result on the left side of Table 1. While Demucs and CREPE are models with high performance, JDC_{pitch} achieves 3 to 4% higher accuracy than the combination in the three metrics. This is presumably because the separated vocal stem from Demucs includes multiple vocal sources (e.g., chorus ensembles) and CREPE predicts discontinuous pitch contours switching between different voices. On the other hand, JDC_{pitch} is more robust to multiple vocal sources because it was trained to extract the main vocal melody from polyphonic music. As a next step, we trained a new neural model, JDC_{note}, using the note-level pseudo labels and evaluated the performance to confirm the efficacy of the repurposed neural network models. The right side of Table 1 shows that JDC_{pitch} still performs better, having more accuracy gaps. Therefore, we used the note-level pseudo labels from JDC_{pitch} for the remaining experiments.

### 4.2. Basic Teacher-Student VS. Noisy Student

In the experience above, we used the note-level pseudo labels from pitch estimation models to train JDC_{note}. The initial pseudo labels can be regarded as "the initial teacher" of JDC_{note} in the view of teacher-student framework. In particular, we used the basic teacher-student model which does not use the random audio augmentation [15]. In this experiment, we compare the basic teacher-student (TS) model to the noisy student (NS) model where we randomly augment the input audio as described in Section 2.2. Table 2 shows the performances of two teacher-student models. The result shows that The NS model achieves consistently higher accuracy in all metrics than the TS model, although the gap is somewhat small in MIR-ST500. This validates that the NS model is an effective approach in the teacher-student framework as concluded in [15].

4.3. Iteration of Self-training

We also investigated the number of self-training iterations in the NS model. We conducted up to 3 iterations of self-training for JDC_{note}. Figure 3 shows the results for the COnP accuracy on MIR-ST500 and Cmedia. The iteration number 0 indicates when the note-level

---

[1] `https://www.music-ir.org/mirex/wiki/2020:Singing_Transcription_from_Polyphonic_Music`

[2] The JDC_{pitch} model was trained with several datasets with pitch labels, which were not used in this paper.
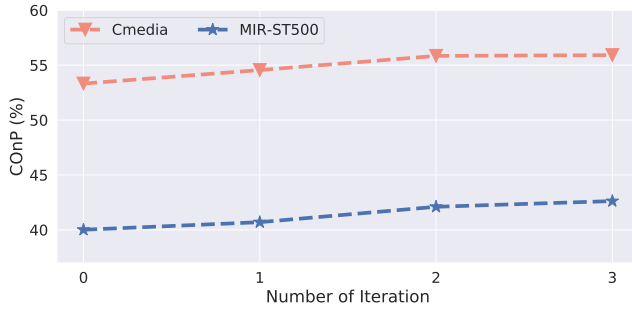
Fig. 3. Effect of iterative training in the noisy student model on Cmedia and MIR-ST500.

Description	
JDC_{note}(U)	Unsupervised model with unlabeled data $\mathcal{D}_{\mathcal{U}}$
JDC_{note}(L)	Supervised model with labeled data $\mathcal{D}_{\mathcal{L}}$
JDC_{note}(L+U)	Semi-supervised model with $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$

Table 3. Description of training setting

pseudo labels converted from estimated pitch contours are evaluated in Section 4.1. The iteration number 1 indicates when JDC_{note} is trained for the first time with the NS model in Section 4.2. In the next two iterations, the performance of JDC_{note} consistently increases although the slopes become saturated. We observed similar trends in the COn and COnPOff metrics. This result indicates that the iterative training using the NS model can reliably improve the model performance.

4.4. Comparison with Supervised and Semi-Supervised Models

While the proposed method in this paper aims to use only unlabeled audio data and achieve meaningful STP performance using pseudo labels from pretrained neural networks, we can also use labeled data together to further improve the performance. Table 3 describes three different versions of JDC_{note} trained with different sets of data. JDC_{note}(U) is the unsupervised model with 3 iterations of self-training from Section 4.3. JDC_{note}(L) is a solely supervised model trained with labeled audio data. JDC_{note}(L+U) is a semi-supervised model trained with both the unlabeled and labeled data. The right side of Table 4 shows the results. The supervised model achieves significantly higher accuracy than the unsupervised model on both test sets. However, the semi-supervised model outperforms the solely supervised model. This indicates that the proposed method has synergy with supervised learning.

5. COMPARISON WITH PREVIOUS WORKS

We finally compare our method to previous works in STP. The left side of Table 4 shows the accuracy metrics from recent works. HZ is a rule-based model submitted to MIREX2020 by Zhuang He and Yin Feng [25]. VOCANO is a semi-supervised STP model based on virtual adversarial training [3]. EFN is a model based on EfficientNet and trained with MIR-ST500 [4]. Note that VOCANO and EFN require singing voice separation as a pre-processing in the inference

Cmedia						
Model	**HZ**	**VOCANO**	**EFN**	JDC_{note}
				(U)	(L)	(L+U)
COnPOff	17.18	28.28	35.13	30.13	35.95	**40.20**
COnP	41.43	48.33	60.77	55.84	62.50	**66.11**
COn	63.63	64.56	**76.40**	65.72	73.88	75.97
MIR-ST500						
Model	**HZ**	**VOCANO**	**EFN**	JDC_{note}
				(U)	(L)	(L+U)
COnPOff	-	-	**45.78**	23.48	40.57	42.23
COnP	-	-	66.63	42.10	67.55	**69.74**
COn	-	-	75.44	58.61	74.94	**76.18**

