#5829

Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music

Sangeun Kum[1], Jongpil Lee[1], Keunhyoung Luke Kim[1], Taehyoung Kim[1], Juhan Nam[2]

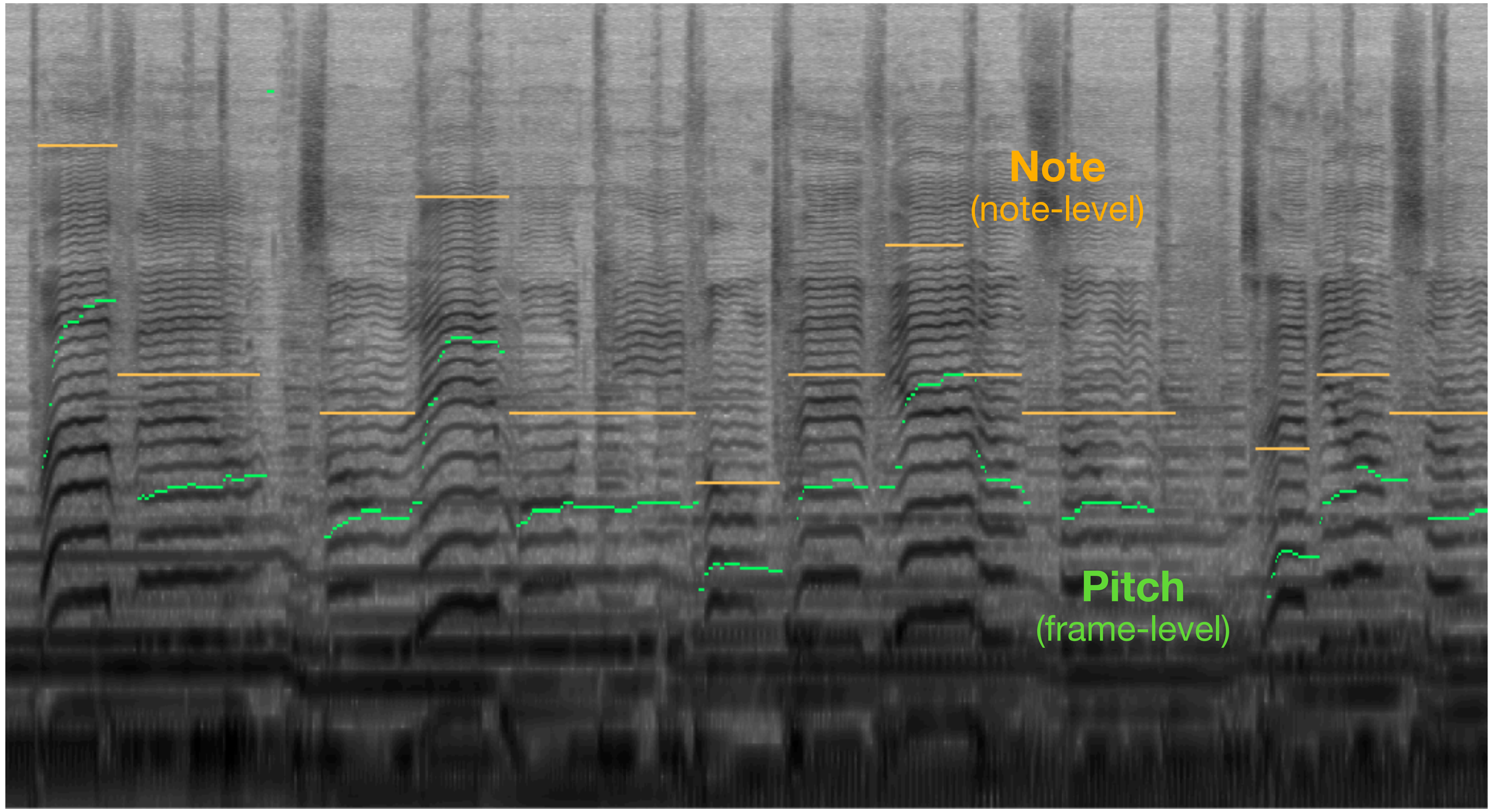
[1] Neutune Research, Seoul, South Korea
[2] Graduate School of Culture Technology, KAIST, Daejeon, South Korea

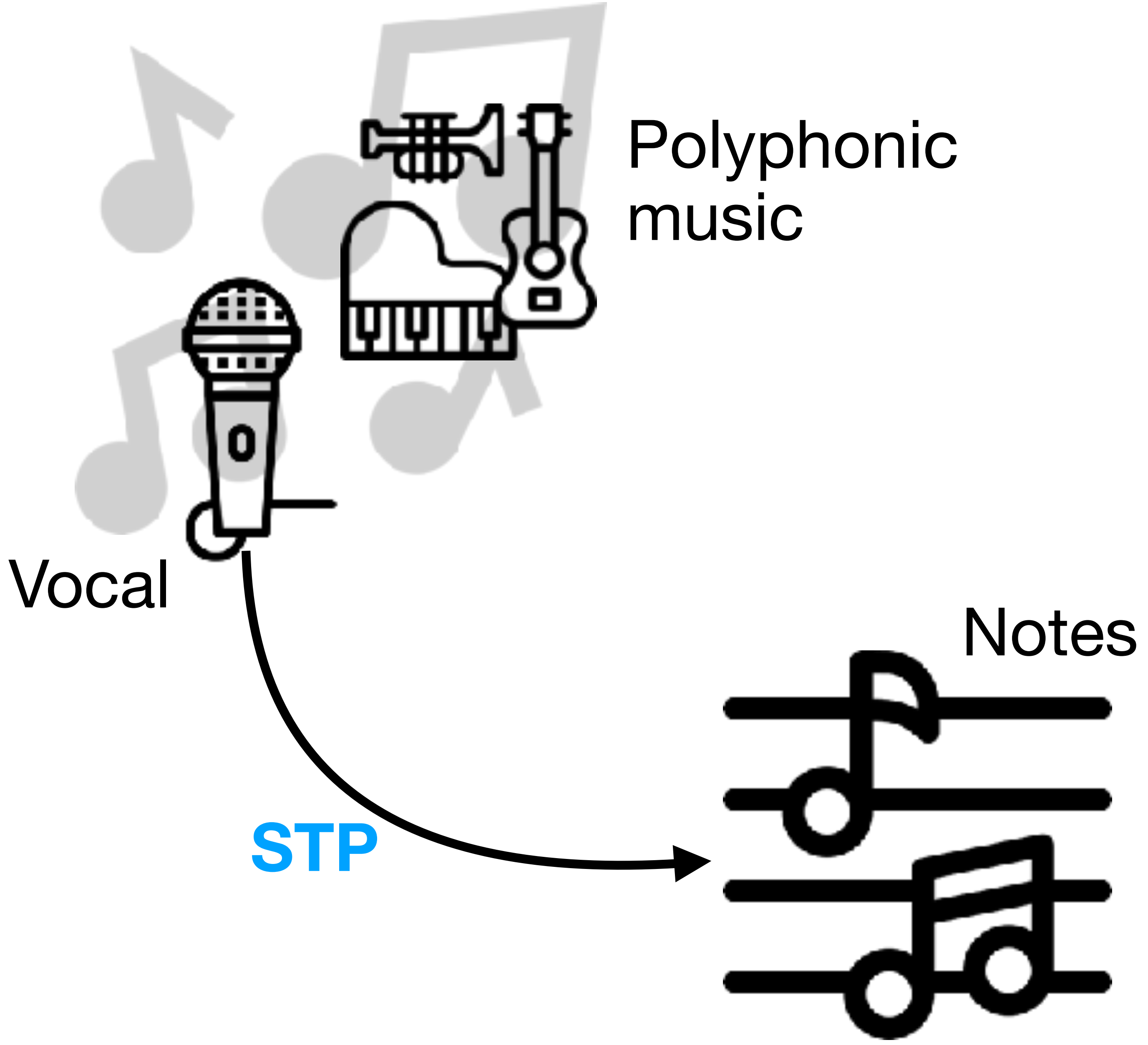
icassp 2022 Singapore

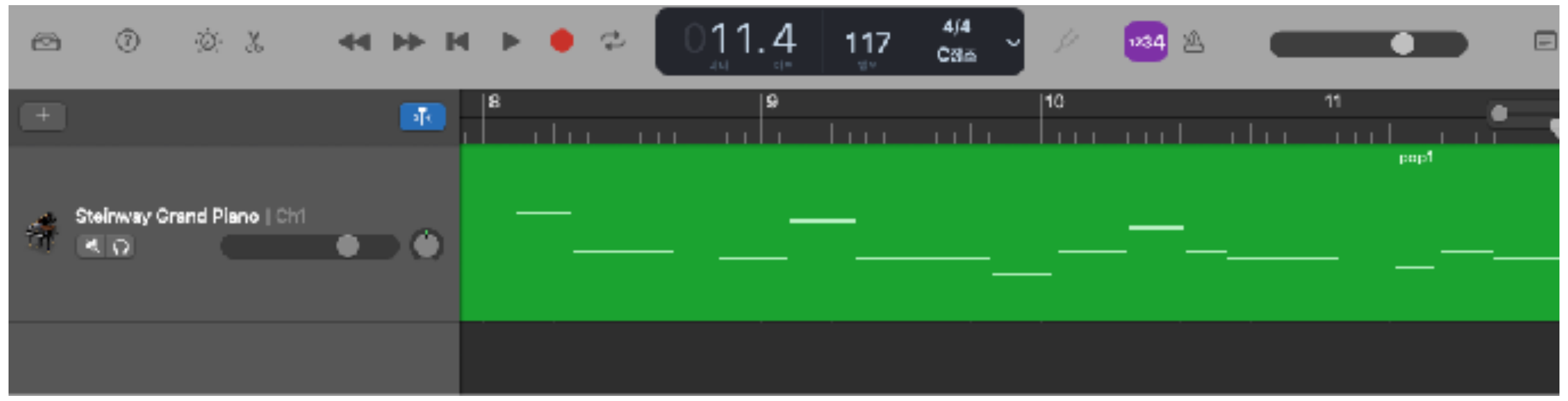
I Singing Transcription from Polyphonic Music

: Definition

- The goal of **Singing Transcription from Polyphonic Music** (**STP**) is to **transcribe** the vocal part of **polyphonic** music into a **series of note.**

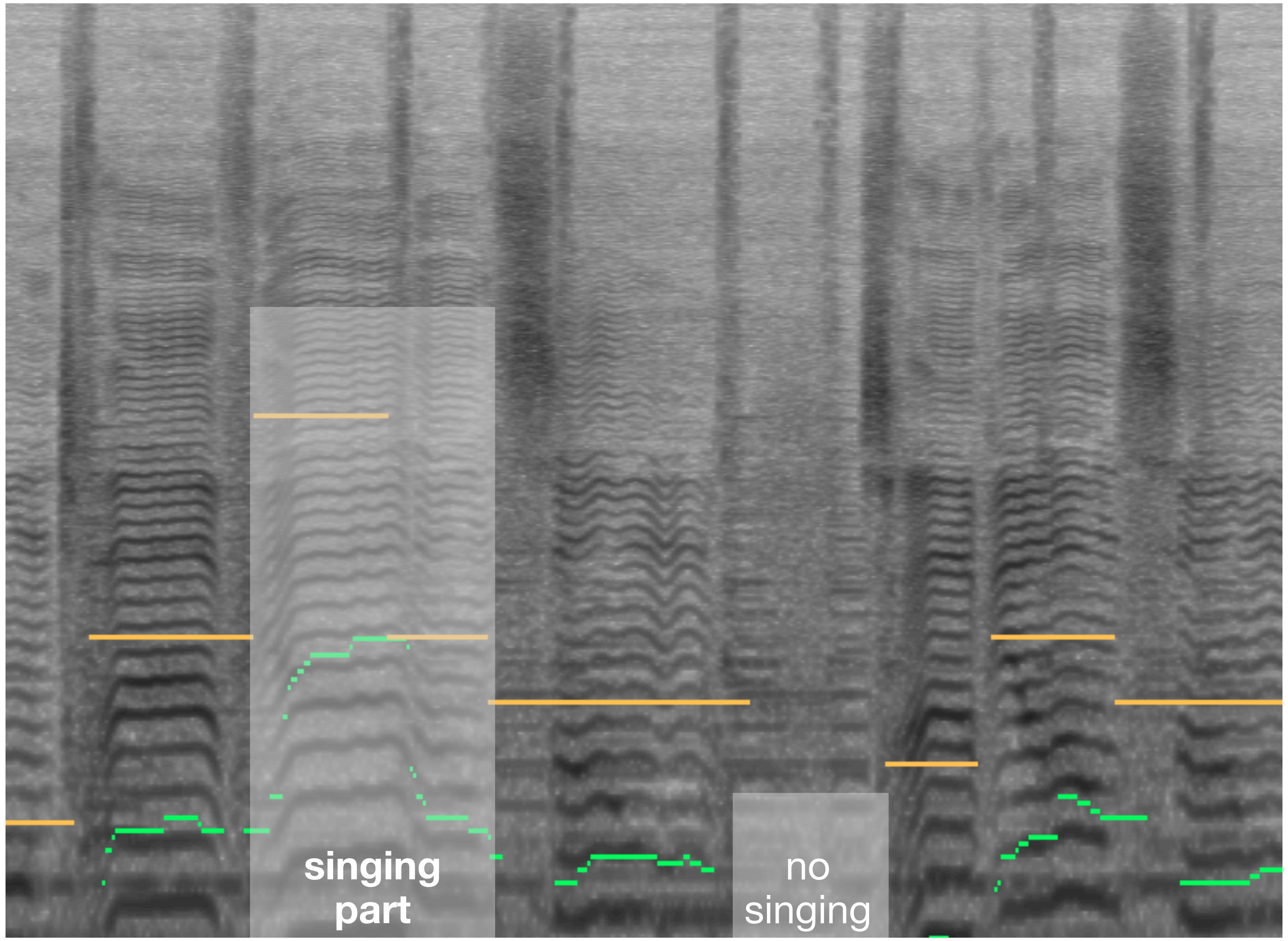






I Singing Transcription from Polyphonic Music

: Motivation

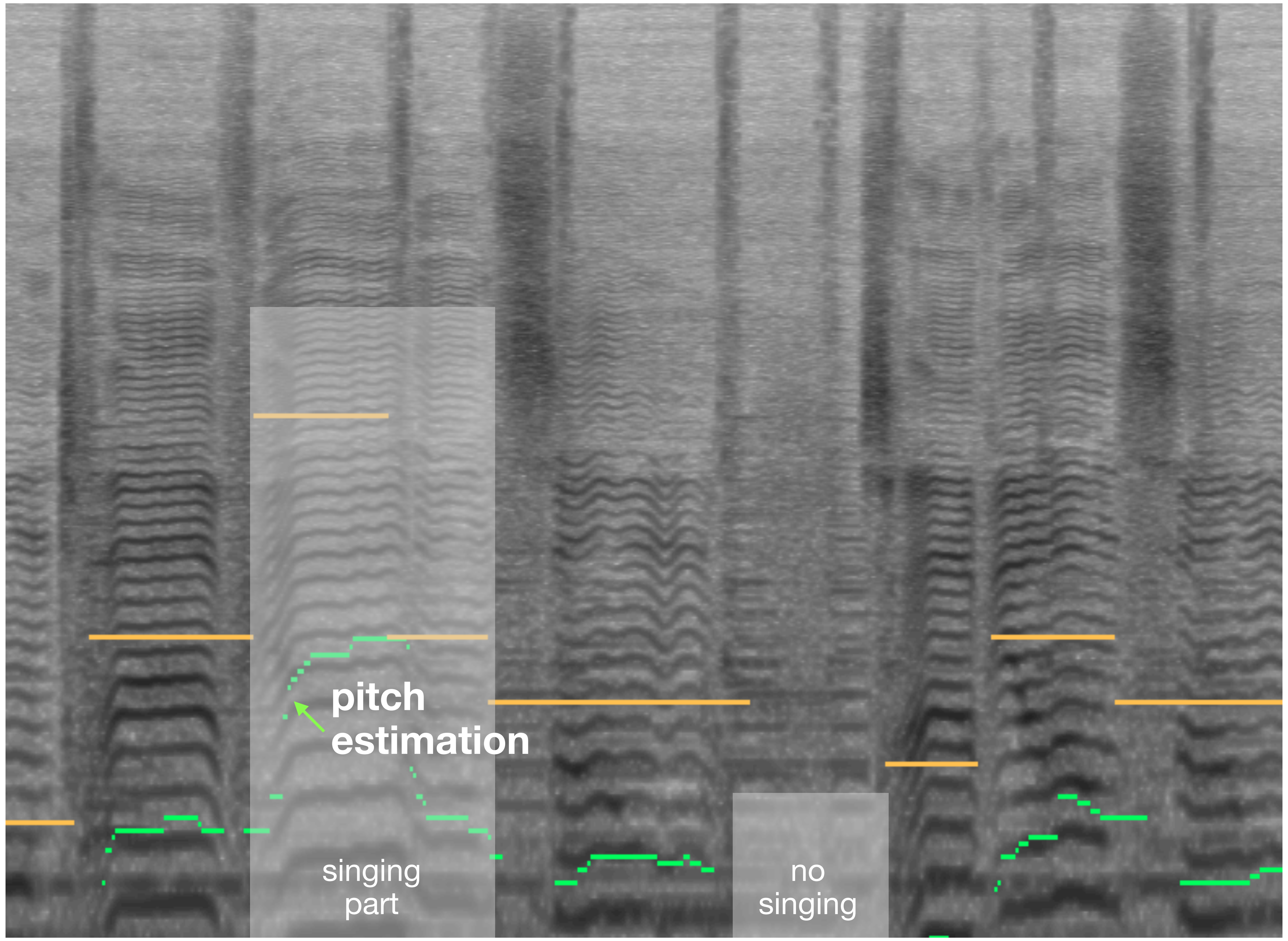


STP includes several sub-task:

