Contrastive and Adversarial Disentanglement for Speaker Representations in Brazilian Portuguese

Ariadne Nascimento Matos
ICMC – USP
ariadnenmtos@usp.br

Arnaldo Candido Junior
IBILCE – UNESP
arnaldo.candido@unesp.br

Moacir Antonelli Ponti
ICMC – USP
ponti@usp.br

Abstract

In this work, we study disentanglement between speaker and environment by combining an adversarial framework with contrastive learning objectives. We investigate supervised contrastive learning (SupCon), which exploits environment labels to structure the environment subspace, and self-supervised SimCLR, which learns invariance from augmented views. Experiments on a controlled synthetic dataset (ST1) and a more realistic corpus (CML-TTS) show that SupCon yields the most discriminative and stable speaker embeddings on ST1, achieving the best verification performance (EER=4.70%, MinDCF=0.24). Overall, our findings emphasize (i) the importance of synthetic benchmarks for diagnosing disentanglement under controlled factor variation and (ii) the effectiveness of combining contrastive and adversarial objectives to encourage speaker representations that are both discriminative and less sensitive to environmental factors.

1 Introduction

Speech signals convey the linguistic structure of the utterance, information about the speaker, recording characteristics, and background content or noise. These elements are entangled in the acoustic signal in complex ways, as they occur in almost all natural speech (Hardcastle et al., 2012). In tasks such as Automatic Speech Recognition (ASR), systems must handle not only linguistic variations, such as prosody and pronunciation, but also maintain performance across diverse acoustic conditions. A key challenge lies in overcoming such extrinsic variability factors (Wang and Gales, 2012). A similar phenomenon is observed in the field of speaker verification, where environmental conditions and noise pose significant challenges to the precision of automatic speaker recognition systems, due to the degradation of the speech signal.

For example, additive noise can mask essential acoustic cues, especially in a crowded room. This masking is harmful to low-energy phonemes like fricatives (“s”,“f”), which encode speaker-specific information but suffer from a reduced effective signal-to-noise ratio (SNR) in noisy environments (Monir et al., 2024). Reverberation similarly degrades the signal by causing sound reflections from walls and surfaces, producing echoes that temporally smear the speech signal. As a result, critical spectral features and formant structures, which are characteristics of an individual vocal tract, become blurred (Al-Karawi, 2015).

Disentanglement techniques aim at separating (disentangling) the underlying factors of variation in a speech signal. Since the success of a machine learning algorithm depends on how the data are represented Bengio et al. (2013), in an ideal scenario, a robust speaker embedding should contain only the essential characteristics of the speaker identity while being invariant to other factors, such as background noise. Thus, disentanglement helps the model learn separate and semantically meaningful representations, aligned with latent factors (Wang et al., 2024). The goal is not to remove information unrelated to the speaker features but to explicitly separate it from the speaker identity within the latent embedding space. This approach offers advantages because it allows the model to learn speaker representations that are inherently invariant to the acoustic environment.

In speaker verification, environment-agnostic embeddings can mitigate environmental noise and thereby enhance system reliability (Lian et al., 2022). Furthermore, these disentangled representations serve as pretrained features for downstream applications, such as voice conversion, preventing transfer of environmental characteristics, leading to more natural voice transformations.

1.1 Goals

This study aims to apply speaker-environment disentanglement to Brazilian Portuguese speech using

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 591–600

a novel synthetic dataset (ST1) and the CML-TTS corpus (Oliveira et al., 2023). Our central goal is to assess whether contrastive objectives, particularly SupCon and SimCLR with adversarial training can provide inductive (contrastive) bias for disentanglement representation. The synthetic dataset is important to allow for controlled experimentation (Casanova et al., 2023). For thorough evaluation, we employ both quantitative metrics from speaker verification tasks and qualitative visualization using t-SNE to assess the degree of disentanglement between speaker and environment representations.

We investigate two training scenarios (i) a baseline without environment dissociation, training with a standard Triplet Loss using only speaker identities; and (ii) an environment dissociation setting that leverages three contrastive objectives: Triplet Loss, Supervised Contrastive Learning (SupCon) (Khosla et al., 2020), and SimCLR (Chen et al., 2020) to analyze their impact on disentanglement performance.

2 Related Work and Contributions

Speaker Representation. In order to enable discriminative speaker embeddings, while disentangling speaker identity from environment factors, Kang et al. (2020) combined ResNet backbones with domain adversarial training and multi-task learning. Later, both ECAPA-TDNN and ResNet were used in this same context (Liu et al., 2023; Nam et al., 2024), with Xing et al. (2024) adapting a different ResNet architecture for the same objective. Along the same lines, Menon et al. (2025) implemented disentanglement learning strategies using ResNet variants alongside ECAPA and ReDimNet (Reshape Dimensions Network) (Yakovlev et al., 2024) as backbone extractors of speaker representations. Concurrently, self-supervised learning approaches have emerged using models like WavLM (Chen et al., 2022) adapted for disentanglement purposes, with Ruggiero et al. (2025) proposing ETA-WavLM to decompose speech representations into speaker-dependent and speaker-independent components.

