#5829

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

Neutune
KAIST

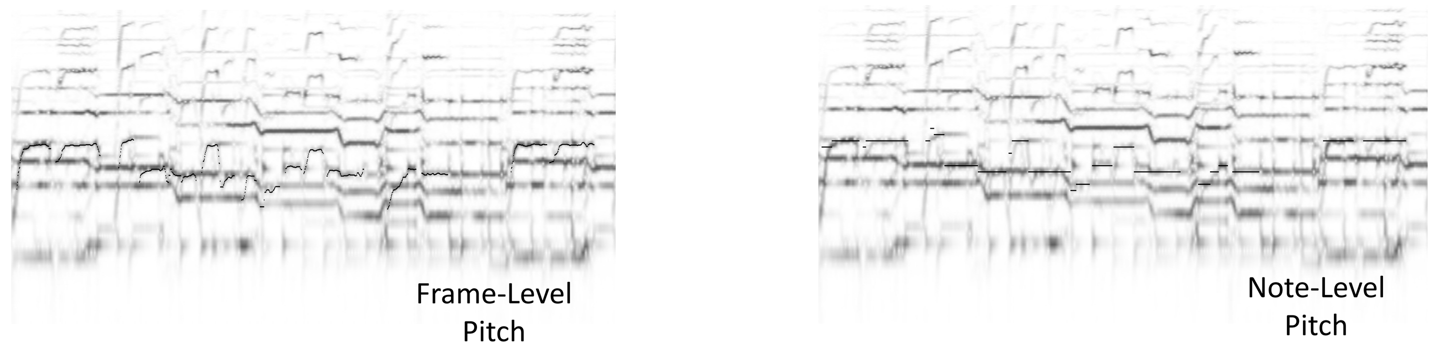
Sangeun Kum[1], Jongpil Lee[1], Keunhyoung Luke Kim[1], Taehyoung Kim[1], Juhan Nam[2]

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

icassp 2022 Singapore

Introduction

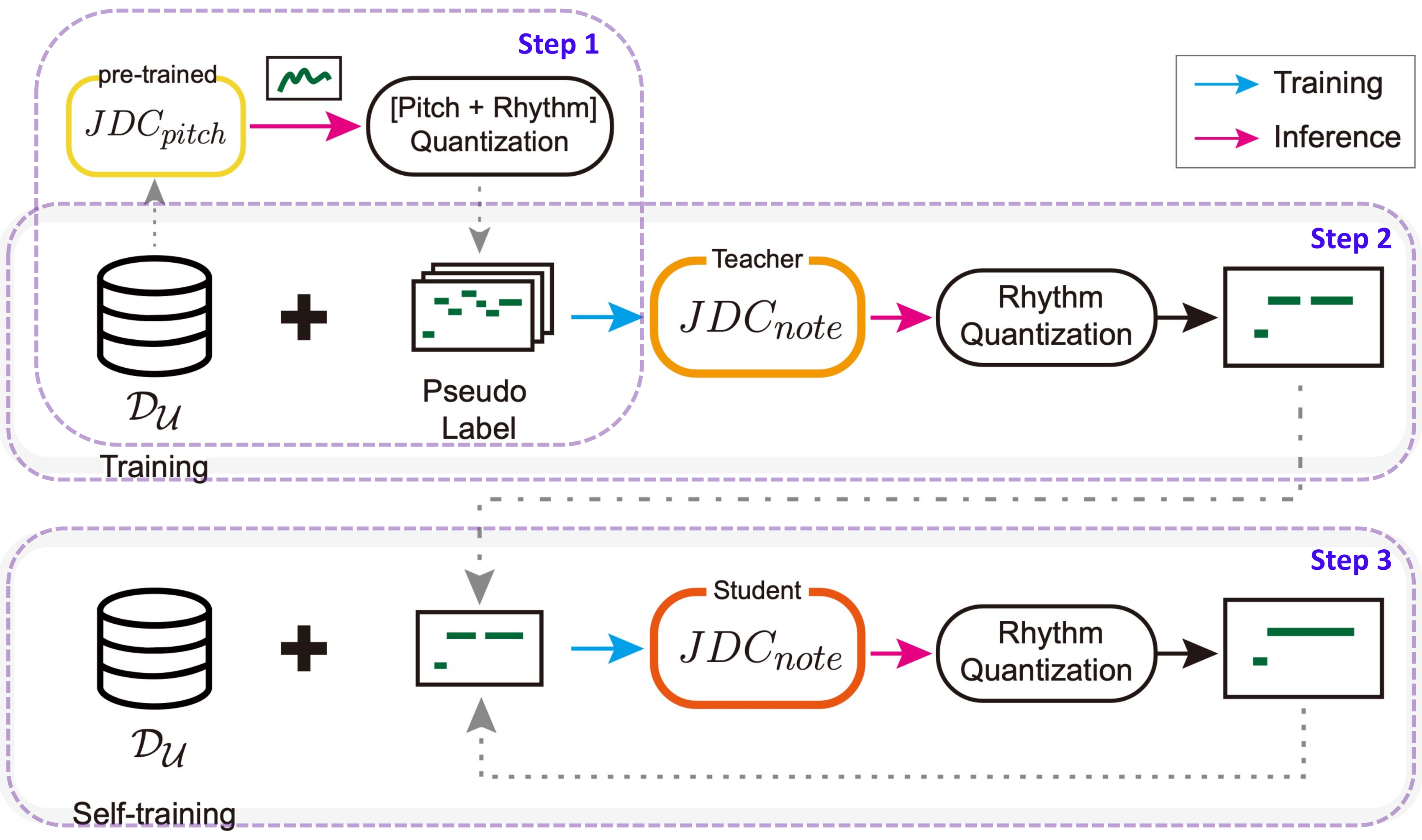
- STP includes several **sub-tasks**:
  1. Singing voice detection
  2. Singing pitch estimation
  3. Note-level segmentation
  4. Onset/offset detection



