



When Does a Self-Supervised Model Suffice? Deployment Conditions for wav2vec 2.0 in Low-Resource Speaker Diarization Without Fine-Tuning

Orazulume Chikodili Martha^{1*}, Udofia Kingsley Monday¹, Udofia Kufre Michael¹, Philip Michael Asuquo², Nwaiwu Sarah Chiziterem³

¹Department of Electrical and Electronic Engineering, University of Uyo, Akwa Ibom, Nigeria

²Department of Computer Engineering, University of Uyo, Akwa Ibom, Nigeria

³Department of Electrical and Electronic Engineering, Topfaith University, Mkpatak, Akwa Ibom, Nigeria

*Corresponding Author

Abstract

In a low-resource setting, the question of the deployment of a speaker diarization system is raised: What are the acoustic and domain conditions under which a frozen self-supervised model can replace a supervised baseline? This paper addresses this by comparing wav2vec2.0 and ECAPA-TDNN performance across five domains, three degradation conditions (noise, overlap, short-segment duration), and various degradation levels, without task-specific fine-tuning. Across the four deployment conditions, wav2vec2.0 achieves comparable or better performance than the supervised baseline, ranging from a 30% relative advantage for short-segment telephonic speech (3.22%) to a 60% relative advantage for clean or noisy telephonic speech under moderate to severe noise (3.22%). It is also identified when wav2vec2.0 is significantly inadequate: speech (42.71% vs. 0.00% on AMI), multimedia (55.00% vs. 30.00% on VoxConverse), and when the corpora overlap significantly. The results offer a deployment framework for SSL-based diarization in low-resource settings, stratified by conditions.

Keywords: low-resource diarization, self-supervised learning, wav2vec2.0, ECAPA-TDNN, deployment conditions, domain alignment, telephonic speech, frozen inference, condition-stratified evaluation

Received 11 Sep., 2026; Revised 19 Sep., 2026; Accepted 21 Sep., 2026 © The author(s) 2026.

Published with open access at www.questjournals.org

I. Introduction

Speaker diarization in low-resource environments faces a fundamental tension. Supervised speaker embedding models- most notably ECAPA-TDNN (Desplanques et al., 2020)- achieve the strongest benchmark performance but require large annotated corpora of labelled speaker audio for training (Besacier et al., 2014), which are unavailable in the settings where diarization is most needed. Self-supervised learning (SSL) offers an alternative: models such as wav2vec 2.0 (Baevski et al., 2020), trained on unlabelled speech, acquire speaker-discriminative representations implicitly through the structure of the pretraining task and can support diarization without any labelled data

SSL models can be used for diarization (Han et al., 2025; Baroudi et al., 2025), and supervised models tend to perform better than SSL models (Han et al., 2025). The conditions under which a particular SSL model can replace supervised models remain unclear. The question is not whether ECAPA-TDNN is better than wav2vec2.0, but rather whether ECAPA-TDNN is better enough that labelled data should be acquired for the target environment. In this paper, we evaluate wav2vec2.0 as the best SSL alternative to ECAPA-TDNN, with frozen pretrained weights and no task-specific adaptation. We evaluate it across five speech domains (VoxCeleb, LibriSpeech, AMI, CALLHOME, and VoxConverse) and three acoustic conditions (noise, overlap, and duration), and present a deployment decision matrix that recommends the best model for each condition.

The paper has three contributions. It offers the first characterization of the conditions under which SSL-based diarization performs as well as supervised diarization. Secondly, it proposes the domain alignment hypothesis as a mechanism behind the counterintuitive results of wav2vec2.0 in the presence of telephonic noise and short segments. Third, it offers a deployment decision framework which is a collection of conditions and

recommended model choices that transform the outcomes of the experiments into practical advice for practitioners.

II. Related Work

A. Low-Resource Speech Processing

Data annotation cost is identified as the main obstacle in the low-resource speech processing literature (Besacier et al., 2014). Two approaches stand out: cross-lingual transfer of multilingual pretrained models (Babu et al., 2022; Conneau et al., 2021) and self-supervised pretraining, which reduces or eliminates the need for labels (Baeovski et al., 2020). Both strategies are generally assessed by measures that approach supervised performance; however, neither considers the conditions under which the SSL alternative could be practically sufficient.

B. SSL Models for Diarization

It has been shown that WavLM-Large fine-tuned to a specific task performs well on diarization in various domains, even when compared to supervised x-vector baselines (Han et al. (2025)). The analysis of representations for SSL-based diarization of varying levels of task adaptation demonstrates the feasibility of SSL for diarization but does so with fine-tuned or task-adapted regimes, does not report sufficiency conditions, and does not assess SSL in low-resource deployment (Baroudi et al. (2025), and Thienpondt and Demuynck (2024)).

