DELULU: Discriminative Embedding Learning Using Latent Units for Speaker-Aware Self-Supervised Speech Foundational Model

Anonymous ACL submission

Abstract

Self-supervised speech models have achieved remarkable success on content-driven tasks, yet they remain limited in capturing speaker-discriminative features critical for verification, diarization, and profiling applications. We introduce DELULU, a speaker-aware self-supervised foundational model that addresses this limitation by integrating external supervision into the pseudo-label generation process. DELULU leverages frame-level embeddings from ReDimNet, a state-of-the-art speaker verification model, to guide the k-means clustering step during pre-training, introducing a strong speaker-discriminative inductive bias that aligns representation learning with speaker identity. The model is trained using a **dual objective** that combines masked prediction and denoising, further enhancing robustness and generalization. DELULU significantly outperforms prior self-supervised learning (SSL) models across a range of speaker-centric tasks, achieving up to **62% relative improvement** in equal error rate (EER) for speaker verification and consistent gains on zero-shot profiling tasks such as gender, age, accent, and speaker counting. Our findings demonstrate that **DELULU is a strong universal encoder for speaker-aware speech processing**, enabling superior performance even without task-specific fine-tuning.

1 Introduction

Speaker information is essential for a wide range of speech-related applications, including speaker verification, diarization, and personalized speech generation (Reynolds et al., 2000; Anguera et al., 2012; Casanova et al., 2022). Despite the recent success of SSL (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) in speech representation, existing models still struggle to capture speaker-specific characteristics effectively. The lack of robust speaker-aware representations poses a fundamental limitation for building systems that rely heavily on identity cues (Qian et al., 2022; Zhang et al., 2023).

Although self-supervised models have achieved strong results across a variety of speech tasks and audio tasks (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Waheed et al., 2024), there is a reason why their performance on speaker-related applications remains limited. The key bottleneck lies in their reliance on pseudo-labels generated through acoustic-only clustering, which are insufficiently aligned with speaker-discriminative structure. In HuBERT (Hsu et al., 2021), the k-means clustering step relies on shallow acoustic features that prioritize phonetic similarity, often suppressing speaker-specific information such as voice quality, prosody, and speaking style. Although WavLM (Chen et al., 2022) introduces additional context modeling and denoising objectives, it inherits the same clustering mechanism and, as a result, continues to struggle with learning robust speaker representations.

We introduce DELULU, a speaker-aware self-supervised foundational model that explicitly incorporates speaker supervision into the pretraining process. DELULU leverages ReDimNet (Yakovlev et al., 2024), a SOTA SV network, to guide k-means clustering with frame-level speaker embeddings rather than purely acoustic features. This external supervision introduces a strong inductive bias into the self-supervised learning pipeline, enabling the model to capture speaker-relevant information more effectively without compromising general acoustic modeling. Furthermore, the model is trained using a dual objective combining masked prediction and denoising which enhances its robustness and generalization.

A stronger speaker-oriented foundation model should yield representations that better encode speaker identity, leading to improved performance on forensic and identity-centric applications both in zero-shot settings and after task-specific fine-

tuning. We show that this holds true: **DELULU** outperforms existing SSL models across speaker verification, profiling (age, gender, accent), and speaker counting tasks, with particularly large gains in zero-shot evaluation. These improvements translate directly to downstream fine-tuning, confirming that introducing speaker-aware supervision at the clustering stage produces more transferable representations. Beyond performance, DELULU illustrates a broader principle: *self-supervised speech models can be strengthened by integrating targeted external supervision to guide pseudo-label formation.* This framework provides a scalable path toward speaker-aware foundation models and can be generalized to other architectures and domains beyond speaker modeling.

Our main contributions are as follows:

- We introduce DELULU, a speaker-aware self-supervised speech model that addresses the fundamental limitation of speaker discriminability in SSL by integrating external speaker supervision into the pseudo-label generation process through ReDimNet guided k-means clustering.
- We design a dual-objective training regime, combining masked prediction and denoising, which encourages the model to encode both content and identity robustly under noisy and clean conditions.
- We achieve state-of-the-art performance across a wide suite of speaker-centric benchmarks, including speaker verification, profiling (age, gender, accent), and speaker counting demonstrating up to 62% relative improvement in EER.
- We establish DELULU as a strong universal encoder for speaker tasks, outperforming prior SSL models by a large margin in zero-shot settings and showing competitive results even in upstream performance comparisons.

