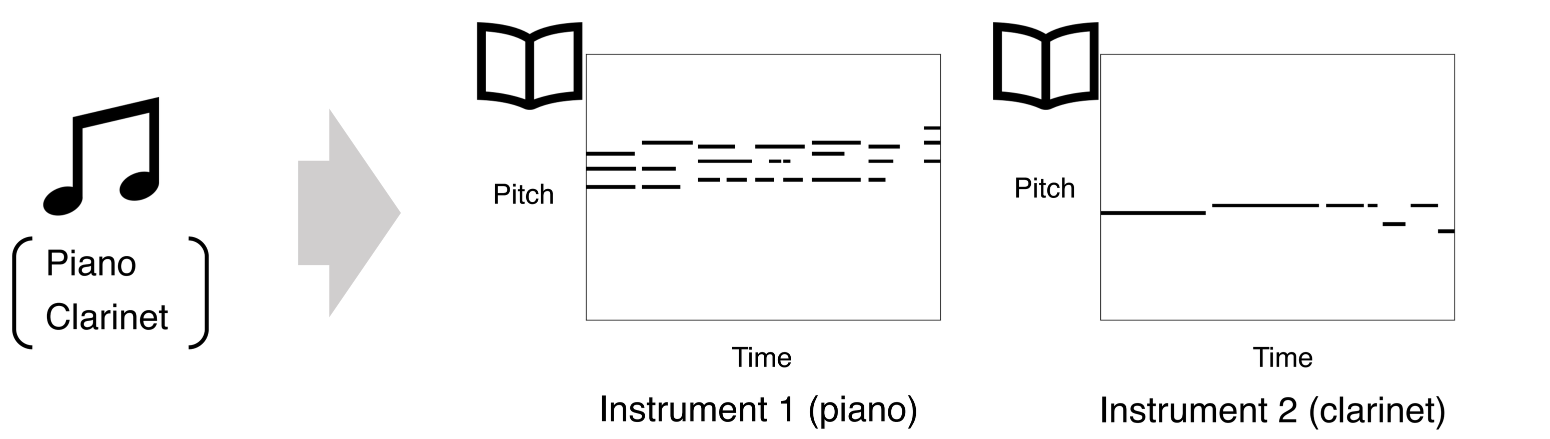
Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams

Keitaro Tanaka[1] Takayuki Nakatsuka[1] Ryo Nishikimi[2] Kazuyoshi Yoshii[2] Shigeo Morishima[3]

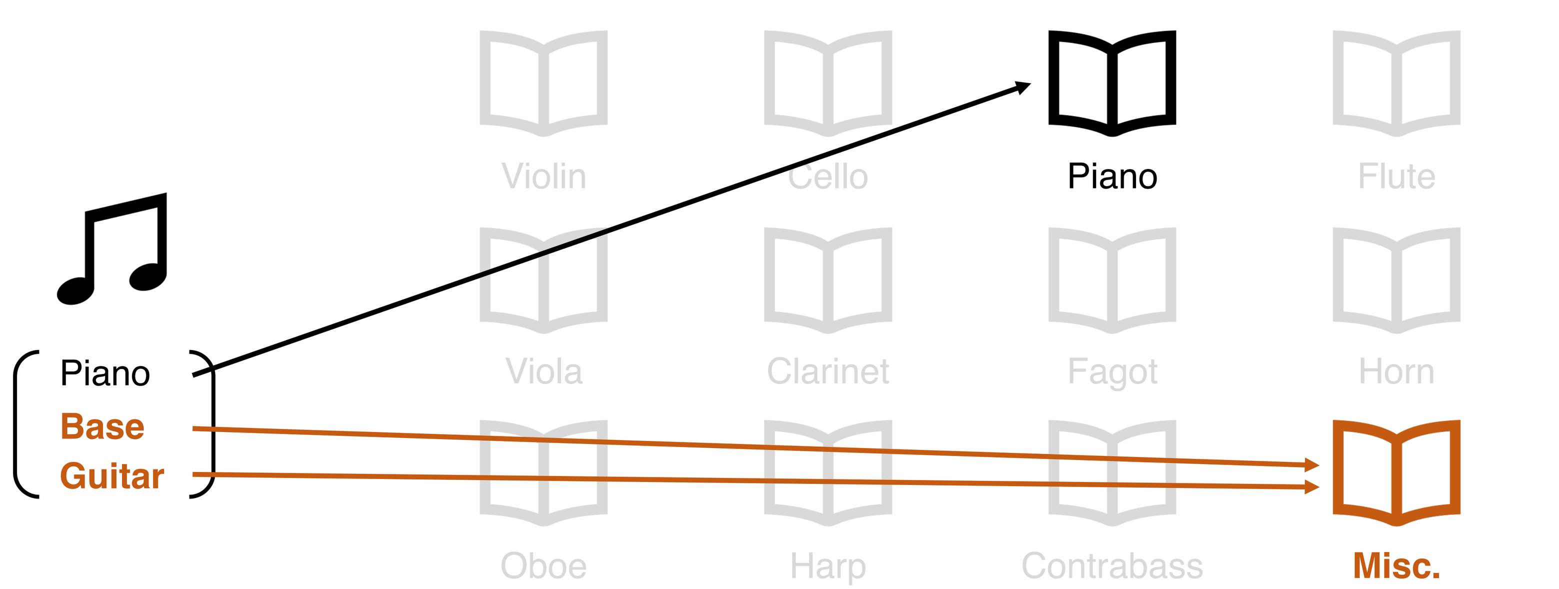
[1] Waseda University [2] Kyoto University [3] Waseda Research Institute for Science and Engineering, Japan

