



# SiamCTC: Learning Speech Representations through Monotonic Temporal Alignment

SooHwan Eom<sup>1</sup>, Mark Hasegawa-Johnson<sup>2</sup>, Chang D. Yoo<sup>\*1</sup>

<sup>1</sup>Korea Advanced Institute of Science and Technology, South Korea

<sup>2</sup>University of Illinois Urbana-Champaign, United States

{sean1105, cd.yoo}@kaist.ac.kr, jhasegaw@illinois.edu

## Abstract

Self-supervised speech representation learning has made significant progress through Siamese networks, which leverage different views of the same input. However, existing methods often require frame-wise alignment between these views, overlooking the broader linguistic context invariance across different speaking styles. We introduce SiamCTC, a framework that integrates Siamese networks with Connectionist Temporal Classification (CTC) to learn speech representations without strict frame-level correspondence. By employing CTC loss to establish flexible, monotonic alignments between differing temporal realizations of the same content, SiamCTC accommodates speed perturbations and other temporal augmentations. This design relaxes frame-wise constraints while preserving temporal coherence and enhancing robustness to speaking-rate variations in downstream tasks. Our experiments demonstrate that SiamCTC leads to more adaptable speech representations, particularly at diverse speaking rates.

**Index Terms:** self-supervised learning, speech representation learning

## 1. Introduction

Self-supervised learning (SSL) for speech representation learns from unlabeled data, using the input signal itself as the supervisory signal. By avoiding manual transcriptions, SSL enables deep neural network training on large-scale raw speech corpora, providing effective pre-training and robust representations for downstream tasks such as automatic speech recognition (ASR) [1] and speaker verification [2].

Speech signals inherently capture a variety of attributes, including speaker characteristics, environmental conditions, and linguistic content. Our primary focus is on extracting linguistic context, specifically, latent phonetic properties that characterize linguistic information. Previous SSL approaches have employed techniques such as in-utterance contrastive learning [3, 4, 5] or masked unit prediction [6]. However, they often neglect linguistic content invariance to varying speech conditions (e.g., speaking rate), limiting robustness when test conditions differ from those seen during training [7].

Motivated by the growing success of self-supervised learning in computer vision, we turn to Siamese-based frameworks [8, 9, 10, 11, 12], which learn from two distinct ‘views’ of the same data sample. These frameworks employ objectives such as contrastive learning, similarity maximization, or masked feature prediction to capture the shared information between views. A key insight of these methods is the existence of a latent structure

that remains invariant under different transformations, which Siamese networks aim to extract.

Several attempts have been made to apply Siamese networks to speech representation learning [13, 14, 15, 16]. However, they typically rely on frame-wise self-supervised labels, creating two main limitations: (1) restricting temporal augmentations like speed perturbation, as these can introduce frame misalignment, and (2) not fully leveraging linguistic content invariance across natural variations in speaking style and rate.

To overcome these limitations, we propose **SiamCTC**, a novel framework for speech representation learning that uses monotonic alignment between latent representations and self-supervised labels. Our approach integrates three key components: (1) a Siamese network to extract invariant speech representations, (2) a Connectionist Temporal Classification (CTC) [17] prediction head to learn monotonic alignments without imposing strict frame-wise constraints, and (3) in-utterance temporal contrastive learning to prevent representation collapse.

By explicitly modeling content invariance within utterances, SiamCTC captures linguistic representations more effectively, even with misaligned views introduced by speed perturbations. Our main contributions are summarized as follows:

- A flexible framework that eliminates frame-wise alignment constraints while maintaining temporal coherence through CTC-based monotonic alignment learning.
- An effective combination of Siamese networks and CTC for robust speech representation learning, improving generalization to varied speaking styles and speeds.
- Comprehensive empirical evaluation across multiple downstream tasks, improving the widely used SSL models such as HuBERT [6] and WavLM [18].

Through extensive experiments, we demonstrate that SiamCTC achieves improved results compared to conventional SSL frameworks, highlighting the benefits of speech invariance through monotonic alignment.

