**Singer Identity Representation Learning Using Self-Supervised Techniques**

Bernardo Torres\(^1\), Stefan Lattner\(^2\), Gael Richard\(^3\)

\(^1\) 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
- Train identity extraction encoders
- Lack of large labelled singing voice datasets

**How well do models trained on speech generalize to singing voice?**

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

---

### Overview and training

- **I - Draw recording**
  - Large dataset of unlabeled 44.1 kHz isolated vocal tracks
  - Not trained
  - Trained
  - Not trained for BYOL only

- **II - Crop**
  - 4s
  - Augment
  - Log-Mel Spec
  - Encoder
  - Projection

- **COLA-like (Bardes et al., 2021)**

- **III - Encode recording**

- **IV - Form batch**
  - Draw, crop and encode other recordings

- **V - Optimize SSL losses**
  - Unit hypersphere
  - Maximize similarity of similar clips
  - Common to all employed SSL techniques

- **VI - Discard projection after training**

- **Self-supervised techniques**

  Common idea: representations from the same recording should be close

  We trained models with the following SSL techniques:

  - **Contrastive** (Chen et al., 2020)
  - **Uniformity-Alignment** (Wang et al., 2020)
  - **VICReg** (Bardes et al., 2021)

  BYOL (Grill et al., 2020)

  1. Weights of bottom branch updated with Exponential Moving Average of top branch
  2. Replace Projection by Predictor

---

### Evaluation

#### Singer identification

**Linear classifier**

- Equal Error Rate (EER)

- Trained on embedding space (frozen encoder)

- Test accuracy of N-fold cross validation

- Rank ground-truth match by similarity with query

---

### Results

**Singer similarity (lower is better)**

- Comparison
  - A big gap still exists for challenging datasets
  - Release of code and trained models

**Evaluation**

- Baselines
  - Work reasonably well except for VocalSet
  - Comparable or superior to baselines

**Trained SSL models**

- Best In-domain: VocalSet

**Quality 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