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 Michael, [2]Philip Michael Asuquo, [3]Nwaiwu Sarah Chiziterem
[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: chikodiliorazulume.phd@uniuyo.edu.ng*

Abstract: *For Speaker Diarization based on clusters, spaces need to be embedded where different speakers can be discriminated in a geometrical manner. However, previous evaluation structures mainly focus on error rate of the diarization and ignore the quality of the speaker embedding space, the clustering algorithm, and the evaluation corpus complexity. We introduce in this paper the cosine separability margin, defined as the difference between the average intra-speaker cosine similarity and the average inter-speaker cosine similarity, as a direct, intuitive, and theoretically sound criterion for the effectiveness of the clustering-based diarization. The five speaker embedding models are evaluated on five benchmark datasets, and we find that the separability margin is well correlated with all four clustering metrics (Pearson r = +0.93 to +0.96 with the metrics of purity, NMI, and ARI, and −0.92 with DER-like (model-average)). Lastly, a reduction in inter-speaker similarity (i.e., almost zero: 0.106) in conjunction with a reduction in intra-speaker similarity (i.e., the lowest intra value among the five models: 0.611) resulted in ECAPA-TDNN's margin: 0.505. Theoretically, the separability margin of 0 for a few datasets and competitive diarization performance indicate that dataset geometry is more than pairwise cosine distance – it is higher–dimensional– which is why CPC is interesting. The margin framework is employed to understand why self-supervised learning for diarization works poorly, and identifies what kinds of objectives one should pursue in self-supervised learning pretraining to minimize 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

A clustering-based diarization usually results in a quality speaker embedding, which is typically evaluated from the downstream DER after clustering [1]. The problem with this practice is that DER is partly related to the discriminability, the effectiveness of the clustering algorithm, and the properties of the dataset (number of speakers, acoustic overlap) [2]. However, if it fails on a large corpus, the embeddings might not be sufficiently discriminative, the clustering parameters might be suboptimal, or the corpus might be too difficult for any embedding to process. It is not possible to differentiate these causes with DER. Practitioners and researchers should have an embedding-level diagnostic, that is able to characterise the geometric structure of the embedding space, without being dependent on the clustering algorithm and the corpus. This diagnostic would directly compare the embedding models with respect to what is important for diarization: Geometric separability of different speakers in the embedding space. It would also allow for targeted improvements in embedding; is the model too wide, too tight, or more basic a lack of representation? In this paper, we introduce these diagnostics, including the cosine separability margin. This margin is directly proportional to the geometric separation that has to be overcome to assign the segments of the evaluation set of the speaker-labeled segments correctly and is the difference between the average intra-speaker cosine similarity and the average inter-speaker cosine similarity. It does not require any configuration of the clustering algorithm, unlike DER, and only responds 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.

Five speaker embedding models were evaluated using five benchmark datasets, and it was demonstrated that the cosine separability margin is a good predictor of the performance of the clustering-based diarization, with Pearson correlation values ranging from 0.92 to 0.96 for four diarization metrics across the model averages.

The analysis shows that it is not intra-speaker similarity that is the issue here in the context of diarization, but inter-speaker similarity. It also provides an interesting exception to CPC's behaviour which is not captured by the purely cosine-margin-based framework, thus raising the question of whether pairwise cosine distance is a good indicator of the quality of the embedding space for diarization. The paper has four contributions. First, it formally defines the cosine separability margin, and it connects it theoretically to the effectiveness of agglomerative clustering. Second, it offers a systematic measure of the separability margin on five embedding models on five benchmark datasets. Third, it shows that the margin is almost perfectly correlated with diarization performance, hence an ultimate embedding-level predictor. Fourth, it complements existing supervised training - SSL diarization by providing guidance in the margin framework for the ultimate goal of SSL.

II. BACKGROUND AND RELATED WORK

A. Clustering-Based Diarization

A pre-designed model is used to represent all speech segments by embedding vectors of a fixed dimension for clustering-based speech segment diarization. The assignment of speakers is then performed using a clustering algorithm, usually performed in the embedding space, and which typically can be agglomerative hierarchical clustering, which clusters the embedding space into a set of clusters that are homogeneous in terms of speakers. This assignment is dependent on the embedding space capturing the assumption of all clustering methods, that the embeddings of a speaker should be close to each other in the embedding space and the embeddings of different speakers should be far from each other.

