VOCAL MELODY EXTRACTION USING PATCH-BASED CNN

Li Su
Institute of Information Science, Academia Sinica, Taiwan

Vocal melody extraction as an object detection and localization problem

- Singing voice: a distinct object with vibrato and sliding behaviors different from other instruments
- Vocal melody extraction: a classification task for classifying whether a time-frequency patch contains a vocal event
- **Problem: how to localize a pitch object?**
- A localization task for simultaneously performing vocal activity detection (VAD) and pitch detection (i.e., suppress unwanted harmonics)

Data representation: combining frequency and periodicity (CFP)

- A pitch object is determined by:
  - A frequency-domain representation indicating its fundamental frequency (f_0) and harmonics (nf_0)
  - A time-domain representation revealing its f_0 and sub-harmonics (f_0/n)

- $\mathbf{x}$: frame-level signal, $\mathbf{F}$: discrete Fourier transform
  - Magnitude spectrum: $\mathbf{z}_0[k] := |\mathbf{Fx}|$
  - Generalized cepstrum (GC): $\mathbf{z}_1[n] := \mathbf{F}^{-1}\sigma_1(|\mathbf{Fx}|)$
  - Generalized cepstrum of spectrum (GCoS): $\mathbf{z}_2[k] := \mathbf{F}\sigma_2\left(\mathbf{F}^{-1}\sigma_1(|\mathbf{Fx}|)\right)$ [Su, 2017]
- Power-scale activation functions: $\sigma_2(z) := z^{0.6}$, $\sigma_1(z) := z^{0.24}$
- Note: a high-pass filter is required in each step for extracting pitch information

- Map $\mathbf{z}_1[n]$ and $\mathbf{z}_2[k]$ into the log-frequency scale: $\tilde{\mathbf{z}}_1[p]$ and $\tilde{\mathbf{z}}_2\,[p]$ from E2 to G5, 48 bands per octave
- Pitch object localization: multiplying $\tilde{\mathbf{z}}_1[p]$ and $\tilde{\mathbf{z}}_2\,[p]$ by the time-domain representation can effectively suppress the harmonic and subharmonic peaks

$$\mathbf{y}[p] = \tilde{\mathbf{z}}_1[p]\,\tilde{\mathbf{z}}_2\,[p]$$

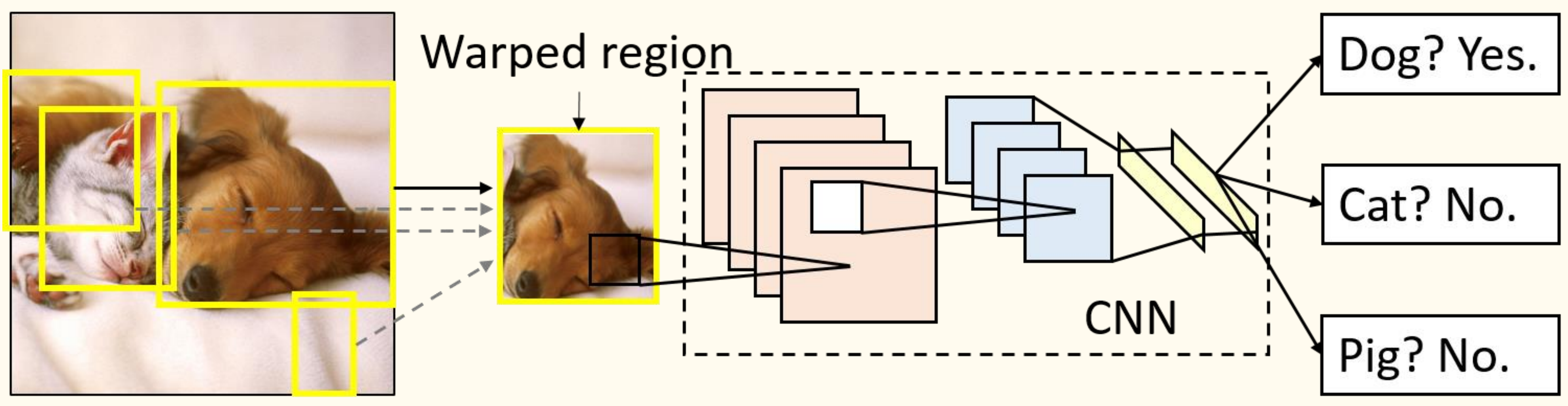
- Perform vocal-nonvocal classification simply by using a localized pitch contour

