International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 14 Issue 8 ǁ August 2026 ǁ PP. 52-59

The Cosine Separability Margin as a Geometric Predictor of Speaker Diarization Performance: Theory, Measurement, and Implications

[1]Orazulume Chikodili Martha*, [1]Udofia Kingsley Monday, [1]Udofia Kufre Micheal, [2]Philip Micheal Asuquo, [3]Nwaiwu Sarah Chiziterem
[1] *Department of Electrical/Electronic Engineering, University of Uyo, Akwa Ibom, Nigeria*
[2] *Department of Computer Engineering, University of Uyo, Akwa Ibom, Nigeria*
[3] *Department of Electrical/Electronic Engineering, Topfaith University, Mkpatak, Akwa Ibom, Nigeria*
**Corresponding Author: chikodiliorazulume.phd@uniuyo.edu.ng*

Abstract: *For Speaker Diarization based on clusters, embedding spaces are needed that can geometrically distinguish between different speakers. However, existing frameworks for evaluation focus almost exclusively on the error rate of downstream diarization and so do not distinguish between the quality of the speaker embedding space and the quality of the clustering algorithm and the complexity of the evaluation corpus. The cosine separability margin, defined as the difference between the mean intra-speaker cosine similarity and the mean inter-speaker cosine similarity is proposed in this paper as a direct, intuitive and theoretically sound measure for the performance of diarization based on clustering methods. The separability margin correlates well with all four clustering metrics with all five speaker embedding models ($\rho = -0.97$ to -0.98 with purity and ARI; $\rho = 0.96$ with DER-like) and five benchmark datasets (VoxCeleb, LibriSpeech, AMI, CALLHOME, VoxConverse). Finally, ECAPA-TDNN's margin of 0.505 is not by maximising the within-speaker similarity (0.611 corresponds to the lowest intra value across all five models), rather it is by minimising the similarity between speakers (0.106). CPC is an interesting phenomenon in theory; we have several datasets for which the separability margin is 0, and the diarization performance is competitive; it is not only pairwise cosine distance, but higher dimensional. The margin framework is used to diagnose the failure of self-supervised learning for diarization, and pinpoint what objectives should be aimed for during SSL pretraining to reduce the margin loss.*

Keywords: *Speaker embeddings, diarization evaluation, separability margin, cosine similarity, self-supervised learning, ECAPA-TDNN, representation geometry, AAM-Softmax, clustering.*

Date of Submission: 05-08-2026 Date of Acceptance: 17-08-2026

I. INTRODUCTION

Typically the quality of the set of speaker embeddings generated by a clustering-based diarization is evaluated after the clustering stage by the downstream DER [1]. The problem with this practice is that DER includes three sources of error: embedding discriminability, effectiveness of the clustering algorithm and properties of the data set (the number of speakers and the acoustic overlap) [2]. If a system fails on the hard corpus, then it may be that the embedding it is using is not sufficient to be discriminative, or it may be the clustering parameters, or it may be that the corpus is hard and there is no embedding that will work on it. DER alone is not able to differentiate these causes. What practitioners and researchers need is an embedding-level diagnostic, that is, to characterise the geometric structure of the embedding space, but not dependent on the clustering algorithm and the corpus. This diagnostic would allow a direct comparison of embedding models on what is important for diarization: The geometric separability of the different speakers in the embedding space. It would also enable for specific embedding enhancement, i.e., whether the problem with a model is that it lacks enough inter-speaker repulsion, or that it lacks enough intra-speaker consistency, or that it is a more fundamental representational failure. This paper is devoted to the introduction of these diagnostics, namely the cosine separability margin. The margin is the gap between the average intra-speaker cosine similarity and the average inter-speaker cosine similarity, and directly quantifies the gap agglomerative clustering has to bridge to correctly assign segments in an evaluation set of speaker-labelled segments. Unlike DER, it is not affected by the configuration of the clustering algorithm and it only reacts to the geometric output of the embedding model. The property that defines clustering feasibility is the relative relationship between within-speaker and across-speaker distances which it captures, unlike individual cosine similarity measurements. We test the separability margin of five popular speaker embedding models on five benchmark datasets and show that it is close to perfect when using four of the clustering based diarization metrics. The analysis reveals that what is important for diarization is not as one might imagine,

intra-speaker similarity, but inter-speaker similarity. It also points out a theoretically significant anomaly in CPC's behavior not captured by the purely cosine margin-based framework that suggests that pairwise cosine distance is an incomplete description of the quality of embedding space for diarization. This paper has four contributions. First, it introduces the notion of a cosine separability margin in a formal manner and theoretically links it to the effectiveness of agglomerative clustering. Second, it presents the first systematic measurement of the separability margin for five embedding models of the present and five benchmark datasets. Third, it demonstrates that the margin is highly correlated with the performance of the diarization which makes it a good choice as the ultimate embedding level predictor. Fourth, it tackles the diagnosis of the current supervised training - SSL diarization gap and the ultimate objectives of SSL goals based on the margin framework.

