MCSSME: Multi-Task Contrastive Learning for Semi-supervised Singing Melody Extraction from Polyphonic Music

Shuai Yu

School of Computer Science and Technology,
Donghua University, China

Abstract

Singing melody extraction is an important task in the field of music information retrieval (MIR). The development of data-driven models for this task have achieved great successes. However, the existing models have two major limitations: firstly, most of the existing singing melody extraction models have formulated this task as a pixel-level prediction task. The lack of labeling data has limited the model for further improvements. Secondly, the generalization of the existing models are prone to be disturbed by the music genres. To address the issues mentioned above, in this paper, we propose a multi-task contrastive learning framework for semi-supervised singing melody extraction, termed as *MCSSME*. Specifically, ***to deal with data scarcity limitation***, we propose a self-consistency regularization (SCR) method to train the model on the unlabeled data. Transformations are applied to the raw signal of polyphonic music, which makes the network to improve its representation capability via recognizing the transformations. We further propose a novel multi-task learning (MTL) approach to jointly learn singing melody extraction and classification of transformed data. ***To deal with generalization limitation***, we also propose a contrastive embedding learning, which strengthens the intra-class compactness and inter-class separability. To improve the generalization on different music genres, we also propose a domain classification method to learn task-dependent features by mapping data from different music genres to shared subspace. MCSSME evaluates on a set of well-known public melody extraction datasets with promising performances. The experimental results demonstrate the effectiveness of the MCSSME framework for singing melody extraction from polyphonic music using very limited labeled data scenarios.

Introduction

Singing melody extraction is a challenging task in the field of MIR. It aims to extract the fundamental frequency contour from polyphonic music. Recently it has become an active research topic with a lot of downstream applications, such as cover song identification (Yu et al. 2019; Serra, Gómez, and Herrera 2010), query-by-humming (Wang and Jang 2015), singing voice separation (Ikemiya, Yoshii, and Itoyama 2015), and music recommendation (Knees and Schedl 2015). Singing melody contour obtained from extraction models can be utilized as an audio feature of musical information to enhance the performance of these downstream tasks.

With the trend of artificial intelligence, deep learning models play an important role in the development of singing melody extraction techniques. A number of deep learning based methods (Su 2018; Hsieh, Su, and Yang 2019; Yu et al. 2021a; Yu, Chen, and Li 2022) have been proposed for supervised singing melody extraction. Despite the remarkable successes, the existing models are facing two major limitations: firstly, most of existing singing melody extraction models have formulated this task as a pixel-level prediction task. Therefore, large amounts of pixel-level labeled data is needed by the supervised models. Obviously, the labeling process is time-consuming and laborious. The lack of labeling data has limited the model for further improvements. Secondly, it has been claimed by prior works (Su 2018; Yu et al. 2021b), music genres can largely affect the performance of singing melody extraction. For example, if one model is trained on a dataset with all of popular songs, and evaluated on a dataset with all of classic songs, the results will not be satisfied.

In an attempt to solve the problem of data insufficiency, semi-supervised learning melody extraction methods have become a cutting-edge direction. A pioneer work (Kum et al. 2020) adopts the teacher-student model to use unlabeled music tracks as training data for the singing melody extraction task, seeking to improve the melody extraction performance. However, this work did not include a data selection process, which many pseudo-labels are wrongly predicted and corrupts the whole training dataset.

The straight-forward solution is to directly apply more advanced semi-supervised models in the field of machine learning to this task. However, it is not practical due to two reasons: firstly, since singing melody extraction is a pixel-level prediction task (that predicts whether a pixel is on a melody contour or not), the neural network (NN) based model is very sensitive to the data disturbance. A small disturbance to the input spectrogram will result in a large decrease in results. We have directly applied *Mean Teachers* (Tarvainen and Valpola 2017), *MixMatch* (Berthelot et al. 2019) to this task, the performances are not ideal. Secondly, since the sensitive characteristic of this task, traditional consistency regularization cannot bring improvements. The

challenge is that the NN model is almost impossible to generate same predictions on transformed data and original data.

To deal with the data scarcity limitation, in this work, we propose a self-consistency regularization (SCR) method. To be specific, we first apply transformations to the raw signal of unlabeled musical audio. Then, we feed the transformed signal into the NN model, features of different levels are fed to prediction layers to generate predictions. In this way, we can justify the predictions from transformed data by themselves. Those predictions from different levels can be verified by checking them whether they are the same or not. Further, we introduce a pre-task that predicts which category of the data transformation belongs to. By introducing the pre-task, the model can extract not only task-specific features, but also general features from the transformations of the signal under a semi-supervised setting.

To deal with the generalization limitation caused by the music genres, we also propose a contrastive learning method to train the unlabeled data. In this process, the contrast learning strengthens the intra-class compactness and inter-class separability, which can typically address the issue of data generalization. Further, we propose to use a domain classifier that induces the model to learn task-dependent features by mapping data from different music genres to shared subspace.

