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

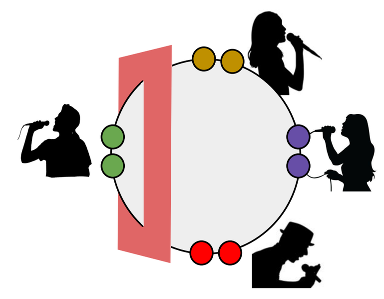
Lack of large labelled singing voice datasets

Can we train better models using **Self-supervised Learning (SSL)**?

Evaluation

Singer identification

Linear classifier



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

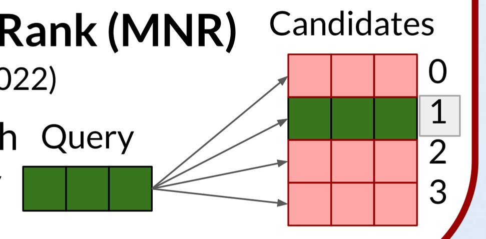
Singer similarity

Equal Error Rate (EER)

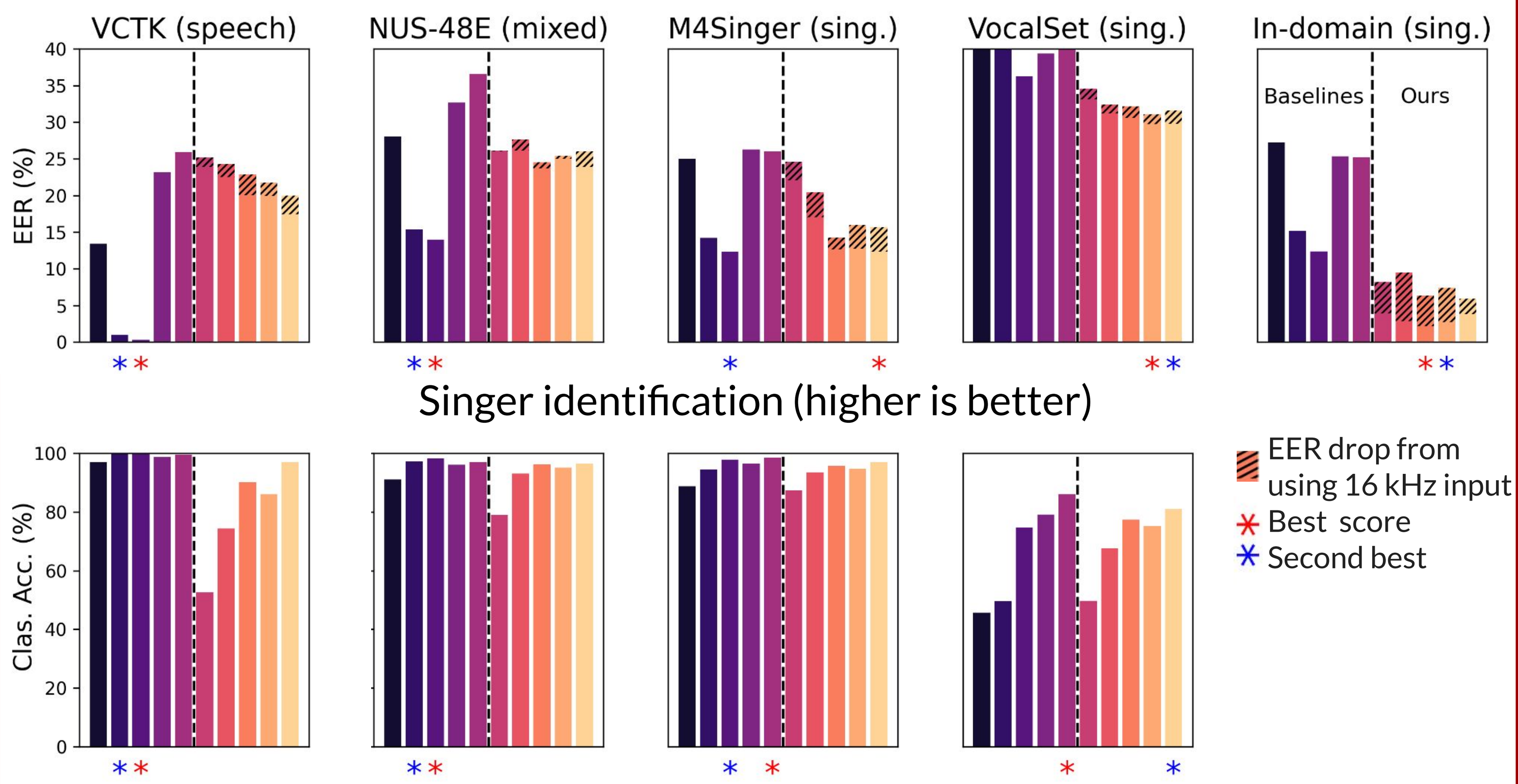
open-set same/different binary classification

