INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

MTANet: Multi-band Time-frequency Attention Network for Singing Melody Extraction from Polyphonic Music

Yuan Gao[1,3], *Ying Hu*[1,3], *Liusong Wang*[1,3], *Hao Huang*[1], *Liang He*[1,2]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing, China
[3] Key Laboratory of Signal Detection and Processing, Xinjiang, Urumqi, China
gya@stu.xju.edu.cn, huying@xju.edu.cn

Abstract

Singing melody extraction is an important task in music information retrieval. In this paper, we propose a multi-band time-frequency attention network (MTANet) for singing melody extraction from polyphonic music, which can generate the feature representation to characterize the fundamental frequency (F0) component. Moreover, a band partition scheme is proposed to fit the position distribution of the F0 component. Further, three hourglass sub-networks are used to capture various multi-band features. Then, a feature fusion module (FFM) is employed to fuse the multi-band features. Visualization analysis shows that the multi-band feature extraction branch can generate the feature representation for characterizing the F0 component effectively. Experimental results show that the MTANet outperforms the existing state-of-the-art methods, while keeping with fewer network parameters. Visualized results intuitively show that the MTANet can reduce the octave and melody detection errors.

Index Terms: Melody extraction, Multi-band, Polyphonic music, Band partition, Music information retrieval

1. Introduction

Singing melody extraction is a challenging task in the field of music information retrieval (MIR) which aims to produce a sequence of frequency values corresponding to the pitch of the singing melody from polyphonic music [1, 2, 3]. The instrumental accompaniments and noises are interwoven with the leading vocal, making the task challenging. Specifically, the accompaniment element like chord progression will naturally contain the leading voice F0 or its harmony, which makes it not toilless to obtain the semantic representation that can effectively distinguish the main melody from the background music. Melody extraction has many downstream applications, such as music transcription [4], query-by-humming [5], and singing voice separation [6].

The basic idea of many deep learning-based methods is to learn a mapping between a matrix representing input audio and another matrix representing the predicted melody line [7, 8]. To learn this mapping adequately, some existing methods utilized effective audio representations and model structures skillfully. For example, Bittner et al. constructed a new input representation named harmonic constant-Q transform (HCQT) by virtue of the harmonic relation and fed it into convolutional neural networks (CNNs) to learn salience representations [9]. Yu et al. proposed a frequency-temporal attention network to mimic human auditory assigning different weights in the time and frequency axis [8]. Hsieh et al. proposed a streamlined encoder-decoder network using a bottleneck layer to estimate the existence of a melody for each time frame [7]. However, the fixed convolutional structure may be along with the locally constrained receptive field, which will be prone to cause octave errors (the prediction overtop the actual pitch by several octaves) [10] and fail to capture long-dependency harmonic relationship [11]. For this reason, several researchers have also tried many unconventional convolutional or non-convolutional methods, such as dilated convolution [12, 13] and self-attention mechanism [14]. Chen et al. proposed a Tone-Combined Frequency and Periodicity (Tone-CFP) representation, which rearranges the tonal harmonics into adjacent bins to capture harmonic relationships and predict the melody line by leveraging self-attention modules [15]. Although the self-attention module brings good performance, it also brings high computing costs.

The highly resonant voices produced by singers are prone to cause that the higher harmonics have larger amplitudes than the F0, which is the main acoustic cause of octave errors caused by algorithmic misjudgments [3]. Therefore, it is vital to obtain a semantic representation that can distinguish the F0 and non-F0 components (e.g., accompaniment or higher harmonics of singers). Meanwhile, we note that the F0 component tends to be distributed in the spectrogram within a certain range, which motivates us to utilize the positional properties in the spectrogram to characterize the F0 and non-F0 components. Given the above considerations, we proposed the band partition scheme and used hourglass sub-network [16] to capture multi-band features that could characterize F0 and non-F0 components discriminately. Through observing the channels in different layers of our model, we found that there is a certain harmonic relationship inside feature. To make full use of the time-frequency relationship within each channel and perform the aggregation of multi-band features across channels, we proposed a feature fusion module that pays more attention to internal information of each channel. The contributions of this paper as follows:

i) We design a multi-band feature extraction branch in MTANet. Visualization analysis shows it can effectively generate the feature representation for characterizing the F0 component.

ii) We design a feature fusion module (FFM) to fuse multi-band features to obtain more salient features characterizing the melody line effectively.

iii) Our proposed MTANet achieves 86.9% of OA on ADC2004 dataset, 89.2% of OA on MIREX05 dataset and 74.6% of OA on MEDLEY DB dataset, outperforming other state-of-the-art CFP-based methods.

