Distilling Temporal Relation for Speech Self-Supervised Learning Model Compression

Qingzheng Wang
qingzhew@andrew.cmu.edu

Shun Li
shunl@andrew.cmu.edu

Toris Ye
huzhengy@andrew.cmu.edu

Fanxing Bu
fanxingb@andrew.cmu.edu

Abstract

Transformer-based speech self-supervised learning (SSL) models have achieved impressive results across various speech tasks. However, their large parameter sizes and high computational requirements limit their usability in environments with constrained resources. To compress the model size, traditional methods often focus on directly matching frame-level representations between the original and compressed models. In contrast, the Speech Temporal Relation (STaR) method [1] introduces a novel approach by distilling temporal relations between frames. This approach achieved state-of-the-art results on the SUPERB benchmark among models with up to 27 million parameters. In this project, we reimplement this distillation method, utilizing speech temporal relation to compress the pre-trained speech self-supervised model HuBERT. The code and the distilled model are in https://github.com/Qingzheng-Wang/STaRHuBERT.

1 Introduction

Transformer-based speech self-supervised learning models have demonstrated remarkable performance across various tasks. However, their large parameter size and high computational demands make them impractical for deployment in resource-constrained environments. To address this issue, we aim to reproduce the Speech Temporal Relation (STaR) distillation method proposed by Jang et al. [1], which compresses the pretrained speech self-supervised model HuBERT [2] by distilling *speech temporal relation*.

In this study, we model temporal relations between speech frames on the spectrogram, as these relationships play a vital role in capturing essential acoustic patterns. The model's output will focus on distilled representations that retain the most critical temporal relationships, ensuring that essential information is preserved even in a simplified form.

For this project, the HuBERT BASE model serves as the teacher model, setting a standard for high-quality speech representations. The STaR framework allows the student model to learn from the teacher model by replicating temporal relations between speech frames. This is an effective strategy for lightweight models, as it transfers key knowledge without needing to match the complexity of the teacher model.

The motivation for this research lies in making advanced speech processing technologies more accessible and affordable. By concentrating on model compression using the STaR framework, we aim to design a model that performs well on tasks like automatic speech recognition (ASR) and speaker identification (SID), but with far fewer computational demands. This will enable broader applications of speech SSL models across devices and systems with limited resources, thus expanding the practical use of cutting-edge speech technology in the real world.

CMU 11785 Introduction to Deep Learning - Group Project Final Report.

2 Related Work

The field of speech self-supervised learning (SSL) has seen significant advancements in recent years, with various models demonstrating impressive performance across a range of speech-related tasks, such as HuBERT [2], wav2vec 2.0 [3], and WavLM [4]. One of the pioneering works in this area is HuBERT, which utilizes a masked prediction approach to learn robust speech representations from unlabeled audio data. The model's ability to capture contextual information from speech frames has set a new standard for performance in tasks such as automatic speech recognition (ASR) and speaker identification (SID) . However, despite its success, HuBERT's large model size and high computational requirements pose challenges for deployment in resource-constrained environments.

While techniques like pruning [5] and quantization [6] offer some compression, knowledge distillation is a more efficient approach as it avoids the excessive computational overhead, which trains a smaller student model to mimic a larger teacher model. Task-specific knowledge distillation has been explored for models like automatic speech recognition (ASR) [7], but it limits generalization to other tasks. Task-agnostic methods such as DistilHuBERT [8] and FitHuBERT [9] reduce HuBERT's size while maintaining competitive performance across various tasks, demonstrating that smaller student models can effectively mimic larger teacher models while balancing efficiency and accuracy highlights the potential of distillation techniques in enhancing speech SSL models, where the former suggests shallow and wide student model design, and the latter suggests deep and narrow one. LightHuBERT [10] applies architecture search, though with high computational costs. Additionally, DPHuBERT [11] combines knowledge distillation with structured pruning to compress models. This approach allows the student architecture to be learned during the distillation process, resulting in a model that outperforms traditional distillation methods across various tasks while requiring less training time and data. ARMHuBERT [12] is another noteworthy advancement in the realm of speech representation learning. This model reuses attention maps and applied both masked and unmasked speech frames for masking distillation, further employing adaptive representation learning to enhance efficiency and adjusting model parameters dynamically bd on input data.

