HARNESS: Lightweight Distilled Arabic Speech Foundation Models

Vrunda N. Sukhadia [1*], **Shammur Absar Chowdhury**[2]
[1]Amazon India, [2]Qatar Computing Research Institute, HBKU, Qatar
sukhadiavrunda@gmail.com, Shchowdhury@hbku.edu.qa

Abstract

Large self-supervised speech (SSL) models achieve strong downstream performance, but their size limits deployment in resource-constrained settings. We present HArnESS, an Arabic-centric self-supervised speech model family trained from scratch with iterative self-distillation, together with lightweight student variants that offer strong accuracy-efficiency trade-offs on Automatic Speech Recognition (ASR), Dialect Identification (DID), and Speech Emotion Recognition (SER). Our approach begins with a large bilingual Arabic-English teacher and progressively distills its knowledge into compressed student models while preserving Arabic-relevant acoustic and paralinguistic representations. We further study PCA-based compression of the teacher supervision signal to better match the capacity of shallow and thin students. Compared with HuBERT and XLS-R, HArnESS consistently improves performance on Arabic downstream tasks, while the compressed models remain competitive under substantial structural reduction. These results position HArnESS as a practical and accessible Arabic-centric SSL foundation for real-world speech applications.
Keywords: Self-supervised model, Distillation, Benchmark resources, Arabic downstream tasks

1. Introduction

Self-supervised learning (SSL) has transformed speech processing by learning transferable representations from large amounts of unlabeled audio. Large SSL models capture rich acoustic and linguistic structure and have shown strong performance across a wide range of speech tasks (Chen et al., 2022; Hsu et al., 2021; Baevski et al., 2022; Mohamed et al., 2022; Chung et al., 2021; wen Yang et al., 2021). These models can be used either as fixed feature extractors or fine-tuned with limited labeled data, making them especially attractive in low-resource settings.

The effectiveness of SSL models, however, depends heavily on the scale, diversity, and balance of the pretraining data. Multilingual SSL models such as XLS-R (Babu et al., 2021) have shown clear advantages for low-resource languages compared with monolingual models trained on high-resource languages such as English (Shi et al., 2023). At the same time, recent evidence suggests that multilingual models may disproportionately favor languages with greater pretraining coverage, which can limit gains for underrepresented languages (Storey et al., 2024). This motivates closer study of language-focused SSL models that better reflect the linguistic and acoustic properties of a target language.

Arabic is a particularly challenging case for speech modeling. It is spoken across 22 countries and exhibits substantial dialectal diversity, with many varieties differing in phonetics, morphology, and lexical usage. In addition, Arabic speech often includes influences from other languages, including English and French (Ali et al., 2021). This diversity makes Arabic speech processing difficult for generic multilingual models, which may not fully capture dialect-sensitive and culturally grounded speech patterns. These challenges motivate Arabic-centric SSL modeling that can better represent spoken Arabic while remaining robust to variation across regions and speaking styles.

At the same time, training and deploying language-focused SSL models remains expensive. Large-scale pretraining requires substantial compute, long training times, and broad unlabeled speech collections. These costs also make deployment difficult in practical and resource-constrained environments, where model size, memory use, and latency matter. Model compression is therefore essential for making such systems more accessible and usable.

Knowledge distillation has emerged as an effective approach for compressing large speech models while preserving much of their performance. In this setting, a smaller student model learns from a larger teacher model, leading to lower memory usage and faster inference with limited degradation in downstream quality. Prior work, including DistillHuBERT (Chang et al., 2022), FitHuBERT (Lee et al., 2022), DPHuBERT (Peng et al., 2023), SKILL (Zampierin et al., 2024), and related methods (Ashihara et al., 2022; Wang et al., 2022), has explored task-agnostic distillation for HuBERT-style models. However, such work has focused primarily on general-purpose compression, with limited attention to Arabic-centric SSL trained from scratch and systematically distilled into lightweight models.

While prior studies have applied self-supervised speech models such as HuBERT and wav2vec 2.0 to Arabic, Arabic-centric SSL remains underexplored, especially in the setting of large-scale

[*]This work was carried out at QCRI.

