CVSM: Contrastive Vocal Similarity Modeling

Christos Garoufis, *Member, IEEE*, Athanasia Zlatintsi, and Petros Maragos, *Life Fellow, IEEE*

***Abstract*—The availability of large, unlabeled datasets across various domains has contributed to the development of a plethora of methods that learn representations for multiple target (downstream) tasks through self-supervised pre-training. In this work, we introduce CVSM (Contrastive Vocal Similarity Modeling), a contrastive self-supervised procedure for music signal representation learning in the audio domain that can be utilized for musical and vocal similarity modeling. Our method operates under a contrastive framework, maximizing the similarity between vocal excerpts and musical mixtures containing the same vocals; we devise both a *label-informed* protocol, leveraging artist identity information to sample the contrastive pairs, and a *label-agnostic* scheme, involving artificial mixture creation from randomly sampled vocal and accompaniment excerpts, which are paired with vocals from the same audio segment. We evaluate our proposed method in measuring vocal similarity both objectively, through linear probing on a suite of appropriate downstream tasks, and subjectively, via conducting a user study consisting of pairwise comparisons between different models in a recommendation-by-query setting. Our results indicate that the representations learned through CVSM are effective in musical and vocal similarity modeling, outperforming numerous baselines across both isolated vocals and complete musical mixtures. Moreover, while the availability of artist identity labels during pre-training leads to overall more consistent performance both in the evaluated downstream tasks and the user study, a label-agnostic CVSM variant incorporating hybrid pre-training with real and artificial mixtures achieves comparable performance to the label-informed one in artist identification and perceived vocal similarity.**

***Index Terms*—Music representation learning, contrastive learning, music similarity, vocal similarity.**

I. Introduction

THROUGHOUT all history, from the musical practices of the ancient world and classical orchestral music to the heavily produced music of our era, musical pieces have always constituted multi-faceted expressions of art. Multiple performers, each using a different instrument or their voices, cooperate to create compound sounds, usually in harmonic and rhythmic consonance. However, each of the co-playing sources exhibits its own distinct characteristics, influencing the musical piece in a specific way, such as defining its rhythm, exhibiting the technical proficiency of the performers, or eliciting a certain mood. Thus, successfully modeling the similarity between particular musical sources is of profound importance, with applications ranging from music analysis [1] to music recommendation [2].

Possibly the most expressive "instrument" in a musical piece, thanks to the range of different mental states it can convey, is the human voice [3]. Perceptually, vocal similarity can be traced to numerous high-level attributes, such as the phonation, the articulation, the timbral clarity, as well as the vocal prosody [4], [5]. These high-level perceptual traits can be connected partially to low-level acoustic features translating into timbral similarity, including the fundamental frequency (F0), formant frequencies (F1-F4), their normalized amplitude ratios, and measures related to the harmonic regularity, volume, and spectral noise [6], [7], [8], [9]. Thus, early attempts to model vocal similarity [5] have focused on extracting, or annotating, descriptive features from vocal clips, and training machine learning classifiers to directly predict them, creating thus an acoustic feature space that could be used for vocal sample retrieval. More recently, it was shown that such feature spaces can be obtained by neural networks that have been directly trained from speech, with a suitable end-to-end objective [9].

However, modeling the singing voice entails a number of challenges compared to spontaneous speech, including higher frequency range, harmonics amplitude, rhythmicality and vowel duration [10], [11], [12]. This, coupled with the presence of instrumental accompaniment, has steered early attempts in modeling vocals in the presence of background music into necessitating either suitable feature selection, or application of a source separation pre-processing step [13], [14]. Another challenge regards the subjectivity of systematically evaluating the (dis)-similarity across different voices. As a result, apart from subjective listening tests [13], vocal similarity has been usually evaluated through proxy tasks, such as singer identification [12], [13], [15], [16], vocal technique recognition [12], [16], gender recognition [12] or vocal pitch estimation [16].

Recently, the availability of large, unlabeled datasets across various domains and the upsurge of deep learning, coupled with the upscaling of hardware resources, have given rise to the field of self-supervised learning (SSL). Under SSL frameworks, neural networks are trained to learn representations which discriminate between different input samples without having access to their target labels. SimCLR [17], which can be viewed as a batch-wide variant of triplet metric learning [18], constitutes

Received 9 July 2025; revised 31 December 2025; accepted 7 February 2026. Date of publication 30 March 2026; date of current version 20 May 2026. This work was supported by the EU's Horizon Europe research and innovation programme under Grant 101070381 (Project: PILLAR-Robots). The associate editor coordinating the review of this article and approving it for publication was Prof. Juhan Nam. *(Corresponding author: Christos Garoufis.)*

Christos Garoufis is with Robotics Institute, Athena Research Center, 15125 Marousi, Greece, and also with the School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece (e-mail: christos.garoufis@athenarc.gr).

Athanasia Zlatintsi is with Robotics Institute, Athena Research Center, 15125 Marousi, Greece (e-mail: nancy.zlatintsi@athenarc.gr).

Petros Maragos is with Robotics Institute, Athena Research Center, 15125 Marousi, Greece, also with the School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece, and also with HERON - Hellenic Robotics Center of Excellence, 15780 Athens, Greece (e-mail: petros.maragos@athenarc.gr).

Digital Object Identifier 10.1109/TASLPRO.2026.3677618

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

one of the most prominent SSL paradigms. In short, SimCLR-based methods aim to create a latent space where projections of different views of the same sample, typically generated by devising a suitable augmentation pipeline, are close to each other, while simultaneously being distant from projections of semantically different samples. Whilst originally developed for computer vision applications, SimCLR has been transferred to other domains [19], [20] or even multimodal settings [21], [22] by appropriate choices of the backbone encoder and the applied augmentations.

In the domain of musical audio, a number of research works have applied contrastive learning in order to model vocal similarity [12], [15] using clean vocal excerpts. However, the contrastive pre-training pipelines followed do not employ instrumental accompaniments, hindering generalization in the case of commercial music where background instruments are also present. Potential detours to this problem would involve coupling song excerpts with isolated vocals during the pre-training stage [23], leading to a latent space that successfully correlates attributes of the isolated vocals with the musical piece, or using jointly clean and mixture pairs [24] during pre-training, reducing the domain gap between musical mixtures and vocal excerpts. Nonetheless, neither of the above strategies guarantees invariance to instrumental accompaniment by disentangling the vocal properties in the learned latent space.

In this work, we present CVSM (Contrastive Vocal Similarity Modeling), a framework that can be utilized for vocal retrieval and vocal similarity modeling both in clean vocals and in-the-wild (i.e., in the presence of background instrumental accompaniment). The presented framework relies on pairing vocal excerpts with mixtures of vocal excerpts and instrumental accompaniment. It is based on MSCOL [23], but improves on it in terms of robustness, specificity and generalizability, adapting it to source-informative representation learning (in our case, the vocals). Our main contributions are the following:

- We leverage the availability of artist identity labels, as their utilization has proven effective in general-purpose music representation learning [25], [26], by proposing a *label-informed* pre-training scheme. In this case, contrastive pairs are created by matching excerpts of isolated vocals with full song excerpts, originating from the same artist,[1] deviating from [25] where both pair components are sampled from the complete musical mixture.
- Inspired by data augmentation techniques used for music source separation [27], [28], we also design a *label-agnostic* pre-training protocol, wherein we generate artificial musical mixtures as anchors by superimposing the positive vocal sample with an accompaniment from a different, randomly sampled musical piece. This way, we intend to create a latent space invariant to instrumental properties of the musical piece [29]; this approach is similar to [30], but applies itself in a contrastive setting. To alleviate the domain gap between the artificial mixtures used for pre-training and original music pieces, we also experiment with i) stochastic application of the proposed augmentation, selecting randomly (on-the-fly) either artificial or real musical mixtures during pre-training and ii) contrastive self-supervised fine-tuning, using solely pairs of real musical mixtures and the corresponding vocals.
- We explore the behavior of both *label-informed* and *label-agnostic* variants, obtaining insights into the extent artificial mixtures and artist-level labels are necessary for learning robust representations for vocal-oriented music understanding tasks, as well as the necessity of having access to high quality separated vocal tracks through using multiple [31], [32] open-sourced music source separation systems for vocal track acquisition.

The framework was pre-trained using the publicly available Music4All [33] dataset, and was evaluated both i) objectively, by training shallow downstream classifiers upon the learned embeddings in the tasks of gender identification and artist identification and similarity, and ii) subjectively, by evaluating the learned latent space of the framework in retrieving and recommending musical pieces with vocals similar to a given query through a listening test. Our findings indicate that CVSM can effectively model musical and vocal similarity across both isolated vocals and complete musical mixtures. In particular, the *label-informed* CVSM variant outperforms its respective baseline [25] in isolated vocals, while performing comparably to it in the presence of instrumental accompaniment. We also show that even without the availability of artist identity labels during pre-training, CVSM can create a latent space that effectively conveys artist identity information, outperforming label-agnostic baselines [19], [23], [34], [35] in artist identification and similarity. In fact, when pre-trained using a combination of real and artificial mixtures, the performance of the *label-agnostic* CVSM variant approaches the one achieved by label-informed pre-training in artist identification. These results are further corroborated by the user study, with the CVSM variant pre-trained with artist-guided sampling scoring the highest among all evaluated models in modeling both overall and vocal similarity in musical mixtures. Interestingly, despite its middling performance in overall perceived similarity, label-agnostic pre-training incorporating a hybrid strategy of utilizing both real and artificial musical mixtures during pre-training performs comparably to label-informed models in encapsulating perceived vocal similarity, suggesting that it can be utilized for vocal-based retrieval applications without prior access to artist metadata, and contributing in paving the way for music understanding applications with respect to specific sources or song-wise elements.

To foster reproducibility of our experimental results, we make our source code as well as pre-trained model weights available at: https://github.com/cgaroufis/CVSM.