Despite the promising results of previous studies, two key limitations remain. First, many approaches fail to consider the limited representational capacity of the student model, instead directly matching the teacher's complex representations for each speech frame by adding linear heads [8, 9, 12]. This often over-constrains the lightweight student, highlighting the need for distillation objectives better suited to the student. Additionally, some studies discard these linear heads after distillation, missing a chance to effectively convey the teacher's knowledge [8, 9, 12, 11]. Second, while pruning reduces model parameters through sparsity, it does not lower computational costs, sometimes resulting in higher overhead compared to methods with predefined model architectures. To overcome these challenges, we focus on distilling speech temporal relations, which simplifies the student model's task without directly mimicking complex teacher representations. No additional parameters are required during distillation, allowing for the creation of a more compact and computationally efficient student model.

3 Method

Speech SSL models predict masked frames as clusters during pre-training, representing speech frames as specific acoustic units. However, directly learning the teacher's detailed frame-level representations can overly constrain the student model, making it difficult to generalize effectively. To address this, the baseline method [1] introduces two flexible distillation strategies: distilling knowledge via the average attention map and capturing temporal relations within speech frames.

3.1 Model

The baseline [1] utilizes HuBERT BASE [2] as the teacher model, which is composed of a feature extractor based on convolutional neural networks (CNN) and several Transformer [13] layers. The student model retains the same overall architecture but reduces the width of the attention layers and feed-forward networks, allowing for more efficient computation and compression. The teacher and student model details are shown in Table 1

Table 1: Model Architectures: HuBERT BASE (Teacher) vs. STaR HuBERT (Student)

Component	Attribute	HuBERT BASE	STaR HuBERT
CNN Encoder	Strides	5, 2, 2, 2, 2, 2, 2	5, 2, 2, 2, 2, 2, 2
	Kernel Width	10, 3, 3, 3, 3, 2, 2	10, 3, 3, 3, 3, 2, 2
	Channel	512	512
Transformer	Layers	12	12
	Embedding Dim.	768	432
	Inner FFN Dim.	3072	976
	Layerdrop Prob.	0.05	0.05
	Attention Heads	8	8
Projection	Dim.	256	256
Num. of Params		95M	22.3M

3.2 Average Attention Map Distillation

Attention maps capture the temporal relationships between key and query vectors in the Transformer [13] layers. Hence, distilling the attention maps provides a natural way to transfer temporal knowledge from the teacher model [1]. The attention map for each head h is given by

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}_h^\top \mathbf{K}_h}{\sqrt{d_h}}\right), \tag{1}$$

where $\mathbf{Q}_h, \mathbf{K}_h \in \mathbb{R}^{d_h \times N}$ are the query and key matrices, d_h is the dimension of the key, and N is the sequence length. The baseline distills knowledge by averaging attention maps across all heads and computing the Kullback-Leibler (KL) divergence between the teacher T's and student S's attention maps. The loss is defined as:

$$\mathcal{L}_{\text{avg-attn}} = \sum_{l=1}^{L}\sum_{t=1}^{N} D_{\text{KL}}\left(\frac{1}{H^{\text{T}}}\sum_{h=1}^{H^{\text{T}}}\mathbf{A}_{h,t}^{l,\text{T}}\middle\|\frac{1}{H^{\text{S}}}\sum_{h=1}^{H^{\text{S}}}\mathbf{A}_{h,t}^{l,\text{S}}\right), \tag{2}$$

where H and L are the number of attention heads and Transformer layers, respectively.

3.3 Temporal Gram Matrix Distillation

Attention maps, though informative, are not directly used during inference, which limits their effectiveness as a distillation signal. In order to provide stronger hints, the baseline introduces the **Temporal Gram Matrix (TGM)**, which captures pairwise temporal relations between speech frames. Given the representation $\mathbf{F} \in \mathbb{R}^{d\times N}$, the TGM is defined as $G_{i,j} = \sum_{k=1}^{d} F_{k,i}F_{k,j}$, where $F_{\cdot,i}$ represents the i-th frame. The TGM distillation uses two loss functions: layer-wise and intra-layer. The **layer-wise loss** measures the mean squared error (MSE) between the TGMs of the teacher and student models across all layers:

$$\mathcal{L}_{\text{layer-wise}} = \sum_{l=1}^{L} \|\mathbf{G}^{l,\text{T}} - \mathbf{G}^{l,\text{S}}\|_2^2, \tag{3}$$

while the **intra-layer loss** captures temporal progression within each layer:

$$\mathcal{L}_{\text{intra-layer}} = \sum_{l=1}^{L} \|\check{\mathbf{G}}^{l,\text{T}} - \check{\mathbf{G}}^{l,\text{S}}\|_2^2. \tag{4}$$

4 Results

4.1 Experimental Setup

Training The student model was trained on the Librispeech's [14] `train-clean-100` dataset for 200 epochs. We selected HuBERT BASE [2] comprising 12 Transformer [13] layers as the teacher model.