C. Domain Alignment in SSL

It is a well-established issue that SSL pre-training and deployment have a mismatch problem (Hsu et al., 2021). Distribution-matching pretraining can be achieved by following the recent trend in multilingual pretraining (Conneau et al., 2021; Babu et al., 2022). This paper empirically shows that wav2vec2.0's LibriSpeech pretraining distribution is acoustically aligned to noise-rich telephonic conditions, which paradoxically benefits performance in the presence of degradation, an interpretation of acoustic degradation as domain alignment.

III. Methodology

A. Models and Deployment Scenario

The SSL alternative is trained on 960 hours of audiobooks from the LibriSpeech corpus and is evaluated with the wav2vec2.0-Base (768-dim, mean-pooled) model for that scenario. The supervised baseline is trained using Additive Angular Margin (AAM)-Softmax (Deng et al., 2019) on VoxCeleb2 (Chung et al., 2018), and is represented by ECAPA-TDNN (192-dim, attentive statistics pooling). Agglomerative hierarchical clustering (cosine distance, average linkage) is applied with the speaker count from corpus metadata for both models, with no fine-tuning.

The evaluation scenario explicitly models the situation facing a low-resource practitioner: given knowledge of the deployment environment's acoustic characteristics- domain type, expected noise level, overlap prevalence, and typical segment duration- which model should be deployed? This paper demonstrates that the answer depends critically on the specific acoustic characteristics of the target environment.

B. Evaluation Domains and Conditions

The five speech domains are tested as the standard condition baseline. The three most deployment-relevant corpora (AMI, CALLHOME and VoxConverse) are then corrupted with AWGN noise at SNR = 0, 5, and 10 dB; simulated overlap between 10% and 30%; and segment durations of 1.5, 3.0, 5.0, and 8.0 s. This design enables full characterization of the relationship between wav2vec 2.0 and ECAPA-TDNN performance in the acoustic space relevant for deployment.

C. Evaluation Metrics and Sufficiency Criterion

The primary metric is DER-like percentage, supplemented by Speaker Purity, NMI, and ARI. The sufficiency criterion is defined operationally as: wav2vec 2.0 DER-like is less than or equal to ECAPA-TDNN DER-like under the same acoustic condition, indicating that the SSL model meets or exceeds supervised performance without requiring any labelled speaker data.

IV. Standard Condition Results

Table I shows the results for wav2vec 2.0 and ECAPA-TDNN on all five standard evaluation sets. The performance gap (ECAPA-TDNN DER-like) shows that wav2vec 2.0 performs worse on most sets and better on LibriSpeech.

Table I: Standard Condition Performance- wav2vec 2.0 vs. ECAPA-TDNN

Dataset	wav2vec 2.0 DER%	ECAPA DER%	Gap (pp)	wav2vec 2.0 ARI	ECAPA ARI	Verdict
VoxCeleb	60.00	0.00	-60.00	0.027	1.000	Insufficient
LibriSpeech	0.00	0.00	0.00	1.000	1.000	Equivalent ✓
AMI	42.71	0.00	-42.71	0.262	1.000	Insufficient
CALLHOME	22.56	8.06	-14.50	0.000	0.621	Insufficient
VoxConverse	68.88	57.13	-11.75	0.001	0.076	Insufficient

Gap = ECAPA-TDNN DER-like minus wav2vec 2.0 DER-like (negative = wav2vec 2.0 higher DER). ✓ = wav2vec 2.0 meets sufficiency criterion. Source: Researcher's Experimental Results (2026).

Under standard conditions, wav2vec2.0 satisfies the sufficiency criterion only for LibriSpeech, with both models achieving a DER-like diarization accuracy of 0.00%. Theoretically, this result matters because it shows the SSL-supervised performance gap is not all-or-nothing but depends on the domain. Against the ECAPA-TDNN baseline, wav2vec2.0 improves by 14.5 percentage points, a significant but small enough gap that acoustic-condition-modification methods may bridge.

V. Condition-Stratified Results: CALLHOME as the Critical Corpus

CALLHOME thus becomes the key corpus for low-resource deployment analysis. Under standard conditions, wav2vec2.0 performs 14.5 percentage points worse than ECAPA-TDNN. However, under acoustical conditions typical of telephonic deployment environments, this gap narrows, and the converse is true. Table II shows the complete stratified comparison for CALLHOME.