Proceedings of SPEAKABLE @ LREC 2026, pages 109–117

training from scratch and deployment-oriented compression. In this work, we focus on both aspects. We introduce **H**uBERT-based **Ar**abic a**n**d **E**nglish **S**elf-**S**upervised Speech (**HArnESS**), an Arabic-centric SSL model family trained from scratch on large-scale bilingual Arabic-English speech, and we study iterative self-distillation to build compact student models that retain strong performance on ASR, SER, and DID.

We adopt bilingual Arabic-English pretraining for two reasons. First, English corpora provide additional acoustic and phonetic diversity at scale, which can stabilize representation learning when Arabic resources are comparatively limited and heterogeneous. Second, Arabic speech in real-world settings often includes borrowed English words and code-switching, particularly in conversational and media domains. Our goal is therefore not to weaken Arabic-centric modeling, but to combine Arabic-focused coverage with the regularization benefits of broader bilingual pretraining.

Following the HuBERT training paradigm, we first train a large teacher model, HArnESS-L, with 24 encoder layers through iterative self-distillation. We then transfer its knowledge to smaller students, yielding HArnESS-S, a shallow variant, and HArnESS-ST, a shallow (S) and thin (T) variant. In addition, we investigate low-rank approximation of the teacher supervision signal to simplify the distillation target space and improve knowledge transfer to compact students.

We evaluate HArnESS-L, HArnESS-S, and HArnESS-ST on three downstream tasks spanning content, dialectal, and paralinguistic information, namely ASR, DID, and SER. We compare them against HuBERT-Large, trained primarily on English, and XLS-R, a multilingual SSL model (Babu et al., 2021). Our results show that Arabic-centric pretraining combined with iterative distillation provides an effective balance between task performance and model efficiency.

Our main contributions are as follows:

1. We introduce HArnESS, an Arabic-centric SSL model family trained from scratch, consisting of HArnESS-L (large), HArnESS-S (shallow), and HArnESS-ST (shallow and thin).

2. We study iterative self-distillation as a strategy for compressing Arabic-centric SSL models into lightweight deployment-oriented students.

3. We investigate compact supervision through low-rank approximation of the teacher signal and analyze its effect on student performance.

4. We benchmark the HArnESS family on ASR, DID, and SER, covering content, speaker-related, and paralinguistic speech tasks.

5. We publicly release the distilled models and benchmark resources to support future research.[1]

2. HArnESS Models

2.1. HArnESS Model

Figure 1 illustrates the HArnESS training pipeline, which follows a HuBERT-style iterative self-distillation procedure. At iteration i, we train a model $\mathbf{M}_i$ using discrete pseudo-labels produced from the previous iteration model $\mathbf{M}_{i-1}$. The core idea is masked-prediction pretraining: a subset of time frames is masked, and the model is optimized to predict the corresponding pseudo-labels.

Iterative self-distillation pipeline. Given an input utterance x, we first obtain frame-level embeddings from $\mathbf{M}_{i-1}$ and convert them into discrete targets via clustering (described below), yielding a pseudo-label sequence $z^{(i-1)} = \{z_t^{(i-1)}\}_{t=1}^{T}$ with $z_t^{(i-1)} \in \{1, \ldots, K\}$. We then train $\mathbf{M}_i$ to predict these targets from contextualized frame representations, where some frames are replaced by a mask token (or masked spans), forcing the model to use broader context.

Training regimen and compression schedule. We perform multiple iterations of refinement. In the first two iterations, we keep the model architecture unchanged to encourage progressively stronger acoustic abstractions. Starting from the third iteration, we distill and compress the model to obtain efficient variants. Specifically, we explore three compression axes: *(a)* reducing Transformer depth (d) to obtain a shallower model; *(b)* reducing model width (encoder dimension, enc_d) to obtain a thinner model; and *(c)* reducing attention capacity by decreasing the number of attention heads ($attn$). This schedule yields a teacher-like encoder in early iterations and a family of compact students in later iterations.