2 Related Work

2.1 Self-Supervised Speech Representation Learning

Self-supervised learning (SSL) has revolutionized speech processing by enabling models to learn rich representations from large-scale unlabeled audio data (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). These models typically employ pretext tasks such as contrastive predictive coding or masked language modeling to capture phonetic and acoustic structures without explicit labels. Wav2vec 2.0 (Baevski et al., 2020) introduced a contrastive loss over quantized latent representations, achieving strong performance on speech recognition benchmarks. HuBERT (Hsu et al., 2021) advanced this by using offline k-means clustering on MFCC features to generate pseudo-labels, focusing on discrete unit discovery that aligns well with phonetic content. WavLM (Chen et al., 2022) further incorporated denoising objectives and utterance mixing to improve robustness to noise and overlapping speech, making it suitable for a broader range of tasks including speaker identification. Despite these advances, standard SSL models often underperform on speaker-centric tasks due to their emphasis on content over speaker identity (Chen et al., 2022). Recent efforts have explored scaling these models to multilingual settings (Pratap et al., 2024) or enhancing them with multi-task learning (Hu et al., 2024), but speaker discriminability remains a challenge.

2.2 Supervised Speaker Representation Learning

Traditional speaker verification systems rely on the fundamental assumption that the human voice is a unique biometric trait (Singh and Raj, 2025). This expected uniqueness underpins decades of research into supervised speaker modeling pipelines. The x-vector system (Snyder et al., 2018) uses time-delay neural networks (TDNNs) with statistical pooling to produce fixed-dimensional utterance embeddings, achieving robust performance on benchmarks like VoxCeleb (Nagrani et al., 2017). Advancements include ECAPA-TDNN (Desplanques et al., 2020), which incorporates channel and context-dependent attention to better capture speaker variability. More recently, ReDimNet (Yakovlev et al., 2024) introduced a reshaped dimensionality network that optimizes for both local and global speaker features, setting new state-of-the-art results on speaker verification tasks. However, these supervised embeddings are often over-specialized for identity discrimination, limiting their generalization to broader speaker profiling tasks such as age and gender estimation or forensic analysis (Baali et al., 2024). For instance, phonetic biases in attention mechanisms can confound speaker-specific cues with content-related artifacts, reducing transferability to non-identity attributes (Baali et al., 2024). These supervised approaches excel in controlled settings but require

large labeled datasets and struggle with domain shifts (Desplanques et al., 2020). Integrating their strengths into SSL pipelines offers a promising direction to enhance unsupervised representation learning, enabling foundational models that produce versatile, speaker-aware representations suitable for diverse downstream applications.

2.3 Self-Supervised Learning for Speaker Tasks

A growing body of work adapts SSL specifically for speaker-related tasks such as verification and diarization, often leveraging contrastive or distillation-based objectives to learn speaker-discriminative representations from unlabeled data (Lepage and Dehak, 2025). Early efforts applied contrastive learning frameworks from computer vision, such as SimCLR adapted for speech (Jiang et al., 2020), which minimizes distances between augmented views of the same utterance while maximizing separations from others. Self-distillation approaches, inspired by DINO (Caron et al., 2021; Ashihara et al., 2024), have shown particular promise. More recently, CoLMbo (Baali et al., 2025b) proposed a speaker language model that integrates prompt-conditioned speaker encoders to generate descriptive speaker profiles. This approach showed strong zero-shot generalization on demographic traits such as dialect and age, demonstrating the potential of using large-scale speaker models beyond classification tasks. Meanwhile, CAARMA (Baali et al., 2025c) introduced an augmentation framework where HuBERT's hidden layers serve as discriminators during adversarial training, improving the speaker discriminability of the learned embeddings. These methods suggest that integrating architectural and training-level bias can enhance SSL's ability to encode speaker traits but they do not intervene directly at the pseudo-label generation step as DELULU does.

DELULU builds on these by directly guiding SSL clustering with supervised speaker features, bridging the gap between general acoustic modeling and targeted speaker discrimination.

2.4 Robustness and Denoising in Speech Models

To handle real-world degradations, several SSL models incorporate denoising objectives. WavLM (Chen et al., 2022) simulates noisy environments during pretraining by overlaying noise and using clean pseudo-labels for prediction. Similar techniques appear in robust speaker models (Chen et al., 2022), which pretrain on noisy data to improve verification in adverse conditions. Multi-modal and noise-aware pretraining (Chen et al., 2023) further enhances resilience, but most focus on phonetic rather than speaker robustness. DELULU's dual-objective training extends these ideas by combining denoising with speaker-guided clustering for improved performance across noisy speaker tasks.