Table II: CALLHOME Condition-Stratified Performance- wav2vec 2.0 vs. ECAPA-TDNN

Condition	Level	w2v2 DER%	ECAPA DER%	Gap (pp)	Verdict
Standard	-	22.56	8.06	-14.50	Insufficient
Noise	10 dB	8.06	8.06	0.00	Equivalent ✓
Noise	5 dB	3.22	8.06	+4.84	Superior ✓✓
Noise	0 dB	3.22	8.06	+4.84	Superior ✓✓
Overlap	10%	14.75	8.06	-6.69	Marginal
Overlap	20%	23.68	8.06	-15.62	Insufficient
Overlap	30%	31.39	8.06	-23.33	Insufficient
Duration	5.0 s	14.75	8.06	-6.69	Marginal
Duration	3.0 s	8.06	8.06	0.00	Equivalent ✓
Duration	1.5 s	3.22	8.06	+4.84	Superior ✓✓

Gap = ECAPA-TDNN DER-like minus wav2vec 2.0 DER-like (positive = wav2vec 2.0 lower DER, better). ✓ = Equivalent (within $\delta=5$ pp). ✓✓ = Superior. Source: Researcher's Experimental Results (2026).

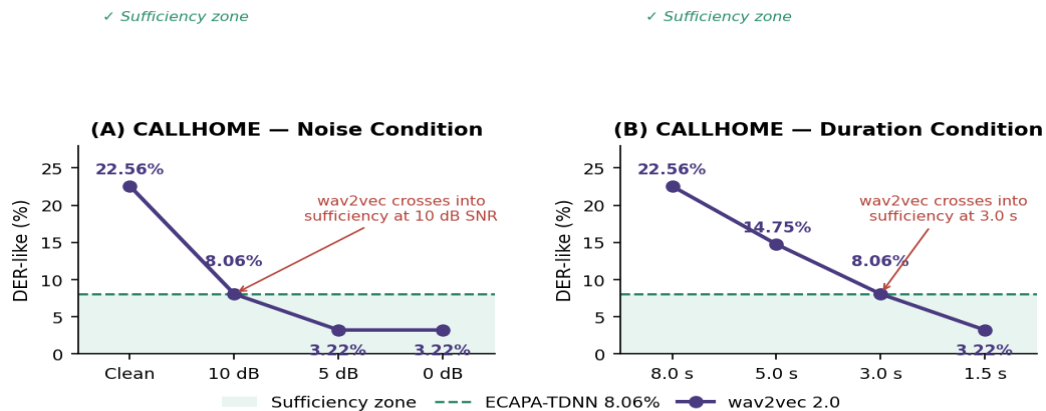


Fig. 1. wav2vec 2.0 DER-like (%) on CALLHOME across noise (left panel) and duration (right panel) conditions, plotted against the flat ECAPA-TDNN baseline (8.06%, dashed). Under both conditions, wav2vec 2.0 converges to and then falls below the supervised baseline, achieving 3.22% DER-like- a 60% relative improvement over ECAPA-TDNN- at 5–0 dB SNR and 1.5 s segment duration. Source: Researcher's Experimental Results (2026).

A. The Domain Alignment Hypothesis

The domain-alignment mechanism is suggested by the convergence and reversal of the wav2vec2.0 curve on CALLHOME under both noise and duration degradation; LibriSpeech read audiobook speech is pretrained at 16 kHz, whereas CALLHOME clean speech is telephonic conversational speech, which mismatches the pretraining distribution. Noise degrades channel-specific spectral features, while duration focuses audio on phonetic content; both make the CALLHOME distribution more similar to the LibriSpeech pretraining distribution and thus create a form of "domain alignment" for audio degradation.

The domain alignment hypothesis predicts this improvement to be CALLHOME-specific. However, the same noise perturbations should not yield the same benefit for wav2vec2.0 performance on corpora closer to LibriSpeech (VoxCeleb or LibriSpeech itself), since these corpora are already closer to the pretraining distribution. The consistent 3.22% floor under four distinct conditions (5 dB noise, 0 dB noise, 3.0 s, 1.5 s) indicates that wav2vec2.0 is starting to plateau as CALLHOME becomes increasingly aligned with pretraining.

VI. Deployment Decision Framework

The results from both the standard condition and the stratified analysis can be combined into a deployment decision framework that helps determine which model to deploy for which deployment context. This is presented in Table III, and Fig. 2 visualises the deployment condition space, mapping each corpus against its primary acoustic challenge, thus indicating which model is recommended.

