information

MDPI

Article

Understanding Self-Supervised Learning of Speech Representation via Invariance and Redundancy Reduction

Yusuf Brima [1,2,*], Ulf Krumnack [1], Simone Pika [2] and Gunther Heidemann [1]

[1] Computer Vision, Institute of Cognitive Science, Osnabrueck University, 49074 Osnabrück, Germany; gheidema@uni-osnabrueck.de (G.H.)

[2] Comparative BioCognition, Institute of Cognitive Science, Osnabrueck University, 49074 Osnabrück, Germany

* Correspondence: ybrima@uos.de

Abstract: Self-supervised learning (SSL) has emerged as a promising paradigm for learning flexible speech representations from *unlabeled* data. By designing *pretext tasks* that exploit statistical regularities, SSL models can capture *useful* representations that are *transferable to downstream tasks*. Barlow Twins (BTs) is an SSL technique inspired by theories of redundancy reduction in human perception. In downstream tasks, BTs representations accelerate learning and transfer this learning across applications. This study applies BTs to speech data and evaluates the obtained representations on several downstream tasks, showing the applicability of the approach. However, limitations exist in disentangling key explanatory factors, with redundancy reduction and invariance alone being insufficient for factorization of learned latents into *modular*, *compact*, and *informative* codes. Our ablation study isolated gains from invariance constraints, but the gains were context-dependent. Overall, this work substantiates the potential of Barlow Twins for sample-efficient speech encoding. However, challenges remain in achieving fully hierarchical representations. The analysis methodology and insights presented in this paper pave a path for extensions incorporating further inductive priors and perceptual principles to further enhance the BTs self-supervision framework.

Keywords: acoustic analysis; Barlow Twins; self-supervised learning; invariance; redundancy reduction; speech representation learning

check for updates

Citation: Brima, Y.; Krumnack, U.; Pika, S.; Heidemann, G. Understanding Self-Supervised Learning of Speech Representation via Invariance and Redundancy Reduction. *Information* **2024**, *15*, 114. https://doi.org/10.3390/info15020114

Academic Editors: Krzysztof Ejsmont, Aamer Bilal Asghar, Yong Wang and Rodolfo Haber

Received: 23 January 2024
Revised: 8 February 2024
Accepted: 9 February 2024
Published: 15 February 2024

1. Introduction

Speech processing plays a pivotal role in diverse applications, spanning speaker identification, diarization, spoken language understanding, speaker segmentation, voice assistants, etc. [1–5]. The extraction of linguistic and para-linguistic features from speech data is essential for ensuring accurate and robust performance within these application domains. Despite the effectiveness of conventional supervised learning methods [2,6], their heavy reliance on labels as *supervisory signals* poses challenges due to the scarcity and cost associated with obtaining such labels [7–9].

Self-supervised learning (SSL) has emerged as a paradigm for learning flexible "universal" representations from *unlabeled* data by exploiting inherent statistical regularities as supervisory signals. A core tenet of SSL is designing *pretext tasks* (i.e., crafting tasks that serve as a context for learning) to train deep learning models to capture intrinsic statistical structures within inputs without the need for human labeling. In speech, abundant redundancies exist within audio content regarding linguistic content, speaker characteristics, emotions, etc. SSL leverages these ubiquitous patterns in speech through extensive use of data augmentation and context-based predictive pretext tasks. These include predicting masked time-frequency spectrogram components from neighboring regions or contrastive learning objectives judging different (distorted) versions of the same underlying utterance as identical [10,11]. Such techniques enable models to focus representations on speaker and/or language information while discarding nuisance variations such as background noise.

A recently proposed cognitive-neuroscience-inspired framework builds upon progress in SSL for speech by aligning with the principles of redundancy reduction characterized by Horace Barlow [12]. Specifically, Barlow Twins (BTs) adopt a *joint embedding architecture* (JEA) trained to produce consistent encoder representations between differently augmented views of the same input [13]. This, in the context of speech, aims to emulate auditory sensory perception efficacy by amplifying speaker-related cues while suppressing irrelevant variations. The integration of core redundancy minimization concepts and the general SSL paradigm holds promise for improving the *sample efficiency*, *flexibility*, and *biological plausibility* of self-supervised speech-encoding techniques to build *robust cognitive schema of auditory representations*.

However, the utility of this framework in achieving *distributed*, *disentangled*, and *invariant* representations remains underexplored. Therefore, this paper undertakes an empirical analysis on three fronts to address open questions:

- **Downstream Task Efficacy**: Quantitative effectiveness on a select speech-processing task;
- **Disentanglement Analysis**: A quantitative assessment of representation decoupling quality and its axis alignment with ground-truth explanatory factors;
- **Objective Variants**: Ablations examine the impact of training components.

This investigation systematically examines the BTs framework, focusing specifically on the representational quality and performance attributable to its redundancy reduction principles. Our study, centered on analytical evaluation rather than competitive benchmarking, provides novel evidence about the framework's suitability for learning useful speech representations. The aim is to substantiate the utility of BTs while identifying potential advancements for sample-efficient speech-encoding models. Although an exhaustive benchmarking of various methods falls beyond this study's scope, our contribution lies in the rigorous assessment of Barlow Twins. We evaluate this framework as a *simple* yet effective approach, particularly in the realms of invariance and redundancy reduction, which are crucial for learning useful speech representations.

We structure the subsequent sections as follows. In Section 2, the Materials and Methods detail our proposed self-supervised speech representation learning framework. Section 3 benchmarks performance across diverse speech tasks and datasets, quantifying emerging representation quality through *disentanglement* analyses and ablation studies of loss objective variants. Section 4 analyzes the result outcomes. Finally, Section 5 encapsulates key contributions, synthesizes insights derived from our study, and provides a succinct summary, offering a conclusive wrap-up to the paper.