1. **Singing voice detection**
2. Singing pitch estimation
3. Note-level segmentation
4. Onset/offset detection

I Singing Transcription from Polyphonic Music

: Motivation

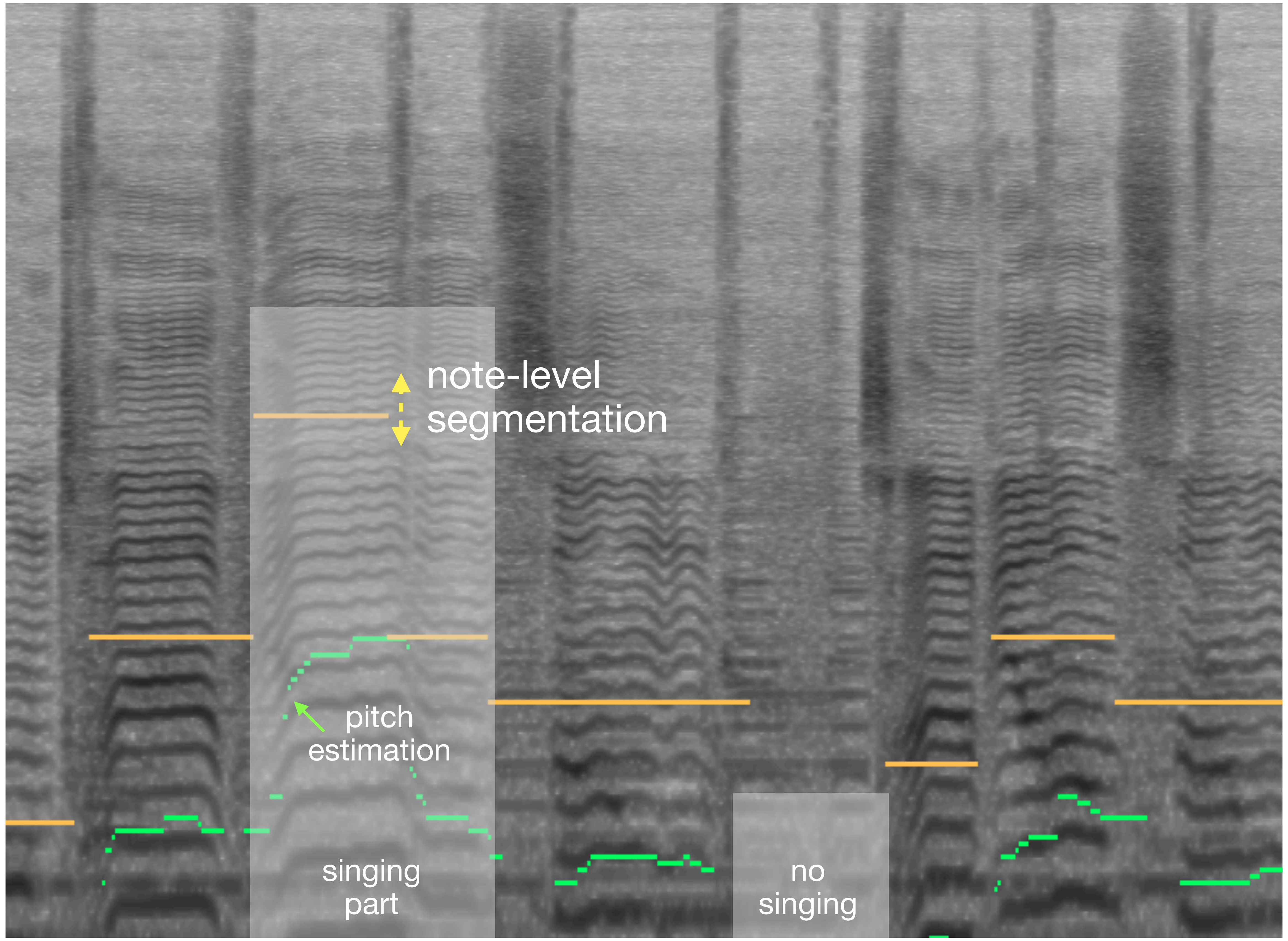


STP includes several sub-task:

1. Singing voice detection
2. **Singing pitch estimation**
3. Note-level segmentation
4. Onset/offset detection

I Singing Transcription from Polyphonic Music

: Motivation

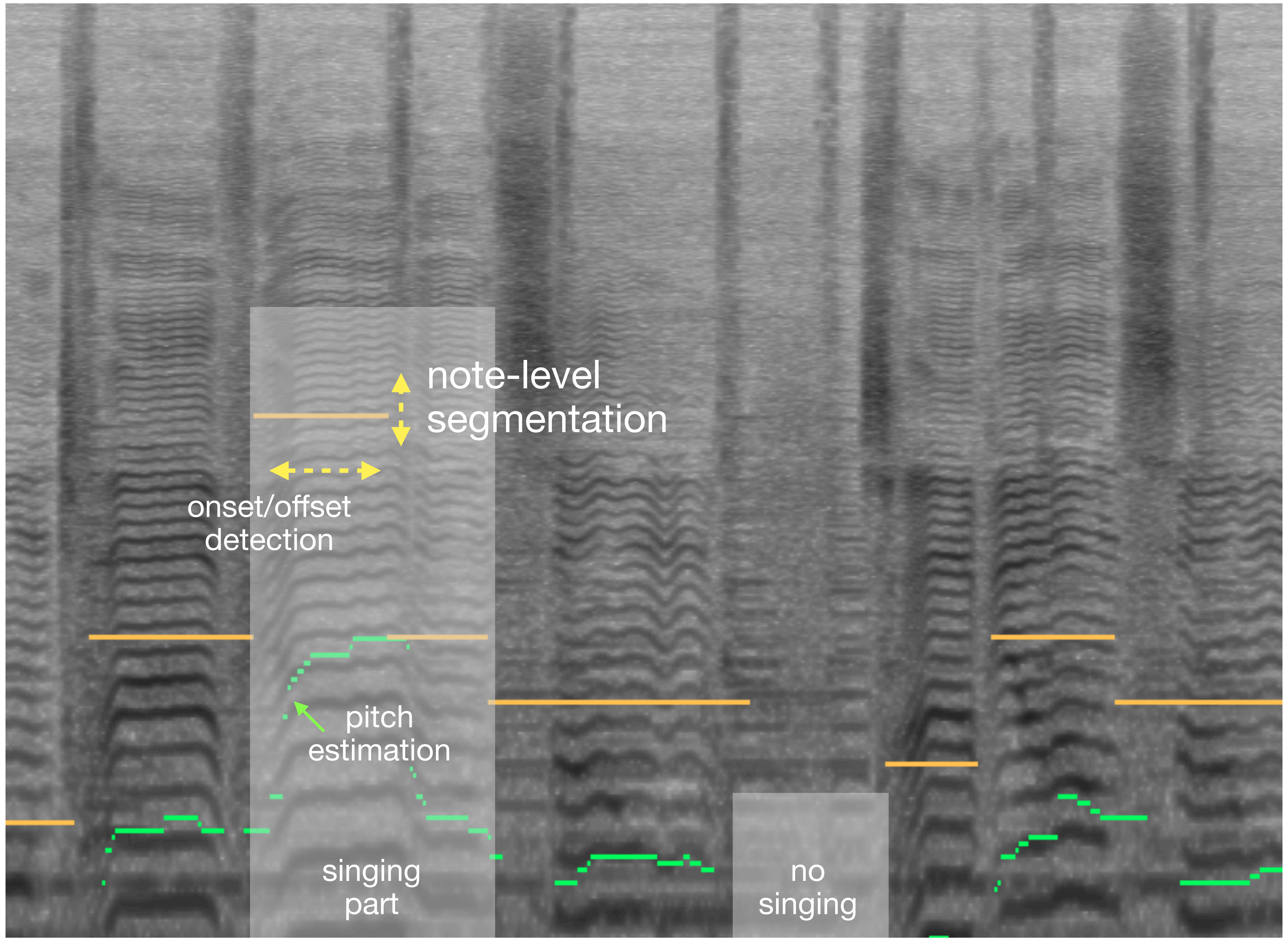


STP includes several sub-task:

1. Singing voice detection
2. Singing pitch estimation
3. **Note-level segmentation**
4. Onset/offset detection

I Singing Transcription from Polyphonic Music

: Motivation

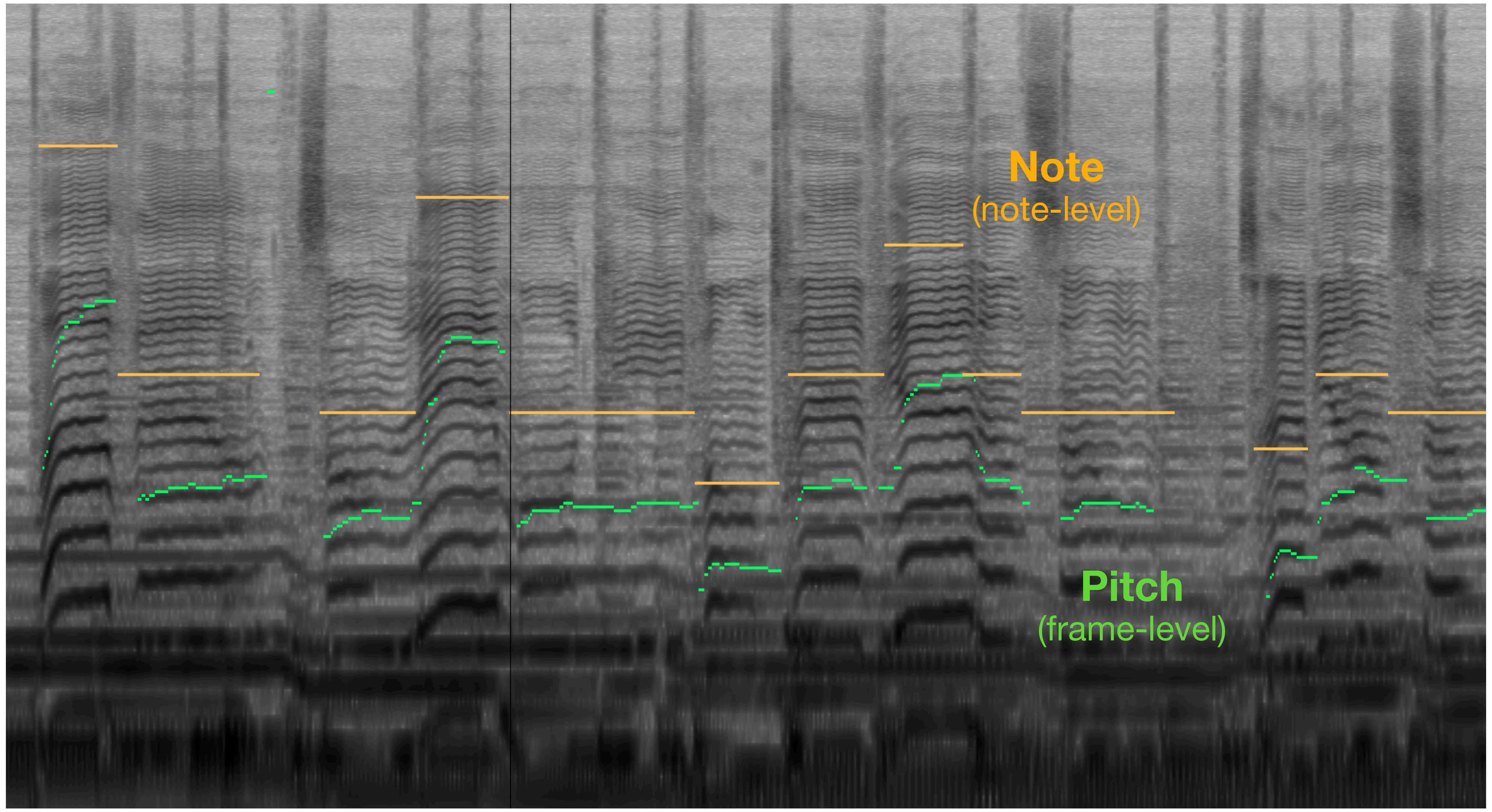


STP includes several sub-task:

1. Singing voice detection
2. Singing pitch estimation
3. Note-level segmentation
4. **Onset/offset detection**

I Singing Transcription from Polyphonic Music

: Motivation

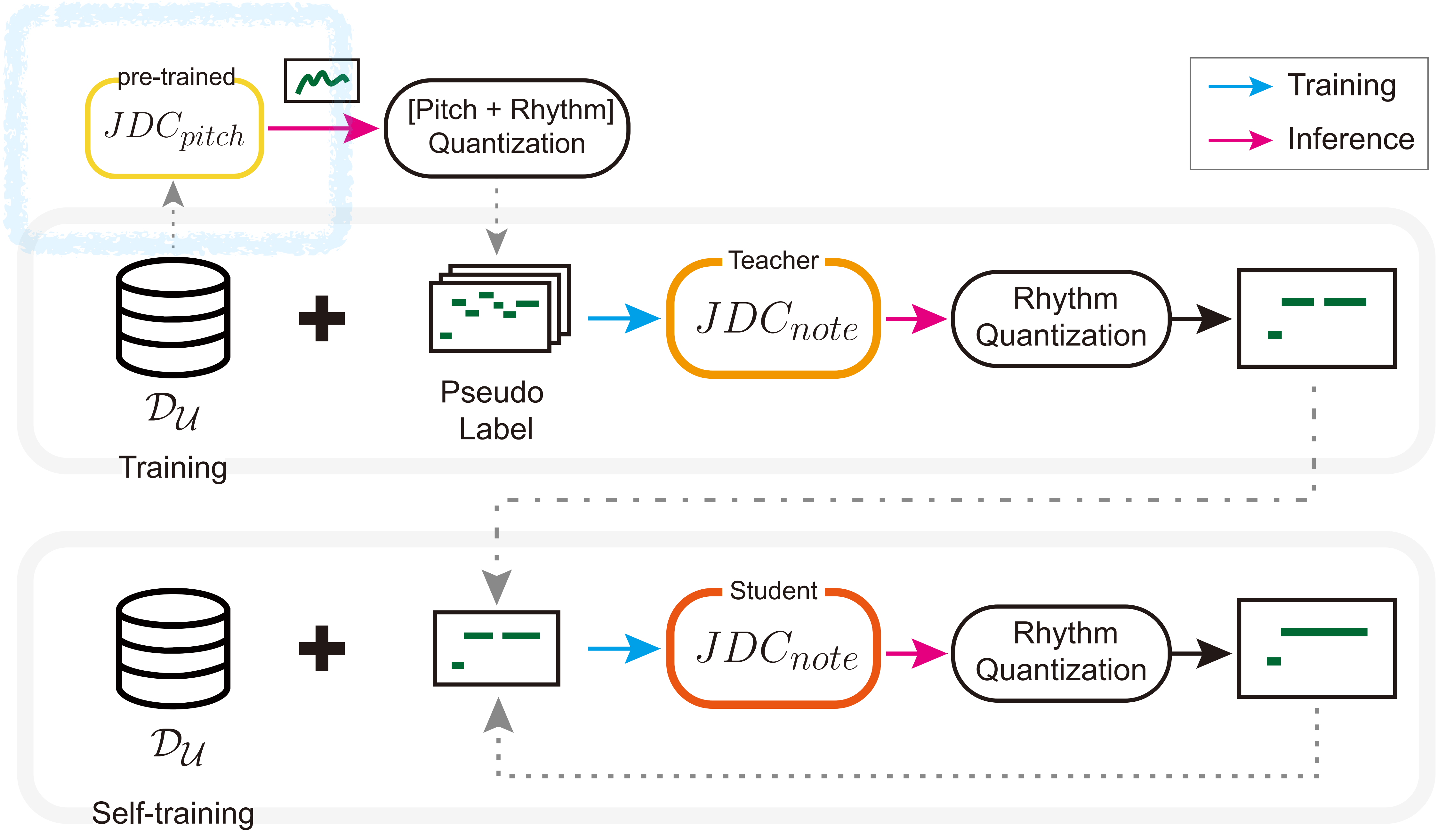


1. Several **sub-tasks**
2. High **variability** of singing voice
   (timbre, expression, formant modulation)
3. **Multiple** instrument sources
4. **Lack** of large-scale
   note-level labeled data for VOCALS

STP is a **challenging** task !!