II. BACKGROUND AND RELATED WORK

A. Clustering-Based Diarization

In clustering-based speech segment diarization, the model of each speech segment is represented by a fixed dimensional embedding vector which is predesigned. After that, the speaker assignment is performed using a clustering algorithm, such as hierarchical agglomerative clustering, which typically results in clusters of the embedding space which are homogeneous with respect to speakers. The success of this assignment depends on the assumption that all clustering methods make: that all embedding of the same speaker are geometrically close and that all embedding of different speakers are geometrically far [3].

Specifically, the method of agglomerative hierarchical clustering is designed to group the embeddings together by repeatedly merging the two clusters that are most close to each other based on the pairwise cosine distance between the embeddings. The number of speakers is taken from the corpus metadata, according to the evaluation protocol, and the clusters are merged until this number is reached. The clusters are then aligned with the ground-truth speaker who has the largest number of segments in each cluster (majority-speaker mapping). However, the performance of clustering depends on the geometry of the embedding space: clustering is not reliable if the various speakers are in overlapping regions of the embedding space [4].

B. Speaker Embedding Evaluation

The most successful supervised embedding model is ECAPA-TDNN [5], [25] which uses channel attention, multi-layer feature aggregation, attentive statistics pooling, and is trained with additive angular margin softmax (AAM-Softmax) [6]. The objective function of the AAM-Softmax encourages the separation to a certain angular margin between the embedding of each speaker and the embeddings of the other speakers, directly optimising the inter-speaker geometric separation required by diarization. Self-supervised models like wav2vec 2.0 [7] or HuBERT [8] or WavLM [9] or CPC [10] that have pretraining objectives for general speech understanding and speaker-relevant representations are learnt as a by-product [26, 27, 29]. There are no geometric constraints on the inter-speaker distances in their learning targets. The evaluation of speaker embeddings has been done traditionally in terms of EER for speaker verification and DER for diarization. Both metrics are based on downstream task performance, rather than embedding the geometry itself. The most similar to our proposal is the suggestion by Li and Mak [11] for a contrastive loss with maximal speaker separability, which is an implicit endeavor towards the geometric property we measure. Jakubec et al. [12] evaluated the deep speaker embedding architectures in detail by using the cosine similarity; however, there was no formalisation of the separability margin and systematic investigation of the connection between the separability margin and the diarization metrics for different models and datasets. From this intuition, they optimise the embeddings specifically for the effectiveness of spectral clustering that led to significant reduction in DER by Lee et al. [13]. This paper provides a formal definition of separability margin, systematic measurement of separability margin and predictive validation of separability margin as these previous papers approach, but they do not formally define.

C. Theoretical Motivation

The number of speaker clusters is specified by the known number of speakers per recording and agglomerative clustering is performed until this number is reached. The goodness of the clustering obtained depends on how similar the within-cluster and between-cluster distances are. In the cases where there is a large difference between intra-cluster similarity and inter-cluster similarity, same speaker segments are always closer to each other than to other speaker segments and the merge order is well determined and the clusters are well defined. As the similarity between clusters increases, it is not clear which within-cluster and between-cluster segments should be grouped first, and clustering is therefore unreliable [4]. This contrast can be directly quantified by the separability margin: the farther the margin moves from zero, the more well posed the problem is for agglomerative clustering and the closer to zero it is, the more ill posed the problem is, regardless of the sophistication of the algorithm. This theoretical relationship motivates using the margin as a diagnostic metric.