## 2. Related Works

### 2.1. Self-Supervised Speech Representation Learning

Self-supervised speech representation learning methods can be broadly classified into three categories based on their pretext task. Generative approaches (e.g., VQ-VAE [21], APC [22]) reconstruct or predict speech signals, potentially preserving non-linguistic attributes such as speaker characteristics. Contrastive methods such as CPC [3] and wav2vec series [4, 5] distinguish positive samples from negative ones. Although effective, they often assume strict temporal correspondence. Predictive approaches (e.g., HuBERT [6], wavLM [18], Data2vec [23]) rely on masked frame prediction objectives. While advancing the

\*Corresponding author.

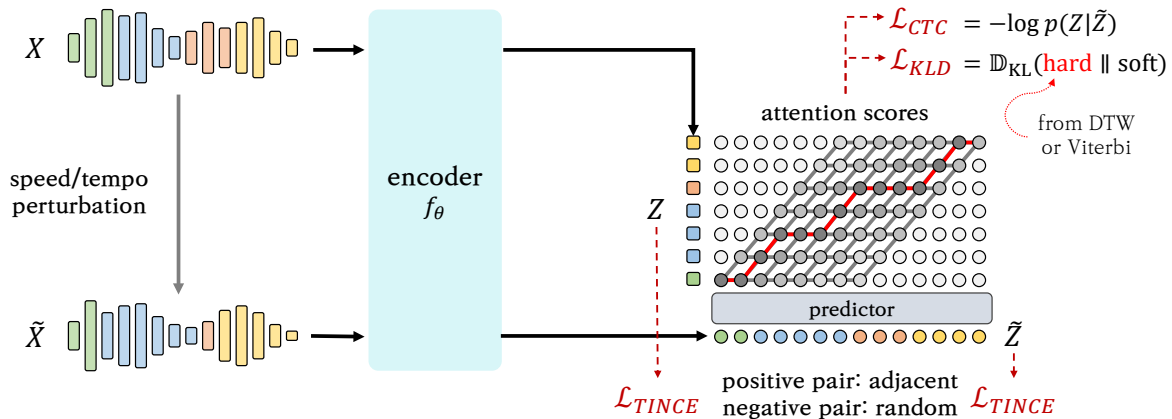


Figure 1: **An overview of SiamCTC framework.** Overview of our proposed SiamCTC framework. The model processes two views of the same input sequence  $X$  and  $\tilde{X}$ , where one undergoes speed/tempo perturbation. Both sequences pass through a shared encoder  $f_\theta$  to produce representation  $Z$  and  $\tilde{Z}$ . The framework optimizes three learning objectives: (1) CTC loss ( $\mathcal{L}_{CTC}$ ) for monotonic alignment learning, (2) KL divergence loss ( $\mathcal{L}_{KLD}$ ) between hard and soft alignments, and (3) Temporal InfoNCE losses ( $\mathcal{L}_{TINCE}$ ) that treat adjacent features as positive pairs while sampling negative pairs randomly within the sequence. Hard alignments are computed via the Viterbi algorithm [19] or DTW [20] and serve as a reference for the soft alignments learned by the model.

state-of-the-art, their frame-wise consistency reliance can cause speed perturbation sensitivity.

## 2.2. Siamese Networks for Self-Supervised Learning

Siamese-based frameworks have become prominent in SSL, particularly in computer vision [8, 10, 12, 11, 9]. These frameworks learn robust, invariant embeddings by comparing augmented input views via contrastive or distillation objectives.

In speech, several recent works [13, 14, 15, 16] have explored Siamese-based approaches. SPIN and R-SPIN [13, 14], inspired by SwaV [24], employ iterative clustering under speaker perturbation but rely on frame-level contrast. C-Siam [15] introduces dual encoders and a contrastive objective that assumes consistent temporal correspondence via predefined framewise alignment. DinoSR [16] combines masked pseudo-label prediction with online clustering and self-distillation, but also depends on frame-level alignment. While promising, their reliance on strict temporal matching can reduce robustness to varying speaking rates. LASER [25] partially addresses this limitation through soft-DTW [26] for direct sequence comparison. In contrast, our approach combines a Siamese framework with CTC-based alignment prediction, offering a more flexible alignment mechanism.

## 3. SiamCTC

This section introduces SiamCTC, a novel framework for learning speech representations invariant to temporal variations by combining a Siamese encoder with a Connectionist Temporal Classification (CTC) [17] alignment head. Auxiliary alignment consistency and temporal contrastive losses further refine alignment and prevent collapse. The overall goal is to train an encoder that captures linguistic representations whose sequential structure remains invariant to temporal perturbations. Figure 1 provides an overview of our approach.

