CNN Based Two-Stage Multi-Resolution End-to-End Model for Singing Melody Extraction

Ming-Tso Chen, Bo-Jun Li, and Tai-Shih Chi, National Chiao Tung University, Hsinchu, Taiwan

Introduction

- Fourier spectrogram uniformly depicts the sound using a particular temporal and spectral resolution.
- The proposed model analyzes the joint spectro-temporal patterns of the sound at various resolutions to decipher pitch.
- The first stage is implemented using the 1-D CNN to similarly behave as a spectrum estimator.
  The second stage is 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 use two 1-D CNNs, whose kernels are with different lengths, in parallel in the first stage.

Pre-training and Experiment setting(1)

Proposed model is an end-to end model and exhibits random permutation on the kernel-index axis according to the learned weights in the 1st stage. We pre-trained a model consisting of only 1-D CNN.

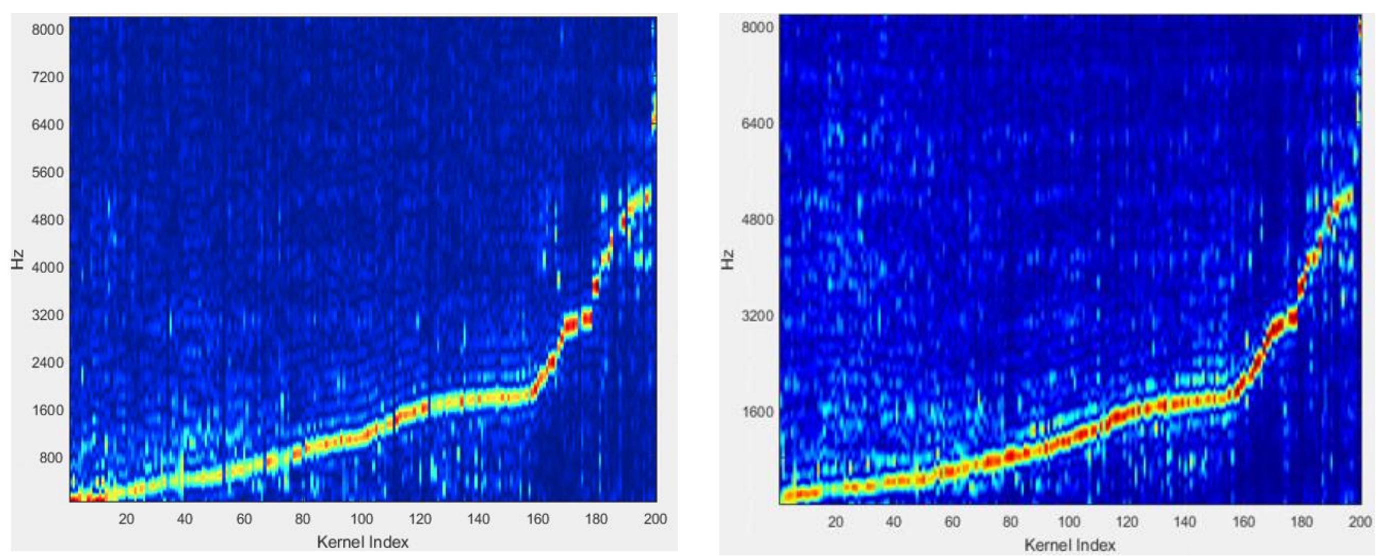


Fig.2. Magnitude response of 1-D CNN with kernel length 64 (left) initial weights from pre-training, (right) the final weights

