International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 9 (September 2026), PP 111-117

Low-Resource Speaker Diarization: A Systematic Review of Progress, Challenges, and Deployment Pathways

Orazulume Chikodili Martha[1]*, Udofia Kingsley Monday[1], Udofia Kufre Michael[1], Philip Michael Asuquo[2], Nwaiwu Sarah Chiziterem[3]
[1]*Department of Electrical and Electronic Engineering, University of Uyo, Akwa Ibom, Nigeria*
[2]*Department of Computer Engineering, University of Uyo, Akwa Ibom, Nigeria*
[3]*Department of Electrical and Electronic Engineering, Topfaith University, Mkpatak, Akwa Ibom, Nigeria*
**Corresponding Author*

Abstract
The task of speaker diarization (identifying who spoke when in a multi-speaker audio recording) is crucial for automatic speech processing but remains underutilized in low-resource settings due to the need for large annotated speaker corpora in order to deploy state-of-the-art systems. In this paper, we present a systematic review of progress, challenges, and deployment strategies in low-resource speaker diarization, from the field's statistical foundations to the latest generation of self-supervised deep learning models. The review categorizes the literature into four key themes: the development of speaker embedding architectures, the problems encountered in the acoustic and domain dimensions in the context of deployment, the evaluation frameworks applied to assess the development, and the new avenues needed for deployment to annotation-free diarization, such as self-supervised learning, cross-lingual transfer, and frozen-inference deployment. The following five open problems are highlighted as the frontier of the field: VoxConverse boundary condition, CPC separability paradox, angular margin gap in SSL objectives, overlap-aware representation learning, and lack of low-resource language benchmarks. Directions for future research are specified for each. The review suggests that self-supervised speaker embeddings, generally trained in an unconstrained mode (but fine-tuned only in the frozen-inference phase), can serve as a viable and scalable base for low-resource diarization in telephonic and read-speech domains, whereas unconstrained multimedia speech is an open challenge.
Keywords: *speaker diarization, low-resource, self-supervised learning, speaker embeddings, deployment, domain mismatch, frozen inference, cross-lingual transfer, evaluation metrics*

Date of Submission: 09-09-2026 Date of acceptance: 18-09-2026

I. INTRODUCTION

In many situations, speech processing is required for recordings with multiple speakers, and determining when each speaker spoke is one of the most practically consequential problems. The applications range from automatic meeting transcription, call-centre analytics and broadcast media indexing, legal and medical documentation, telephone forensics, to accessibility services for the hearing impaired (Anguera et al., 2012; Park et al., 2022). Although speaker diarization has been extensively studied over the years and has made great progress recently with deep learning, it is still a challenging problem in low-resource scenarios such as settings with limited labelled speaker data, limited computational infrastructure, or underrepresented languages.

Most speaker diarization methods rely onspeaker embeddings: these are low-dimensional numerical representations of speech segments that encode speaker identity. Supervised embedding models like ECAPA-TDNN (Desplanques et al., 2020) work well on standard benchmarks, but they require training on large, carefully annotated speaker corpora. These are available for only a few high-resource languages and domains, and most deployment domains and languages are not covered. However, there is also a theoretically informed alternative, namely self-supervised learning (SSL), which trains speech representations directly from the unlabelled speech data, without relying on speaker labels at all (van den Oord et al., 2018; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022).

This review fills a gap in the literature, although there are some systematic reviews of speaker diarization (Anguera et al., 2012; Park et al., 2022); none of these reviews has focussed on the low-resource deployment problem as an organising principle and covered SSL models, condition-stratified performance, cross-lingual transfer, and failure mode analysis within a unified analysis. The review has four contributions. First, it offers the most extensive survey of existing SSL models for speaker diarization across acoustic conditions, followed by a detailed performance analysis. Second, it presents a framework to help determine

deployment viability as a function of model performance and acoustic challenge type. Third, it captures and describes five concrete open problems, with given directions for research. Fourth, it explicitly takes up the low-resource language dimension.

