

USAD: Universal Speech and Audio Representation via Distillation

Heng-Jui Chang, Saurabhchand Bhati, James Glass, Alexander H. Liu

MIT CSAIL

Cambridge, MA, USA

hengjui@mit.edu

Abstract—Self-supervised learning (SSL) has revolutionized audio representations, yet models often remain domain-specific, focusing on either speech or non-speech tasks. In this work, we present Universal Speech and Audio Distillation (USAD), a unified approach to audio representation learning that integrates diverse audio types—speech, sound, and music—into a single model. USAD employs efficient layer-to-layer distillation from domain-specific SSL models to train a student on a comprehensive audio dataset. USAD offers competitive performance across various benchmarks and datasets, including frame and instance-level speech processing tasks, audio tagging, and sound classification, achieving near state-of-the-art results with a single encoder on SUPERB and HEAR benchmarks.¹

Index Terms—speech and audio representation learning, self-supervised learning, knowledge distillation

I. INTRODUCTION

In recent years, self-supervised learning (SSL) methods—learning frameworks that utilize unlabeled data without explicit supervision—have significantly advanced representation learning for audio processing. Speech SSL models like wav2vec 2.0 [1], HuBERT [2], and WavLM [3] have become the foundation of many applications like automatic speech recognition (ASR), speaker identification, and phoneme classification. In parallel, SSL approaches developed for audio event classification and music understanding, such as SSAST [4], BEATs [5], and MERT [6], have successfully been shown to be effective in non-speech tasks.

In practice, the use of audio representation has extended beyond simple downstream tasks. These representations have become the proxy for modern multimodal systems to understand the world and to interact with humans through audio. For example, many audio-enabled large language models (LLM) are trained to understand audio via representations [7]–[10], where researchers have found the quality of the representation to be critical to audio LLM performance [11]. Furthermore, audio representation learning has been applied to new encoding techniques that empower complex systems to understand and generate speech [12]–[15].

Despite these advancements, open challenges remain in audio representation learning. In this work, we are particularly interested in the problem of learning representations of general audio—the union of speech, sound, and music. While these

different types of audio are essentially signals with various patterns, they have been treated as separate domains in the audio processing literature (see Section II), with specialized models tailored specifically to each. On the road to artificial general intelligence, audio is an inevitable barrier that needs to be solved regardless of the context. Having disjoint representations for different types of audio increases the complexity of solutions towards the ultimate goal [9], [16], [17].

To address this problem, we propose Universal Speech and Audio Distillation (USAD), a unified model for speech and audio representation. USAD is trained via layer-to-layer distillation [18] from pre-trained teacher models in speech and audio, using a mixture of multi-domain audio data. Our results show that USAD is competitive across various benchmarks in speech, sound, and music processing, often approaching state-of-the-art performance despite being a single, general-purpose audio encoder. Additionally, we conduct ablation studies on key method elements, including teacher model choice, student model size, distillation data, and computational budget. To summarize, our contribution can be highlighted as follows:

- We propose USAD, a unified audio representation model that jointly learns from speech, sound, and music.
- USAD distills domain-specific experts to an encoder.
- Results show USAD learns a unified embedding space across domains with a reasonable computational budget.
- Notably, USAD encodes audio into representations that closely match the performance of oracle teacher models and state-of-the-art expert models in diverse audio tasks, regardless of input domain.

II. RELATED WORKS

Speech Representation Learning. To utilize large-scale unlabeled speech datasets, well-known SSL models like wav2vec 2.0 [1] and HuBERT [2] rely on predicting pseudo labels (derived from contrastive learning or offline clustering) of partially masked speech to train the representation encoder. These methods have shown that learning through pseudo labels significantly improves downstream performance on speech tasks compared to a fully supervised method without SSL. WavLM [3] further improves the technique by introducing a denoising task to SSL training, thus producing richer representations useful across a broader range of speech applications such as ASR, keyword spotting, and speaker recognition.