Related Works

Self-supervised learning (SSL) methods have gained traction in speech processing for their ability to learn representations without manual annotations [10,14–20]. In this context, the complexities of existing SSL techniques, such as wav2vec 2.0 [21] and HuBERT [22], often involve specialized negative sampling, stop gradients, and *intricate training recipes*. These complexities, while contributing to the effectiveness of these methods, can also pose challenges to their flexibility and adaptability.

Wav2vec 2.0 and HuBERT represent state-of-the-art SSL techniques in speech processing. Wav2vec 2.0 employs a contrastive learning approach, where the model learns to distinguish between positive and negative samples by maximizing agreement between positive pairs and minimizing it between negative pairs [21]. This requires careful handling of negative samples and intricate training recipes to ensure convergence and effectiveness.

HuBERT, on the other hand, focuses on a masked language modeling approach combined with contrastive learning, leveraging hierarchical structures for representation learning [22]. The model involves complex strategies such as the predictive masking of hidden units and k-means clustering to enhance the quality of speech embeddings. These methods, while successful, introduce challenges related to the need for specialized negative sampling and the delicate balance required during training.

In contrast, the Barlow Twins (BTs) framework offers a conceptually *simpler* SSL approach, relying only on data augmentation (*multi-view creation)* and a redundancy reduction and invariance objective [13]. This simplicity is achieved through maximizing the cross-correlation between augmented views of inputs while minimizing cross-sample cross-correlation.

Motivated by this revelation of simplicity versus complexity trade-offs, this study seeks not to outperform the state of the art, but rather to conduct an extensive empirical analysis quantifying the utility of adopting the BTs framework for speech representation learning. Through an evaluation of diverse downstream speech tasks and datasets, we center our investigation on assessing the quality of learned representations along pertinent axes of generalization, disentanglement, and factorial representation of key speech factors.

2. Materials and Methods

2.1. Learning Framework

The Barlow Twins (BTs) framework, as depicted in Figure 1, employs a *joint embedding architecture* (JEA) to learn invariant representations. Specifically, it uses an encoder network f_θ to project augmented views of speech within a mini-batch—denoted as X^A and X^B—into a *shared latent space*, producing latent representations Z^A and Z^B, respectively (see Figure 1). The key idea is that differently augmented views of the same underlying speech sample should have similar latent variables, while views from different samples should be de-correlated in the latent space.

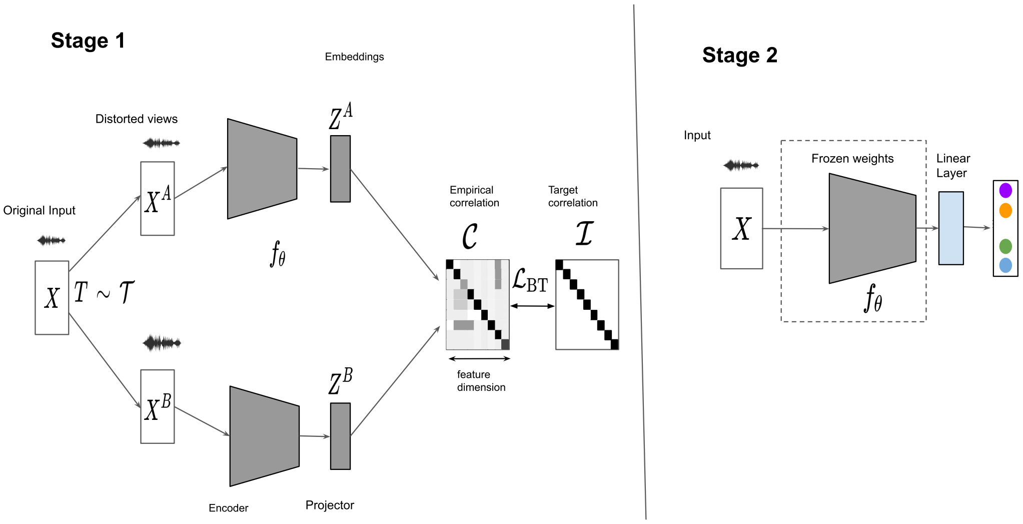


Figure 1. The BTs framework for learning invariant speech representations. **Stage 1:** An encoder f_θ process augments views X^A and X^B of the same speech input X and projects them into a shared latent space. The BTs' loss (Equation (1)) enforces redundancy reduction between latents from different samples while maximizing correlation for positive pairs (two views of the same sample). This causes the encoders to produce invariant representations capturing speaker identity while reducing sensitivity to augmentations. **Stage 2:** The learned latent representations Z^A and Z^B can then be used for downstream speech-processing tasks to evaluate the model's generalization capability.

This is formalized through the two-component optimization objective in Equation (1). First, to encourage representations for a positive pair, that is, two augmented views of the same sample, to be similar, the cross-correlation matrix C_{ij} in Equation (2) between the latent variables Z^A and Z^B should have diagonal elements close to 1 (the *invariance* part of the loss function). Second, the redundancy term enforces de-correlation between latent variables from different samples. This has the combined effect of making Z^A and Z^B

invariant for positive pairs, while also reducing redundancy across the mini-batch. Training the encoder f_θ with this learning objective thus produces a representation space with useful properties for downstream tasks.

$$\mathcal{L}(C;\lambda) \triangleq \underbrace{\sum_i (1-C_{ii})^2}_{\text{invariance}} + \lambda \underbrace{\sum_i \sum_{j\neq i} (C_{ij})^2}_{\text{redundancy reduction}} , \tag{1}$$

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}. \tag{2}$$

In Figure 2, we illustrate the creation of two views crucial to our SSL approach. The left column showcases View 1, offering both the time-domain representation (top row) and the corresponding time-frequency spectrogram (second row), both derived from the first perturbed version of the original audio signal. On the right column, View 2 mirrors this representation, providing a parallel set of time-domain and spectrogram views. These views, capturing variations within the input signal, form the foundation for our SSL framework, enabling the model to glean invariant information while attenuating irrelevant variations.