2. Proposed Method

Fig.1 is the overall architecture of the MTANet that has two branches. In the top branch, the inputs are firstly partitioned into the sub-bands features to fit the position distribution of the F0 and non-F0 components initially. Further, they are fed into three hourglass sub-networks and concatenated to generate the fea-

10.21437/Interspeech.2023-2494

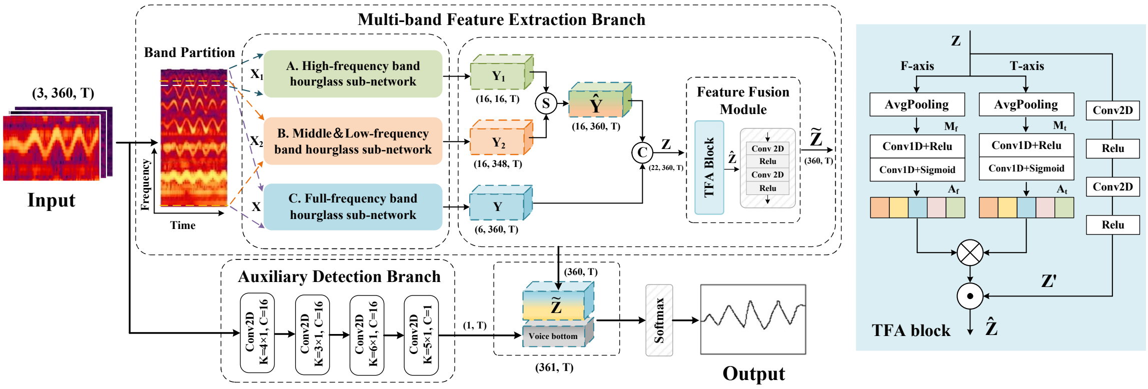


Figure 1: *The overall architecture of multi-band time-frequency attention network (MTANet). The top branch contains three hourglass sub-networks and a feature fusion module (FFM) embedded with the TFA block and conv-2D unit. The bottom branch is an auxiliary detection branch (ADB). The Ⓢ denotes Con in the formula (1), the Ⓒ denotes the concatenation along the channel dimension.*

ture representation that can distinguish both components. Then, these features are fused by the feature fusion module (FFM) embedded with the time-frequency attention (TFA) block. The bottom branch named auxiliary detection branch (ADB) contains stacked convolution layers with consecutive downsampling operations, and outputs the feature with the size of 1 × T guiding the presence of melody. Finally, the outputs of two branches are concatenated along frequency dimension and then undergo a softmax operation to estimate the melody.

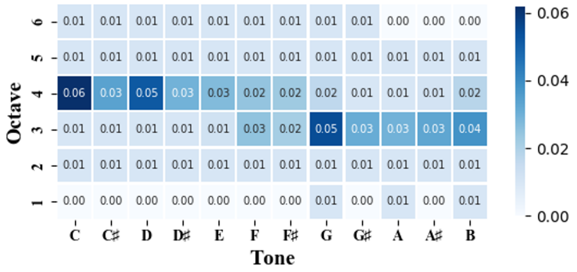
2.1. Band Partition of CFP Representation

We also choose CFP representation as the inputs because of its effectiveness [17]. The CFP representation, $X \in \mathbb{R}^{3\times F\times T}$, contains three parts: a power-scaled spectrogram, a generalized cepstrum [18] and a generalized cepstrum of spectrum [19]. F denotes the number of frequency bins of CFP and is set with 360 here, T denotes the time frame. The input features are firstly split into two sub-bands to be processed discriminatively. To partition the feature band more reasonably for processing features discriminately, we analyze the statistics distribution of F0 of the samples in the training data. As shown in Fig.2, we find that the F0 of the main melody is mainly distributed within the third and fourth octaves for the frequency range of our selected samples from 32 Hz (C1) to 2050Hz (B6). When the frequency exceeds 1760Hz (348th bins), only the background music and the higher harmonic components of the melody exist. Considering the complexity of the locating the F0 in the actual scene, we only preliminary partition to guide the network to learn the positional distribution of the two components.