3 Architecture

Based on the architecture depicted in Figure 1, DELULU follows a dual-objective design that integrates self-supervised learning with speaker-discriminative guidance through an external teacher model.

3.1 DELULU Encoder

The DELULU encoder follows the wav2vec 2.0 architecture (Baevski et al., 2020), consisting of a convolutional feature extractor, a Transformer encoder, a projection layer, and a code embedding layer. The convolutional encoder is composed of seven 512-channel layers with strides $[4, 2, 2, 2, 2, 2, 2]$ and kernel widths $[10, 3, 3, 3, 3, 2, 2]$, differing from HuBERT (Hsu et al., 2021) by adjusting the stride pattern to ensure temporal alignment with the teacher model's frame-level outputs. For 16 kHz input speech, this produces a latent feature sequence with a 16 ms frame rate (256$\times$ down-sampling factor). These latent features are then passed through a stack of Transformer blocks that model long-range dependencies, after which a projection layer reduces dimensionality and prepares features for clustering and loss computation. Overall, the DELULU encoder serves as the **student model** that learns to predict both denoised and clustered representations.

3.2 Teacher-Guided Clustering

Unlike conventional self-supervised speech models that rely purely on acoustic clustering, DELULU leverages **ReDimNet** (Yakovlev et al., 2024), a state-of-the-art speaker verification model, to guide the pseudo-label generation process. Instead of using pooled utterance-level embeddings, we extract *prepooled frame-level features* from ReDimNet to preserve temporal resolution and ensure alignment with DELULU's encoder outputs. These features are then used for k-means clustering with $k = 256$

clusters. This approach ensures that the discrete targets encode speaker-specific characteristics such as voice quality, prosody, and speaking style, rather than prioritizing only phonetic content.

3.3 Training

DELULU is trained using a **dual-task objective** that combines masked prediction and denoising.

Denoising Loss: To enhance robustness, we augment training with a denoising objective. Real-world noise Cross-talk,Environmental is simulated and added to clean speech at SNRs uniformly sampled from [15, 25] dB. Although this represent moderate level of noise since the purpose is non-aggressive denoising to help improve the robustness. Both clean x and noisy $\tilde{x}$ waveforms are passed through the encoder, with attention masks from the clean sample applied to maintain alignment. The denoising loss minimizes:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{h}_{\text{clean}}^{(t)} - \mathbf{h}_{\text{noisy}}^{(t)}\|^2 \quad (1)$$

where $\mathbf{h}_{\text{clean}}$ and $\mathbf{h}_{\text{noisy}}$ are the clean and noisy embeddings.

Masked Prediction Loss: Following the masked language modeling paradigm, a portion of the input time steps are randomly masked. The model is trained to predict the cluster assignments (derived from ReDimNet-guided k-means) at the masked positions. Given the model's output $\mathbf{C}_i$ at masked position i and the corresponding target cluster ID, the cross-entropy loss is computed as:

$$\mathcal{L}_{\text{mask}} = -\sum_{i \in \mathcal{M}} \log P(c_i^* \mid \mathbf{C}_i) \quad (2)$$

where $\mathcal{M}$ denotes the set of masked positions, $\mathbf{C}_i$ represents the logits over the $k = 256$ possible clusters, and c_i^* is the ground-truth cluster assignment obtained from the teacher model.

The final training objective combines both losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}} + \lambda \cdot \mathcal{L}_{\text{denoise}} \quad (3)$$

where λ is a hyperparameter balancing the two objectives.

4 Experimental Setup

We build our system on top of the official `torchaudio` self-supervised learning recipes. For clustering, we adopt the MiniBatchKMeans algorithm from `scikit-learn`, with a mini-batch size of 10,000 frames. Initialization is performed with k-means++ using 20 random starts to ensure stability, and cluster assignments are set to $k = 256$.

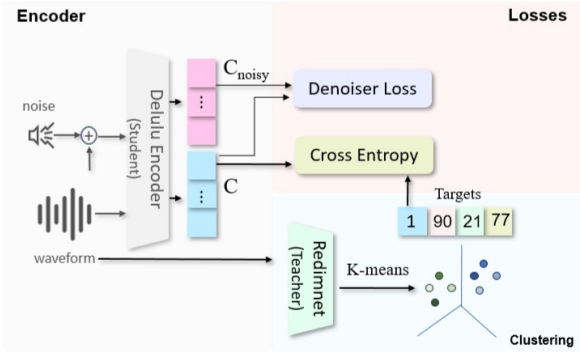

Figure 1: Illustration of DELULU architecture. The student encoder is trained with (a) a denoising objective aligning clean and noisy features, and (b) a masked prediction objective, where frame-level features from the teacher (ReDimNet) are clustered into discrete speaker-aware targets for cross-entropy training.