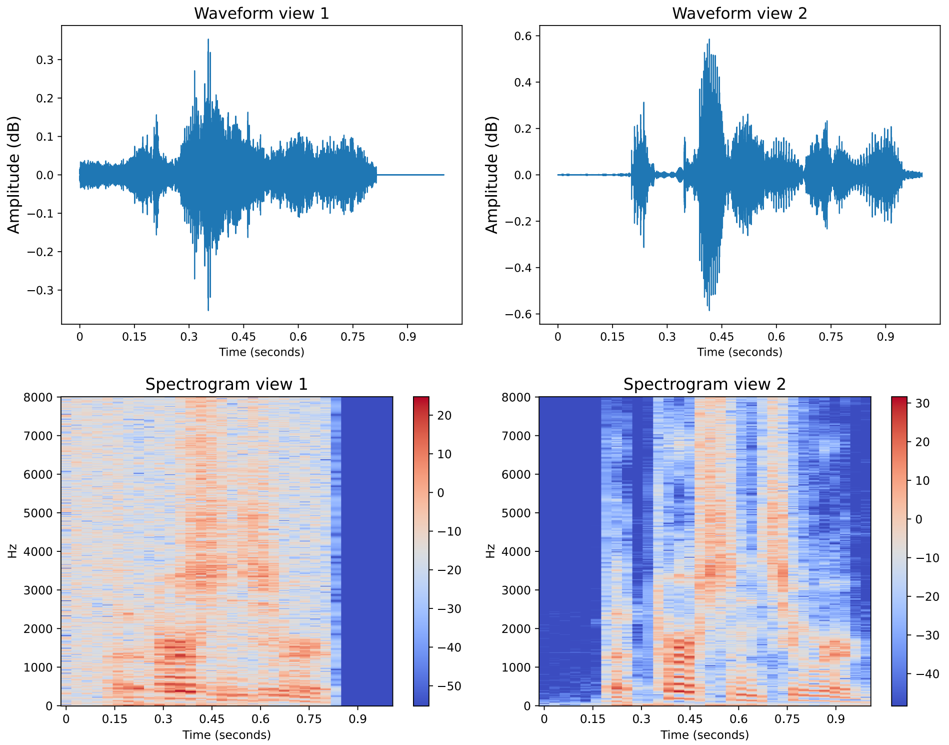


Figure 2. (**Left column**) View 1 provides a dual representation, featuring the time-domain signal (**top row**) and its corresponding time-frequency spectrogram **(second row**), both derived from the first perturbed version of the original audio signal. (**Right column**) View 2 presents a similar pair of representations. The higher harmonic partials present in the first view are not visibly present in the second view; however, the underlying information content remains invariant.

2.2. Datasets

We utilize a diverse collection of speech datasets, summarized in Table 1, to train representation models (upstream) and evaluate practical applications (downstream). For upstream representation learning, we leverage *VoxCeleb-1* , *LibriSpeech-100*, and *LibriSpeech-360*, which provide a wide variety of speakers and types of speech. VoxCeleb-1 contains over 100,000 utterances from 1211 celebrities, while LibriSpeech-100 and LibriSpeech-360 consist of readings of audiobook excerpts by 128 and 921 speakers, respectively. This diversity of training data is crucial for learning robust and generalizable speech representations.

We assess the learned representations on different downstream tasks, allowing us to assess the practicality of the learned representation in various tasks. The *Google Speech Commands* dataset provides a collection of spoken words for keyword spotting, a task gaining importance in the context of voice controlling smart devices. The *Emotional Speech Dataset (ESD)* aims at inferring speaker emotions from voice data, supporting potential applications in human–machine interaction. Finally, the *World Leaders at the US Congress (WLUC)* dataset, consisting of world leader speeches, will be used to test the suitability of the representations for speaker and gender identification. Performance on these downstream tasks indicates how informative and transferable the upstream representations are for speech-processing objectives.

Table 1. Summary of upstream and downstream datasets. "Upstream" tasks refer to self-supervised training, in our case, optimizing for the BTs learning objective of redundancy reduction and invariance of the multi-view representations, while "Downstream" tasks include keyword spotting, emotional tone recognition, speaker identification, and gender recognition.

Source	Dataset Name	# Samples	# Classes	Duration (h)	Usage
[23]	VoxCeleb-1	148,642	1211	340.39	Upstream
[24]	LibriSpeech-100	14,385	128	100	Upstream
[24]	LibriSpeech-360	104,935	921	360	Upstream
[25]	Speech Commands	7985	2	2.18	Downstream
[26]	ESD	7000	2	5.52	Downstream
[27]	WLUC	7500	5	2.05	Downstream

By learning representations on diverse upstream datasets and testing generalization capability through varied downstream tasks, we comprehensively evaluate the models' capabilities. The multi-dataset, multi-task framework provides a rigorous methodology for representation learning and evaluation in speech processing.

2.3. Experimental Setup

In our experimental setup, we followed established practices for SSL and utilized a ResNet-50 backbone for pre-training, as proposed in the original paper and consistent with other SSL frameworks. This choice of backbone architecture is well-established in the literature and provides a robust foundation for learning representations from the audio datasets outlined in Table 1. The pre-training process involved 50 epochs for each upstream model, with a mini-batch size of $n = 64$ due to computational constraints and a latent dimensionality of $m = 2028$, ensuring a comprehensive exploration of the feature space. Additionally, our audio pre-processing, including standardized sampling rates and the generation of log-scaled spectrograms, laid the groundwork for effective model training and subsequent evaluation.

