Semi-Supervised Learning Using Teacher-Student Models for Vocal Melody Extraction

Sangeun Kum[1], Jing-Hua Lin[2], Li Su[2], Juhan Nam[1]

[1] Graduate School of Culture Technology, KAIST, South Korea
[2] Institute of Information Science, Academia Sinica, Taiwan

ISMIR MTL2020

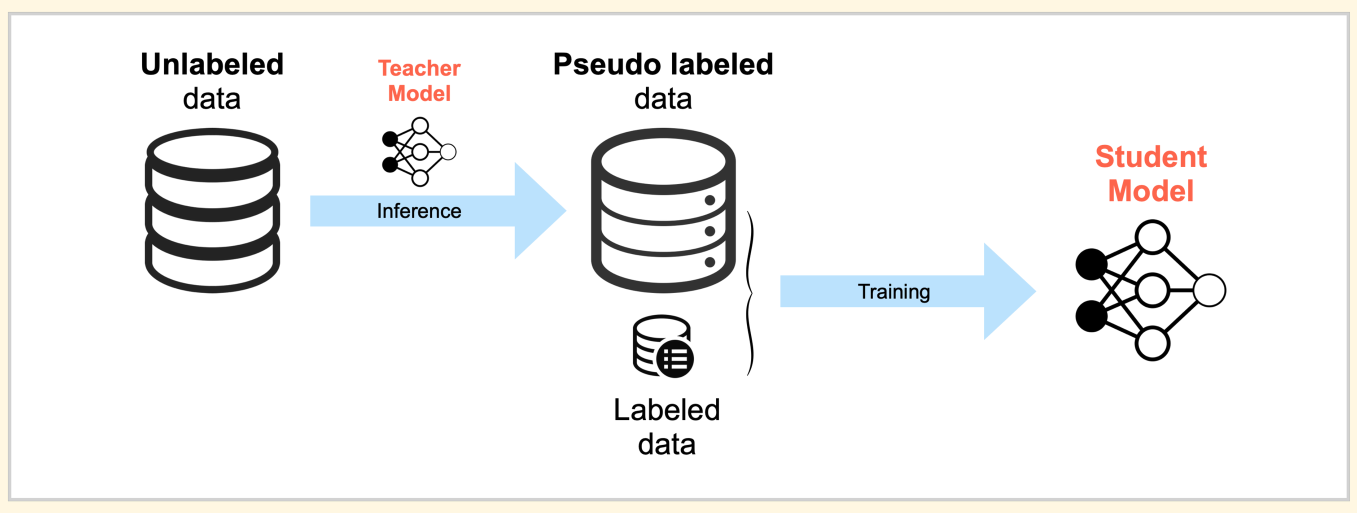
Summary

Problem

- Not enough labeled dataset.
- Pitch labeling is extremely laborious and costly

Contribution

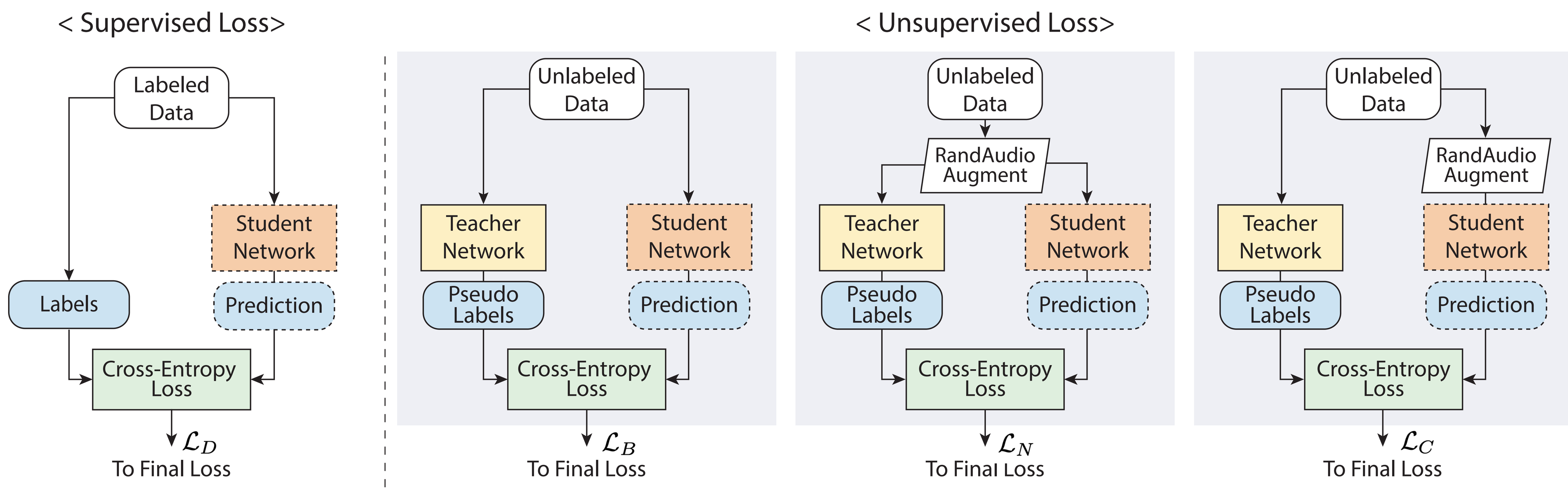
- We present the Semi-Supervised Learning (SSL) methods for vocal melody extraction leveraging large-scale unlabeled music datasets.
- We compare three setups of teacher-student models along with various audio data augmentation techniques. We show the model with the consistency regularization is most effective.
- We investigate effective SSL strategies by exploring joint training, the size of unlabeled data, and the number of self-training iterations.



