INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Vocal Pitch Extraction in Polyphonic Music using Convolutional Residual Network

Mingye Dong[1], *Jie Wu*[2], *Jian Luan*[2]

[1]University of Science and Technology of China
[2]Microsoft Search Technology Center Asia, Xiaoice

sa517062@mail.ustc.edu.cn, Jie.Wu@microsoft.com, jianluan@microsoft.com

Abstract

Pitch extraction, also known as fundamental frequency estimation, is a long-term task in audio signal processing. Especially, due to the presence of accompaniment, vocal pitch extraction in polyphonic music is more challenging. So far, most of deep learning approaches use log mel spectrogram as input, which neglect the phase information. In addition, shallow networks have been applied on waveform directly, which may not handle contaminated vocal data well. In this paper, a deep convolutional residual network is proposed. It analyzes and extracts effective feature from waveform automatically. Residual learning can reduce model degradation due to the skip connection and residual mapping. In comparison to reported results, the proposed approach shows 5% and 4% improvement on overall accuracy(OA) and raw pitch accuracy(RPA) respectively.

Index Terms: vocal pitch extraction, polyphonic music, convolutional residual network, raw waveform

1. Introduction

Pitch extraction, also referred to fundamental frequency (f0) estimation is a task to estimate the lowest frequency partial from audio signal, which has been a popular research topic for many years [1]. Extraction of the pitch contour is important in many tasks, such as speech processing [2], recognition [3], singing processing [4] and music information retrieval [5]. Pitch extraction plays a key role in singing signal processing, where pitch is a core component of melody. Especially, due to the presence of accompaniment, vocal pitch extraction in polyphonic music is more challenging.

Many heuristic based pitch extraction algorithms have been studied, which usually calculate the saliency function of pitch candidates or separate the singing signal from polyphonic music. Salamon [6] proposed to extract melody based on pitch contours tracking and characterization. In [7], source-filter model is used to do pitch salience estimation. Meanwhile, machine learning methods based on classification are also applied to this task. Ellis and Poliner proposed a support vector classifier to predict pitch label from STFT features [8]. Bittner used random forest classifier to do extract pitch based on the highly handcrafted features [9].

Recently, many deep learning approaches have been proposed to learn the pitch directly from data. Deep Neural Network (DNN) was proposed to perform f0 estimation based on log-spectrograms [10]. Kum [11] proposed to use multi-column deep neural networks (MCDNN) to predict pitch from spectrogram. Bittner [12] described a fully convolutional neural network (CNN) for multi-f0 tracking based on harmonic constant-Q transform (HCQT). Meanwhile, a patch-based CNN was proposed to extract vocal melody from the combined frequency and periodicity (CFP) representation [13]. Even though, these acoustic features, such as log-spectrogram, HCQT and CFP are designed to capture the characteristic of the waveform, there always exists phase information neglection problem. Several papers have been proposed to operate directly on the raw time-domain waveform using deep learning approaches. In [14], raw waveform was directly used as input to multi-layered neural network to perform end-to-end learning. At the same time, raw waveform was also directly fed into the CLDNN (Convolutional, LSTM, DNN) [15] to do voice activity detection (VAD). Kim [16] proposed a pitch tracking algorithm in monophonic audio signal based on the CNN and raw waveform. However, these networks applied on waveform directly are usually shallow, which may not handle the contaminated vocal data well in polyphonic music. Convolutional neural network with deep layers has show significant performance in many areas, such as image recognition and computer vision [17]. However, it is hard to stack deeper layers due to the gradient vanishing problem. To address the problem, He [18] proposed a residual neural network (ResNet) with an identity shortcut connection. ResNet allows the training of over hundred layers with increasing accuracy and can greatly improve the training efficiency and reduce model degradation due to the skip connection and residual mapping.

Following the success and significant performance of ResNet and raw time-domain waveform, in this paper, we propose to use a deep convolutional residual network to analyze and extract effective feature from waveform automatically to do vocal pitch extraction in polyphonic music.

The rest of the paper is as follows: Section 2 presents the residual neural network. The proposed approach is described in Section 3. Section 4 is devoted to providing the experimental setup and analyzing the results. Section 5 discusses and concludes the paper and Section 6 gives the future work.