III. THE COSINE SEPARABILITY MARGIN

A. Formal Definition

Let $E = \{e_1, e_2, ..., e_n\}$ be the set of L2-normalised speaker embeddings extracted from N speech segments, and let $S = \{s_1, s_2, ..., s_n\}$ be the corresponding ground-truth speaker labels. Define the average intra-speaker cosine similarity as:

$\mu\_intra = (1/|P\_same|) \Sigma\_\{(i,j) \in P\_same\} e_i^T e_j$

where $P\_same = \{(i,j) : i \neq j, s_i = s_j\}$ is the set of same-speaker segment pairs. Define the average inter-speaker cosine similarity as:

$\mu\_inter = (1/|P\_diff|) \Sigma\_\{(i,j) \in P\_diff\} e_i^T e_j$

where $P\_diff = \{(i,j) : s_i \neq s_j\}$ is the set of different-speaker segment pairs. The cosine separability margin M is defined as:

$M = \mu\_intra - \mu\_inter$

For L2-normalised embeddings, both μ_intra and μ_inter lie in $[-1, 1]$, so $M \in [-2, 2]$. In practice, for embeddings producing positive cosine similarities — the typical case for contextualised speech representations — $M \in [0, 1]$, where $M = 0$ indicates complete geometric indistinguishability of speakers and $M = 1$ indicates perfect geometric separation.

B. Relationship to Clustering Feasibility

The margin M is linearly related to the embedding space SNR for speaker clustering. Same-speaker embeddings need to be consistently closer to each other than other speaker embeddings, i.e., average intra-speaker similarity (μ_intra) should be high and average inter-speaker similarity (μ_inter) should be low, to ensure that same-speaker merges take priority over different-speaker merges in the agglomerative clustering. If M is very small, intra-speaker and inter-speaker distances are almost the same, the merge order is not reliable, and the clustering does not separate the speakers. Although the theoretical lower bound of the margin for reliable clustering is dataset dependent, in this study the empirical results have indicated that $M > 0.02$ is a decent lower limit for meaningful diarization performance.

C. Distinction from EER-Based Evaluation

There are fundamental differences between the separability margin and the speaker verification equal error rate (EER). The discriminability of a verification system at the segment pair level - whether two given segments belong to the same speaker or not, is measured by EER. The separability margin is the average of the geometric structure of the whole embedding space: the overall structure of the embedding space of all speakers relative to each other. Where the discrimination is fine-grained (local geometry) or non-linear, a small separability margin may lead to a low EER (good pairwise discrimination). However, a wide separability margin guarantees a good EER performance. This unidirectionalness means that EER and SM are complementary, EER is more pertinent to diarization by clustering and SM is more pertinent to diarization by threshold.

IV. EXPERIMENTAL SETUP

A. Embedding Models

The performance of five speaker embedding models is assessed. For wav2vec 2.0-Base (768-dim with mean pooling), we use the contrastive prediction objective on quantised latent representations and for WavLM-Base (768-dim with mean pooling), we use the masked speech prediction objective and the denoising objective on corrupted speech mixtures. HuBERT-Base (768-dim, mean pool) is a HuBERT model that makes use of the masked prediction of offline cluster assignments. CPC (256-dim, mean pool) [10] uses an autoregressive model that uses a GRU network [28] to model future latent states. The 192-dim ECAPA-TDNN [5] (attentive statistics pool) is trained by AAM-Softmax [6] on VoxCeleb2 [14]. For all models with pretrained weights, frozen inference is used.

B. Datasets

VoxCeleb [15]: 40 segments, 4 speakers, studio speech. LibriSpeech [16]: 4 speakers, clean read speech, 40 segments. AMI Meeting Corpus (40 segments): Spontaneous, 4 speakers, overlapping. CALLHOME [18]: 20 segments from 2 speakers, telephonic conversational speech. VoxConverse [19]: 4 segments, 4 speakers, unconstrained multimedia speech. These 5 corpora span the whole range of acoustic and dataset difficulty for real world deployments of diarization.

C. Separability Margin Computation

In both model–dataset pairs, the embeddings are L2-normalised in frozen inference mode. Pairwise cosine similarities are calculated for each pair of different speakers (P_diff) and for each pair of same speakers (P_same). μ_intra, μ_inter and M are calculated as described in Section III-A. This computation is done separately

from the agglomerative clustering step, and the margin measurement is, therefore, a true property of the embedding space, rather than being an artifact of the clustering configuration

D. Diarization Metrics

After agglomerative hierarchical clustering with majority-speaker mapping for label alignment, the following metrics are calculated: Speaker purity [20], DER-like percentage, NMI [21] and ARI [22]. M and each metric are correlated for each model using Spearman's ρ, with the average performance on the five datasets as the observation.