I Singing Transcription from Polyphonic Music

: Contribution



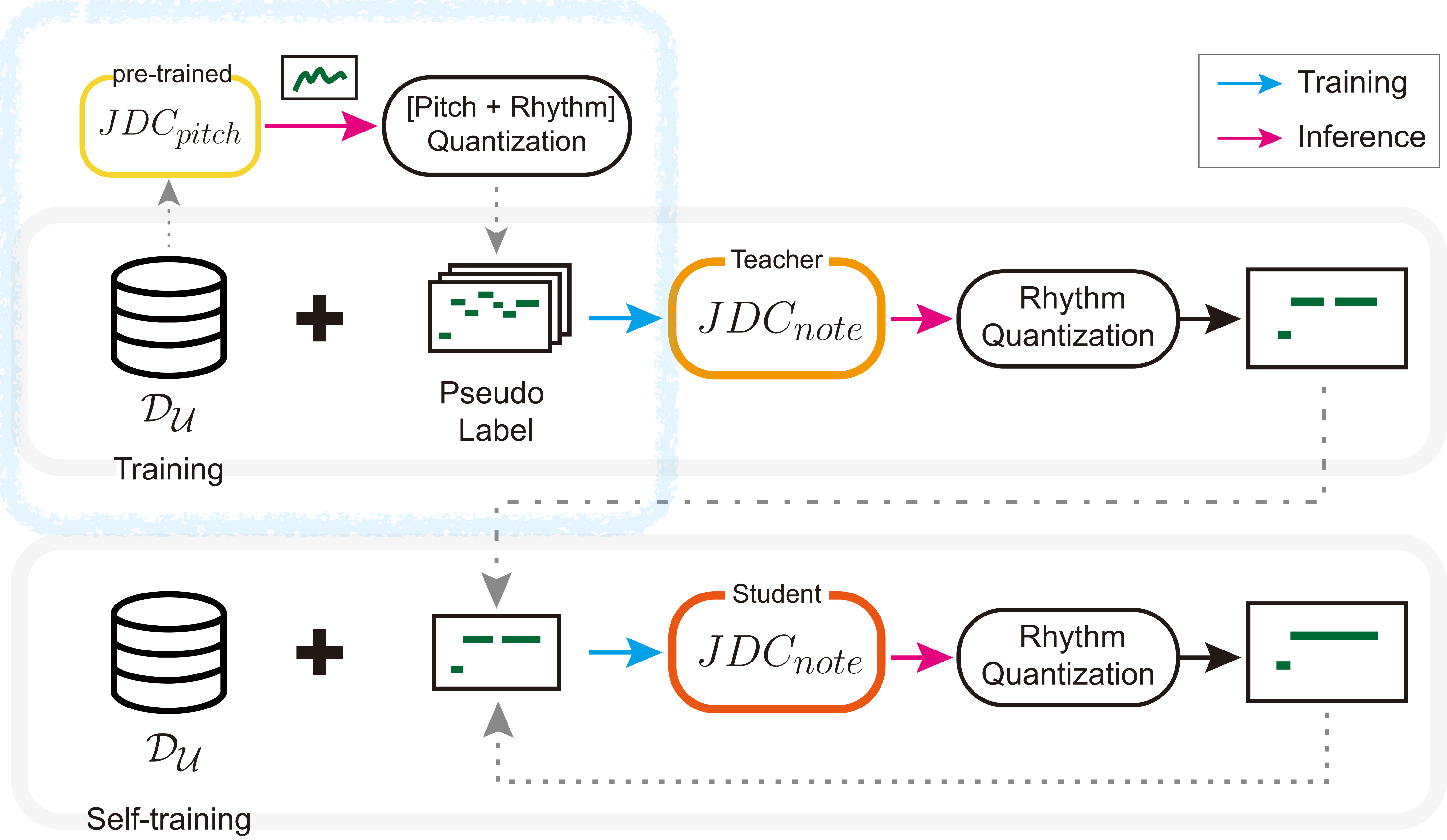
STP is a **challenging** task !!

Methods

1. Using pseudo labels from **pre-trained pitch estimation model.**
2. Convert the **frame-level** pseudo label to **note-level**
3. Training **STP** model using **joint detection and classification model (JDC)**
4. Self-training in an teacher-student framework

I Singing Transcription from Polyphonic Music

: Contribution



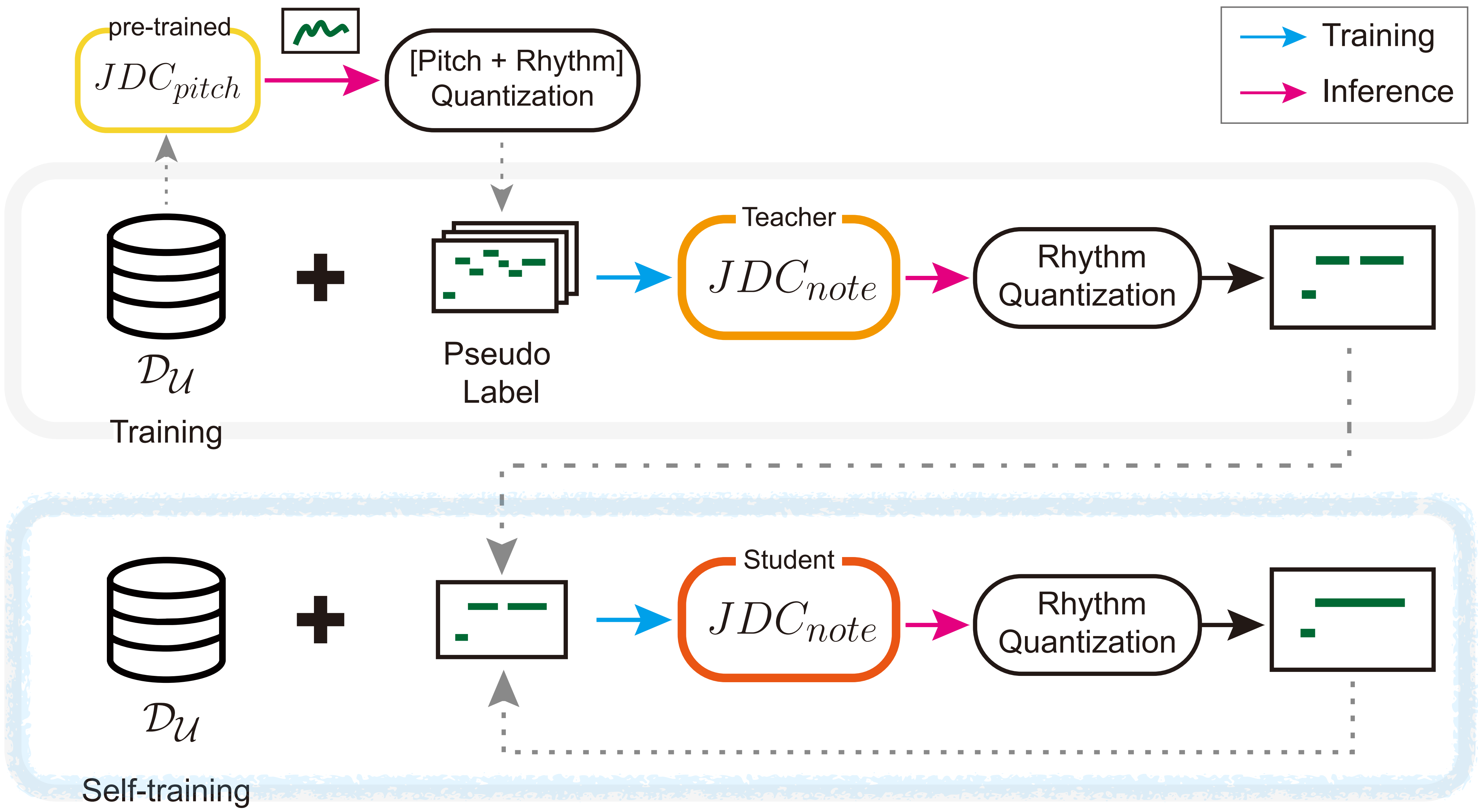
STP is a **challenging** task !!

Methods

1. Using pseudo labels from **pre-trained pitch estimation model.**
2. Convert the **frame-level** pseudo label to **note-level**
3. Training **STP** model using **joint detection and classification model (JDC)**
4. Self-training in an teacher-student framework

I Singing Transcription from Polyphonic Music

: Contribution



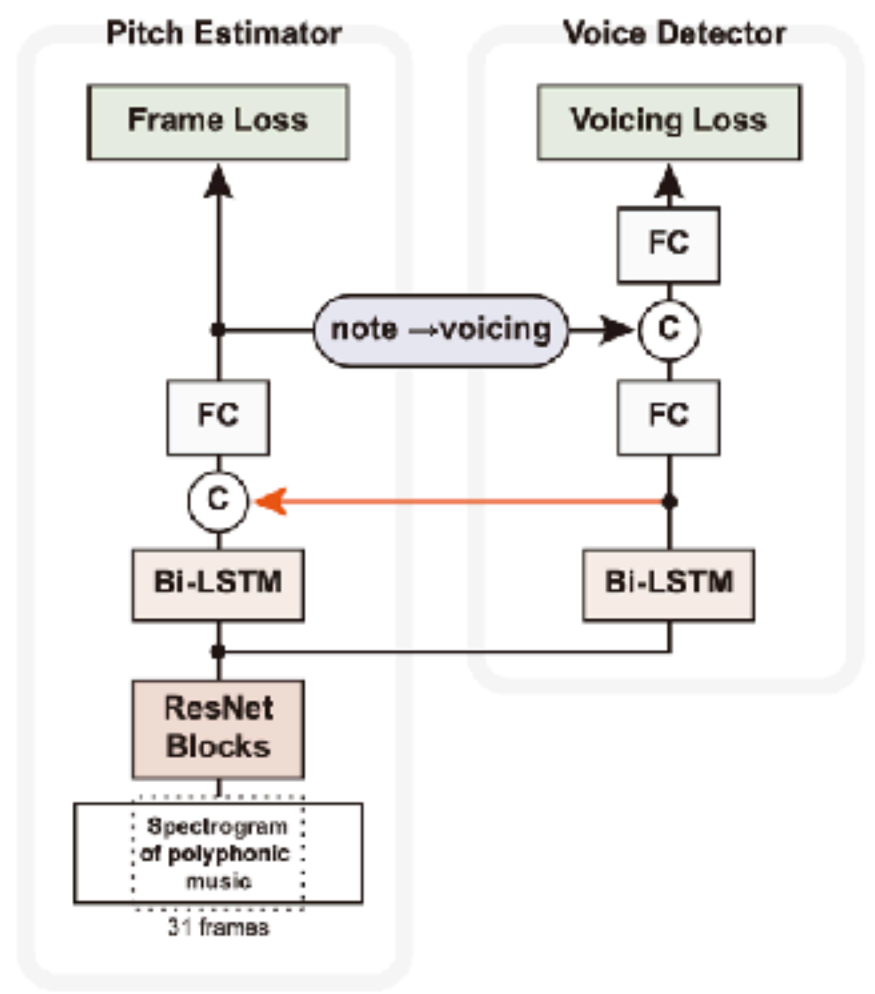


Fig. 2. The model architecture for JDC_{note}. "C" indicates feature concatenation.

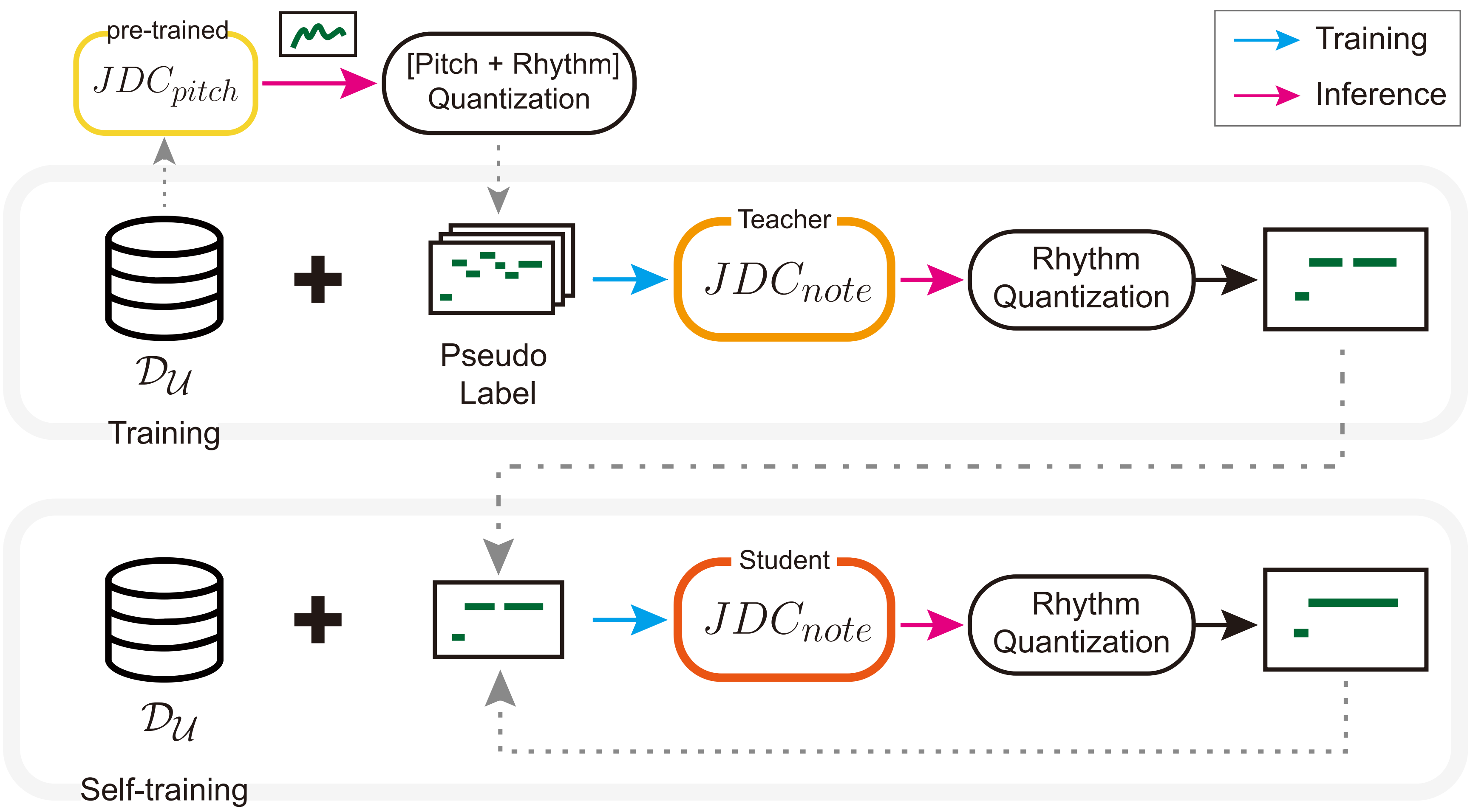
STP is a **challenging** task !!