Model architecture. HArnESS consists of a convolutional (CNN) feature extractor followed by a stack of Transformer encoder layers. Similar to HuBERT encoders, the CNN front-end comprises 7 temporal convolution layers that transform raw audio into latent frame features. The Transformer encoder contains d layers with hidden dimension, emb_d. Each layer includes multi-head self-attention (MHA) with $attn$ heads and a position-wise feed-forward network (FFN). A linear prediction head

[1] https://huggingface.co/QCRI/distillHarness

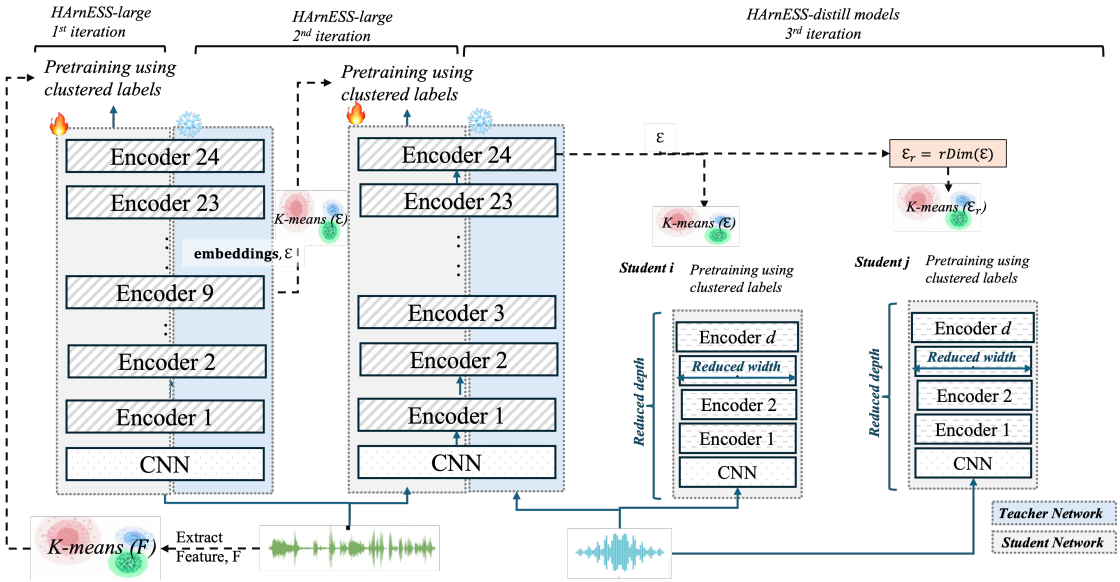


Figure 1: Overview of the iterative self-distillation framework used to build the HArnESS model family.

maps contextualized frame representations to a categorical distribution over K cluster IDs.

Training objective. HArnESS is trained with a HuBERT-style masked prediction objective. For each utterance, we mask a subset of time frames (or spans) in the input representation and train $\mathbf{M}_i$ to predict the corresponding discrete pseudo-labels generated from $\mathbf{M}_{i-1}$. We use standard cross-entropy classification over the cluster IDs. To improve stability, we compute the loss on both masked and unmasked frames and combine them with a fixed weighting: the masked-frame loss encourages contextual reasoning over surrounding speech, while the unmasked-frame loss provides an additional learning signal that helps prevent training collapse and improves convergence.

Pseudo-label generation. To obtain discrete targets for iteration i, we extract frame-level embeddings from the previous model $\mathbf{M}_{i-1}$ and cluster them using K-means, producing a pseudo-label sequence $z^{(i-1)} = \{z_t^{(i-1)}\}_{t=1}^{T}$ with $z_t^{(i-1)} \in \{1, \ldots, K\}$. Unless otherwise stated, we use last-layer embeddings, which provide the most abstract and stable representations.[2]

PCA for supervision signal compression. We also investigate PCA as a way to compress the teacher supervision signal before clustering. Instead of clustering the full teacher embedding, we optionally project it to a lower-dimensional space and then derive pseudo-labels from the projected representation. Concretely, given an embedding vector $h_t \in R^D$ from $\mathbf{M}_{i-1}$, we apply PCA to obtain a compressed representation $\tilde{h}_t \in R^{D'}$ ($D' \ll D$) and then cluster $\tilde{h}_t$.