II. Related Work

Traditionally, vocal similarity in musical pieces has been modeled via tags denoting the presence or absence of vocals, or inherent attributes of the vocalist [36]. Hence, acoustic features and, more recently, time-frequency representations have been extracted from mixed audio excerpts and fed to shallow machine learning classifiers, or deep neural networks, to directly predict those attributes [37], [38]. Thanks to recent developments in the task of music source separation [28], [32], [39], which have been

[1] We note that since artist labels are utilized to guide the contrastive sampling, the proposed label-informed scheme is not strictly self-supervised.

fueled by the release of diverse datasets [40], [41], [42], isolating the vocals from the mixed audio [43], [44], [45] has emerged as a promising alternative, leading to more robust estimation of vocal tags [43], [45]. However, these methods introduce computational overhead, through the deployment of an auxiliary network for vocal track isolation.

Recent advances in SSL have led to the creation of large-scale music foundation models [34], [46], trained to learn general-purpose representations of audio through large amounts of unlabeled data; contrastive learning, in particular, has been identified as a promising avenue towards this goal. The majority of contrastive approaches follow the SimCLR [17] scheme of projecting batches of paired input views into a shared latent space, which has been shown to outperform other self-supervised approaches [47], [48] in music tagging tasks [49]. COLA [19] has set a simple, yet effective paradigm for transferring SimCLR into the audio domain, sampling data pairs by cropping different excerpts from the same audio sample. As the encoder backbone, [19] adapts EfficientNet-B0 [50], initially developed for image classification, to audio understanding tasks by accepting spectrogram inputs.

The aforementioned data sampling strategy usually forms the basis of more complex augmentation chains. These chains often include additional augmentations such as gain amplification, frequency masking/filtering, reverberation effects, time warping or pitch shifting operations applied either on time-frequency input representations [29], [51], [52], or on the waveform itself [53], [54]. The pair creation process may also be assisted by auxiliary supervisory pseudo-labels; these have included editorial metadata, such as artist or album information [25], [55] and playlist co-occurrence statistics [55], [56]. Numerous of the above methods deviate from [19] in the choice of encoder backbone; while the Efficient-Net encoder is indeed a popular choice [25], [49] experimented with a Res-Net [57] backbone, while [51] opted for a SWin-Transformer [58] and [35] employed a SlowFast NF-Net [59].

Learning the identity of artists, either using complete musical pieces [25], [26] or isolated singing voices [12], [15], has been recently employed as an indirect way to model musical or vocal similarity [13], [14]. Such approaches rely either on supervised learning, where neural networks correlate audio excerpts to artist identity labels [26], or on contrastive learning, building upon the SimCLR framework by pairing elements from the same audio segment [12], [15] or artist [25]. However, the majority of those operate either in the space of singing voices [12], [15], which hinders their ability to model the vocals in the presence of background music, or in complete musical mixtures [25], [26], being thus unable to disentangle attributes pertaining to the singing voice. A potential solution to this could involve the utilization of both complete musical mixtures and isolated sources during the training process [24], [30]. Given that particular attributes of musical pieces, such as the tempo, the primary melody, or the target elicited emotion, are tied to either specific instruments, or the vocals [23], [44], [45], a number of frameworks that associate, through a contrastive process, segments of musical pieces with isolated sources have been developed. For instance, information related to the tempo can be captured through percussive components [60], [61], whereas associating the mixture segments with randomly chosen isolated sources can lead in effective general-purpose representation learning of musical signals [23]. Moreover, encoders pre-trained in musical source association have been employed for evaluating the plausibility of generated instrumental accompaniments in [62].

However, the majority of those methods operate using either raw, or weakly augmented, views of the input audio segments; a less explored family of augmentations, inspired by mixup [63], involves creating additive mixtures of audio signals as pre-training inputs. Its most basic variant, involving generation of synthetic audio excerpts via direct superimposition of distinct audio segments, has been applied in various works [35], [47], [64] as a general-purpose input augmentation. Yet, despite the wide usage, and success, of generating artificial mixtures by superimposing source excerpts of different origins in tasks such as frame-wise pitch estimation [65], [66] and music source separation [28], [32], [67], this avenue has only concurrently been explored in contrastive setups [68], for the task of music sample identification. Moreover, artificial increase of the training dataset by generating synthetic mixtures has been reported to introduce a domain gap during inference [69], [70]; as such, researchers have attempted to bypass it by either performing some small-scale finetuning on datasets from the target domain [71] or introducing an alignment stage, with regard to either tempo, or pitch, before superimposing the various audio segments [11], [72].

III. Proposed Method

In this work, we propose CVSM, an extension of MSCOL [23] for vocal similarity modeling; we present its overview in Fig. 1. Both MSCOL and CVSM were designed for learning representations of musical audio by associating audio excerpts with isolated sources of the audio through a batch-wise contrastive loss objective. However, while MSCOL was developed for general-purpose music representation learning, here we focus solely on the case of vocals. To this end, we extend MSCOL by modifying the contrastive pair generation process as well as the followed training scheme, both by incorporating artificial mixture generation [27] and employing artist-level pair sampling [24], [25], in order to increase its robustness and invariance to non-vocal elements.

A. Contrastive Pair Generation

Label-Informed Sampling: The approach we follow for sampling batches of contrastive pairs is built upon the segment-wise sampling procedure followed by COLA [19]. In more detail, we couple complete musical mixtures with isolated vocal excerpts[2], with the mixture and vocal excerpts originating from the same artist. Employing artist identity labels to guide the pair selection process helps in projecting audio segments with similar vocals close to each other; moreover, in contrast to segment-level sampling, the learned similarity is not tied to song-specific attributes. No augmentations are applied in this stage, so as to capture both timbre-related and pitch-related information about the singing voice [12].

[2]Throughout the paper, we use the terms *excerpts* for network input audio, *segments* for the audio slices used to crop the audio excerpts, and *previews/clips* for complete audio files.

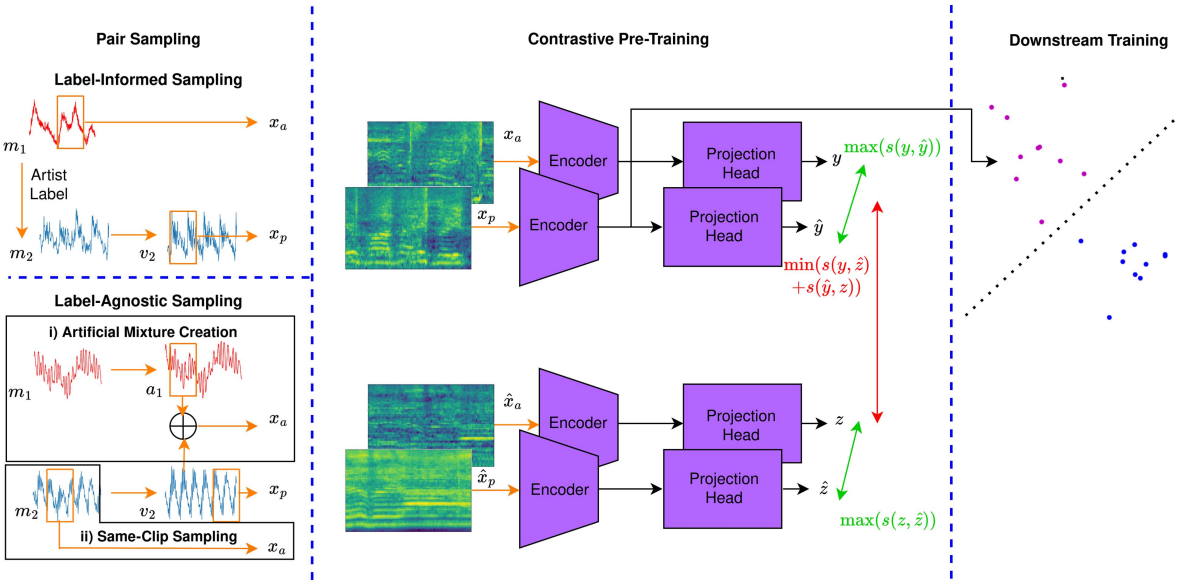

Fig. 1. Overview of our proposed framework for learning audio representations. Contrastive pairs $\{x_a, x_p\}$ are generated either using (top left) label-informed sampling, where pairs of musical mixtures and isolated vocals are sampled from the same artist, or (bottom left) in a label-agnostic manner, by i) creating artificial mixtures by summing the accompaniments and vocals of different song excerpts (artificial mixture creation) or ii) sampling excerpts from the complete song, and coupling them with time-shifted vocal excerpts (same-clip sampling). These contrastive pairs are then used to pre-train an encoder backbone with a contrastive loss objective (right); m_i, v_i and a_i denote complete musical mixtures, isolated vocals and isolated accompaniment, respectively.

Label-Agnostic Sampling: The availability of artist labels guides sampling towards pairs which, while containing vocals from the same artist, do disentangle vocal attributes from non-vocal information, since they may originate from different songs. However, when following the above procedure in a label-agnostic setting, the mixtures and isolated vocals are highly correlated in terms of various non-vocal properties, since they are sampled from the same segment. Thus, in this case the contrastive pairs are created by generating artificial mixtures, consisting of a vocal excerpt superimposed with a randomly sampled instrumental accompaniment, and coupling them with isolated vocals from the same audio segment. Using randomly selected accompaniments to generate artificial mixtures has shown to increase the robustness of networks in frame-wise singing voice understanding tasks [65], [66]. Furthermore, compared to pairing the vocal sample with its original accompaniment [23], this strategy i) not only acts as *data augmentation*, increasing the network's ability to generalize [27], [28], but also ii) helps in forming a latent space *invariant to non-vocal elements* of musical pieces [29]. Similar to above, no further augmentations are applied to either the isolated vocals or the artificial mixtures.