Pre-training is conducted on 960 hours of LibriSpeech audio using 4 NVIDIA H100 GPUs, with each GPU processing 87.5 seconds of audio per batch. The model is trained for a total of 400k updates. Optimization follows the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.98$) with an initial learning rate of 5e-4. A linear warmup is applied for the first 32k steps, followed by a polynomial decay back to zero. To regularize training, we apply a weight decay of 0.01 and clip gradients at a maximum norm of 10.0.

Computational complexity. Compared to the HuBERT baseline, which requires approximately 23 minutes per epoch on our setup, the introduction of the denoising objective increases training time to about 38 minutes per epoch. Since the denoising loss is applied only during training, inference cost remains identical to HuBERT.

5 Ablation Study

To quantify the contribution of each component in DELULU's design, we perform systematic ablation experiments on the VoxCeleb1-O dataset, evaluating upstream speaker verification performance using EER, with all models trained for 400k updates on LibriSpeech 960h under identical optimization settings for fair comparison. Examining clustering feature sources, MFCC features with $k = 100$ yield a baseline EER of 37.73%,

Features	Clusters	Stride	EER (%)	Rel. Impr.
MFCC	100	20	37.73	-
HuBERT (Stage 2)	500	20	34.05	9.8%
ReDimNet	256	16	**13.53**	**60.2%**
ReDimNet	500	16	14.16	58.4%
ReDimNet	1024	16	14.16	58.4%
ReDimNet	256	15	14.16	58.4%

Table 1: Ablation study on VoxCeleb1-O upstream speaker verification (EER %). Lower is better. Results demonstrate that ReDimNet-guided clustering is the primary factor in DELULU's performance, with optimal cluster size k=256 and stride=16 for temporal alignment.

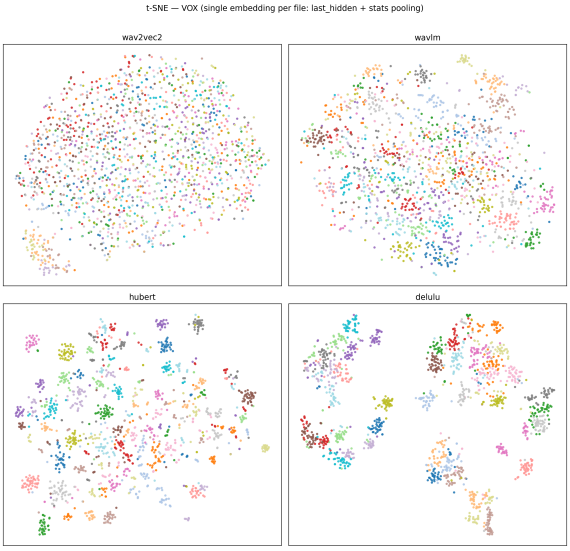

Figure 2: t-SNE visualization of embeddings from 40 VoxO1 speakers. DELULU yields compact, well-separated clusters with lower intra-speaker variability and higher inter-speaker separation than baseline SSL models.

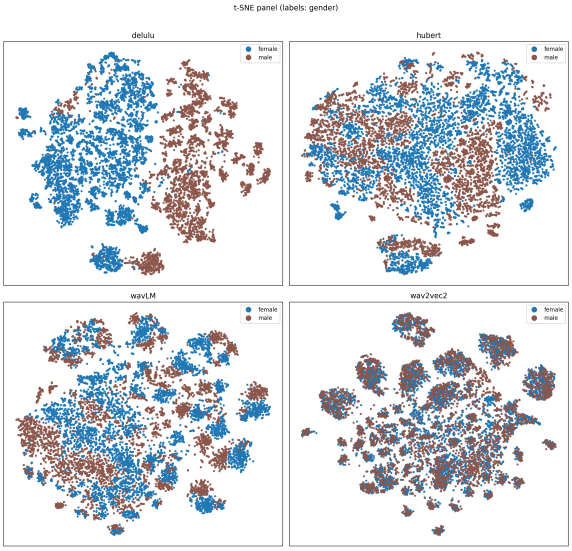

Figure 3: t-SNE visualization of speaker embeddings by gender on the EARS dataset. DELULU shows clear speaker clusters and strong gender separation.