To facilitate optimal learning, we applied a consistent pre-processing pipeline to all audio samples. Initially, we standardized the sampling rate of the samples to 16 kHz. Subsequently, each audio segment underwent partitioning into contiguous 1- second intervals, ensuring uniform input lengths for subsequent processing.

A pivotal aspect of the feature extraction process involved the generation of log-scaled spectrograms. By employing a window size of 64 milliseconds with a 32-millisecond hop size, we captured 513 mel-frequency bins spanning the audible frequency range of 0 to 8 kHz. The resulting spectrograms, denoted as $X \in \mathbb{R}^{513\times126}$, encapsulated both frequency and temporal information. These spectrograms formed the foundation of our neural network architecture, serving as input tensors $X_B \in \mathbb{R}^{n\times1\times513\times126}$, where n represents the mini-batch size. This audio pre-processing ensured a standardized and informative representation, crucial for effective model training and subsequent evaluation.

3. Results

3.1. Effect of Upstream and Downstream Dataset Sizes

Our results in Table 2 demonstrate that downstream task performance generally improves with more in-domain data, as evidenced by the increasing accuracy with larger dataset fractions. However, we achieve substantial gains even with very small downstream sets (5–10%) by transferring self-supervised upstream representations, validated via linear evaluation. This showcases the transferability of the learned features without the need for extensive manual annotations.

Intriguingly, we find that LibriSpeech-100, the smallest upstream corpus, drives the strongest downstream gains—achieving over 80% accuracy in speaker and gender recognition with just 50% of the target data. More notably, with full downstream sets, it exceeds the performance of the larger upstream datasets on all four tasks. This reveals that rather than sheer dataset size, quality is more crucial for representation generalization—aspects at which LibriSpeech-100 excels due to expert voice actors used and minimal noise.

Table 2. Top test performance evaluation across 4 downstream tasks—Speaker Recognition (SR), Gender Recognition (GR), Keyword Spotting (KWS), and Emotional Tone Recognition (ER)—utilizing varied fractions of these respective downstream datasets.

	Fraction (%)	Supervised	LibriSpeech-100	LibriSpeech-360	VoxCeleb1
SR	5	34.21	39.47	28.95	36.84
	10	54.67	64.00	54.67	48.00
	50	75.20	83.73	77.60	68.00
	100	84.53	84.93	81.20	75.07
GR	5	66.67	70.00	63.33	66.67
	10	75.00	71.67	75.00	68.33
	50	79.67	88.67	87.00	78.33
	100	62.67	90.17	88.67	84.67
KWS	5	47.50	52.50	62.50	45.00
	10	51.25	52.50	50.00	50.00
	50	50.76	55.33	52.28	47.72
	100	50.32	75.03	60.08	53.99
ER	5	48.57	54.29	51.43	68.57
	10	61.43	47.14	44.29	50.00
	50	46.57	46.29	50.86	48.86
	100	83.00	59.00	51.00	46.29

We show that smaller upstream datasets, while limited in volume, can unlock substantial transfer potential if the data are diverse, high-quality, and relevant to target domains. Specifically, LibriSpeech-100, despite its modest size, drives the strongest performance owing to its inclusion of a variety of professional speakers and minimal artifacts. This suggests curation may supersede the scale of the raw dataset.

However, while transferred features accelerate downstream learning, sufficient in-domain supervision remains indispensable for maximizing overall performance. This is evidenced by accuracy gaps when using 100% training data between self-supervised and supervised paradigms. Therefore, effectively pre-trained representations complement, rather than replace, target task annotations.

Additionally, we find that task complexity and similarity across domains modulate the transferability of representations. Simpler objectives like speaker recognition mature faster with less task-specific data. But complex tasks, like emotion recognition, necessitate more in-domain data. Likewise, the affinity between pre-training and targets boosts feature usability—VoxCeleb-1 specializes in speaker cues.

Thus, high-quality, diverse self-supervised pre-training can unlock substantial value from modest downstream supervision, but task complexity, dataset relevance, domain similarity, and in-domain data size interact to determine performance gains. Carefully navigating these factors is key to optimizing representation transfer from upstream tasks to domain-specific problems.

3.2. Can Enforcing Redundancy Reduction and Invariance Result in Disentanglement?

To explore the disentanglement of latent variables in the learned representations in upstream models, we employ various disentanglement metrics, including Mutual Information Gap (MIG) [28]; Joint Entropy Minus Mutual Information Gap (JEMMIG) [29]; Disentanglement, Completeness, and Informativeness MIG (DCIMIG) [30]; Attribute Predictability Score (SAP) [31]; and Modularity Score [32]. By focusing on factors such as accent, identity, and gender, we aim to quantify and evaluate the degree to which our models disentangle these specific attributes from the overall representation. These metrics provide valuable insights into the *modularity*, *compactness*, and *informativeness* of trained BTs models, shedding light on the nuanced aspects of the learned latent space.

To visually assess the nature of learned representations, Figure 3 compares representations of both randomly initialized and trained Barlow Twins networks. For the trained network, we observe a nearly perfect correlation along the diagonal of the cross-correlation matrix, indicating invariance between augmented views of the same input speech sample. Additionally, the off-diagonal elements are pushed closer to zero, demonstrating redundancy reduction between latents from different samples. In contrast, the untrained network shows no clear regularity. Employing such visualization techniques provides valuable insights into the structure of the learned representation space. Our analysis verifies that the BTs objective successfully enforces invariance and redundancy reduction between two augmented views of the speech input. This is a crucial step in quantifying the model's capacity to capture intricate information in speech data while attenuating nuisance variation.