Contrastive Learning for Speech. Recent advances in disentangled representation learning for speaker recognition often leverage contrastive learning. One line of work combines Contrastive Predictive Coding (CPC) (van den Oord et al., 2018), a self-supervised method that enforces temporal consistency by treating frames within an utterance as positive pairs, with Factorized Variational Autoencoders (FVAE) (Deng et al., 2017). For instance, Xie et al. (2024) integrated CPC with FVAE and adversarial learning for to explicitly disentangle speaker identity from environment components as well as other linguistic factors.

Another approach employs a Disentangled Sequential Variational Autoencoder (DSVAE) (Tu et al., 2024) augmented with a SimCLR-like contrastive loss. This framework separates speaker identity information from linguistic content, where the contrastive objective encourages invariance to environmental variations, thereby improving verification robustness in noisy conditions.

Contrastive learning has also been integrated with WavLM pre-trained models, in which the contrastive loss guides the model to produce embeddings that emphasize speaker identity while suppress environmental effects (Xue et al., 2024).

2.1 Contributions

Our main contributions are two-fold:
(1) Extending the framework proposed by Nam et al. (2024) to investigate two contrastive learning (SupCon and SimCLR) and analyze how they affect the separation of speaker identity and environmental background factors.
(2) Explicitly compare disentanglement capabilities (via visualization), not only speaker verification metrics.

To our knowledge, this is the first study on the disentanglement of speaker identity and environment background in Portuguese.

3 Methodology

3.1 Datasets

We conducted our experiments on prepared speech datasets that provide fine-grained control over speaker and environment factors. Many prior studies also used controlled or synthetic speech because, with spontaneous speech, evaluation depends on annotated labels rather than well-defined factors of variation. Because those labels are often unbalanced across speakers or contexts, the resulting metrics can become unreliable (Zhang et al., 2023).

Synthetic dataset: We synthesized speech using XTTS v2 (Casanova et al., 2024), a multilingual text-to-speech model that supports zero-shot voice cloning. Our speaker set was derived from Brac-

cent (Batista, 2019), comprising 69 speakers for training and 28 for testing.

For each Braccent speaker, we generated speech for 50 sentences drawn from classic Brazilian literature. Table 1 reports the total duration of the synthetic dataset (ST1). For each audio file, we applied the following acoustic conditions from MUSAN (Snyder et al., 2015): point-source noise, hd-classical, sound-bible and fma-western-art.

Table 1: Total of hours for synthetic datasets generated with the XTTS v2 model.

Split	Hour	speakers	utterances
Train	~ 34	69	20,272
Test	~ 17	28	10,220

CML-TTS: CML-TTS (Oliveira et al., 2023) is a public, multilingual speech corpus derived from Multilingual LibriSpeech (MLS) and adapted for training TTS (text-to-speech) models. The corpus consists of seven languages (Dutch, French, German, Italian, Portuguese, Polish, and Spanish) with 3, 176 hours of audiobooks from Project Gutenberg text and recorded by LibriVox volunteers. We used the Portuguese subset containing 69 hours from 48 speakers (31 male, 17 female).

3.2 Disentanglement Framework

Our model architecture follows Nam et al. (2024) (depicted in Figure 1) but replaces Voxceleb2 and video-session metadata with our synthetic data and CML-TTS, with background environments. The model comprises four modules: batch construction with data augmentation, the ECAPA-TDNN speaker verification backbone, a disentangling autoencoder, and a discriminator for adversarial learning. Also, unlike the previous work, which optimized only a Triplet Loss objective, we expand the contrastive-learning setup by adding Supervised Contrastive (SupCon) and SimCLR losses, to investigate whether novel contrastive objectives improve speaker-identity preservation while suppressing environmental factors.

While the ECAPA-TDNN backbone acts as a feature extractor from the input, the autoencoder focuses on disentangling the resulting speaker embeddings with environment. To achieve this, we use three discriminators: a speaker discriminator, responsible for capturing speaker information, an environment discriminator trained to capture environmental information, and a third discriminator, which encourages the model to remove environment from speaker representation. In addition, we add contrastive-learning losses for the environment discriminator, which allows to backpropagate the desired latent space properties.

Batch construction: We follow the original batch formulation: each mini-batch contains three utterances (x_{i1}, x_{i2}, x_{i3}) from the same speaker, where i indexes the speaker and j the environment. The anchor-positive pair consists of segments from the same speaker and the same environment/scene (as presented on Figure 1 by $x_{i,2}$ and $x_{i,1}$), and the negative pair consists of segments from the same speaker but a different environment (presented on Figure 1 by $x_{i,3}$). Still, we tailored the sampling for our study: the anchor and positive come from the same environment, while the negative comes from the same speaker in a different environment. For augmentation, we applied music, noise, and reverberation from MUSAN. The anchor and positive received identical augmentations, whereas the negative received a different augmentation. We believe this setup helps the environment contrast while controlling for speaker identity.