Contact (Author's Order): phys.keitaro1227@ruri.waseda.jp
t59nakatsuka@fuji.waseda.jp
nishikimi@sap.ist.i.kyoto-u.ac.jp
yoshii@kuis.kyoto-u.ac.jp
shigeo@waseda.jp

Backgrounds

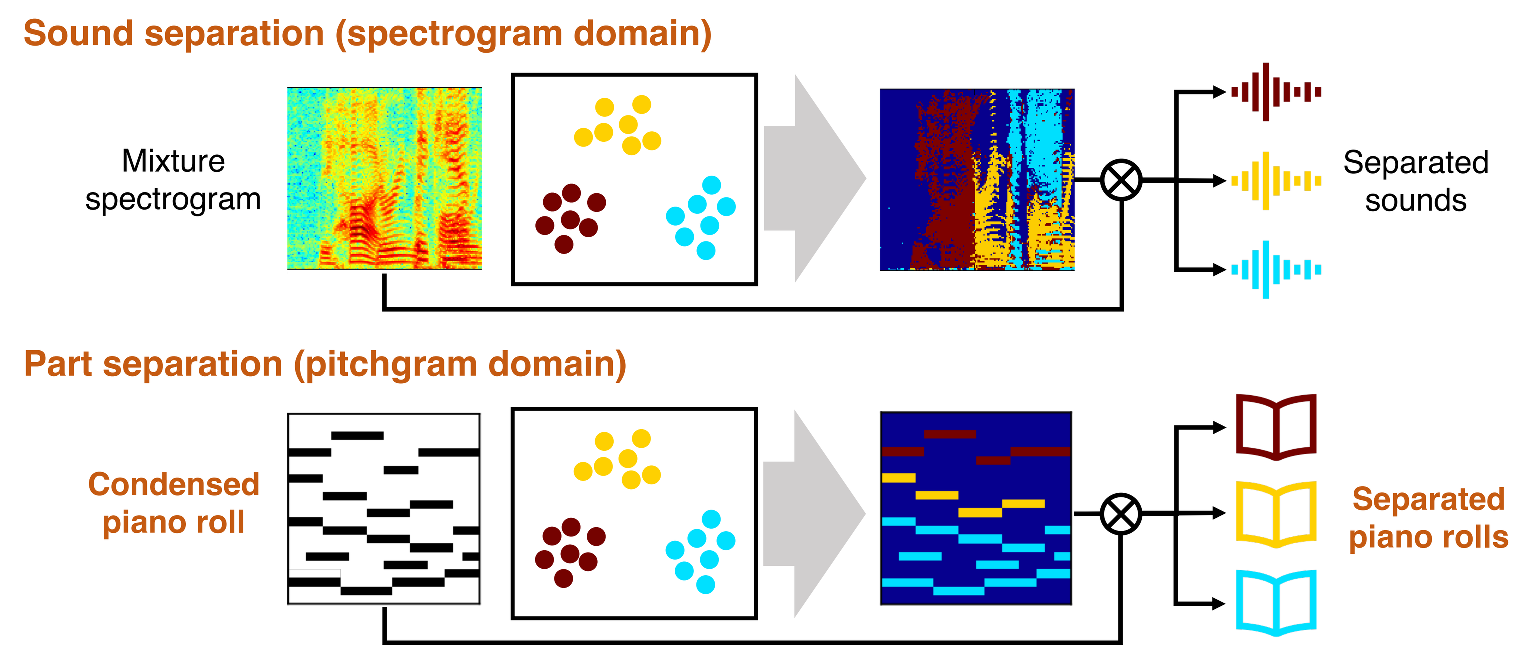


- Transcribe arbitrary musical instruments
- Deal with music signals played by any harmonic instruments
- Specify the number of instruments in advance
- Estimate piano rolls of multiple instrument parts

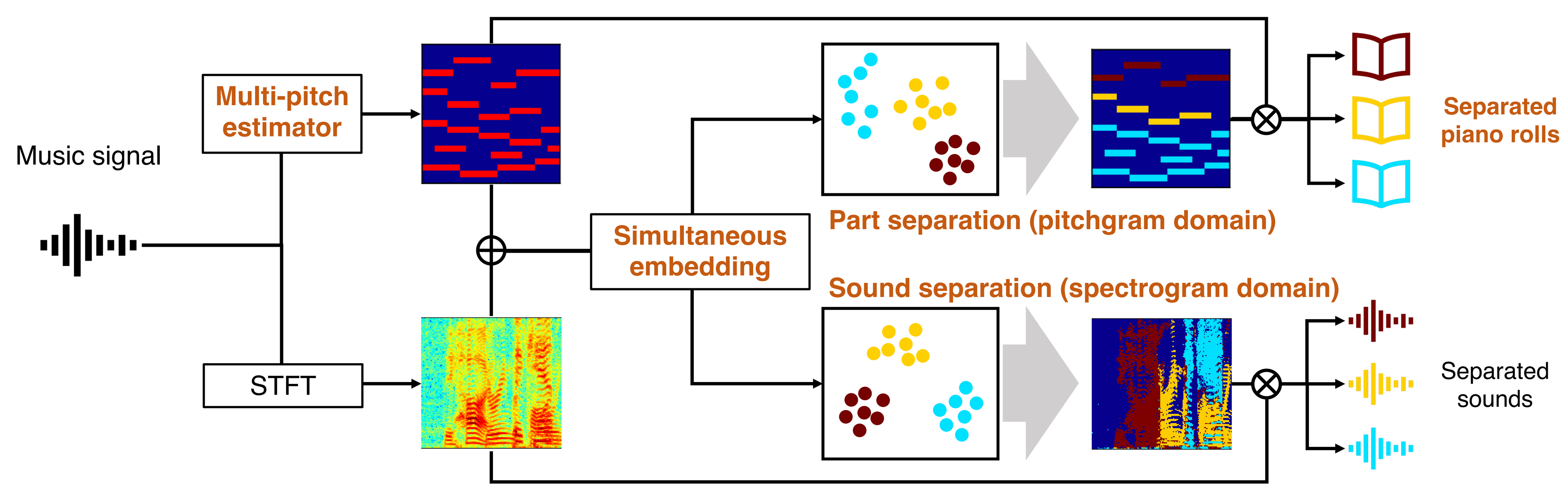


- Most conventional methods based on supervised learning of DNNs can deal with only predefined instruments included in training data
- Thus, it is impossible to transcribe undefined instruments that are not included in the training data

Method (Key Ideas)



- Deep clustering separates a speech mixture to an arbitrary number of speakers based on the characteristics of voices
- Application of deep clustering to part separation in the task of multi-instrument music transcription