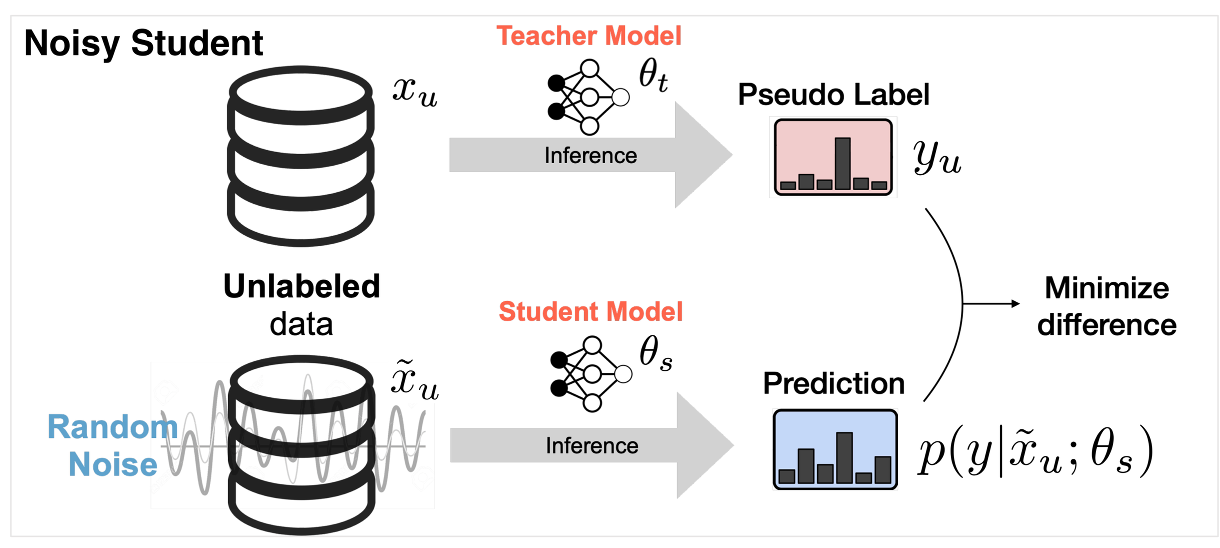
Teacher-Students Models

Self-training in the teacher-student framework :

1. The **teacher** model is first trained with labeled data
2. The **student** model is trained with artificial labels generated from the teacher model using unlabeled data.
3. We repeat the same pseudo labeling and the training with a new student model



(a) Basic Teacher-Student (b) Noisy Teacher-Student (c) Noisy Student



$$\mathcal{L}_C = \mathcal{L}_D + \frac{1}{M}\sum_{u=1}^{M} H(y_u, p(y|\tilde{x}_u;\theta_s))$$