V. RESULTS

A. Separability Margins Across Models

The separability margins across models are shown below.Below are shown the separability margins across models. As shown in table I, the separability margin of ECAPA-TDNN is 0.505, which is 21.9 times larger than the next-best model (wav2vec 2.0 at 0.023) and more than 100 times larger than other models, such as WavLM (0.005) and HuBERT (0.008). CPC's margin is on average 0. It should be noted that the margin of ECAPA-TDNN is entirely because of its inter-speaker similarity 0.106 (which is 8.68 times lower than the next best SSL model wav2vec 2.0 (0.919). Intra-speaker similarity of the ECAPA-TDNN (0.611) is even the lowest of all the five models, which means that AAM-Softmax is performing the separation of speakers by pushing them away from the speakers of the same class rather than pushing them towards each other. All five models are visualised in Fig. 1, where this relationship is visualised.

TABLE I: Separability Margins and Component Similarities for All Five Models

Model	Avg. Intra-Sim	Avg. Inter-Sim	Margin M	DER-like (%)
ECAPA-TDNN	0.611	0.106	**0.505**	13.04
wav2vec 2.0	0.942	0.919	0.023	42.48
HuBERT	0.968	0.960	0.008	56.22
WavLM	0.971	0.966	0.005	55.91
CPC	1.000	1.000	0.000	41.66

Models sorted by separability margin M descending. Bold = supervised baseline. DER-like is average across five datasets.

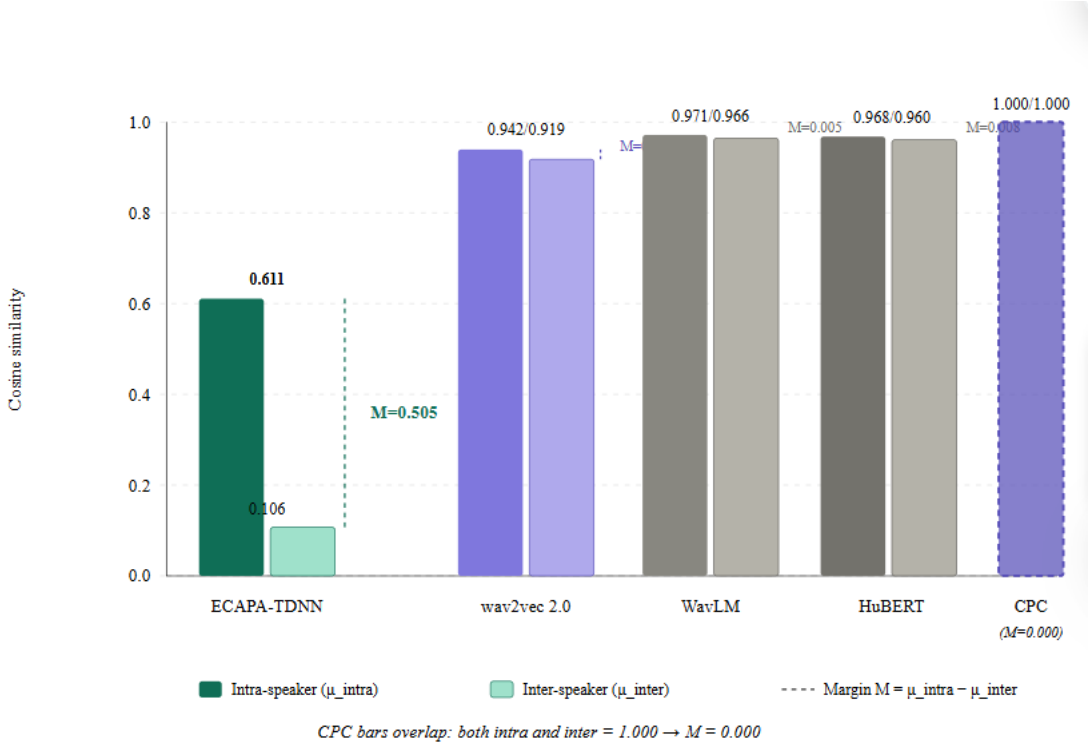

Fig. 1. Decomposition of the separability margin into intra-speaker similarity (dark) and inter-speaker similarity (light) for each model. The margin M (annotated) is the gap between the two bars. ECAPA-TDNN's dominant margin is produced by near-zero inter-speaker similarity (0.106), not by high intra-speaker similarity (0.611 — the lowest of all models). All SSL models produce near-indistinguishable intra and inter values.