2. Deep Residual Network

Deep residual network, also called ResNet is a very deep neural network with skip connections by passing input from one layer

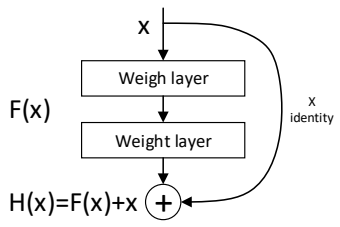

Figure 1: *Architecture of a residual block [18]*

http://dx.doi.org/10.21437/Interspeech.2019-2286

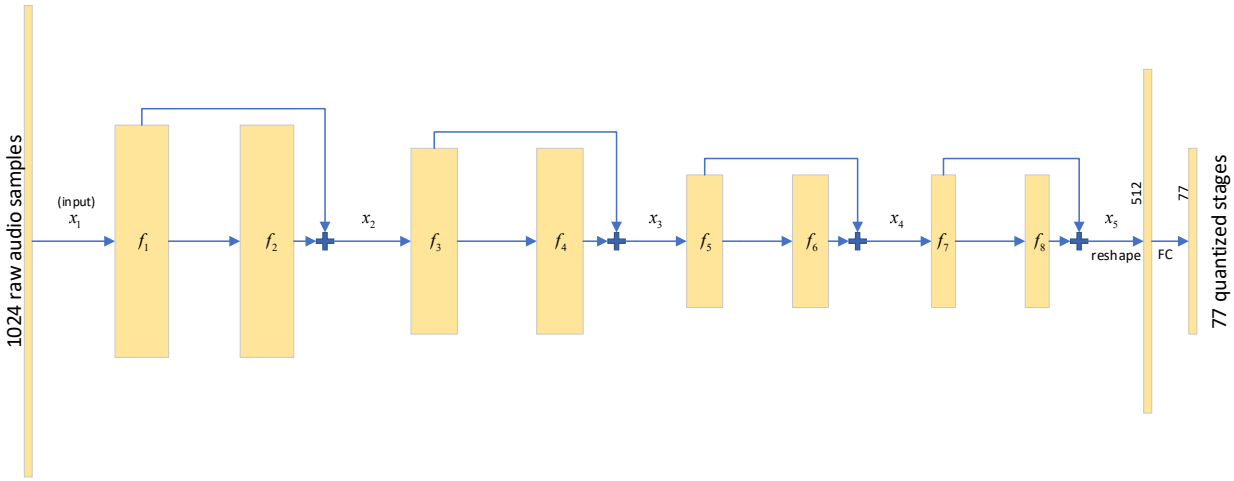

Figure 2: *The architecture of the proposed approach. Running eight convolution layers directly on the raw waveform produces an output vector representing the vocal pitch classification, which is then used to derive an accurate pitch value.*

to one or more later layers proposed in [18]. It consists a series of residual blocks. The diagram of one building block is shown in Figure 1, it can be expressed as:

$$H(x) = x + F(x), \tag{1}$$

where x is the input, $H(x)$ is the output of the stacked two layers, $F(\cdot)$ represents the weight layer, which can be convolution layer, pooling layer, batch normalization (BN) layer [19] and so on. x represents the identity skip connection. In traditional neural network, gradient is basically the product of a series of matrix. When the layers go deep, gradient goes small till to vanishing, which leads to a wrong performance. To overcome the gradient vanishing problem in deep network, the identity skip connection in the residual network is the ultimate solution, since in the skip connection, information could be directly propagated from one layer to any other layers in both forward and backward pass. Therefore, the skip connection can increase the training efficiency.

ResNet allows deeper network training and can generate models with better performance. It achieves impressive results on CIFAR-10, CIFAR-100, and SVHN [20]. ResNet's powerful presentation capabilities are not only reflected in the image classification direction [18], but also greatly improve the performance of many other computer vision applications, such as object detection [21] and facial recognition.

3. Architecture

This paper proposes a deep convolutional residual network which directly analyzes on the raw waveform to do vocal pitch extraction in polyphonic music. The overall architecture of the proposed approach is presented in Figure 2.

1024 samples extracted from the raw waveform are used as input. The network consists of eight convolutional layers and each of them is followed by a Batch-Normalization (BN) layer. Residual mapping is added for every two stacked convolutional layers. To simply interpret the network architecture, the Rectified Linear Unit (ReLU) and Batch Normalization notations are ignored in Figure 2. Now, the calculation can be expressed as the following equations:

$$x_2 = f_1(x_1) + f_2(f_1(x_1)), \tag{2}$$

$$x_3 = f_3(x_2) + f_4(f_3(x_2)), \tag{3}$$

$$x_4 = f_5(x_3) + f_6(f_5(x_3)), \tag{4}$$

$$x_5 = f_7(x_4) + f_8(f_7(x_4)), \tag{5}$$

where x_1, x_2, x_3 and x_4 are the input of two stacked layers, x_5 is the output of the residual network, f_1, f_2...f_8 are the activation functions of these convolutional layers.