Methods

1. Using pseudo labels from **pre-trained** **pitch estimation model.**
2. Convert the **frame-level** pseudo label to **note-level**
3. Training **STP** model using **joint detection and classification model (JDC)**
4. Self-training in an **teacher-student framework**

I Singing Transcription from Polyphonic Music

: Contribution



Cmedia						
Model	HZ	VOCANO	EFN	JDC_{note}		
				(U)	(L)	(L+U)
COnPOff	17.18	28.28	35.13	30.13	35.95	40.20
COnP	41.43	48.33	60.77	55.84	62.50	66.11
COn	63.63	64.56	**76.40**	65.72	73.88	75.97

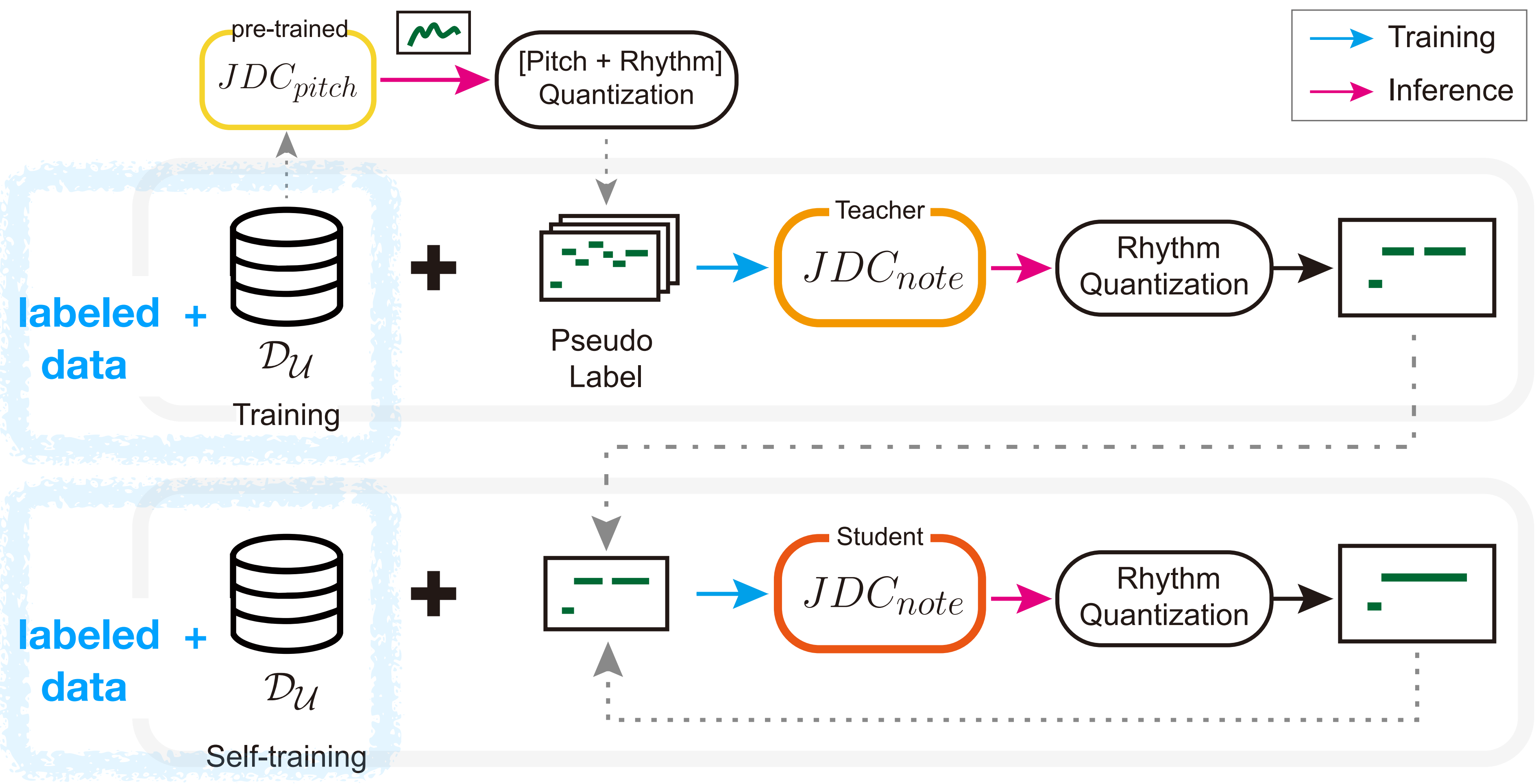
1. Using vocal pitch estimation model to predict frame-level pitch and **convert** it to note-level pseudo label.

2. Model (Using **only unlabeled** data)
   : Comparable results to the previous work (**no** use source separation)

3. Model (with **additional labeled** data)
   : Better performance than the model trained with only labeled data.

I Singing Transcription from Polyphonic Music

: Contribution



Cmedia						
Model	HZ	VOCANO	EFN	JDC_{note} (U)	JDC_{note} (L)	JDC_{note} (L+U)
COnPOff	17.18	28.28	35.13	30.13	35.95	**40.20**
COnP	41.43	48.33	60.77	55.84	62.50	**66.11**
COn	63.63	64.56	**76.40**	65.72	73.88	75.97

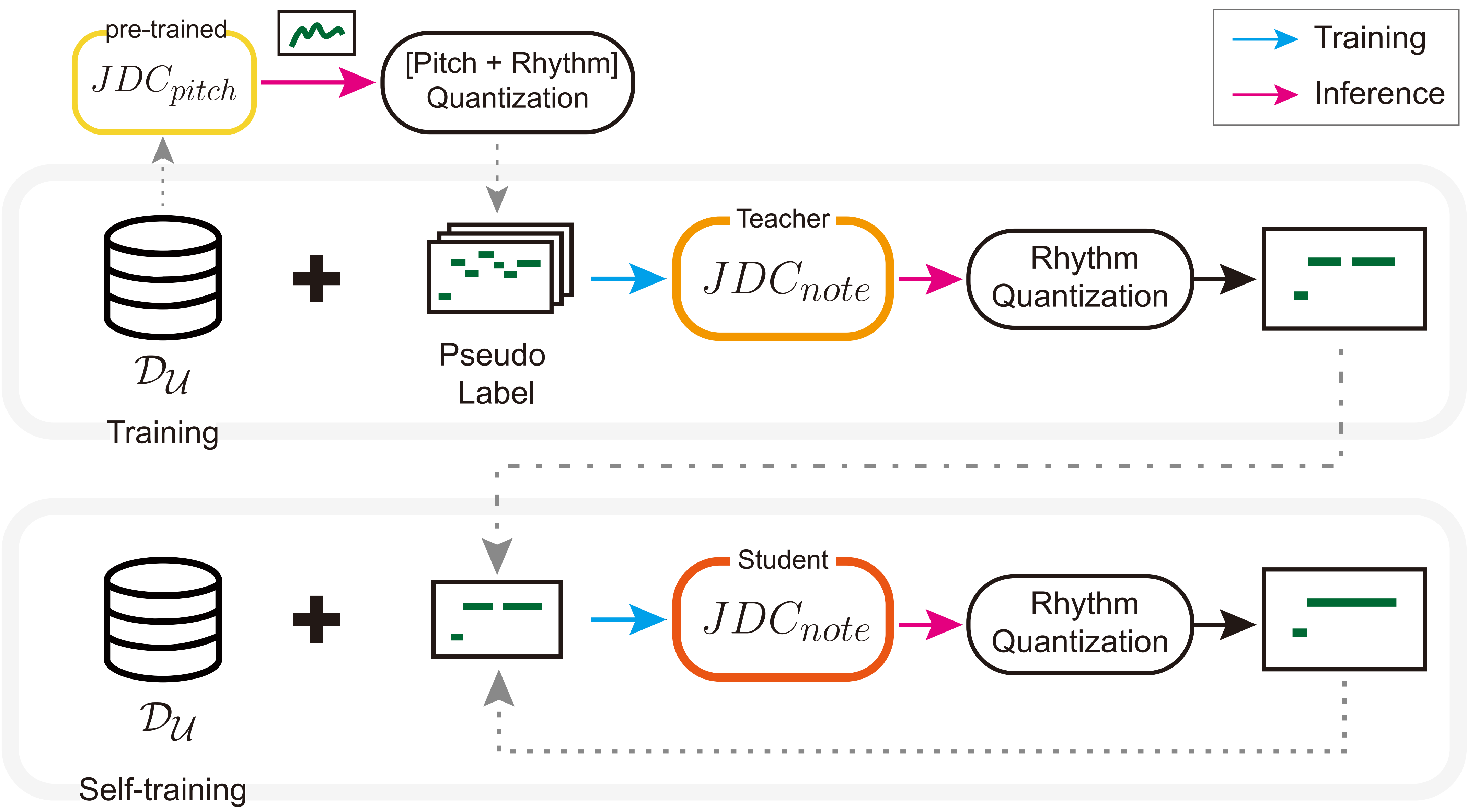
1. Using vocal pitch estimation model to predict frame-level pitch and **convert** it to note-level pseudo label.

2. Model (Using **only unlabeled** data)
   : Comparable results to the previous work (**no** use source separation)

3. Model (with **additional labeled** data)
   : Better performance than the model trained with only labeled data.

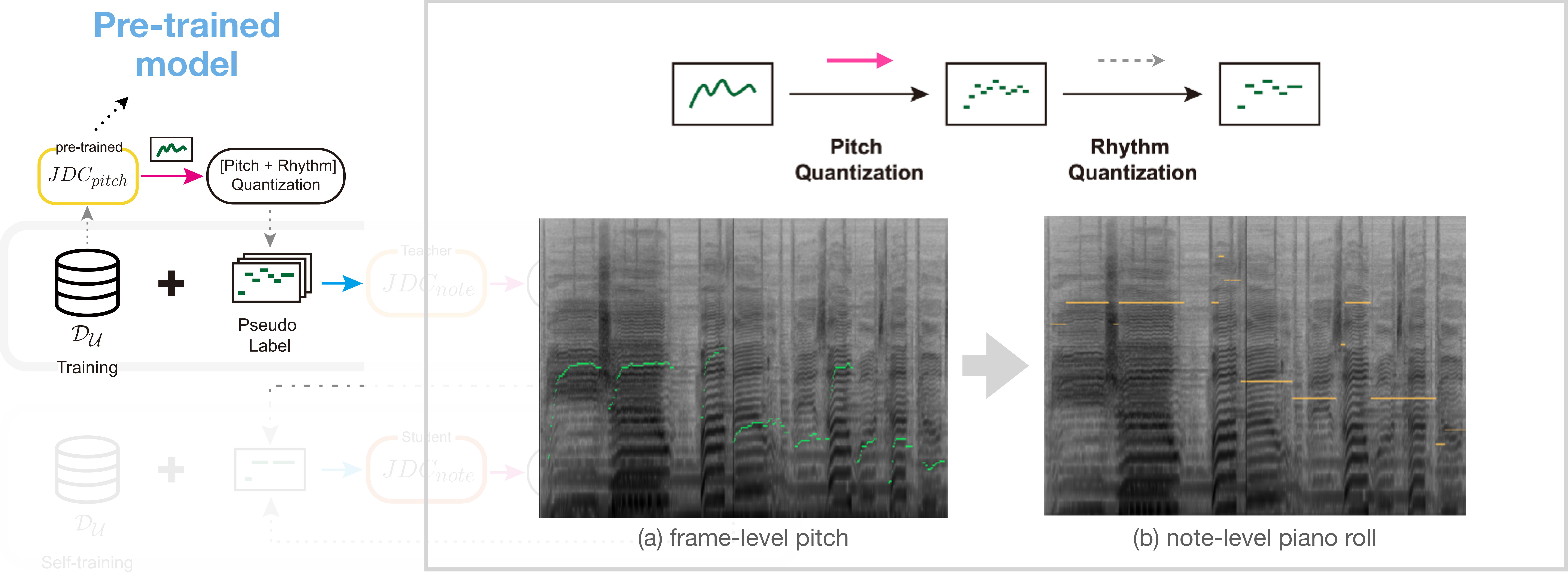
[Step 1] Making pseudo labels

: Using vocal pitch estimation model from Unlabeled dataset



[Step 1] Making pseudo labels

: Using pre-trained vocal pitch estimation model



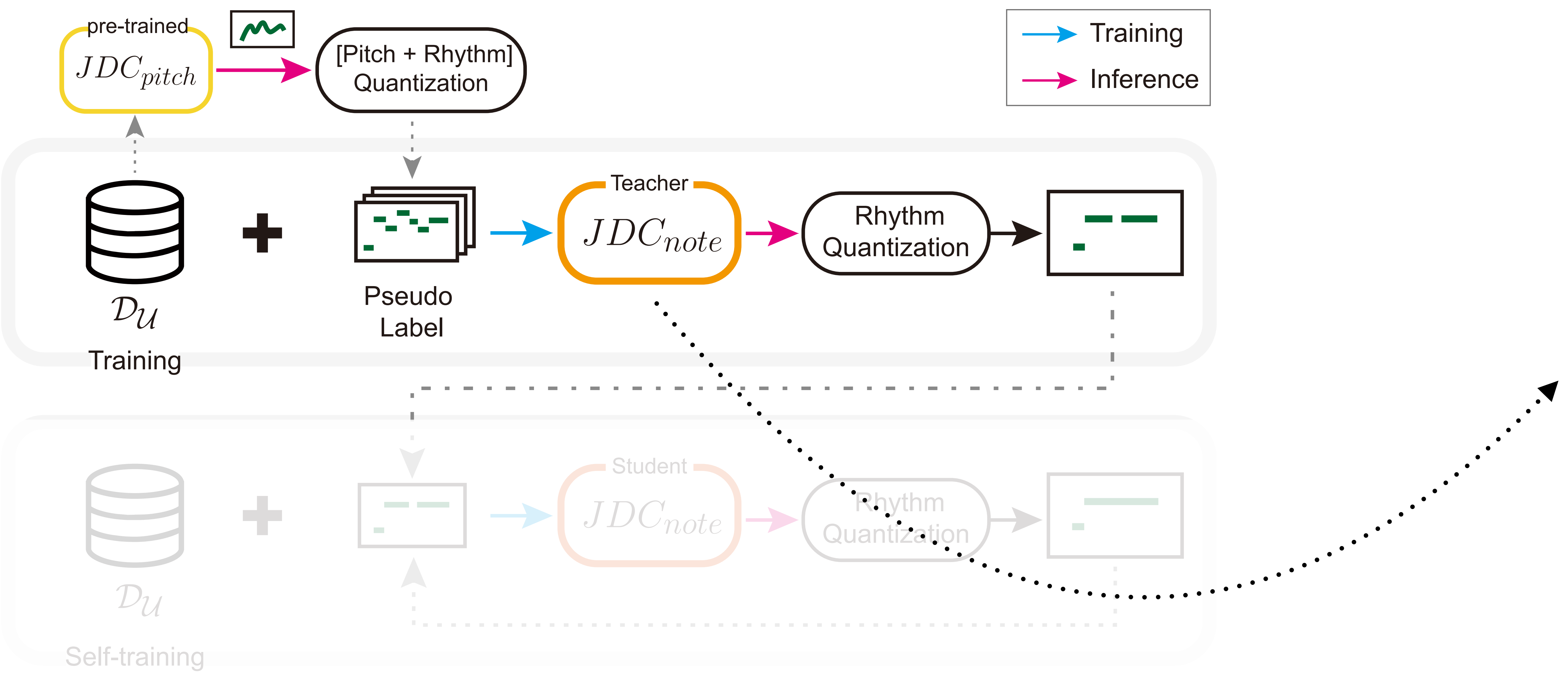
(a) frame-level pitch

(b) note-level piano roll

[Step 2] Training note transcription model

: Training STP model using pseudo label

First build a **new neural network model** for STP and train it using the note-level pseudo labels obtained from the first stage.