while higher-level acoustic representations from a pretrained HuBERT model with $k = 500$ improve it to 34.05%, indicating modest gains from context-aware features; in contrast, frame-level ReDimNet embeddings (dimension 2304) with $k = 256$ reduce EER to 13.53%, a 60% relative improvement over HuBERT, confirming that external speaker-discriminative supervision during clustering drives DELULU's performance. For cluster size using ReDimNet features, we evaluate $k \in 256, 500, 1024$, obtaining EERs of 13.53%, 14.16%, and 14.16% respectively, showing $k = 256$ offers the optimal trade-off between granularity and generalization, as 256 discrete speaker-aware units suffice for identity variation at this scale. We also varied the DELULU encoder's stride, to find the optimal stride. Since the training paradigm requires frame-level synchrony between the teacher and student, the RedimNet stride was modified to be identical to that of the encoder. A stride of 16ms achieves 13.52% EER, while both lower (15ms or less) and higher (20ms or greater) strides resulted in EERs greater than 14%, hence we chose 16ms as our final stride. Overall, these ablations highlight speaker-discriminative clustering as the key factor, with ReDimNet-guided pseudo-labels providing a 60% relative improvement over HuBERT's acoustic-only approach, optimal $k = 256$ balancing discriminability and generalization, and stride=16 ensuring effective supervision, validating the introduction of external speaker supervision into pseudo-label generation as a powerful strategy for speaker-aware representations.

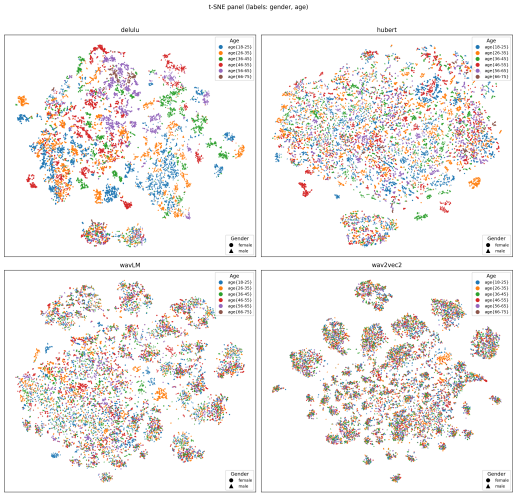

Figure 4: t-SNE visualization of EARS embeddings by gender and age groups (Table 3). DELULU clusters primarily by speaker identity, with demographic traits forming secondary structure and no apparent bias.

6 Evaluation

We evaluate DELULU across multiple speaker-centric benchmarks described below to assess its effectiveness in capturing speaker-discriminative information. To verify our hypothesis that, as a more speaker-aware foundation model, DELULU may be expected to result in representations that are naturally better organized by speaker characteristics, and deliver better performance both in zero-shot "upstream" settings where they are used as-is, and in "downstream" settings where the model is further fine-tuned to the task, we perform the evaluations (comparing it to other baseline models) in both settings, with a primary focus on speaker verification for the downstream task. We also visualize the derived repersentations for further confirmation of our hypothesis. We report on these tests below.

For fair comparison, all baseline models (wav2vec 2.0, HuBERT, and WavLM) are pre-trained on the same 960 hours of LibriSpeech data, ensuring that differences in performance arise from architectural and training objective choices rather than data scale.

6.1 Upstream Evaluation: Speaker Verification

We first evaluate the models in an upstream setting, where representations are extracted directly from the pretrained encoders without any fine-tuning. This zero-shot evaluation reveals the inherent speaker-discriminative properties encoded during self-supervised pretraining.

Protocol. We evaluate on two widely-used speaker verification benchmarks: VoxO1 (Nagrani et al., 2017) and SITW (McLaren et al., 2016) (see Appendix A). For each model, we extract utterance-level embeddings by mean pooling over the temporal dimension of final layer representations. Verification trials are scored using cosine similarity, and we report EER, where lower is better.

Results. Table 2 shows that DELULU achieves substantial improvements over all baseline SSL models. On VoxO1, DELULU obtains an EER of 13.53%, representing a 62% relative improvement over HuBERT (34.05%) and a 60% improvement over WavLM (35.93%). Similar gains are observed on SITW, where DELULU achieves 25.40% EER compared to HuBERT's 42.60% and WavLM's 44.00%. These results demonstrate that ReDimNet guided clustering successfully embeds strong speaker-discriminative structure into the learned representations.