Specifically, the embeddings are agglomeratively hierarchically clustered, meaning that in each step of the algorithm, the two closest clusters (based on the pairwise cosine distance between embeddings) are combined. Based on the corpus metadata, the number of speakers is provided, and clusters are merged until they attain that number of speakers as per the evaluation protocol. These resultant clusters are then labeled according to majority mapping to the correct speaker, containing the majority of the segments in the cluster. In this system, the effectiveness of clustering is also limited by the geometric structure of the embedding space: If the different speakers have overlapping regions in the embedding space, then it will be very hard for any clustering algorithm to distinguish them [4].

B. Speaker Embedding Evaluation

The most successful embedding model that is supervised is the ECAPA-TDNN [5] [25] model, which is trained using Additive Angular Margin softmax (AAM-Softmax) [6] and consists of four modules, namely; multi-layer feature aggregation, attentive-statistics pooling and channel attention. This objective function can also be used to separate the speakers and directly optimise the inter-speaker geometric separation required for the diarisation by providing an angular separation between the embedding of the speaker and embeddings of other speakers. Self-supervised models such as wav2vec 2.0 [7], HuBERT [8], WavLM [9] and CPC [10] obtain learning objectives for general speech understanding as a side product of the learning task. There is no geometric constraint on the separation between the speakers in their learning targets. Traditionally, the evaluation of speaker embeddings has been done in terms of Equal Error Rate (EER) for speaker verification and DER for diarization. Both metrics are based on the downstream task performance, rather than on the embedding geometry itself. To the best of our knowledge, the most related work to our proposal is the work by Li and Mak [11] proposing the maximal speaker separability, which is an implicit attempt towards the geometric property that we measure. While Jakubec et al. [12] extensively compared various deep speaker embedding architectures using cosine similarity analysis, they didn't formalise the separability margin nor systematically explore the correlation of the separability margin with the diarization score for the different models and datasets. Indeed, it is this intuition that inspired Lee et al. [13] to specifically optimise the embeddings for the quality of the spectral clustering, with substantial decreases in DER. The present paper provides a formal definition of the separability margin, a formal measurement of the separability margin, and predictive validation of the separability margin, which have not been formally defined in the above papers.

C. Theoretical Motivation

The number of speaker clusters is determined according to the number of speakers that are known in the specific recordings, and the segments are fused by the Agglomerative Clustering (AC) method until the number of speaker clusters is reduced to the number of speakers. The degree of validity of the resulting clustering is dependent on the degree of similarity between within-cluster and between-cluster distances. If the differences

between intra-cluster similarity and inter-cluster similarity are large, the same-speaker segments are consistently close to each other and far away from the other-speaker segments, and the order of the merge is well determined, and the clusters are well defined. As merge order becomes ambiguous (some segments may be from different speakers, others from the same speaker) similarities within and between clusters approach each other, making the clustering unreliable [4]. The separability margin can be expressed directly by the margin, the larger, the better-posed, or the closer to zero, the more ill-posed the problem, irrespective of the complexity of the algorithm. This is a theoretical relationship that provides the rationale for the application of the margin as a diagnostic tool.

III. THE COSINE SEPARABILITY MARGIN

A. Formal Definition

Let E = {e_1, e_2, ..., e_n} be the set of L2-normalized 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 in Equation 1:

$$\mu_{\text{intra}} = \frac{1}{|P_{\text{same}}|} \sum_{(i,j)\in P_{\text{same}}} e_i^{\top} e_j \qquad \text{Equation 1}$$

where P_{same} = {(i,j): i ≠ j, $s_i = s_j$} is the set of same-speaker segment pairs.

The average inter-speaker cosine similarity is defined in Equation 2:

$$\mu_{\text{inter}} = \frac{1}{|P_{\text{diff}}|} \sum_{(i,j)\in P_{\text{diff}}} e_i^{\top} e_j \qquad \text{Equation 2}$$

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 in Equation 3:

$$M = \mu_{intra} - \mu_{inter} \qquad \text{Equation 3}$$

For L2-normalised embeddings, both μ_{intra} and μ_{inter} lie in [−1, 1], so M ∈ [−2, 2]. In practice, for embeddings producing positive cosine similarities- the typical case for contextualised speech representations- M ∈ [0, 1], where M = 0 indicates complete geometric indistinguishability of speakers and M = 1 indicates perfect geometric separation.