Experiment settings and data

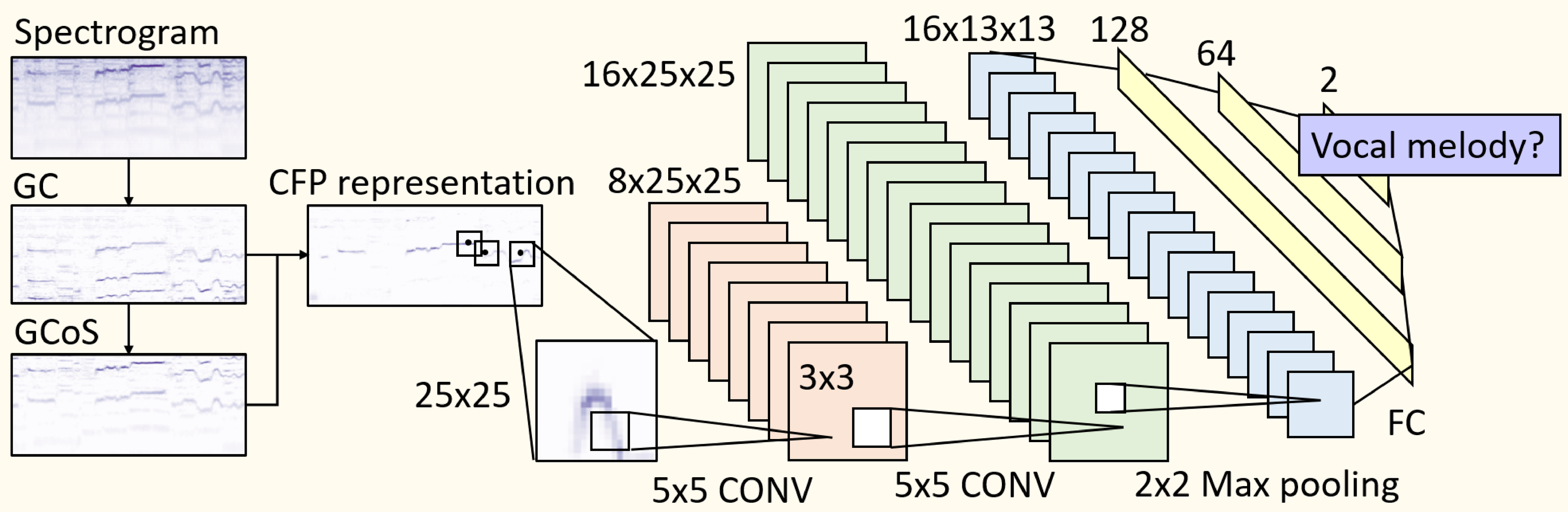
- **CFP-Max:** directly employing the CFP representation by simply taking the pitch index corresponding to the maxima of the frame
- **CNN-MaxIn:** from patches having an output probability > 0:5, taking the frequency index where the CFP representation reaches its maximum
- **CNN-MaxOut:** taking the frequency index corresponding to the largest output probability

