CNN BASED TWO-STAGE MULTI-RESOLUTION END-TO-END MODEL FOR SINGING MELODY EXTRACTION

Ming-Tso Chen, Bo-Jun Li, and Tai-Shih Chi

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan

ABSTRACT

Inspired by human hearing perception, we propose a two-stage multi-resolution end-to-end model for singing melody extraction in this paper. The convolutional neural network (CNN) is the core of the proposed model to generate multi-resolution representations. The 1-D and 2-D multi-resolution analysis on waveform and spectrogram-like graph are successively carried out by using 1-D and 2-D CNN kernels of different lengths and sizes. The 1-D CNNs with kernels of different lengths produce multi-resolution spectrogram-like graphs without suffering from the trade-off between spectral and temporal resolutions. The 2-D CNNs with kernels of different sizes extract features from spectro-temporal envelopes of different scales. Experiment results show the proposed model outperforms three compared systems in three out of five public databases.

Index Terms— Melody extraction, multi-resolution, convolution neural network, end-to-end learning, music information retrieval

1. INTRODUCTION

Melody consists of a pitch track which one might hum to recognize a clip of polyphonic music. Melody extraction is one of the popular topics in the research field of music information retrieval (MIR) [1]. Melody contains important information of music and could be further used in many applications such as query-by-humming [2], version identify [3], and audio source separation [4][5].

The most commonly used data representation of music in MIR related work is the spectrogram transformed via the short time Fourier transform (STFT). Recently, there has been an increasing focus on directly using the raw waveform as the data representation for an end-to-end learning model [6][7][8][9]. In general, the features obtained from the Fourier spectrogram are effective enough for many applications. However, the Fourier spectrogram is generated using a fixed time window such that it uniformly depicts the sound using a particular temporal and spectral resolution. On the other hand, psychoacoustic studies show that people detect pitch using information embedded in different resolutions. For the low frequency region, people analyze the sound using a frequency resolution high enough to resolve each individual harmonic to decipher pitch. For the high frequency region, people analyze the sound using a temporal resolution high enough to decipher the periodicity pitch, which is the reciprocal of the time-domain period of the sound. Based on this duplex behavior of human pitch perception, we have built a high-performance composite neural network (NN) for singing melody extraction by combining a CNN-based NN in the spectrogram domain and a NN in the time domain [10]. Similarly, a novel representation with combined information of frequency and periodicity was used to extract the melody [11].

In addition to the multi-resolution property, human hearing perception is a multi-stage process. Based on neurophysiological data, a two-stage auditory model was proposed in [12]. The first stage estimates the spectrum using a bank of constant-Q filters by mimicking the frequency selectivity of the cochlea. The second stage mimics the function of the auditory cortex (A1), which analyzes the spectral-temporal envelope of the sound, using a bank of 2-D spectro-temporal modulation filters. Inspired by these perceptual properties, we build a CNN-based two-stage multi-resolution end-to-end model for singing melody extraction in this paper. Unlike the approach [10], which combines a pure spectral NN at a fixed resolution and a pure temporal NN, the proposed model in this paper analyzes the joint spectro-temporal patterns of the sound at various resolutions to decipher pitch. In the proposed model, the first stage was implemented using the 1-D CNN to similarly behave as a spectrum estimator. The second stage was implemented using the 2-D CNN to analyze the joint spectral-temporal contents of the sound. In order to extract information embedded in different resolutions, we used two 1-D CNNs, whose kernels are with different lengths, in parallel in the first stage.

The rest of this paper is organized as follows. In Section 2, we describe the architecture of the proposed two-stage model. In Section 3, we give details of the data preparation and the configuration of the proposed model. Experiment results on singing melody extraction are demonstrated in Section 4 and the conclusion is given in Section 5.

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 107-2221-E-009-132-MY3.