Thus, we partition X along the frequency axis into two subbands features $X_1 = X[:, 345 : 360, :]$, $X_2 = X[:, 1 : 348, :]$. Meanwhile, we set an overlap of 4 bins for X_1 and X_2 to ensure the harmonic continuity of the features. To offset the information loss caused by partition, we set the full band hourglass network whose input is X as a supplement. Then, the outputs of the three subnetworks are Y_1, Y_2 and Y, respectively. For recovering the size of the features in the frequency dimension, we concatenate the Y_1 and Y_2 to generate the $\hat{Y}$ as follow:

$$\begin{aligned}\hat{Y} = Con(&Y_2[:, 1 : 344, :];\\ &\lambda Y_2[:, 345 : 348, :] + (1 - \lambda)Y_1[:, 1 : 4, :];\\ &Y_1[:, 5 : 16, :])\end{aligned} \quad (1)$$

where *Con(·)* denotes a concatenation operation along frequency dimension. For guiding the network to autonomously learn the weight of two multi-band features around the breakpoint, we set a trainable parameter λ with an initial value of 0.5.



Octave	C	C♯	D	D♯	E	F	F♯	G	G♯	A	A♯	B
6	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
5	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
4	0.06	0.03	0.05	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.02
3	0.01	0.01	0.01	0.01	0.01	0.03	0.02	0.05	0.03	0.03	0.03	0.04
2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01

Figure 2: *Statistics of F0 acoustic locations in the training data. The value within cell represents the ratio that one case where F0 appears in certain tone and octave compared with all cases.*

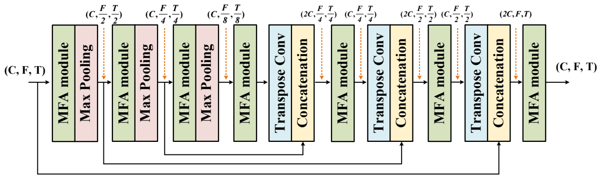

(C, F, T)
MFA module
Max Pooling
$(C, \frac{F}{2}, \frac{T}{2})$
MFA module
Max Pooling
$(C, \frac{F}{4}, \frac{T}{4})$
MFA module
Max Pooling
$(C, \frac{F}{8}, \frac{T}{8})$
MFA module
Transpose Conv
Concatenation
$(2C, \frac{F}{4}, \frac{T}{4})$
MFA module
$(C, \frac{F}{4}, \frac{T}{4})$
Transpose Conv
Concatenation
$(2C, \frac{F}{2}, \frac{T}{2})$
MFA module
$(C, \frac{F}{2}, \frac{T}{2})$
Transpose Conv
Concatenation
$(2C, F, T)$
MFA module
(C, F, T)

Figure 3: *The architecture of hourglass sub-network. The kernel sizes of transposed convolution and max pooling are both 2×2.*

2.2. Hourglass Sub-network

Based on good feature extraction capability of hourglass subnetwork [20, 21], we replace the conventional convolution with multi-scale feature aggregation (MFA) modules embedded with multilevel dilated convolution to solve the octave error problem caused by the fixed receptive field. To compensate for the loss of detailed information in higher-level semantic information, we concatenate features of the same size, and the dashed arrows indicate the dimension tracking here. Due to the T-F dependency of the spectrogram is easier to be learned through the neural network compared with that of the raw waveform signal, we use the multiple cascaded hybrid dilated CNN (HDC) blocks [22, 23] to extract the fine-grained T-F dependency. The details of the MFA module are shown in Fig.4. To cope with the problem that the dilated convolution can only learn information from surrounding pixels which can hardly learn global information, we aggregate the output features of all the previous layers