Speaker-verification backbone: We extracted embeddings with an ECAPA-TDNN speaker verification model. The architecture uses SE-Res2 (1D Res2) blocks with squeeze-and-excitation. It splits channels into multiple paths, processes each path through residual connections, and applies channel-wise attention. For preprocessing, we applied pre-emphasis (coefficient 0.97) to each waveform, computed STFT spectrograms with 25 ms windows, 10 ms hop, and a Hamming window, and converted them to 80-dimensional log-mel spectrograms. These log-mel features serve as the model input. The ECAPA-TDNN backbone outputs a feature vector of size 3072. We kept all hyperparameter as found by Nam et al. (2024).

Autoencoder: The Autoencoder takes ECAPA-TDNN embeddings as input and, in the encoder, splits the latent vector $e_{i,j}$ into two lower-dimensional embeddings: a speaker representation e^{spk} and an environment representation e^{env}. The disentanglement occurs with complementary losses: a speaker classification loss $\mathcal{L}_{spk}$, adversarial terms that discourage speaker cues in e^{env} and environment cues in e^{spk} and a correlation penalty $\mathcal{L}_{corr}$ which reduces statistical dependence between the two embeddings. The decoder reconstructs an approximation of the original embedding $e_{i,j}$ from e^{spk} and e^{env}, optimized with $\mathcal{L}_1$ recon-

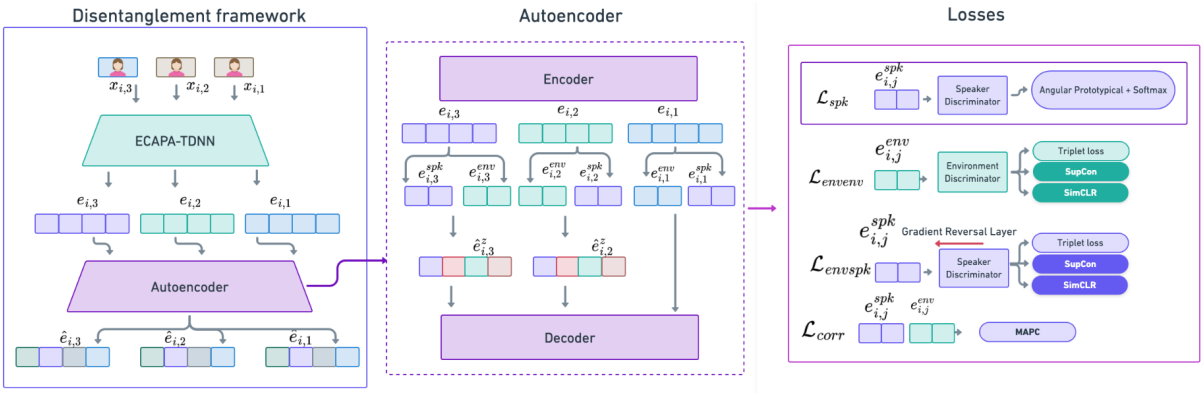

Figure 1: Overview of the disentanglement framework, consisting of a backbone (ECAPA-TDNN) responsible for extracting embeddings and an autoencoder where speaker and environment representations are disentangled. After passing through the encoder, the embeddings are split into two halves ($e_{i,j}^{env}$ and $e_{i,j}^{spk}$), and different loss functions ($\mathcal{L}_{spk}, \mathcal{L}_{envenv}, \mathcal{L}_{envspk}, \mathcal{L}_{corr}$) are applied to disentanglement.

struction loss (Equation 1).

$$\mathcal{L}_{\text{recons}} = \sum_{j=1}^{3} \left(|e_{i,j} - \hat{e}_{i,j}| \right) \tag{1}$$

We applied three discriminators: a speaker discriminator (S) and two environment ones (E^E and E^S).

Speaker Discriminator (S): For the speaker discriminator, we optimize a joint $\mathcal{L}_{spk}$ classification loss that combines Softmax and Angular Prototypical objectives. During end-to-end training with the ECAPA-TDNN backbone, both terms operate on the speaker embedding $e_{i,j}^{spk}$. The Softmax term drives class separation by mapping $e_{i,j}^{spk}$ to the N speaker classes. The Angular Prototypical term shapes angular geometry by maximizing cosine similarity within class and minimizing it across classes. We set $M = 3$, i.e for each speaker class, we sample one query embedding $e_{i,j}^{spk}$ and two support embeddings $e_{i,1}^{spk}$, $e_{i,2}^{spk}$.