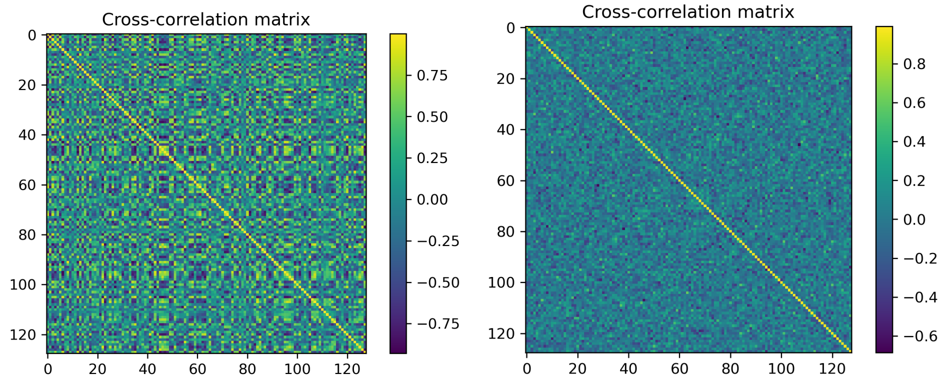


Figure 3. The empirical cross-correlation between the 128 features of the latent representations Z^A and Z^B for paired augmented views, contrasting the untrained state (**left**) with the trained state (**right**) within the BTs framework. These matrices visually represent the relationships between different views of the same speech input for the current mini-batch. The comparison allows us to observe the transformation in cross-correlation patterns following the self-supervised learning process, highlighting the model's ability to capture invariance (higher correlation of diagonal elements of the trained network's matrix) and de-correlation of off-diagonal elements.

Analyzing the suite of disentanglement metrics in Table 3, we assess if simply enforcing redundancy reduction and invariance through the core BTs learning objective can factorize learned representations along explanatory attributes without further constraints.

The consistently low scores across crucial metrics like MIG (0.020 max) and SAP (0.037 max) indicate that this training alone de-correlates but does not fully factorize key explanatory factors. While higher modularity scores (0.696 max) confirm the clustering of semantic information, specificity along individual latent dimensions remains insufficient. This highlights the need to combine complementary techniques that impose stricter decomposition in order to realize fully compact and decoupled representations.

Table 3. Disentanglement metrics with standard deviation for BTs models over 50 evaluation runs each. Values are presented as mean ± standard deviation. BT-n denotes models with a latent dimensionality of n. BT-LS-100 and BT-LS-360 indicate models trained on LibriSpeech-100 and LibriSpeech-360 datasets, respectively. BT-VC-1 represents the model trained on the VoxCeleb-1 dataset. This table summarizes disentanglement quantification for Barlow Twins models trained on various datasets and with differing latent capacities to elucidate the impact of these factors.

	Compactness		**Holistic**		**Modularity**
Model	**MIG**	**SAP**	**DCIMIG**	**JEMMIG**	**Mod. Score**
BT-16	0.004 ± 0.0022	0.033 ± 0.0016	0.013 ± 0.0004	0.191 ± 0.0120	0.659 ± 0.0043
BT-32	0.004 ± 0.0011	0.010 ± 0.0020	0.013 ± 0.0004	0.212 ± 0.0154	0.707 ± 0.0040
BT-64	0.008 ± 0.0012	0.007 ± 0.0015	0.012 ± 0.0003	0.182 ± 0.0089	0.652 ± 0.0031
BT-128	0.020 ± 0.0019	0.019 ± 0.0020	0.010 ± 0.0004	0.231 ± 0.0047	0.664 ± 0.0029
BT-512	0.003 ± 0.0019	0.003 ± 0.0012	0.010 ± 0.0004	0.266 ± 0.0252	0.675 ± 0.0022
BT-1024	0.007 ± 0.0010	0.025 ± 0.0018	0.016 ± 0.0004	0.238 ± 0.0175	0.681 ± 0.0021
BT-2048	0.005 ± 0.0015	0.014 ± 0.0020	0.012 ± 0.0005	0.294 ± 0.0082	0.691 ± 0.0018
BT-LS-100	0.007 ± 0.0012	0.016 ± 0.0024	0.013 ± 0.0005	0.141 ± 0.0091	0.685 ± 0.0021
BT-LS-360	0.006 ± 0.0012	0.037 ± 0.0029	0.015 ± 0.0005	0.109 ± 0.0070	0.696 ± 0.0032
BT-VC-1	0.007 ± 0.0012	0.018 ± 0.0025	0.008 ± 0.0005	0.113 ± 0.0063	0.619 ± 0.0044

However, abysmal gains along the compactness axes for higher-dimensionality models like BT-2048 (MIG: 0.005) over BT-16 (MIG: 0.004) showcase the diminishing effect of dimensionality in contrast to BT-128 (MIG: 0.020), which potentially enables more granular decoupling. Furthermore, training with a wider variety of data, such as in BT-LS-360, improves both clustering and the retention of compactness (MIG: 0.006, Modularity: 0.696), elucidating the value of using diverse training corpora.