The paradigm of object detection and localization: object proposals + CNN

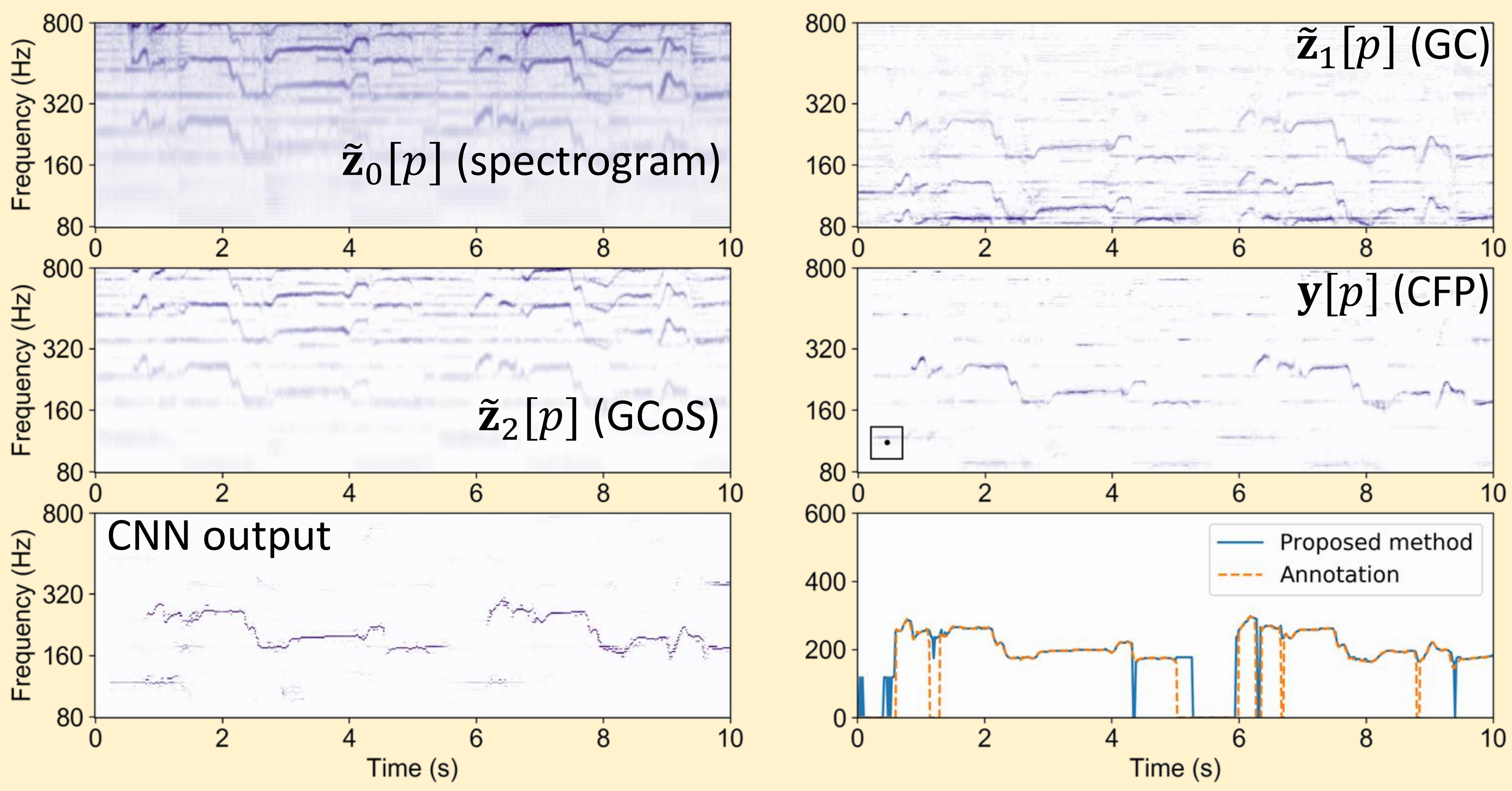
- Convolutional neural networks (CNN)
- R-CNN for image object detection [Girshick *et al.*, 2014]:



- Patch-based CNN for vocal melody extraction:



	Feature	Localization	Classification
R-CNN for object detection	CNN	Selective search	Support vector machine (SVM)
Patch-based CNN for vocal melody detection	CFP	CFP and peak picking	CNN + fully connected layers



Testing data	ADC2004 (vocal)					MIREX05 (vocal)				
Method	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
CFP-Max	61.2	71.7	76.8	-	-	46.3	70.7	75.5	-	-
CNN-MaxIn	**74.3**	**76.7**	**78.4**	90.1	41.3	73.2	81.2	82.2	**95.1**	**41.1**
CNN-MaxOut	72.4	74.7	75.7	90.1	41.3	74.4	**83.1**	**83.5**	**95.1**	**41.1**
DSM (th=0.3)	68.0	68.4	70.9	78.2	**25.5**	**76.3**	70.4	71.2	80.1	13.6
DSM (th=0.1)	70.8	77.1	78.8	**92.9**	50.5	69.6	76.3	77.3	93.6	42.8
Testing data	iKala					MedleyDB (vocal)				
Method	OA	RPA	RCA	VR	VFA	OA	RPA	RCA	VR	VFA
CFP-Max	46.9	69.7	72.6	-	-	38.3	55.6	62.4	-	-
CNN-MaxIn	74.3	76.5	77.8	**94.2**	33.0	54.7	58.7	63.6	78.4	55.1
CNN-MaxOut	**74.6**	**76.9**	**77.7**	**94.2**	33.0	55.2	59.7	63.8	78.4	55.1
DSM (th=0.3)	72.7	67.9	69.7	82.4	**18.8**	**66.7**	**61.7**	**64.7**	70.2	**21.9**
DSM (th=0.1)	67.4	73.4	74.6	94.1	46.6	66.2	72.0	74.8	**88.4**	48.7