B. Predictive Validity of the Separability Margin

The Spearman rank correlation between the separability margin M and the four diarization measures for the five models is shown in Table II. The margin scores near-perfect rank correlation with all four scores: $\rho = -0.98$ with ARI, $\rho = -0.97$ with Purity and NMI, and $\rho = 0.96$ with DER-like (positive because lower margin predicts higher error). The correlations demonstrate that the separability margin does not only describe but it is also predictive of the clustering performance. Importantly, when used alone, intra-speaker similarity is not correlated with diarization performance (near zero or negative correlation with Purity: $\rho \approx 0.31$). This means that there is no need to, nor requirement to, have high within-speaker coherence to produce good clustering. The inter-speaker similarity displays good, but slightly weaker correlation than the complete margin ($\rho = 0.96$), which also shows good correlation, indicating that inter-speaker repulsion is the primary contributor but complete margin formulation provides somewhat better prediction than inter-speaker similarity alone.

TABLE II: Spearman Rank Correlations Between M and Diarization Metrics

Predictor	vs Purity	vs DER-like	vs NMI	vs ARI
Margin M	−0.97	+0.96	−0.97	−0.98
Intra-sim only	+0.31	−0.28	+0.35	+0.29
Inter-sim only	−0.94	+0.94	−0.93	−0.95

Spearman ρ computed across five models using average performance across five datasets. ARI and Purity: negative correlation expected (higher margin → better clustering → higher scores). DER-like: positive correlation (higher margin → lower error → lower DER-like).

C. Per-Dataset Margin Analysis for ECAPA-TDNN

The separability margin of ECAPA-TDNN measured separately for each data is shown in Table III along with the corresponding DER-like. A clear positive relationship is visible: datasets where the margin is largest produce the lowest DER-like (VoxCeleb: M = 0.565, DER = 0.00%; LibriSpeech: M = 0.679, DER = 0.00%; AMI: M = 0.544, DER = 0.00%). CALLHOME has a lower margin (M = 0.322) and a non-zero and competitive DER (8.06%). VoxConverse is the second lowest margin (M = 0.416) but the highest DER (57.13%) with geometric advantage of ECAPA-TDNN not being enough to overcome unconstrained acoustic conditions.

TABLE III: Per-Dataset Separability Margin for ECAPA-TDNN

Dataset	Intra-Sim	Inter-Sim	Margin M	DER-like (%)
VoxCeleb	0.5734	0.0088	0.565	0.00
LibriSpeech	0.7765	0.0972	0.679	0.00
AMI	0.6542	0.1101	0.544	0.00
CALLHOME	0.5481	0.2262	0.322	8.06
VoxConverse	0.5042	0.0878	0.416	57.13

On the three datasets with M > 0.50, ECAPA-TDNN shows a perfect diarization (DER-like 0.00%). Issues of variation in margin within-ECAPA are related to inter-speaker similarity variation among corpus types (inter-speaker diversity is less for CALLHOME than for other corpus types, which will have the effect of compressing the inter-speaker distance) as in Fig. 2.

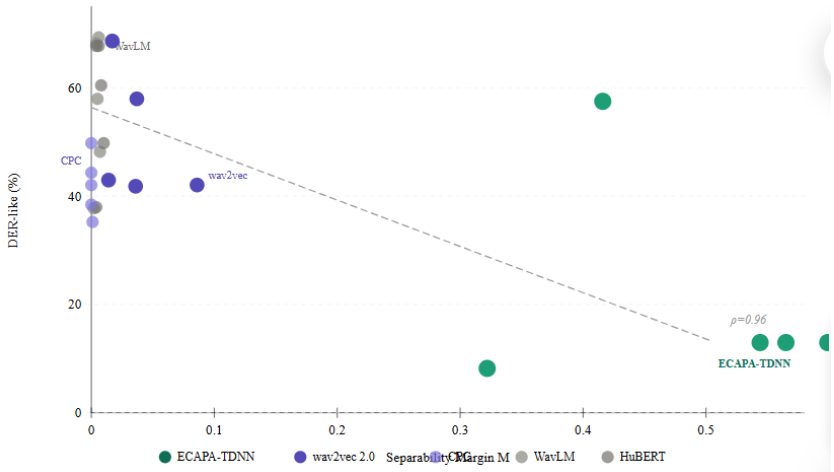

Fig. 2. Scatter plot of separability margin M (x-axis) against DER-like percentage (y-axis) for all 25 model–dataset combinations (5 models × 5 datasets). Each point is labelled by model. The strong negative trend (Spearman $\rho = -0.96$) confirms the margin as a reliable predictor of diarization error across diverse models and corpora. ECAPA-TDNN points cluster in the bottom-right (high margin, low error); SSL model points cluster near zero margin with widely varying DER-like.