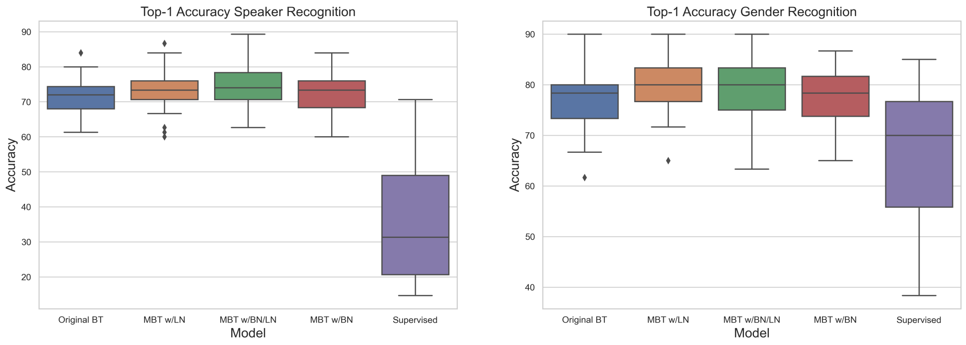
Therefore, while invariance and redundancy reduction induce minor factorization of informative factors of variation and the disposal of irrelevant variation, additional explicit constraints must complement these objectives to achieve fine-grained disentanglement along speech factors like speaker traits, accents, emotions, and linguistic content throughout the latent feature hierarchy. Our analysis quantitatively demonstrates this limitation while revealing pathways to potentially facilitate targeted factorization through greater model capacity and diverse training data.

3.3. Ablation of Loss Function Variants

In this section, we conduct an extensive ablation study on variants of the BTs loss functions, evaluating their impact on learned representations. Figures 4 and 5 present the results of this investigation, comparing the original Barlow Twins (BTs) with several modified versions. Specifically, we analyze the Modified Barlow Twins with Latent Normalization (MBT w/LN), Modified Barlow Twins with Batch Normalization and Latent Normalization (MBT w/BN/LN) (column-wise), and Modified Barlow Twins with Batch Normalization (MBT w/BN). The ablation concludes with a benchmark using the standard supervised method.

Building on this setup, the empirical results are illustrated in Figure 4a. This plot reveals the Top-1 accuracy of different models in the context of speaker recognition. We first note the original BTs model, which exhibits a median accuracy of around 70%, paired with a relatively symmetrical interquartile range (IQR) and outliers that suggest variations in performance. In contrast, the MBT w/LN model demonstrates a similar median but with a notably tighter IQR, indicating more consistent results. A slight deviation is observed in the MBT w/BN/LN model, which has a marginally lower median accuracy and a larger IQR, pointing to greater variability in its performance. A notable divergence is seen in the MBT w/BN model, characterized by a much wider range and a lower median accuracy, with outliers indicating instances of particularly low performance. Interestingly, the supervised model markedly stands out with a a significantly lower median accuracy and a large IQR, underscoring its consistent underperformance relative to the other models. The presence of outliers, especially noticeable in the original BT and MBT w/BN models, suggests instances where the models either excel or fall dramatically short. Overall, the MBT models incorporating layer normalization (LN) appear to strike a desirable balance between achieving high accuracy and ensuring result consistency, while the supervised model exhibits considerable limitations in accuracy for this specific task.

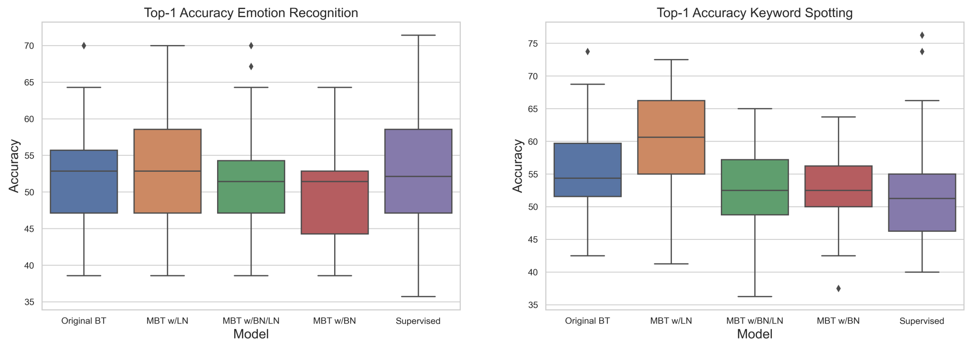
In Figure 4b, we can see the Top-1 accuracy of various models in gender recognition tasks. The original BTs model's median accuracy is situated just above 70%, with a relatively broad IQR, indicating some variability in its performance. Notably, there are a few outliers that fall significantly below the lower quartile, which may point to specific instances where the model underperforms. Moving on to the MBT w/LN model, we notice a higher median accuracy and a narrower IQR, suggesting that this model not only performs better on average but also does so more consistently. The MBT w/BN/LN demonstrates a median accuracy comparable to MBT w/LN, but with a slightly wider IQR, indicating a bit more inconsistency in its results. In contrast, the MBT w/BN model exhibits a lower median accuracy and the widest IQR of all the MBT models, showing substantial variability in performance. Lastly, the supervised model shows a significantly lower median accuracy, below 60%, and a very wide IQR, which implies that while it can occasionally perform well, it is generally less reliable than the other models. The presence of outliers in the original BT and MBT w/LN models suggests that there are occasional deviations in performance, which could be due to a variety of factors such as model overfitting, anomalies in the test data, or limitations inherent to the models themselves. Overall, the MBT w/BN/LN model seems to offer the best balance between accuracy and reliability for gender recognition tasks.



(**a**) (**b**)

Figure 4. (**a**) Top-1 accuracy for speaker recognition, comparing five base models over 50 experimental runs, highlighting the performance and stability of these techniques. (**b**) Top-1 accuracy for gender recognition from speech, using the same base models, which shows a similar performance trend, indicating task-specific model effectiveness and the nuanced nature of gender features in speech data.