### 3.1. Siamese Encoder with Temporal Perturbation

Given an input speech utterance  $X$ , we create two views: the original sequence  $X$  and its temporally perturbed version  $\tilde{X}$ . Both sequences are processed by a shared encoder  $f_\theta$  to obtain

their respective representations:

$$Z = f_\theta(X), \quad \tilde{Z} = f_\theta(\tilde{X}). \quad (1)$$

Our approach learns flexible inter-view alignments without strict, predefined frame-level correspondence. This design enables effective use of temporal augmentations, improving robustness to natural variations in speech rate.

### 3.2. Monotonic Alignment Learning

Connectionist Temporal Classification (CTC) [17] aims to maximize the log-likelihood between two sequences with unknown alignments, a common scenario in automatic speech recognition [27, 28]. Specifically, CTC sums the likelihoods of all valid alignment paths, leveraging a forward-backward procedure for efficient computation.

In our framework, the original representation  $Z$  serves as the target sequence or ‘pseudo-label’ for the CTC loss. The perturbed view representation  $\tilde{Z}$  is passed through a linear prediction layer  $\phi$  to produce predicted representations. Frame-wise logits for CTC are then derived from the attention scores between  $Z$  and  $\phi(\tilde{Z})$ . These logits, forming an attention score matrix, are used in the CTC loss:

$$\mathcal{L}_{CTC} = -\log \sum_{\pi \in \mathcal{B}^{-1}(Z)} p(\pi | \phi(\tilde{Z})), \quad (2)$$

where  $\mathcal{B}$  is the CTC many-to-one mapping and  $\pi$  represents valid alignment path. This allows CTC to learn alignments without requiring pre-defined discrete frame-wise labels.

For practical implementation, since CTC typically expects the input length to be longer than or equal to the target length, we *transpose* our logits when necessary to fit this requirement (e.g., under speed-up perturbation). Furthermore, to handle the mandatory `<blank>` symbol in CTC decoding, we reserve a fixed `<blank>` column in our attention logits.

### 3.3. Temporal Contrastive Loss

A trivial solution to the CTC objective is to collapse all frames into a single representation, thereby minimizing alignment cost without preserving fine-grained distinctions. To avoid this and ensure that distinct representations are learned for temporally

distinct speech segments, we incorporate a temporal InfoNCE (TINCE) loss:

$$\mathcal{L}_{\text{TINCE}}(h) = \mathbb{E}_{i, \mathcal{N}} \left[ -\log \frac{e^{s(\tilde{h}_i, h_{i+1})/\tau}}{e^{s(\tilde{h}_i, h_{i+1})/\tau} + \sum_{j \in \mathcal{N}} e^{s(\tilde{h}_i, h_j)/\tau}} \right], \quad (3)$$

where  $h = [h_1, h_2, \dots]$  is the input feature sequence,  $\tilde{h} = \psi(h)$  is a linear prediction of anchor frame  $h_i$ ,  $\tau$  is a temperature hyperparameter,  $s(a, b) = a^\top b / \|a\| \|b\|$  is the cosine similarity, and  $\mathcal{N}$  is a set of  $M$  negative samples drawn from positions at least  $K$  steps away from  $i$ . The core idea is that adjacent frames (positive pairs) should be more similar than randomly sampled non-adjacent frames (negative pairs). By pushing apart non-adjacent features, TINCE helps prevent the representation from collapsing into a single vector and preserves fine-grained adjacency between frames, which we refer to as ‘‘local context.’’

As a final loss, we compute the TINCE loss separately for both branches  $Z$  and  $\tilde{Z}$  and then take the average:

$$\mathcal{L}_{\text{TINCE}} = \frac{1}{2} (\mathcal{L}_{\text{TINCE}}(Z) + \mathcal{L}_{\text{TINCE}}(\tilde{Z})) \quad (4)$$

### 3.4. Alignment Consistency Loss

While CTC can learn monotonic alignments, its alignment quality may degrade without sufficient guidance. To address this, we utilize KL divergence loss  $\mathcal{L}_{\text{KLD}}$  between the hard alignment path ( $p_{\text{hard}}$ ), obtained via Viterbi [19] or DTW [20], and the attention scores ( $p_{\text{soft}}$ ):

$$\mathcal{L}_{\text{KLD}} = \mathbb{D}_{\text{KL}}(p_{\text{hard}} \| p_{\text{soft}}). \quad (5)$$

This loss constrains the learned attention to remain close to a meaningful alignment reference. In practice, we find that  $\mathcal{L}_{\text{KLD}}$  has a relatively minor effect, as the TINCE loss already discourages representation collapse and implicitly guides the soft alignment away from extreme solutions.