- Major obstacle to **Singing Transcription** from **Polyphonic music (STP)** = Lack of large-scale **note-level** labeled data for **VOCALS**

>> Contribution

1. To obtain effective pseudo-labels, we **use** vocal pitch estimation model to predict frame-level label and **convert** it to note-level label.
2. The proposed method (pseudo labeling, teacher-student framework, and JDC network) can achieve comparable results to the previous work using **only** unlabeled data, even if there is **no** source separation algorithm.
3. With additional labeled data, it achieves better performance than the model trained with only labeled data.



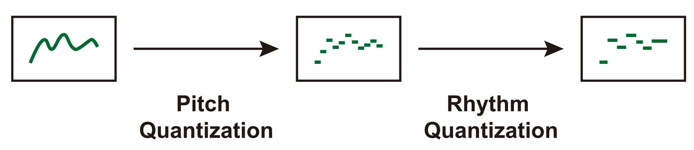
Method

Dataset

Labeled Dataset
- Cmedia (100): test
- MIR-ST500 (500) [1] : training

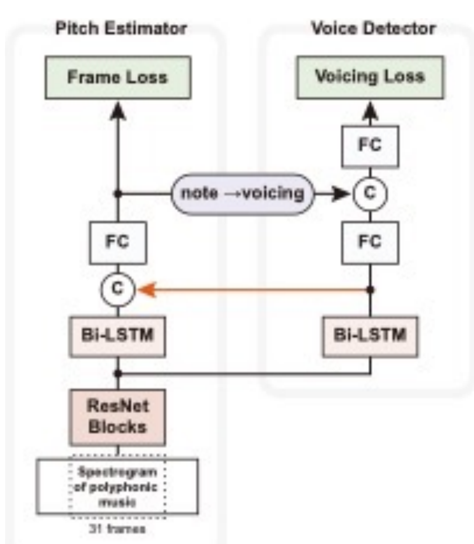
Unlabeled dataset
- In-house (2000): training
- FMA (168,000): training

[Step 1] Making pseudo labels using vocal pitch estimation model from Unlabeled dataset



Pitch + Rhythm Quantization
- Rounds the continuous pitch to semi-tone
- Smoothing the quantized pitch with a series of three median filters
- Remove small fragments