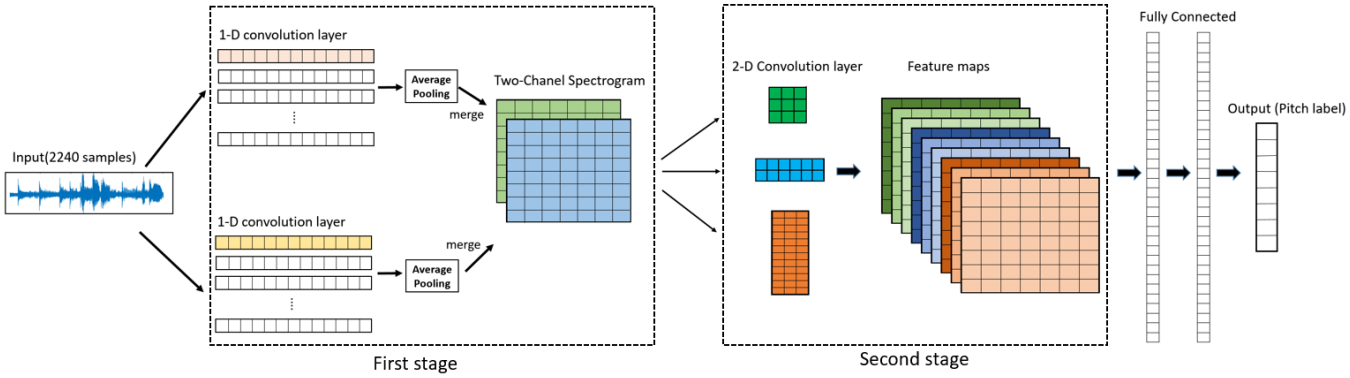


Fig. 1. Architecture of the proposed model.

2. MODEL ARCHITECTURE

The proposed model is shown in Fig. 1. The first stage consists of two paralleled 1-D CNNs with kernels of different lengths. The input waveform through the first stage produces two spectrogram-like graph representations resolved by two frequency resolutions. Then, the second stage utilizes 2-D CNNs to extract the joint spectro-temporal features from the spectrogram-like representations. The 'Inception' module [13] is used to expand the width of the model to simulate multi-resolution analysis on the graph using 2-D kernels with different sizes. Finally, the output features of the second stage are cascaded for the following two fully connected layers, each of which has 1024 units, to predict the output pitch state of the input signal. ReLu is used in all units as the activation function in the proposed model.

2.1. Settings of the first stage

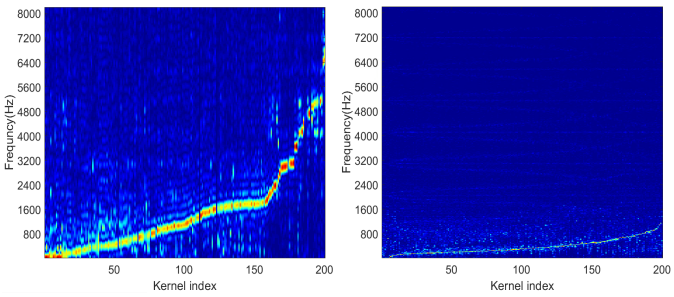
We used the 1-D CNN as an alternative to Fourier transform for spectrum estimation. The window size used in the short-term Fourier transform (STFT) determines the time/frequency resolution of the analysis and there exists the trade-off between the time and the frequency resolutions. Similarly, the length of the 1-D CNN kernels, which can be thought as the impulse responses of filters, determines the frequency bandwidth of the analysis bands. Therefore, using 1-D CNN kernels with the same length is just like performing frequency analysis using a bank of filters whose bandwidth is limited by a pre-set minimum value. Therefore, we used two 1-D CNNs in parallel with different kernel lengths to produce two spectrogram-like graphs using two limitations on frequency resolution. Each 1-D CNN can be thought as an independent channel here.

In our model, two different lengths of the kernels of the paralleled 1-D CNNs were set to 960 (60 ms) and 64 (4 ms) and the stride size was set to 16 (1 ms). There were 200 kernels in each 1-D CNN network. After the sound passing through the 1-D convolutional layer, the average pooling was performed to somehow downsample the output of the convolutional layer. In our pilot simulations only using each of the 1-D CNNs, we found average pooling can slightly improve the performance of the network with the kernel length of 960 over maximum pooling. However, the performance boost using average pooling rather than maximum pooling was not observed in the CNN with the kernel length of 64. Therefore, we used average pooling with the pool size of 10 samples in all of our experiments.