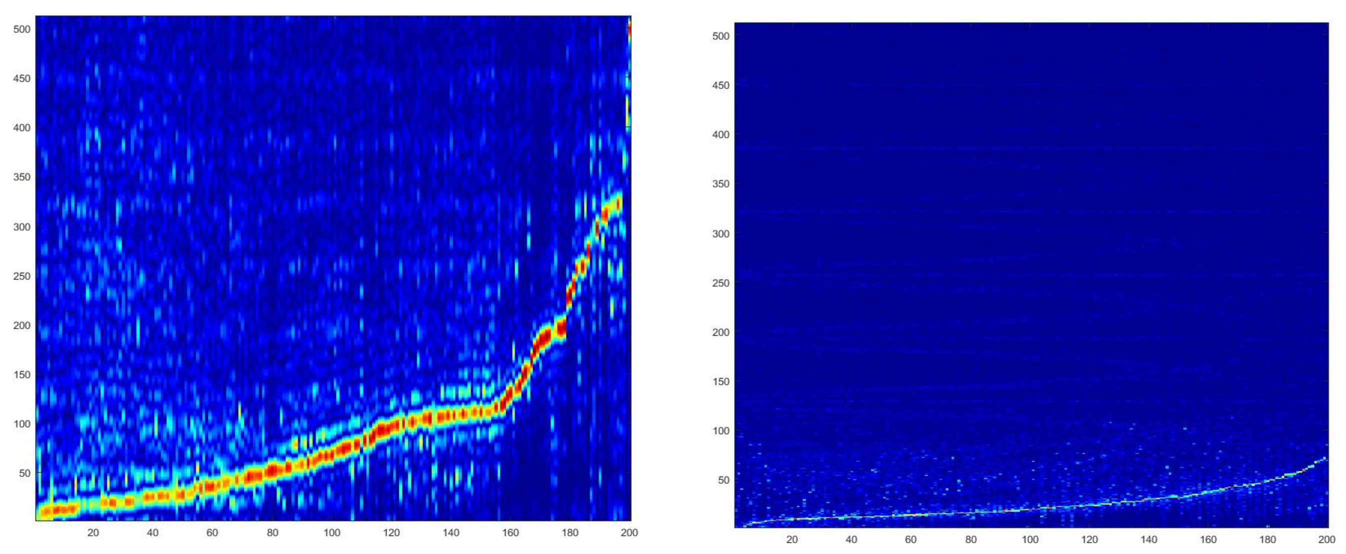


Fig.3. Magnitude response of 1-D CNN (left) kernel length 64, (right) kernel length 960

Architecture

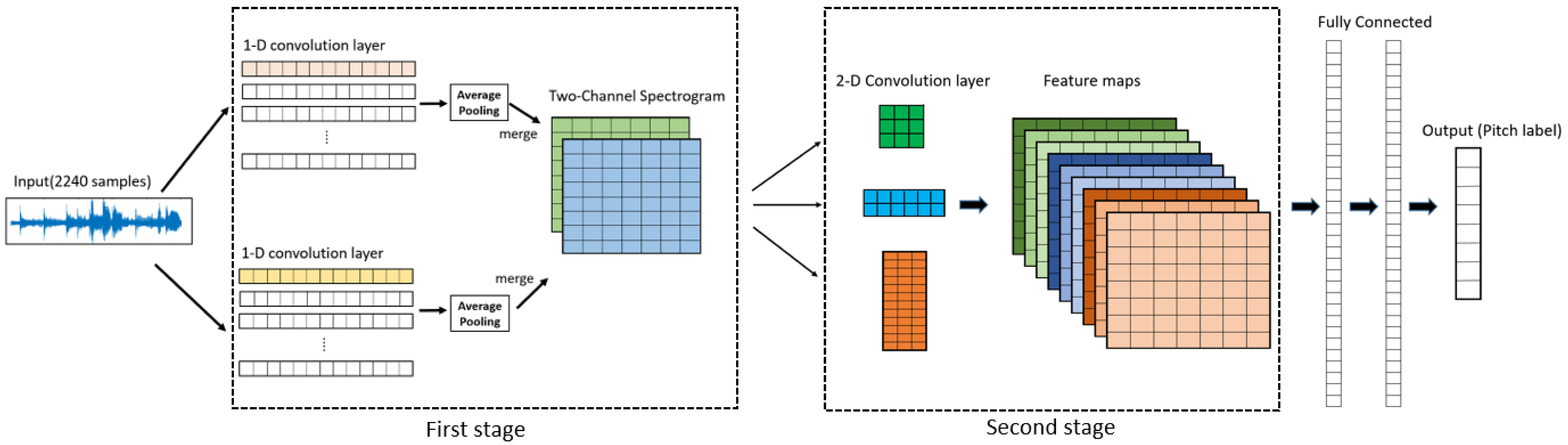


Fig.1. The proposed architecture

- The first stage consists of two paralleled 1-D CNNs with kernels of different lengths which can thought as impulse responses of filters, determines the frequency bandwidth of the analysis bands.
- The 'Inception' module is used to expand the width of the model to simulate multi-resolution analysis on the graph using 2-D kernels with different sizes which extract useful spectro-temporal patterns, which might include the harmonic structure, temporal continuity, and other melody related patterns.

Pre-training and Experiment setting(2)

- The proposed second stage will automatically produce suitable 2-D kernels with pre-training strategy.

