Singer Identity Representation Learning using Self-Supervised Techniques

Bernardo Torres[1], Stefan Lattner[2], Gael Richard[1]

[1]LTCI, Telecom Paris, Institut Polytechnique de Paris.

[2]Sony Computer Science Laboratories Paris

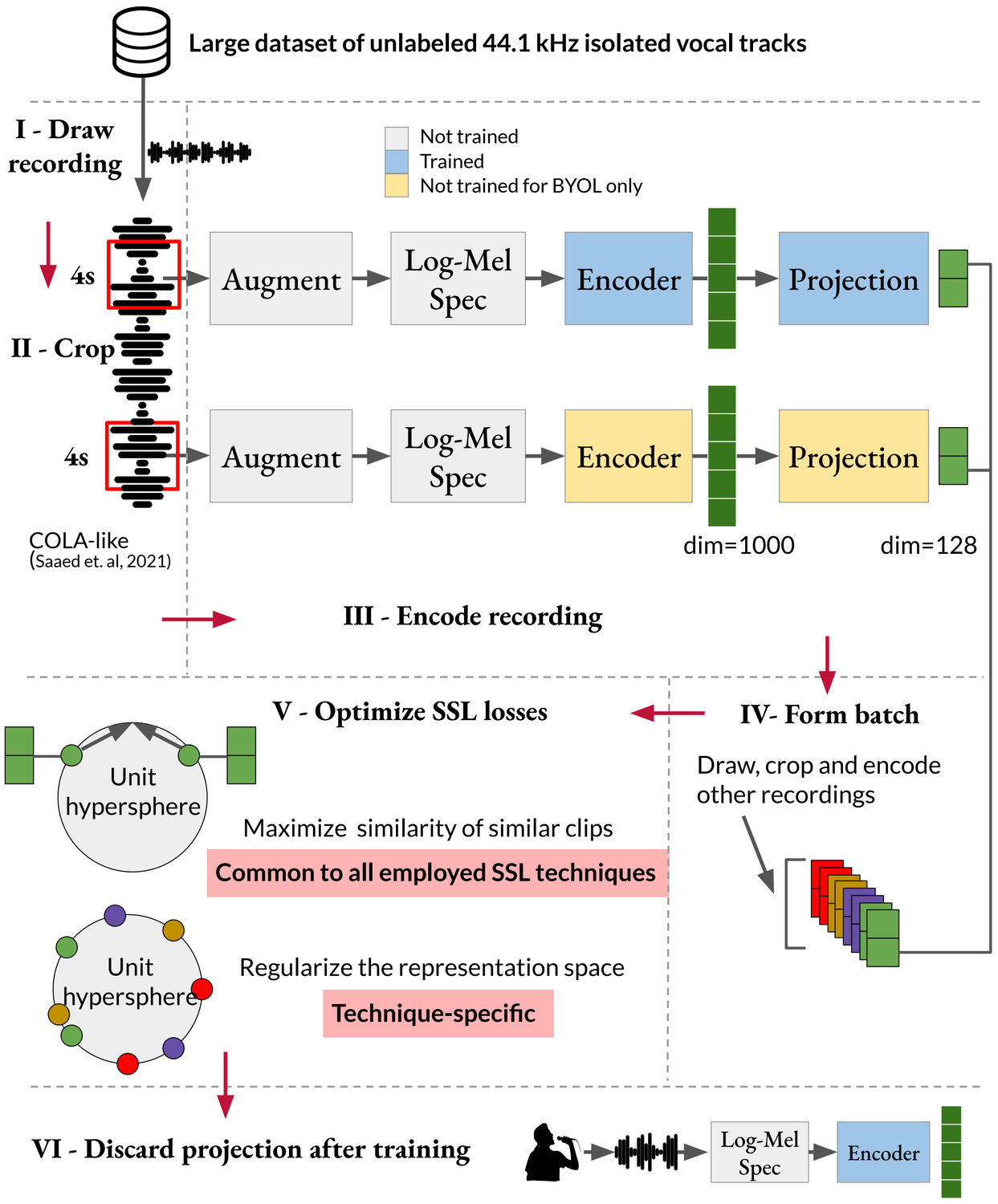
Introduction

Goal: obtain time-invariant identity representations from singing voice

Existing models from speech literature → How well do models trained on speech **generalize to singing voice**?