This serves two purposes. First, dimensionality reduction can remove noisy or redundant directions, which may improve clustering robustness. Second, it produces a simpler target space that better matches the capacity of compressed students, especially when model width (enc_d) is reduced. In this sense, PCA does not compress the student input directly; rather, it simplifies the discrete targets used during distillation.

Initialization and iteration-specific supervision. For the initial iteration $i = 1$, we bootstrap pseudo-labels by extracting MFCC features from raw speech x and clustering them to obtain $z^{(0)}$. For the next iteration ($i = 2$), we generate pseudo-labels from intermediate representations of $\mathbf{M}_0$, using the 9th Transformer layer embeddings for clustering. For all subsequent iterations ($i \geq 3$), we generate pseudo-labels using last-layer embeddings from $\mathbf{M}_{i-1}$, which provide the most abstract and stable representations.

For training $\mathbf{M}_i$ ($i \geq 1$), we explore two weight initialization strategies: *(a)* random initialization (uniformly sampled weights), and *(b)* blocked-averaging initialization, where groups of student layers are initialized by averaging corresponding blocks of layers from $\mathbf{M}_{i-1}$. Blocked averaging provides a smoother transition across iterations and often improves stability, particularly when compressing depth/width in later iterations.

[2]We also explored averaged embeddings from selected layers and observed no consistent gains.

3. Experimental Setups

3.1. Pre-training Data

Iterations 1–2 (bilingual pretraining). We pretrain HArnESS on a mixture of publicly available Arabic and English speech corpora, including QASR (Mubarak et al., 2021), MGB3 (Ali et al., 2017), LibriSpeech (Panayotov et al., 2015), Common Voice (Arabic/English) (Ardila et al., 2020), and GigaSpeech (Chen et al., 2021), among others. The base pretraining pool consists of approximately **4K hours of Arabic** and **3.56K hours of English** speech. We further expand the training data through augmentation to reach approximately **23K hours** in total. Of this augmented portion, around **300 hours** come from additive background-noise augmentation, while the remainder is primarily produced through SpecAugment-based transformations.

To improve dialectal and cultural coverage, we also incorporate spoken content from 15 Arabic-speaking countries crawled from YouTube, covering diverse Arabic dialects. We provide a coarse dialect breakdown of the Arabic data by major region in Table 1, grouping samples into MSA, Gulf, Levantine, Egyptian, Maghrebi, mixed, and unlabeled categories where exact dialect labels are unavailable. All official development and test partitions are excluded from pretraining to avoid data leakage.

Category	Sub-category / Dialect	Duration (Hrs)
Original Clean Data		**7,566.00**
	English Subset	3,565.00
	Arabic Subset	4,001.00
	MSA / General Arabic	3,603.28
	Levantine	107.69
	Egyptian	109.20
	Gulf	77.13
	Maghrebi	69.11
	Other	34.59
Augmented Data		**15,434.00**
	Speed Perturbation ($0.9\times, 1.1\times$)	15,134.00
	Noise Augmentation (Arabic)	300.00
Total Training Volume		**23,000.00**

Table 1: Comprehensive breakdown of the 23,000-hour training corpus, including language distribution, dialectal variety, and augmentation strategies.

Iteration 3 (Arabic-only distillation). Our primary goal is to obtain lightweight Arabic-centric models. Accordingly, for the distillation/compression iteration, we use approximately **1,100 hours** of Arabic speech drawn from the QASR training data. For K-means training in this phase, we randomly sample **30%** of the iteration-3 data (approximately **300 hours**) to reduce clustering cost while maintaining linguistic diversity.

3.2. Downstream Tasks and Data

Benchmarking SSL speech encoders for English is supported by standardized suites such as SUPERB (wen Yang et al., 2021). In contrast, Arabic speech lacks an analogous standardized benchmark. To address this gap, we evaluate HArnESS across three representative Arabic tasks: **ASR** (content recognition), **dialect identification (DID)** (speaker information), and **speaker emotion recognition (SER)** (paralinguistic analysis).

ASR. We fine-tune on a **300-hour** subset of QASR and evaluate on the MGB2 (Ali et al., 2019) test set. To assess out-of-domain generalization, we additionally report performance on the MGB3 test set.