¹Models: <https://huggingface.co/MIT-SLS/USAD-Base>

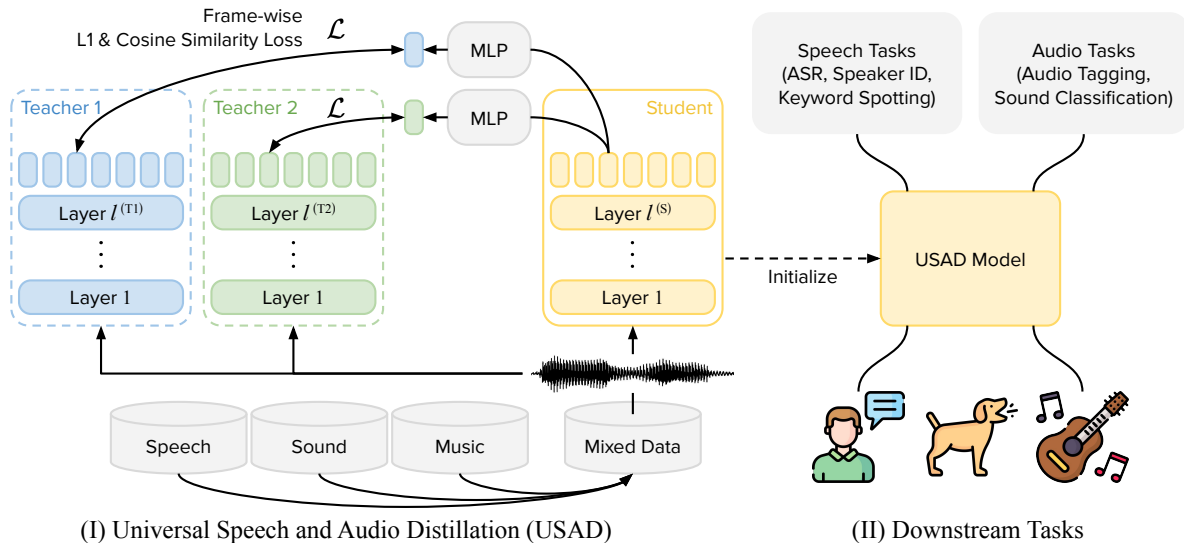


Fig. 1: The proposed USAD. In stage (I), a student model distills knowledge from two domain-specific teachers through sparse layer-to-layer distillation with mixed audio data. In stage (II), the USAD model can be used in various speech and audio tasks.

Sound and Music Representation Learning. Similar to prior works in speech, sound/music SSL frameworks are designed to learn semantically meaningful representations from large-scale, unlabeled datasets. Reminiscent of speech SSL methods, prior works in this area found the combination of pseudo labeling and masked prediction effective [4], [19]–[21]. Not surprisingly, the resulting representation enhances audio processing tasks like sound and music classification. The inspiration for this work stemmed from the similarities shared by speech and non-speech representation learning frameworks.

Self-supervised Learning with Distillation. Distillation has been studied extensively in audio SSL and can be divided into self-distillation (SD) and knowledge distillation (KD), where the difference lies in the teacher model. For SD [22]–[29], teacher and student models have the same size and are randomly initialized. SD relies on providing different views to the two models (e.g., masked and unmasked) and designing an update policy for the teacher model (e.g., moving average). Meanwhile, KD relies on pre-trained teacher models as the source of knowledge. Methods like DistilHuBERT [30], DPHuBERT [31], DASS [32], and CoLLD [18] utilize this technique to compress large models. DistilHuBERT focuses explicitly on reducing model complexity by predicting hidden representations from a teacher model layer by layer, achieving comparable accuracy at significantly lower computational cost. Multi-teacher ensemble distillation has also been explored previously, but is limited to speech [33]. Another work uses decoupled KD and task arithmetic to learn a unified compressed speech and music encoder [34].

While USAD falls in the category of KD, USAD differs from existing works: (1) USAD simultaneously distills from two teacher models from different domains; (2) USAD aims at generalization rather than compression of the student model; (3) existing approaches remain confined to speech tasks, without considering cross-domain audio representation learning.

III. METHODS

A. Unified Speech and Audio Distillation

This paper aims to develop a unified encoder to extract representations across speech, sound, and music domains. Motivated by the success of single-domain SSL and knowledge distillation, we propose Unified Speech and Audio Distillation (USAD) to distill domain-specific SSL pre-trained models into a shared encoder. The framework is depicted in Fig. 1. First, we select two pre-trained SSL models (T1 and T2) that were trained on different domains. A randomly initialized student model and multi-layer perceptron (MLP) prediction heads are used. During training, all models receive the same mixed dataset inputs, comprising speech, sound, and music. Specific student layer ($l^{(S)}$) outputs are fed into two MLP heads to predict corresponding hidden layer ($l^{(T1)}$ and $l^{(T2)}$) representations from the teacher models. As all models use transformer encoders [35], the student layers predict the teacher's feed-forward network (FFN) features, which is effective as demonstrated previously [18], [22]. We use the frame-wise reconstruction loss from DistilHuBERT [30], detailed further in the following section.

B. Sparse Layer-to-layer Distillation

Unlike most distillation works, USAD adopts two teacher models to learn general representations across multiple audio domains. Designing an efficient distillation method to train the student model is thereby a key factor to make USAD useful in practice. We propose sparse layer-to-layer (L2L) distillation that improves the efficiency of the existing (dense) L2L distillation methods with a simplified objective function.