B. Network Backbone and Projection Head

As the encoder, we make use of the EfficientNet [50] family of models. Its combination of small parameter footprint and solid learning capability renders it a suitable choice for contrastive learning pipelines [15], [19], [23], [25], [62], which are dependent on larger batch sizes [17], [19], [53]. Since EfficientNets consist of 2D convolutions, the generated input waveforms have to be transformed into the time-frequency domain; thus, their mel-spectrograms are computed, with 64 bands, a window length of 25 ms and a hop size of 10 ms.

EfficientNet architectures consist of distinct blocks (stages), each of which processes its input through a series of depthwise-separable convolutions [73] incorporating inverted residual connections and downsamples it through a two-dimensional pooling operation. The output tensor of the final convolutional stage is flattened through an average pooling operation to acquire a global feature vector, which can be used in conjunction with a classification head for downstream tasks. We opted for the EfficientNet-B0 encoder, amounting to a total of 5.3 M parameters, including 9 stages and 18 convolutional layers, and leading to an 1280-dimensional embedding.

For the projection head, we follow [19] by applying a linear layer, with a size of 512, on top of the encoder, followed by Layer Normalization [74] and a tanh() activation. We note that this projection head is used only during pre-training, for measuring the similarity between semantically similar embeddings, and discarded during downstream training.

C. Training Scheme and Loss Function

For pre-training, we generate contrastive pairs, following the procedure outlined in Section III-A, and train the network in identifying the vocal excerpt included in each mixture (real or artificial) in the respective batch. For this purpose, first the bilinear similarity, $s(y, \hat{y})$ [19], is computed between all anchor and positive embeddings in each batch. These similarities are first transformed into logits, via a softmax operation, and then used to train the projection head, using the normalized binary cross-entropy loss for all elements in each batch $S = S_a \cup S_p$ (where S_a contains the anchors, S_p the positives, and $\hat{y} \in S_p$):

$$\mathcal{L} = -\sum_{y \in S_a} \log \frac{\exp(s(y, \hat{y}))}{\sum_{z \in S_p} (\exp(s(y, z)))}, \tag{1}$$

where $s(y, \hat{y}) = y^T W \hat{y}$ denotes the aforementioned bilinear similarity between the anchor embeddings y and the positive embeddings $\hat{y}$ computed through a learnable linear layer W. We note that, in contrast to MSCOL [23], since we are mostly interested in identifying timbral differences between vocal excerpts, rather than acquiring information about high-level attributes through the existence or absence of vocals, we do not use the modified cross-entropy loss presented in [23].

Whereas in the *label-informed* case the same training protocol is applied throughout pre-training, we empirically noted that, for the *label-agnostic* pre-training, the learned embeddings did not generalize well in practice, especially for shorter audio clips. We hypothesize that the core reason for this is the domain gap [71] that incurs between the artificial data used for pre-training and the real data used in practice. To alleviate the gap, we experiment with the following strategies:

- *Hybrid Pre-Training:* In this case, we simultaneously expose the network to real and artificial mixtures during pre-training. To this end, real and artificial mixtures for each vocal segment are generated stochastically, at probabilities $1 - p$ and p, so that each batch of contrastive pairs contains both real and artificial musical mixtures.
- *In-Domain Finetuning:* Here, we introduce a second training stage in the self-supervised training procedure, after pre-training with artificial mixtures. During this stage, we do not use artificial mixtures as anchor examples, feeding instead the network solely with pairs of song excerpts and vocals isolated from the same segment.

IV. Experimental Setup

A. Data and Preprocessing

As our primary dataset, we employed Music4All [33], a publicly available dataset consisting of metadata (such as artist names and song titles), lyrics, genre information and user listening statistics for a large-scale music catalog, as well as 30 sec audio previews (clips) of the included songs. In total, the dataset includes 109,269 songs from 16,269 different artists, at diverse sampling rates. In Fig. 2, we depict the exact distribution of the artist identities in the dataset according to the number of audio previews available for each artist (left), as well as the percentage of the available previews corresponding to each category (right). From the left subfigure, we observe that the distribution of artist labels in Music4All is not balanced; only 2.08 % of the artists are represented with more than 50 audio previews each, whereas 82.92 % of the artists have less than 10 previews each. The subfigure on the right further supports this point, since comparable portions of the dataset previews were recorded by artists with less than 10 previews (32,598 total previews, or 29.83%) and artists with more than 50 previews (26,507 previews total, or 24.26%).

Similar to other works utilizing large singing voice datasets for music understanding tasks [12], [23], an open-source framework for music source separation, open-unmix [31], was chosen for extracting the vocal segments and the instrumental accompaniment, facilitating both the contrastive pre-training objective and the creation of the artificial mixtures. All audio clips, prior to pre-training, were downsampled to 16 kHz, for computational efficiency purposes, and split into 5 sec segments. Vocal segments with a mean amplitude lower than 0.01 (amounting to 20.53% of the full dataset) were discarded.

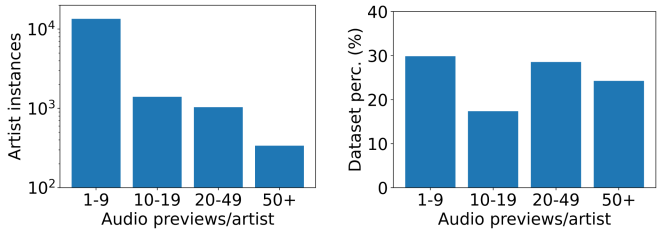

Fig. 2. Dataset statistics for Music4All: the number of artist identities (left) and the percentage of audio previews in the dataset (right), grouped according to the number of audio previews available for each artist.

B. Training and Evaluation Protocols

We pre-train CVSM embeddings with Music4All, following the protocol delineated in Section III. During pre-training, the dataset is split into training, validation and testing subsets, using an 8:1:1 ratio, so that there is no artist leakage between the different subsets. The backbone encoder was pre-trained for a maximum of 8,000 steps (approx. 160 epochs for the subset of Music4All containing vocals), with each step consisting of 64 mini-batches of 128 contrastive pairs. Taking into account data loading considerations, a complete training cycle lasted approximately 5 days in a single Nvidia GeForce 2080. For the in-domain finetuned backbone, following preliminary experiments, we interrupt pre-training with artificial mixtures at 6,000 steps, and only use real mixture-vocal pairs for the final 2,000 steps. We used Adam [75] as the optimizer, with an initial learning rate of 0.001. The pre-training progress was monitored by measuring the pre-text loss in intervals of 10 steps; the learning rate was halved in case the running average of the validation loss did not improve over 1,000 steps.

In order to evaluate the capability of the proposed framework to model vocal similarity, we freeze the model's encoder and benchmark its performance on the following vocal understanding tasks, as per the concurrent literature [12], [15], [16]:

- *Gender Identification:* The goal here is to correctly classify the biological gender (male/female) of the singing artist. For this task, we re-use Music4All, using only artists for whom gender information is retrievable following [76]. For the evaluation, we employ an artist-stratified 10-fold cross-validation procedure, to gauge the timbral content of the embeddings on unseen data. As our metric, we employ the classification accuracy (Acc., %).
- *Artist Identification:* Similar to the previous case, the goal is to correctly classify the artist identity given an embedding vector. We utilize the testing subset of Music4All, using the available previews from $M = 50$ randomly sampled artists from the testing subset (for which no label information has been available during pre-training), and performing 5 repetitions of the experiment with differently sampled artists. During each repetition, an 8:1:1 split of

the respective dataset into training, validation and testing data was applied. In contrast to the case of gender identification, the task target regards discrimination between audio excerpts of artists that have been accessed during downstream training; thus, we opt for a random data split, instead of an artist-stratified one. Since the distribution of artist labels in Music4All is not balanced, we report on both the identification accuracy (Acc., %) as well as the macro-F1 score, which can be calculated as the average of the per-class F1 scores.

- *Artist Similarity:* Here, we directly probe the learned embeddings, computing the similarity between pairs that belong to the same, or different, artists. Similar to the previous case, we utilize the testing subset of Music4All for this task. Following [15], we compute the Equal Error Rate (EER) and the Mean (MNR) and Median (MedNR) Normalized Rank metrics, which are used for retrieval purposes [15], [77], [78] and denote the ability of a system to identify input sample pairs as similar (from the same origin) or dissimilar (from different origins), penalizing low similarities between samples from the same origin. For EER calculation, we use $K = 5000$ sets of similar and dissimilar pairs; for MNR and MedNR, we use a batch size of $N = 50$, and $K = 100$ trials.

We note that the network backbone is always kept frozen, and the performance is measured through training a linear classifier upon the learned embeddings (or, in the case of artist similarity, directly probing the learned latent space *after* the projection head, but *prior* to the bilinear similarity computation). All experiments were repeated for two different configurations; i) for complete musical mixtures (in the presence of instrumental accompaniment), using all 1 sec excerpts with vocals present, as well as ii) on isolated vocals. For the gender and artist identification tasks, network performance is measured over each 30 sec clip, by aggregating the estimates for each excerpt that contains vocals into a single prediction. The downstream linear classifiers are trained over a maximum of 200 epochs, using again Adam [75], a learning rate of 5e-4, and early stopping by monitoring the clip-wise identification accuracy, with a patience of 6 epochs. For the artist similarity task, we use the cosine similarity as the similarity function, measured between pairs of averaged embeddings of all valid (containing vocals) 1 sec excerpts within each clip; the averaged embeddings are L2-normalized prior to the distance calculation. In all cases, paired t-tests were conducted across all experiment repetitions, between CVSM variants and the appropriate (label-informed or label-agnostic) best-performing baseline.

C. Baselines

We compare CVSM to four identity-agnostic baselines, which learn excerpt-level audio representations:

- COLA [19], trained with time-shifted song excerpt pairs.
- MSCOL [23], trained to associate song excerpts with the corresponding vocals in a single-source setup.
- MERT [34], trained via a masked language modeling (MLM) self-supervised setup, with target pseudo-labels provided by an acoustic and a musical teacher model.