SER. We use KSUEmotion (Meftah et al., 2021), collected from 23 speakers with six emotion classes. The dataset is split into train (3.30 h), dev (0.83 h), and test (1.0 h).[3]

DID. We use the ADI5 dataset with five region-based dialect classes (MSA, Egyptian, Levantine, North African, and Gulf) and the official train/dev/test splits.

Metrics. We report word error rate (WER) for ASR and classification accuracy (Acc) for DID and SER.

3.3. Pre-training Hyperparameters

We train HArnESS using the `fairseq` codebase (Ott et al., 2019). Table 2 summarizes the key hyperparameters. Unless stated otherwise, we only change the supervision source and model capacity (Table 3).

3.4. Pre-training Procedure

Model configurations for the upstream encoders are summarized in Table 3.

Iterations 1–2 (HArnESS-L). For the first two iterations, we train the large model (**HArnESS-L**; 24 Transformer layers) on the 23k-hour bilingual mixture. Iteration 1 is trained for 500k steps and iteration 2 for 700k steps. For iteration 1 pseudo-labels, we cluster 39-dimensional MFCC features with $K{=}1000$ clusters. For iteration 2 pseudo-labels, we extract latent representations from the **9th Transformer layer** of the iteration-1 model and cluster them with $K{=}1000$ clusters to obtain refined targets.

Iteration 3 (compressed students). For iteration $i{=}3$, we train compressed models using **HArnESS-S** and **HArnESS-ST**, both with a 4-layer Transformer encoder and reduced capacity (Table 3).

[3]We will release our split for reproducibility.

Total pre-training audio (Iter 1–2)	23k hours (Arabic/English ≈ balanced)
K-means training subset (Iter 1–2)	300 hours
Distillation audio (Iter 3)	1,100 hours (Arabic-only)
K-means training subset (Iter 3)	30% ≈ 300 hours
# clusters (K)	1000
Feature type for $i{=}0$ targets	MFCC (39-dim)
Embedding layer for $i{=}1$ targets	Layer 9 embeddings of $\mathbf{M}_0$
Embedding layer for $i{\geq}2$ targets	Last-layer embeddings of $\mathbf{M}_{i-1}$
Iteration 1 steps / GPUs / batch	500k / 24×H100 / 62.5s audio per GPU
Iteration 2 steps / GPUs / batch	700k / 24×H100 / 62.5s audio per GPU
Iteration 3 steps / GPUs / batch	300k / 8×H100 / 75s audio per GPU
Mask probability (p_{mask})	0.80
Mask span length (frames)	10
PCA dimension (D'; when enabled)	512

Table 2: Key pre-training hyperparameters.

Models	**XR**	**HuL**	**H-L**	**H-S**	**H-ST**	**H-ST (PCA)**
Supervision	–	–	$L9_{\text{emb}}$ $(i=1)$	$L23_{\text{emb}}$ $(i=2)$		PCA($L23_{\text{emb}}$) $(i=2)$
CNN Encoder						
Strides	5, 2, 2, 2, 2, 2, 2					
Kernel Width	10, 3, 3, 3, 3, 2, 2					
Channels	512					
Transformer						
Depth (l)	24	24	24	4	4	4
Emb. Dim (emb_d)	1024	1024	1024	1024	512	512
FFN Dim (d_{ffn})	4096	4096	4096	2048	2048	2048
Attn. Heads (h_{attn})	16	16	16	16	16	16
Projection						
Dim. (d_p)	768	768	768	768	768	768
Params						
in M	300	316	316	65	28	28

Table 3: SSL Model Architecture Comparison. XR: XLS-R, HuL: HuBERT-Large, H-L: HArnESS-Large, H-S: HArnESS-Shallow, H-ST: HArnESS-Shallow and Thin. Dim. dimension, Emb.: Embedding. $L{*}_{emb}$: Embedding from layer * (e.g. 23) of model from iteration i.

Pseudo-labels are generated by clustering **last-layer** embeddings from the iteration-2 HArnESS-L teacher with $K{=}1000$ clusters. Thus, HArnESS-S and HArnESS-ST correspond to the third-iteration distilled models trained on 1,100 hours of Arabic speech.