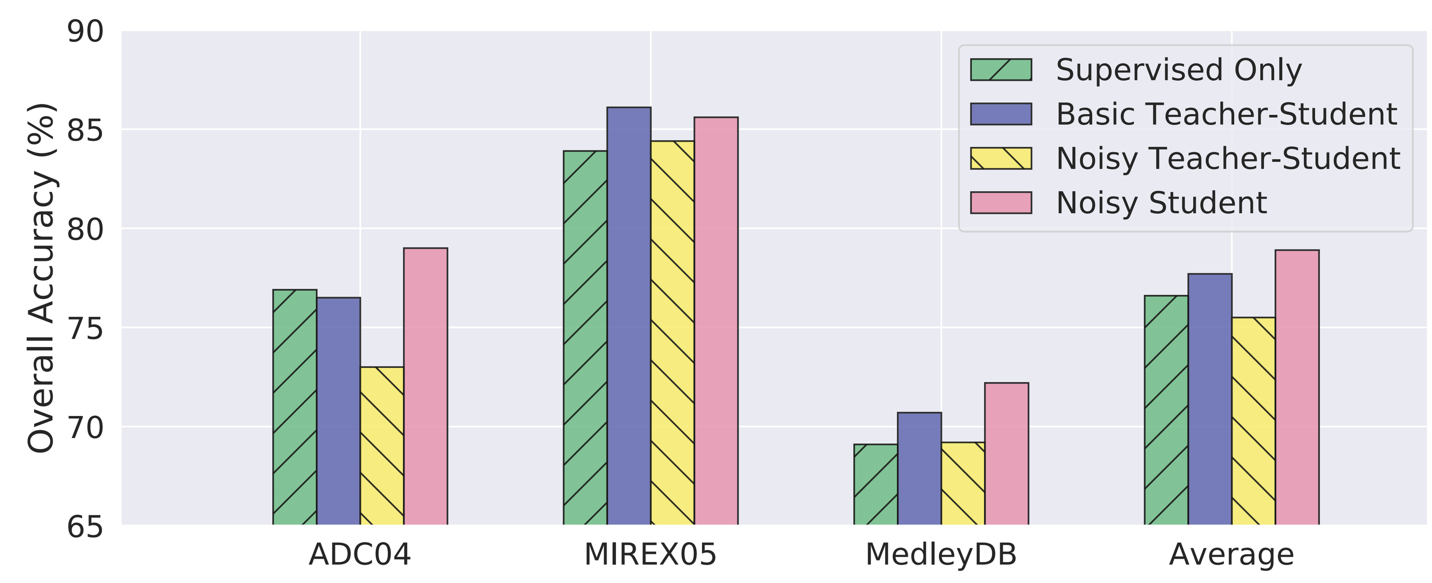


Fig1. Comparison with supervised-learning model and three student models on three test sets.

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

Data Selection

	Dataset	Number of Tracks	Total Length
Training (Labeled)	RWC	100	6h 47m
	MedleyDB	61	2h 39m
	iKala	262	2h 6m
Training (Unlabeled)	In-house	535	6h 21m
	FMA_small	3,521 / 8,000	25h / 60h
	FMA_medium	10,639 / 25,000	89h / 208h
	FMA_large	40,505 / 106,574	337h / 888h
Test	ADC04	12	4m
	MIREX05	9	4m
	MedleyDB	12	43m
	AST218	218	14h 53m

Table1. Description of datasets. In FMA, the two numbers indicate tracks with vocal (the vocal ratio above 0.3). We use our own Singing Voice Detector to include only vocal songs.