L2L distillation, proposed by CoLLD [18], is originally designed to train the student network to predict the output of *each and every* layer of the teacher model. While L2L distillation demonstrates strong results, the exhaustive distillation scheme inevitably leads to increased computational cost. To mitigate

this issue, we introduce sparse L2L distillation that distills only K layers, exploiting the fact that similarity between consecutive transformer layers is expected to be high [36]. For a 12-layer student with $K = 4$, distillation occurs at $l^{(S)} \in \{3, 6, 9, 12\}$, reducing loss compute by approximately 75%. Formally, given student and teacher model depths $L^{(S)}$, $L^{(T1)}$, and $L^{(T2)}$, the k -th student layer $l_k^{(S)} = \lfloor \frac{kL^{(S)}}{K} \rfloor$ is matched to teacher layers $l_k^{(T1)} = \lfloor \frac{kL^{(T1)}}{K} \rfloor$ and $l_k^{(T2)} = \lfloor \frac{kL^{(T2)}}{K} \rfloor$, where $k = 1, \dots, K$. The student’s output at $l_k^{(S)}$ passes through two separate MLP heads, which predict the corresponding FFN features of each teacher.

In addition to sparsity introduced in the learning target, we improve the training objective of L2L in the same spirit. The original design of CoLLD utilizes contrastive loss that requires both positive and negative samples on a per-frame basis, making loss computation expensive. To reduce the cost, we eliminate the need for negative samples by adopting L1-cosine similarity from DistilHuBERT [30]. For predicting teacher T1’s FFN features from student layer $l_k^{(S)}$, we define the student’s MLP output: $\tilde{z}_{k,t}^{(T1)} = f_k^{(T1)}(\mathbf{h}_{k,t}^{(S)}) \in \mathbb{R}^D$. The training objective minimizes L1 distance and maximizes cosine similarity between predictions and target features:

$$\mathcal{L}_{k,t}^{(T1)} = \frac{1}{D} \left\| \tilde{z}_{k,t}^{(T1)} - z_{k,t}^{(T1)} \right\|_1 - \log \sigma \left[\cos \left(\tilde{z}_{k,t}^{(T1)}, z_{k,t}^{(T1)} \right) \right],$$

where σ is the sigmoid function. We aggregate losses across layers and frames:

$$\mathcal{L} = \frac{1}{2KT} \sum_{k=1}^K \sum_{t=1}^T \mathcal{L}_{k,t}^{(T1)} + \mathcal{L}_{k,t}^{(T2)},$$

where T is the number of audio frames. This setup significantly improves computational efficiency compared to CoLLD, preserving comparable performance.

C. Audio Feature Extraction: Frame vs. Patch

A critical distinction between speech and audio SSL is the feature extraction strategy. Phonetic durations are typically on the order of 100 ms, necessitating a temporal resolution of at least 10Hz, and making speech processing tasks require fine-grained features. Thus, speech models utilize frame-wise features—audio segments downsampled along the time axis (e.g., 50Hz in models like WavLM [3] and Whisper [37]).

In contrast, general audio signals exhibit complex temporal and frequency patterns, primarily targeting instance-wise classification tasks. Therefore, audio SSL approaches often use lower temporal resolution encoders or patch-based features [4], [5], [25], [26], [28]. Typically, these methods begin with a 128-dimensional Mel-spectrogram with a stride of 10 ms, cropped into non-overlapping 16×16 patches, and flattened into a sequence of 256-dimensional vectors. Although the patch sequence has a 50Hz framerate, the actual temporal resolution is 6.25Hz, smoothing out local details and blurring speech-specific information.

While patch-based features preserve frequency detail, this approach degrades speech task performance (see Section IV). Moreover, despite having the same framerate, frame- and

TABLE I: Preprocessed dataset information. The dataset sizes might differ from the original ones due to preprocessing.

Dataset	Clips	Duration (hours)	Proportion of Mix126k-B
Speech (English)			
LibriVox [38]–[40]	13,974,853	56,167.4	27.5%
VoxPopuli [41]	3,051,826	24,084.4	6.0%
GigaSpeech [42]	4,631,558	6,973.2	9.1%
Common Voice 17 [43]	1,117,555	1,765.3	2.2%
Fisher [44]	1,168,612	1,708.5	2.3%
VoxLingua107 [45]	15,860	48.8	0.03%
Sound			
AudioSet [46]	1,932,298	5,320.6	7.6%
SoundNet [47]	9,765,588	25,553.8	38.4%
LAION-Audio-630k [48]	1,433,319	3,535.7	5.6%
Music			
Music4All [49]	327,807	910.6	1.3%
Total			
Speech	23,960,264	90,747.6	47.1%
Sound + Music	13,459,012	35,320.7	52.9%
Mix126k-B			
= Speech + (Sound + Music) × 2	50,878,288	161,388.9	100.0%

TABLE II: Hyperparameters of USAD.

Hyperparameter	Small			
	Ablation	Small	Base	Large
Model				
Parameters	24M	24M	94M	330M
Hidden Dimension	384	384	768	1024
Layers	12	12	12	24
Training				
Learning Rate	5e-4	8e-4	1e-3	1.5e-3
LR Warmup	4k	8k	8k	8k
Updates	150k	400k	400k	400k
Batch Size (seconds)	200	800	800	800
GPUs	1	4	4	4

patch-based teacher features are misaligned, complicating the student’s joint representation learning. Thus, we propose distilling from frame-based audio teachers to optimize performance across both speech and audio domains.