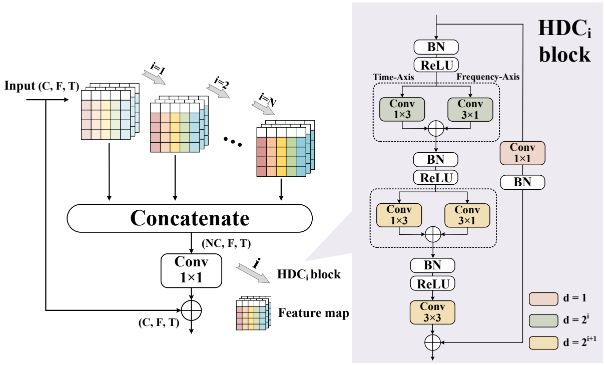


Figure 4: *The diagram of the multi-scale feature aggregation (MFA) module. d denotes the dilation factor of CNN.*

along the channel dimension in the last layer [24] and recover the same number of channels as the input by Con1×1.

For the HDC block, the coefficient *i* (index starts at 1) in HDC$_i$ block is an index correlated with the dilation factor. Each HDC block contains one traditional CNN with the kernel size of 3×3 (Conv 3×3) and two asymmetric convolution blocks which are more computationally efficient than the traditional 2-D convolution [25, 26]. The first asymmetric convolution block adopts the dilation rate of $d = 2^i$. The second asymmetric convolution block and Conv 3×3 adopt the same dilation rate, $d = 2^{i+1}$. Following the increase of *i*, the receptive field of the HDC$_i$ block is increased by feat of an exponentially growing dilation rate. For middle & low-frequency, high-frequency and full-frequency band hourglass sub-networks, the number of output channels of each HDC block is set to 16, 10 and 6. Meanwhile, the MFA modules are equipped with 4, 3 and 2 HDC blocks same as [27] to meet the frequency span and role needs of the three subnetworks processing features.

2.3. Feature Fusion Module

The feature fusion module (FFM) is a contribution of this work, which includes a time-frequency attention (TFA) block and Conv2D unit. Inspired by the idea of time-frequency attention mechanism [28, 8], we devise this module to fuse various multi-band features as shown in Fig.1. Given the input feature map $Z \in \mathbb{R}^{C\times F\times T}$, two operations of row average pooling along the frequency and time axis are employed to calculate the distribution of magnitudes along the time axis and temporal relationships for all frequency bins to generate M_f and M_t, respectively.

Then, we choose the 1-D convolution to perform frequency and temporal attention since it does well in learning the relationship along the frequency bins or time axis [8]. Relu and sigmoid are utilized to perform nonlinear activation and obtain feature maps $A_f \in \mathbb{R}^{C\times F\times 1}$ and $A_t \in \mathbb{R}^{C\times 1\times T}$, the process of operation can be written as:

$$\begin{aligned} A_f &= unsqueeze(\sigma(k_{f2} * (\delta(k_{f1} * M_f))) \\ A_t &= unsqueeze(\sigma(k_{t2} * (\delta(k_{t1} * M_t))) \end{aligned} \tag{2}$$

where k_{fi} and k_{ti} denote the kernels of convolutional layer. $*$ denotes the convolutional operation, σ and δ the ReLU and sigmoid activation functions and $unsqueeze$ denotes the dimension expansion operation. For convenience, Fig.1 only draws the process of generating A_f and A_t for one channel of Z.

Then, the attention maps are obtained by matrix multiplication and the output of the TFA block $\hat{Z}$ can be calculated as:

$$\hat{Z} = (A_f \otimes A_t) \odot Z' \tag{3}$$

where $\otimes$ and $\odot$ denote the matrix multiplication of the last two dimensions and element-wise product, respectively. Finally, the Conv2D unit is employed to perform linear recombination of inter-channel features and reduction of the channel dimension.