The contribution of this paper is summarized as follow:

- A self-consistency regularization (SCR) method is proposed to address the data scarcity limitation. We use predictions from different levels of the model to justify the pseudo labels and select agreed data.
- A novel MTL approach is proposed to jointly learn two tasks: singing melody extraction and augmentation [1] category classification, alleviating the data scarcity limitation.
- To improve the generalization of our proposed model, we also propose a contrastive embedding learning to unlabeled data. To address the issue that music genres influence the performance of the singing melody extraction, we propose a domain-adversarial classifier that induces the model to learn task-dependent features by mapping data from different music genres to shared subspace.
- We use MIR-1K dataset and part of music tracks of the MedleyDB dataset as labeled data for training the model and we evaluate the performance on the well-known ADC 2004, MIREX 2005, iKala and another part of MedleyDB. The experimental results demonstrates the superiority of our method compared with other state-of-the-art ones.

Related Works

Singing Melody Extraction

Deep learning models for the singing melody extraction task undergone various model architectures throughout its history. (Kum, Oh, and Nam 2016) proposes a multi-column deep neural network to learn a nonlinear mapping between frame and melody. Subsequently, many convolutional neural network (CNN) based approaches have been developed to better capture spectral-temporal information (Lu, Su et al. 2018; Su 2018; Chen, Li, and Chi 2019). For example, (Chen, Li, and Chi 2019) proposes a two-stage CNN model to simulate the cochlea and mimic the auditory cortex in analyzing the spectral-temporal envelope. In addition, the use of musical prior knowledge and structural priors has further broadened the design of melody extraction models. For example, voiced and unvoiced frames in audio can be independently recognized (Hsieh, Su, and Yang 2019), or jointly detected with classification tasks (Kum and Nam 2019). The relationship between frequencies can be further captured through multi-dilation or attention networks (Gao, You, and Chi 2020; Yu et al. 2021a), or harmonic constant-Q transform (HCQT) (Bittner et al. 2017). The separate prediction of octave and pitch-class is proposed in (Chen et al. 2022) to further enhance the octave accuracy and chroma accuracy of the melody extraction. These models further improve the melody extraction performance.

Semi-supervised Learning

Although dealing with the lack of labeled data is a critical task in the era of deep learning, merely few studies (Kum et al. 2020; Yu et al. 2021b) have been conducted for the singing melody extraction task. As far as we know, Kum et al. (Kum et al. 2020) used a pretrained teacher model to generate pseudo labels on unlabeled data so that the model can train labeled data and unlabeled data. However, the model is likely to make wrong prediction on unlabeled data, which may lower the performance of this task. Actually, consistency regularization (Laine and Aila 2017; Tarvainen and Valpola 2017; Athiwaratkun et al. 2019; Xie et al. 2020) has become a main research direction in the field of semi-supervised learning. Therefore, it would be important if we investigate existing consistency regularization algorithms in the task of singing melody extraction and tailor a new consistency regularization method for semi-supervised singing melody extraction. Yu et al. (Yu et al. 2021b) proposed a few-shot learning algorithm to address the imbalance distribution of the samples due to the scarce of labeling data. Unfortunately, this algorithm can not utilize the large-scale unlabeled music datasets and can not be used in the scenario of semi-supervised learning. In this paper, we propose a self-consistency regularization method to perform singing melody extraction using a large-scale unlabeled music dataset.

Self-supervised Contrastive Learning

Contrastive learning currently is popular in many artificial intelligence based applications. A number of self-supervised contrastive learning models have achieved remarkable successes in the field of computer vision, such as SimCLR (Chen et al. 2020), MoCo (He et al. 2020), BYOL (Grill et al. 2020), Siamese (Chen and He 2021). In the field of music information retrieval and audio-based applications, a number of pioneer works (Zhao et al. 2022; Zhu et al.

[1] In this work, the terms transformation and augmentation have the same meaning. We use the two terms alternatively throughout the paper.