The vocal pitch extraction in polyphonic music is considered as a classification task, since the 1-dimensional pitch feature is too limited in regression task. We linearly quantify the frequency range into 76 states between 87.309Hz and 784.00Hz, which covers four octaves. Besides, we add one state to indicate the silence and unvoiced singing segments. Finally, the dimension of output is 77. Moreover, conventional one-hot output label is smoothed to extend the learning range. In this way, different weights are added to their corresponding bands. For example, if 100Hz is quantized to the 10th band, then the 10th band is 0.8, while the 9th and 11th band are both 0.1.

In order to mimic the critical bands of human subjective listening perception, the frequency range is quantified on mel-scale. The binary cross entropy between the target vector and the predicted one is used as the loss function:

$$\pounds(y, \hat{y}) = \sum_{i=1}^{77}(-y_i log\hat{y} - (1 - y_i)log(1 - \hat{y_i})), \tag{6}$$

Finally, x_5 is reshaped into a 512-dimensional vector and then mapped into the 77-dimensional output vector using a fully-connected feed-forward layer. The objective state in the predicted 77-dimensional vector is the one who has the maximum weight, and then we convert the intermediate value of the objective state to its corresponding frequency value.

4. Experiment

4.1. Experimental Setup

In our experiments, the dataset MIREX is used as training and evaluation data, since it has the ground truth annotation for ev-

Table 1: *Details of the parameters used for the proposed architecture.*

name	kernel size/stride	outputsize
input	-	1024
conv1	256/8	128*512
conv2	32/1	128*512
conv3	32/4	32*256
conv4	8/1	32*256
conv5	8/4	8*128
conv6	2/1	8*64
conv7	2/1	8*64
conv8	2/1	8*64
fc	-	77

ery 20 milliseconds which is extracted from the clean singing data. MIR-1k [22] is a Chinese pop song dataset that includes 1000 songs with a total length of 133 minutes. The raw audio contains two channels, the left is clean singing signal and the right is the accompaniment signal. To perform at different signal-to-noise (SNR), we extract these two channels separately and then add the accompaniment to the clean singing signal at 0dB and +5dB.

900 songs randomly-selected from the dataset are used as the training and validation data, the remaining 100 songs are used as the testing data. Both the training and testing data are a mix of the clean singing signal and accompaniment at 0dB and +5dB. To reduce the training complexity, the raw audio is downsampled to 8kHz.

In our proposed approach, the input is 1024 samples, so each frame of the input contains 128 milliseconds audio clips. There are eight convolutional layers and a BN layer behind each convolutional layer in the network. The output is the 77-dimensional quantized vector on mel-scale. The number of units, the filter size and the stride in each convolutional layer are listed in the Table 1. The deep convolutional residual network (CRN) model is optimized by ADAM [23] optimizer with a learning rate of 0.0001.

In order to verify the performance of our proposed approach (CRN-Raw), we compare it with two recently developed deep learning based approaches DSM-HCQT and DNN-Raw. We implement these three approaches for comparison:

- **DSM-HCQT**: deep salience map approach in which the input HCQT is a 3 dimensional array of 360*50*6 indexed by harmonic, frequency and time. The network contains five convolutional layers. The threshold parameter is 0.1, which achieves the best performance. More detail referred to [12].
- **DNN-Raw**: the input is raw waveform samples of 480 dimension, and the network has 3 fully-connected feed-forward layers with 4096 units per layer [14].
- **CRN-Raw**: the proposed approach which uses the deep convolutional residual network to directly operate on raw waveform.

4.2. Evaluation Metrics

A brief introduction of the evaluation metrics [24] used in this paper is shown below:

- **Voicing Recall Rate (VR)**: the proportion of frames labeled as voiced in ground truth to that estimated as voiced frames by the approach

Table 2: *Vocal pitch extraction results of different approaches at 0dB*

Method	OA	**RPA**	RCA	**VR**	VFA
DSM-HCQT	73.3	74.2	78.2	89.1	32.3
DNN-Raw	74.5	72.2	82.3	87.0	18.6
CRN-Raw	78.4	**78.6**	87.0	**91.7**	12.0

Table 3: *Vocal pitch extraction results of different approaches at +5dB*

Method	OA	**RPA**	RCA	**VR**	VFA
DSM-HCQT	78.3	80.5	83.6	92.1	26.6
DNN-Raw	78.5	77.2	86.9	89.3	7.5
CRN-Raw	82.2	**81.7**	90.9	**93.3**	4.6