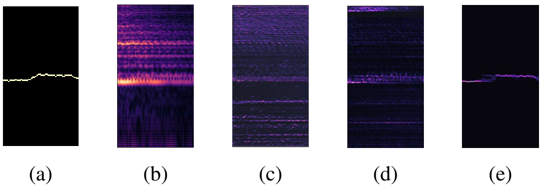


Figure 5: *Visualization analysis. Diagram (a) is the ground truth, (b) is the amplitude spectrogram from input X. Diagrams (c) and (d) are extracted from Y and $\hat{Y}$ in Fig.1. Diagram (e) is the output of the feature fusion module, $\tilde{Z}$.*

2.4. Visualization Analysis

To explore the effectiveness of each design within the multi-band feature extraction branch, we make a visualization analysis. As shown in Fig.5, (c) is an intermediate representation outputted from the full-frequency hourglass band sub-network. We can see that the full-frequency band hourglass sub-network can capture the low-frequency information though the input spectrogram has tiny energy in the low-frequency band. Compared with (c), (d) includes more F0 components. This shows the effectiveness of band partition. As shown in diagrams (a) and (e), the output of FFM can characterize the F0 components clearly, which shows the fusion ability of FFM for the various frequency band features.

3. EXPERIMENTS

3.1. Experimental Setup

We choose all 1,000 Chinese karaoke clips from the MIR-1K[1] and 35 vocal tracks from the Medley DB [29] as the training data. Then, the 12 clips from ADC2004, 9 clips from MIREX05[2], and 12 clips from MedleyDB are selected as the testing data which all contain human singing melodies. Note the testing data are not overlapped with the training data.

All data are re-sampled at 8 kHz. The STFT is calculated with the window size of 768 samples and hop size of 80 samples. We divide the training clips into fixed length segments of T = 128 frames, which is 1.28 seconds. To satisfy the frequency resolution for singing melody extraction, the number of CFP frequency bins is set to 360, with 60 bins per octave and 6 octaves in total. The frequency range spans from 32 Hz to 2,050 Hz covering C1 to B6.

Our model implemented with Pytorch was trained and tested in NVIDIA RTX 2080Ti GPUs. We choose the binary cross entropy as the loss function, and the Adam optimizer [30] with a learning rate of 0.0001 and batch size 8. Following the convention in the literature, 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) from mir_eval library [31]. For verifying the octave accuracy, we also use the raw octave accuracy (ROA) [15] as a supplementary metric.

[1] https://sites.google.com/site/unvoicedsoundseparation/mir-1k
[2] https://labrosa.ee.columbia.edu/projects/melody

Table 1: *The comprehensive results of the MTANet and compared methods on three datasets.*

Dataset	ADC2004						
Metrics	Params	VR	VFA↓	RPA	RCA	ROA	OA
MCDNN[1]	5.6M	76.8	13.8	71.3	72.8	73.4	74.1
MSNet[7]	0.5M	88.6	15.3	78.6	79.3	85.8	79.7
FTANet[8]	3.4M	85.8	**7.4**	79.0	79.1	84.5	81.5
TONet[15]	214M	86.1	15.3	82.3	82.7	84.9	82.4
MTANet	0.3M	**91.5**	11.3	**86.5**	**86.6**	**89.3**	**86.9**
Dataset	MIREX 05						
Metrics	Params	VR	VFA↓	RPA	RCA	ROA	OA
MCDNN[1]	5.6M	72.5	7.9	69.5	69.9	70.7	77.7
MSNet[7]	0.5M	86.4	11.9	78.2	78.6	83.7	81.8
FTANet[8]	3.4M	85.7	5.4	80.2	80.2	84.3	85.4
TONet[15]	214M	88.4	7.9	82.1	82.9	86.8	85.8
MTANet	0.3M	**91.8**	**4.2**	**85.5**	**85.5**	**88.5**	**89.2**
Dataset	MEDLEY DB						
Metrics	Params	VR	VFA↓	RPA	RCA	ROA	OA
MCDNN[1]	5.6M	51.3	12.2	44	46.1	46.4	64.5
MSNet[7]	0.5M	62.6	14.5	52.7	54.6	57.3	68.1
FTANet[8]	3.4M	63.4	**10.7**	57.1	58.0	60.4	72.2
TONet[15]	214M	65.7	12.1	56.8	58.9	61.3	71.4
MTANet	0.3M	**74.1**	15.5	**65.8**	**67.3**	**67.6**	**74.6**