The remainder of this paper is organised as follows. Section II offers background information on the diarization pipeline and the concept of low-resource. The evolution of speaker embedding architectures is discussed in Section III. In Section IV, SSL models for diarization are reviewed. In Section V, deployment challenges are analysed. Section VI reviews evaluation frameworks. Crosslingual considerations are dealt with in Section VII. Open problems are listed in Section VIII, and Section IX concludes.

II. BACKGROUND

2.1 The Speaker Diarization Pipeline

All speaker diarization systems follow this pipeline: Voice activity detection (VAD) identifies regions of speech; Speaker segmentation: Segments speech into analysis units; Speaker embedding Extraction:Maps each segment onto a fixed-dimension embedding; Clustering: Clusters embeddings from the same speaker; Speaker Assignment: Labels each segment. The clustering stage does not need labelled speaker data during inference in the clustering-based systems, which are the most common approach in low-resource environments (Anguera et al., 2012). The quality of the diarization is heavily reliant on the embedding stage: if the embeddings of the same speaker are geometrically similar, and embeddings of different speakers are dissimilar, then Agglomerative Hierarchical Clustering (AHC) can reliably separate speakers. Average intra-speaker and inter-speaker cosine similarity (Orazulume et al., 2026, manuscript in preparation) is an important geometric indicator of diarization performance.

2.2 Defining Low-Resource

There is no consensus on how the term low-resource is applied in the diarization literature. A four-dimensional definition, which focuses on deployment, is used for this review. One is annotation scarcity: the target domain does not have enough speaker-labelled data to train or fine-tune a speaker embedding model. Second, language scarcity: the target language has limited annotated speech corpora, a lack of pretrained models, and a lack of benchmarks. Third, computational scarcity: the deployment environment doesn't have Graphics Processing Unit (GPU)infrastructure. Fourth, domain scarcity: there is no training corpus in the target acoustic domain. These dimensions tend to be co-occurring, and annotation scarcity is the most researched, while the others are studied comparatively little.

III. EVOLUTION OF SPEAKER EMBEDDING ARCHITECTURES

3.1 Statistical Foundations: GMM-UBM and i-Vectors

The first successful speaker diarization systems used Gaussian Mixture Models (GMM) trained on Mel-Frequency Cepstral Coefficients (MFCC) and Hidden Markov Models (HMM) for temporal dynamics (Reynolds et al., 2000). The speakers were represented by GMMs developed from a Universal Background Model (UBM). GMM-UBM systems were effective in controlled environments, but faced acoustic variability and needed lots of labelling for the training of UBM. The first major architectural change, the i-vector framework (Dehak et al., 2011), mapped utterances to a low-dimensional total variability space via factor analysis. Although i-vectors made significant progress over GMM-HMM systems and remained state-of-the-art until the mid-2010s, their linear factorization could not capture nonlinear speaker variability in naturalistic speech samples.

3.2 Deep Learning Era: x-Vectors and ECAPA-TDNN

By taking advantage of the nonlinear representational capacity of neural networks, Time Delay Neural Networks trained for speaker classification yielded x-vectors (Snyder et al., 2018), which have continuously outperformed i-vectors. ECAPA-TDNN (Desplanques et al., 2020) improved this template in three ways: by adding channel attention to selectively prioritize regions of time-frequency space; by adding multi-scale residual connections to combine information at different temporal scales; and by adding propagation and aggregation mechanisms to boost information flow among different layers of the network. Trained on VoxCeleb2, with Additive Angular Margin (AAM)Softmax(Deng et al., 2019), ECAPA-TDNN obtains state-of-the-art results on speaker verification and diarization benchmarks (Dawalatabad et al., 2021). One of its main drawbacks is its reliance on large annotated corpora, including VoxCeleb2, with over 1 million utterances from over 6,000 speakers, mostly in English, from media interviews.