IV. EXPERIMENTS

A. Setup

Self-supervised Models. For speech SSL models, we include DPWavLM [31], WavLM Base+ [3], and data2vec 2.0 Speech [23], where DPWavLM is a compressed WavLM Base+ model. For audio SSL models, we consider SSAST [4], BEATs [5], EAT [25], SSLAM [26], and ATST [28], where SSAST and ATST both have a frame-based version, but ATST has a framerate of 25Hz. Additionally, we pre-train two data2vec 2.0 models of the same architecture as the speech version, using both audio and mixed data for comparison.

Data. To maximize domain coverage, we combine multiple datasets containing English speech, sound, and music, as shown in Table I. First, we segment sound and music into 10-second clips, discarding segments shorter than two seconds or longer than 30 seconds for training efficiency, and remove silence clips. All clips are resampled to 16kHz.

TABLE III: Results on SUPERB, AS-20K, and ESC-50. KS1 refers to Speech Commands v1 with 10 types of keywords, a silence class, and an unknown class [53], while KS2 includes all 35 classes in Speech Commands v2. LibriSpeech and AudioSet are denoted as LS and AS, respectively. The best results in **bold** and the second and third best results are underlined.

Model	Params	Data	Feature Type	Frame Rate	Frame-level Speech Tasks			Instance-level Speech Tasks					Instance-level Audio Tasks [♣]		SUPERB Score [↑]			
					PR PER _↓	ASR WER _↓	SD DER _↓	KS1 Acc _↑	KS2 [♣] Acc _↑	IC Acc _↑	SID Acc _↑	ER Acc _↑	AS-20K mAP _↑	ESC-50 Acc _↑	Speech Frame	Speech Instance	Audio	Avg
Speech SSL																		
DPWavLM [31]	24M	LS	frame	50Hz	8.2	10.2	5.5	96.3	97.7	<u>98.6</u>	82.1	65.2	23.5	78.8	751.3	868.7	-499.8	373.4
WavLM Base+ [3]	95M	Mix94k	frame	50Hz	<u>3.9</u>	<u>5.6</u>	3.5	<u>97.4</u>	<u>98.4</u>	99.0	<u>89.4</u>	68.7	24.0	77.0	946.1	<u>945.9</u>	-611.0	427.0
data2vec 2.0 Speech [23]	94M	LS	frame	50Hz	3.7	4.9	6.5	96.8	98.2	<u>98.4</u>	80.8	64.1	26.0	73.7	813.6	870.8	-791.5	297.6
Audio SSL																		
SSAST Patch [4]	89M	AS + LS	patch	50Hz	83.9	97.2	12.9	96.0	98.0	25.4	64.2	59.6	31.0	88.8	-1390.5	613.8	310.3	-155.4
SSAST Frame [4]	89M	AS + LS	frame	50Hz	46.9	22.8	6.8	96.7	98.1	49.3	80.8	60.5	29.2	85.9	312.7	725.3	82.5	373.5
BEATs iter3 [5]	90M	AS	patch	50Hz	36.4	25.9	5.2	<u>97.7</u>	98.3	53.4	57.1	64.5	38.3	<u>95.6</u>	386.1	717.2	<u>903.5</u>	669.0
EAT [25]	88M	AS	patch	50Hz	55.0	25.9	4.7	92.8	98.3	53.7	45.0	62.5	<u>40.2</u>	95.9	331.5	650.9	<u>959.9</u>	647.4
SSLAM [26]	88M	AS	patch	50Hz	56.4	27.8	4.6	98.8	98.1	51.6	42.6	62.6	40.9	95.9	294.8	655.3	993.3	647.8
ATST Frame [28]	86M	AS	frame	25Hz	20.4	18.8	4.7	95.1	98.1	85.4	69.8	64.4	<u>39.0</u>	91.1	597.6	806.9	616.9	673.8
data2vec 2.0 Audio	94M	AS	frame	50Hz	22.8	15.1	4.5	95.1	98.0	85.3	61.1	64.1	<u>35.1</u>	91.0	658.8	779.1	536.4	658.1
Speech & Audio Mixed SSL																		
data2vec 2.0 Mix	94M	Mix126k-B	frame	50Hz	6.3	8.4	4.8	97.1	98.2	98.2	72.1	66.8	33.1	87.8	824.9	872.0	284.7	660.5
Proposed (Teachers: WavLM Base+ & ATST Frame)																		
USAD Small	24M	Mix126k-B	frame	50Hz	7.8	9.5	4.9	96.8	<u>98.4</u>	95.5	73.5	66.3	34.5	89.3	796.2	871.7	411.0	<u>692.9</u>
USAD Base	94M	Mix126k-B	frame	50Hz	5.1	7.7	4.2	97.1	98.5	98.3	<u>88.6</u>	<u>68.0</u>	35.7	91.1	868.9	<u>938.0</u>	554.2	<u>787.0</u>
USAD Large	330M	Mix126k-B	frame	50Hz	<u>4.0</u>	<u>6.5</u>	<u>3.9</u>	97.1	98.5	<u>98.4</u>	91.2	<u>68.4</u>	37.4	<u>92.7</u>	<u>913.4</u>	948.3	693.3	851.7

♣ The SSL models are fine-tuned with downstream task data.