Environment Discriminator (E^E):

The environment discriminator is optimized with an environment loss $\mathcal{L}_{env}$, instantiated by a single choice of loss among three methods: Triplet Loss, Supervised Contrastive Learning (SupCon), or SimCLR (see Section 4.1). In our experiments, we select one of these formulations per configuration: Triplet Loss is used when we explicitly construct anchor, positive and negative, SupCon when environment labels are available to define positives and negatives within a batch and SimCLR with a label-free baseline based on augmented views of the same sample. Regardless of the chosen formulation, $\mathcal{L}_{env}$ encourages a representation space where embeddings from the same environment are close and embeddings from different environments are apart. This structure supports the adversarial component that removes environmental cues from the final speaker embedding.

Environment Discriminator with Adversarial Training (E^S): While E^E denotes the environment discriminator, E^S extracts residual environment information from the speaker representation $e_{i,j}^{\text{spk}}$ using a Gradient Reversal Layer (GRL). During backpropagation, the GRL multiplies the gradient by $-\lambda$, making the environment discriminator task more difficult. Adversarial training drives the speaker representation to discard environment cues. We denote the GRL-mediated environment loss by $\mathcal{L}_{\text{env spk}}$.

We added Mean Absolute Pearson Correlation (MAPC) regularization with loss $\mathcal{L}_{corr}$ (see Eq.2). We computed Pearson correlations between each speaker dimension and the environment embedding, took absolute values, and averaged them. We minimize $\mathcal{L}_{corr}$ to reduce dependence between the two representations.

$$L_{corr} = \frac{|Cov(e_{i,j}^{spk}, e_{i,j}^{env})|}{\sigma(e_{i,j}^{spk} \cdot \sigma(e_{i,j}^{env}))} \tag{2}$$

4 Training Strategies and Objectives

A multi-objective optimization strategy trains the disentanglement framework. The overall learning objective combines three key losses: a speaker discrimination loss ($\mathcal{L}_{spk}$) to preserve speaker identity, an environment discrimination loss ($\mathcal{L}_{env}$) to capture acoustic conditions, and an adversarial

loss ($\mathcal{L}_{\text{env spk}}$) to remove environmental information from the final speaker embedding.

4.1 Contrastive Objectives for Environment Disentanglement

As mentioned before, we extended prior work by implementing $\mathcal{L}_{env}$ through three distinct contrastive learning: Triplet Loss, Supervised Contrastive Learning (SupCon), and SimCLR with the goal to analysis of how different environment representation learning strategies affect final speaker embedding quality and disentanglement.

Environment Triplet Loss (Scenario 1): We first applied a Triplet Loss on ST1 dataset without environmental dissociation, where positive and negative samples shared similar environments. This setup enabled assessment of spontaneous speaker-environment entanglement when environmental cues remain available as discriminative features.

Environment Triplet Loss (Scenario 2): In scenario 2, we applied the Triplet Loss by sampling the anchor and positive come from the same environment and the negative from a different environment. We forme all triplets following this rule. The margin is set to m=0.3 and the scale factor to 0.30. The margin enforces a minimum gap between anchor-positive and anchor-negative distances, while the scale factor as a temperature, higher values increase contrast between samples and lower values smooth the loss.

Environment with SupCon and SimCLR: within our adversarial framework, we applied the environment factor using the contrastive objective (Equation 3). For each utterance we generated two randomly augmented views and encode them into embeddings, yielding N embeddings in a mini-batch. For a given anchor embedding $\mathbf{e}_{i,j}$, we computed cosine similarities to other embeddings in the batch. In Equation 3, s_{ip} denotes the similarity between the anchor i and a *positive* example $p \in P(i)$, whereas s_{ia} denotes the similarity between the anchor i and any other candidate $a \neq i$ in the batch (including both positives and negatives), which appears in the denominator. The index a therefore ranges over all embeddings except the anchor itself. The key difference between SupCon and SimCLR is the definition of the positive set $P(i)$. In SupCon, we use environment labels so that $P(i)$ contains all indices whose samples share the same environment labels as the anchor (excluding i), pulling together embeddings from the same acoustic environment.

In SimCLR, the learning is self-supervised and $P(i)$ contains only the index of the other augmented view of the same utterance. All remaining samples in the batch act as negatives, so the model learns environment-related invariance induced by the augmentation pipeline without requiring environment labels.

$$\mathcal{L} = -\sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(s_{ip}/T\right)}{\sum\limits_{\substack{a=1 \\ a \neq i}}^{N} \exp\left(s_{ia}/T\right)} \tag{3}$$

4.2 Overall Training Objective

We define the total loss ($\mathcal{L}_{total}$ (Eq. 4)) as the sum of the reconstruction loss and all auxiliary terms. The loss trains the encoder to learn environment-invariant representations via the GRL-based adversarial objective and the MAPC correlation penalty, whereas $\mathcal{L}_{\text{env spk}}$ trains only the E^S discriminator to predict the environment. Each loss term is scaled by a weighting coefficient (λ), specifically, we use (λ_E) for the environment discriminator loss, (λ_{adv}) for the adversarial loss, and (λ_c) for the correlation loss.