3.5. Downstream Training

For downstream evaluation, we keep the SSL encoder frozen and use it strictly as a feature extractor. We obtain frame-level representations from all Transformer layers, average them to form an utterance-level representation, and train task-specific downstream models on top of these fixed features only. No gradients are propagated into the SSL encoder during downstream training.

3.5.1. DID and SER Architecture

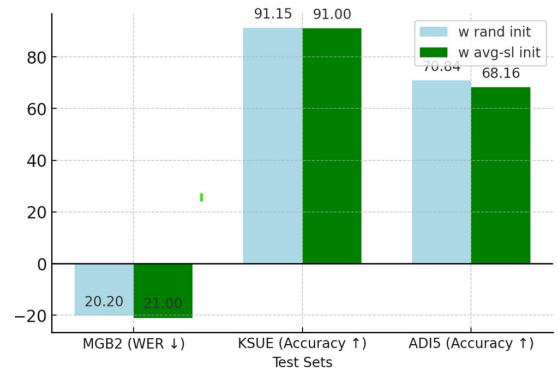
For DID and SER, we train a lightweight classifier on top of the frozen SSL features with a batch size of 4 for 10k steps. The classifier consists of three temporal convolution layers with kernel size 5, ReLU activations, and dropout of 0.4, followed by self-attention pooling, a feed-forward layer, and a final softmax layer. All hidden dimensions are set to 80. This setup isolates the quality of the learned SSL representations by keeping the encoder fixed and limiting the trainable parameters to the downstream classifier.

3.5.2. ASR Architecture

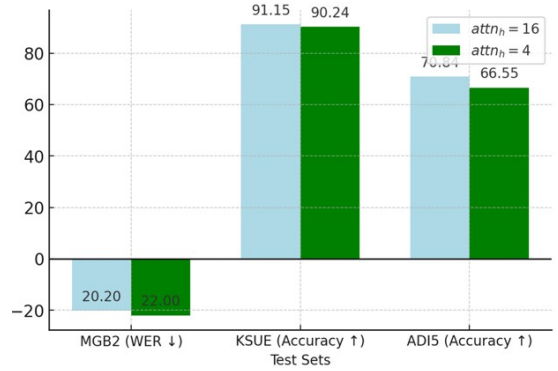
For ASR, we train an encoder–decoder model with a joint CTC/attention objective using the ESPnet toolkit.[4] The encoder consists of two Conformer layers and the decoder consists of two Transformer layers, each with 8 attention heads and 2048 linear units. We train for 70 epochs.

[4] Using ESPnet toolkit.

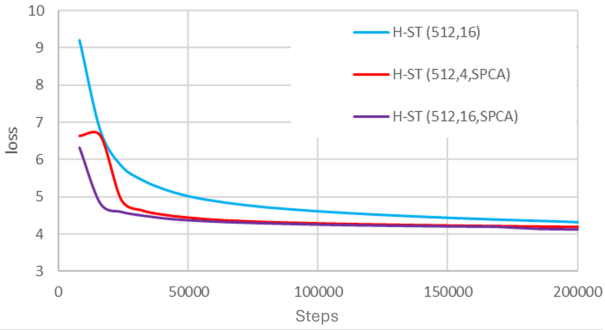
4. Results



(a) Weight Initialization



(b) Number of attention heads



(c) Effect of dimension reduction

Figure 2: Ablation results for the compressed student models. a) Effect of student initialization strategy. (b) Effect of reducing the number of attention heads. (c) Effect of applying PCA to teacher embeddings before clustering when generating pseudo-labels for student training. H-ST(emb_d, $attn_h$), SPA means PCA applied for supervision.