### 3.5. Overall Objective

We combine these losses into a single training objective:

$$\mathcal{L} = \mathcal{L}_{\text{CTC}} + \alpha \mathcal{L}_{\text{KLD}} + \beta \mathcal{L}_{\text{TINCE}}, \quad (6)$$

where  $\alpha$  and  $\beta$  are balancing coefficients. We set 1.0 for both coefficients. This composite objective simultaneously encourages robust monotonic alignment, alignment consistency, and non-collapsing temporal structure.

## 4. Experimental Details

### 4.1. Datasets

We use LibriSpeech [29] for both pre-training and downstream fine-tuning. LibriSpeech is a widely used corpus of approximately 1,000 hours of read English audiobooks derived from LibriVox, sampled at 16 kHz. It is partitioned into subsets designated as ‘‘clean’’ or ‘‘other,’’ reflecting differences in recording quality and speaker accents. In this work, we focus on the *train-clean-100* split to facilitate rapid experimentation and training efficiency under limited computational resources. For evaluation, we use the *test-clean* subset.

### 4.2. Model Training

We use two pre-trained self-supervised speech encoders as our base: HuBERT [6] and WavLM [18]. Both models were originally trained on 960 hours of LibriSpeech [29] corpus, and

Table 1: *Performance comparison on phoneme recognition (PR) and automatic speech recognition (ASR) downstream tasks. We report Phoneme Error Rate (PER%) and Word Error Rate (WER%), with lower values indicating better performance. The numbers in parentheses show the performance gap relative to the base model.*

Model	PR PER(%) ↓	ASR WER(%) ↓
HuBERT [6]	5.41	6.42
HuBERT+Spin [13]	4.39 (-1.02)	6.34 (-0.08)
HuBERT+LASER [25]	4.61 (-0.80)	6.18 (-0.24)
HuBERT+SiamCTC	4.32 (-1.09)	6.23 (-0.19)
WavLM [18]	4.84	6.21
WavLM+Spin [13]	4.18 (-0.66)	5.88 (-0.33)
WavLM+LASER [25]	4.28 (-0.56)	5.92 (-0.29)
WavLM+SiamCTC	3.96 (-0.88)	5.73 (-0.48)

are publicly available via S3PRL.<sup>12</sup> The model is then further trained for 5,000 updates, with a total batch size of 8. We employ AdamW optimizer [30] with a peak learning rate of 2e-5, linearly warmed up over the first 1,000 updates and then linearly decayed thereafter. A maximum gradient norm of 1.0 is applied to stabilize training.

The encoder architecture is based on HuBERT [6] and WavLM [18], and outputs 256-dimensional features. The alignment head uses a temperature of 2.0. For hard alignment, we rely on the Viterbi algorithm [19]. For temporal contrastive learning, we use  $-1$  as `<blank>` log probability, sample 20 negative examples from positions at least 5 frames away and apply a contrastive temperature of 0.2. For data augmentation, we apply speed perturbations using factors  $\{0.8, 0.9, 1.0, 1.1, 1.2\}$  and pitch perturbations using factors  $\{-2, -1, 0, 1, 2\}$ .

### 4.3. Baselines

We compare our approach with two widely used self-supervised speech representation encoders, HuBERT [6] and WavLM [18]. In addition, we compare ours with their SPIN [13] and LASER [25] variants, which are fine-tuned on the same base models as ours. All models are evaluated under identical conditions using the SUPERB benchmark for fair comparison.<sup>3</sup>

### 4.4. Evaluation Metrics

We report Phoneme Error Rate (PER%) and Word Error Rate (WER%) on the LibriSpeech *test-clean* for phoneme recognition and ASR tasks, respectively. In addition, we conduct ablation studies to examine the contribution of each loss component. We also test model robustness against speaking rate variations in speed-perturbed versions of the evaluation set.