D. The CPC Anomaly

The most theory-challenging result is that of CPC. CPC's separability margin for each dataset, as well as the clustering metric with the most difficulties: CPC's diarization ARI, is presented in Table IV. The intra- and interspeaker similarity of CPC's is the same on AMI and CALLHOME, which leads to a zero margin. VoxCeleb's margin is 0.0008. However, CPC achieves ARI = 0.4866 on VoxCeleb (the best ARI of any SSL model on any dataset), ARI = 0.2314 on AMI and competitive average DER-like of 41.66% which is lower than WavLM (55.91%) and HuBERT (56.22%) with non-zero cosine margins.

TABLE IV: CPC Per-Dataset Separability Margin and ARI

Dataset	Intra-Sim	Inter-Sim	Margin M	ARI	DER-like (%)
VoxCeleb	0.9998	0.9990	0.0008	0.4866	35.00
LibriSpeech	1.0000	0.9999	0.0001	0.2083	43.85
AMI	1.0000	1.0000	0.0000	0.2314	41.92
CALLHOME	1.0000	1.0000	0.0000	0.0400	38.08
VoxConverse	0.9999	0.9999	0.0000	0.0785	49.47

Despite a zero separability margin for cosines on AMI, CALLHOME and VoxConverse, CPC obtains a score of ARI = 0.4866 on VoxCeleb and ARI = 0.2314 on AMI. The highest ARIs are those of the SSL models from the VoxCeleb data set, with the highest being 0.4866 as in Figure 3.

This CPC anomaly shows that the separability margin measured with cosine is a necessary but not a sufficient quality characterisation of embedding space for clustering-based diarization. The autoregressive GRU-based architecture of CPC may encode speaker structure in higher dimensional distributional patterns, such as local temporal relationships, a phonetic-speaker interaction, or distributional manifold geometry, which are not apparent from pairwise cosine distance but are still somewhat exploited by the clustering algorithm using the complete set of pairwise distances. This finding leads to examining more advanced embedding quality measures than pairwise cosine distance, such as measures of distributional divergence between speaker-specific distributions, or separability analysis on manifolds.

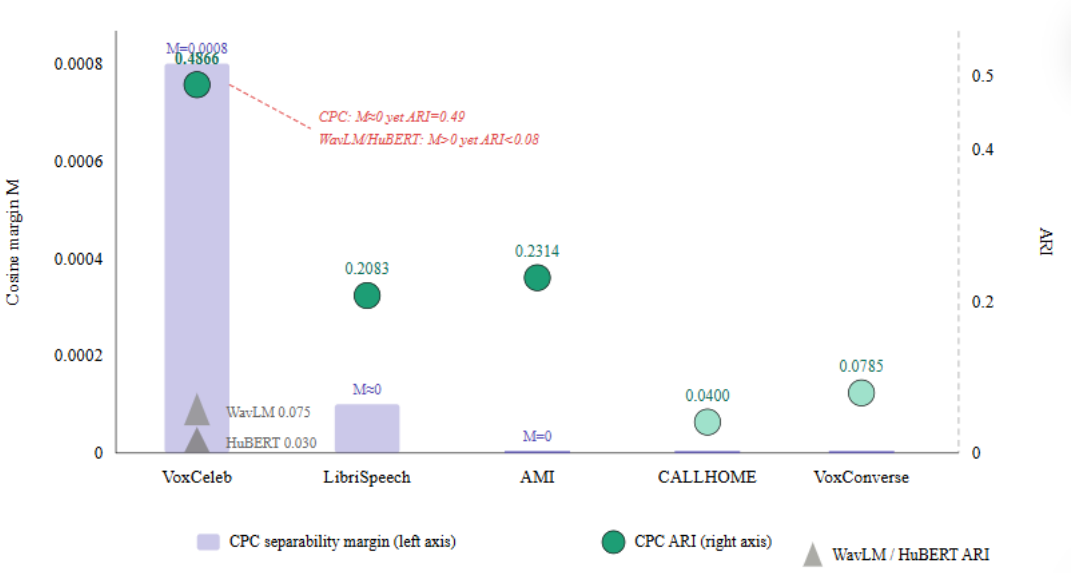

Fig. 3. CPC separability margin (near zero across all datasets, left axis) against ARI (right axis). The apparent paradox — competitive ARI values (particularly 0.4866 on VoxCeleb) despite zero cosine margin — demonstrates that standard cosine geometry does not fully characterise the embedding information available to the clustering algorithm. WavLM and HuBERT, with non-zero cosine margins, are shown for comparison; both achieve lower ARI than CPC on VoxCeleb.