Table 2: *Results of Ablation Study on the ADC2004 dataset.*

Methods				ADC2004					
Ver.	input_A	input_B	input_C	VR	VFA↓	RPA	RCA	ROA	OA
(i)	X_1	X_2	X	**91.5**	11.3	**86.5**	**86.6**	**89.3**	**86.9**
(ii)	$X_1^{'}$	$X_2^{'}$	X	87.4	12.5	83.9	83.9	84.6	85.4
(iii)	X	X	X	87.5	19.2	80.8	81.6	84.9	80.8
(iv)	X_1	X	X	86.4	14.2	79.0	79.2	84.0	80.2
(v)	X	X_2	X	90.5	24.0	81.1	81.5	87.7	80.1
(vi)	(i) w/o Overlap			86.0	8.0	83.9	83.9	84.6	85.4
(vii)	(i) w/o FFM			80.6	**7.2**	76.4	76.7	79.3	79.5
(viii)	(i) w/o ADB			89.5	14.7	84.1	84.9	86.7	84.3

3.2. Comprehensive Performance Comparison

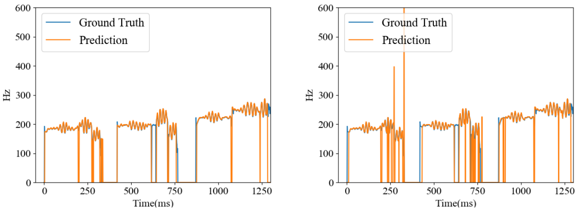
To investigate the effectiveness of the proposed MTANet, we compare it with four representative methods including the MCDNN [1], MSNet [7], FTANet [8], and TONet [15] on three datasets as shown in Tab.1. The CFP representation is employed as the input and the TONet is the state-of-the-art method. We carefully tuned the hyperparameters of four methods to ensure that they reached peak performances on our training dataset. The results show that the MTANet achieves the best scores except the VFA on three test sets. Generally, OA is considered more important than other metrics [3]. Compared with the similar structure of MSNet, the improvement of the performance fully verifies the effectiveness of the band partition scheme and overall structure deployment. Meanwhile, we also achieve good results on the basis of a large reduction in the parameters compared with TONet with multi-head self-attention, which makes our model more convenient for embedded design.

We also adopt a visualization approach to explore what types of errors are solved by our model as shown in Fig.6. We choose MSNet [7] to compare due to its structural similarity and popularity. Specifically, we plot the predictive frequencies over the time and ground truths by the MTANet and MSNet on one opera song: "opera male5.wav" from the ADC2004. We can observe that there are fewer octave errors (i.e., vertical jumps in the contours) in (a) than (b). Furthermore, there are fewer melody detection errors around 250ms and 750ms (i.e., predicting a melody frame as a non-melody one) in (a) than (b).

3.3. Ablation Study

As shown in Tab.2, we conducted seven ablations to verify the effectiveness of each design in the proposed MTANet. Due to the page limit, we only selected the ADC2004 dataset for ablation studies. More detailed results, pre-trained models and code implementations are available online[3]. (i) denotes our proposed complete MTANet. (ii) denotes the model where we partition X averagely and feed them into two sub-band hourglass subnetwork (i.e., $X_1^{'} = X[:, 1 : 182, :]$, $X_2^{'} = X[:, 178 : 360, :]$). To further explore the importance of the sub-band feature, we conducted the test on models (iii) with X as the input of each sub-band hourglass sub-network (i.e., without band partition). Then, we replace X with X_1 as the input of high-frequency band hourglass sub-network as (iv), replace X with X_2 as the input of mid&low-frequency band hourglass sub-network as (v). Meanwhile, (vi) denotes the model without the overlap of 4 bins on the basis of (i) for exploring the importance of boundary information between partitions. (vii) and (viii) denote the model without the feature fusion module and auxiliary detection branch, respectively.

Figure 6: *Visualized comparison between MTANet and MSNet on opera male5.*



(a) Opera male5 with MTANet (b) Opera male5 with MSNet

We make five observations. First, we compare (ii), (iii) with (i). Most metrics decrease by 1.0-6.1%, which reflects the effectiveness of partition by referring to the position distribution of the F0 of the actual singing voice. Second, the comparison between (iv), (v) and (iii) indicates that focusing on the information in a certain frequency band will cause performance loss. Then, we also discover that ROA in (v) is better than that in (iv). One possible reason is that the features emphasizing non-F0 components are concatenated in the top of feature map, which may have some negative effect on the subsequent F0 localization. Third, the results of (vi) show that all metrics reduce especially ROA when removing the overlap, which confirms our viewpoint that partition may destroy the harmonic information around the breakpoint. Fourth, the results of (vii) show that all metrics reduce significantly when removing the FFM, which notes that FFM can effectively fuse multi-band features to characterize the F0 component. Lastly, the results of (viii) verify the auxiliary detection role of ADB.