The training process used the AdamW optimizer, with an initial learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a batch size of 2. We applied a warmup cosine scheduler with a warmup proportion of 0.05.

Student Model The student model shares the same architecture as the teacher model, HuBERT BASE [2], comprising a CNN feature extractor, with configurations consistent with those of the teacher model, followed by 12 Transformer layers [13]. However, the student model has a reduced attention layer width and feed-forward network (FFN) width, set to 432 and 976, respectively. The student model details are illustrated in Table 1.

Loss Selection The distillation loss combines both layer-wise and intra-layer losses, with the total training loss defined as their sum. As noted in the original STaR paper [1], this combination can achieve optimal performance. Due to limited time and computational resources, we did not implement average attention map distillation loss or conduct ablation studies on the distillation losses in our experiments.

Evaluation To assess the student model's task-agnostic characteristics, we evaluated it using the SUPERB benchmark [15], which encompasses over ten speech-related tasks, including Phoneme Recognition (PR), Automatic Speech Recognition (ASR), Query-by-Example (QbE), Speaker Identification (SID), Automatic Speaker Verification (ASV), Speaker Diarization (SD), Emotion Recognition (ER), Spoken Intent Classification (IC), Spoken Slot Filling (SF), and Speech Enhancement (SE).[1] These tasks evaluate the distilled model's performance across content, speaker, paralinguistics, semantics, and generation dimensions.

Table 2: Evaluation Results on SUPERB Benchmark. Metrics include the number of parameters, number of Multiply-Accumulates (MACs), phoneme error rate (%), word error rate (WER, %) without language model, maximum term weighted value (MTWV), accuracy (Acc, %), equal error rate (EER, %), diarization error rate (DER, %), F1 score (F1, %), concept error rate (CER, %), scale-invariant signal-to-distortion ratio (SI-SDR, dB), short-time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ).

Models	Computation		Content			Speaker			Semantics			Paral.	Generization		
	Params Millions ↓	MACs Giga ↓	PR PER ↓	ASR WER ↓	QbE MTWV ↑	SID Acc ↑	ASV EER ↓	SD DER ↓	IC Acc ↑	SF F1 ↑	SF CER ↓	ER Acc ↑	SE SI-SDR ↑	SE STOI ↑	SE PESQ ↑
HuBERT BASE	94.70	1669	5.41	6.42	0.0736	81.42	5.11	5.88	98.34	88.53	25.20	64.92	9.36	93.90	2.58
STaRHuBERT (original)	22.31	463.5	9.17	10.92	0.0626	72.26	5.66	5.97	97.21	84.89	30.77	61.12	-	-	-
STaRHuBERT (ours)	22.31	463.5	9.81	9.97	0.0584	69.71	6.44	4.54	96.73	86.75	27.10	63.13	8.040	85.84	2.05

4.2 Detailed Benchmark Results

The SUPERB benchmark results are shown in Table 2. In this section, we highlight key discrepancies between the reproduced STaRHuBERT model ("ours") and the results reported in the original STaRHuBERT paper. Rather than providing a metric-by-metric or benchmark-by-benchmark breakdown, we focus on notable outliers and areas of either underperformance or overperformance relative to the original results.

Notable Overperformance: One area where our reproduced model exceeds the original STaRHuBERT is ASR. Specifically, the original model reports a WER of 10.92%, whereas our replication achieves a WER of 9.97%. This improvement suggests that our training or fine-tuning procedure may have yielded a slightly more robust phoneme and word-level representation than the original. Beyond ASR, in the semantic-oriented task SF, our model also shows gains. Specifically, while the original STaRHuBERT achieved an F1 score of 84.89% in SF, our model attains 86.75%. Similarly, the CER of SF drops from 30.77% in the original to 20.10% in our reproduction. These enhancements indicate that our approach better captures semantic details and downstream linguistic nuances.