Comparison with Upper Bound: SOTA Model As contextual reference, Table 4 also reports representative published results from strong task-specific systems for Arabic ASR, DID, and SER. These results come from separate studies and are not directly comparable to our models because they differ in architecture, supervision, training data, and experimental protocol. We therefore use them only to contextualize performance, not to claim direct state-of-the-art results. For ASR, we report results alongside Fanar ASR (Fanar et al., 2025), a specialized system trained on more than 10K hours of MSA and dialectal Arabic speech. Under our much more constrained setup, where ASR fine-tuning uses only 300 hours of MSA data, HArnESS-L remains within about 5 WER points on MGB2 and MGB3, and HArnESS-S within about 10 points. For DID and SER, HArnESS-L also shows strong results relative to the published reference numbers, achieving 84.98% on ADI5 compared with 82.5% (Kulkarni and Aldarmaki, 2023), and 94.66% on KSUEmotion compared with 85.53% (on ResNet-based architecture). Although these comparisons are only approximate, they indicate that HArnESS-L is competitive with strong specialized systems, while the distilled students preserve much of the teacher's performance with substantially smaller capacity.

HArnESS-L *vs.* Existing SSLs for Arabic Compared with HuBERT-L and XLS-R, HArnESS-L performs better across the evaluated Arabic tasks, suggesting that Arabic-centric pretraining is beneficial for downstream Arabic speech processing. The compressed HArnESS variants also outperform the multilingual XLS-R baseline on several tasks, indicating that iterative distillation preserves useful task-relevant structure even under heavy compression.

Effects of structural compression and design choices. Figure 2 presents three ablation studies on student design, covering initialization, structural compression, and supervision compression. For iteration $i = 3$, we first examined the effect of student weight initialization and observed only minor differences in downstream performance (Figure 2). This indicates that initialization plays a limited role at this stage, and that performance depends more strongly on the distilled supervision signal.

We then evaluated the effect of reducing model depth. HArnESS-S achieves 79.4% structural compression relative to HArnESS-L while maintaining strong performance across tasks. Despite this compression, it still outperforms the multilingual and English SSL baselines, highlighting the effectiveness of Arabic-centric distillation. Compared with HArnESS-L, however, HArnESS-S shows a 4.7 absolute increase in WER, a 3.51-point drop in SER accuracy, and a 14.4-point drop in DID accuracy. The larger degradation on DID suggests that dialect-related cues are harder to preserve in shallower architectures.

Next, we reduced the number of attention heads from HArnESS-S ($attn = 16$) to HArnESS-S*

Models	ASR (WER ↓)		SER (Acc ↑)	DID (Acc ↑)
	MGB2	MGB3	KSUEmotion	ADI5
Published task-specific reference results (context only)				
Reference result	10.24 (Fanar et al., 2025)	21.31 (Fanar et al., 2025)	85.53% (Abouzeid et al., 2025)	82.5% (Kulkarni and Aldarmaki, 2023)
Our downstream evaluation with frozen SSL encoders				
HuBERT-L (*English*)	22.6*	51.2*	91.92%	64.14%
XLS-R (*Multilingual*)	22.60*	51.80*	73.32%	42.35%
HArnESS-L (*Bilingual: Arabic-English*)	**15.50***	**41.60***	**94.66%**	**84.98%**
Compressed HArnESS students distilled with ≈1000h Arabic-only data				
HArnESS-S ($\Delta S = 79.4\%$)	*20.20**	*52.80**	*91.15%*	*70.84%*
HArnESS-ST ($\Delta S = 93.7\%$)	23.20*	58.20*	89.02%	69.77%
HArnESS-ST$^{\Xi}$ ($\Delta S = 93.7\%$)	22.50*	55.60*	87.34%	61.64%

Table 4: Performance comparison on ASR, SER, and DID. ASR downstream results in our setup are obtained by training on a **300h QASR** subset, results denoted by a *. The top block lists previously published task-specific reference results from separate studies and is included only for context. These numbers are not from a unified baseline and are not directly comparable to the models evaluated in our setup. L denotes the large teacher model, S the shallow student, and ST the shallow-thin student. ΔS denotes overall structural compression relative to HArnESS-L.

Test Sets	emb_d=1024	emb_d=512	emb_d=256
MGB2 (WER ↓)	20.2	23.20	22.3
KSUEmotion (Acc ↑)	91.15%	89.02%	79.42%
ADI5 (Acc ↑)	70.84%	69.77%	53.41%
ΔS	70.43%	91.14%	96.52%

Table 5: Performance Comparison for different embedding dimensions. ΔS: Overall structural compression.