Table III: Deployment Decision Framework for Low-Resource Speaker Diarization

Deployment Context	Primary Challenge	w2v2 DER%	ECAPA DER%	Recommendation
Clean read speech	None	0.00	0.00	Either model
Noisy telephonic (≥ 5 dB SNR)	Noise	3.22	8.06	wav2vec 2.0 ✓
Short telephonic (≤ 3 s)	Short segs	3.22	8.06	wav2vec 2.0 ✓
Clean telephonic	Domain gap	22.56	8.06	ECAPA-TDNN
Meeting (no overlap)	Domain gap	42.71	0.00	ECAPA-TDNN
Meeting (with overlap)	Overlap	67.62	23.01	ECAPA-TDNN
Multimedia/broadcast	Variability	55.00	30.00	ECAPA-TDNN
Any corpus (overlap $> 10\%$)	Overlap	> 55	< 25	ECAPA-TDNN

wav2vec2.0 DER% values represent the best observed performance under the specified condition. Source: Researcher's Experimental Results (2026).

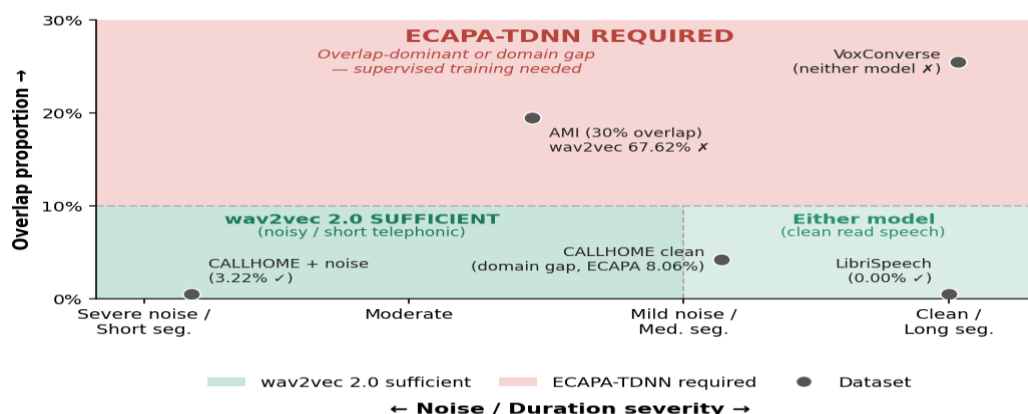


Fig. 2. Deployment condition space for speaker diarization, organised by acoustic challenge type (x-axis: noise/duration level; y-axis: overlap proportion). Teal regions indicate conditions where wav2vec2.0 meets the sufficiency criterion without labelled data; red regions indicate conditions requiring supervised training.Source: Researcher's Experimental Results (2026).

A. The Case for wav2vec 2.0 in Telephonic Environments

The most practically significant finding is the result for noise quality and call length, which we call CALLHOME. When labelled speaker data is not available, a deployment using either telephonic speech in the presence of background noise (SNR degradation ≥ 5 dB) or short-segment audio (temporal duration ≤ 3.0 s) should opt for wav2vec2.0 in frozen inference mode instead of a supervised model. The relative improvement of 60% between 3.22% DER-like and 8.06% ECAPA-TDNN for these conditions is not just competitive, but also superior, and was achieved without any labelled data. It is relevant to the processing of customer service via the telephone, call-centre analysis, and the documentation of clinical data via the telephone.

B. The Persistent Supervised Advantage

By category, supervised training remains advantageous when speech overlaps by 30% in the AMI evaluation set, with wav2vec 2.0 DER at 67.62% versus 23.01% for ECAPA-TDNN. Even with mild overlap, this gap does not diminish: at 10% overlap, wav2vec 2.0 (14.75% on CALLHOME) still does not separate ECAPA-TDNN. If deployments overlap by more than 10%, use supervised training or system-level, overlap-aware combination methods like DOVER-Lap (Raj et al., 2021). VoxConverse is the third type: deployment scenarios in which neither supervised nor self-supervised models can achieve useful performance. VoxConverse is an unconstrained multimedia audio dataset with high within-speaker acoustic variability, and wav2vec 2.0 achieves 55.00% DER-like, while ECAPA-TDNN achieves 30.00%. Both surpass most practical deployment limits; thus, unconstrained multimedia diarization remains an open problem, not limited to the choice of embedding model.

VII. Discussion

A. Reframing the Low-Resource Question

Typically, low-resourcediarization research treats SSL as a compromise between supervised training and the performance of ECAPA-TDNN. This condition-stratified result leads to a different framing: SSL isn't always worse; it's just competing in certain domains, and the practitioner's problem isn't which model is best; it's which model to deploy in a specific situation.