Mean Normalized Rank (MNR)
(Lattner, 2022)

Rank ground-truth match by similarity with query



Results



Summary of results for singing voice:

Baselines

Speech supervised
- GE2E
- F-ResNet
- H/ASP

↓ compared to evaluation on speech data

Work reasonably well;except for VocalSet

Speech SSL
- Wav2Vec-base
- XLSR-53

Bad on similarity, well on identification

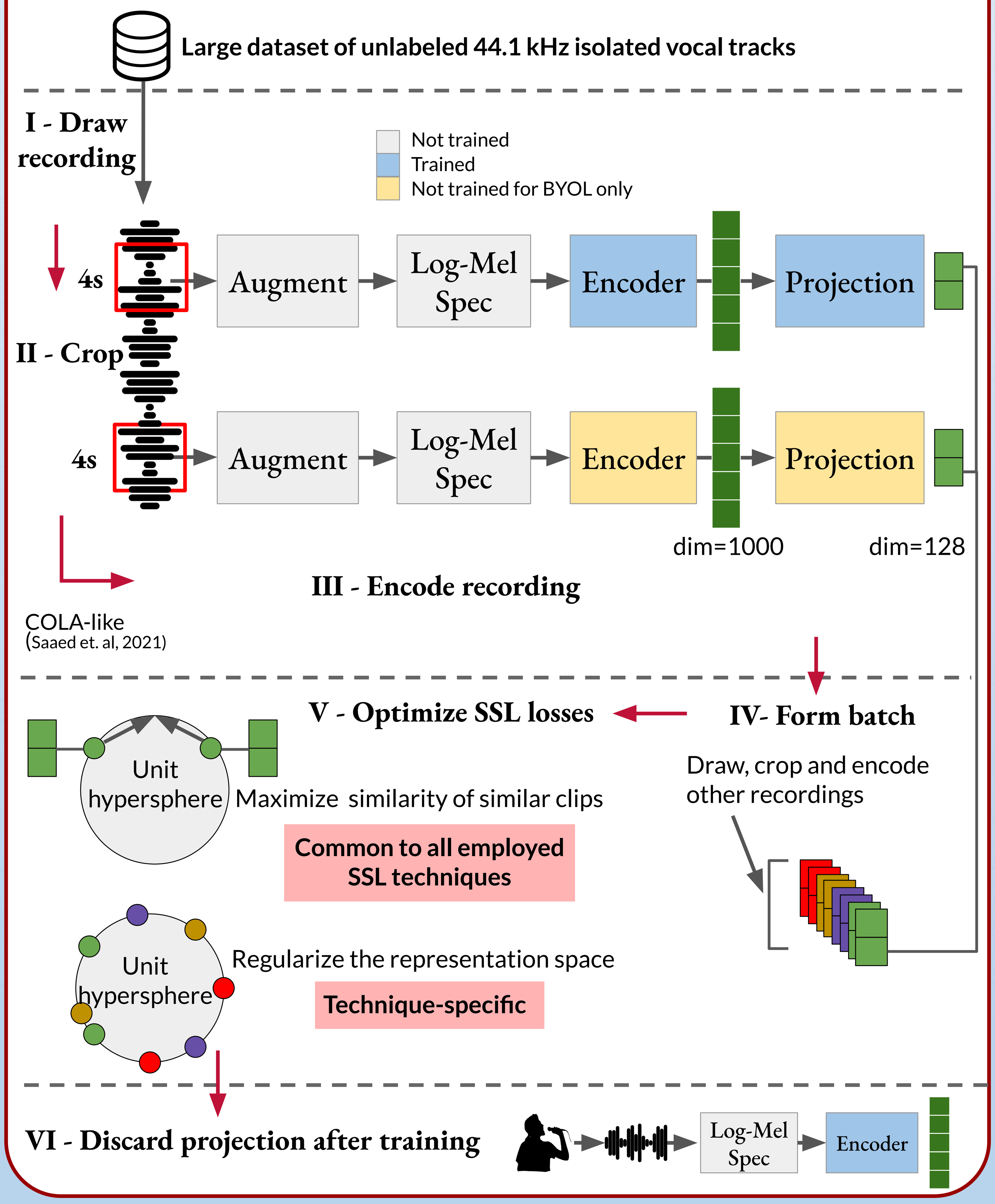
Trained SSL models
- VICReg
- UNIF
- CONT
- CONT-VC
- BYOL