Train identity extraction encoders → Lack of large labelled singing voice datasets → Can we train better models using **Self-supervised Learning (SSL)**?

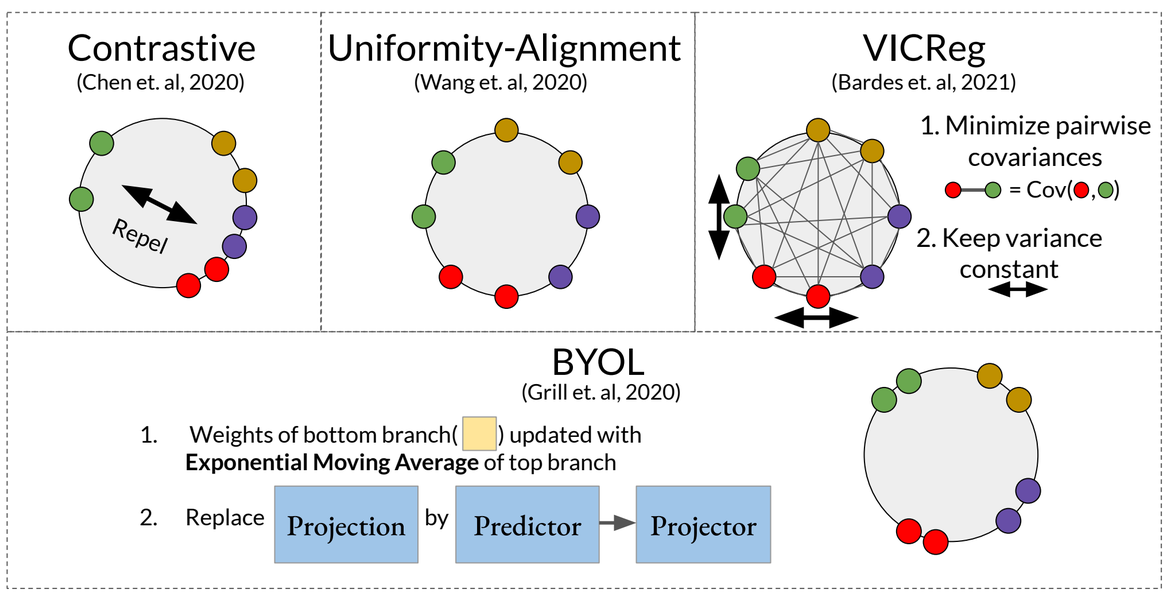
Overview and training



Self-supervised techniques

Common idea: representations from the same recording should be close

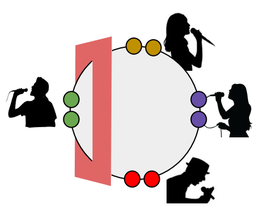
We trained models with the following SSL techniques:



Evaluation

Singer identification

Linear classifier



Trained on embedding space (frozen encoder)
Test accuracy of N-fold cross validation

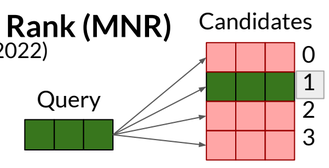
Singer similarity

Equal Error Rate (EER)

same/different binary classification

Mean Normalized Rank (MNR)
(Lattner, 2022)

Rank ground-truth match by similarity with query



Evaluation on out-of-domain public datasets

Corpus	Language	#Hours	#Singers	Type	
VCTK	English	44	110	Speech	Yamagishi et. al, 2019
NUS-48E	English	1.91	12	Speech/Singing	Duan et.al, 2013
VocalSet	English	10.1	20	Singing	Wilkins et al., 2018
M4Singer	Chinese	29.77	20	Singing	Zhang et. al, 2022