6.2 Downstream Evaluation: Speaker Verification

To assess transferability beyond upstream frozen representations, we evaluate **DELULU** on a downstream speaker verification task using VoxCeleb1-O. Since our objective is to evaluate DELULU's transferability rather than optimize for maximum downstream performance, we employ a minimal fine-tuning architecture that adds only a simple classification head on top of the frozen encoder, as described in Appendix B and then compute EER to measure verification accuracy.

As shown in Table 4, DELULU achieves 5.63% EER with $k = 256$ clusters, representing a 24.5% relative improvement over HuBERT Stage 2 (7.45%) and a 56.7% gain over the MFCC baseline (13.00%). Larger vocabularies ($k = 500, 1024$) yield marginal improvements, confirming that $k = 256$ provides the optimal balance between discriminability and generalization. These results demonstrate that the speaker-discriminative structure learned during pretraining transfers effectively to supervised fine-tuning, further validating DELULU as a strong foundational encoder for speaker verification.

6.3 Demographic Subgroup Analysis

To understand how speaker-discriminative information is distributed across demographic groups fol-

Dataset	WavLM	Hubert	DELULU	Wav2Vec2
VoxO1	35.93	34.05	**13.53**	43.17
SITW	44.00	42.60	**25.40**	42.20

Table 2: EER (%) across upstream speaker verification benchmarks (↓ better)

Category	Subgroup	WavLM	Hubert	DELULU	Wav2Vec2
Gender	Female (59 spks)	41.69	41.06	**28.11**	45.60
	Male (43 spks)	40.76	41.47	**28.54**	44.70
Age	F (18–25), 13 spks	41.93	39.67	**31.01**	45.03
	F (26–35), 13 spks	40.60	41.61	**30.90**	43.90
	F (36–45), 7 spks	43.37	44.26	**29.38**	47.40
	F (46–55), 14 spks	42.17	40.48	**29.56**	45.78
	F (56–65), 10 spks	42.71	42.02	**31.28**	47.78
	F (66–75), 2 spks	43.66	42.25	**40.12**	48.32
	M (18–25), 14 spks	41.55	43.06	**35.01**	45.87
	M (26–35), 10 spks	40.51	40.60	**30.90**	42.97
	M (36–45), 10 spks	39.40	39.47	**24.53**	45.35
	M (46–55), 4 spks	41.27	42.55	**29.63**	44.25
	M (56–65), 5 spks	42.53	43.32	**31.37**	47.53

Table 3: EER (%) across demographic subgroups in EARS dataset on upstream speaker verficiation (↓ better)

lowing the SVERITAS (Baali et al., 2025a) benchmark, we analyze model performance on the EARS (Richter et al., 2024) dataset (details in Appendix C), stratified by gender and age subgroups.

Protocol. Following the upstream protocol, we extract mean-pooled embeddings and compute EER for speaker verification trials within each demographic subgroup. This analysis reveals whether models learn speaker representations uniformly across age and gender categories or exhibit demographic biases.

Results. Table 3 presents EER across gender and age subgroups. DELULU consistently outperforms baselines across all demographics, with particularly strong improvements for middle-aged speakers (36-55 years). For male speakers aged 36-45, DELULU achieves 24.53% EER compared to HuBERT's 39.47%. The consistent improvements across subgroups indicate that DELULU's speaker-discriminative representations generalize effectively across demographic variations without introducing systematic biases.

6.4 Zero-Shot Speaker Profiling Tasks

Beyond verification, we assess **DELULU**'s capacity to encode diverse speaker attributes through zero-shot evaluation on multiple profiling tasks from the DynamicSUPERB benchmark (Huang et al., 2024). We evaluate six speaker-related tasks, including age, gender, accent, speaker counting, and spoof detection (details in Appendix D).

Protocol. We extract layer-wise representations from each model and train K-Nearest Neighbors (KNN) classifiers without any fine-tuning. All evaluations are conducted in a fully zero-shot setting; no supervised training is used for adaptation of the models. Given that these models are trained without explicit supervision, we expect performance to be poor, possibly approaching random levels; however, all SSL-based models achieve meaningful results, with DELULU consistently outperforming other models. For each task, we perform k-fold cross-validation and report Macro-F1 scores averaged across folds, along with standard deviations to indicate variability and robustness.