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spk}} + \lambda_E \cdot \mathcal{L}_{\text{env env}} + \\ \lambda_{adv} \cdot \mathcal{L}_{\text{env spk}} + \lambda_c \cdot \mathcal{L}_{corr} + \mathcal{L}_{\text{recons}} \end{aligned} \tag{4}$$

5 Experimental Setup

5.1 Implementation Details

ECAPA-TDNN: the backbone produces a 3072 dimensional embedding. This embedding was processed through an encoder that reduces its dimensionality to 512. Subsequently, its is partitioned into two distinct components: a speaker embedding ($e_{i,j}^{spk}$) and an environment embedding ($e_{i,j}^{spk}$), each comprising 256 dimensions. For the ECAPA-TDNN baseline, we employ the same architecture proposed by Desplanques et al. (2020).

Autoencoder: The autoencoder architecture incorporates a batch normalization layer followed by a fully connected layer with symmetric input and output dimensions. For the speaker discriminator (S), we employed a single fully connected layer with cross-entropy loss, where the output dimension corresponds to the number of speaker classes. For the environment discriminators (E^E) and (E^S), we implemented a two-layer MLP (first layer: 256-D, second layer:128-D) with batch normalization, ELU activation functions, and a final fully connected layer.

Training: All experiments were conducted on a computational cluster equipped with NVIDIA RTX A5000 GPUs and CUDA version 12.4. Training was performed on two distinct datasets (CML-TTS and synthetic dataset - ST1), with each dataset undergoing 300 epochs of training using Adam optimizer with initial learning rate of 0.001 and a batch size of 128. The loss function incorporated weighted components, with $\lambda_{adv} = 0.5$ for the adversarial loss, $\lambda_E = 0.1$ and $\lambda_{corr} = 0.02$. The discriminator was updated 5 times per training iteration and the weight decay $5e - 5$.

5.2 Evaluation Metrics

We evaluated the learned representation using two standard speaker-verification metrics, Equal Error Rate (EER) and minimum Detection Cost Function (minDCF), because they provide to verify whether disentanglement improves speaker-related structure while reducing sensitivity to environmental factors. EER summarizes the operation where the false acceptance and false rejection rates are equal, and lower values indicate fewer errors when deciding whether two utterances belong to the same speaker. The metric minDCF complements EER by measuring performance under an application driven tradeoff between misses and false alarms, using fixed prior and cost settings; it is obtained by sweeping the decision threshold and taking the minimum cost and lower minDCF therefore indicates better performance in realistic decision.

6 Results and Discussion

We first evaluated the speaker verification performance of the different training strategies using the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF). While these metrics provide a quantitative assessment for ASR, they alone cannot reveal the structural properties of the learned embedding space with respect to environmental factors. Table 2 reports results for all methods on synthetic set and CML-TTS with contrastive methods.

ECAPA-TDNN (baseline): Although ECAPA-TDNN baseline achieves a lower EER with 1.5% and MinDCF=0.06 than the disentanglement model for ST1, this comparison may not be appropriate due to architectural differences between the disentanglement framework and ECAPA-TDNN. The baseline is trained to optimize speaker discrimination on the in-domain objective, and encode nuisance factors such as channel and background acoustics. In ST1, these environmental cues may correlate with speakers, acting as shortcuts that reduce EER. In contrast, the disentanglement discourages the encoding environment dependent information.

Scenario 1 (without environment dissociation) with Triplet Loss: we first train the model with the ST1 dataset using the standard Triplet Loss without environmental dissociation i.e., we form triplets from similar environments, the model learns embeddings that entangle environmental acoustics with speaker identity. This design allows us to evaluate the effect of triplet construction when environmental conditions are not explicitly differentiated. This setup achieves good metrics (EER=5.08, minDCF=0.32), but the visualization (Figure 2) reveals poor disentanglement. Therefore, EER/minDCF improvements alone do not constitute evidence of an embedding space that separates speaker identity from environment.

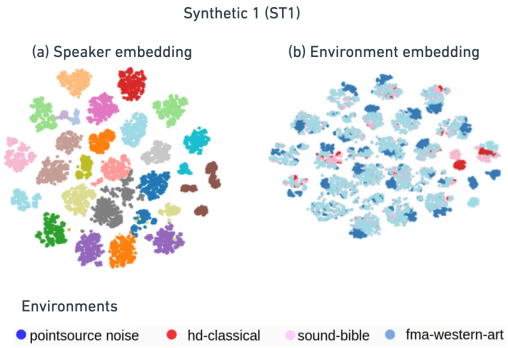

Figure 2: t-SNE of speaker and environment embeddings under Triplet loss without environment-discriminative triplets.