2.2. Settings of the second stage and pre-training

The second stage adopted 2-D CNNs to extract useful spectro-temporal patterns, which might include the harmonic structure, temporal continuity, and other melody related patterns. Therefore, the input of this stage should be like a spectrogram-like graph to have proper spectro-temporal characteristics. However, the proposed model is an end-to-end learning method such that the output graph of the first stage exhibits random permutation on the kernel-index axis according to the learned weights in the first stage. To address this problem, we pre-trained a different model consisting of only one 1-D CNN for melody extraction. Fig. 2 shows the properly ordered magnitude responses of the kernels with lengths of 64 and 960 from this pre-trained model. It clearly shows 1-D kernels with the length of 960 only focus on low-frequency regions with high frequency resolution. In contrast, those 1-D kernels with the length of 64 can analyze the frequency regions up to 5 kHz with relatively low frequency resolution. After the pre-training, the 1-D kernels of the proposed 2-stage model were initialized using these properly ordered pre-trained kernels shown in Fig. 2.

Three different sizes, including 100×4 (freq$\times$time), 70×8, and 40×12, were selected as the sizes of 2-D kernels in the second stage to capture the spectral-temporal structures respectively. In our model, we used 24 2-D kernels in total (8 kernels for each size) in the second stage without pooling



(a) kernel length of 64 points (b) kernel length of 960 points

Fig. 2. Magnitude responses of the pre-trained 1-D CNN kernels with different lengths. Note, the kernels are ordered based on the frequencies of the maximum responses.

layers. In other words, this stage can be thought as containing three paralleled 2-D CNNs with different kernel sizes. Therefore, the input of the second stage contains two spectrogram-like graphs from the two 1-D CNN channels such that the proposed second stage will automatically produce suitable 2-D kernels during training for either the first or the second 1-D CNN channel.

3. EXPERIMENTS

3.1. Data representation and pitch labels

Because the proposed model is an end-to-end learning model, the time domain waveforms are taken as the input representation. We consider melody extraction as a classification problem and follow the previous studies [10][14] to quantize continuous frequency to the suitable pitch label from D2 to F#5 (from 73 to 740 Hz) with a step of 50 cents (1/2 semitone). In addition to these 82 states, one more state of "non-melody" is included for those frames without pitch. There are 83 states in total for our model output conditions and the training criteria are cross-entropy and softmax function for generating probability of each state. In our experiments, a time-domain sequence of 140 ms (2240 sample points for 16 kHz sampling frequency) was used as the input signal to predict the pitch state of the center section of 20 ms. There was an overlap of 120 ms between consecutive input sequences.

3.2. Dataset and evaluation metrics

To train the proposed model and evaluate its performance, we used five popular datasets in the research field of music information retrieval, including MIR-1K [15], iKala [16], MedleyDB [17], ADC2004, and MIREX05. The first two datasets consist of vocal melody and the others consist of mixed melody from vocal and instrumental sounds. Because we only focused on extracting vocal melody, the clips with vocal melody in MedleyDB dataset were picked based on labeled styles of the clips [17]. Only 12 and 9 clips were respectively selected from ADC2004 and MIREX05 datasets based on the criteria in [11][14]. For training, we used 740 and 200 vocal clips from MIR-1k and iKala dataset, respectively. The rest clips of the MIR-1K and iKala datasets were used for evaluation. In our experiments, singing voice and background music were mixed with equal energy, i.e., SNR = 0 dB. All clips extracted from these datasets were resampled to 16 kHz. The melody pitch tracks provided by these datasets were also resampled to provide pitch values at certain instants using the interpolation method of the mir_eval toolkit released at ISMIR 2014 [18].