TABLE I
OVERVIEW OF THE COMPARED METHODS REGARDING THE INCLUSION OF VOCALS, ARTIST IDENTITY INFORMATION, ARTIFICIAL MIXTURES AND IN-DOMAIN FINETUNING IN THE PRE-TRAINING PIPELINE; ALL METHODS ARE CONTRASTIVELY PRE-TRAINED

Method	Vocals	Artist ID	Artif. Mix	Finetune
COLA [19]	✗	✗	-	-
MSCOL [23]	✓	✗	✗	-
MERT [34]	✗	✗	-	-
MULE [35]	✗	✗	-	-
COLA-ART [25]	✗	✓	-	-
CVSM-A	✓	✗	✓	✗
CVSM-AH	✓	✗	✓	✗
CVSM-AF	✓	✗	✓	✓
CVSM-ART	✓	✓	✗	-

- MULE [35], which is built upon a convolutional backbone and trained with a contrastive objective, involving data-driven augmentations, in an in-house dataset.

As well as the identity-informed baseline:

- COLA-ART, where we follow the procedure used in [25], associating song excerpts with the same artist label.

For COLA and MSCOL, as well as COLA-ART, we trained the baseline encoders using the same pre-training dataset and under the same protocol (backbone architecture and training scheme) as CVSM; for MERT, we make use of the publicly available checkpoint obtained via pre-training in Music4All,[3] and use embeddings obtained via 1 sec audio excerpts to ensure a fair comparison. Finally, for MULE, we utilize the model weights from the official repository[4] for embedding extraction; we note, however, that the comparison is not inherently fair due to MULE being trained out-of-domain.

We experiment with the following CVSM variants:

- CVSM-A, incorporating label-agnostic pre-training with excerpt-level creation of *artificial* mixtures of vocals and instrumental accompaniment, without exposing the model to any real musical mixtures.
- CVSM-AH, where the network is pre-trained with the *hybrid* scheme of viewing both real and artificial musical mixtures during pre-training. After experimentation, the artificial pair creation probability was set to $p = 0.5$.
- CVSM-AF, where we *finetune* CVSM-A in-domain using solely real mixture-vocal pairs.
- CVSM-ART, which has been pre-trained utilizing artist identity information, as delineated in Section III-A.

An overview of various properties of these models is presented in Table I; we note that for $p = 0$ and $p = 1$, CVSM-AH reduces to MSCOL and CVSM-A, respectively.

V. RESULTS AND DISCUSSION

A. Main Results

The results in all three downstream tasks, for all tested configurations, are displayed in Table II – top for mixture input, bottom

[3]https://huggingface.co/m-a-p/MERT-v0-public
[4]https://github.com/PandoraMedia/music-audio-representations/

TABLE II
EXPERIMENTAL RESULTS ON THE TASKS OF GENDER IDENTIFICATION, ARTIST IDENTIFICATION AND ARTIST SIMILARITY ON MUSIC4ALL, (TOP: MUSICAL MIXTURE INPUT, BOTTOM: ISOLATED VOCALS), FOR ALL EXAMINED MODEL CONFIGURATIONS AND BASELINES

Configuration	Gender ID	Artist ID		Artist Sim.		
	Acc. (%) ↑	Acc. (%) ↑	macro-F1 (%) ↑	EER ↓	MNR ↓	MedNR ↓
COLA [19]	81.01 ± 3.68	59.67 ± 5.02	47.02 ± 3.68	28.10 ± 4.46	21.43 ± 3.40	11.82 ± 3.73
MSCOL [23]	85.24 ± 3.80	70.40 ± 3.71	59.32 ± 8.35	25.30 ± 4.71	18.21 ± 3.11	9.08 ± 3.03
MERT [34]	81.61 ± 2.84	65.29 ± 5.65	55.85 ± 7.50	35.34 ± 3.98	27.97 ± 3.44	17.94 ± 4.82
MULE [35]	85.14 ± 4.31	70.33 ± 3.10	61.28 ± 4.88	27.10 ± 4.53	19.16 ± 2.98	10.52 ± 3.32
COLA-ART [25]	**87.15 ± 3.44**	76.58 ± 1.92	69.31 ± 3.43	17.92 ± 4.07	10.33 ± 2.04	4.43 ± 1.34
CVSM-A	**85.65 ± 3.31**	73.26 ± 4.14	59.40 ± 7.05	29.54 ± 4.10	19.89 ± 2.82	8.54 ± 2.89
CVSM-AH	85.40 ± 3.33	**77.66 ± 2.19**	**66.30 ± 4.77***	**22.54 ± 4.07***	**15.23 ± 2.61***	**6.12 ± 2.04***
CVSM-AF	85.48 ± 3.12	72.79 ± 2.60	60.27 ± 3.76	23.42 ± 4.44*	16.19 ± 2.58*	7.27 ± 2.20*
CVSM-ART	87.12 ± 3.41	**78.65 ± 2.24**	**70.00 ± 3.30**	**17.08 ± 3.89***	**9.82 ± 1.90***	**4.03 ± 1.04***

Configuration	Gender ID	Artist ID		Artist Sim.		
	Acc. (%) ↑	Acc. (%) ↑	macro-F1 (%) ↑	EER ↓	MNR ↓	MedNR ↓
COLA [19]	83.89 ± 3.38	54.85 ± 4.66	40.75 ± 4.83	28.50 ± 4.79	21.05 ± 2.93	12.20 ± 3.51
MSCOL [23]	85.27 ± 2.87	70.67 ± 4.55	56.58 ± 4.23	23.58 ± 4.50	16.51 ± 2.62	8.12 ± 2.52
MERT [34]	81.67 ± 1.91	62.88 ± 7.02	47.78 ± 9.50	35.28 ± 3.53	29.06 ± 3.89	18.63 ± 5.15
MULE [35]	85.64 ± 3.63	69.12 ± 4.15	57.15 ± 2.36	29.12 ± 4.01	22.49 ± 3.31	9.43 ± 2.73
COLA-ART [25]	85.05 ± 2.83	71.06 ± 2.38	57.45 ± 3.30	24.74 ± 4.64	17.15 ± 2.73	8.07 ± 2.52
CVSM-A	85.23 ± 3.26	**78.20 ± 3.74***	65.10 ± 4.00*	22.32 ± 4.25*	**14.70 ± 2.37***	**6.22 ± 1.93***
CVSM-AH	**86.06 ± 3.20**	76.33 ± 3.64*	66.09 ± 3.57*	**22.04 ± 3.78***	14.96 ± 2.21*	6.32 ± 2.00*
CVSM-AF	84.91 ± 3.61	77.67 ± 3.30*	**66.67 ± 5.46***	22.34 ± 4.52*	15.06 ± 2.25*	6.54 ± 1.92*
CVSM-ART	**87.46 ± 3.13***	**78.57 ± 3.17***	**68.52 ± 3.92***	**16.84 ± 4.00***	**9.80 ± 1.85***	**3.96 ± 1.11***

for vocal input. In both sub-tables, the first row corresponds to the COLA [19] setup, the second to MSCOL [23], the third to MERT [34], the fourth to MULE [35] and the fifth to the COLA-like artist identity-informed sampling scheme presented in [25]. Lines 6-8 correspond to the label-agnostic CVSM variants, whereas the performance of the label-informed CVSM variant is presented on the last line. The best results across both label-informed and label-agnostic methods are typeset in bold, and statistically significant ($p < 0.05$) improvements are denoted with an asterisk. We observe that in both scenarios, the label-agnostic CVSM variants outperform the COLA [19] and MERT [34] baselines in the majority of downstream tasks (the exception being artist similarity for CVSM-A), while performing comparably to MSCOL [23] and MULE [35] on musical mixtures, and better on isolated vocals. This improvement is statistically significant at the $p < 0.05$ level for all CVSM variants for artist identification and similarity with vocal input, and CVSM-AH and CVSM-AF for artist similarity with mixture input. Similarly, the proposed model pre-trained on the association between mixtures and isolated vocals from the same artist (CVSM-ART) was competitive to the label-informed baseline [25] for the case of musical mixtures (Table II, top) – for the task of artist similarity, the improvement is deemed statistically significant ($p < 0.05$) – while outperforming it for isolated vocals (Table II, bottom) in all downstream tasks at the $p < 0.05$ statistical significance level. We note that, in accordance to the literature [25], [55], availability of artist labels during pre-training leads to improvement in all downstream tasks.

Upon comparing the two sub-tables, we observe that for the classification tasks (gender and artist identification), the best performance reached on musical mixtures (Table II, top) is comparable to that achieved with isolated vocals (Table II, bottom). This indicates that CVSM is suitable for end-to-end application in vocal understanding tasks, without necessitating a vocal isolation pre-processing step. However, this holds less true in the artist similarity task, denoting that vocal isolation is still recommended for successful retrieval applications.