- **Voicing False Alarm Rate (VFA)**: the proportion of frames labeled as unvoiced in the ground truth to that are estimated as voiced by the approach.
- **Raw Pitch Accuracy (RPA)**: the proportion of voiced frames where the estimated pitch is within $\pm\frac{1}{4}$ tone (50 cents) of the ground truth pitch.
- **Raw Chroma Accuracy (RCA)**: the proportion of voiced frames in which the estimated pitch and the ground truth pitch are mapped into a single octave. It gives a measure of the pitch accuracy ignoring the octave errors.
- **Overall Accuracy (OA)**: the proportion of frames estimated correctly by the approach considering the pitch extraction accuracy and the voice activity detection. For voiced frames, it is correctly estimated within the $\pm\frac{1}{4}$ tone range of the ground truth pitch and for unvoiced frames, it is correctly estimated as the ground truth.

4.3. Results

The proposed approach (CRN-Raw) is compared with two recently developed deep learning based approaches DSM-HCQT and DNN-Raw.

Table 2 and Table 3 show the evaluation results of these three methods on signals at 0dB and +5dB SNR respectively. Both at 0dB and +5dB, our proposed approach achieves the best performance, which confirms its effectiveness in vocal pitch extraction from polyphonic music. Here we just take the results of Table 2 as an example to compare these three approaches individually. Firstly, we would like to see the performance of deep residual network in vocal pitch extraction. In Table 2, the CRN-Raw achieves a 5% higher score than DNN-Raw in VR, and a 6.6% lower score in VFA. It shows that the CRN has an advantage over DNN in the voice activity detection(VAD) task. At the same time, the scores of CRN-Raw are higher than those of DNN-Raw both in RPA and PCA, which confirms that CRN can improve the accuracy of model training to predict the pitch state. Moreover, it also shows smoothed one-hot label can increase RPA and RCA. The above results indicate CRN has a advantage in voiced/unvoiced classification and pitch state estimation, which contributes a higher score in OA. On the whole, these results confirm the effectiveness of deep residual network in vocal pitch extraction task.

Compared with DSM-HCQT in Table 2, we can see that CRN-Raw achieves a 3% higher score in VR and a 20% lower score in VFA. As for RPA, RCA and OA, CRN-Raw also ob-

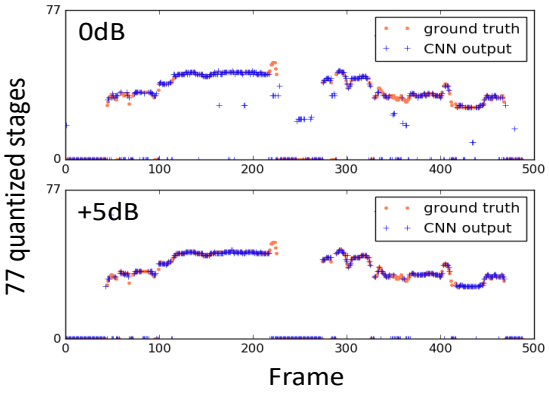

Figure 3: *The trajectory of pitch quantized states for estimated and ground truth*

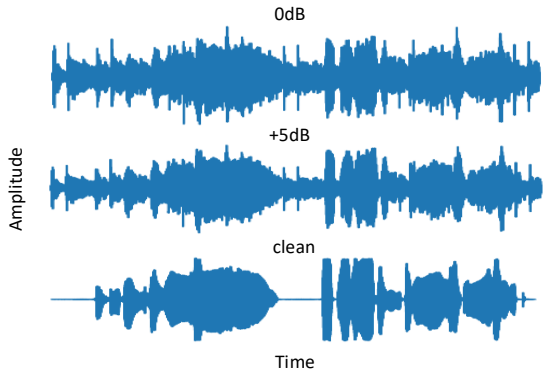

Figure 4: *The three waveforms show the effect of accompaniment on the overall amplitude at different SNRs.*

tains the higher scores. These results show that the raw time-domain waveform is significantly better than HCQT. In other words, even the HCQT feature is designed to capture the characteristic of the waveform, the phase information neglection problem is still serious, which would reduce the estimation accuracy to some extent.

In addition, we note that the CRN-Raw achieves a better performance at +5dB from Table 2 and Table 3. Compared with the other two algorithm test results, the proposed method improves the test result on the 0dB test set more than the +5dB test set, which also indicates that the method has stronger anti-accompaniment interference ability.