Scenario 2 (with environment dissociation) using Triplet Loss: with environment dissociation using Triplet Loss (anchor and positive same environment and negative distinct) and, presented better performance than scenario 1. Figure 3 shows the effect of enforcing environment dissociation while keeping the speaker fixed within each triplet. Compared to scenario 1, scenario 2 yields a clearer separation across environment categories, but the EER increases to 11.5%. This suggests that, in scenario 1, the model may partially rely on environment-related cues to separate speakers. In scenario 2, the objective is stricter: the embedding must be speaker-discriminative while environment-invariant. This additional constraint makes speaker separation harder, which is consistent with the less

	ST1		CML-TTS	
Contrastive Method	**EER (%)**	**MinDCF**	**EER (%)**	**MinDCF**
Triplet (scenario 1)	5.08	0.32	—	—
Triplet (scenario 2)	11.55	0.75	**14.80**	**0.95**
simCLR	6.00	0.38	20.70	0.98
Supcon	**4.70**	**0.24**	18.50	0.92

Table 2: Comparison of contrastive methods (Triplet Loss, SimCLR, SupCon on ST1 and CML-TTS, reported as EER(%) and MinDCF. Best values per column are in bold. Rows with "—" means the configuration was not evaluated.

compact speaker clusters observed for ST1. On CML-TTS, enforcing environment dissociation improves the separation between environment categories, but increases the EER to 14.80%. This behavior shows the same trade-off described previously between speaker discriminability and environment invariance.

6.1 SupCon and SimCLR methods

SupCon - ST1: SupCon produced an embedding space organization comparable to the structure observed with Triplet Loss (Figure 3(c)). It achieved the best speaker verification performance on ST1 among the evaluated methods (Triplet Loss and SimCLR), with EER=4.70% and MinDCF=0.34. This result is consistent with supervised contrastive learning, which leverages multiple positives and negatives per anchor. This method encourages small intra-speaker distances and larger inter-speaker distances, which can improve the within-speaker structure in the embedding space. The lower EER suggests that samples from the same speaker were pulled closer, while reducing sensitivity to nuisance variability such as environmental conditions.

SupCon - CML-TTS: On CML-TTS, SupCon yielded visible speaker clusters in the projection, however they are more fragmented, or showed overlap, as we observed in Figure 3 (g). In the environment projection, samples appeared more diffusely distributed and mixed throughout the space, with weaker global separation between environments (Figure 3 (h)). This pattern is consistent with embeddings that are closer to environment-invariant, since utterances from different environments are not mapped into clearly distinct regions. However, the quantitative results (EER=18.5%) suggest a trade-off, as mentioned before: some of the factors being suppressed, likely environment-related cues, were also contributing to speaker discrimination in this dataset. The model appears to have been partially relying on environmental information to separate speakers.

SimCLR - ST1: SimCLR achieved performance close to SupCon on ST1 (EER=6%, MinDCF=0.38). However, because SimCLR is self-supervised, the positive pairs are defined as two augmented views of the same sample, and the invariance learned depend on the augmentation set. If the augmentations do not sufficiently reduce environment-related factors, some of this variability may remain in the embedding. In addition, negatives are formed from other samples within the same batch, and the method does not directly enforce grouping across samples from the same speaker recorded under different conditions, which can limit robustness compared to SupCon.

SimCLR - CML-TTS: On CML-TTS, SimCLR also produced speaker groupings in the projection similar to that observed for SupCon. The environment projection showed weaker global separation between environments. In terms of verification, SimCLR obtained EER=20.7%, reinforcing the trade-off.

This highlights the importance of applying discriminative objectives such as SupCon, which explicitly encourage tight within-speaker clustering and clear between-speaker separation, helping preserve speaker-relevant structure even as the representation more robust to environment variation.

6.2 Effect of Acoustic Environment on EER

Table 3 shows that SupCon performance degrades as environmental complexity increases. SupCon achieves lowest EER in hd-classical (4.76%), whose background is harmonic and more stable over time, and therefore tends to interfere less with speaker discriminative cues. In western music