Delving into more detail in the model performance across the evaluated downstream tasks, we first investigate the results for the case of full musical mixtures (Table II, top). We first note that the incorporation of vocal excerpts in label-agnostic training pipelines leads to improved performance in all examined tasks. Indeed, CVSM outperforms both COLA [19] and MERT [34], which have been trained through full mixtures, while the performance yielded by MSCOL [23], which has been trained through mixture-vocal pairs, and MULE [35], relying on large-scale contrastive pre-training and a different network backbone [59], is comparable to some CVSM variants. On the contrary, both label-informed models yield comparable performance in all tasks, suggesting that the variation of contrastive pairs obtained through artist identity sampling is sufficient. Regarding the different tasks, the advantage of the label-informed models (CVSM-ART and COLA-ART) is clearer in gender identification, where all label-agnostic CVSM variants perform to a similar level to MSCOL [23]. However, examination of the results in artist identification and retrieval reveals a more complex picture; while solely incorporating artificial mixtures

of vocals and instrumental accompaniment (CVSM-A) does improve artist identification compared to MSCOL [23], it comes at the cost of a more fragmented latent space, as implied by the higher metrics in artist similarity. Both hybrid real-artificial pre-training (CVSM-AH) and in-domain finetuning (CVSM-AF) help in resolving this issue, leading to better metrics in both identification and similarity than MSCOL. Among these two methods, CVSM-AH appears more effective; in fact, in artist identification, its results approach the ones of the label-informed approaches, despite not accessing identity labels during pre-training. The performance yielded by CVSM-AH in comparison to both MSCOL [23] and CVSM-A implies that in contrast to the majority of augmentations devised in [53], increasing the probability of generating artificial pairs does not guarantee better downstream performance. Finally, the advantage that the label-informed models hold in artist similarity is relatively significant, amounting to a larger than 4% overall decrease in both EER and MNR, and approximately 2% overall decrease in MedNR, compared to their label-agnostic counterparts.

The overall picture changes slightly for isolated vocal input, (see Table II, bottom). Here, i) the availability of vocals during pre-training, as well as ii) the application of appropriate augmentations, emerge as crucial factorsaffect model performance. In more detail, CVSM-ART outperforms COLA-ART [25] in all examined downstream tasks, highlighting the importance of incorporating isolated vocal excerpts into pre-training; this trend is maintained on the label-agnostic models, with CVSM variants and MSCOL [23] scoring higher than COLA [19], MERT [34] and MULE [59]. In addition, in both artist identification and artist similarity tasks, all CVSM variants improve over MSCOL [23], indicating that the proposed scheme of correlating artificial musical mixtures with vocal excerpts succeeds in isolating vocal-related attributes. Contrary to musical mixture inputs, the performance of CVSM-A is relatively similar to that of the two more sophisticated variants, probably due to the exposure of CVSM-A in clean vocals (as opposed to artificial mixtures) during pre-training. Finally, the performance of CVSM-ART in the gender identification (requiring a coarsely structured latent space) and artist similarity tasks suggests that in terms of latent space structure, availability of artist labels plays a crucial role; on the other hand, as we also observed for complete mixtures, the label-agnostic CVSM variants achieve performance close to CVSM-ART in artist identification.

Overall, while CVSM-ART held a clear performance edge over [25] in vocal-oriented downstream tasks, this mostly diminishes for musical mixtures, leading in comparable model performance. Thus, taking into account the computational overhead introduced by acquiring separated vocal tracks, we believe that utilizing the label-informed CVSM pre-training scheme is superior over [25] in cases where either the downstream usage concerns vocal excerpts, or the vocal tracks are independently available, and as such, do not induce additional computational load. On the other hand, there is no such additional computational overhead introduced by CVSM variants compared to MSCOL [23], since instrumental accompaniments are computed by subtracting the (already separated) vocals from the available musical mixtures, rendering their training advantageous for most downstream use cases.

TABLE III
EXPERIMENTAL RESULTS ON THE TASKS OF ARTIST IDENTIFICATION AND ARTIST SIMILARITY ON MUSIC4ALL, (TOP: MUSICAL MIXTURE INPUT, BOTTOM: ISOLATED VOCALS), WHEN USING HT-DEMUCS-OBTAINED VOCAL EXCERPTS FOR PRE-TRAINING

Configuration	Artist ID	Artist Sim.	
	Acc. (%) ↑	EER ↓	MNR ↓
MSCOL [23]	74.27 ± 4.19	23.18 ± 4.11	16.26 ± 2.54
CVSM-AH	77.02 ± 2.12	19.68 ± 4.05	13.61 ± 2.39
CVSM-ART	**78.75 ± 2.51**	**14.24 ± 3.94**	**8.56 ± 1.72**

Configuration	Artist ID	Artist Sim.	
	Acc. (%) ↑	EER ↓	MNR ↓
MSCOL [23]	73.46 ± 3.86	22.18 ± 4.35	15.20 ± 2.58
CVSM-AH	75.73 ± 1.85	19.26 ± 4.20	12.95 ± 2.55
CVSM-ART	**81.67 ± 4.01**	**14.64 ± 3.75**	**8.73 ± 1.97**

B. Additional Experiments

In order to gain further insights into the performance of CVSM, we conducted a series of experiments focusing at various stages of the complete training pipeline by: i) varying parameters of the pre-training scheme, such as the model used to obtain the single-source excerpts, the ratio of real and artificial mixtures during pre-training, and the proportion of artist-labeled excerpts and ii) examining the performance of the pre-trained embeddings under differing fine-tuning setups.

Network Pre-Training: The best-performing label-agnostic CVSM variant, CVSM-AH, relies on obtaining isolated source excerpts through an external source separation model, as well as combining real and artificial musical mixtures during pre-training. Considering the recent advances in the field of music source separation, we also pre-trained vocal similarity models, namely MSCOL [23], CVSM-AH and CVSM-ART, using vocal excerpts obtained from a more recent music source separation framework, HT-Demucs [32], to investigate in which extent they can contribute to improved model performance. We also experimented with the ratio of artificial mixtures in each batch, using the values $p = \{0.25, 0.5, 0.75\}$, since while the achieved performance on vocal excerpts appears to improve for higher values of p, downstream understanding tasks on musical mixtures seem to favor intermediate values of p. Moreover, since artist identity information is not always easily obtainable, we also pre-trained CVSM models with smaller percentages of audio samples with artist identity information (by making an appropriately sized portion of artist identities available), treating the rest of the audio segments as unlabeled.

The results of the various models, pre-trained through vocal excerpts obtained via HT-Demucs [32], in the tasks of artist identification (Acc., %) and artist similarity (EER, MNR) are reported in Table III – top for musical mixtures, bottom for isolated vocals. Our primary observation is that the relative ordering of the compared models is mostly independent of the separation backbone utilized. In more detail, both label-agnostic and label-informed CVSM variants outperform MSCOL [23], with the label-informed variant retaining its performance advantage. The improved separation backbone does, indeed, lead to a boost in downstream performance; we note, however, that it is

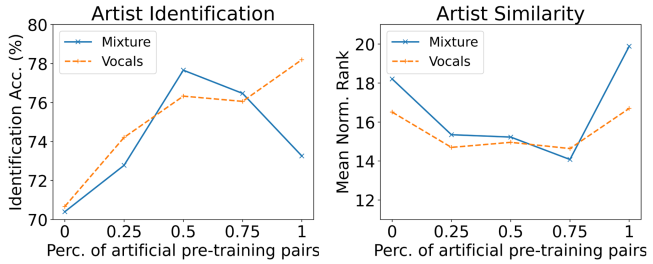

Fig. 3. Performance on artist identification (left) and artist similarity (right), depending on the proportion of artificial mixtures used during pre-training.

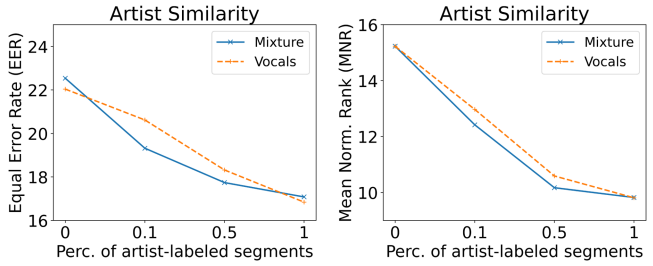

Fig. 4. Performance on the task of artist similarity, in terms of EER (left) and MNR (right), depending on the proportion of excerpts with available artist information used during network pre-training.

more consistent for direct artist similarity modeling compared to artist identification.

In Fig. 3, we display the results in the artist identification (Acc. (%), left) and artist similarity (MNR, right) tasks, for both mixture and vocal input, depending on the proportion of artificial mixtures, p, used for pre-training. We observe that for artist identification, the best performance for mixture input is achieved for equal proportions of real and artificial mixtures, $p = 0.5$, whereas vocal performance improves upon increasing the amount of artificial data, indicating that artificial mixtures successfully disentangle vocal attributes, at the cost of generalizing to real mixtures. On the other hand, for artist similarity, the performance appears less dependent on the exact pre-training protocol, especially for vocal input, where the only underperforming variant corresponds to $p = 0$ (MSCOL [23]).

Finally, regarding the proportion of audio excerpts with known artist identity during pre-training, we present the results for the artist similarity task (since it was the downstream task with the largest differential between label-informed and label-agnostic methods), in terms of EER and MNR, in Fig. 4. The results reveal that there is a gradual improvement for both metrics upon increasing the amount of artist-labeled data, regardless of whether the similarity task is performed on musical mixtures (blue line) or isolated vocals (orange dashed line). Interestingly, even a small proportion (10%) of artist-labeled data contributes to approximately half of the performance differential between the label-agnostic (CVSM-AH) and fully label-informed (CVSM-ART) methods, indicating the suitability of our approach even in label-scarce settings.

Downstream Fine-Tuning: The results presented on Section V-A. were obtained, assuming the availability of song clips of sufficient length (30 sec). However, in practice (i.e. for real-time applications), successfully inferring information from

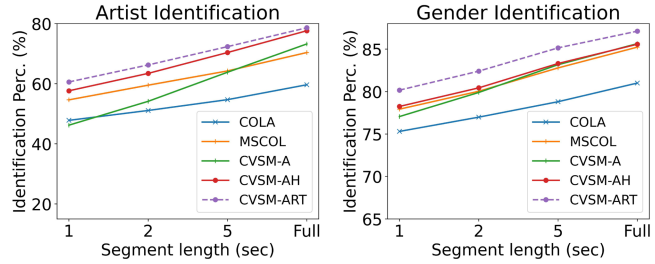

Fig. 5. Performance on the tasks of artist identification (left) and gender identification (right), depending on the input length (in sec).

shorter audio segments is also necessary. To this end, we experimented with varying the *clip length* used for aggregating the per-excerpt predictions into a single one. In addition to the above, we investigated the performance of the embeddings CVSM generates under a *low-resource setup*, as is common practice in the literature [51], [53]. For this purpose, we re-trained linear classifiers on top of the frozen encoders in the task of artist identification, using the same artist splits as in the previous experiments, but with a decreased portion of data available for training and validation. Note that throughout these experiments, we mainly want to compare contrastively pre-trained models that integrate vocal information in their training scheme; as such, we exclude MERT [34], MULE [35] and COLA-ART [25] from those, keeping COLA [19] as a baseline reference. We also omit results for the CVSM-AF variant, since it exhibited a similar trend to CVSM-AH.