Results. Table 5 summarizes the zero-shot results across six profiling tasks. DELULU achieves the highest performance on all tasks, considerably outperforming other SSL models in zero-shot settings. The second-best model varies by task—for instance, WavLM performs competitively on gender classification, while HuBERT fares better on accent detection. On spoof detection, DELULU achieves 57.20% F1 compared to HuBERT's 53.51% and WavLM's 51.44%; on accent detection, it reaches 78.38% F1, surpassing HuBERT (62.86%) and

Features	Clusters	EER (%)	Rel. Impr.
MFCC	100	13.00	–
HuBERT (Stage 2)	500	7.45	–
ReDimNet	256	**5.63**	**24.5%**
ReDimNet	500	6.28	15.7%
ReDimNet	1024	5.99	19.6%

Table 4: EER (%) across downstream speaker verification (VoxO1) (↓ better)

Task	WavLM	Hubert	DELULU	Wav2Vec2
Spoof Detection	51.44	53.51	**57.20**	52.88
Speaker Count	62.71	64.83	**67.13**	64.20
Accent Detection	77.76	62.86	**78.38**	58.60
HowFarSpk	69.37	70.58	**73.36**	57.84
Gender Detection	95.75	93.97	**96.18**	92.73
Age Estimation	32.69	29.43	**36.00**	31.99

Table 5: Zero-shot Macro-F1 score (%) across profiling tasks (↑ better).

wav2vec 2.0 (58.60%). Even on more challenging tasks such as age estimation (36.00% F1), DELULU exhibits clear gains despite the absence of task supervision. These findings demonstrate that the speaker-guided clustering in DELULU enables robust and discriminative representations that generalize across diverse zero-shot profiling scenarios.

6.5 Representation Analysis

We analyze the learned embedding space using t-SNE visualizations to assess whether DELULU produces structured and discriminative speaker representations. In Figure 2, we expect embeddings from the same speaker to form compact clusters, distinct from those of other speakers. Indeed, DELULU's representations exhibit clear and well-separated speaker clusters, indicating strong identity preservation and reduced intra-speaker variability compared to baseline models. In Figure 3, we expect representations to separate naturally by gender. The visualization confirms this behavior, showing distinct groupings corresponding to male and female speakers, while maintaining tight clustering within each group. Finally, in Figure 4, we examine whether age-related patterns emerge within the embedding space. As expected, DELULU produces smooth transitions across age groups, suggesting that the model implicitly encodes demographic cues such as vocal maturity and pitch characteristics alongside dominant speaker identity. These results demonstrate that DELULU learns a representation space that reflects both speaker individuality and meaningful demographic structure, confirming the model's capacity for fine-grained, interpretable speaker encoding.

7 Conclusion

In this work, we introduce **DELULU**, a speaker-aware self-supervised foundational model that enhances speaker representation learning by guiding the clustering process with frame-level embeddings from ReDimNet, a SOTA speaker verification network. Trained using the dual objectives of masked prediction and denoising, DELULU learns representations that are both content-rich and identity-aware.

Across benchmarks in speaker verification, profiling, and counting, DELULU consistently outperforms existing SSL models, establishing it as a strong universal encoder for speaker-centric speech processing. Since the core idea lies in providing a relevant external reference to guide pseudo-label generation, we expect this strategy to generalize to other foundational model architectures and to yield gains across a broader range of downstream tasks. Future work will explore replacing clustering with direct distillation-based objectives to further simplify and strengthen the pretraining pipeline.

Limitations

While DELULU demonstrates substantial improvements in speaker verification, profiling, and

zero-shot speaker-aware tasks, several limitations warrant further investigation.

Extending DELULU to large-scale, multi-domain datasets with higher speaker and environmental variability remains an open challenge. Also, incorporating an external model (ReDimNet) into the clustering process introduces additional computational overhead and memory requirements, potentially limiting accessibility for research groups with restricted computational resources. Finally, while DELULU excels in identity-aware modeling, its downstream adaptability to non-speaker-centric applications, such as emotion or intent recognition, has yet to be comprehensively explored. Future work will focus on addressing these limitations through more efficient training strategies, broader data coverage, and cross-domain generalization studies.

Ethics Statement

The development of DELULU is guided by a strong commitment to ethical responsibility and privacy preservation. As a speaker-aware foundational model, DELULU possesses the potential to be misused for surveillance, impersonation, or identity inference without consent. We emphasize that its use must strictly comply with ethical research standards, data protection laws, and institutional review protocols. All datasets employed in this study (LibriSpeech, VoxCeleb, Dynamic-Superb, and SVeritas) consist of publicly available and consented speech samples intended solely for research purposes. No personally identifiable or private data was used or generated. Furthermore, while DELULU improves representation learning for speaker-related tasks, it does not perform speaker identification on unconsented audio, nor is it intended for forensic or monitoring applications. In line with the ACL Ethics Policy, we advocate for transparent deployment of DELULU, ensuring that its advancements contribute positively to responsible and equitable speech technology research.