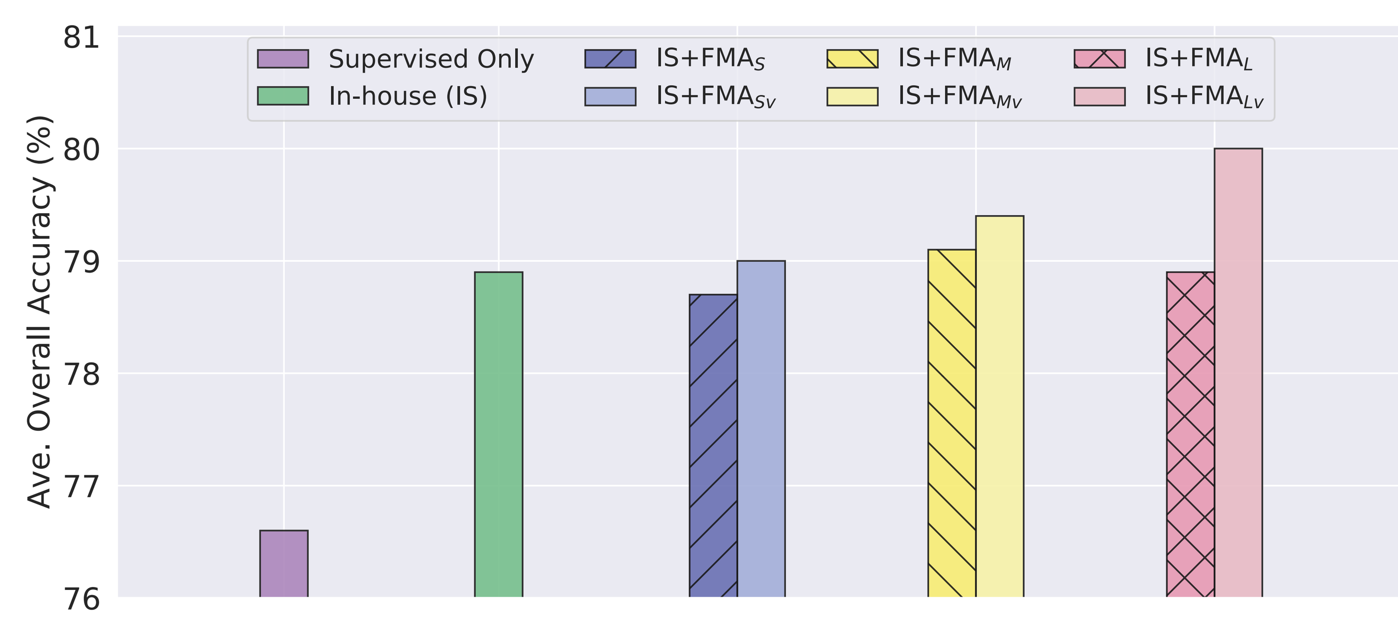
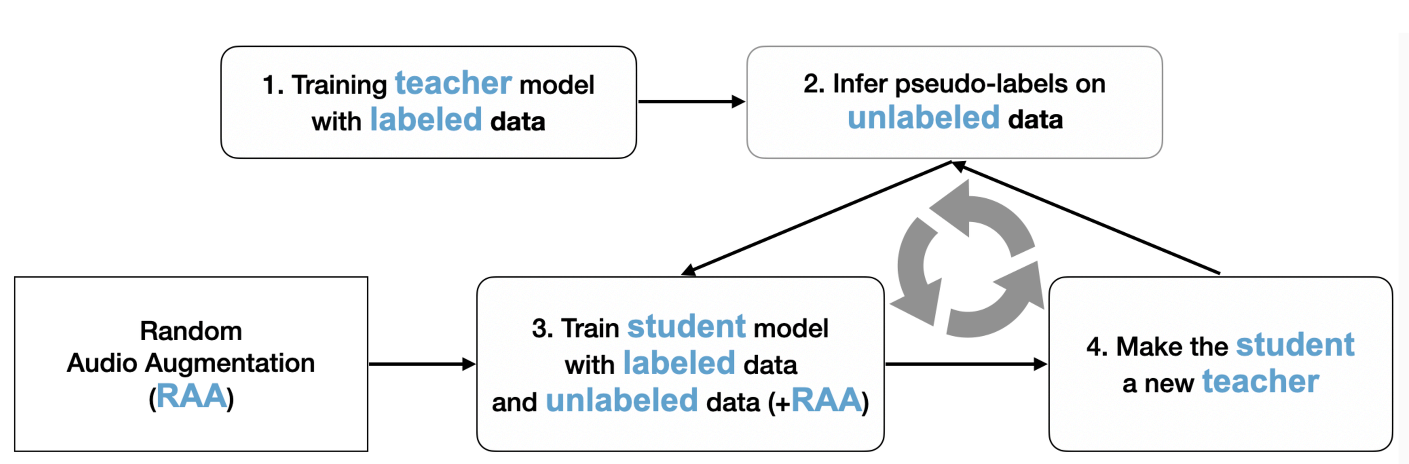


Fig2. Comparison with Noisy Students on varied sizes of unlabeled datasets.
The subscript 'v' denotes a selected subset of FMA whose vocal ratio exceeds a threshold.