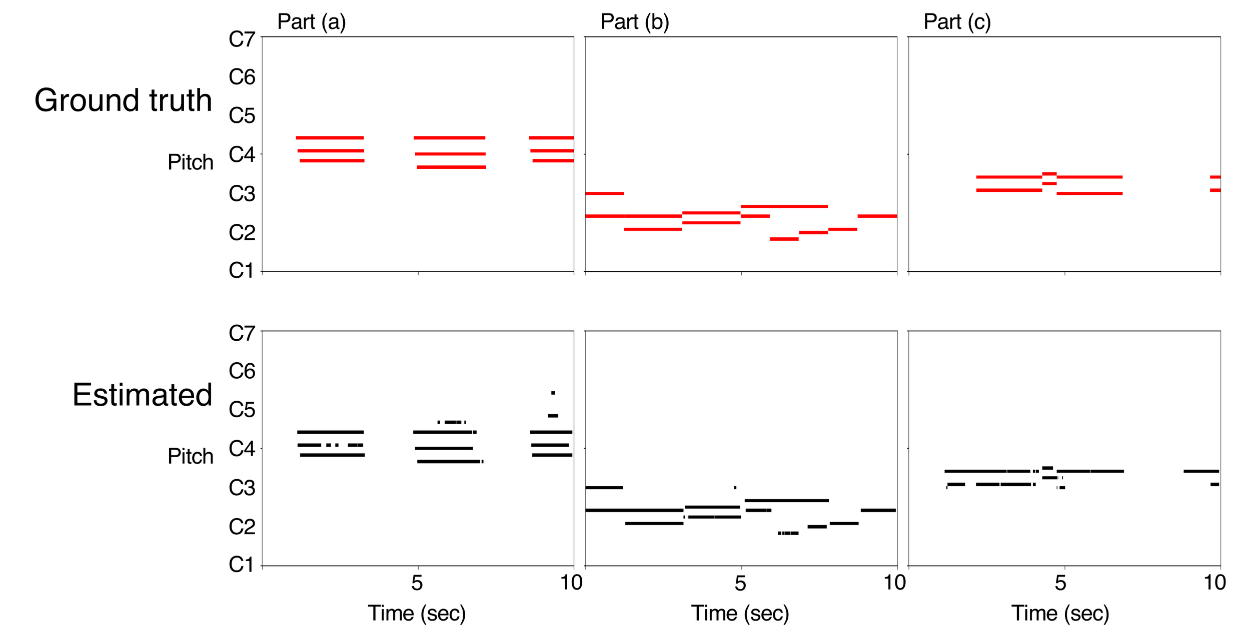


- Make effective use of complementary relationships between part separation in two domains by joint part separation
- Use a pitchgram as a proxy of a condensed piano roll
- Estimate the pitchgram using an existing multi-pitch estimator