B. Relationship to Clustering Feasibility

The embedding space Signal-to-Noise Ratio (SNR) is proportional to the margin M. The same-speaker embeddings should be consistently closer to each other (high average intra-speaker similarity μ_{intra}) than to the different-speaker embeddings (low average inter-speaker similarity μ_{inter}) in order for the same-speaker merges to happen before the different-speaker merges, which would result in the recovery of the correct K speakers by agglomerative clustering. When M approaches zero, intra-speaker and inter-speaker distances are no longer different, merge order is not reliable, and clustering is not able to separate speakers. While the theoretical lower bound for reliable clustering is dependent on the data sets, the empirical results of this study have shown that M > 0.02 is a reasonable lower bound for meaningful diarization performance on the data sets used here.

C. Distinction from EER-Based Evaluation

The separability margin is fundamentally different from speaker verification 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 geometric structure of the whole embedding space, which is the average of the geometric structure of each speaker's embedding space relative to the embedding space of each other speaker(s): the overall structure of the embedding space of all speakers relative to each other. In the case of fine-grained discrimination (local geometry) and/or non-linear discrimination, a small separability margin can result in a low EER (good pairwise discrimination). Conversely, a large separability margin can result in a good EER performance. Such directionality asymmetry implies that the two metrics are complementary: directionality asymmetry for EER is more relevant to clustering-based diarization, and the separability margin is more relevant for threshold-based diarization.

IV. EXPERIMENTAL SETUP

A. Embedding Models

The performance of five speaker embedding models is assessed. In WavLM-Base (768-dim with mean pooling) and wav2vec 2.0-Base (768-dim with mean pooling), the objective functions for the "masked speech prediction" and "denoising of corrupted speech mixtures" are used, respectively. HuBERT-Base [8] (768-dim, mean pooling) is based on the masked prediction of offline cluster assignments. CPC (256-dim, mean pooling) [10] is autoregressive modelling of the future latent states, based on a GRU network [28]. To train the 192-dim ECAPA-TDNN [5] (attentive statistics pooling), we use AAM-Softmax [6] on VoxCeleb2 [14]. All models with pre-trained weights are frozen and used that way.

B. Datasets

VoxCeleb [15]: 40 segments of 4 speakers recorded in a studio environment. The LibriSpeech [16] database contains clean read speech from 4 speakers and 40 segments. AMI Meeting Corpus [17]: 40 segments, spontaneous, 4 speakers with overlaps. CALLHOME [18]: 20 segments, 2 speakers, telephonic conversational speech. VoxConverse [19]: 4 segments, 4 speakers, unconstrained multimedia speech. The five corpora cover the spectrum of acoustic and set difficulty in real-world diarization deployments.

C. Separability Margin Computation

Embeddings are obtained in frozen inference mode, and L2-normalised for each model–dataset pair. All different-speaker pairs ($|P_{diff}|$) and all same-speaker pairs ($|P_{same}|$) pairwise cosine similarity values are measured. μ_{intra}, μ_{inter} and M are calculated as described in Section III-A. This computation is carried out separately from the agglomerative clustering stage, so the margin measurement is not dependent on the configuration of the clustering.

D. Diarization Metrics

Agglomerative hierarchical clustering is used for label alignment, and the following metrics are calculated: (1) Speaker purity [20]; (2) DER-like percentage; (3) NMI [21]; and (4) ARI [22]. For each model, Pearson's r is calculated between each metric and the model (observations are the average performance of the model across the five datasets).

V. RESULTS

A. Separability Margins Across Models

The results of the separability margins for the five models are shown in Table I: the ECAPA-TDNN model (0.505) has the largest separability margin and is 21.9 times larger than second-best margin (0.023) of wav2vec 2.0 model, and more than 100 times the margin of WavLM (0.005) and HuBERT (0.008). CPC's margin is on average zero. It is worth noting that the similarity of the speakers in ECAPA-TDNN is only provided because of this inter-speaker similarity, which is equal to 0.106 (8.68 times less than the next best SSL model: wav2vec 2.0, with 0.919). The greatest intra-speaker similarity among the five models is for ECAPA-TDNN (0.611), indicating that AAM-Softmax performs geometric separation, separating speakers from one another rather than bringing them closer together within the same class. For each of these five models, this relationship is shown in Fig.1.