Baselines

Pre-trained, publicly available speech models → Supervised speaker verification; General purpose SSL

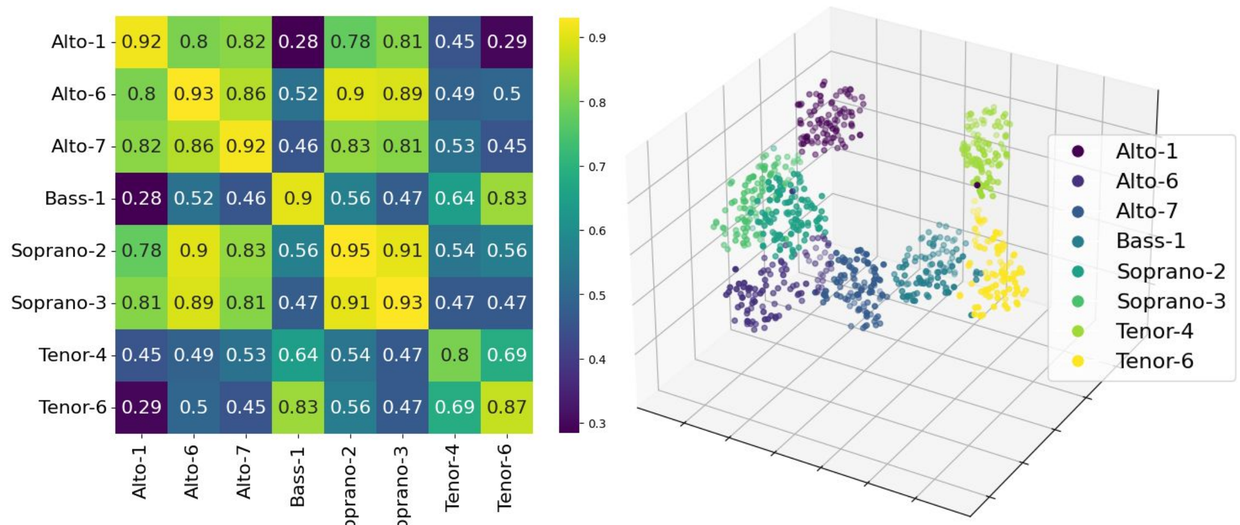
Model	#Params	SR	Dim.	Backbone	
GE2E	1.4M	16	256	LSTM	Wan et. al, 2018
F-ResNet	1.4M	16	512	ResNet-34	Chung et.al, 2020
H/ASP	8.0M	16	512	ResNet-34	Kwon et al., 2022
Wav2Vec-base	95M	16	12X768	Wav2Vec 2.0	Baevski et. al, 2020
XLSR-53	300M	16	24X1024	Wav2Vec 2.0	Conneau et. al, 2021
Ours	5.0M	44.1	1000	EfficientNet-B0	

Results

Speech baselines on singing voice

- ↓ compared to speech data.
- Still work reasonably well; except for VocalSet
- SSL baselines: performed bad on similarity, well on identification

Trained SSL identity encoders



Left: Average similarity score between singers over 100 4s clip draws for each singer (M4Singer dataset)
Right: T-SNE visualization for the same embeddings in 3D (original dimensionality is 1000)

The trained SSL models were **comparable or superior** to baselines

Comparison of SSL techniques → Best on out-of-domain: **BYOL**; Best In-domain: **Contrastive**

Conclusion

- **Trained** identity encoders using **Self-Supervised Learning (SSL)**
- **Dataset**: large unlabeled singing voice isolated recordings
- **Comparison** with publicly available pre-trained speech models
- **Evaluation** on singer identification and similarity tasks
- A big gap still exists for challenging datasets
- Release of code and trained models