(*attn* $= 4$), which yields an additional 26.15% structural compression, reducing the model from 65M to 48M parameters. This change has only a limited effect on ASR and SER, but causes a larger drop on DID (Figure 2), again indicating that dialect-sensitive information is more susceptible to architectural compression.

Finally, we examined embedding-dimension reduction (Table 5). At extreme compression ($\Delta S = 96.52\%$ relative to HArnESS-L), performance drops sharply across tasks. This result suggests that overly aggressive dimensionality reduction weakens the representational capacity of the student and substantially limits downstream performance.

Effect of compressing the supervision signal. We also study whether simplifying the teacher supervision signal improves student training. Specifically, in iteration $i = 3$, we compare knowledge distillation with and without applying PCA to the teacher embeddings before clustering. As shown in Figure 2c, supervision derived from PCA-reduced embeddings converges faster than supervision from the original embeddings. This suggests that reducing redundancy in the teacher feature space produces a cleaner supervision signal, leading to more stable and efficient optimization while preserving effective knowledge transfer.

5. Conclusion

In this work, we introduced HArnESS, an Arabic-centric self-supervised speech model family designed to better capture the diversity of Arabic dialectal speech. Using an iterative self-distillation framework, we transferred knowledge from a large bilingual teacher model to compact shallow and shallow-thin student models while preserving Arabic-relevant speech representations. Experiments on Arabic ASR, SER, and DID show that HArnESS is competitive with, and in some cases stronger than, multilingual baselines such as HuBERT and XLS-R. The compressed HArnESS variants further offer an attractive efficiency-performance trade-off, making them suitable for more resource-constrained settings. Our downstream evaluation relies on frozen encoders, providing a controlled assessment of representation quality, but it does not fully reflect the gains that may emerge under end-to-end fine-tuning. Future work will extend the comparison to fine-tuned settings and broader baselines. We will publicly release the lightweight models and benchmarking resources to facilitate future research.

6. Bibliographical References

Ali Abouzeid, Bilal Elbouardi, Mohamed Maged, and Shady Shehata. 2025. Arabemonet: A lightweight hybrid 2d cnn-bilstm model with attention for robust arabic speech emotion recognition.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition.

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting Arabs: bridging the gap in dialectal speech recognition. *Communications of the ACM*, pages 124–129.

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, and Tomohiro Tanaka. 2022. Deep versus wide: An analysis of student architectures for task-agnostic knowledge distillation of self-supervised speech models. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2022. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091. IEEE.

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

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Fanar, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *arxXiv:2501.13944*.

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

Ajinkya Kulkarni and Hanan Aldarmaki. 2023. Yet another model for Arabic dialect identification. In *Proceedings of ArabicNLP 2023*, pages 435–440, Singapore (Hybrid). Association for Computational Linguistics.

Yeonghyeon Lee, KANGWOOK JANG, Jahyun Goo, Youngmoon Jung, and Hoi-Rin Kim. 2022. Fithubert: Going thinner and deeper for knowledge distillation of speech self-supervised learning. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, pages 3588–3592. ISCA.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Yifan Peng, Yui Sudo, Shakeel Muhammad, and Shinji Watanabe. 2023. Dphubert: Joint distillation and pruning of self-supervised speech models. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 62–66.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. Ml-superb: Multilingual speech universal performance benchmark.

Edward Storey, Naomi Harte, and Peter Bell. 2024. Language bias in self-supervised learning for automatic speech recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 37–42. IEEE.

Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, and Haizhou Li. 2022. Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert. *Proc. Interspeech 2022*, pages 1686–1690.

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, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark.

Luca Zampierin, Ghouthi Boukli Hacene, Bac Nguyen, and Mirco Ravanelli. 2024. Skill: Similarity-aware knowledge distillation for speech self-supervised learning. *arXiv preprint arXiv:2402.16830*.

7. Language Resource References

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

Ali Hamid Meftah, Mustafa A. Qamhan, Yasser Seddiq, Yousef A. Alotaibi, and Sid Ahmed Selouani. 2021. King saud university emotions corpus: Construction, analysis, evaluation, and comparison. *IEEE Access*, 9:54201–54219.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera Speech Resource. A Large Scale Annotated Arabic Speech Corpus. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2274–2285, Online. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. 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.