Table 4. Comparison of singing transcription from polyphonic music. The best score in each row is highlighted in bold.

phase. The three previous works reported the results on Cmedia but the result on MIR-ST500 is available only for EFN because the dataset was released recently and EFN was introduced as a baseline. Compared to our proposed method, the unsupervised model, JDC_{note}(U), achieves high accuracy than HZ and VOCANO in all metrics on the Cmedia test set. This validates that the proposed method is superior to the semi-supervised method in VOCANO or the rule-based approach in HZ. However, JDC_{note}(U) has lower accuracy than EFN and the gap is larger on the MIR-ST500 test set because EFN was trained with the same MIR-ST500 dataset (but the training split). For fairness, we can compare EFN to JDC_{note}(L). The result shows that JDC_{note}(L) is better than EFN in COnPOff and CONP on Cmedia but the result is reverse on MIR-ST500. Given that JDC_{note}(L) was also trained with the same MIR-ST500 training set, the two models seem to be comparable to each other. However, JDC_{note}(U+L) pushes the accuracy levels higher, achieving best performances in COnPOff and CoOnP on Cmedia and in COnP and CON on MIR-ST500.

6. CONCLUSIONS

We presented a method for STP that uses pre-trained vocal pitch estimation models and unlabeled datasets. The method converts the frame-level pseudo labels to note-level and augments the label quality through self-training in the teacher-student framework. Through the ablation study, we showed that the model trained through the proposed method can achieve comparable results to the previous works, and with additional labeled data, it achieves better performance than the model trained with only labeled data. Since the test sets cover only Chinese music in this paper, we plan to evaluate the method on various genres of music in different cultural backgrounds as future work.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] Ryo Nishikimi, Eita Nakamura, Satoru Fukayama, Masataka Goto, and Kazuyoshi Yoshii, "Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism," in *Proc ICASSP*, 2019, pp. 161–165.

[2] Luwei Yang, Akira Maezawa, Jordan BL Smith, and Elaine Chew, "Probabilistic transcription of sung melody using a pitch dynamic model," in *Proc. ICASSP*, 2017, pp. 301–305.

[3] Jui-Yang Hsu and Li Su, "VOCANO: A note transcription framework for singing voice in polyphonic music," in *Proc. ISMIR*, 2021.

[4] Jun-You Wang and Jyh-Shing Roger Jang, "On the preparation and validation of a large-scale dataset of singing transcription," in *Proc. ICASSP*, 2021, pp. 276–280.

[5] Emilia Gómez and Jordi Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.

[6] Emilio Molina, Ana Maria Barbancho-Perez, Lorenzo Jose Tardon-Garcia, Isabel Barbancho-Perez, et al., "Evaluation framework for automatic singing transcription," in *Proc. ISMIR*, 2014.

[7] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters, "Dali: a large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm.," in *Proc. ISMIR*, 2018.

[8] Bilei Zhu, Fuzhang Wu, Ke Li, Yongjian Wu, Feiyue Huang, and Yunsheng Wu, "Fusing transcription results from polyphonic and monophonic audio for singing melody transcription in polyphonic music," in *Proc. ICASSP*, 2017, pp. 296–300.

[9] Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[10] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P Bello, "Deep salience representations for f0 estimation in polyphonic music," in *Proc. ISMIR*, 2017.

[11] Wei-Tsung Lu and Li Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *Proc. ISMIR*, 2018, pp. 521–528.

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

[13] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang, "A streamlined encoder/decoder architecture for melody extraction," in *Proc. ICASSP*, 2019, pp. 156–160.

[14] Sangeun Kum and Juhan Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, pp. 1324, 2019.

[15] Sangeun Kum, Jing-Hua Lin, Li Su, and Juhan Nam, "Semi-supervised learning using teacher-student models for vocal melody extraction," in *Proc. ISMIR*, 2020.

[16] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proc. SCIPY*, 2015, vol. 8, pp. 18–25.

[17] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. NeurIPS*, 2019, pp. 5050–5060.

[18] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proc. CVPR*, 2020, pp. 10687–10698.

[19] Ladislav Mošner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Björn Hoffmeister, "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *Proc. ICASSP*, 2019, pp. 6475–6479.

[20] Kangkang Lu, Chuan-Sheng Foo, Kah Kuan Teh, Huy Dat Tran, and Vijay Ramaseshan Chandrasekhar, "Semi-supervised audio classification with consistency-based regularization," *Proc. Interspeech*, pp. 3654–3658, 2019.

[21] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. Latent Variable Analysis and Signal Separation*. 2017, pp. 323–332, Springer.

[22] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A dataset for music analysis," in *Proc. ISMIR*, 2017.

[23] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "mir_eval: A transparent implementation of common mir metrics," in *Proc. ISMIR*, 2014.

[24] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[25] Zhuang He and Yin Feng, "Singing transcription from polyphonic music using melody contour filtering," *Applied Sciences*, vol. 11, no. 13, pp. 5913, 2021.