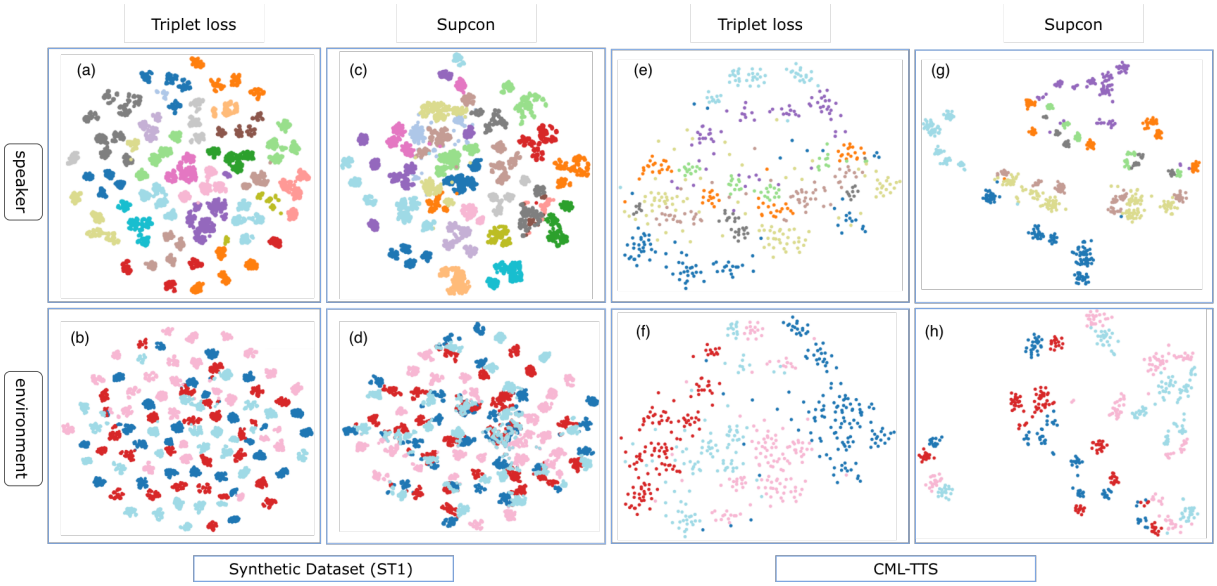

Figure 3: t-SNE visualizations of embeddings learned with a Triplet-loss and SupCon contrastive methods for ST1 and CML-TTS dataset.

(8.57%), rather than being similar to hd-classical, the background often contains simultaneous instruments. The most challenging condition is sound bible (11.43%), which includes non-stationary sound effects that vary over time and can overlap speech across time-frequency.

Environment	EER (%)
hd-classical	**4.76**
point-source	7.69
fma-western	8.57
sound-bible	11.43

Table 3: Equal Error Rate (EER) of SupCon across acoustic environments.

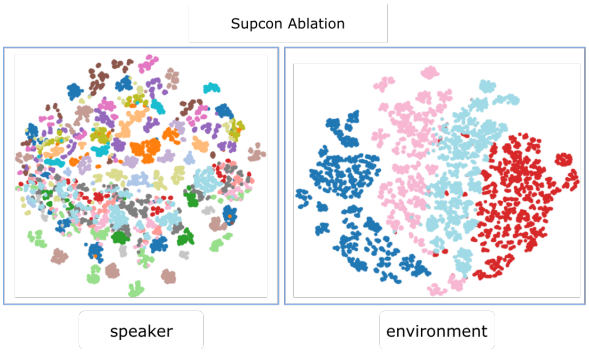

Figure 4: t-SNE visualization of speaker and environment embeddings using ST1 dataset with ablation of SupCon loss.

7 Ablation Studies

To evaluate the contribution of each loss term, we conducted an ablation study on the ST1 dataset using the SupCon configuration, which yielded our best performance. In each experiment, we removed exactly one loss term while keeping the architecture, training schedule, and remaining losses unchanged. Figure 4 illustrates the corresponding t-SNE projections (speaker and environment), and Table 4 reports the speaker verification metrics.

Overall, removing any component leads to a clear degradation in verification performance, indicating that the losses play complementary roles in shaping a representation that is both speaker-discriminative and robust to environmental variation. In Figure 4, removing the environment SupCon term ($\mathcal{L}_{\text{env env}}$) makes the environment projection more clearer separable, while the higher error rates (EER=11.4%, minDCF=0.70). A similar result is observed when removing the other loss terms, highlighting the importance of these additional objectives as inductive biases that promote factor separation and improve disentanglement.

8 Conclusions

The results of this study highlight two central takeaways for speaker representation learning. First, they underscore the importance of controlled syn-

$\mathcal{L}_{adv}$	$\mathcal{L}_{\text{env env}}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{corr}$	**EER** ↓	**MinDCF** ↓
x	✓	✓	✓	11.3	0.76
✓	x	✓	✓	11.4	0.70
✓	✓	x	✓	12.0	0.80
✓	✓	✓	x	11.7	0.80

Table 4: Ablation study on ST1: speaker-verification performance (EER and minDCF) when removing one loss term at a time from the proposed model.

thetic datasets, in which speaker identity and environmental factors are explicitly defined, for the reliable evaluation of disentanglement methods. In such settings, exemplified by the ST1 dataset, supervised contrastive learning (SupCon) achieves both stronger disentanglement and superior speaker verification performance. Crucially, this controlled environment enables a more causal interpretation of the results, ensuring that observed performance gains stem from genuine factor separation rather than from spurious correlations present in real-world data.