3.3 Self-Supervised Learning: From CPC to WavLM

The SSL template was proposed in 1998 as Contrastive Predictive Coding (CPC; van den Oord et al., 2018): a convolutional encoder extracts latent representations, an autoregressive model constructs a context

vector, and a contrastive loss learns to predict future latent representations from the context. With 960 hours of speech from LibriSpeech audiobooks, Wav2vec 2.0 (Baevski et al., 2020) took the paradigm forward by introducing a transformer architecture and quantisationmodule trained with a masked contrastive objective. HuBERT (Hsu et al., 2021) adopts iterative masked prediction and offline k-means cluster targets. WavLM (Chen et al., 2022) introduces a denoising task to the masked prediction, resulting in a representation that is robust against acoustic degradation. An overview of the most frequently investigated properties of the five models used for diarization is given in Table I.

Table I: Speaker Embedding Models for Low-Resource Diarization — Key Properties

Model	Type	Pretraining Objective	Emb. Dim.	Key Advantage	Key Limitation
CPC	SSL	Contrastive future prediction	256	Lightweight; competitive despite zero margin	Weak cosine separability
wav2vec 2.0	SSL	Masked contrastive	768	Domain-alignment advantage in telephonic speech	High intra-speaker similarity
HuBERT	SSL	Iterative masked prediction	768	Fine-grained acoustic discrimination	Phonetically sensitive
WavLM	SSL	Masked pred. + denoising	768	Theoretically noise-robust	Denoising benefit not confirmed at base scale
ECAPA-TDNN	Supervised	AAM-Softmax classification	192	Highest separability margin; best standard DER	Requires large labelled corpus

Source: Desplanques et al. (2020); van den Oord et al. (2018); Baevski et al. (2020); Hsu et al. (2021); Chen et al. (2022).

IV. SELF-SUPERVISED MODELS FOR LOW-RESOURCE DIARIZATION

4.1 Frozen Inference Paradigm

The most relevant result obtained by recent SSL-based diarization research is that it is possible to obtain meaningful diarization results without fine-tuning or adaptation on target-domain data by using weights of a pretrained SSL model in frozen mode during inference. In the frozen inference mode, the speech segments are fed into the pretrained model, and the embeddings obtained are clustered using AHC with cosine distance and average linkage. This paradigm does not need any labelled data during the inference stage and is feasible for computing on relatively low-power hardware. Results from the systematic evaluation of wav2vec 2.0 on five benchmark datasets across five models demonstrate that in "frozen inference mode", wav2vec 2.0 achieves perfect diarization performance (0.00% DER-like) on LibriSpeech read speech and 3.22% DER-like on CALLHOME under noise and short-duration conditions, which is comparable to the supervised ECAPA-TDNN baseline under specific acoustic conditions (Orazulume et al., 2026, manuscript in preparation).

4.2 Condition-Stratified Performance

It is not generally true that the SSL models perform better or worse than supervised baselines; it depends on the conditions. ECAPA-TDNN consistently outperforms all SSL models on the five benchmark datasets, with an average DER-like of 13.04%, while the average of the SSL models was 41.66% to 56.22%. In the meantime, wav2vec 2.0 achieves a higher DER-like of 3.22% on CALLHOME under telephonic noise conditions (5 dB and 0 dB SNR) and short segment duration (1.5 s) than ECAPA-TDNN (8.06%). The reason for the reversal is that the audiobook speech in LibriSpeech used for the pretraining distribution of wav2vec 2.0 is closer to noisy telephonic speech than the training distribution (i.e., corpus) of VoxCeleb2, which is used for ECAPA-TDNN. This domain alignment hypothesis is used for principled model selection in low-resource settings. The condition-stratified deployment recommendation is summarised in Table II.

Table II: Condition-Stratified Deployment Framework for Low-Resource Diarization

Deployment Context	Primary Challenge	Recommended Model	Empirical Basis
Clean read speech	None	Either model	0.00% DER-like for both models on LibriSpeech
Noisy telephonic (>=5 dB SNR degradation)	Noise	wav2vec 2.0	3.22% vs. 8.06% ECAPA-TDNN; 60% relative advantage
Short telephonic segments (<=3 s)	Short duration	wav2vec 2.0	3.22% vs. 8.06% ECAPA-TDNN at 1.5 s
Clean telephonic	Domain gap	ECAPA-TDNN	22.56% vs. 8.06%; domain gap persists without noise

Deployment Context	Primary Challenge	Recommended Model	Empirical Basis
Meeting speech (>10% overlap)	Overlap	ECAPA-TDNN	SSL models degrade more under overlap
Multimedia / online video	Acoustic variability	Neither reliable	No model achieves <49% DER-like on VoxConverse

Source: Orazulume et al. (2026, manuscript in preparation). Values represent the best observed DER-like under frozen inference conditions.