VI. DISCUSSION

A. Why Inter-Speaker Similarity is the Decisive Property

The results demonstrate that intra-speaker similarity is nearly useless to predict the diarization performance ($\rho \approx 0.31$), while inter-speaker similarity has good predictive power ($\rho = 0.94$), thus clarifying the longstanding ambiguity of speaker embedding evaluation. Intra-speaker similarity is important for consistent speaker representation to ensure that the embeddings of one speaker are close together. However, compactness is not sufficient: all speakers can be compact, but the clusters do not overlap, then clustering will not work. Between cluster terms (inter-speaker similarity) are the more important terms and the distance between within cluster and between cluster are used to determine the feasibility of clustering.

This is the reason for the higher accuracy of ECAPA-TDNN. Interestingly, the objective function used for training it, AAM-Softmax, is explicitly designed to ensure that angles between speaker classes are insufficiently separated [6] and therefore explicitly optimises inter-speaker distance, which is the distance which our analysis here has shown to be important. The objective is not to maximize intra-speaker coherence and in fact, the intra-speaker similarity of ECAPA-TDNN is 0.611 which is the lowest among all five models. The training goal emphasizes discrimination, rather than compaction, which means that the embedding space is designed in such a way that the decision boundary is easy to determine, even when there are variations within speakers. This insight has an immediate implication for SSL training: directly cope with the inter-speaker repulsion deficit as identified by the separability margin analysis for SSL contrastive objectives as proposed by Lepage and Dehak [23] and Miara et al. [24].

B. The Threshold Effect

When results are computed for each dataset, ECAPA-TDNN shows the following: margins above around 0.50 always lead to perfect diarization (0.00% DER-like) for the three structural and reading corpora VoxCeleb, LibriSpeech, and AMI. Below this threshold, however, the performance of diarization is competitive but not perfect (DER-like 8.06%) below this threshold as on CALLHOME (M = 0.322). ECAPA-TDNN only achieves 57.13% DER-like with a significant margin of 0.416, VoxConverse overturns this trend. This exception validates the idea that the separability margin is a measure of the quality of the embedding's geometry, and not of the difficulty of the corpus, as even ECAPA-TDNN is not enough for VoxConverse's within-speaker acoustic variability. The margin framework thus proposes two types of failure mode: SSL failure mode in which inter-speaker repulsion is not enough; VoxConverse failure mode in which intra-speaker consistency is not enough, in the case of extreme acoustic variability.

C. Beyond Cosine Distance

The CPC anomaly implies that one of the intrinsic constraints of using cosine distance as a complete embedding space geometry characterization is that. CPC does not actually need to have a low cosine margin to achieve competitive clustering; the embeddings, rather, have to encode speaker identity in a higher dimensional distributional structure that is not captured by these average pairwise cosine similarities (intra and inter). Despite the fact that CPC's speaker information is contained in patterns that are not explicitly reflected in the average intra- and inter-speaker similarity that defines the margin, still, a zero value for the cosine margin means that CPC can partially exploit that information through the complete collection of pairwise distances, using agglomerative clustering.

These could be related to higher-order moments of the intra and/or inter-speaker similarity distributions (e.g., variance, skewness) rather than to means, or block structure in the local neighbourhood structure revealed by the full set of pairwise distances despite equal global mean speaker similarity, or distributional geometry of speaker-specific clusters (e.g., linear subspace structure, Riemannian manifold curvature) that cosine distance averages out. Future work that could be done is distinguishing between these, which would need more detailed analysis of the geometry of the space of CPCs.

D. Implications for Embedding Design

The separability margin framework provides actionable guidance in embedding model development. The primary take-home message for SSL models is that they are modeled for inter-speaker repulsion, and current SSL objectives (contrastive prediction, masked unit prediction, denoising), all result in intra-speaker similarities near 1, meaning that the within-speaker consistency is high. The only deficit is inter-speaker separation with SSL margins of 0.000 - 0.023, essentially making the clustering problem geometrically ill-posed. This simplest motivated intervention is to add the angular margin losses to SSL training, like the AAM-Softmax objective of ECAPA-TDNN, which will implicitly penalise small distance between speakers without the need for labelled speaker identities through pseudo-labels or self-distillation [26, 27].

An additional implication is on the assessment of new embedding models in diarization. The current practice is to submit models for full evaluation of the diarization system and this process is time consuming and it hides the embedding quality from the diarization system design decisions. Evaluating the separability margin using a standard evaluation set gives a direct and system independent diagnostic which can predict the diarization performance to be $\rho > 0.96$. We believe the separability margin should be considered a standard embedding quality measure in speaker embedding evaluation protocols, in addition to EER.