## 5. Experiment Results

### 5.1. Main Results

Table 1 summarizes our results in phoneme recognition (PR) and automatic speech recognition (ASR) on LibriSpeech, com-

<sup>1</sup>[https://huggingface.co/s3prl/converted\\_ckpts/resolve/main/hubert\\_base\\_ls960.pt](https://huggingface.co/s3prl/converted_ckpts/resolve/main/hubert_base_ls960.pt)

<sup>2</sup>[https://huggingface.co/s3prl/converted\\_ckpts/resolve/main/wavlm\\_base.pt](https://huggingface.co/s3prl/converted_ckpts/resolve/main/wavlm_base.pt)

<sup>3</sup><https://github.com/s3prl/s3prl>

Table 2: Ablation study on different loss components for phoneme recognition. We evaluate performance in terms of Phoneme Error Rate (PER%) using combinations of CTC, KL divergence (KLD), and Temporal InfoNCE (TINCE) loss. Lower PER indicates better performance.

Loss			PR
CTC	KLD	TINCE	PER(%) ↓
✓	✗	✗	5.26
✓	✓	✗	5.16
✓	✗	✓	4.48
✓	✓	✓	4.32

paring HuBERT [6] and WavLM [18] baselines with variants enhanced by SPIN [13], LASER [25], or our proposed SiamCTC. We report phoneme error rate (PER%) and word error rate (WER%), with lower values indicating better performance. The values in parentheses show the improvement relative to the respective baseline.

For HuBERT-based systems, SiamCTC achieves the lowest PER (4.32%), surpassing both SPIN (4.39%) and LASER (4.61%). Although LASER slightly outperforms SiamCTC in WER (6.18% vs. 6.23%), our method consistently produces a strong overall result.

In the WavLM-based setting, SiamCTC again shows the largest improvement in PER (down from 4.84% to 3.96%) compared to SPIN (4.18%) and LASER (4.28%). Notably, SiamCTC also attains the best WER (5.73%) among the WavLM variants. These findings suggest that our explicit monotonic alignment objective provides robust benefits in capturing linguistic content, even under varying speaking rates.

## 5.2. Ablation Study

Table 2 presents an ablation study examining the effect of each loss component on phoneme recognition performance with HuBERT base model. We evaluate different configurations of CTC loss, alignment consistency loss (KLD), and Temporal InfoNCE (TINCE) loss using the Phoneme Error Rate (PER%), where lower values indicate better performance.

A baseline model employing only CTC achieves a PER of 5.26%. Adding the KLD term to guide alignment consistency reduces PER to 5.16%, indicating that aligning the learned attention with a hard alignment reference provides moderate improvements. In contrast, replacing KLD with TINCE yields a more substantial gain, lowering the PER to 4.48%. This result suggests that temporal contrastive learning effectively preserves local context and mitigates representation collapse, even without explicit hard alignment guidance.

Finally, combining CTC, KLD, and TINCE further increases performance to 4.32%, demonstrating that alignment consistency and temporal contrastive learning complement each other. In general, these findings confirm that both KLD and TINCE play a significant role in improving the robustness and accuracy of the learned speech representations.

## 5.3. Analysis

We evaluate the fine-tuned HuBERT and its SiamCTC variant on speed factors {0.8, 0.9, 1.0, 1.1, 1.2} to assess the robustness of our framework against different speaking rates. Figure 2 shows the Phoneme Error Rate (PER) for each factor, where lower values indicate better performance.

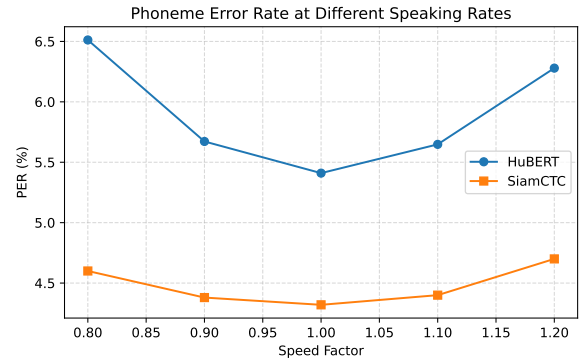


Figure 2: Phoneme Error Rate (PER) on LibriSpeech test-clean for different speed factors. The blue circle line represents HuBERT, and the orange square line represents SiamCTC.