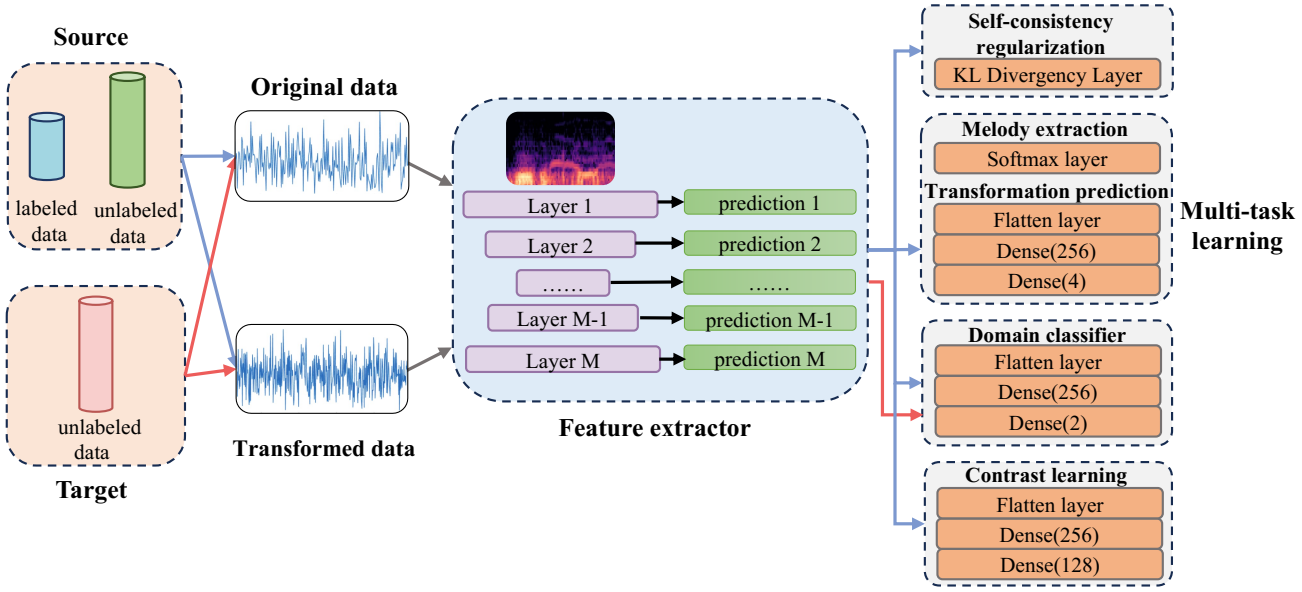


Figure 1: The framework of the proposed MCSSME. The blue and red arrows denote data flows of the source and target domains, respectively. The black arrow denotes the shared data streams. The feature extractor is pyramid-shaped, which has M layers. Each layer is forced to make predictions of melody extraction for thereafter self-consistency regularization. We define the outputs of the M-th layer as the final prediction.

2021; Choi et al. 2022; Li et al. 2023) have applied self-supervised contrastive learning to the specific tasks. Zhao et al. (Zhao et al. 2022) proposed to combine Moco and Swin Transformer to learn music representation for music genre classification. Choi et al. (Choi et al. 2022) proposed a self-supervised method to extract music representation based on Siamese for general MIR purpose. Li et al. (Li et al. 2023) propose a multi-task contrastive learning method on sleep stage prediction. Despite the successes these works, they are not intended for pixel-level tasks in MIR. For singing melody extraction task, currently there is no research works using self-supervised learning to obtain more accurate results. In this work, to address the generalization limitation, we propose a contrastive embedding learning to unlabeled data. In this process, the embedding strengthens the intra-class compactness and inter-class separability, which can typically address the issue of data generalization.

Methodology

The overview of the proposed framework is presented in Fig.1. Inputs are from two domains: popular songs and classic songs. Transformations are applied to the raw waveform of music signal. We choose to use MSNet (Hsieh, Su, and Yang 2019) as the feature extractor of MCSSME. A multi-task module is designed to predict transformed labels and singing melody extraction. At the same time, we perform self-consistency regularization, contrast learning and domain classification between the two groups of the generated music representations. We will introduce each component in the following subsections.

Semi-supervised Learning Setup

In this paper, raw waveform music signals are from both labeled and unlabeled data. For the source domain data, the music signals are denoted as $D^s = \{D_l^s, D_u^s\}$. $D_l^s = \{(x_1, y_1), (x_2, y_2), ..., (x_M, y_M)\}$ and $D_u^s = \{u_1, u_2, ..., u_N\}$ denotes labeled music data and unlabeled music data, respectively. M and N are the number of labeled data and unlabeled data. T denotes the total number of whole source dataset, and $M + N = T$[2]. The learning objective function is constructed in the following form:

$$\min_{\theta}\{L_l^s(D_l^s, \theta) + \lambda L_u^s(D_u^s, \theta)\}, \quad (1)$$

where L_l^s is the loss of supervised learning and L_u^s is the loss of unsupervised learning, respectively. λ is a non-negative parameter controlling the weight between L_l and L_u. θ represents the parameters of our proposed framework.

Self-consistency Regularization for Singing Melody Extraction

The aim of self-consistency regularization (SCR) is to better utilize the pseudo labels generated from unlabeled data and alleviate the data scarcity. Since lots of popular singing melody extraction model are pyramid-shaped, such as MSNet (Hsieh, Su, and Yang 2019), FTANet (Yu et al. 2021a). Inspired by these pyramid-shaped models, we design a self-consistency regularization method as shown in Fig. 2.

[2]Please note that D^s, D_l^s, D_u^s, M, N and T are only defined on the source domain data.

We perform self-consistency regularization on both labeled and unlabeled data, the augmented data (i.e., transformed data) is fed into the feature extractor to extract task-specific features. We then fuse the feature maps from different levels to generate predictions within a feature fusion module (FFM). The feature fusion module accepts a feature map from low-resolution outputs and a feature map from current stage. Then the feature map from low-resolution performs upsampling to have the same shape with the feature map from current stage. Finally, a element-wise addition is performed to fuse the two feature maps. This process can be described:

$$FFM(F_{low}, F_{curr}) = F_{curr} \oplus \alpha \cdot up(F_{low}), \quad (2)$$

where F_{low} and F_{curr} denote the feature maps from low-level outputs and current stage, respectively. $\oplus$ denotes the element-wise addition, up$(\cdot)$ denotes the process of upsampling and α is weight parameter controlling how much low-level information is fused. Then the outputs from FFM are fed to the prediction layer to obtain predictions.