4.3 The Angular Margin Gap

A cosine separability margin gap of 0.505 between ECAPA-TDNN (0.505) and all SSL models (0.000-0.023) is due to the cosine separability margin gap in the training objective of ECAPA-TDNN, which explicitly penalises embedding falling within the angular margin boundary for each speaker class (Deng et al., 2019). There is no similar geometric constraint on speaker separability in SSL pre-training objectives like contrastive prediction, masked prediction, or denoising. Recent studies have investigated the combination of SSL pretraining with angular margin fine-tuning, yielding encouraging results towards bridging the gap, yet still needing a small amount of labelled speaker data for fine-tuning (Lepage and Dehak, 2024; Miara et al., 2024).

V. DEPLOYMENT CHALLENGES

5.1 Within-Speaker Acoustic Variability

The most basic challenge in speaker diarization in naturalistic environments is within-speaker acoustic variability, where the same speaker can produce speech that varies widely across acoustic conditions in a single recording, resulting in a spread-out cluster of embeddings for a speaker. The worst of the five models and five robustness conditions is 30.00% (ECAPA-TDNN at 1.5 s segment duration), as reported by VoxConverse (Chung et al., 2020), which is far from deployment quality. The resistance of VoxConverse includes four failure modes: within-speaker acoustic variability, domain mismatch, and speaker ordering inversion ("Orazulume et al. manuscript in preparation" is able to outperform ECAPA-TDNN with only a near-zero separability margin), and HuBERT misleading clustering (with the only negative ARI in the five-model, five-dataset evaluation).

5.2 Domain Mismatch

Degradation in performance that occurs when a model trained on one domain is deployed to another is the most consistently reported deployment challenge (Anguera et al., 2012; Park et al., 2022). In the case of speaker embeddings, domain mismatch comes down to a decrease in the cosine separability margin. There are two main sources relevant to low-resource settings: (1) language mismatch – models are pre-trained on English and show degraded performance on other languages, and (2) channel mismatch – models are trained on studio-quality recordings and show degraded performance on far-field or telephone-channel recordings. Multilingual pretraining can significantly boost cross-lingual transfer (Babu et al., 2022; Conneau et al., 2021).

5.3 Overlapping Speech

Clustering-based systems are most adversely affected by overlapping speech or simultaneous speech from two or more speakers. In overlap, a segment's embedding blends speakers' identities, degrading cluster purity. In the majority of clustering-based systems, overlapping segments are assigned to one speaker, and the other speaker's contribution is ignored. Increased performance can be achieved by using overlap-aware diarization architectures (Raj et al., 2021), which require much more annotated training data; thus, they are not strictly applicable in low-resource scenarios.

5.4 Short-Duration Segments

The duration of the segments is not directly proportional to the accuracy of the diarization and depends on the model. Shorter segments do not hurt for ECAPA-TDNN on VoxConverse: The DER decreases monotonically from 8.0 s (60.45%) to 1.5 s (30.00%). For wav2vec 2.0 on CALLHOME, a similar pattern holds (22.56% at 8.0 s to 3.22% at 1.5 s). This seemingly paradoxical outcome is due to the power of within-speaker variability over the amount of information: shorter segments are more acoustically consistent, leading to more compact embeddings (Orazulume et al., 2026, manuscript in preparation).

VI. EVALUATION FRAMEWORKS

6.1 Diarization Error Rate and Its Limitations

The most common evaluation method is the Diarization Error Rate (DER), which indicates the percentage of time when speaker attribution is wrong (Anguera et al., 2012). DER is a metric with three important shortcomings in low-resource areas. Firstly, it is also susceptible to VAD errors and treats VAD

quality as embedding quality. Second, it needs accurate segments, which are not necessarily available in low-resource areas. Third, it does not depend on the geometric characteristics of the embedding space: two systems having the same DER could have very different separability margins, and this would have a different consequence for the robustness of deployment.