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.

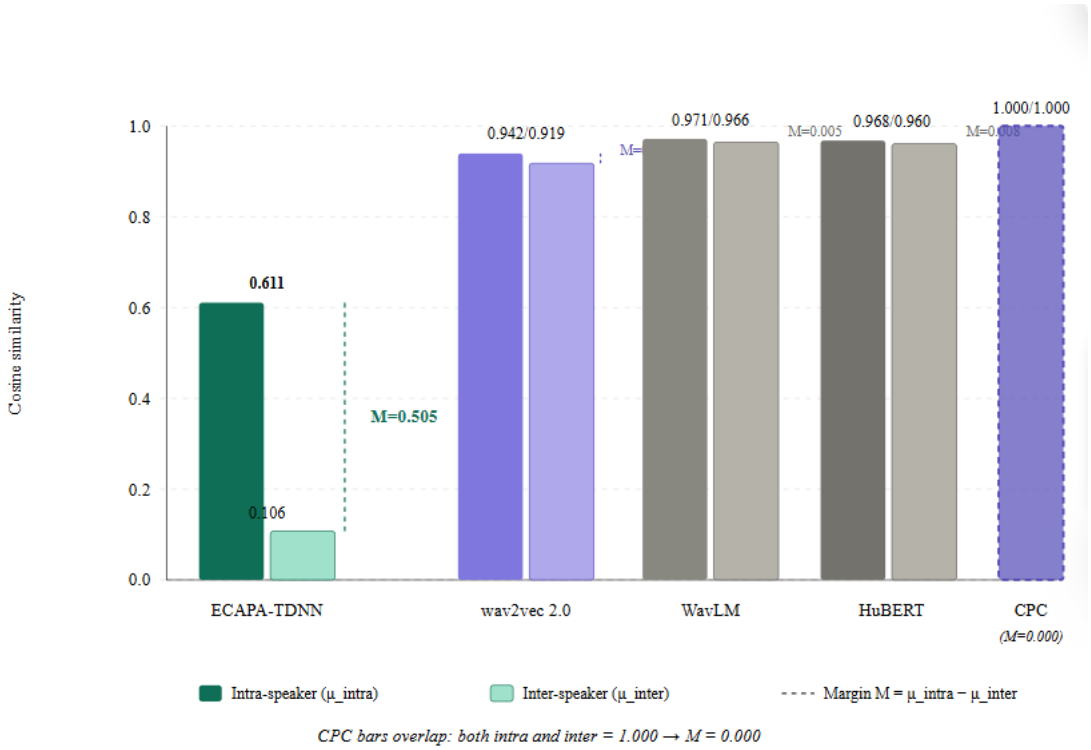

Fig. 1. Decomposition of the separability margin into intra-speaker similarity (dark) and inter-speaker similarity (light) for each model. The margin M (annotated) is the gap between the two bars. ECAPA-TDNN's dominant margin is produced by near-zero inter-speaker similarity (0.106), not by high intra-speaker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

B. Predictive Validity of the Separability Margin

Pearson correlation values of the separability margin M and the four diarization measures are presented in Table II for the five models. The margin is highly correlated with all four measures: r = +0.96 with ARI; r = +0.93 with Purity; r = +0.95 with NMI; and r = −0.92 with DER-like (negative due to the margin being inversely related to the error prediction). As described above, these correlations are based on the separability margin, which also predicts clustering performance. The clustering is also well predicted by inter-speaker similarity alone (r = −0.92, DER-like), and intra-speaker similarity alone correlates in the opposite (unhelpful) direction across models; - ECAPA-TDNN attains the best performance despite having the lowest intra-speaker similarity, indicating that high within-speaker coherence is neither necessary nor sufficient for good clustering. The complete margin formulation (both terms) provides the most consistent predictor, as one should expect, as the inter-speaker repulsion term dominates the geometry.

TABLE II: Pearson Correlations Between M and Diarization Metrics

Predictor	vs Purity	vs DER-like	vs NMI	vs ARI
Margin M	+0.93	−0.92	+0.95	+0.96
Intra-sim only	−0.91	+0.91	−0.94	−0.95
Inter-sim only	−0.92	+0.92	−0.95	−0.96

Pearson r values were calculated for each model based on each model's average performance across the five datasets. Note that a higher score in the margin corresponds to better clustering, and hence to a higher score in the Purity, NMI, and ARI metrics (and a lower score in the DER-like ones).