	RPA	RCA	OA
CNN-960	75.84	78.94	77.46
CNN-64	47.45	51.01	60.70
Multi-CNN	77.79	81.05	78.34

Table 1. Three main performance scores of three compared 1-D CNN based systems using MIR-1K dataset.

There are five standard metrics for melody extraction evaluation, including the voicing recall rate (VR), the voicing false alarm rate (VFA), raw pitch accuracy (RPA), raw chroma accuracy (RCA) and overall accuracy (OA). To compute the scores, we also used the library of mir_eval.

4. RESULTS

The study showed applying the Viterbi algorithm as post-processing does not improve performance of singing melody extraction very much [10]. Therefore, in this study, no post-processing module such as HMM [20], the Viterbi algorithm [21] or the dynamic programming algorithm was used with the proposed model to further smooth the estimated pitch contour. The output of the proposed model was used to directly compare with the ground truth for calculating the scores. All scores shown in this section are in percent (%).

To show that providing multi-resolution information is beneficial to melody extraction, we first evaluated the 1-D CNN networks during pre-training. Only the MIR-1K dataset was used in this evaluation, where 740 vocal clips were used for training and the remaining 260 vocal clips were used for

	VR	VFA	RPA	RCA	OA
Multi-CNN	88.89	20.33	77.79	81.05	78.34
Proposed model	88.27	16.65	79.27	81.67	80.46
No pre-training	88.55	18.05	78.95	81.59	79.83

Table 2. Performance scores of the proposed 2-stage model initialized with pre-trained 1-D CNN kernels ('Proposed model') or random kernels ('No pre-training'). The 'Multi-CNN' system only contains the first stage of the proposed model.

	VR	VFA	RPA	RCA	OA
Proposed	**88.25**	17.20	**79.32**	**81.58**	**80.33**
Hybrid [10]	80.97	14.74	70.30	73.88	74.67
MCDNN [14]	77.49	**11.29**	69.74	72.46	75.28
Melodia [19]	84.78	30.04	69.87	72.37	69.89

(a) MIR-1K

	VR	VFA	RPA	RCA	OA
Proposed	**89.47**	16.15	**81.17**	**82.41**	**82.05**
Hybrid [10]	83.65	17.30	74.50	76.97	77.21
MCDNN [14]	77.25	**9.46**	71.23	73.89	77.59
Melodia [19]	81.97	26.76	72.64	74.77	72.83

(b) iKala

	VR	VFA	RPA	RCA	OA
Proposed	64.63	18.51	54.27	59.80	58.59
Hybrid [10]	56.65	**9.88**	50.20	55.03	56.54
MCDNN [14]	50.19	10.15	45.38	49.28	58.37
Melodia [19]	**81.47**	17.24	**71.72**	**74.86**	**73.48**

(c) ADC2004

	VR	VFA	RPA	RCA	OA
Proposed	**87.15**	12.65	**79.66**	**80.84**	**82.31**
Hybrid [10]	81.91	7.37	74.36	76.22	80.67
MCDNN [14]	75.75	**5.99**	70.10	71.60	78.36
Melodia [19]	87.44	24.60	78.46	79.73	77.40

(d) MIREX05

	VR	VFA	RPA	RCA	OA
Proposed	**86.19**	43.33	**65.61**	**71.54**	60.04
Hybrid [10]	81.36	41.37	62.99	69.13	60.27
MCDNN [14]	77.16	**37.10**	60.09	66.06	**61.84**
Melodia [19]	82.56	46.44	57.37	67.35	54.99

(e) MedleyDB

Table 3. Performance scores of the proposed model and other compared systems on five different test datasets. The proposed model, the hybrid model [10], and the MCDNN [14] were all trained using 740 and 200 clips of MIR-1K and iKala datasets. The rest clips of MIR-1K and iKala were used for the evaluation.

test. Note that there were no 2-D CNNs activated during this evaluation. Table 1 shows the three main metrics of the three compared system, CNN-960, CNN-64 and the Multi-CNN. The CNN-960 and CNN-64 refer to the 1-D CNN system with kernel lengths of 960 and 64, respectively, and the Multi-CNN refers to the system with these 2 paralleled 1-D CNNs. Clearly, the CNN-64 performs the worst since its short kernels couldn't provide high resolution at the low frequency region to resolve the fundamental frequency as shown in Fig. 2(a). However, it does provide complementary information to the CNN-960 system such as the harmonic structure in higher frequency region. Therefore, paralleling the two 1-D CNNs indeed improves the performance in terms of these three main metrics.