After obtaining the predictions from different levels, KL divergence is employed to calculate the difference between predictions from FFM and final prediction. The loss function can be formulated:

$$L_{scr} = \frac{1}{|D_u|}\frac{1}{|st|}\sum_{i \in D_u}\sum_{j \in st} KL(F_{final}, F_{stage_{ij}}), \quad (3)$$

where $st \in \{1, 2, 3, 4\}$. We also proposed to use L_{scr} as a threshold to select unlabeled data adding them into the training set.

Signal Transformation Classification

In this work, we propose to use signal transformation on the unlabeled raw music signals, the transformed raw music signals are then used to improve the representation capability of the network via recognizing the transformations. To this end, we perform four signal transformation on the raw music signals: Noising, Filtering, Rotation, Shuffle.

Noising: We add white noise onto the raw music signals s_i. The generated signals can be denoted: s'_n.

Filtering: Each of music signal s_i is fed into the Savitzky-Golay Filter to reduce noise, which generates transformed signal s'_f.

Rotation: The raw music signal s_i is rotated within a fixed interval. Let $s_i = \{s_{i_1}, s_{i_2}, ..., s_{i_t}\}$, where $t \in \{1, 2, 3, ..., R\}$. R denotes the number of samples in s_i. Then we perform rotation on s_i, which generates $s'_i = \{s_{i_n}, s_{i_n+1}, ..., s_{i_1}, s_{i_2}, ..., s_{i_n-1}\}$. n is chosen randomly from $(1, R)$. Finally, we can obtain the rotated raw music signals s'_r.

Shuffle: We randomly shuffle the samples in the signal s_i, and the generated shuffled music signals can be denoted s'_s. Note that we can get the shuffled order number, which can be used to recover the shuffled prediction of melody contour.

We perform the above transformations on both labeled and unlabeled data to obtain transformed music signals $S_T = \{s'_n, s'_f, s'_r, s'_s\}$. Such transformations do not change the dimension of music signals. We also record the labels

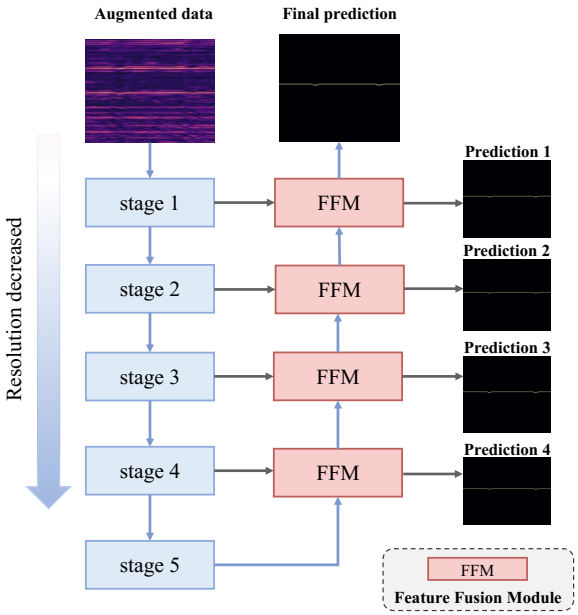


Figure 2: Detailed architecture of the proposed self-consistency regularization. The blue arrows denote the upsample or downsample process.

of the signal transformation in order to identify the transformation of tasks. The feature extractor would be required to understand the latent structure of the signal for recognizing the four transformations. Therefore, multi-task learning is applied to attach transformation classification task to our proposed singing melody extraction framework in order to not only learn task-specific features, but also the general features from the transformations of music signal. The feature extractor is trained to recognize the four transformation tasks by the following loss function:

$$L_{tr} = \frac{1}{|C|}\sum_{c \in C}\sum_{s_i \in D} L_{CE}(G_\theta(ST(s_i, c)), c), \quad (4)$$

where C stands for the set of the above mentioned four kinds of transformations. $ST(s_i, c)$ denotes performing c signal transformation on s_i, and G_θ is the neural network with parameter θ. The loss L_{ce} denotes the standard cross entropy loss.

Domain Classification

In this work, we use popular music tracks as the source domain data and classic music tracks as the target domain data. The target domain data does not attend other components of MCSSME, such as SCR, melody extraction, transformation prediction and contrastive learning. To improve the robustness of MCSSME, we also perform transformation on the target domain data, and then the data is passed through the shared feature extractor. Finally, the learned feature map is fed into a three-layer domain classifier to perform domain classification. Let W denotes the raw signals data in both domains and their transformed data. The loss function is cal-

culated:

$$L_{dc} = \frac{-1}{|W|} \sum_{s_i \in W} L_{CE}(DC(s_i), y_{d_{si}}), \quad (5)$$

where $y_{d_{si}}$ denotes the domain label of the example s_i, $DC(.)$ denotes our proposed domain classification module.