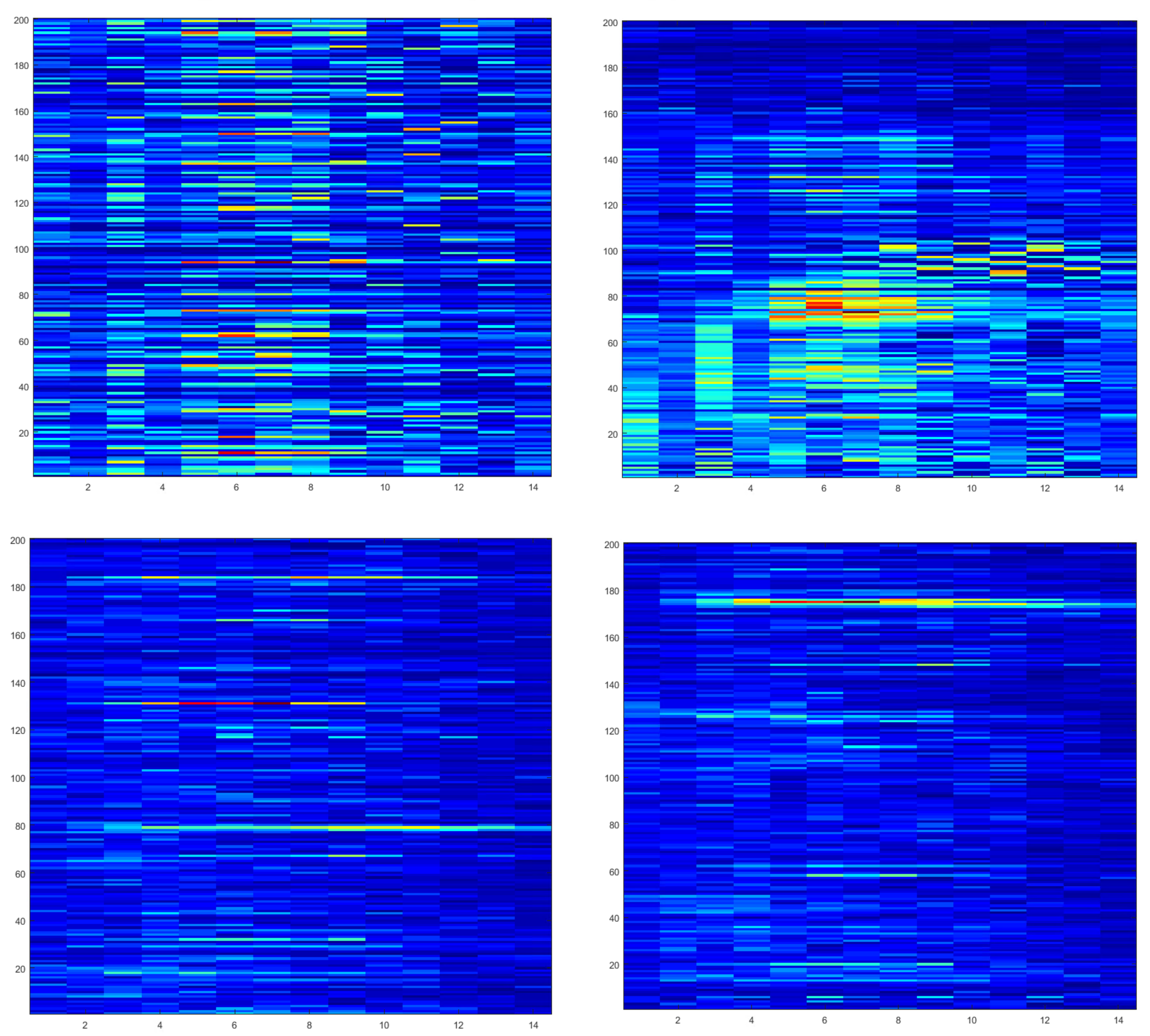
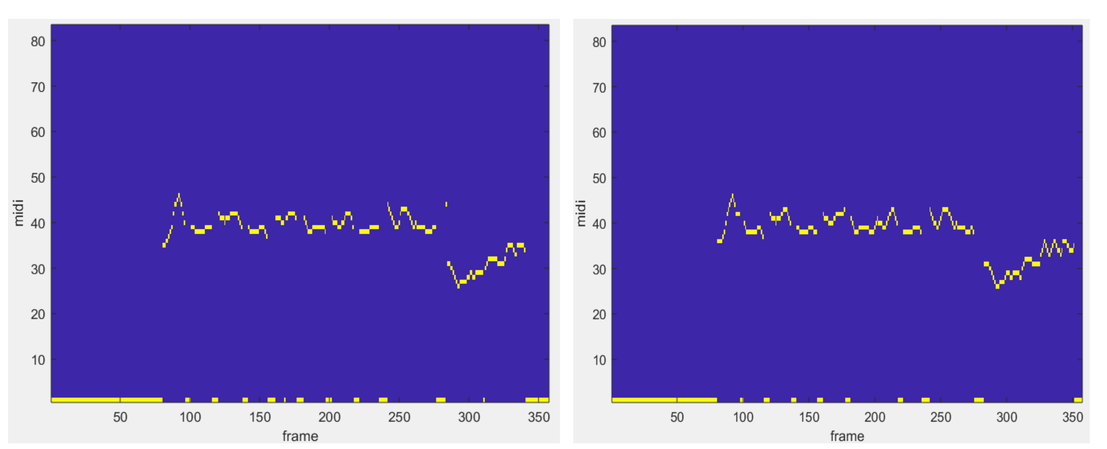


Fig.4. The spectrogram like graph with different kernel and with/without pre-training method

Result

- 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 couldn't detect instru-mental melody very well.

Fig.5. Midi pitch (left) predict (right) truth



	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

	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