Underperformance Cases: Despite these improvements, there are areas where our model falls short of the original. In particular, SID accuracy decreases from 72.26% to 69.71%, suggesting that our reproduced model has a harder time discerning subtle speaker characteristics. Additionally, in ASV, our model records a slightly higher EER than the original (6.44% vs. 5.66%), indicating that we struggle more with tasks requiring robust speaker-dependent embeddings. For tasks like QbE and semantics-based tasks IC, our model shows marginal reductions in performance. Although these

[1] https://github.com/s3prl/s3prl

Table 3: Loss Selection [1]

Methods	Overall ↑	General. ↑	Ranking ↓
LibriSpeech 100h distillation			
FitHuBERT [9]	74.5	695	-
MaskHuBERT	76.3	789	-
$\mathcal{L}_{\textbf{last-attn}}$	75.7	750	-
$\mathcal{L}_{\text{avg-attn}}$	76.5	766	4.4
$\mathcal{L}_{\text{layer-wise}}$	77.8	829	2.7
$\mathcal{L}_{\text{intra-layer}}$	77.7	820	3.2
$\mathcal{L}_{\text{layer-wise}} + \mathcal{L}_{\text{intra-layer}}$	77.8	831	2.2
$\mathcal{L}_{\text{layer-wise}} + \mathcal{L}_{\text{intra-layer}} + \mathcal{L}_{\text{avg-attn}}$	77.7	827	2.6
LibriSpeech 960h distillation			
$\mathcal{L}_{\text{layer-wise}}$	79.4	887	2.1
$\mathcal{L}_{\text{layer-wise}} + \mathcal{L}_{\text{intra-layer}}$	79.5	887	1.7
$\mathcal{L}_{\text{layer-wise}} + \mathcal{L}_{\text{intra-layer}} + \mathcal{L}_{\text{avg-attn}}$	78.8	871	2.3

differences are relatively small, they point to potential inconsistencies in how the model handles the alignment of acoustic features with semantic or lexical cues.

Overall Discrepancies: Our reproduced STaRHuBERT model's results closely mirror the findings from the original paper, reinforcing the validity of our implementation. Most metrics display only minor variations, suggesting that any discrepancies are likely attributable to nuances in training conditions, data preprocessing, or random initialization. In general, these deviations are relatively small and fall within a plausible range of experimental variance. While we note slight underperformance in certain speaker-related tasks (e.g., SID, ASV) and small performance boosts in some semantic-oriented metrics (e.g., ASR, SF), these differences are not substantial. Overall, our reproduction demonstrates a strong alignment with the original STaRHuBERT results, underscoring the reproducibility and robustness of the original work.

5 Discussion and Future Work

Building on this closely matched replication, our current STaRHuBERT model—trained for 200 epochs—reaffirms the reliability of the original approach and invites further exploration. Minor performance differences, though minimal, suggest opportunities for refining loss functions, adopting more targeted parameter reduction methods, and broadening the scope of tasks and datasets. In doing so, we can deepen our understanding of how to optimize lightweight speech models, push the limits of speech representation learning, and ultimately improve generalization and efficiency in both familiar and novel evaluation scenarios.

5.1 Loss Design and Optimization

The design of the distillation loss function plays a critical role in determining a student model's performance. As shown in Table 3, the original STaR method combines layer-wise and intra-layer temporal Gram Matrix (TGM) losses but demonstrates limited improvements with average attention map losses. This indicates the need for further refinements in the loss design. In particular, there are several promising avenues to enhance distillation performance:

Refined Weighting of Loss Terms: Rather than equally weighting each component, adaptive weighting schemes could be explored. For instance, emphasis might initially be placed on temporal structure, gradually shifting to layer-wise alignment as the model matures. Dynamic weighting or reinforcement learning-based methods could yield more balanced and effective distillation signals.

Task-Specific Loss Modulation: Downstream tasks vary in their reliance on different representation aspects. A multi-phase distillation strategy could tailor losses to early and later layers based on targeted downstream objectives, ensuring better generalization across speaker, content, semantic, and paralinguistic tasks.