To balance speech and non-speech data, audio and music datasets are upsampled by a factor of two. This balanced dataset is named **Mix126k-B**, where “126k” indicates the pre-upsampled duration in hours. Additionally, we create a smaller dataset for ablation studies by downsampling LibriVox (LV) to match AudioSet (AS), denoted as **LV-AS**. Note that LV is LibriSpeech [38], Libri-Light [39], and Multilingual LibriSpeech [40] combined with no overlapping utterances.

USAD Model and Training. USAD is implemented with fairseq [50] and trained on four NVIDIA A6000 GPUs. First, the input waveform is transformed into a 128-dimensional Mel-spectrogram (25 ms window, 10 ms stride) and normalized following SSAST [4]. If any of the teacher models use frame-wise features, we add a convolutional feature extractor with a stride of two. For patch-wise features, a patch embedding module similar to those used in standard audio SSL approaches is employed. The features are then normalized [51]. If the two teachers have different types of features, we sum the extracted features since they share the same framerate. The features are fed to a 5-layer convolutional positional encoding as in data2vec, then input to a transformer encoder with relative positional encoding [52]. Each prediction head consists of two fully-connected layers with a ReLU activation in between. We use a linear learning rate scheduler with warmup. Detailed hyperparameters for USAD are listed in Table II. When ATST is one of the teachers, we apply mean pooling with a kernel size and stride of two before calculating the loss, as ATST has a lower framerate.

Downstream Tasks. We evaluate USAD on SUPERB [54]–[56] and HEAR [57]. These benchmarks freeze the model to be evaluated, extract the hidden layer representations, and train a lightweight prediction head for each task. Additionally, we fine-tune the encoders for audio tagging (AS-20K) [46] and

sound classification (ESC-50) [58], where AS-20K is a subset of AudioSet containing 20k recordings with balanced labels over all 527 possible sound classes and 5-fold cross-validation is employed for ESC-50.

To offer an overall performance metric across multiple tasks, we compute the SUPERB Score [59] for each model f :

$$\text{SUPERB_Score}(f) = \frac{1000}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{s_{\tau}(f) - s_{\tau}(\text{baseline})}{s_{\tau}(\text{SOTA}) - s_{\tau}(\text{baseline})},$$

where \mathcal{T} is the set of tasks and $s_{\tau}(\cdot)$ is a model’s measured performance of task τ , e.g., $s_{\tau} = \text{WER}$ and $\tau = \text{ASR}$. The baselines are fbank features for SUPERB and randomly initialized SSAST for audio tasks [4].

B. Downstream Speech Evaluation

We first evaluate models on SUPERB by adopting three frame-level tasks: phoneme recognition (PR), automatic speech recognition (ASR), and speaker diarization (SD); and four instance-level tasks: keyword spotting (KS), intent classification (IC), speaker identification (SID), and emotion recognition (ER). As shown in Table III, the USAD Small model surpasses the similar-size DPWavLM on most SUPERB tasks. The performance gap relative to the WavLM teacher narrows as the USAD model size increases, surpassing the SUPERB Score in instance-level speech tasks for USAD Large. USAD outperforms all audio SSL models listed in Table III, especially in frame-level tasks, highlighting the advantage of distillation from a speech SSL teacher. Therefore, USAD demonstrates strong capabilities in speech processing tasks.

C. Downstream Audio Evaluation

Fine-tuning for sound event detection and classification has been a standard evaluation for audio SSL models, so

TABLE IV: HEAR 2021 Benchmark [57]. All metrics are higher-better. VL107 refers to VoxLingua107. The best and second-best results are in **bold** and underlined.