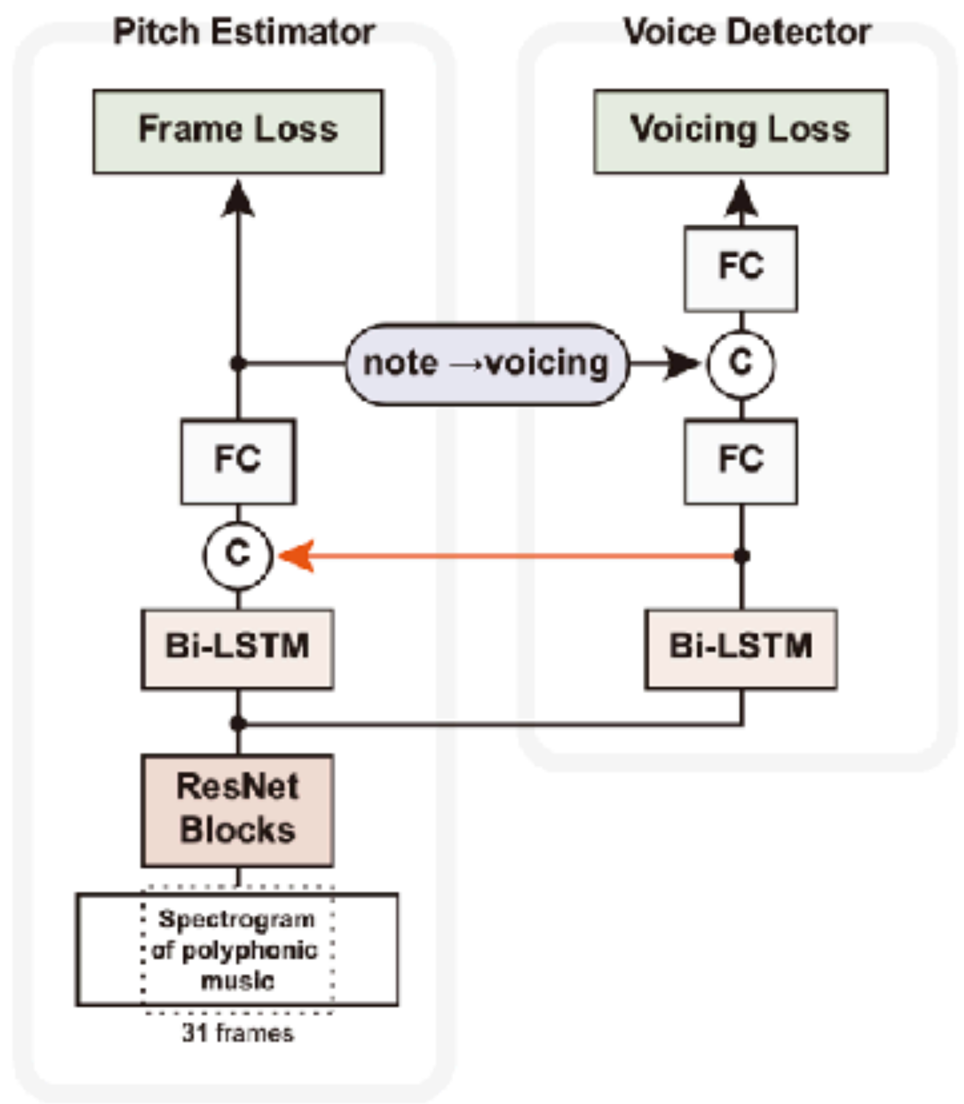
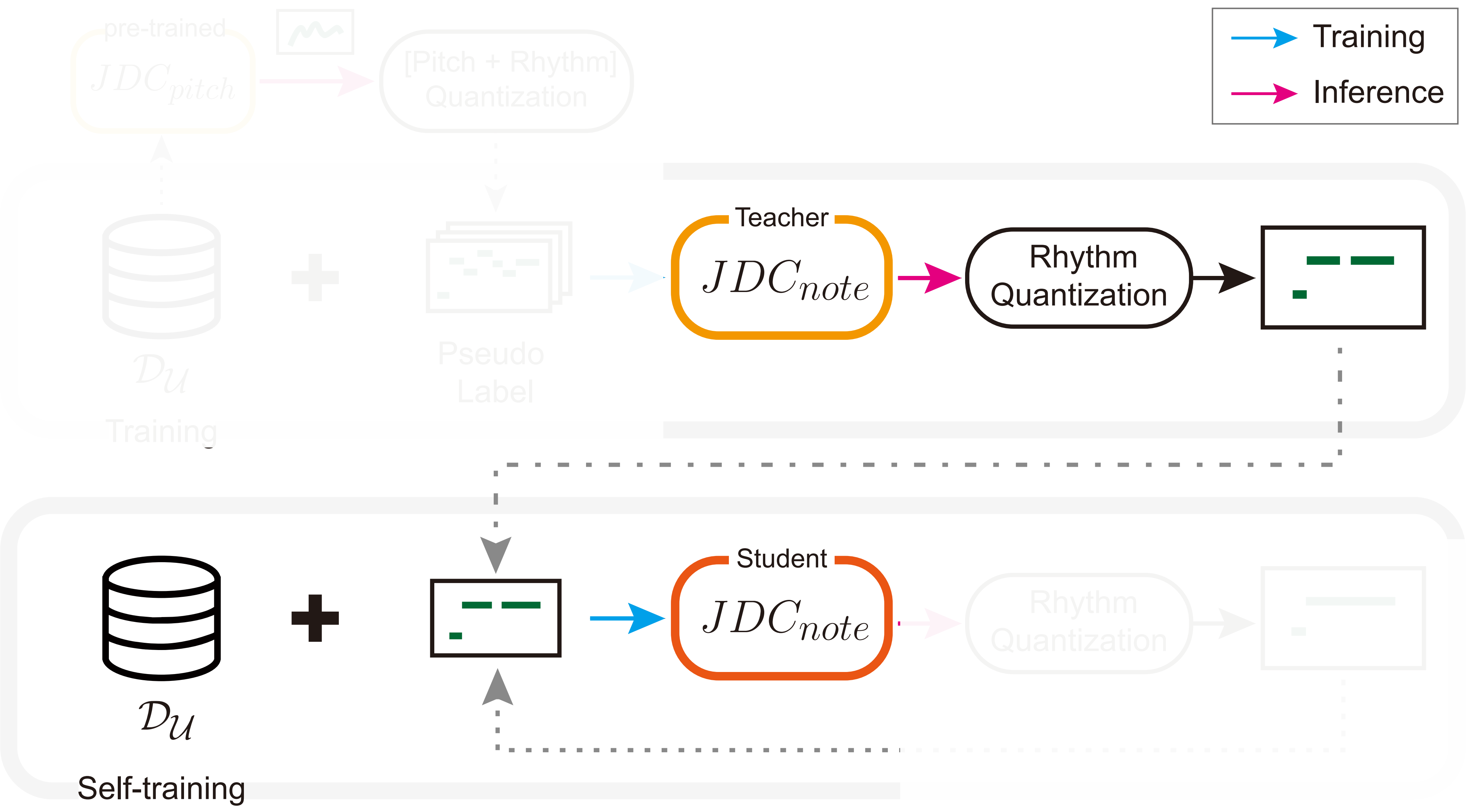
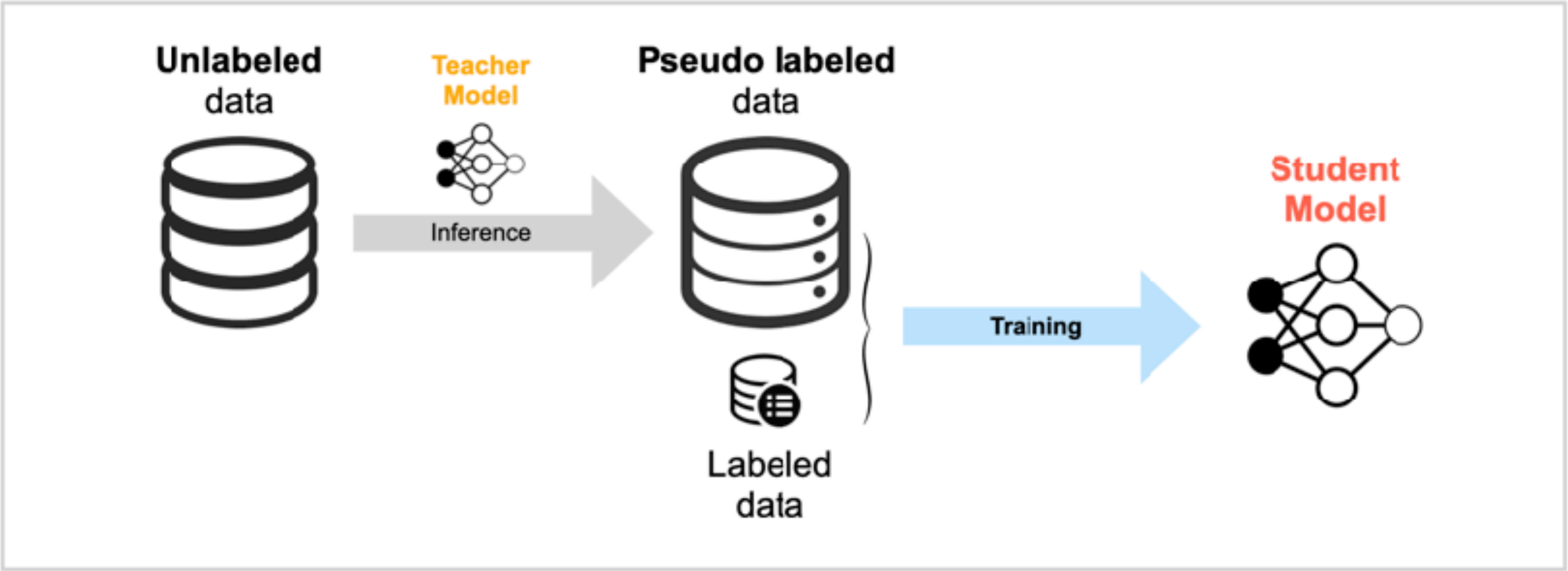


Fig. 2. The model architecture for JDC_{note}. "C" indicates feature concatenation.

[Step 3] Self-training in the Teacher-Student framework

: Noisy Student

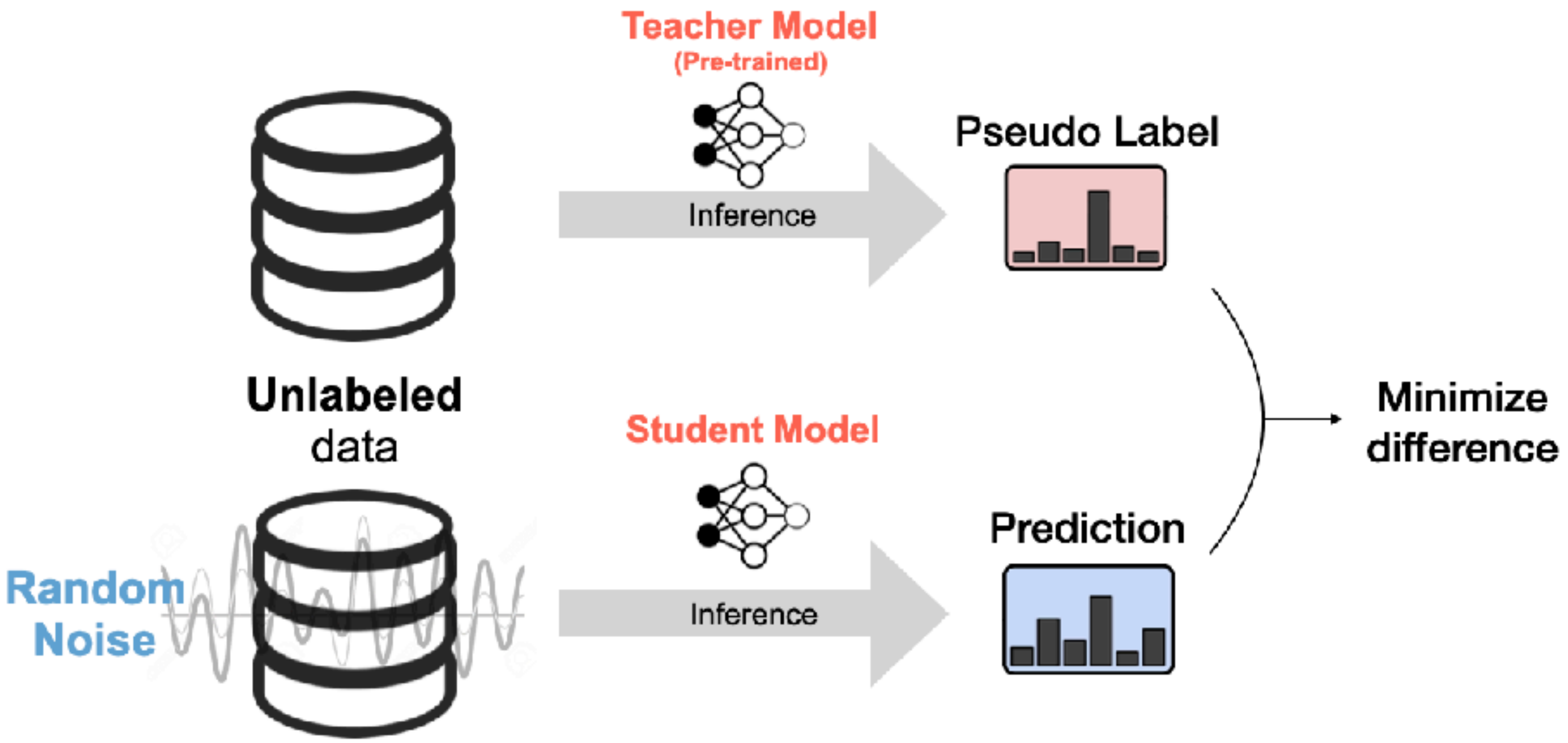




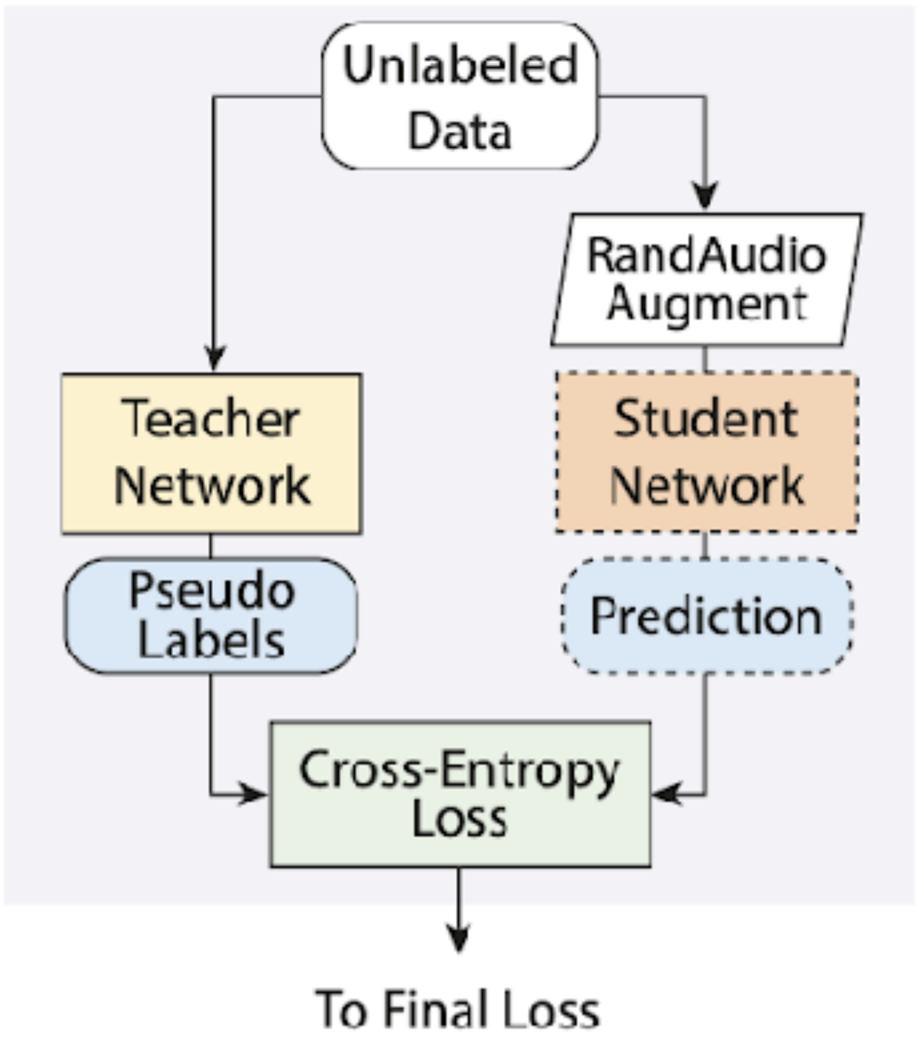
Train the **student model** using **pseudo labeled** data from teacher model (& labeled data)