Contrastive Learning for Singing Melody Extraction

The raw music signals $S = \{s_1, s_2, ..., s_N\}$ are transformed by the signal transformation module mentioned above to generate T signal pairs $S_T = \{(s'_{1_r}, s'_{1_n}, s'_{1_f}, s'_{1_s}), ..., (s'_{T_r}, s'_{T_n}, s'_{T_f}, s'_{T_s})\}$. We first perform STFT on the raw signals and then we use MSNet (Hsieh, Su, and Yang 2019) as feature extractor to extract feature embeddings h and $h_t = \{h_r, h_n, h_f, h_s\}$. We propose to use these pairs to perform contrastive learning for better music representation. For each h_i, we calculate the similarity between h_i and $h' \in h_t$, the transformed features are considered as positive pairs and others as negative pairs. Similar to SimCLR (Chen et al. 2020), we choose to use cosine similarity to calculate the similarity between h_i and h':

$$sim(h_i, h') = cos(h_i, h') = \frac{h_i^T h'}{|h_i||h'|}, \quad (6)$$

For labeled data D_l^s and unlabeled data D_u^s, we consider two different strategies. For unlabeled data D_u^s, we choose the raw signal and the corresponding transformed data as positive pairs. The contrastive learning on unlabeled data can be described:

$$L_{cl} = -log \frac{exp(sim(h_i, h'_{i_t})/\tau)}{\sum_{k \neq i} exp(sim(h_i, h'_{k_t})/\tau)} \quad (7)$$

where τ is a temperature parameter, we set τ to 0.5 as default.

For labeled data D_l, inspired by SupCon (Khosla et al. 2020) we choose not only the corresponding transformed data as positive pairs, but also the other raw signals belong to the same class, the loss function is as follow:

$$L_{cl} = -log \frac{\sum_{j \in \gamma_i} exp(sim(h_i, h_j)/\tau)}{\sum_{k \in \gamma \setminus \gamma_i} exp(sim(h_i, h_k)/\tau)} \quad (8)$$

where γ_i denotes the set that contains the embeddings from transformed data of s_i and other embeddings from signals has the same label with s_i, γ contains all labeled data and its transformations, and $\setminus$ denotes the remove operation of the set.

Experiments

Datasets

We train and evaluate our proposed MCSSME framework on several public datasets, the descriptions of the datasets we used are listed in Table 1. For the training data, we first choose 1000 popular music tracks from MIR-1K (Hsu and Jang 2010) and 30 popular music tracks from MedleyDB (Bittner et al. 2014) with melody annotated. Then we also choose 2000 popular music tracks from FMA dataset (Defferrard et al. 2017) without labels. To train the proposed domain classifier, we also introduce 20 classic music tracks from MedleyDB dataset and 24 classic music tracks from RWC dataset (Goto et al. 2002). For the testing data, we use four well-known testing datasets for this task: 12 tracks from ADC2004, 9 tracks from MIREX05[3], 12 tracks from MedleyDB and 262 tracks from iKala (Chan et al. 2015).

	Dataset	Number	Dur.	S/T
Training (Labeled)	MIR-1K	1000	2h13m	S
	MedleyDB	30	1h58m	S
Training (Unlabeled)	FMA	2000	15h	S
Training (Unlabeled)	MedleyDB	20	1h20m	T
	RWC	24	2h58m	T
Testing	ADC2004	12	4min	-
	MIREX 05	9	4min	-
	Medley DB	12	48min	-
	iKala	262	2h6m	-

Table 1: The detailed descriptions of the datasets for training and testing the proposed framework MCSSME. "S/T" denotes the the dataset belongs to source domain or target domain.

Experiment Setup

Following the convention in the literature (Salamon et al. 2014), we use the following metrics for performance evaluation: overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), voicing recall (VR) and voicing false alarm (VFA). We use mir eval library [19] with the default setting to calculate the metrics. For each metric other than VFA, the higher score, the higher performance. In the literature, OA is often considered more important than other metrics.

The proposed framework is implemented using PyTorch [4]. All experiments are conducted on a machine with two NVIDIA RTX 3090 GPUs. For a fair comparison, we train the baseline models using the same training data. Following (Hsieh, Su, and Yang 2019), we choose to use a set of input representations. It contains three parts: (1) the generalized cepstrum(GC) (Kobayashi and Imai 1984), (2) the generalized cepstrum of spectrum (GCoS) (Su 2017), (3) the Combined Frequency and Periodicity (CFP) spectrum (Su 2018). In this work, the audio files are resampled to 8 kHz and merged into one mono channel following (Yu et al. 2021a). Data representations are computed with a Hanning window of 768 samples and hop size of 80 samples. To adapt the pitch ranges required in singing melody extraction, following (Hsieh, Su, and Yang 2019), we set hyper-parameters in computing the CFP for our model. For vocal melody extraction, the number of frequency bins is set to 320, with 60 bins per octave, and the frequency range is from 31 Hz (B0)

[3] https://labrosa.ee.columbia.edu/projects/melody

[4] https://pytorch.org

Dataset Methods	ADC2004					MIREX 05				
	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
DSM	62.8	61.4	64.8	70.2	19.2	68.4	70.2	72.4	80.7	34.2
MSNet	70.1	71.3	73.2	75.6	21.3	81.7	76.7	76.9	83.6	18.6
MD+MR	71.2	72.4	73.8	73.1	24.7	80.8	77.2	77.8	81.3	24.8
Teacher-student	73.1	72.8	74.8	77.6	**16.8**	82.1	78.3	79.2	82.4	19.2
FTANet	72.4	73.8	75.2	77.3	24.9	84.4	79.7	80.0	83.8	**5.1**
HGNet	71.3	72.8	73.1	74.9	23.7	80.1	77.4	78.3	80.5	21.7
MCSSME (ours)	**76.7**	**78.1**	**78.9**	**80.8**	20.3	**85.6**	**83.8**	**84.2**	**87.3**	13.7

Table 2: The performances of the proposed MCSSME and baseline methods on the ADC2004 and MIREX 05 datasets, the values in the table are percentile.