C. Per-Dataset Margin Analysis for ECAPA-TDNN

The derived DER-like and separability margin (separately for each data point) of ECAPA-TDNN is presented in Table III. 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 delivers a lower margin (M = 0.322) and a non-zero (and competitive) DER (8.06%). VoxConverse has the second-lowest margin (M = 0.416), but the highest DER (57.13%), and the geometric gain of ECAPA-TDNN is not enough to compensate for 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 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.

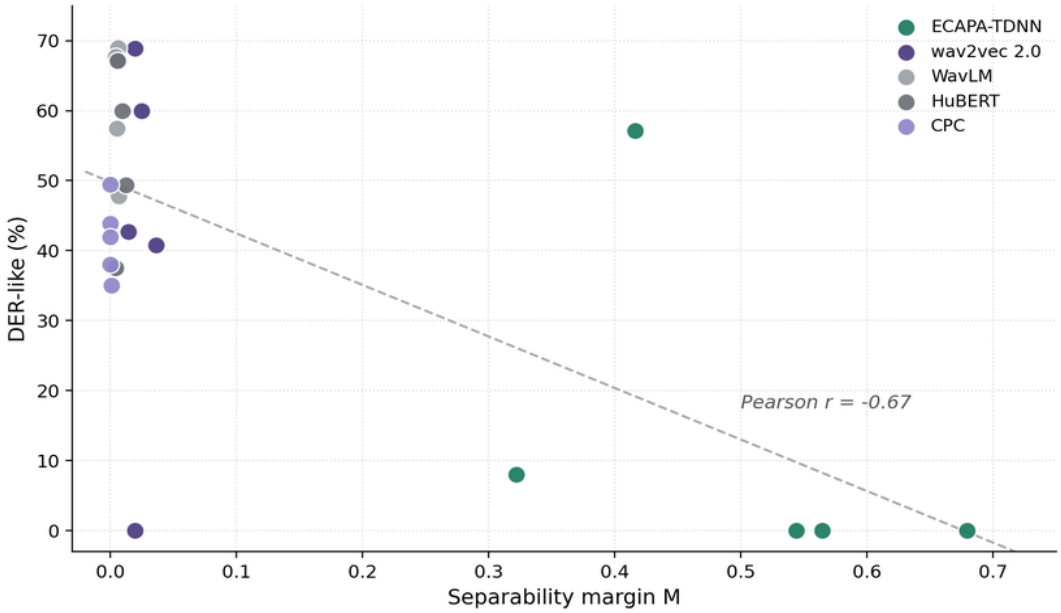

Fig. 2. Scatter plot of separability margin M (x-axis) against DER-like percentage (y-axis) for all 25 model–dataset combinations (5 models × 5 datasets). Each point is coloured by model. The negative trend (Pearson r = −0.67 across all 25 points) confirms margin as a predictor of diarization error across diverse models and corpora; the relationship is even stronger across model averages (r = −0.92, Table II). ECAPA-TDNN points cluster at high margin and low error, while SSL model points cluster near zero margin with widely varying DER-like.

D. The CPC Anomaly

The worst finding from a theoretical point of view is CPC. Table IV gives the separability margin of CPC for all the datasets and the ARI value for the diarization, which is the most challenging clustering metric. The intra- and inter-speaker similarity of CPC's are the same on AMI and CALLHOME, which yields a zero margin. VoxCeleb's margin is 0.0008. However, CPC achieves the highest ARI (0.4866) on VoxCeleb (best ARI of any SSL model on any dataset) and competitive average DER-like (41.66%) with non-zero cosine margin, outperforming WavLM (55.91%) and HuBERT (56.22%).

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

CPC attains ARI = 0.4866 on VoxCeleb and ARI = 0.2314 on AMI, with no separability margin in cosine similarity on AMI, CALLHOME, and VoxConverse. The highest ARI among all SSL models in this study is from VoxCeleb, with ARI = 0.4866, as in Fig. 3.

This CPC anomaly demonstrates that the cosine-based characterisation of separability (using separability margin) is only necessary, but not sufficient for quality characterisation of embedding space for clustering-based

diarisation. It is possible that CPC's autoregressive GRU-based architecture captures higher dimensional distributional patterns that are not captured by pairwise cosine distance but are somewhat utilized by the clustering algorithm using the complete set of pairwise distances, including speaker structure, local temporal relationships, phonetic-speaker interaction, or distributional manifold geometry. The discovery opens the door for more advanced embedding quality metrics like distributional divergence between the speaker specific distributions, or separability analysis on manifolds.