[Step 3] Self-training in the Teacher-Student framework

: Noisy Student



(c) Noisy Student

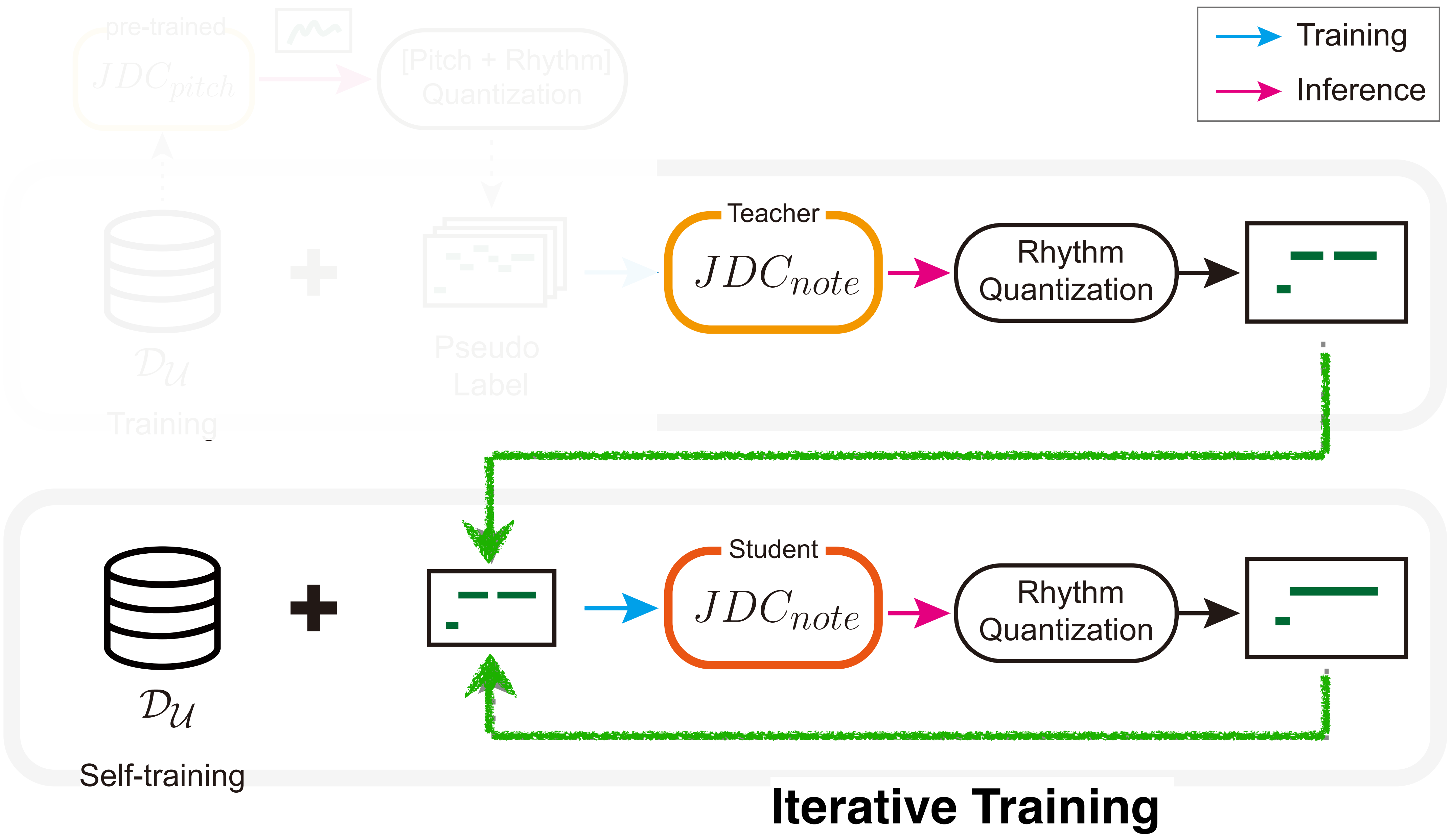


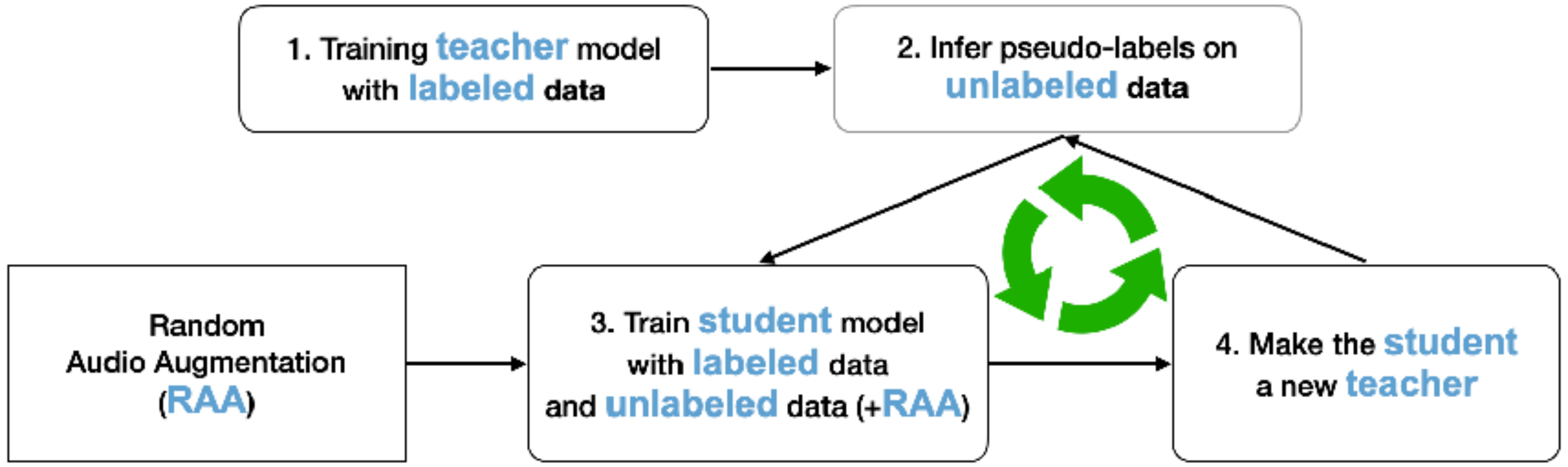
We train STP model using the Noisy Student model [2]
to encourage the model to produce **consistent** output.

[2] Kum, S., Lin, J. H., Su, L., & Nam, J.. Semi-supervised learning using teacher-student models for vocal melody extraction. ISMIR, 2020

[Step 3] Self-training in the Teacher-Student framework

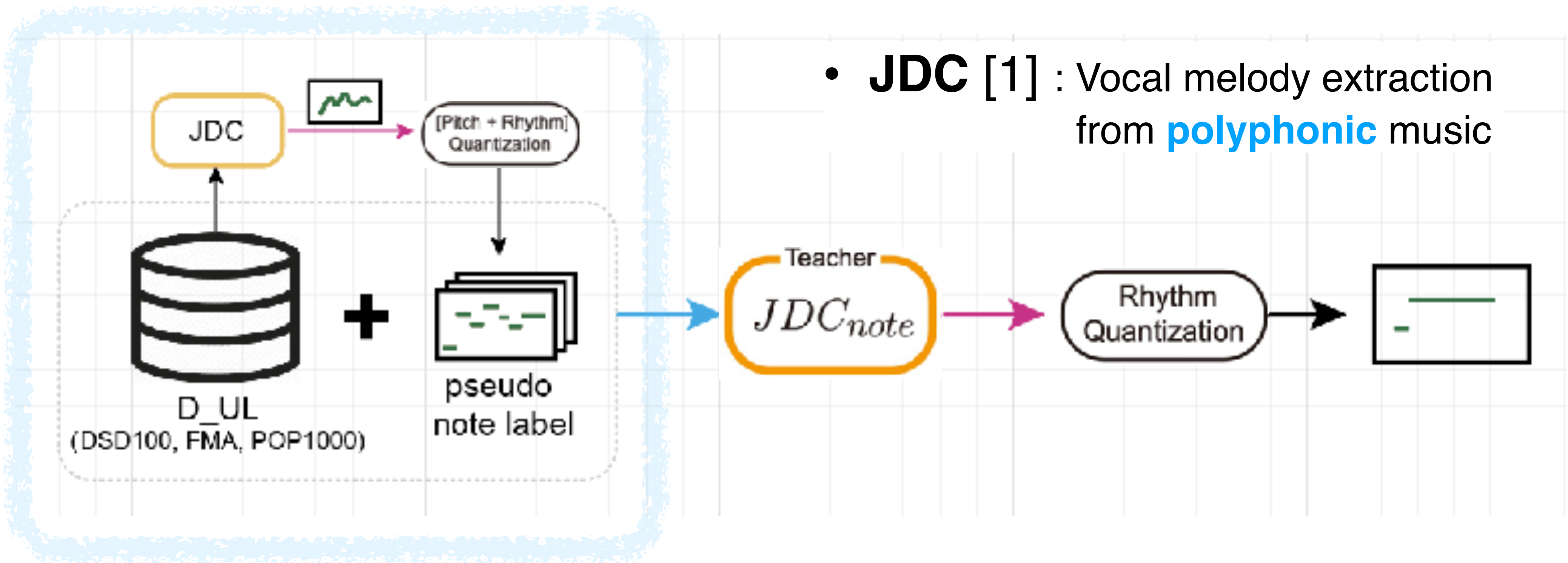
: Iterative Training



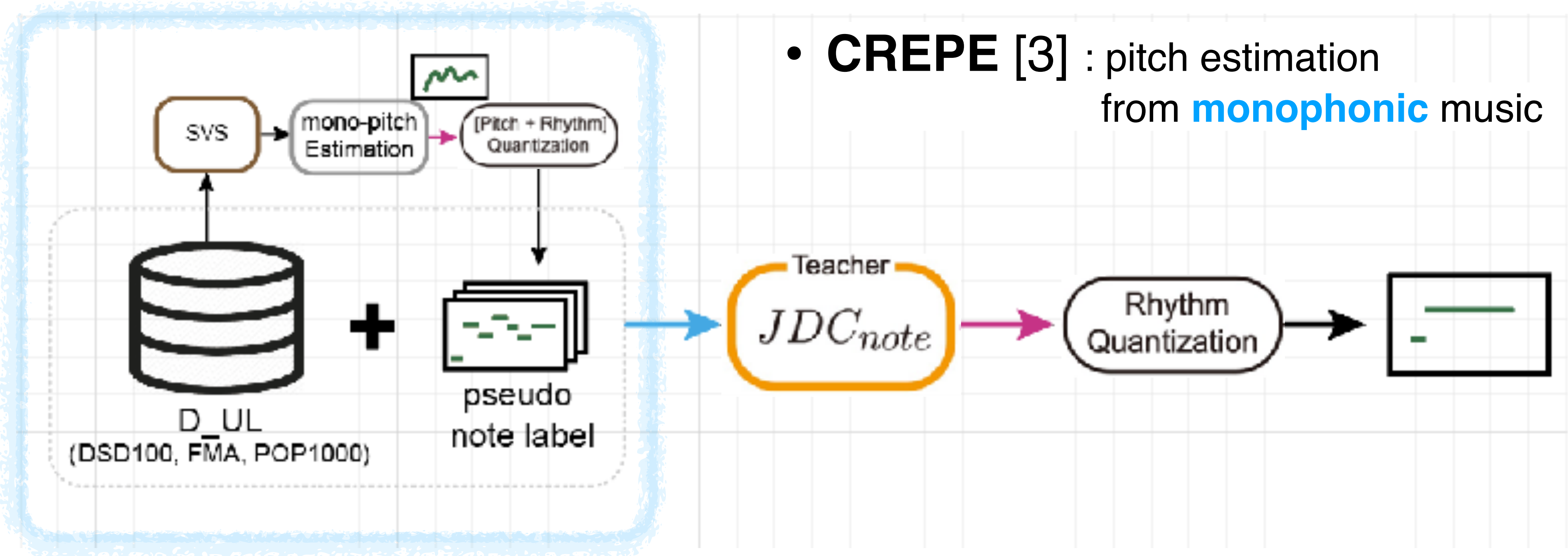


I Experiment 1: Comparison of Pitch Estimation Models

: Pitch estimation methods for obtain pseudo label



- **JDC** [1] : Vocal melody extraction from **polyphonic** music



- **CREPE** [3] : pitch estimation from **monophonic** music

	Initial Pseudo Labels		JDC_{note} (Teacher)	
Repurposed Models	Demucs + CREPE	JDC_{pitch}	Demucs + CREPE	JDC_{pitch}
COnPOff	22.43	25.44	24.71	28.97
COnP	45.01	48.48	48.64	53.32
COn	57.65	61.94	62.32	64.74

+

JDC_{pitch} > **source separation** (Demucs) **+ CREPE**

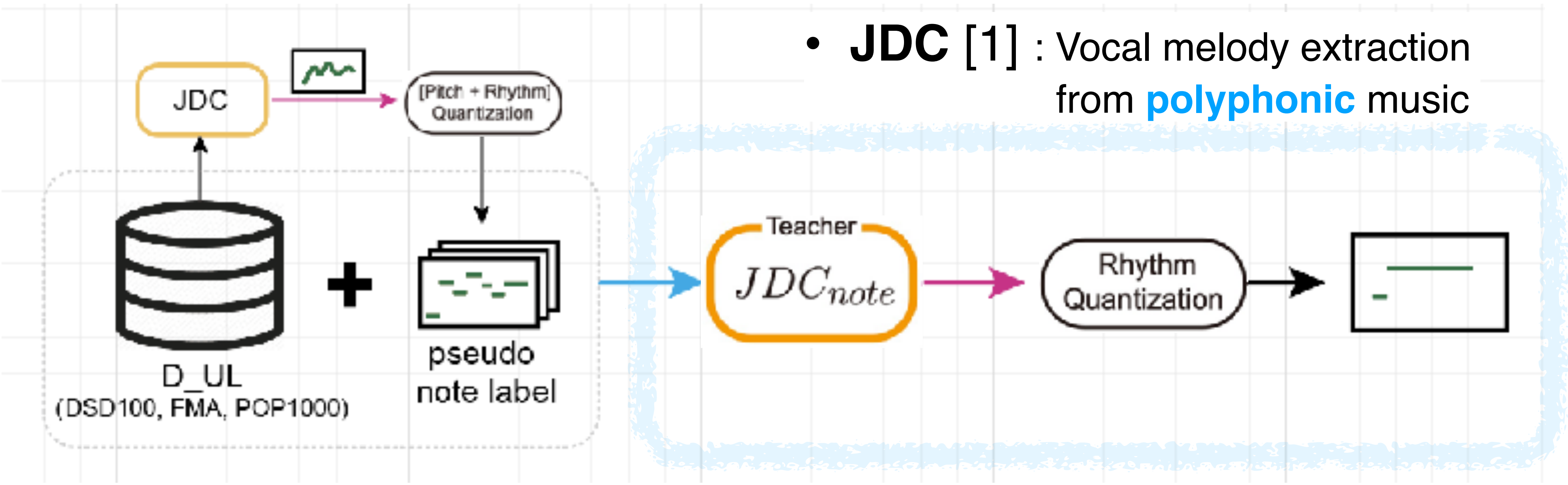
: Separation algorithms **cannot separate only the main vocal melody**
: Polyphonic vocals are still remained → CREPE = Low performance

[3] Kim, J. W., Salamon, J., Li, P., & Bello, J. P. Crepe: A convolutional representation for pitch estimation, ICASSP , 2018

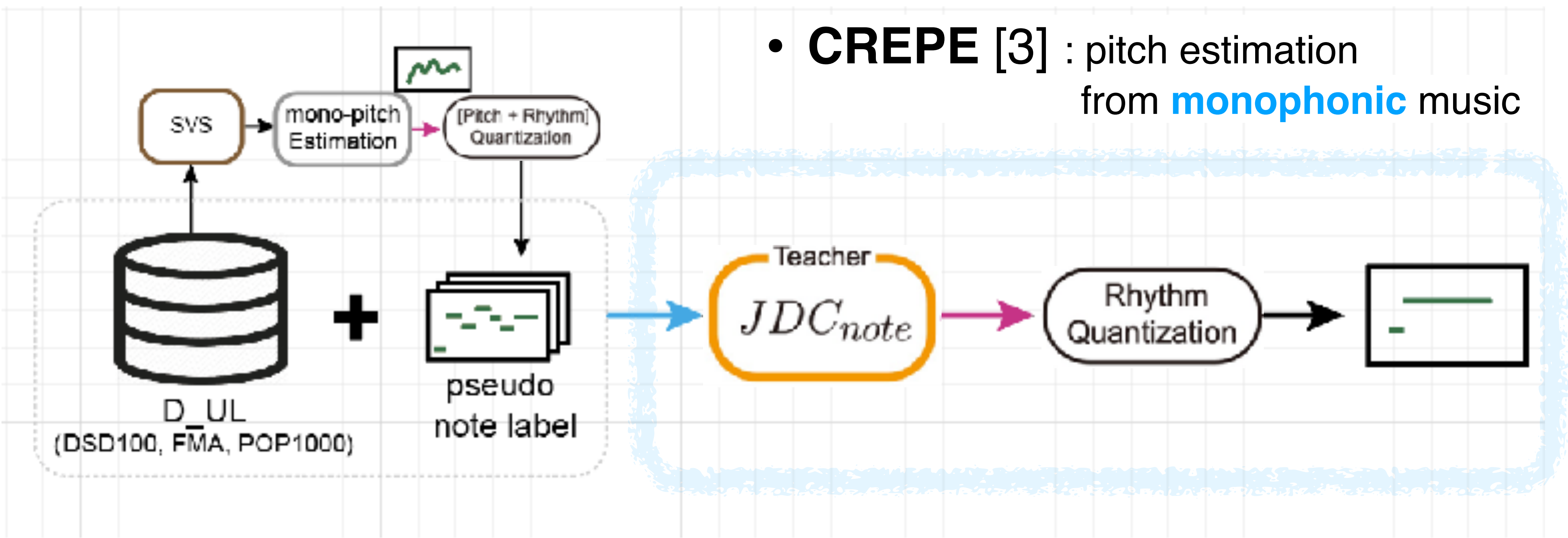
I Experiment 1: Comparison of Pitch Estimation Models

: Pitch estimation methods for obtain pseudo label

- **JDC** [1] : Vocal melody extraction from **polyphonic** music



- **CREPE** [3] : pitch estimation from **monophonic** music



	Initial Pseudo Labels		JDC_{note} (*Teacher*)	
Repurposed Models	Demucs + CREPE	JDC_{pitch}	Demucs + CREPE	JDC_{pitch}
COnPOff	22.43	25.44	24.71	28.97
COnP	45.01	48.48	48.64	53.32
COn	57.65	61.94	62.32	64.74