Model	GTZAN																			Speech		NSynth		Vocal Imitation	VL107 Top 10	Avg			
	Beijing BEEHIVE			CREMA-D		DCASE 2016		ESC-50		FSD50K		Gunshot		Music Genre		Libri Speech		Maestro 5h		Mridangam		Commands					Pitch		
	AUROC	Acc	Acc	Onset FMS	Acc	mAP	Acc	Acc	Acc	Acc	Acc	Onset FMS	Acc	Acc	Acc	Acc	Stroke	Tonic	5h	full	5h	50h	Pitch Acc				Pitch Acc	mAP	Acc
Teacher																													
WavLM	61.2	90.2	76.6	88.7	73.3	43.7	93.2	83.4	96.1	77.2	12.8	96.7	89.6	95.9	96.9	34.8	67.4	14.9	73.8	71.9									
ATST Frame	64.6	<u>95.8</u>	76.7	95.7	<u>89.0</u>	55.7	94.3	<u>88.3</u>	100.0	78.1	24.4	97.5	96.9	92.6	95.1	<u>68.6</u>	82.0	<u>22.3</u>	66.9	78.1									
WavLM + ATST	63.4	94.9	79.2	<u>95.0</u>	87.8	<u>56.0</u>	94.0	87.7	100.0	79.4	29.8	97.5	96.5	96.1	97.1	64.2	81.9	21.8	69.4	78.5									
Proposed																													
USAD Small	78.6	94.5	78.2	89.5	81.8	51.1	93.2	86.6	98.5	77.0	25.3	97.3	94.3	96.2	97.2	55.6	77.7	20.0	73.6	77.2									
USAD Base	84.7	<u>95.8</u>	80.0	93.6	82.2	52.2	94.0	86.3	100.0	78.7	26.7	97.3	95.7	96.6	<u>97.6</u>	57.0	81.6	19.5	76.0	78.7									
USAD Large	<u>86.5</u>	94.1	<u>79.5</u>	93.9	83.4	53.0	97.6	87.4	100.0	<u>79.1</u>	<u>38.4</u>	<u>97.4</u>	96.1	<u>97.0</u>	97.5	57.0	<u>83.2</u>	18.5	<u>75.3</u>	<u>79.7</u>									
SOTA [♣]	87.8	97.5	75.2	92.5	96.7	65.5	<u>96.7</u>	90.8	<u>99.2</u>	78.5	46.9	97.5	<u>96.6</u>	97.6	97.8	87.8	90.0	22.7	72.0	83.7									

♣ Frame-level tasks. ♣ Each task’s best result as of May 28, 2025, according to the HEAR leaderboard (<https://hearbenchmark.com/hear-leaderboard.html>).

we follow EAT to fine-tune USAD models [25].² Following SSAST [4], we average the weights from 10k, 15k, . . . , 40k updates checkpoints. As shown in Table III, USAD models outperform all speech and some audio SSL models in audio tasks, despite the gap between USAD and the ATST Frame teacher remaining mainly because USAD is designed to perform well on both speech and general audio tasks. Patch-based approaches dominate sound event classification, but USAD offers steady improvements by scaling up model capacity, surpassing the ATST Frame teacher’s Audio SUPERB Score with the Large model. Moreover, because speech and audio SSL require different masking strategies, data2vec 2.0 Mix is more challenging to learn general domain representations, resulting in inferior performance compared with USAD. Combined with the findings in Section IV-B and the average SUPERB Score, USAD achieves the best overall results, demonstrating the success of the proposed methods.

D. Joint Speech and Audio Evaluation

This section compares USAD with the teacher models on HEAR, a comprehensive benchmark that encompasses 19 tasks covering speech, sound, and music. In Table IV, the WavLM teacher performs poorly because most tasks relate to sound and music, showing the necessity of building a joint speech and audio representation model. Next, we set a topline by concatenating the representations of the two teacher models, denoted as “WavLM + ATST.” Compared to the topline, USAD shows superior performance in speech tasks like CREMA-D (speech emotion classification) and VoxLingua107 (language identification). In contrast, the degradation in sound and music tasks suggests that USAD models may focus more on distilling information from the WavLM teacher. The average performance is improved when USAD is scaled to larger sizes, surpassing the topline and closing the gap with state-of-the-art results. To summarize, the HEAR benchmark offers a comprehensive indicator of the effectiveness of USAD, while also highlighting potential advancements.

²<https://github.com/cwx-worst-one/EAT>

TABLE V: Phoneme error rate (PER) and AS-20K results with USAD Small Ablation distilled from different teacher models. The default setting for USAD is underlined.

Speech Teacher	Audio Teacher	PR PER↓	AS-20K mAP↑
WavLM Base+	–	7.9	27.7
data2vec 2.0 Speech	–	6.6	24.5
–	SSAST Patch	88.7	19.8
–	SSAST Frame	45.0	26.4
–	BEATs iter3	55.5	33.6
–	EAT	58.5	33.8
–	SSLAM	60.5	34.2
–	ATST Frame	28.3	34.8
–	data2vec 2.0 Audio	25.8	32.8
WavLM Base+	SSAST Patch	11.0	28.5
WavLM Base+	SSAST Frame	11.9	29.8
WavLM Base+	BEATs iter3	11.2	30.6
WavLM Base+	EAT	10.0	29.7
WavLM Base+	SSLAM	9.9	29.7
WavLM Base+	<u>ATST Frame</u>	8.7	30.6
WavLM Base+	data2vec 2.0 Audio	9.5	30.3
data2vec 2.0 Speech	ATST Frame	7.4	29.2
data2vec 2.0 Speech	data2vec 2.0 Audio	8.2	29.5

E. Teacher Model Selection

This section discusses the effect of the teacher models for USAD training. All models follow the USAD Small Ablation setup in Table II, trained with the LV-AS dataset mentioned in Section IV-A. In Table V, USAD with one teacher performs well in either the speech or audio domain, implying the necessity of distilling from multiple SSL teachers specializing in different domains. Next, we explore combinations of speech and audio teachers. Compared with frame-based audio teachers, patch-based models degrade PER while offering similar AS-20K results, corroborating our hypothesis that frame-based feature extraction is more suitable to jointly learn speech and audio representations because the learning targets from the two teachers are aligned (Section III-C). A shared phenomenon is the apparent trade-off between PR and AS-20K results, indicating that the choice of teacher models remains a crucial factor in USAD training.