References

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, 20(2):356–370.

Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. 2024. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10166–10170. IEEE.

Massa Baali, Abdulhamid Aldoobi, Hira Dhamyal, Rita Singh, and Bhiksha Raj. 2024. PDAF: A phonetic debiasing attention framework for speaker verification. In *SLT*.

Massa Baali, Sarthak Bisht, Francisco Teixeira, Kateryna Shapovalenko, Rita Singh, and Bhiksha Raj. 2025a. SVeritas: Benchmark for robust speaker verification under diverse conditions. In *Findings of the Association for Computational Linguistics: EMNLP*.

Massa Baali, Shuo Han, Syed Abdul Hannan, Purusottam Samal, Karanveer Singh, Soham Deshmukh, Rita Singh, and Bhiksha Raj. 2025b. CoLMbo: Speaker language model for descriptive profiling. In *ASRU*. IEEE.

Massa Baali, Xiang Li, Hao Chen, Syed Abdul Hannan, Rita Singh, and Bhiksha Raj. 2025c. CAARMA: Class augmentation with adversarial mixup regularization. In *Findings of the Association for Computational Linguistics: EMNLP*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. 2023. Joint prediction and denoising for large-scale multilingual self-supervised learning. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*.

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*, 29:3451–3460.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and 1 others. 2024. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP*, pages 12136–12140. IEEE.

Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. 2020. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991*.

Theo Lepage and Reda Dehak. 2025. Self-supervised frameworks for speaker verification via bootstrapped positive sampling. *arXiv preprint arXiv:2501.17772*.

Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The speakers in the wild (sitw) speaker recognition database. In *Interspeech*, pages 818–822.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*, page 2616.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. 2022. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International conference on machine learning*, pages 18003–18017. PMLR.

Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*.

Rita Singh and Bhiksha Raj. 2025. Human voice is unique. *arXiv preprint arXiv:2506.18182*.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333. IEEE.

Abdul Waheed, Hanin Atwany, Bhiksha Raj, and Rita Singh. 2024. What do speech foundation models not learn about speech? *arXiv preprint arXiv:2410.12948*.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. 2024. Reshape dimensions network for speaker recognition. In *Proc. Interspeech 2024*, pages 3235–3239.

Ziyang Zhang, Wu Guo, and Bin Gu. 2023. Introducing self-supervised phonetic information for text-independent speaker verification. In *Proc. Interspeech*, pages 4698–4702.

A Upstream Evaluation

SV Benchmarks

- **VoxCeleb1-O** (Nagrani et al., 2017): The original VoxCeleb1 test set containing 37,720 verification trials across 40 speakers.
- **SITW** (McLaren et al., 2016): Speakers in the Wild, a challenging dataset with 6,445 trials featuring diverse acoustic conditions and speaking styles.

B Downstream Model

Model Architecture. Our downstream model consists of the pretrained SSL encoder (frozen) followed by temporal mean pooling and an L2 normalization layer. A single linear embedding head projects the pooled representations to speaker logits for classification. This minimal design isolates the quality of the pretrained representations by limiting the capacity of the downstream classifier.

Training Setup. We train the embedding head on VoxCeleb development set using cross-entropy loss with the AM-Softmax objective (Wang et al., 2018). The SSL encoders remain frozen throughout training, ensuring that performance differences reflect the quality of pretrained features rather than fine-tuning capacity. We train for 30 epochs with a batch size of 32 and learning rate of 1e-3.

C Demographic Subgroup Analysis

Description of the EARS Dataset. The EARS (Expressive Anechoic Recordings of Speech) dataset contains speakers spanning ages 18-75, with balanced gender representation. We evaluate on 102 speakers (59 female, 43 male) across six age brackets.

D Zero-Shot

Profiling Tasks.

- **Spoof Detection**: Binary classification task detecting synthesized or manipulated speech versus genuine human speech.
- **Speaker Counting**: Predicting the number of unique speakers in an audio segment.
- **Accent Detection**: Classifying the regional accent of the speaker.The dataset includes nine distinct distance classes.
- **HowFarSpk**: Determining the spatial distance between speaker and microphone. The dataset includes three distinct distance classes.
- **Gender Recognition**: Binary classification of speaker gender.
- **Age Classification**: Multi-class prediction of speaker age group.