Second, our experiments demonstrate the effectiveness of combining contrastive learning with adversarial training as an inductive bias for disentanglement. Contrastive objectives such as SupCon and SimCLR impose a structured geometry on the embedding space, while adversarial training actively suppresses environmental information within the speaker representation. Together, these components promote embeddings that are simultaneously discriminative of speaker identity and invariant to environmental conditions. The ablation study further confirms that each loss term plays a complementary and essential role in maintaining this balance between discrimination and invariance.

Looking ahead, future work will explore alternative disentanglement strategies and the development of richer synthetic benchmarks with a broader set of controllable factors. Such benchmarks will facilitate more fine-grained and diagnostic evaluations of disentanglement quality, supporting the principled design of robust speaker representations.

Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by FAPESP grant #2019/07665-4 and IBM Corporation, and was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of Oct 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44. The authors are also thankful for CNPq 315158/2023-9 fellowship support.

References

Khamis Al-Karawi. 2015. Automatic speaker recognition system in adverse conditions — implication of noise and reverberation on system performance. *International Journal of Information and Electronics Engineering*, 5.

Nathalia Alves Rocha Batista. 2019. Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. Master's thesis, Universidade Estadual de Campinas, Campinas.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. pages 4978–4982.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Proc. Interspeech 2023*, pages 1244–1248.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Zhiwei Deng, Rajitha Navarathna, Peter Carr, Stephan Mandt, Yisong Yue, Iain Matthews, and Greg Mori. 2017. Factorized variational autoencoders for modeling audience reactions to movies. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6014–6023.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 3830–3834. ISCA.

William J Hardcastle, John Laver, and Fiona E Gibbon. 2012. *The handbook of phonetic sciences*. John Wiley Sons.

Woo Hyun Kang, Sung Hwan Mun, Min Hyun Han, and Nam Soo Kim. 2020. Disentangled speaker and nuisance attribute embedding for robust speaker verification. *IEEE Access*, 8:141838–141849.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiachen Lian, Chunlei Zhang, and Dong Yu. 2022. Robust disentangled variational speech representation learning for zero-shot voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6572–6576. IEEE.

Tianchi Liu, Kong Aik Lee, Qiongqiong Wang, and Haizhou Li. 2023. Disentangling voice and content with self-supervision for speaker recognition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Aditya Menon, Raj Gohil, Kumud Tripathi, and Pankaj W. 2025. Laspa: Language agnostic speaker disentanglement with prefix-tuned cross-attention. In *Interspeech*, pages 3623–3627.

Nasser-Eddine Monir, Paul Magron, and Romain Serizel. 2024. A phoneme-scale assessment of multichannel speech enhancement algorithms. *Trends in Hearing*, 28:23312165241292205.

KiHyun Nam, Hee-Soo Heo, Jee-weon Jung, and Joon Son Chung. 2024. Disentangled representation learning for environment-agnostic speaker recognition. *arXiv preprint arXiv:2406.14559*.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *Text, Speech, and Dialogue*, pages 188–199, Cham. Springer Nature Switzerland.

Giuseppe Ruggiero, Matteo Testa, Jurgen Van De Walle, and Luigi Di Caro. 2025. Eta-WavLM: Efficient speaker identity removal in self-supervised speech representations using a simple linear equation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2494–2504, Vienna, Austria. Association for Computational Linguistics.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *CoRR*, abs/1510.08484.

Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien. 2024. Contrastive self-supervised speaker embedding with sequential disentanglement. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:2704–2715.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696.

Yongqiang Wang and Mark J. F. Gales. 2012. Speaker and noise factorization for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):2149–2158.

Yuying Xie, Michael Kuhlmann, Frederik Rautenberg, Zheng-Hua Tan, and Reinhold Haeb-Umbach. 2024. Speaker and style disentanglement of speech based on contrastive predictive coding supported factorized variational autoencoder. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 436–440. IEEE.

X. Xing, M. Xu, and T. F. Zheng. 2024. A joint noise disentanglement and adversarial training framework for robust speaker verification. In *Proceedings of Interspeech 2024*, pages 707–711.

Yuhang Xue, Ning Chen, Yixin Luo, Hongqing Zhu, and Zhiying Zhu. 2024. Clessr-vc: Contrastive learning enhanced self-supervised representations for one-shot voice conversion. *Speech Communication*, 165:103139.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. 2024. Reshape dimensions network for speaker recognition. In *Interspeech 2024*, pages 3235–3239.

Olivier Zhang, Olivier Le Blouch, Nicolas Gengembre, and Damien Lolive. 2023. An extension of disentanglement metrics and its application to voice. *Proc. Interspeech 2023*, pages 2878–2882.