The performance of CVSM embeddings of different origins, aggregated through various clip durations is visualized in Fig. 5, for mixture inputs and the tasks of artist (left) and gender (right) identification. In both tasks, aggregation of the per-excerpt scores for longer inputs leads to higher performance in the artist identification task, as it stabilizes network predictions. However, we observe that the performance of all variants does not rise consistently in relation to the excerpt length. In more detail, in the task of artist identification, the embeddings obtained from CVSM-A record lower identification performance than the COLA embeddings obtained without any vocal exposure, when subjected to single excerpts (1 sec. duration). On the other hand, aggregating embeddings from a successively larger amount of excerpts yields significantly better performance than COLA, even improving over MSCOL and approaching CVSM-AH. We hypothesize that this is related to the more informative, yet "noisier" latent space of CVSM-A, since it has not encountered any real mixtures during pre-training. A similar trend can be deduced for the task of gender identification; while the CVSM-A embeddings, obtained prior to in-domain finetuning, record the best scores among label-agnostic models after prediction aggregation, their relative performance to both MSCOL and CVSM-AH drops when reducing the duration of available audio, reaching a negative difference of -1.5% for 1 sec. excerpts. Finally, the performance advantage of the embeddings obtained via artist-informed pre-training, CVSM-ART, is consistent throughout all clip lengths; in fact, its gap to label-agnostic models in artist identification is larger for shorter inputs, reaching approximately 3% for 1 sec excerpts.

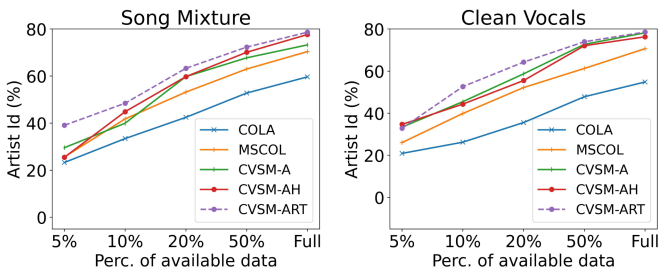

Fig. 6. Frozen embedding performance on the task of artist identification, subject to a reduced data regime, for mixture (left) or vocal (right) input.

For the low-resource experiments, the results are visualized in Fig. 6, for both mixture input (left) or isolated vocals (right). In both cases, we observe that while for an adequate amount of available data the gap between CVSM-ART and the label-agnostic CVSM variants remains close, CVSM-ART performs significantly better under low-resource settings; this result is aligned with the higher effectiveness of CVSM-ART in the artist similarity task. Interestingly, among label-agnostic models, the performance gap between MSCOL and CVSM tends to decrease for smaller data percentages, suggesting that the dilution of the latent space with artificial examples requires a higher amount of labeled examples to unlock its performance advantage. A possible explanation for this could lie in the lack of musicality in the generated examples (since they are formed by random vocal-accompaniment superimposition), which could be potentially resolved through appropriate multi-track sampling for mixture generation [62], [79].

C. Latent Space Visualizations

To visualize the extent CVSM manages to effectively create a latent space with disentangled singer attributes, we plot the T-SNE [80] projections of the clipwise average embeddings of complete musical mixtures (probed from the encoder's output layer), for a subset of the Music4All validation split, conditioned on either the *gender of the singer*, or the *artist identity*. To measure the quality of the generated clusters, we measure the per-cluster average silhouette score [81] as well as the per-cluster average ratio between the mean intra-cluster and the mean inter-cluster distances, averaged across 5 T-SNE runs, for MSCOL, CVSM-A, CVSM-AH and CVSM-ART.

In Fig. 7, we display the results for the case of gender labels, with clips corresponding to male singers displayed in blue dots, while those of female singers colored in orange. We observe that there is high overlap between male and female voices for MSCOL, suggesting that external factors, such as a song's tempo or instrumentation, are influencing the structure of the learned latent space. On the other hand, for CVSM-A and CVSM-AH, the male- and female-voiced clips occupy slightly more segregated areas in the T-SNE plot. This distinction is equally discernible for CVSM-AH (bottom left), indicating that despite exposure to contrastive pairs sampled from the same song, the learned latent space retains vocal-specific properties. However, none of the label-agnostic models create separate clusters between male and female voices, which we attribute to the absence of explicit supervisory signals. This distinction is a bit more visible for the label-informed method, resulting in a slightly more structured latent space. In addition, despite a relatively clear (with minimal

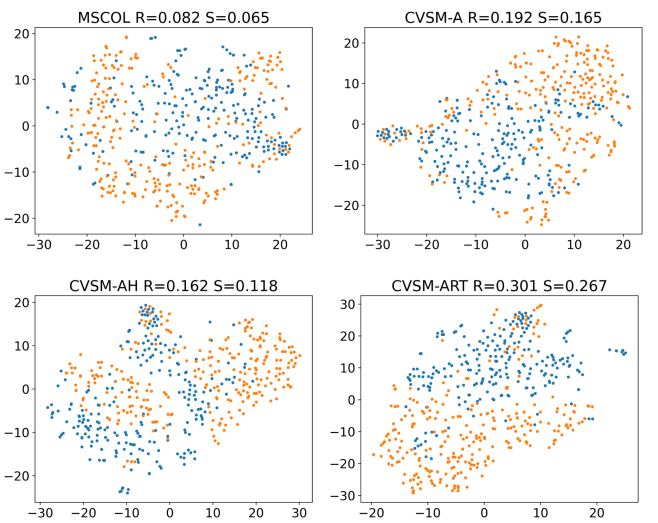

Fig. 7. T-SNE projections of clip-wise average embeddings from various models; blue dots correspond to male singers, orange to female.

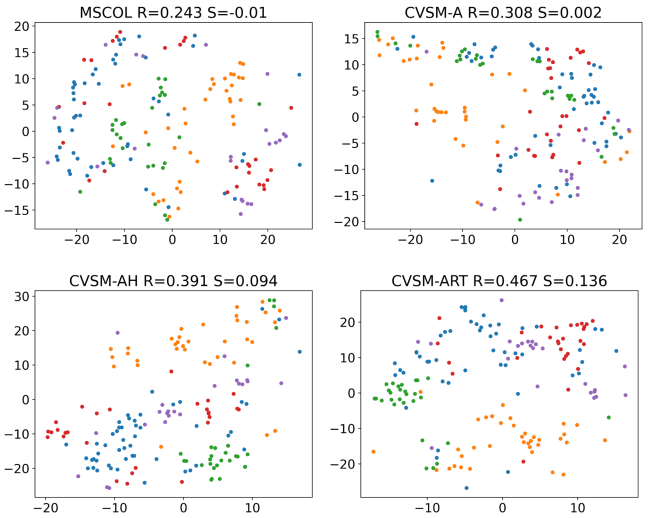

Fig. 8. T-SNE projections of clip-wise average embeddings from various models; dots of the same color correspond to the same artist.

overlap) border between the two clusters, no separate areas are formed in the T-SNE plot. Sub-clusters may also appear, such as the small encircled male-majority area in the bottom left of the CVSM-ART plot (bottom right).

Similarly, in Fig. 8, we present the embedding T-SNE projections with respect to the artist identity of the input audio (differently colored dots correspond to different artists), with the same model correspondence as in Fig. 7. Among the various models, we observe that again, the one pre-trained with the artist-informed protocol achieves the best distinction among artist classes, with some of those occupying distinct subspaces in the T-SNE plot. These results are consistent with those in artist similarity, wherein models utilizing artist identity labels during pre-training yielded much better scores. Among label-agnostic models, the performance achieved is relatively similar, with the largest separation achieved in the case of CVSM-AH; however, the computed metrics are significantly lower than the ones yielded by label-informed models. For the rest of the label-agnostic models, we generally observe higher overlap between different artist classes.

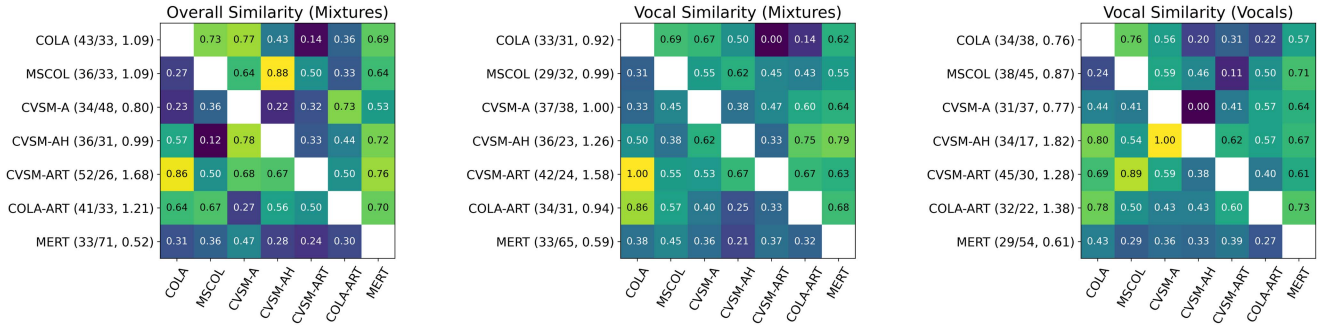

Fig. 9. Pairwise comparison results between the models evaluated on the listening test, regarding overall similarity between musical mixtures (left), vocal similarity between musical mixtures (center) and similarity between isolated vocals (right); the overall scores for each model (in terms of both wins/losses and Bradley-Terry scores) are displayed in parentheses at the left of each row.