$JDC_{note}(Teacher)$ > Initial Pseudo Labels

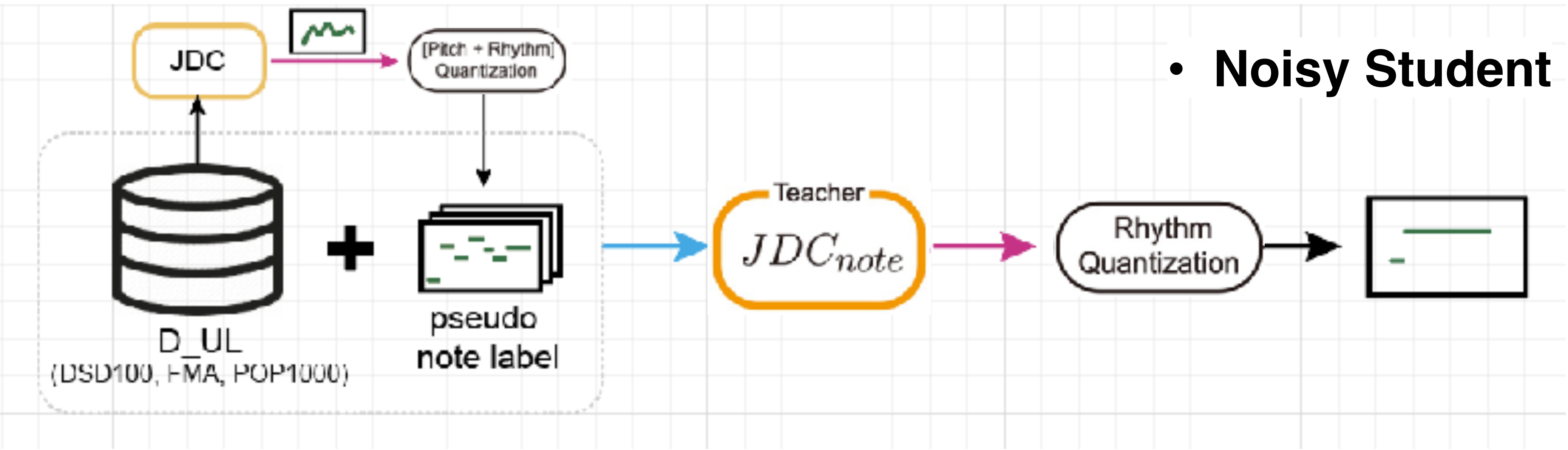
: Confirm the efficacy of the **repurposed** neural network models

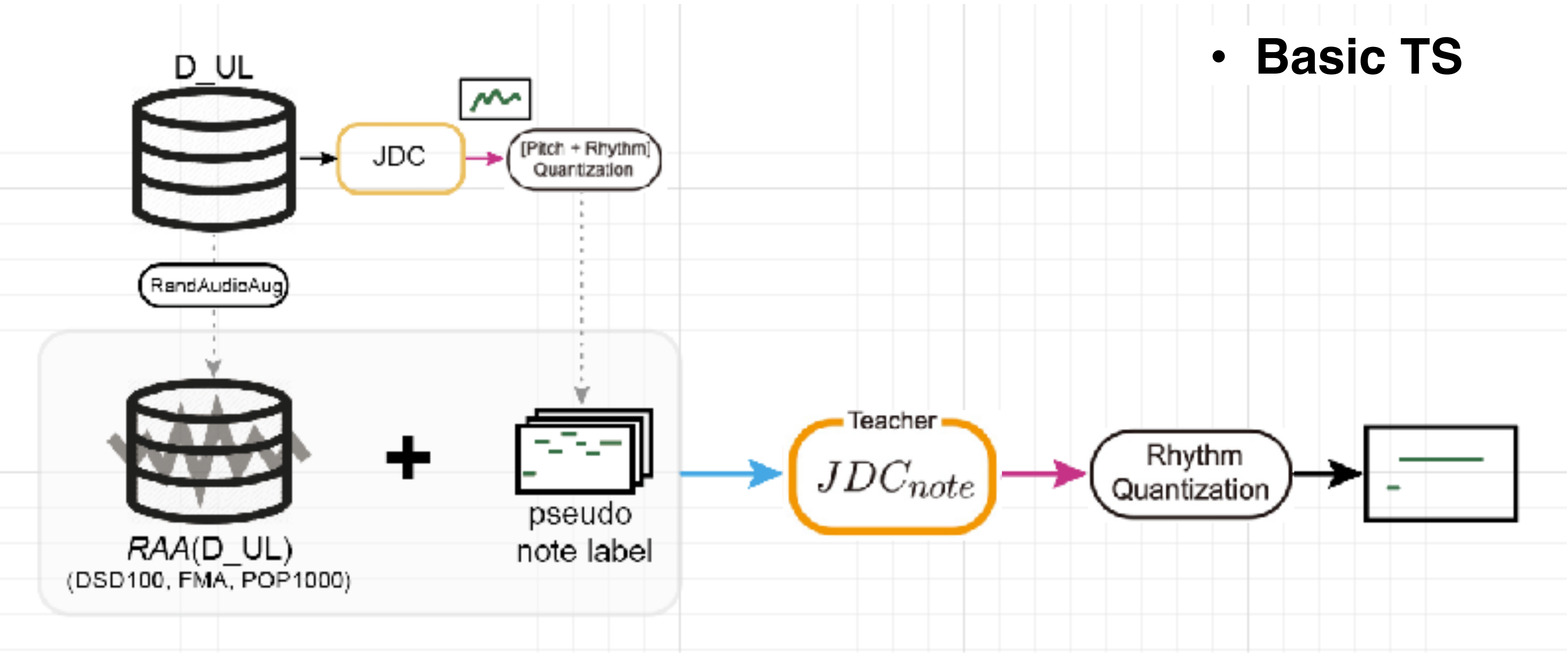
: Confirm the efficacy of the **JDC network** for STP

[3] Kim, J. W., Salamon, J., Li, P., & Bello, J. P. Crepe: A convolutional representation for pitch estimation, ICASSP , 2018

I Experiment 2: Teacher-Student Framework

: Basic Teacher-Student VS. Noisy Student





	Cmedia		MIR-ST500	
Models	TS	NS	TS	NS
COnPOff	28.97	29.62	22.12	22.62
COnP	53.32	54.55	40.01	40.70
COn	64.74	65.61	56.90	57.87

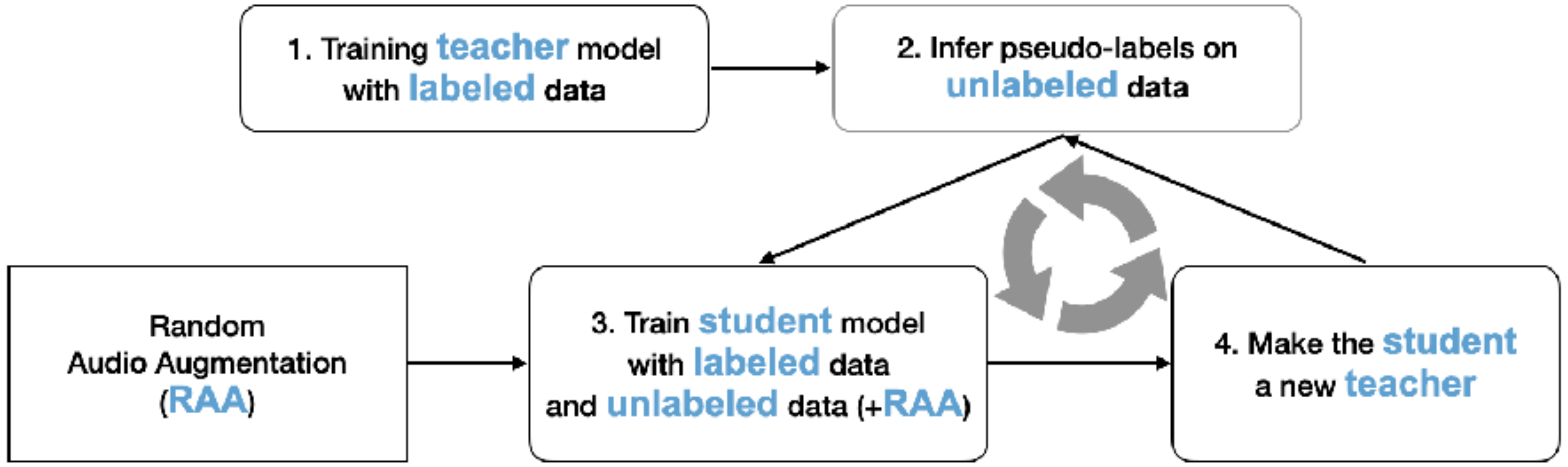
+ +

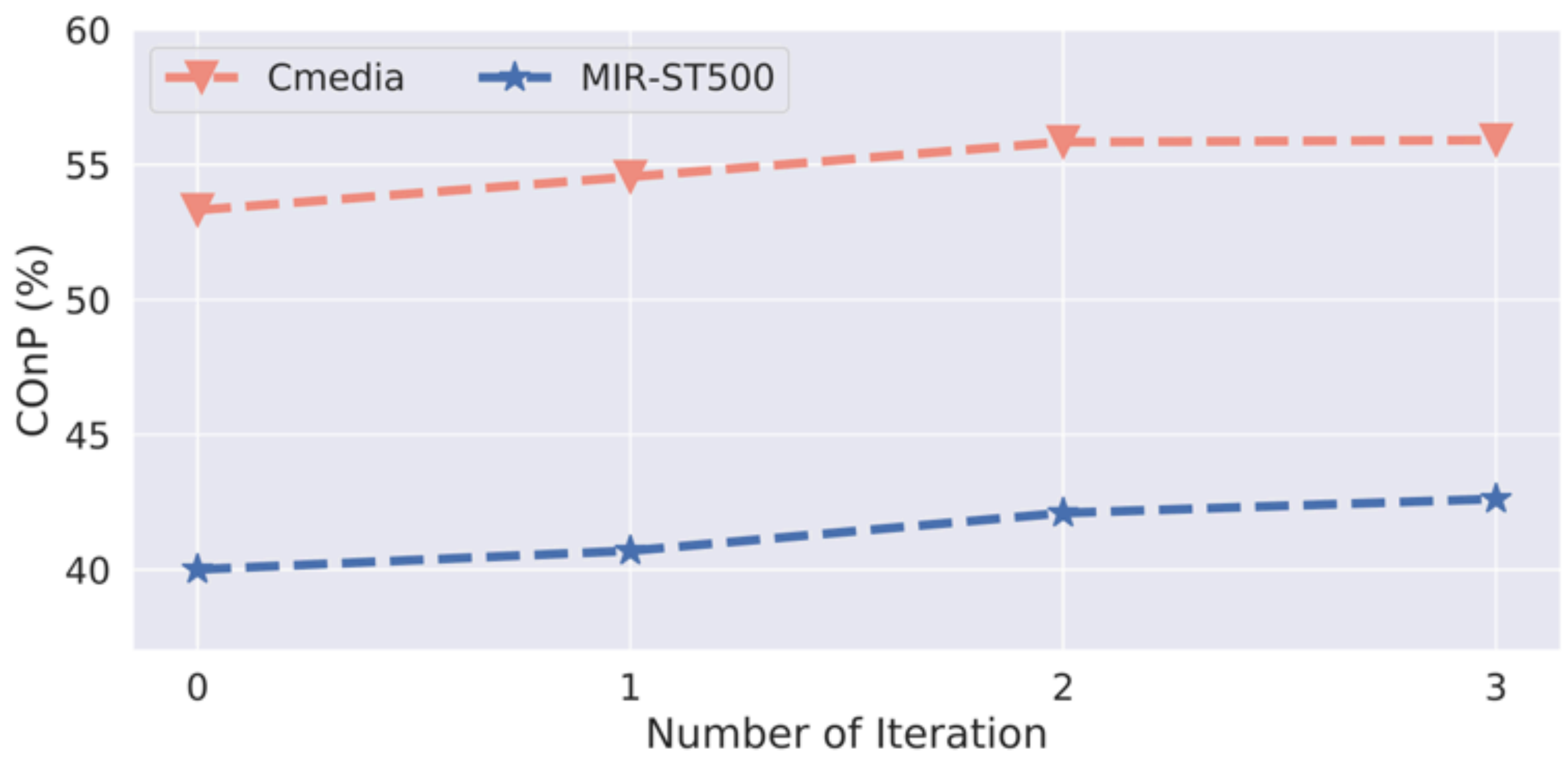
Noisy Student > Basic TS

: The student produce **consistent** outputs that **minimize the difference** from the teacher even though the **input is perturbed**

Experiment 3: Teacher-Student Framework

: Iterative Training





Iterative Training

: The performance continuously **increases** up to 2 iterations

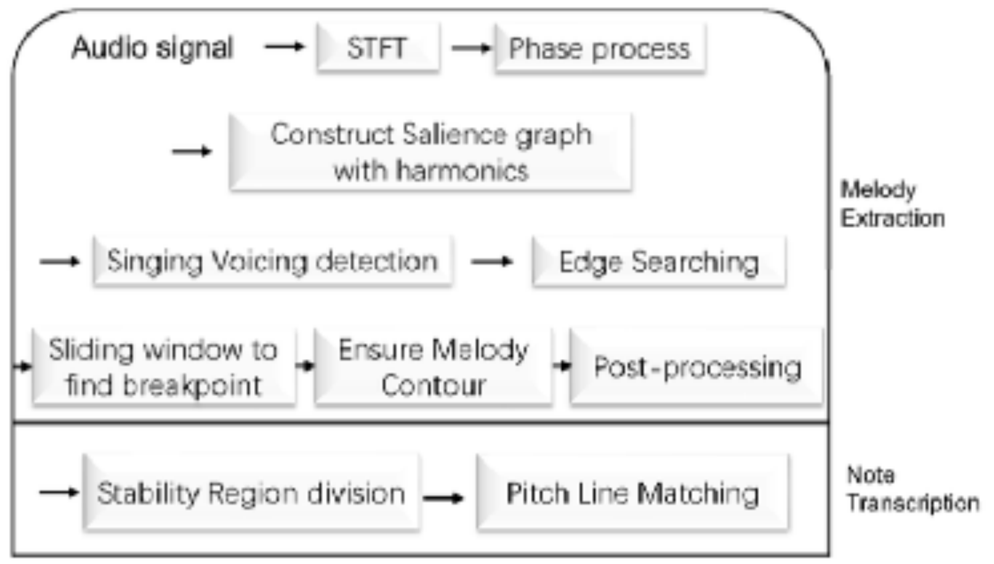
I Comparison with Supervised and Semi-Supervised Models

: Unsupervised, Supervised, and Semi-supervised

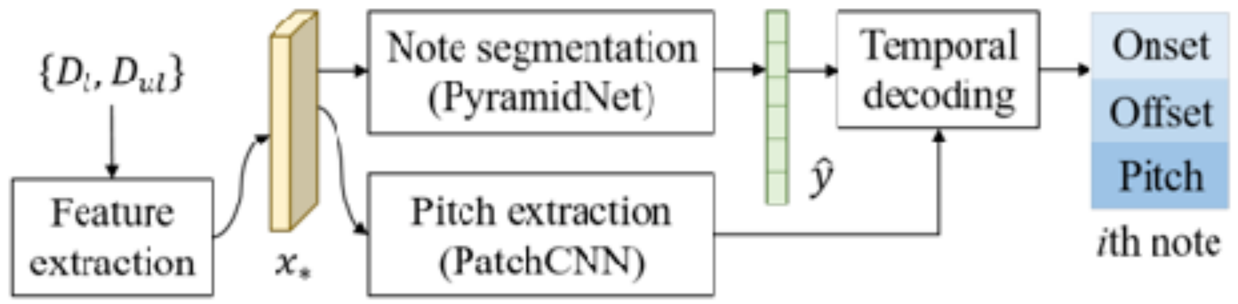
- **HZ** [4]**:** Rule-based model
- **VOCANO** [5]**:** Semi-supervised model
- **EFN** [6]**:** Supervised model

	Source Separation Required	**Method**	**Data**
HZ [4]	x	Rule-based	-
VOCANO [5]	o	Semi-supervised	Labeled + Unlabeled
EFN [6]	o	Supervised	Labeled
Proposed	x	Teacher-student framework	Labeled + Unlabeled