4. Conclusions

In this paper, we propose a multi-band time-frequency attention network (MTANet) for singing melody extraction. Our proposed band partition scheme is proven to be able to effectively exploit the positional distribution of the F0 component to further capture various multi-band features. Meanwhile, the multi-band features are fused well with the help of the feature fusion module (FFM). Experimental results show the MTANet achieves promising performances. The visualized result intuitively shows that the MTANet can effectively reduce the octave errors and the melody detection errors. We will focus on validating the performance of MTANet on more datasets and explore more ways to partition efficiently in future work.

5. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (U1903213), Multi-lingual Infomation Technology Reasearch Cente of Xinjiang (ZDI145-21).

[3] Codes are available in https://github.com/Annmixiu/MTANet

6. References

[1] S. Kum, C. Oh, and J. Nam, "Melody extraction on vocal segments using multi-column deep neural networks." in *ISMIR*, 2016, pp. 819–825.

[2] W. T. Lu, L. Su *et al.*, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning." in *ISMIR*, 2018, pp. 521–528.

[3] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[4] N. Kroher and E. Gómez, "Automatic transcription of flamenco singing from polyphonic music recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 901–913, 2016.

[5] C.-C. Wang and J.-S. R. Jang, "Improving query-by-singing/humming by combining melody and lyric information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 798–806, 2015.

[6] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 574–578.

[7] T.-H. Hsieh, L. Su, and Y.-H. Yang, "A streamlined encoder/decoder architecture for melody extraction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 156–160.

[8] S. Yu, X. Sun, Y. Yu, and W. Li, "Frequency-temporal attention network for singing melody extraction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 251–255.

[9] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for f0 estimation in polyphonic music." in *ISMIR*, 2017, pp. 63–70.

[10] S. Yu, X. Chen, and W. Li, "Hierarchical graph-based neural network for singing melody extraction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 626–630.

[11] S. Yu, Y. Yu, X. Sun, and W. Li, "A neural harmonic-aware network with gated attentive fusion for singing melody extraction," *Neurocomputing*, vol. 521, pp. 160–171, 2023.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[13] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] K. Chen, S. Yu, C.-i. Wang, W. Li, T. Berg-Kirkpatrick, and S. Dubnov, "Tonet: Tone-octave network for singing melody extraction from polyphonic music," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 621–625.

[16] H. Liu, L. Xie, J. Wu, and G. Yang, "Channel-wise subband input for better voice and accompaniment separation on high resolution music," *arXiv preprint arXiv:2008.05216*, 2020.

[17] L. Su and Y.-H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1600–1612, 2015.

[18] T. Kobayashi and S. Imai, "Spectral analysis using generalised cepstrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1235–1238, 1984.

[19] L. Su, "Between homomorphic signal processing and deep neural networks: Constructing deep algorithms for polyphonic music transcription," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 884–891.

[20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 483–499.

[21] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *ISMIR*, 2018, pp. 289–296. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/138_Paper.pdf

[22] Y. Hu, X. Zhu, Y. Li, H. Huang, and L. He, "A multi-grained based attention network for semi-supervised sound event detection," *arXiv preprint arXiv:2206.10175*, 2022.

[23] Y. Chen, Y. Hu, L. He, and H. Huang, "Multi-stage music separation network with dual-branch attention and hybrid convolution," *Journal of Intelligent Information Systems*, pp. 1–22, 2022.

[24] Y. Lee, J.-w. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and gpu-computation efficient backbone network for real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[26] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1911–1920.

[27] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[28] Y. Zhao and D. Wang, "Noisy-reverberant speech enhancement using denseunet with time-frequency attention." in *Interspeech*, vol. 2020, 2020, pp. 3261–3265.

[29] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *ISMIR*, vol. 14, 2014, pp. 155–160.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics." in *ISMIR*, 2014, pp. 367–372.