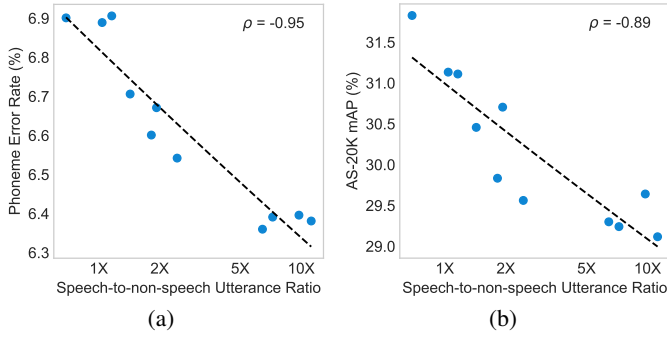


Fig. 2: Training dataset speech-to-non-speech ratio vs. downstream speech (phoneme recognition) and audio (AS-20K) performance. All results are USAD Base with WavLM Base+ and data2vec 2.0 Audio teachers trained on a lower budget.

TABLE VI: Phoneme recognition and AS-20K with USAD Small Ablation and different distillation strategies.

Method	PR PER↓	AS-20K mAP↑
Proposed		
L1-cosine + no masking + $K = 4$ Layers	9.5	30.3
Distillation Objective		
InfoNCE + mask-prediction [18]	10.4	29.1
InfoNCE + no masking	11.3	31.9
L1-cosine + mask-prediction	8.9	28.3
Layer-to-layer Distillation		
$K = 12$ Layers	10.2	29.0
$K = 6$ Layers	9.6	26.5
$K = 3$ Layers	9.4	29.9

F. Data Distribution

We inspect the effect of USAD training data distribution by training 11 USAD models with different combinations of the datasets listed in Table I. As shown in Fig. 2, the ratio between speech and non-speech recordings strongly correlates with downstream performance. That is, with more speech clips present in the training data, USAD’s capability in modeling speech increases (Fig. 2a) but degrades for non-speech tasks (Fig. 2b). Nonetheless, the drop in phoneme recognition performance when reducing speech data is less significant than the improvements on AS-20K, suggesting that training USAD with a balanced dataset of speech and non-speech offers more generalized representations across various domains. Consequently, we propose upsampling the sound and music data in the USAD training data to match the speech data size (Mix126k-B in Table I).

G. Distillation Strategies

This section discusses the effect of distillation strategies in USAD. As shown in Table VI, the InfoNCE loss from CoLLD yields worse results. Under identical conditions, we found replacing InfoNCE with L1-cosine similarity increases training speed by 25%, suggesting the L1-cosine similarity loss is more suitable for USAD. With different distillation objectives, we found that mask-prediction helps PR performance but degrades

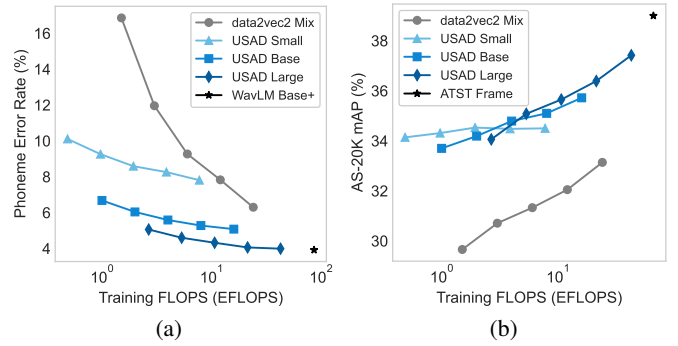


Fig. 3: Training FLOPS vs. downstream speech (phoneme recognition) and audio (AS-20K) performance.

AS-20K. We suspect that the difference in masking strategies for speech and non-speech SSL is the cause of this issue. Moreover, we inspect the effect of the layer-to-layer distillation technique by varying K , i.e., the number of layers to which the distillation loss is applied (Section III-B). The results show that distilling from more hidden layers degrades downstream performance ($K = 12$ and 6), implying that some teacher layer representations are unhelpful for the student to learn from. Meanwhile, $K = 3$ and 4 behave similarly, indicating that a potential improvement may be to find the optimal combination of teacher layers to distill from.