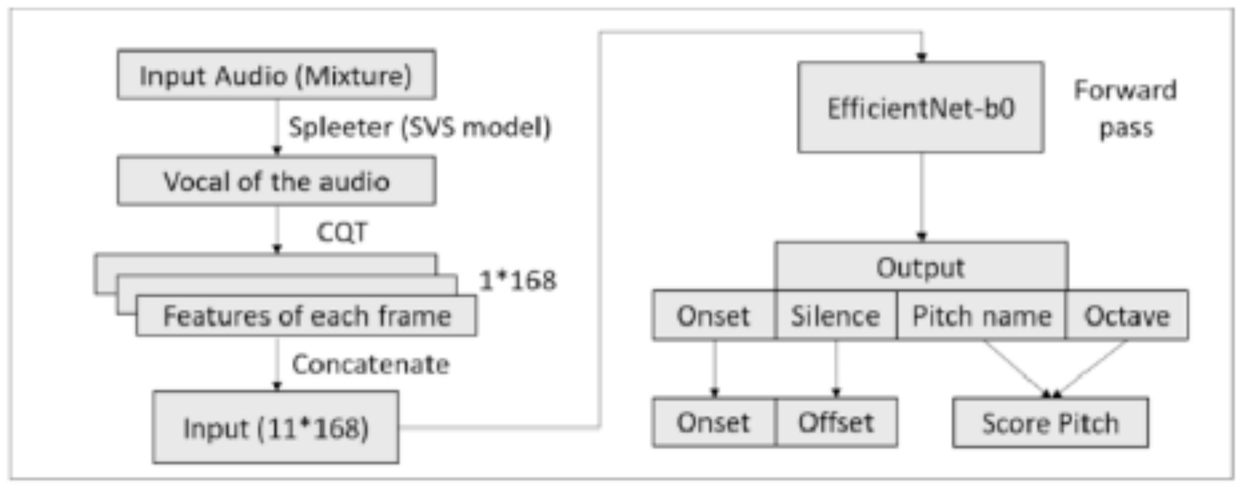
HZ



VOCANO



EFN



[4] He, Z. & Feng, Y., "Singing transcription from polyphonic music using melody contour filtering," Applied Sciences, 2021
[5] Wang, J., & Jang, J., "On the preparation and validation of a large-scale dataset of singing transcription," in Proc. ICASSP, 2021
[6] Hsu, J. & Su, L., "VOCANO: A note transcription framework for singing voice in polyphonic music," in Proc. ISMIR, 2021

I Comparison with Supervised and Semi-Supervised Models

: Unsupervised, Supervised, and Semi-supervised

Description	
JDC$_{note}$(U)	Unsupervised model with unlabeled data $\mathcal{D}_{\mathcal{U}}$
JDC$_{note}$(L)	Supervised model with labeled data $\mathcal{D}_{\mathcal{L}}$
JDC$_{note}$(L+U)	Semi-supervised model with $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$

Cmedia

Model	HZ	VOCANO	EFN	JDC$_{note}$ (U)	JDC$_{note}$ (L)	JDC$_{note}$ (L+U)
COnPOff	17.18	28.28	35.13	30.13	35.95	**40.20**
COnP	41.43	48.33	60.77	55.84	62.50	**66.11**
COn	63.63	64.56	**76.40**	65.72	73.88	75.97

MIR-ST500

Model	HZ	VOCANO	EFN	JDC$_{note}$ (U)	JDC$_{note}$ (L)	JDC$_{note}$ (L+U)
COnPOff	-	-	**45.78**	23.48	40.57	42.23
COnP	-	-	66.63	42.10	67.55	**69.74**
COn	-	-	75.44	58.61	74.94	**76.18**

- **HZ** [4]: **Rule-based** model
- **VOCANO** [5]: **Semi-supervised** model
- **EFN** [6]: **Supervised** model

EFN > $JDC_{note}(U)$ **> VOCANO > HZ**

: This validates that the proposed method is superior to the semi-supervised method in VOCANO or the rule-based approach in HZ.

I Comparison with Supervised and Semi-Supervised Models

: Unsupervised, Supervised, and Semi-supervised

	Description
JDC$_{note}$(U)	Unsupervised model with unlabeled data $\mathcal{D}_{\mathcal{U}}$
JDC$_{note}$(L)	Supervised model with labeled data $\mathcal{D}_{\mathcal{L}}$
JDC$_{note}$(L+U)	Semi-supervised model with $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$

Cmedia

Model	HZ	VOCANO	EFN	JDC$_{note}$		
				(U)	(L)	(L+U)
COnPOff	17.18	28.28	35.13	30.13	35.95	**40.20**
COnP	41.43	48.33	60.77	55.84	62.50	**66.11**
COn	63.63	64.56	**76.40**	65.72	73.88	75.97

MIR-ST500

Model	HZ	VOCANO	EFN	JDC$_{note}$		
				(U)	(L)	(L+U)
COnPOff	-	-	**45.78**	23.48	40.57	42.23
COnP	-	-	66.63	42.10	67.55	**69.74**
COn	-	-	75.44	58.61	74.94	**76.18**

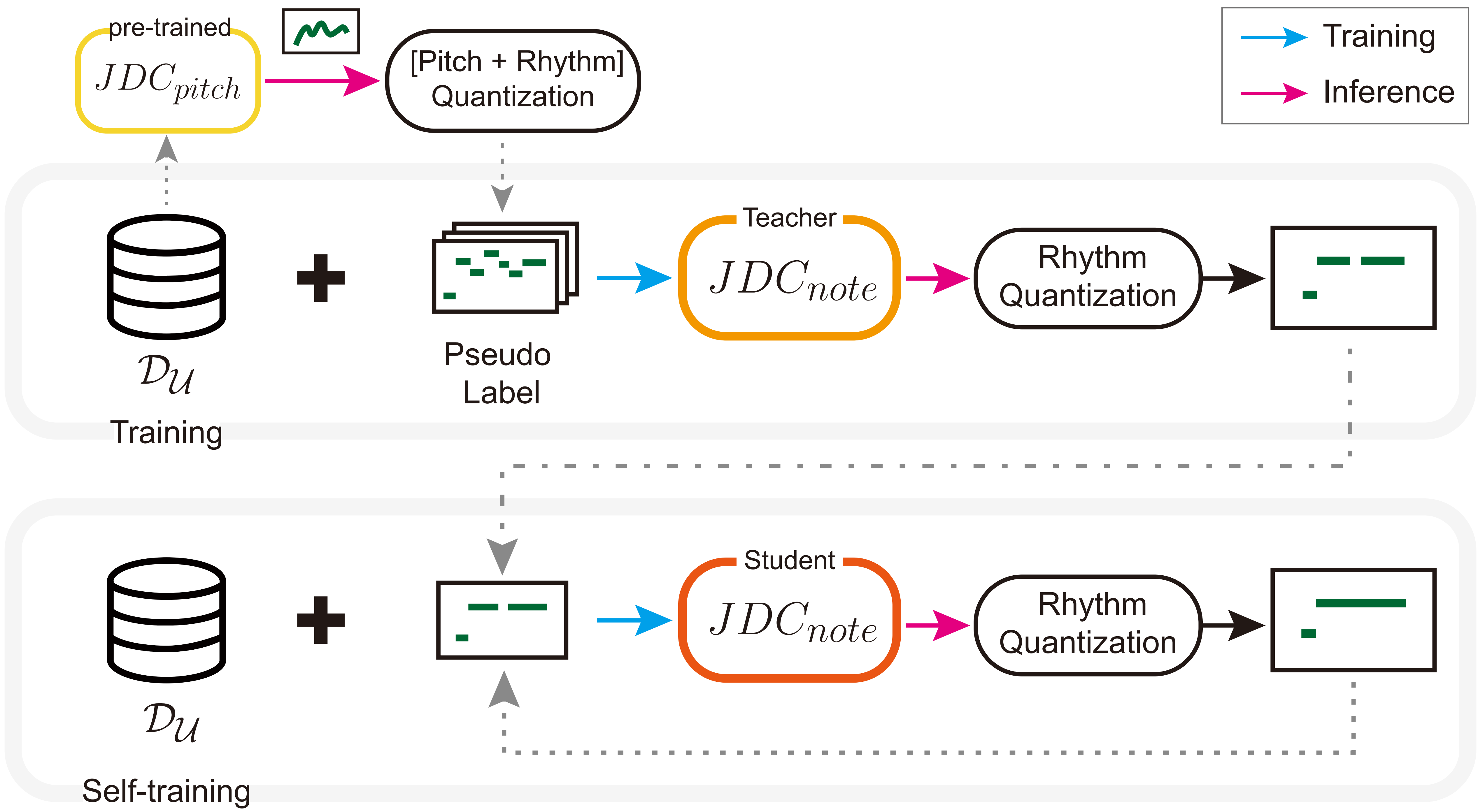
- **HZ** [4]**:** **Rule-based** model
- **VOCANO** [5]**:** **Semi-supervised** model
- **EFN** [6]**:** **Supervised** model

$JDC_{note}(L+U)$ > EFN > $JDC_{note}(L)$ > VOCANO > HZ

1. Given that $JDC_{note}(L)$ was also trained with the same training set that was used in EFN, the two models seem to be comparable to each other.
2. $JDC_{note}(L+U)$ pushes the accuracy levels higher, achieving best performances.

I Singing Transcription from Polyphonic Music

: Conclusion



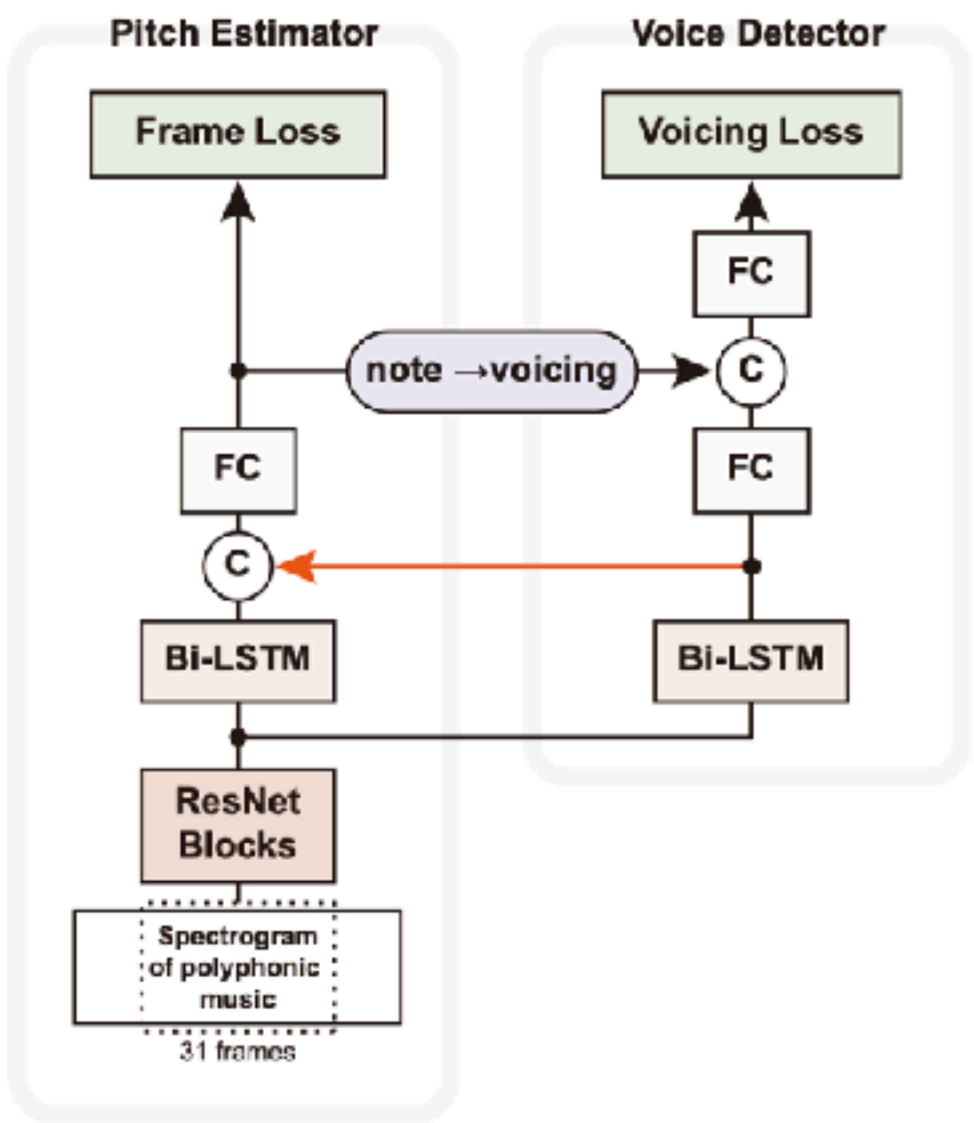
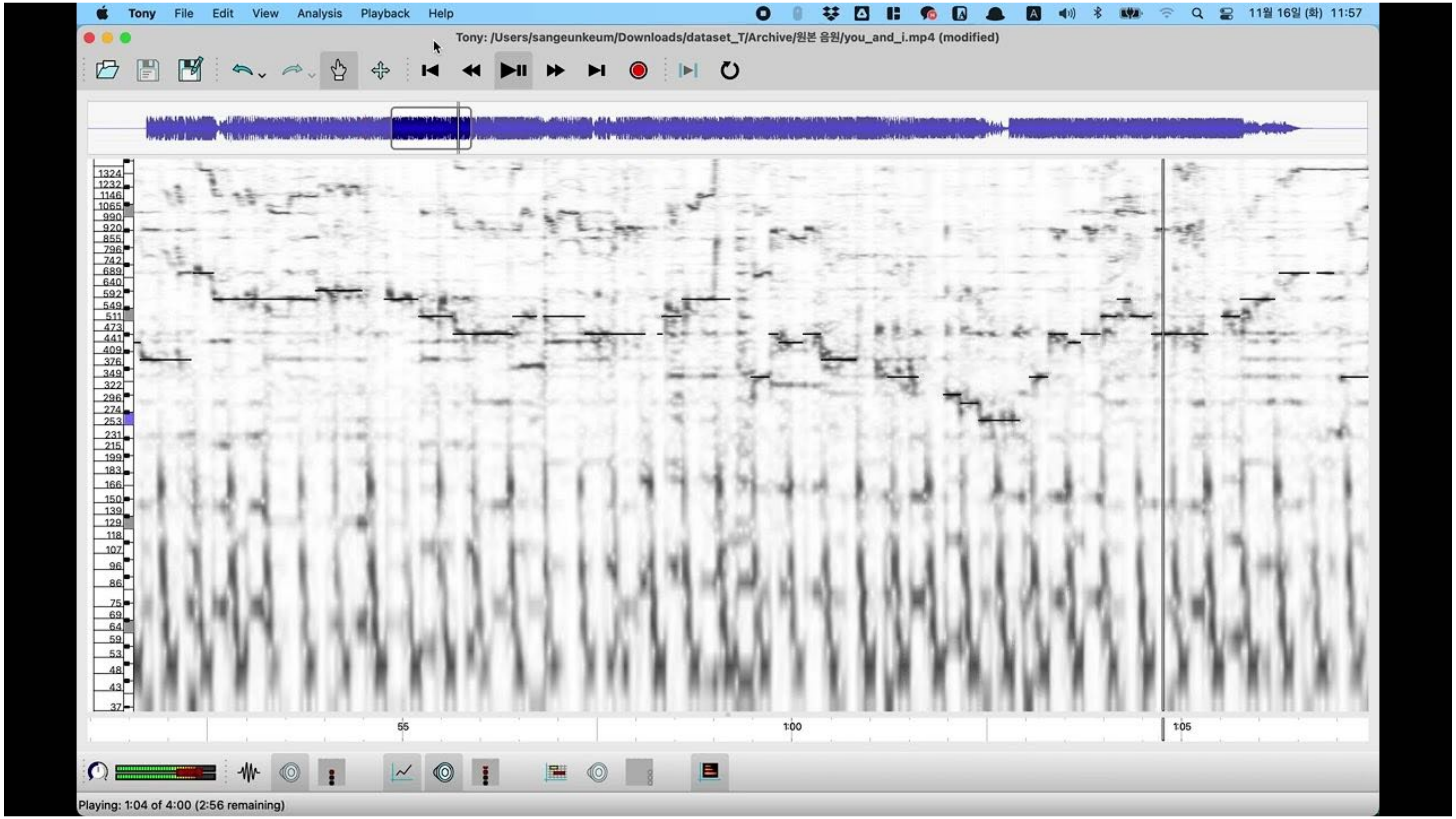


Fig. 2. The model architecture for JDC_{note}. "C" indicates feature concatenation.

I Demo


Tony File Edit View Analysis Playback Help
11월 16일 (화) 11:57
Tony: /Users/sangeunkeum/Downloads/dataset_T/Archive/원본 음원/you_and_i.mp4 (modified)
55
1:00
1:05
Playing: 1:04 of 4:00 (2:56 remaining)

#5829

Pseudo-Level Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music

Sangeun Kum[1], Jongpil Lee[1], Keunhyoung Luke Kim[1], Taehyoung Kim[1], Juhan Nam[2]

[1] Neutune Research, Seoul, South Korea
[2] Graduate School of Culture Technology, KAIST, Daejeon, South Korea

icassp 2022 Singapore

Neutune KAIST