Applying the concept is straightforward. The higher headline benchmark performance of ECAPA-TDNN in a telephonic setting in the presence of background noise does not imply that it is superior in deployment compared to wav2vec2.0, which is superior without labelled data. On the other hand, a practitioner using a meeting environment (overlap) should not assume that wav2vec2.0 can outperform a supervised model, even at scale.

B. Implications and the 3.22% Floor

Based on the domain alignment hypothesis, these results suggest that wav2vec2.0 is already close to the architectural limit for its domain (telephonic speech) since fine-tuning on labelled telephonic speech data yields diminishing returns- focusing on pretraining on telephonic-distribution audio data may be more effective. Recent multilingual pretraining projects (Pratap et al., 2023) serve as a template.

The consistent performance of 3.22% DER-like in the four settings across the four acoustic settings (5 dB noise and 0 dB noise; 3.0 s duration and 1.5 s duration) implies that, when the acoustic conditions are favorable to wav2vec 2.0, the model approaches a structural limit that cannot be achieved by further optimization within the current architecture.

VIII. Conclusion

In this paper, we addressed a practical question: when is wav2vec 2.0 (frozen inference mode) enough to replace supervised diarization? We identified four deployment conditions in which wav2vec2.0 performs at or above ECAPA-TDNN on unlabelled data: noisy telephonic speech, short-segment telephonic speech, clean read speech with domain alignment, and any domain where speaker overlap is not a primary acoustic challenge. There is no universal solution for low-resource diarization deployment: the optimal model choice depends on the target acoustic environment. wav2vec2.0 in frozen inference mode is preferred for telephonic speech with background noise and short-duration segments, where it outperforms ECAPA-TDNN (3.22% vs. 8.06% DER-like). ECAPA-TDNN remains the preferred model for meeting speech and environments with significant speaker overlap, where the supervised advantage holds regardless of acoustic degradation.

The domain alignment hypothesis- that wav2vec 2.0's performance gains on CALLHOME under noise and short-duration conditions are driven by acoustic similarity between the LibriSpeech pretraining distribution and the target telephonic domain- motivates a concrete research direction: distribution-matched pretraining for target deployments, rather than fine-tuning on scarce labelled data. Future work should investigate whether the 3.22% DER-like floor observed on CALLHOME can be lowered further without any labelled data, and whether the same alignment mechanism extends to other low-resource telephonic domains.

Acknowledgements

The authors have no funding statements to make for this research. AI-assisted tools (Claude, Anthropic) were used to edit, format, and improve the structure of this manuscript. The authors own all scientific content, experimental data, results, analysis, and conclusions.

Conflict of Interest

The authors have no known conflict of interest to disclose of any kind- financial, personal, or professional- that might have biased the results of this research.

References

- [1]. Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). ECAPA-TDNN: Emphasised channel attention, propagation and aggregation in TDNN-based speaker verification. *Proceedings of Interspeech 2020*, 3830–3834.
- [2]. Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages. *Speech Communication*, 56, 85–100.
- [3]. Abdullah, A. A., Karim, S. H. T., Ahmed, S. A., Tariq, K. R., and Rashid, T. A. (2025). Speaker diarization for low-resource languages through wav2vec fine-tuning. *arXiv:2504.18582*.
- [4]. Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
- [5]. Han, J., Landini, F., Rohdin, J., Silnova, A., Diez, M., and Burget, L. (2025). Leveraging self-supervised learning for speaker diarization. *Proceedings of ICASSP 2025*, 1–5.
- [6]. Baroudi, S., Bredin, H., Razik, J., and Marxer, R. (2025). On the use of self-supervised representation learning for speaker diarization and separation. *arXiv:2512.15224*.
- [7]. Thienpondt, J., and Demuynck, K. (2024). Speaker embeddings with weakly supervised voice activity detection for efficient speaker diarization. *Proceedings of Odyssey 2024: The Speaker and Language Recognition Workshop*, 131–136.
- [8]. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of Interspeech 2021*, 2426–2430.
- [9]. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., et al. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of Interspeech 2022*, 2278–2282.
- [10]. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (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.
- [11]. Chung, J. S., Nagrani, A., and Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. *Proceedings of Interspeech 2018*, 1086–1090.
- [12]. Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of IEEE/CVF CVPR*, 4690–4699.
- [13]. Raj, D., Xu, P., Chen, Z., Watanabe, S., Stolcke, A., Khudanpur, S., and Povey, D. (2021). DOVER-Lap: A method for combining overlap-aware diarization outputs. *Proceedings of IEEE SLT 2021*, 881–888.
- [14]. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., et al. (2023). Scaling speech technology to 1,000+ languages. *arXiv:2305.13516*.