H. Training Efficiency

To demonstrate the training efficiency of USAD compared to SSL approaches, we plot the training FLOPS against downstream performance. In Fig. 3, a distinct capacity–efficiency trade-off is shown: in phoneme recognition, all USAD variants converge at a similar rate, and extra parameters reduce the final PER, implying that the Small model already captures local acoustic cues. In contrast, AS-20K performance scales almost linearly with model size because the task depends on long-range, semantically rich information and benefits from larger receptive fields and extra representational bandwidth, while the larger model capacity also regularizes the weak, noisy labels. Furthermore, USAD Small surpasses data2vec 2.0 Mix with under one EFLOPS, whereas USAD Large attains 36–37% mAP using an order-of-magnitude less compute than ATST Frame. Overall, the results demonstrate the training efficiency of the proposed USAD and suggest that further improvements can be achieved by scaling the model size.

V. CONCLUSION

We present Universal Speech and Audio Distillation (USAD), a unified framework that learns general-purpose audio representations via distillation from domain-specific teachers. USAD bridges speech and non-speech domains with strong performance on SUPERB, HEAR, AudioSet, and ESC-50. Sparse layer-to-layer distillation with a frame-wise L1-cosine similarity loss enables competitive and compute-efficient models. Future work includes improving robustness, extending to multilingual speech, and applying USAD to realistic tasks like audio large language models.

REFERENCES

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
- [2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, 2021.
- [3] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, 2022.
- [4] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *AAAI*, 2022.
- [5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *ICML*, 2022.
- [6] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *ICLR*, 2024.
- [7] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," in *ICLR*, 2024.
- [8] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. MA, and C. Zhang, "SALMONN: Towards generic hearing abilities for large language models," in *ICLR*, 2024.
- [9] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv*, 2024.
- [10] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *EMNLP*, 2024.
- [11] Y. Zhang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv*, 2023.
- [12] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speeche tokenizer: Unified speech tokenizer for speech language models," in *ICLR*, 2024.
- [13] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv*, 2023.
- [14] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv*, 2024.
- [15] H.-J. Chang, H. Gong, C. Wang, J. Glass, and Y.-A. Chung, "Dc-spin: A speaker-invariant speech tokenizer for spoken language models," in *Interspeech*, 2025.
- [16] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *ASRU*, 2023.
- [17] Z. Wang, X. Xia, X. Zhu, and L. Xie, "U-sam: An audio language model for unified speech, audio, and music understanding," in *Interspeech*, 2025.
- [18] H.-J. Chang, N. Dong, R. Mavlyutov, S. Popuri, and Y.-A. Chung, "CoLLD: Contrastive layer-to-layer distillation for compressing multilingual pre-trained speech encoders," in *ICASSP*, 2024.
- [19] A. Baade, P. Peng, and D. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," in *Interspeech*, 2022.
- [20] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," in *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, 2022.
- [21] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *NeurIPS*, 2022.
- [22] 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 *ICML*, 2022.
- [23] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *ICML*, 2023.
- [24] A. H. Liu, H.-J. Chang, M. Auli, W.-N. Hsu, and J. R. Glass, "Dinosr: Self-distillation and online clustering for self-supervised speech representation learning," in *NeurIPS*, 2023.
- [25] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," in *IJCAI*, 2024.
- [26] T. Alex, S. Atito, A. Mustafa, M. Awais, and P. J. Jackson, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," in *ICLR*, 2025.
- [27] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *IJCNN*, 2021.
- [28] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *TASLP*, 2024.
- [29] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *ICASSP*, 2023.
- [30] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP*, 2022.
- [31] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "Dphubert: Joint distillation and pruning of self-supervised speech models," in *Interspeech*, 2023.
- [32] S. Bhati, Y. Gong, L. Karlinsky, H. Kuehne, R. Feris, and J. Glass, "Dass: Distilled audio state space models are stronger and more duration-scalable learners," in *SLT*, 2024.
- [33] K.-P. Huang, T.-h. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, and H.-y. Lee, "Ensemble knowledge distillation of self-supervised speech models," in *ICASSP*, 2023.
- [34] F. Ritter-Gutierrez, Y.-C. Lin, J.-C. Wei, J. H. Wong, E. S. Chng, N. F. Chen, and H.-y. Lee, "Distilling a speech and music encoder with task arithmetic," in *Interspeech*, 2025.
- [35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [36] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.
- [37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.
- [38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.
- [39] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.
- [40] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech*, 2020.
- [41] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.
- [42] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Interspeech*, 2021.
- [43] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.
- [44] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, 2004.
- [45] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *SLT*, 2021.
- [46] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.
- [47] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.
- [48] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023.
- [49] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *IWSSIP*, 2020.
- [50] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT*, 2019.
- [51] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv*, 2016.
- [52] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *NAACL*, 2018.
- [53] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv*, 2018.
- [54] S.-w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Interspeech*, 2021.

- [55] H.-S. Tsai *et al.*, "SUPERB-SG: Enhanced speech processing universal PERFORMANCE benchmark for semantic and generative capabilities," in *ACL*, 2022.
- [56] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang *et al.*, "A large-scale evaluation of speech foundation models," *TASLP*, 2024.
- [57] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, 2022.
- [58] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *ACM MM*, 2015.
- [59] T.-h. Feng *et al.*, "Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *SLT*, 2022.