To extend our investigation beyond speaker representation, we further explore the impact of these loss function variants on emotion recognition and keyword-spotting tasks. Figure 5a showcases the results for emotion recognition accuracy, while the second plot illustrates accuracy in keyword spotting. In this plot, we compare the Top-1 accuracies across different models for emotion recognition. The original BTs model shows a median accuracy slightly above 55%, with a broad IQR which indicates a fair amount of variability in performance. The model also exhibits outliers, suggesting that some predictions are notably different from the rest. The MBT w/LN model presents a higher median accuracy near 60% and a slightly narrower IQR, implying more consistent performance than the original BTs. The MBT w/BN/LN has a similar median to the MBT w/LN but with an even tighter IQR, which may indicate a higher level of consistency in its emotion recognition capabilities. In contrast, the MBT w/BN shows a lower median accuracy and a wider IQR, indicating less reliability. Lastly, the supervised model shows a median accuracy comparable to MBT w/BN, but with the widest IQR of all the models, signifying the most variability in its accuracy. This analysis suggests that while no model excels at emotion recognition with high accuracy, the MBT w/LN and MBT w/BN/LN models perform more consistently than the others, with the supervised model being the least consistent and potentially overfitting or not generalizing well to the emotion recognition task.



(**a**) (**b**)

Figure 5. (**a**) Boxplot of Top-1 accuracy in emotion recognition across five different base models over 50 experimental runs, showing the consistency and variability in model performances. (**b**) Boxplot of Top-1 accuracy in a keyword-spotting task for the same base models and number of runs, illustrating the impact of model architecture on task-specific accuracy.

Shifting our focus to the parallel task of keyword spotting, the corresponding part of Figure 5b offers an intriguing comparison. Examining this subfigure, we observe the performance of various models in the task of Top-1 accuracy in keyword spotting. The original BTs model has a median accuracy of just above 50%, with a moderate IQR, suggesting a decent consistency in performance. However, there is a noticeable lower outlier that could indicate an occasional significantly lower deviation from the median. The MBT w/LN model presents a higher median accuracy, around 65%, and a tighter IQR, which points to a more consistent and accurate performance in spotting keywords. MBT w/BN/LN shows a median accuracy comparable to MBT w/BN but with a slightly broader IQR, implying a bit more variability. The MBT w/BN model indicates a lower median accuracy, near 50%, and a wider IQR, signifying less reliable performance. Finally, the supervised model exhibits a median accuracy similar to MBT w/BN, but with the widest IQR of all the models, reflecting substantial inconsistency in its keyword-spotting capability. Outliers in the original BT, MBT w/BN/LN, and supervised models suggest that certain keywords may be particularly challenging for these models. In summary, while all models show potential for keyword spotting, MBT w/LN demonstrates the best combination of high accuracy and consistency, with the supervised model appearing to be the least stable.

While minor advantages of latent normalization were observed across the downstream tasks, our broader analysis of these tasks paints a more nuanced picture. In this ablation study, we evaluated variants of the BTs objective on several speech-processing tasks. Our results suggest that incorporating normalization into the loss can potentially improve model accuracy and reliability, depending on the specific downstream task. However, further investigation is needed to determine if these trends hold more broadly, as the benefits were not conclusively demonstrated for all tasks. While variants like the Modified Barlow Twins with Latent Normalization showed promise, claiming definitive improvements would require more extensive experimentation and analysis. This study provides an initial path suggesting that modifying the Barlow Twins objective may yield benefits, motivating further research into enhanced SSL techniques.

4. Discussion

This study provides an extensive empirical analysis of the Barlow Twins (BTs) framework for SSL in speech representation. Our findings affirm the framework's effectiveness, while also highlighting critical areas for further development and exploration.

4.1. Generalization in Downstream Tasks

The BTs framework demonstrates notable success in generalization across various downstream tasks, such as speaker recognition, gender detection, emotion recognition, and keyword spotting. Remarkably, models trained on LibriSpeech-100 achieved over 80% accuracy in speaker identification with only half of the labeled data, suggesting that a dataset having a curated quality may be more crucial than its size. This insight opens up opportunities for optimizing dataset selection in speech-processing tasks, focusing on quality and diversity rather than volume alone.

However, the transferability of learned representations is influenced by task complexity and domain alignment. Simpler tasks, such as speaker recognition, benefit more rapidly from pre-trained models, while more complex tasks like emotion detection necessitate greater amounts of domain-specific data. This variation underscores the need for task-specific fine-tuning and adaptation of pre-trained models, particularly when dealing with complex or nuanced speech-processing tasks.

4.2. Disentanglement of Latent Representations

Our disentanglement analysis reveals a significant area for improvement in the BT framework. Despite achieving redundancy reduction and invariance, the framework falls short in optimally disentangling key explanatory factors in speech, as indicated by the low MIG and SAP scores. This limitation points to the necessity of integrating additional mechanisms or constraints to enhance the disentanglement capabilities of the framework. The potential for leveraging greater model capacity, as shown by the improved compactness in higher-dimensional models, and the benefits of diverse training data, suggest paths forward. Future research should focus on developing and incorporating novel architectural models and inductive priors that can facilitate more effective and targeted factorization of speech attributes.

4.3. Inconsistencies across Different Tasks

The ablation studies conducted as part of this research provide valuable insights but also highlight inconsistencies across different tasks. While improvements were observed in certain scenarios, such as emotion recognition and keyword spotting with latent space normalization, these were not uniformly seen across all tasks. This inconsistency calls for a more nuanced understanding of how different components of the loss function and other architectural choices affect various speech-processing tasks. Further investigation and experimentation are needed to establish more definitive conclusions about the efficacy of these modifications.