VI. Subjective Evaluation

Experimental Protocol: To further validate these results, we also assessed the ability of CVSM to model vocal attributes through a subjective listening test. In more detail, we probed the latent space of each network and retrieved, by means of cosine similarity, the most similar musical piece to a given query. Then, participants were presented with pairs of retrieved musical pieces from different networks, along with the query, and tasked with responding to the following questions:

- *Overall Similarity:* Which of the two retrieved musical pieces is more similar to the initial query in terms of overall musicality (encompassing timbral similarity, rhythmic similarity, and general feeling)?
- *Vocal Similarity:* Which of the two retrieved pieces is more similar to the initial query regarding the vocals?

In total, 37 people took part in the survey, recruited through our social circles, work environments, and community mailing lists, and informed about the survey's purpose and procedure prior to their participation; no data were recorded apart from the anonymized demographics and the questionnaire responses. The participants (23 male, 13 female, 1 other) had an average age of 30.92 years ($\pm$ 6.10 years), and were generally familiar with artificial intelligence and its applications (4.16 $\pm$ 0.94 in a 5-point Likert scale). Despite this familiarity and the positive relationship of the participants with music (81.08% of the participants responded to be listening to music daily), their familiarity with music recommender systems was highly variant (3.30 $\pm$ 1.27 in a 5-point Likert scale).

During the listening test, pairwise comparisons were conducted between two randomly selected models. The model pool consisted of all models presented in. Table I, with the exceptions of CVSM-AF and MULE [35]. Each participant was presented with a total of $N = 20$ different triplets (10 of complete musical mixtures and 10 of isolated vocals), sampled randomly, for each participant, out of an initial pool of $M = 500$ queries, each of a 15 sec duration. Instances where the same song was retrieved by the selected models were mostly excluded,[5] with the exception of a few cases which were left in the study as controls. To assess the performance of all individual models whilst simultaneously avoiding concerns about multiple comparisons, apart from direct model comparison scores, we also report the Bradley-Terry score [82] for each model, as computed from the respective pair-wise winrates after 10 algorithm iterations. Statistical significance compared to random model choice (i.e, with the null hypothesis, H_0, being that the observed Bradley-Terry scores occur from random comparisons) was computed by estimating the Bradley-Terry score probability distribution. To do so, we performed 10,000 inter-model comparison simulations, following the settings of the respective comparison matrix (7 models in total, and equal number of total model comparisons).

Results and Discussion: The complete results of the pairwise comparisons conducted during the listening tests are portrayed in Fig. 9. The two subfigures on the left and the center correspond to the perceived similarity between complete musical mixtures (the left one regarding the overall similarity, and the center one on the vocal similarity), while the right one on the respective results on isolated vocals. Each row and column on the heatmap corresponds to a different model, with the percentage displayed in each cell denoting the winrate of the model in its row against the model in its column; the overall score for each model (both in terms of wins/losses and overall Bradley-Terry score) is displayed in parentheses at the left of each row. From the results, we observe the following:

- In general, models with artist identity information during pre-training appear to outperform their identity-agnostic counterparts in detecting similarity in musical mixtures. This is apparent from the positive overall winrates for both CVSM-ART and COLA-ART models; in the case of CVSM-ART, the achieved scores (52/26 overall similarity and 42/24 vocal similarity, with no pairwise comparisons involving it not favoring it) supersede random chance at the $p < 0.05$ statistical significance level ($p = 0.005$ for musical mixtures and $p = 0.012$ for isolated vocals).
- The label-agnostic models, while mostly lagging behind the label-informed ones, display slight deviations depending on whether the target similarity concerns the overall properties of the musical piece, or just the vocals. In more detail, CVSM-A performs the worse among all contrastively pre-trained models in modeling overall similarity, whereas surprisingly, COLA [19] yields the best results amongst all label-agnostic models. On the other hand, CVSM-AH outperforms all other label-agnostic models on vocal similarity, with the rest of the contrastive models (COLA [19], MSCOL [23], and CVSM-A) yielding

[5] Since the probability of two models retrieving the same recommendation for a given query is not uniform across all models, not all inter-model comparisons were conducted with the same frequency.

effectively similar results. The performance of COLA [19] in overall similarity, compared to models that incorporate vocals into their training pipeline, implies that vocal similarity modeling "sacrifices" perceptually important instrumental information. Conversely, the relative improvement achieved by both CVSM-A and CVSM-AH in vocal similarity modeling, compared to the overall one, suggests that the proposed pre-training scheme with artificial mixtures partially succeeds in creating a latent space with perceptually important vocal information. We note, however, that none of the overall Bradley-Terry scores are significantly different to random chance.

- In the case where the query and the retrieved pieces consist of isolated vocals, the results of the user study largely coincide with those obtained from evaluating vocal similarity in complete musical mixtures. In particular, COLA-ART among label-informed models and CVSM-AH among label-agnostic ones exhibit the best performance – for the case of CVSM-AH, this improvement yields a statistical significance score of $p = 0.003$; with the exception of CVSM-ART, other contrastive models display performance close to random chance.
- The aforementioned results remain mostly robust when controlling on the music listening frequency of the participants. Considering only the subset of users (30) who reported to be listening to music daily, when evaluated on musical mixtures, CVSM-ART achieves the highest overall Bradley-Terry scores (1.38 on overall similarity and 1.49 on vocal similarity), while MSCOL (1.23)/COLA (1.22) and CVSM-AH (1.49) lead among label-agnostic models on overall and vocal similarity respectively. For the case of vocal similarity on isolated vocals, CVSM-AH (2.07) recorded the highest overall Bradley-Terry score.
- Finally, under both examined scenarios, all contrastive models outperform MERT [34], indicating that contrastive pre-training paradigms might be more suitable for similarity modeling and retrieval applications.

As we have mentioned, the subjective test uses primarily sample triplets where the compared models provide different recommendations, mostly excluding the instances where a model pair retrieves the same recommendation. Thus, as an additional similarity measure across the various models, we calculated the percentage of instances among the query set where the same recommendation is obtained for each model pair. These similarities are depicted, in heatmap format, in Fig. 10, where each row and each column corresponds to a different model; the left heatmap presents the similarities for mixture input, the right heatmap on isolated vocals. We note that under both settings, the highest similarity to the CVSM-ART model, which demonstrated consistent performance in both evaluation setups, is achieved by CVSM-AH, followed by MSCOL [23]. This suggests that enriching an unlabeled pre-training dataset with artificial mixtures of vocals and accompaniment may contribute in approximating the latent space of a label-informed model. Interestingly, while CVSM-A shares a relatively small portion of recommendations with the other models for mixture input, presumably due to the lack of exposure in real musical mixtures, its similarity with CVSM-AH in the case of vocal input is particularly high, with the two models retrieving the same recommendation in 64% of the input queries. Finally, we observe that in both cases, MERT [34] yields the lower number of shared recommendations across all models, presumably due to the different training scheme followed in [34], not involving contrastive losses, as well as the larger excerpt length used for pre-training [34].

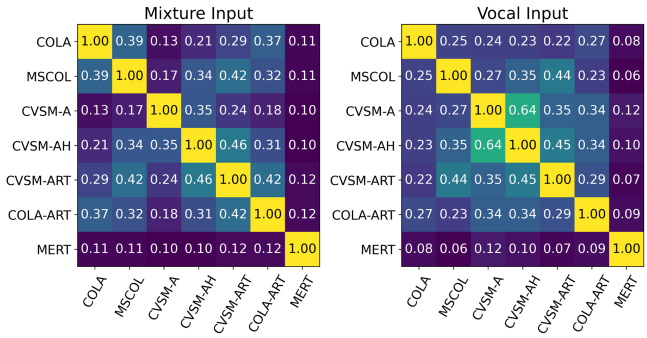

Fig. 10. Pairwise comparison of the recommendations given by the evaluated models, among the queries presented in the subjective test; higher values correspond to higher recommendation similarity.

VII. Conclusion

In this work, we explored the applicability of contrastive learning into learning representations of musical audio with respect to attributes of the singing voice, and presented CVSM, a framework that learns such representations by maximizing the similarity between musical mixtures and vocal excerpts. We devised both a *label-informed* pre-training scheme, which leverages artist labels during contrastive pair sampling, and a *label-agnostic* protocol based on generating artificial mixtures through superimposing isolated vocals and randomized instrumental accompaniment. Our results, validated both through linear probing on downstream tasks encapsulating aspects of vocal similarity and a user study, suggest that the proposed method is effective in conveying vocal similarity, performing at least comparably to the respective (pending on artist label availability during pre-training) baselines. We also note that while the label-informed scheme exhibited more consistent performance across both downstream testing and the user study, a hybrid, label-agnostic pre-training scheme with a combination of real and artificial musical mixtures performed competitively to it both in the artist identification task and in perceived vocal similarity. This work contributes in paving the way towards modeling musical similarity with respect to particular sources, or other fine-grained musical attributes.

Since the obtained results indicate that identity-informed pre-training leads to more consistent downstream performance and similarity modeling, additional work should be carried out towards bridging the gap between label-informed and label-agnostic protocols, by incorporating an identity estimation step during contrastive sampling. Moreover, artificial mixture integration into label-informed pre-training should be further investigated, since our preliminary experiments (not reported here) yielded a performance drop in downstream testing. Finally, it would be interesting to adapt the proposed strategy in other music information retrieval tasks involving frame-wise information

from vocals, such as automatic lyrics transcription [10] and vocal fundamental frequency estimation [83].