6.2 Clustering-Based Metrics and Embedding Geometry

The clustering-based metrics (Speaker Purity (Manning et al., 2008), Normalised Mutual Information (Strehl and Ghosh, 2002), Adjusted Rand Index (Hubert and Arabie, 1985)) are complementary measures. Purity is a measure of contamination within each cluster, NMI is a measure of shared information between the predicted and the true labels, and ARI reflects the number of pairs of labels that are in agreement but not by chance, negative scores occur when the performance is worse than random. The cosine separability margin directly measures embedding quality without depending on the performance of the clustering algorithm. Pearson correlations of 0.92-0.96 between the separability margin and diarization performance indicate that the separability margin can be used to predict diarization performance (Orazulume et al., 2026, manuscript in preparation). Combined use of all four metrics: DER-like, Purity, NMI and ARI is recommended for low resource evaluations.

VII. CROSS-LINGUAL TRANSFER AND UNDERREPRESENTED LANGUAGES

7.1 Cross-Lingual Transfer from SSL Models

To a great extent, the acoustic properties (speaker-discriminative prosody, vocal tract characteristics) of speech encoded by SSL models are language-universal, thus reflecting salient features of human speech more broadly. The relative decrease in DER with wav2vec 2.0 fine-tuned on Kurdish compared to a baseline (Abdullah et al., 2025) is 7.2%, while the cluster purity is 13% higher, indicating that the cross-lingual transfer is effective even for phonologically dissimilar languages. The XLS-R model (Babu et al., 2022) is initialized with a huge amount of speech data (436,000 hours) across 128 languages using the wav2vec 2.0 architecture, and performs well on low-resource language benchmarks with zero-shot and few-shot transfer.

7.2 The Benchmark Gap for Underrepresented Languages

All five corpora employed for diarization evaluation are in English: VoxCeleb, LibriSpeech, AMI, CALLHOME, and VoxConverse. African, Middle Eastern, Southeast Asian, and Indigenous languages are largely unrepresented. This benchmark gap has two impacts: First, it makes it difficult to determine whether deployment pathways planned for English transfer to other languages, and second, it creates a positive feedback loop that draws research resources to already well-resourced languages. Diarization benchmarks for underrepresented languages must be developed first to enable equitable use of speech technology worldwide.

VIII. OPEN PROBLEMS AND FUTURE RESEARCH DIRECTIONS

8.1 The VoxConverse Boundary Condition

All speaker embedding models under all acoustic conditions are at the current limit when the speakers use VoxConverse. These advances lie in three directions: domain-aligned pre-training on corpora that contain multimedia content from the same user as in YouTube; regularisation of within-speaker variability, namely explicitly penalising the inconsistencies across sound conditions; and consistency across the segments while maintaining that they are not too short. VoxConverse is the reference standard for evaluating the progress for unconstrained multimedia speaker diarization.

8.2 The CPC Separability Paradox

This is the CPC Separability Paradox.This is the CPC Separability Paradox. With such near-zero cosine separability margins, the most difficult result in recent SSL diarization research, CPC delivers competitive diarization performance. This suggests that pairwise cosine similarity is not sufficient to exploit speaker-discriminative information in CPC representations, and that agglomerative clustering can learn geometric patterns not reflected in the margin. Theories of the distribution mechanism are important, and this could be studied in higher-dimensional distributional structure.

8.3 The Angular Margin Gap

The separability margin gap between ECAPA-TDNN (0.505) and all SSL models (0.000 – 0.023) is directly due to the angular margin objective. The main representational problem for SSL diarization studies is to close this gap without using any labelled speaker information. The most straightforward direction is to formulate angular-margin-inspired pretraining objectives on unlabelled speech, such as creating positive and negative speaker pairs based on acoustic similarity in the recording.

8.4 Overlap-Aware SSL Representation Learning

All existing SSL models are based on the principle of interpreting overlapping speech as acoustic noise, rather than as a problem to be explicitly modelled. An essential improvement would be SSL objectives that are based on separating overlapping speakers, such as training supervised separation signals using multi-channel recordings or synthesising overlap examples with known speaker identities using the temporal turn-taking structure.