VII. CONCLUSION

In this paper, a theoretically motivated criterion, the cosine separability margin, has been proposed which can then be directly measured to predict the clustering-based speaker diarization performance. The margin exhibits near-perfect Spearman rank correlation scores with all four diarization clustering metrics for five embedding

models and five benchmark datasets ($\rho = -0.97$ to -0.98 with purity and ARI, $\rho = 0.96$ with DER-like), which validate the margin as the key embedding-level geometric property for diarization.

Four conclusions come to the fore. The number one geometric parameter of diarization is inter-speaker similarity, not intra-speaker similarity: models with high intra but high inter (all SSL models) do not cluster, and ECAPA-TDNN's good performance is due to driving inter-speaker similarity close to zero. Second, the geometric property of the training objective of ECAPA-TDNN, which explains the categorical advantages of ECAPA-TDNN over SSL models without having to resort to any architectural arguments. Third, the CPC anomaly (competitive diarization despite zero cosine margin) also shows that cosine distance is not sufficient measure of embedding space quality for clustering-based diarization and the need for more comprehensive geometric diagnostics. Fourth, the margin framework is a proper one that reflects the challenge of SSL: SSL pretraining tasks should involve inter-speaker repulsion using angular margin losses or other geometric constraints.

In order to complement DER in model development of diarization, a new system-independent and predictively valid standard diagnostic measure, Cosine Separability Margin is proposed that is fast to compute. .

REFERENCES

[1] X. Anguera et al., "Speaker diarization: A review of recent research," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 2, pp. 356–370, 2012.
[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," Comput. Speech Lang., vol. 72, p. 101317, 2022.
[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.
[4] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Proc. NeurIPS, vol. 14, 2002, pp. 849–856.
[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification," in Proc. Interspeech, 2020, pp. 3830–3834.
[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF CVPR, 2019, pp. 4690–4699.
[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, vol. 33, 2020, pp. 12449–12460.
[8] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 3451–3460, 2021.
[9] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE J. Sel. Topics Signal Process., vol. 16, no. 6, pp. 1505–1518, 2022.
[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv:1807.03748, 2018.
[11] Z. Li and M.-W. Mak, "Speaker representation learning via contrastive loss with maximal speaker separability," in Proc. IEEE ICASSP, 2022, pp. 7542–7546.
[12] M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," Eng. Appl. Artif. Intell., vol. 127, p. 107232, 2024.
[13] E. P. C. Lee, G. Sun, C. Zhang, and P. C. Woodland, "Spectral clustering-aware learning of embeddings for speaker diarisation," in Proc. IEEE ICASSP, 2022, pp. 7927–7931.
[14] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in Proc. Interspeech, 2018, pp. 1086–1090.
[15] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in Proc. Interspeech, 2017, pp. 2616–2620.
[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in Proc. IEEE ICASSP, 2015, pp. 5206–5210.
[17] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in Proc. MLMI, 2005, pp. 28–39.
[18] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English speech," LDC, 1997.
[19] J. S. Chung et al., "Spot the conversation: Speaker diarisation in the wild," in Proc. Interspeech, 2020, pp. 299–303.
[20] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.
[21] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," J. Mach. Learn. Res., vol. 3, pp. 583–617, 2002.
[22] L. Hubert and P. Arabie, "Comparing partitions," J. Classification, vol. 2, pp. 193–218, 1985.
[23] T. Lepage and R. Dehak, "Additive margin in contrastive self-supervised frameworks to learn discriminative speaker representations," arXiv:2404.14913, 2024.
[24] M. Miara, M. Ravanelli, and F. Landini, "Towards supervised performance on speaker verification with self-supervised learning," in Proc. Interspeech, 2024.
[25] N. Dawalatabad et al., "ECAPA-TDNN embeddings for speaker diarization," in Proc. Interspeech, 2021, pp. 3560–3564.
[26] J. Han et al., "Leveraging self-supervised learning for speaker diarization," in Proc. IEEE ICASSP, 2025, pp. 1–5.
[27] S. Baroudi, H. Bredin, J. Razik, and R. Marxer, "On the use of self-supervised representation learning for speaker diarization and separation," arXiv:2512.15224, 2025.
[28] X. Liu et al., "Self-supervised learning: Generative or contrastive," IEEE Trans. Knowl. Data Eng., vol. 35, no. 1, pp. 857–876, 2021.
[29] A. Mohamed et al., "Self-supervised speech representation learning: A review," IEEE J. Sel. Topics Signal Process., vol. 16, no. 6, pp. 1179–1210, 2022.