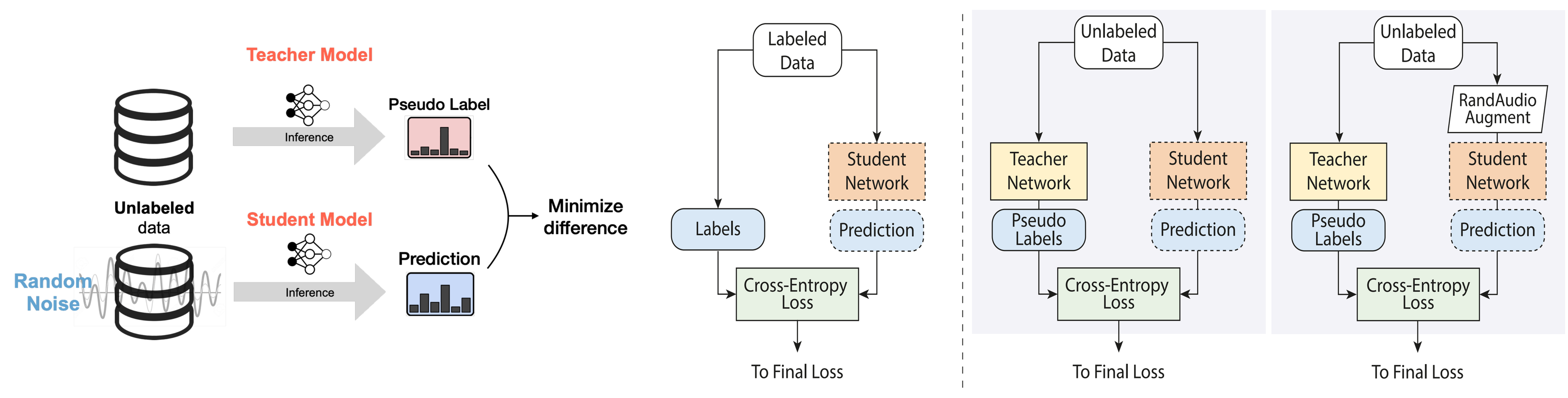
[Step 2] Training Model



- The model architecture for STP is based on the joint detection and classification (**JDC**) model [2]
- Training teacher model for singing transcription (JDC_{pitch}) using pseudo label from JDC_{note}

[Step 3] Teacher-Student Framework [3] for singing transcription

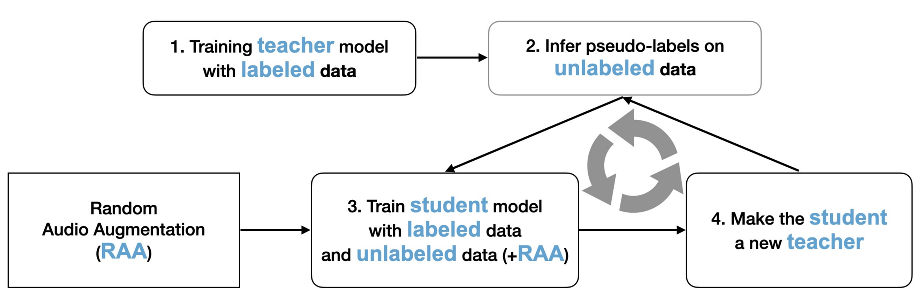
- Noisy Student



(a) Basic Teacher-Student

(b) Noisy Teacher-Student

- Iterative Training



Experiments

1. Comparison of Pitch Estimation Models

	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**
  : Vocal melody extraction from **polyphonic** music
- **Demucs**
  : music source separation
- **CREPE** [4]
  : pitch estimation from **monophonic** music

JDC > Demucs + CREPE

: Separation algorithms **cannot separate only the main vocal melody** and polyphonic vocals are still remained
→ Low performance

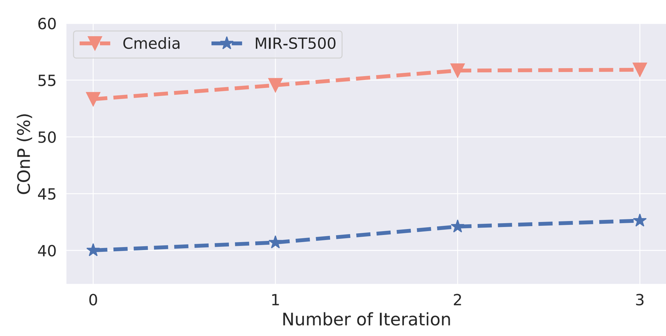
2. Basic Teacher-Student VS. Noisy Student

	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

3. Iteration of Self-Training



Iterative Training

: The performance continuously increases up to 2 iterations

4. Comparison with Supervised and Semi-Supervised Models

	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 [5]: Rule-based model
VOCANO [6]: Semi-supervised model
EFN [1] : Supervised model

EFN > **JDC(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

JDC(L+U) > EFN > JDC(L) > VOCANO > HZ

1. Given that **JDC(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(U+L)** pushes the accuracy levels higher, achieving best performances.

5. Demo video

1. https://tinyurl.com/yyxxsbyl
2. https://tinyurl.com/y4nnqs2k

Conclusion

- We presented a method for STP that uses **pre-trained vocal pitch estimation models** and **unlabeled datasets**.
- The method **converts the frame-level pseudo labels to note-level** and augments the label quality through **self-training** in the teacher-student framework.
- The unsupervised model trained through the proposed method can **achieve comparable results** to the previous works
- With **additional** labeled data, it **achieves better performance** than the model trained with only labeled data.

Reference

[1] Wang, J., & Jang, J., "On the preparation and validation of a large-scale dataset of singing transcription," in Proc. ICASSP, 2021
[2] Kum, S., & Nam, J., Joint detection and classification of singing voice melody using convolutional recurrent neural networks. Applied Sciences, 2019
[3] Kum, S., Lin, J. H., Su, L., & Nam, J.. Semi-supervised learning using teacher-student models for vocal melody extraction. ISMIR, 2020
[4] Kim, J. W., Salamon, J., Li, P., & Bello, J. P. Crepe: A convolutional representation for pitch estimation, ICASSP , 2018
[5] He, Z. & Feng, Y., "Singing transcription from polyphonic music using melody contour filtering," Applied Sciences, 2021
[6] Hsu, J. & Su, L., "VOCANO: A note transcription framework for singing voice in polyphonic music," in Proc. ISMIR, 2021