Incorporation of Additional Constraints or Regularizations: Beyond TGMs and attention maps, new loss components—such as smoothness constraints, spectral norm regularization, or contrastive objectives—could stabilize and enrich the distilled model. Experimentation with such constraints may yield more robust and informative embeddings, even under tight parameter budgets.

5.2 Advanced Parameter Reduction Techniques

The student model shares the same architecture as the teacher model, HuBERT BASE, and simply reduced attention layer width and feed-forward network width. We consider integrating structured pruning to remove less critical parameters and explore lightweight architecture search methods for targeted compression. For instance, future work could explore developing and evaluating a joint distillation and structured pruning approach, such as the one proposed in [11], which selectively prunes less important parameters instead of uniformly reducing the number of parameters across all layers.

5.3 Improving Training Stability and Efficiency

While extending the number of training epochs from 20 to 200 helped our model better align with the original performance, additional steps may improve both convergence and stability:

Curriculum Distillation: Introducing a staged training process, where simpler alignment tasks precede more complex relational objectives, may stabilize learning and help the student incrementally assimilate the teacher's knowledge.

Hyperparameter Tuning and Data Augmentation: Systematic exploration of hyperparameters—such as learning rates, batch sizes, and warmup steps—and the integration of data augmentation techniques (e.g., SpecAugment, noise injection) could optimize convergence and enhance the final quality of the student representations.

6 Conclusion

Our reimplementation of the STaR distillation method closely aligns with the original reported performance, confirming its stability and adaptability. Although some variations emerged in speaker-related and semantic tasks, these discrepancies remained small and manageable. The results demonstrate that temporal relation-based distillation not only can be replicated reliably but also retains competitive performance under reduced computational footprints. Given the model's versatility, its potential extends beyond the tasks evaluated here. Future research may explore novel loss functions, more nuanced parameter reduction techniques, and broader application domains, positioning STaR as a fertile ground for continued innovation in lightweight speech SSL models.

7 Work Division

Table 4 outlines the division of work within our team.

Table 4: Work Division

Name	Tasks
Qingzheng Wang	Model training; report writing for the method and results sections.
Shun Li	Literature review and peer review answering; report writing for the introduction, related work, discussion, and conclusion sections. Slides creation.
Toris Ye	SUPERB benchmarking; report writing for the results and discussion sections; video editing.
Fanxing Bu	SUPERB benchmarking.

References

[1] Kangwook Jang, Sungnyun Kim, and Hoirin Kim. Star: Distilling speech temporal relation for lightweight speech self-supervised learning models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10721–10725. IEEE, 2024.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

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

[5] Cheng-I Jeff Lai, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass. Parp: Prune, adjust and re-prune for self-supervised speech recognition. *Advances in Neural Information Processing Systems*, 34:21256–21272, 2021.

[6] Naigang Wang, Chi-Chun Charlie Liu, Swagath Venkataramani, Sanchari Sen, Chia-Yu Chen, Kaoutar El Maghraoui, Vijayalakshmi Viji Srinivasan, and Leland Chang. Deep compression of pre-trained transformer models. *Advances in Neural Information Processing Systems*, 35:14140–14154, 2022.

[7] Euntae Choi, Youshin Lim, Byeong-Yeol Kim, Hyung Yong Kim, Hanbin Lee, Yunkyu Lim, Seung Woo Yu, and Sungjoo Yoo. Masked token similarity transfer for compressing transformer-based asr models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[8] Heng-Jui Chang, Shu-Wen Yang, and Hung-yi Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit BERT. *CoRR*, abs/2110.01900, 2021.

[9] Yeonghyeon Lee, Kangwook Jang, Jahyun Goo, Youngmoon Jung, and Hoirin Kim. FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Learning. *arXiv e-prints*, page arXiv:2207.00555, July 2022.

[10] Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, and Haizhou Li. Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert. *arXiv preprint arXiv:2203.15610*, 2022.

[11] Yifan Peng, Yui Sudo, Shakeel Muhammad, and Shinji Watanabe. Dphubert: Joint distillation and pruning of self-supervised speech models. *arXiv preprint arXiv:2305.17651*, 2023.

[12] Kangwook Jang, Sungnyun Kim, Se-Young Yun, and Hoirin Kim. Recycle-and-distill: Universal compression strategy for transformer-based speech ssl models with attention map reusing and masking distillation. *arXiv preprint arXiv:2305.11685*, 2023.

[13] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[14] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[15] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.