8.5 Low-Resource Language Benchmarks

The current state of speaker diarization benchmarks for underrepresented languages is a prerequisite for enabling the field to benefit the communities in which diarization is most needed for low-resource languages. The following design principles should be followed when developing benchmarks: the benchmark should represent actual deployment in the target language community; it should represent a degree of diversity of speakers within the linguistic community; and it should be reproducible using the same programming tools, with reported inter-annotator agreement.

IX. CONCLUSION

This review has offered a systematic synthesis of the literature on low-resource speaker diarization, ranging from the speaker embedding architectures used in GMM/UBM systems to i-vectors, x-vectors and ECAPA-TDNN, and then on to the newest generation of self-supervised speaker embedding systems. The main conclusion is that self-supervised speaker embeddings used in frozen inference mode can serve as a valid baseline for low-resource diarization in telephonic and read-speech domains, but they are not universally applicable to all deployment scenarios. This review is the first to provide condition-stratified, evidence-based recommendations on the deployment viability of low-resource diarization models, with wav2vec 2.0 as the appropriate SSL model for noisy telephonic conditions and ECAPA-TDNN's benefits under overlap. The research frontier includes five open problems: the VoxConverse boundary condition, the CPC separability paradox, the angular margin gap, overlap-aware representation learning, and the benchmark gap for underrepresented languages. The field's ability to meet the needs of the communities for which low-resource diarization is most important will be determined by progress on these five problems.

REFERENCES

[1]. 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. https://arxiv.org/abs/2504.18582.

[2]. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 356–370.

[3]. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Lozhkov, A., Collobert, R., Wiseman, S., Han, J., Edunov, S., Auli, M., and Hsu, W. N. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. Proceedings of Interspeech 2022, 2278–2282.

[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]. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (pp. 28–39). Springer.https://doi.org/10.1007/11677482_3

[6]. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Zhou, S., and Qian, Y. (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.

[7]. Chung, J. S., Huh, J., Nagrani, A., Afouras, T., and Zisserman, A. (2020). Spot the conversation: Speaker diarisation in the wild. Proceedings of Interspeech 2020, 299–303.

[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]. Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., and Na, H. (2021). ECAPA-TDNN embeddings for speaker diarization. Proceedings of Interspeech 2021, 3560–3564.

[10]. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788–798.

[11]. Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4690–4699.

[12]. 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.

[13]. Han, J., Landini, F., Rohdin, J., Silnova, A., Diez, M., and Burget, L. (2025). Leveraging self-supervised learning for speaker diarization. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1–5.

[14]. 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.

[15]. Hubert, L., and Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.

[16]. Lepage, T., and Dehak, R. (2024). Additive margin in contrastive self-supervised frameworks to learn discriminative speaker representations. arXiv. https://arxiv.org/abs/2404.14913.

[17]. Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

[18]. Miara, M., Ravanelli, M., and Landini, F. (2024). Towards supervised performance on speaker verification with self-supervised learning. Proceedings of Interspeech 2024.
[19]. Nagrani, A., Chung, J. S., and Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. Proceedings of Interspeech 2017, 2616–2620.
[20]. Orazulume, C. M., Udofia, K. M., Udofia, K. M., Asuquo, P. M., and Nwaiwu, S. C. (2026). The cosine separability margin as a geometric predictor of speaker diarization performance: Theory, measurement, and implications. International Journal of Research in Engineering and Science, 14(8), 52–59. https://www.ijres.org/.DOI: 10.5281/zenodo.21979536
[21]. Orazulume, C. M., Udofia, K. M., Udofia, K. M., Asuquo, P. M., and Nwaiwu, S. C. (2026). Self-supervised vs. supervised speaker embeddings for low-resource multi-speaker diarization. Manuscript in preparation.
[22]. Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 5206–5210.
[23]. Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech and Language, 72, Article 101317.
[24]. 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.
[25]. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1–3), 19–41.
[26]. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 5329–5333.
[27]. Strehl, A., and Ghosh, J. (2002). Cluster ensembles: A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3, 583–617.
[28]. van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748.