Across all speed factors, SiamCTC consistently achieves a lower PER than HuBERT. At the original speed (1.0), HuBERT obtains a PER of 5.41%, whereas SiamCTC achieves 4.32%. When the audio is slowed down, HuBERT’s PER rises above 5.60%, peaking at 6.52% at 0.8 $\times$ . In contrast, SiamCTC exhibits a smaller performance drop, staying below 4.6%. For faster speech, HuBERT performance again degrades to 6.28% at 1.2 $\times$ , whereas SiamCTC remains relatively stable (4.7% at 1.2 $\times$ ). These results suggest that learning temporal perturbation invariance produces more robust representations across various speaking speeds and styles.

## 6. Conclusion

We proposed SiamCTC, a self-supervised learning framework that merges Siamese encoding with a Connectionist Temporal Classification (CTC) based alignment objective as its core mechanism to handle temporal perturbations in speech. This is complemented by temporal InfoNCE (TINCE) loss and an alignment consistency loss to prevent representation collapse and refine alignment quality. SiamCTC provides flexible alignment without relying on strict frame-level matching, thus capturing content invariance across varying speaking rates. Experiments on the LibriSpeech dataset show that SiamCTC demonstrates notable performance gains over established baselines, suggesting its effectiveness in preserving linguistic content even under speed perturbations. Our findings emphasize the benefits of learning flexible monotonic alignments in a self-supervised manner, paving the way for more robust and adaptive speech representation learning.

## 7. Limitation

While SiamCTC has shown promising performance, we observe that the downstream results can be sensitive to hyper-parameters such as augmentation strategies, negative pair sampling, and attention logit temperature. In particular, we find that using a lower temperature, which produces more peaked logits, is critical for alignment seeking. Future work would address these sensitivities with more robust or adaptive strategies.

While current SiamCTC is built on pre-trained models for efficiency, training SiamCTC from scratch may further reveal its full potential—especially if combined with broader augmentation techniques like masking. Exploring discrete speech units (e.g., via Vector Quantization [21]) for applications like spoken language modeling [31], beyond current continuous representations, can be another future direction.

## 8. Acknowledgements

This work was partly supported by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD) and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

## 9. References

- [1] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7694–7698.
- [2] C.-I. Lai, "Contrastive predictive coding based feature for automatic speaker verification," *arXiv preprint arXiv:1904.01575*, 2019.
- [3] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.
- [4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.
- [5] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.
- [6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021.
- [7] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Interspeech 2021*, 2021, pp. 721–725.
- [8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.
- [9] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9650–9660.
- [10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 271–21 284.
- [11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [12] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 750–15 758.
- [13] H.-J. Chang, A. H. Liu, and J. Glass, "Self-supervised Fine-tuning for Improved Content Representations by Speaker-invariant Clustering," in *Proc. Interspeech*, 2023.
- [14] H.-J. Chang and J. Glass, "R-spin: Efficient speaker and noise-invariant representation learning with acoustic pieces," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 642–662.
- [15] S. Khorram, J. Kim, A. Tripathi, H. Lu, Q. Zhang, and H. Sak, "Contrastive siamese network for semi-supervised speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7207–7211.
- [16] A. H. Liu, H.-J. Chang, M. Auli, W.-N. Hsu, and J. Glass, "Dinosr: Self-distillation and online clustering for self-supervised speech representation learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 58 346–58 362.
- [17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376.
- [18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," 2021.
- [19] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.
- [20] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–83, 1999.
- [21] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [22] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.
- [23] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 1298–1312.
- [24] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9912–9924.
- [25] A. Meghanani and T. Hain, "Laser: Learning by aligning self-supervised representations of speech for improving content-related tasks," in *Interspeech 2024*, 2024, pp. 2835–2839.
- [26] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 894–903.
- [27] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, p. II–1764–II–1772.
- [28] S. Eom, E. Yoon, H. S. Yoon, C. Kim, M. Hasegawa-Johnson, and C. D. Yoo, "Adamer-ctc: Connectionist temporal classification with adaptive maximum entropy regularization for automatic speech recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 707–12 711.
- [29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.
- [31] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023, p. 1–5.