Acknowledgment

The authors would like to thank Panagiotis P. Filntisis for his very useful feedback on the content of the paper, as well as the anonymous reviewers for their constructive feedback, significantly improving the manuscript. The publication of the article in OA mode was financially supported by HEAL-Link.

This project is funded by the European Union under Horizon Europe (grant No. 101136568 - HERON).

References

[1] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, pp. 303–319, Apr. 2011.

[2] Y. Deldjoo, M. Schedl, and P. Knees, "Content-driven music recommendation: Evolution, state of the art, and challenges," *Comput. Sci. Rev.*, vol. 51, 2024, Art. no. 100618.

[3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, 2003, Art. no. 770.

[4] J. Sundberg, "The acoustics of the singing voice," *Sci. Amer.*, vol. 236, no. 3, pp. 82–91, 1977.

[5] N. Obin and A. Roebel, "Similarity search of acted voices for automatic voice casting," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1642–1651, Sep. 2016.

[6] T. K. Perrachione, K. T. Furbeck, and E. J. Thurston, "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," *J. Acoustical Soc. Amer.*, vol. 146, no. 5, pp. 3384–3399, 2019.

[7] Y. Lee, P. Keating, and J. Kreiman, "Acoustic voice variation within and between speakers," *J. Acoustical Soc. Amer.*, vol. 146, no. 3, pp. 1568–1579, 2019.

[8] J. Kreiman, Y. Lee, M. Garellek, R. Samlan, and B. R. Gerratt, "Validating a psychoacoustic model of voice quality," *J. Acoustical Soc. Amer.*, vol. 149, no. 1, pp. 457–465, 2021.

[9] S. Liu, M. Babel, and J. Zhu, "A comparison of voice similarity through acoustics, human perception and deep neural network (DNN) speaker verification systems," in *Proc. Interspeech*, Kos, Greece, 2024, pp. 3674–3678.

[10] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, 2010, Art. no. 546047.

[11] C. Zhang et al., "PDAugment: Data augmentation by pitch and duration adjustments for automatic lyrics transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 454–461.

[12] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1614–1623, 2022.

[13] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 638–648, Mar. 2010.

[14] E. J. Humphrey et al., "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 82–94, Jan. 2019.

[15] B. Torres, S. Lattner, and G. Richard, "Singer identity representation learning using self-supervised techniques," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Milan, Italy, 2023, pp. 448–456.

[16] Y. Yamamoto, "Toward leveraging pre-trained self-supervised frontends for automatic singing voice understanding tasks: Three case studies," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Taipei, Taiwan, 2023, pp. 1745–1752.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[18] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, Copenhagen, Denmark, 2015, pp. 84–92.

[19] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3875–3879.

[20] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Proc. Mach. Learn. Health*, 2020, pp. 238–253.

[21] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[22] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5.

[23] C. Garoufis, A. Zlatintsi, and P. Maragos, "Multi-source contrastive learning from musical audio," in *Proc. SMC*, Stockholm, Sweden, 2023, pp. 162–169.

[24] D. Desblancs, G. Meseguer-Brocal, R. Hennequin, and M. Moussallam, "From real to cloned singer identification," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, San Francisco, CA, USA, 2024, pp. 327–334.

[25] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from discogs," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 825–33.

[26] J. Park, J. Lee, J. Park, J.-W. Ha, and J. Nam, "Representation learning of music using artist labels," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Paris, France, 2018, pp. 717–724.

[27] S. Uhlich et al., "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 261–265.

[28] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep Res-U-Net for music source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 342–349.

[29] M. C. McCallum, M. E. Davies, F. Henkel, J. Kim, and S. E. Sandberg, "On the effect of data-augmentation on local embedding properties in the contrastive learning of music audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Seoul, South Korea, 2024, pp. 671–675.

[30] K. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Delft, The Netherlands, 2019, pp. 295–302.

[31] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, 2019, Art. no. 1667.

[32] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5.

[33] I. A. P. Santana et al., "Music4all: A new music database and its applications," in *Proc. Int. Conf. Syst., Signals Image Process.*, Niteroi, Brazil, 2020, pp. 399–404.

[34] Y. Li et al., "MERT: Acoustic music understanding model with large-scale self-supervised training," in *Proc. Int. Conf. Learn. Representations*, Wien, Austria, 2024, pp. 1–13.

[35] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. F. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 256–263.

[36] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam, "Semantic tagging of singing voices in popular music recordings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1656–1668, 2020.

[37] G. Marques, F. Gouyon, T. Langlois, and M. A. Domingues, "Three current issues in music autotagging," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Miami, FL, USA, 2011, pp. 795–800.

[38] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, New York, NY, USA, 2016, pp. 805–811.

[39] K. Schulze-Forster, G. Richard, L. Kelley, C. S. Doire, and R. Badeau, "Unsupervised music source separation using differentiable parametric source models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1276–1289, 2023.

[40] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," doi: 10.5281/zenodo.3338373.

[41] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[42] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "MoisesDB: A dataset for source separation beyond 4-Stems," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Milan, Italy, 2023, pp. 619–626.
[43] B. Sharma, R. K. Das, and H. Li, "On the importance of audio-source separation for singer identification in polyphonic music," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 2020–2024.
[44] J. De Berardinis, A. Cangelosi, and E. Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 310–317.
[45] C. Garoufis, A. Zlatintsi, and P. Maragos, "Pre-training music classification models via music source separation," in *Proc. Eur. Signal Process. Conf.*, Lyon, France, 2024, pp. 411–415.
[46] M. Won, Y.-N. Hung, and D. Le, "A foundation model for music informatics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Seoul, South Korea, 2024, pp. 1226–1230.
[47] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Self-supervised learning for general-purpose audio representation," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.
[48] J. Anton, H. Coppock, P. Shukla, and B. W. Schuller, "Audio Barlow twins: Self-supervised audio representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5.
[49] G. Meseguer-Brocal, D. Desblancs, and R. Hennequin, "An experimental comparison of multi-view self-supervised methods for music tagging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Seoul, South Korea, 2024, pp. 1141–1145.
[50] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 6105–6114.
[51] H. Zhao, C. Zhang, B. Zhu, Z. Ma, and K. Zhang, "S3T: Self-supervised pre-training with swin transformer for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, Singapore, 2022, pp. 606–610.
[52] J. Choi, S. Jang, H. Cho, and S. Chung, "Towards proper contrastive self-supervised learning strategies for music audio representation," in *Proc. IEEE Int. Conf. Multimedia Expo*, Taipei, Taiwan, 2022, pp. 1–6.
[53] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 673–681.
[54] M. Vásquez and J. Burgoyne, "Tailed U-Net: Multi-scale music representation learning," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 67–75.
[55] G. Meehan and J. Pauwels, "Evaluating contrastive methodologies for music representation learning using playlist data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Hyderabad, India, 2025, pp. 1–5.
[56] P. Alonso-Jiménez et al., "Pre-training strategies using contrastive learning and playlist information for music classification and similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
[58] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.
[59] L. Wang et al., "Towards learning universal audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, 2022, pp. 4593–4597.
[60] M. Heydari and Z. Duan, "Singing beat tracking with self-supervised front-end and linear transformers," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 617–624.
[61] D. Desblancs, V. Lostanlen, and R. Hennequin, "Zero-note samba: Self-supervised beat tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2922–2934, 2023.
[62] R. Ciranni et al., "COCOLA: Coherence-oriented contrastive learning of musical audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Hyderabad, India, 2025, pp. 1–5.
[63] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018, pp. 1–13.
[64] S. Srivastava et al., "Conformer-based self-supervised learning for non-speech audio tasks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, 2022, pp. 8862–8866.
[65] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1118–1128, 2020.
[66] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "PESTO: Pitch estimation with self-supervised transposition-equivariant objective," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Milan, Italy, 2023, pp. 535–544.
[67] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 816–820.
[68] H. Cheston, J. Van Balen, and S. Durand, "Automatic identification of samples in hip-hop music via multi-loss training and an artificial dataset," 2025, *arXiv:2502.06364*.
[69] J. Tremblay et al., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, 2018, pp. 969–977.
[70] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vancouver, BC, Canada, 2017, pp. 23–30.
[71] I. Simon, J. Gardner, C. Hawthorne, E. Manilow, and J. H. Engel, "Scaling polyphonic transcription with mixtures of monophonic transcriptions," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 44–51.
[72] A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Augmentation methods on monophonic audio for instrument classification in polyphonic music," in *Proc. Eur. Signal Process. Conf.*, 2021, pp. 156–160.
[73] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1251–1258.
[74] J. L. Ba, J. R. Kiros, and G. Hinton, "Layer normalization," in *Proc. Adv. Neural Inf. Proc. Syst. Deep Learn. Symp.*, Barcelona, Spain, 2016, pp. 1–12.
[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–11.
[76] A. Ferraro, X. Serra, and C. Bauer, "Break the loop: Gender imbalance in music recommenders," in *Proc. Conf. Hum. Inf. Interaction Retrieval*, Canberra, ACT, Australia, 2021, pp. 249–254.
[77] S. Lattner, "SampleMatch: Drum sample retrieval by musical context," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 781–788.
[78] S. Yang et al., "SUPERB: Speech processing universal PERformance benchmark," in *Proc. Interspeech*, Brno, Czechia, 2021, pp. 1194–1198.
[79] A. Riou, S. Lattner, G. Hadjeres, M. Anslow, and G. Peeters, "Stem-JEPA: A joint-embedding predictive architecture for musical stem compatibility estimation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, San Francisco, CA, USA, 2024, pp. 625–633.
[80] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008, pp. 2579–2605.
[81] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[82] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
[83] H. Cuesta, B. McFee, and E. Gómez, "Multiple F0 estimation in vocal ensembles using convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 302–309.