The second evaluation was to see whether the 2-D CNN can further provide complementary information from spectro-temporal features of the sound for melody extraction. As in the first evaluation, only the MIR-1K dataset was used. Table 2 demonstrates the 5 performance metrics of the three compared systems. The Multi-CNN system contained 2 paralleled 1-D CNNs as shown in Table 1. The proposed model cascaded the Multi-CNN system with three 2-D CNNs, while the Multi-CNN was initialized with properly ordered 1-D kernels from pre-training. The 'No pre-training' system refers to the proposed model but was randomly initialized. From this table, one can deduce that adding 2-D CNNs can further boost the overall performance a bit. In addition, spectro-temporal features extracted from the properly ordered spectrogram-like graph are effective in distinguishing melody frames from non-melody frames such that the proposed model has the lowest VFA rate.

The last evaluation was to see the general performance of the proposed model on different datasets. The 740 and 200 vocal clips from MIR-1k and iKala dataset were used for training the proposed model and the rest clips of these two datasets together with clips in ADC2004, MIREX05 and MedleyDB datasets were used for test. The performance of the proposed model for each test dataset are separately shown in Table 3 with the performance of compared methods, including the deep learning methods [10][14] and the expert system Melodia [19]. The numbers in boldface are the best scores in each of the 5 metrics. As shown in Table 3, the proposed model performs the best in terms of the OA score on MIR-1k, iKala, and MIREX05 datasets but not on ADC2004 and MedleyDB datasets. The reason is that the proposed model was trained using singing melody such that it probably couldnt detect instrumental melody very well. Interestingly, Melodia gets the best OA score on ADC2004. It seems that there are many opera songs in ADC2004 dataset such that Melodia is better than other compared methods in detecting pitch from fast-changing pitch motions such as the vibrato. Nevertheless, these results clearly demonstrate the proposed two-stage multi-resolution model produces remarkable results on singing melody extraction.

5. CONCLUSION

We built a two-stage multi-resolution CNN-based model for melody extraction. The proposed end-to-end model directly uses the time-domain polyphonic music signals for melody extraction such that pre-processing and post-processing are not needed. Two paralleled 1-D CNNs produce two spectrogram-like graphs in the multi-resolution fashion. Three sets of 2-D CNN kernels of different sizes encode 2-D spectro-temporal patterns from different scales. Experiment results show the proposed model does extract more information in distinguishing melody such that it produces better scores than most compared systems.

6. REFERENCES

[1] J. Salamon, E. Gómez, D. 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.

[2] 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.

[3] P. Foster, S. Dixon, and A. Klapuri, "Identifying cover songs using information-theoretic measures of similarity," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 993–1005, 2015.

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

[5] A. Jansson, E. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 323–332.

[6] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6964–6968.

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

[8] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *INTERSPEECH*, 2015, pp. 1–5.

[9] Z. Zhu, J. H. Engel, and A. Hannun, "Learning multiscale features directly from waveforms," 2016, arXiv:1603.09509.

[10] H. Chou, M.-T. Chen, and T.-S. Chi, "A hybrid neural network based on the duplex model of pitch perception for singing melody extraction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 381–385.

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

[12] T.-S. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1–9.

[14] S. Kum, C. Oh, and J. Nam, "Melody extraction on vocal segments using multi-column deep neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval (ISMIR)*, 2016, pp. 819–825.

[15] C.-L. Hsu and J.-S. Roger Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[16] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 718–722.

[17] 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 *Proceedings of the 15th International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 66–70.

[18] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the 15th International Society for Music Information Retrieval (ISMIR)*, 2014.

[19] 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.

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

[21] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2158–2168, 2014.