Comparable or superior to baselines

Best on out-of-domain: **BYOL**

Best In-domain: **Contrastive**

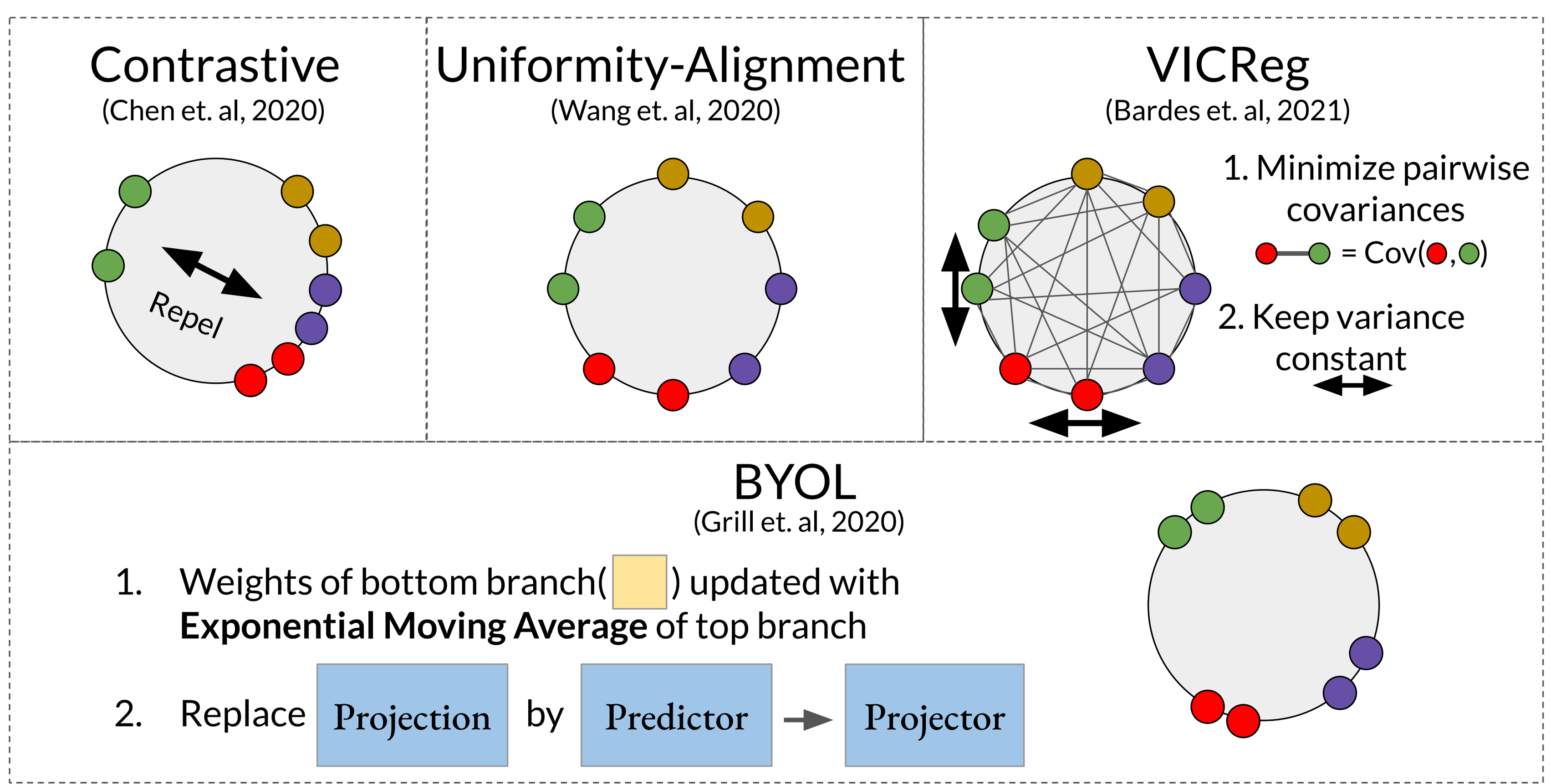
Overview and training



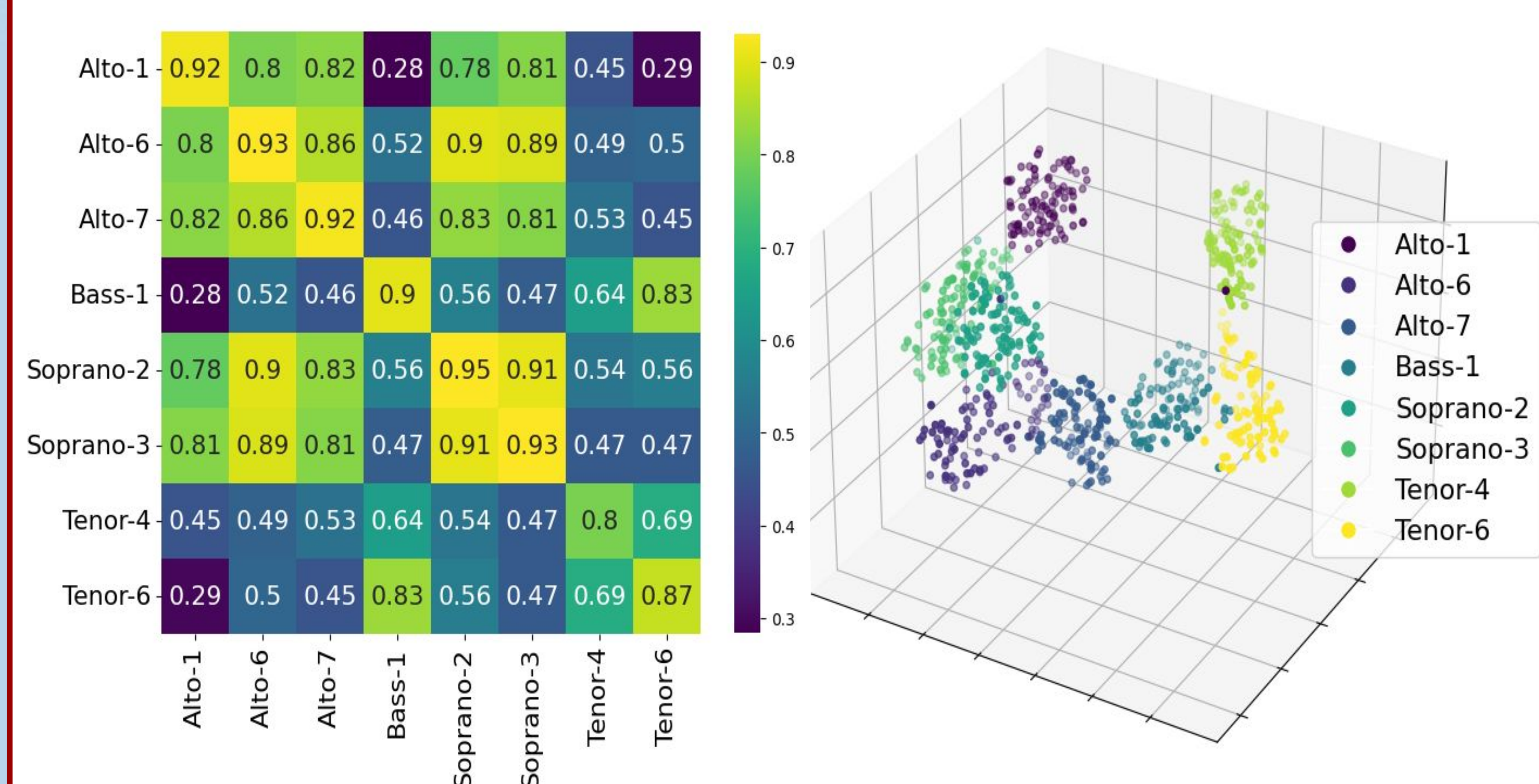
Self-supervised techniques

Common idea: representations from the same recording should be close

We trained models with the following SSL techniques:



Qualitative analysis



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)

Conclusion

- **Trained** identity encoders using **Self-Supervised Learning (SSL)**
- **Dataset**: unlabeled isolated singing voice 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