Results

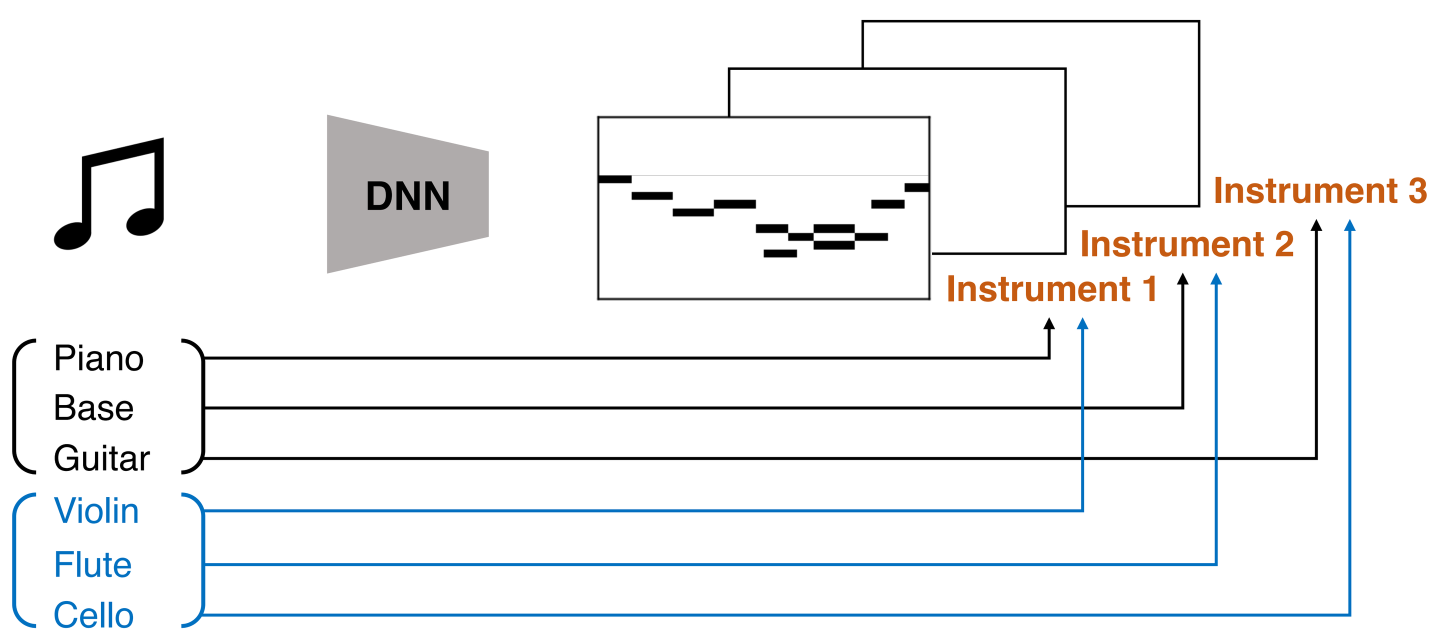
Instrument	Closed condition						Open condition					
	Wu+@ICASSP2019			Proposed			Wu+@ICASSP2019			Proposed		
	P	R	F	P	R	F	P	R	F	P	R	F
Piano	51.28	46.50	45.87	62.02	39.61	44.07	52.51	48.04	47.37	61.87	38.90	43.64
Base	73.75	58.79	64.04	39.72	50.78	42.24	74.27	59.66	64.67	40.59	51.88	43.23
Guitar	46.64	36.72	37.69	52.91	35.45	39.46	44.59	37.12	37.25	53.45	36.50	40.32
Strings	55.27	56.79	52.74	66.35	48.74	52.40	53.21	56.97	52.05	65.31	48.40	52.04
Synth pad	43.72	44.80	42.07	49.65	35.12	38.70	44.42	46.89	43.91	51.99	36.58	40.81
Reed	28.53	33.90	29.27	29.87	37.37	31.53	26.92	31.72	27.53	28.87	35.46	30.04
Brass	35.24	25.12	24.50	37.10	30.23	29.53	37.66	25.67	25.89	36.78	30.64	30.26
Organ	-	-	-	-	-	-	20.14	19.01	16.89	36.62	28.57	29.11
Pipe	-	-	-	-	-	-	22.62	27.13	23.02	38.37	39.49	35.22
Synth lead	-	-	-	-	-	-	20.58	17.44	17.59	29.41	25.11	24.98

- The proposed method can transcribe undefined instruments as well as predefined instruments used for training
- Accuracies on undefined instruments have improved in the proposed method



- The proposed method can successfully achieve multi-instrument music transcription
- Note that every part is played by a polyphonic instrument

Approach



- Use a DNN capable of timbre-based clustering
- Specify the number of instruments at run-time

Method (Special Notes)

- Overall optimization after training each part

$\boldsymbol{X}^{\mathrm{pi}}$	ground truth condensed pitchgram
$\widehat{\boldsymbol{X}}^{\mathrm{pi}}$	estimated condensed pitchgram
$\boldsymbol{V}^{\mathrm{pi,ti}}$	two latent spaces for a pitchgram and a spectrogram
$\widehat{\boldsymbol{M}}^{\mathrm{pi,ti}}$	two correct masks for a pitchgram and a spectrogram
$\alpha,\ \beta$	parameters to decide the weights of two losses