Dataset Methods	MedleyDB					iKala				
	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
DSM	64.2	50.3	51.7	61.2	44.2	68.1	74.3	74.8	77.2	21.9
MSNet	66.9	47.2	48.4	53.2	**12.6**	77.6	79.7	80.4	80.8	12.7
MD+MR	67.1	48.6	49.9	53.8	21.3	78.0	80.2	81.3	81.4	29.3
Teacher-student	68.1	49.0	49.6	58.3	29.7	76.7	76.4	78.2	79.1	36.9
FTANet	68.8	50.2	51.4	63.2	27.3	80.3	82.4	84.0	84.7	25.6
HGNet	65.3	45.4	46.2	51.7	24.9	78.7	80.0	80.6	81.5	24.9
MCSSME (ours)	**73.4**	**56.8**	**57.4**	**64.2**	16.2	**84.7**	**85.8**	**86.2**	**88.3**	**9.2**

Table 3: The performances of the proposed MCSSME and baseline methods on the MedleyDB and iKala datasets, the values in the table are percentile.

to 1250 Hz (D#6).

Experimental Results

Comparison with State-of-the-art Methods We compare our framework with six state-of-the-art (SOTA) methods for singing melody extraction: (1)DSM (Bittner et al. 2017), (2) MSNet (Hsieh, Su, and Yang 2019), (3) MD+MR (Gao, You, and Chi 2020), (4) Teacher-student (Kum et al. 2020) (5) FTANet (Yu et al. 2021a), (6) HGNet (Yu, Chen, and Li 2022). To demonstrate the effectiveness of our proposed method, we train the proposed framework MCSSME and compare our method with other baseline methods. The quantitative results are shown in Table 2 and Table 3. It is observed that with assisted unlabeled music data, our proposed MCSSME achieves the best performance on four public testing sets in general. For comparison with other baselines, when focusing on OA, the proposed method outperforms FTANet by 5.9% in ADC2004, by 1.4% in MIREX 05, by 6.7% in Medley DB and by 5.5% in iKala, relatively. When comparing with semi-supervised methods on OA, the proposed method outperforms Teacher-student by 4.9% in ADC2004, by 4.3% in MIREX 05, by 7.8% in Medley DB and by 10.4% in iKala, relatively. It is worthy to mention that the effectiveness of our method is from the self-consistency regularization and contrast learning compared with the performance of a semi-supervised model Teacher-student (Kum et al. 2020).

Qualitative Analysis To investigate the quality of music representation learned from our proposed MCSSME, we visualize the learned representation via t-SNE. We use 12 popular music tracks to perform t-SNE, as observed in Fig. 3, the left is the distribution of the raw signals. The right is the distribution of learned music representations. representations are well clustered. Owing to the proposed MCSSME framework, the predictions of our method have more smooth contours and the examples with the same frequency are closer to each other.

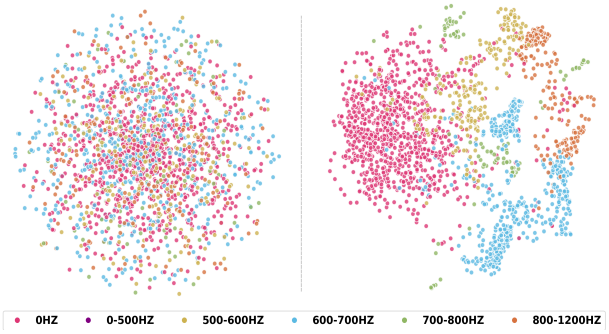


Figure 3: Visualization of the learned music representation via t-SNE. The left is the distribution of the raw signals. The right is the distribution of learned music representations. Different colors denote various examples with different frequencies.

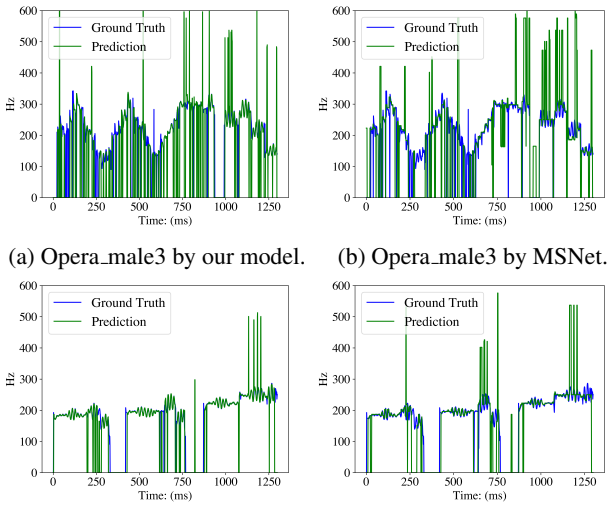
Ablation Study To investigate the effectiveness of the key components in our framework, we conduct ablation studies