The pitch quantized state trajectories of the CRN-Raw predicted output and the ground truth are shown in Figure 3. Compared to the +5dB, there exist more classification errors at 0dB,which results in a lower VR and a higher VFA score. The contour of predicted quantized pitch states is quite near to the ground truth state contour, which contributes to a high RCA value both in Table 2 and Table 3. Moreover, even if the pitch state is predicted to the adjacent state of its corresponding ground truth state, human might not perceive the difference in subjective listening. The values of objective evaluation metrics in Table 2 and Table 3 show that it is more difficult to extract vocal pitch at a lower SNR. Waveforms of polyphonic music at different SNRs are shown in Figure 4. We can see that the waveform at 0dB is more complicated than that at +5dB, since the amplitude at 0dB varies widely and the accompaniment brings more distortion to the clean singing signal. In conclusion, the results show that the energy of accompaniment has a great influence on the vocal pitch extraction task where the energy is greater. However, the waveforms under different SNR conditions are still similar in contour to the clean singing signal waveform. Therefore, the pitch information can be extracted by modeling directly on the raw waveform.

5. Discussion and Conclusions

In this paper, a deep convolutional residual network is proposed to directly analyze on raw time-domain waveform to do vocal pitch extraction in polyphonic music. The objective evaluation metrics show that our proposed approach achieves the best performance compared to another two recently developed deep learning based approaches DNN-Raw and DSM-HCQT, which confirms the significant capability of residual network in improving the accuracy and the importance of waveform in convolutional network.

As we all know, vocal pitch is a core component of melody, which is extremely important in singing voice synthesis. Much effort has been contributed to estimate pitch in monophonic music. However, it is hard and expensive to collect these monophonic music in real application. Many singers have published much polyphonic music. Collecting such polyphonic music is quite easy to some extent. Therefore, it is necessary to investigate how to improve the performance of extracting vocal pitch from polyphonic music.

Due to the presence of accompaniment, vocal pitch extraction from polyphonic music is very challenging since the prosody tendency of accompaniment is similar to that of the monophonic music. In addition, the pitch range of singing signals usually has eight octaves, which is greatly larger than that of speech signal. Therefore, we propose to use deep convolutional residual network to learn the inner relationship between the temporal variation. Furthermore, the residual network allows deeper layers without gradient-vanishing problem and can also improve the training efficiency and increase the estimation accuracy due to its identity skip connection and residual mapping. Meanwhile, the raw time-domain waveform is directly used as input of the network, to avoid the information neglection problem in feature transition. The experiments show that the first layer of a convolutional network can directly learn characteristic from the raw waveform precisely. What's more, we find that a large convolution kernel in the first layer can contribute to the learning from raw waveform. This deep learning method of direct modeling on the raw waveform can also be applied to other pitch extraction related tasks, which may have unexpected effects.

6. Future Work

In our experiments, we found that different music instruments may show quite different impact to vocal pitch extraction task. For example, in the case of guitar or piano, even if with a higher SNR, the result is still poor. Consequently, the performance of the proposed approach can be further improved on robustness to different kinds of instruments. It may bring significant improvement to increase the proportion of these types of data in training data, or customize specific. Besides, we will apply smoothing or dynamic programming algorithms to decrease the isolated pitch state errors.

7. References

[1] P. De La Cuadra, A. S. Master, and C. Sapp, "Efficient pitch detection techniques for interactive music." in *ICMC*, 2001.

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[3] L. R. Rabiner, B.-H. Juang, and J. C. Rutledge, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[4] X. Rodet, "Synthesis and processing of the singing voice," in *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*. Citeseer, 2002, pp. 15–31.

[5] J. S. Downie, "Music information retrieval," *Annual review of information science and technology*, vol. 37, no. 1, pp. 295–340, 2003.

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

[7] J. J. Bosch, R. M. Bittner, J. Salamon, and E. Gómez, "A comparison of melody extraction methods based on source-filter modelling." in *ISMIR*, 2016, pp. 571–577.

[8] D. P. Ellis and G. E. Poliner, "Classification-based melody transcription," *Machine Learning*, vol. 65, no. 2-3, pp. 439–456, 2006.

[9] R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello, "Melody extraction by contour classification." in *ISMIR*, 2015, pp. 500–506.

[10] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks." in *ISMIR*, 2016, pp. 737–743.

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

[12] 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.

[13] L. Su, "Vocal melody extraction using patch-based cnn," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 371–375.

[14] P. Verma and R. W. Schafer, "Frequency estimation from waveforms using multi-layered neural networks." in *INTERSPEECH*, 2016, pp. 2165–2169.

[15] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[17] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[21] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1476–1481, 2017.

[22] MIR-1k:, https://sites.google.com/site/unvoicedsoundseparation/mir-1k.

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

[24] G. E. Poliner, D. P. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.