5. Conclusions

This work provides a thorough empirical evaluation of the Barlow Twins framework for self-supervised representation learning for speech data. Our findings validate the efficacy of this approach in achieving generalization across diverse downstream tasks, providing a new representation scheme that can be applied to various practical applications. Especially in situations hindered by data sparsity, such general representations may be a useful tool for obtaining a resource-constrained system.

Our experiments underscore the importance of dataset quality over size. However, the results also reveal limitations in disentangling key explanatory factors within the learned representations, despite redundancy reduction and invariance constraints. Additional techniques are needed to enable fine-grained factorization of key explanatory factors in speech. The ablation studies isolated gains attributable to invariance, but inconsistent advantages were found to motivate enhancements in the framework. To address these limitations and advance Barlow Twins' application to speech, incorporating perceptual principles, speech-specific pretext tasks, and comparative benchmarking are proposed as directions of future work. In summary, this investigation substantiates the sample efficiency and emerging utility of Barlow Twins while paving the path for continued progress through our rigorous assessment methodology. Key challenges exist in realizing fully decoupled hierarchical representations, motivating tailored pretext tasks and constraints to enable more granular speech attribute factorization. Overall, this work not only validates the potential of self-supervised learning for speech processing but also opens avenues for enhancing these techniques toward more efficient and robust representations.

Author Contributions: Conceptualization, S.P. and Y.B.; methodology, Y.B.; software, Y.B.; validation, U.K. and Y.B.; formal analysis, Y.B. and U.K.; investigation, Y.B.; resources, Y.B.; data curation, Y.B.; writing—original draft preparation, Y.B.; writing—review and editing, Y.B. and U.K.; visualization, Y.B.; supervision, G.H. and S.P.; project administration, Y.B.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by funds of the research training group (RTG) in "Computational Cognition" (GRK2340) provided by the Deutsche Forschungsgemeinschaft (DFG), Germany, and an EU-Consolidator grant (772000, TurnTaking).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are publicly available and accessible.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Togneri, R.; Pullella, D. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits Syst. Mag.* **2011**, *11*, 23–61.
2. Tirumala, S.S.; Shahamiri, S.R. A review on deep learning approaches in speaker identification. In Proceedings of the 8th International Conference on Signal Processing Systems, Auckland, New Zealand, 21–24 November 2016; pp. 142–147.
3. Lukic, Y.; Vogt, C.; Dürr, O.; Stadelmann, T. Speaker identification and clustering using convolutional neural networks. In Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Salerno, Italy, 13–16 September 2016; pp. 1–6.
4. Trong, T.N.; Hautamäki, V.; Lee, K.A. Deep Language: A comprehensive deep learning approach to end-to-end language recognition. In *Odyssey*; Harper and Row Publishers Inc.: San Francisco, CA, USA, 2016; Volume 2016, pp. 109–116.
5. Adaloglou, N.; Chatzis, T.; Papastratis, I.; Stergioulas, A.; Papadopoulos, G.T.; Zacharopoulou, V.; Xydopoulos, G.J.; Atzakas, K.; Papazachariou, D.; Daras, P. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Trans. Multimed.* **2021**, *24*, 1750–1762.
6. Bhangale, K.B.; Mohanaprasad, K. A review on speech processing using machine learning paradigm. *Int. J. Speech Technol.* **2021**, *24*, 367–388.
7. Mohamed, A.; Lee, H.Y.; Borgholt, L.; Havtorn, J.D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.W.; Livescu, K.; Maaløe, L.; et al. Self-supervised speech representation learning: A review. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1179–1210.
8. Kemp, T.; Waibel, A. Unsupervised training of a speech recognizer: Recent experiments. In *Proc. EUROSPEECH*; 1999. Available online: https://isl.anthropomatik.kit.edu/pdf/Kemp1999.pdf (accessed on 22 January 2024).
9. Lamel, L.; Gauvain, J.L.; Adda, G. Lightly supervised and unsupervised acoustic model training. *Comput. Speech Lang.* **2002**, *16*, 115–129.
10. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
11. Chung, Y.A.; Hsu, W.N.; Tang, H.; Glass, J. An unsupervised autoregressive model for speech representation learning. *arXiv* **2019**, arXiv:1904.03240.
12. Barlow, H. Redundancy reduction revisited. *Netw. Comput. Neural Syst.* **2001**, *12*, 241.
13. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 12310–12320.
14. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876.
15. Liu, S.; Mallol-Ragolta, A.; Parada-Cabaleiro, E.; Qian, K.; Jing, X.; Kathan, A.; Hu, B.; Schuller, B.W. Audio self-supervised learning: A survey. *Patterns* **2022**, *3*, 100616.
16. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
17. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 1597–1607.
18. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.

19. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
20. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
21. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
22. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460.
23. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2020**, *60*, 101027.
24. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
25. Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv* **2018**, arXiv:1804.03209.
26. Emotional voice conversion: Theory, databases and ESD. *Speech Commun.* **2022**, *137*, 1–18.
27. American Rhetoric Online Speech Bank. World Leaders Address the U.S. Congress. 2011. Available online: https://www.americanrhetoric.com/speechbank.htm (accessed on 22 January 2024).
28. Chen, R.T.; Li, X.; Grosse, R.B.; Duvenaud, D.K. Isolating sources of disentanglement in variational autoencoders. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
29. Do, K.; Tran, T. Theory and evaluation metrics for learning disentangled representations. *arXiv* **2019**, arXiv:1908.09961.
30. Sepliarskaia, A.; Kiseleva, J.; de Rijke, M. How to not measure disentanglement. *arXiv* **2019**, arXiv:1910.05587.
31. Kumar, A.; Sattigeri, P.; Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv* **2017**, arXiv:1711.00848.
32. Ridgeway, K.; Mozer, M.C. Learning deep disentangled embeddings with the f-statistic loss. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.