Dataset Methods	ADC2004			MIREX 05		
	OA	RPA	RCA	OA	RPA	RCA
w/o DC	73.1	74.6	75.2	85.3	83.4	84.0
w/o SCR	74.0	74.8	75.6	81.9	80.8	81.5
w/o CL	74.7	76.3	77.1	83.2	81.6	82.3
w/o TR	75.3	77.4	77.9	84.0	82.1	82.9
MCSSME	**76.7**	**78.1**	**78.9**	**85.6**	**83.8**	**84.2**

Table 4: Results of Ablation Study on ADC2004 and MIREX 05 dataset. The values in the table are percentile. "w/o DC" and "w/o SCR" denote without domain classifier and self-consistency regularization respectively. "w/o CL" stands for without contrastive learning. "w/o TR" stands for without signal transformation.

and the quantitative results are presented in Table 4. We first remove the domain classifier and direct train the source domain data. As observed in Table 4, the performances of OA decreased by 4.7% in ADC2004 and 0.3% in MIREX 05. There is no surprise that MIREX 05 decreased less than in ADC2004, because the music tracks in MIREX 05 are popular music. We then remove the SCR module, the performances of OA decreased by 3.5% in ADC2004 and 4.3% in MIREX 05. The observation indicates that the use of self-consistency regularization helps improve the performance of singing melody extraction. Next, we remove the contrast learning module, the performances of OA decreased by 2.6% in ADC2004 and 2.8% in MIREX 05. Finally, we remove the signal transformation, the performances of OA decreased by 1.8% in ADC2004 and 1.9% in MIREX 05. The results show that the proposed multi-task learning can indeed help alleviate the limitation of data scarcity. Overall, the key components of our framework MCSSME are tightly incorporated and collaboratively devote to remarkable results.

Case Study To investigate what types of errors are solved by the proposed model, a case study is performed on two opera songs: "opera male3.wav" and "opera male5.wav" in the ADC2004 dataset. We choose MSNet (Hsieh, Su, and Yang 2019) to compare with due to its effectiveness and popularity. As depicted in Fig. 4, we can observe that there are fewer octave errors in diagram (a) and (c) than in diagram (b) and (d). Moreover, from 1000-1200 ms in the diagram (d), we can find some errors that predict the wrong frequency bin near the right one, which are correctly predicted in diagram (c). Through the visualization of the predicted melody contour, we can say that the performance gains of the proposed model can be attributed to solving the octave errors and other errors. However, we can also observe that there seem to be more the melody detection errors (i.e., predicting a non-melody frame as a melody one) than in MSNet (Hsieh, Su, and Yang 2019). We analyzed our proposed model, we found the reason lies in the signal transformation, it will decrease the number of non-melody frames, we leave this as a future research topic.



(a) Opera_male3 by our model. (b) Opera_male3 by MSNet.

(c) Opera_male5 by our model. (d) Opera_male5 by MSNet.

Figure 4: Visualization of singing melody extraction results on two opera songs using different models.

Conclusion

In this paper, we propose a multi-task contrastive learning framework for semi-supervised singing melody extraction, MCSSME. Specifically, to deal with data scarcity limitation, we propose a self-consistency regularization (SCR) method to select right pseudo labels on the unlabeled data. Transformations are applied to the raw signal of polyphonic music, which makes the network to improve its representation capability via recognizing the transformations. We further propose a novel multi-task learning (MTL) approach to jointly learn singing melody extraction and classification of transformed data. To deal with generalization limitation, we also propose a contrastive embedding learning, which strengthens the intra-class compactness and inter-class separability. To improve the generalization on different music genres, we also propose a domain classification method to learn task-dependent features by mapping data from different music genres to shared subspace. MCSSME evaluates on a set of well-known public melody extraction datasets with promising performances. The experimental results demonstrate the effectiveness of the MCSSME framework for singing melody extraction from polyphonic music using very limited labeled data scenarios. This work has provided another verification of the feasibility for integrating contrastive learning strategy with semi-supervised framework to improve the learning ability of semi-supervised models.

References

Athiwaratkun, B.; Finzi, M.; Izmailov, P.; and Wilson, A. G. 2019. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *Proc. ICLR*.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proc. NeurIPS*, 5050–5060.

Bittner, R. M.; McFee, B.; Salamon, J.; Li, P.; and Bello, J. P. 2017. Deep Salience Representations for F0 Estimation in Polyphonic Music. In *Proc. ISMIR*, 63–70.

Bittner, R. M.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; and Bello, J. P. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proc. ISMIR*, 155–160.

Chan, T.; Yeh, T.; Fan, Z.; Chen, H.; Su, L.; Yang, Y.; and Jang, J. R. 2015. Vocal activity informed singing voice separation with the iKala dataset. In *Proc. ICASSP*, 718–722.

Chen, K.; Yu, S.; Wang, C.; Li, W.; Berg-Kirkpatrick, T.; and Dubnov, S. 2022. Tonet: Tone-Octave Network for Singing Melody Extraction from Polyphonic Music. In *Proc. ICASSP*, 621–625.