→ **Effective SSL requires a large amount of unlabeled data with a similar distribution for labeled data.**

Iterative Training



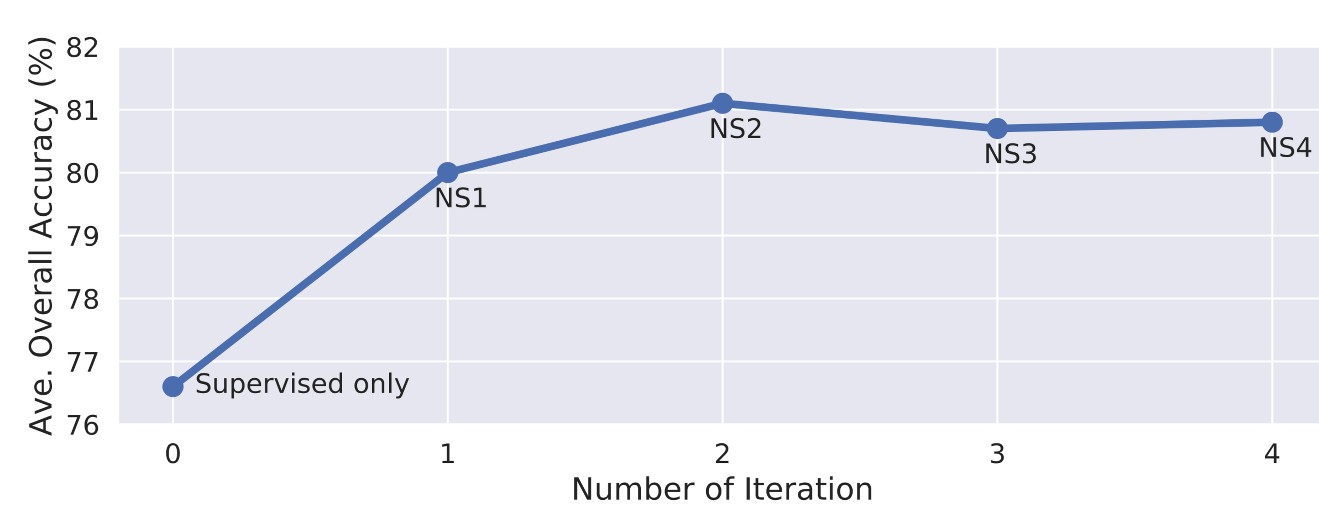


Fig3. Effect of iteration training for Noisy Students

→ **The performance continuously increases up to 2 iterations achieving the highest average OA**

Comparison with State-of-the-Arts

Methods	ADC04	MIREX05	MedleyDB	AST218
PatchCNN [1]	76.9 / 72.9	69.7 / 73.8	44.0 / 59.3	42.3 / 59.7
DSM [2]	89.2 / 72.2	87.7 / 80.1	80.6 / 75.4	38.9 / 68.3
SegNet [3]	88.7 / 83.3	82.6 / 80.0	70.6 / 75.5	41.5 / 68.1
JDC [4]	**90.6 / 83.5**	**91.4 / 87.4**	72.7 / 78.1	55.8 / **75.4**
Baseline	78.7 / 76.8	79.9 / 81.5	57.2 / 70.7	**56.3** / 69.7
Proposed (NS)	90.4 / 82.2	90.4 / 85.9	**76.3 / 79.2**	54.2 / 74.2

Table2. Vocal melody extraction results in terms of (**RPA / OA**) of the proposed and other methods on various test sets. (Baseline: supervised only)

Conclusion

- This study provides a **framework of semi-supervised learning** using the **teacher-student model** for vocal melody extraction.
- The **Noisy Student model** is the most effective and robust to real-world music.
- **Large-scale unlabeled data** is effective when they are properly selected.
- **Iterative training** for the teacher-student model helps improve performance.
- The effectiveness of the proposed method by evaluating it on **artificial large-scale test data** generated from automatically annotated multitrack data.
- Our method **can be extended to other MIR tasks** that suffer from the **lack of labeled data** such as automatic music transcription and chord recognition.

[1] L. Su, "Vocal melody extraction using patch-based CNN," in *Proc. ICASSP*, 2018, pp. 371–375.
[2] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for f0 estimation in polyphonic music," in *Proc. ISMIR*, 2017.
[3] T.-H. Hsieh, L. Su, and Y.-H. Yang, "A streamlined encoder/decoder architecture for melody extraction," in *Proc. ICASSP*. IEEE, 2019, pp. 156–160.
[4] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

Sangeun Kum(keums@kaist.ac.kr)

https://github.com/keums/melodyExtraction_SSL