- Train the multi-pitch estimator

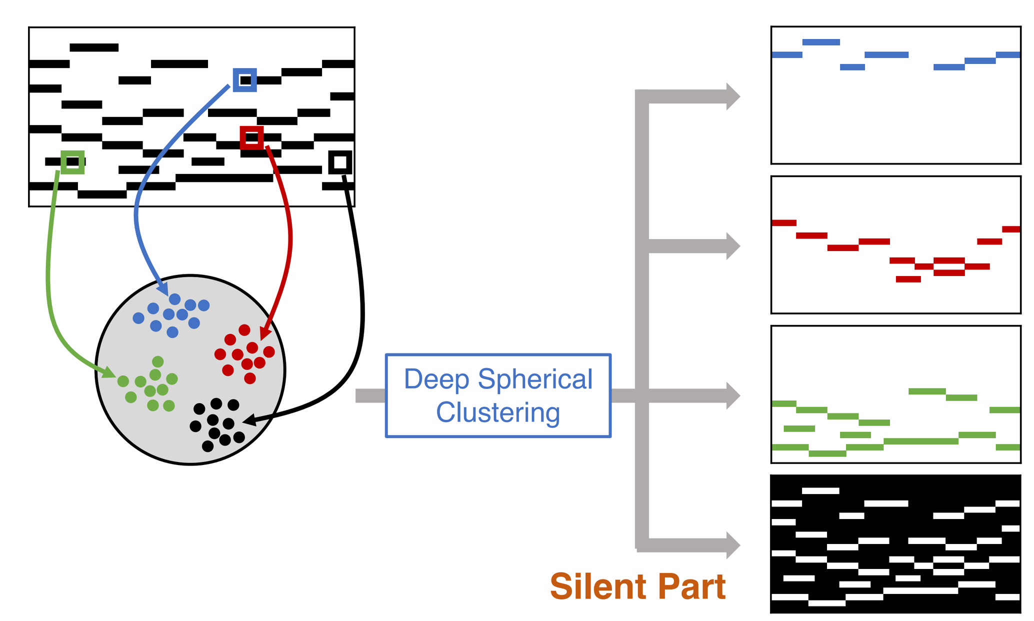
$$\mathcal{L}_{DS} = -\widehat{\boldsymbol{X}}^{\mathrm{pi}} \log\left(\boldsymbol{X}^{\mathrm{pi}}\right) - (1 - \widehat{\boldsymbol{X}}^{\mathrm{pi}}) \log\left(1 - \boldsymbol{X}^{\mathrm{pi}}\right)$$

- Train the simultaneous embedding

$$\mathcal{L}_{DC}{}^{\mathrm{pi,ti}} = \left\| \boldsymbol{V}^{\mathrm{pi,ti}} \boldsymbol{V}^{\mathrm{pi,ti}^T} - \widehat{\boldsymbol{M}}^{\mathrm{pi,ti}} \widehat{\boldsymbol{M}}^{\mathrm{pi,ti}^T} \right\|_F^2$$

- Optimize the whole network

$$\mathcal{L} = \mathcal{L}_{DS} + \alpha \mathcal{L}_{DC}{}^{\mathrm{pi}} + \beta \mathcal{L}_{DC}{}^{\mathrm{ti}}$$



- A pitchgram contains false estimates
- Prepare a silent part in separating it

Conclusions & Future work

- Multi-instrument music transcription method based on deep clustering
- The pitchgram and spectrogram are jointly embedded into features spaces
- k-means clustering with a specified number of instruments is conducted
- Undefined instruments can be dealt with as well as predefined instruments
- Explore other timbre representations as alternatives to the spectrogram

References

- "Polyphonic Music Transcription with Semantic Segmentation" (Wu et al., ICASSP 2019)
- "Deep Clustering: Discriminative Embeddings for Segmentation and Separation" (Hershey et al., ICASSP 2016)
- "Deep Salience Representations for F0 Estimation in Polyphonic Music" (Bittner et al., ISMIR 2017)

Acknowledgement

This paper is supported by the JST ACCEL Grant Number JPMJAC1602, JSPS KAKENHI Grant Number JP16H01744, and JP19H04137.