Chen, M.-T.; Li, B.-J.; and Chi, T.-S. 2019. CNN Based Two-stage Multi-resolution End-to-end Model for Singing Melody Extraction. In *Proc. ICASSP*, 1005–1009.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proc. ICML*, volume 119, 1597–1607.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proc. CVPR*, 15750–15758.

Choi, J.; Jang, S.; Cho, H.; and Chung, S. 2022. Towards Proper Contrastive Self-Supervised Learning Strategies for Music Audio Representation. In *Proc. ICME*, 1–6.

Defferrard, M.; Benzi, K.; Vandergheynst, P.; and Bresson, X. 2017. FMA: A Dataset for Music Analysis. In *Proc. ISMIR*, 316–323.

Gao, P.; You, C.-Y.; and Chi, T.-S. 2020. A Multi-Dilation and Multi-Resolution Fully Convolutional Network for Singing Melody Extraction. In *Proc. ICASSP*, 551–555.

Goto, M.; Hashiguchi, H.; Nishimura, T.; and Oka, R. 2002. RWC Music Database: Popular, Classical and Jazz Music Databases. In *ISMIR*, volume 2, 287–288.

Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. Á.; Guo, Z.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Proc. NeurIPS*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. CVPR*, 9726–9735.

Hsieh, T.-H.; Su, L.; and Yang, Y.-H. 2019. A streamlined encoder/decoder architecture for melody extraction. In *Proc. ICASSP*, 156–160.

Hsu, C.; and Jang, J. R. 2010. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Trans. Speech Audio Process.*, 18(2): 310–319.

Ikemiya, Y.; Yoshii, K.; and Itoyama, K. 2015. Singing voice analysis and editing based on mutually dependent F0 estimation and source separation. In *Proc. ICASSP*, 574–578.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Knees, P.; and Schedl, M. 2015. Music retrieval and recommendation: A tutorial overview. In *Proc. SIGIR*, 1133–1136.

Kobayashi, T.; and Imai, S. 1984. Spectral analysis using generalised cepstrum. *TASLP*, 32(6): 1235–1238.

Kum, S.; Lin, J.; Su, L.; and Nam, J. 2020. Semi-supervised learning using teacher-student models for vocal melody extraction. In *Proc. ISMIR*, 93–100.

Kum, S.; and Nam, J. 2019. Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks. *Applied Sciences*, 9(7).

Kum, S.; Oh, C.; and Nam, J. 2016. Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks. In *Proc. ISMIR*, 819–825.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *Proc. ICLR*.

Li, Y.; Luo, S.; Zhang, H.; Zhang, Y.; Zhang, Y.; and Lo, B. 2023. MtCLSS: Multi-Task Contrastive Learning for Semi-Supervised Pediatric Sleep Staging. *IEEE J. Biomed. Health Informatics*, 27(6): 2647–2655.

Lu, W. T.; Su, L.; et al. 2018. Vocal Melody Extraction with Semantic Segmentation and Audio-symbolic Domain Transfer Learning. In *Proc. ISMIR*, 521–528.

Salamon, J.; Gómez, E.; Ellis, D. P.; and Richard, G. 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2): 118–134.

Serra, J.; Gómez, E.; and Herrera, P. 2010. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Proc. Advances in Music Information Retrieval*, 307–332. Springer.

Su, L. 2017. Between homomorphic signal processing and deep neural networks: Constructing deep algorithms for polyphonic music transcription. In *Proc. APSIPA ASC*, 884–891.

Su, L. 2018. Vocal melody extraction using patch-based CNN. In *Proc. ICASSP*, 371–375.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 1195–1204.

Wang, C.; and Jang, J. R. 2015. Improving Query-by-Singing/Humming by Combining Melody and Lyric Information. *IEEE/ACM Trans. Audio Speech Language Processing*, 23(4): 798–806.

Xie, Q.; Dai, Z.; Hovy, E. H.; Luong, T.; and Le, Q. 2020. Unsupervised Data Augmentation for Consistency Training. In *Proc. NeurIPS*, 6256–6268.

Yu, S.; Chen, X.; and Li, W. 2022. Hierarchical Graph-Based Neural Network for Singing Melody Extraction. In *Proc. ICASSP*, 626–630.

Yu, S.; Sun, X.; Yu, Y.; and Li, W. 2021a. Frequency-Temporal Attention Network for Singing Melody Extraction. In *Proc. ICASSP*, 251–255.

Yu, S.; Yu, Y.; Chen, X.; and Li, W. 2021b. HANME: Hierarchical Attention Network for Singing Melody Extraction. *IEEE Signal Process. Lett.*, 28: 1006–1010.

Yu, Z.; Xu, X.; Chen, X.; and Yang, D. 2019. Temporal Pyramid Pooling Convolutional Neural Network for Cover Song Identification. In *Proc. IJCAI*, 4846–4852.

Zhao, H.; Zhang, C.; Zhu, B.; Ma, Z.; and Zhang, K. 2022. S3T: Self-Supervised Pre-Training with Swin Transformer For Music Classification. In *Proc. ICASSP*, 606–610.

Zhu, H.; Niu, Y.; Fu, D.; and Wang, H. 2021. MusicBERT: A Self-supervised Learning of Music Representation. In *Proc. ACM MM*, 3955–3963.