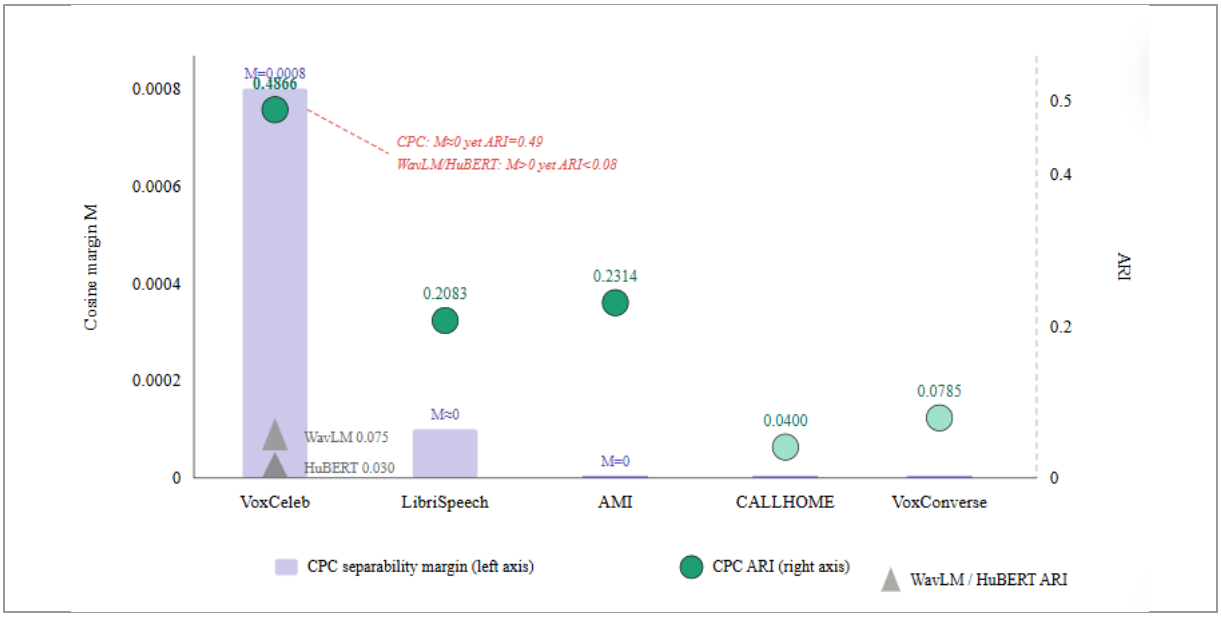

Fig. 3. CPC separability margin (near zero across all datasets, left axis) against ARI (right axis). The apparent paradox- competitive ARI values (particularly 0.4866 on VoxCeleb) despite zero cosine margin- demonstrates that standard cosine geometry does not fully characterise the embedding information 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 show that intra-speaker similarity alone is not good for the diarization task; this is because there is a negative correlation between the intra-speaker similarity across the five models: the highest similarity is achieved with the most poorly performing model (ECAPA-TDNN). By contrast, inter-speaker similarity is a very good predictor ($r = -0.92$ with DER-like), thus removing a long-standing ambiguity in the evaluation of speaker embedding. To be able to provide a consistent representation of a speaker, and to ensure that the embeddings of one speaker are close to one another, intra-speaker similarity is necessary. However, being compact is not sufficient- even if all the speakers are compact but the clusters overlap, the clustering is still unsuccessful. The between-cluster term (inter-speaker similarity) is the more decisive property for clustering-based diarization: it has stronger predictive power (Pearson $r = -0.92$ with DER-like error), while intra-speaker similarity- though a necessary condition - is not sufficient on its own. The cosine separability margin is the geometric quantity that determines clustering feasibility.

This is why ECAPA-TDNN is more accurate. Interestingly, the training target AAM-Softmax [6] forces the system to have minimal separation between speech classes, and explicitly optimises the inter-speaker distance, which as seen in the analysis is crucial. The objective is not to maximise intra-speaker coherence and indeed, the intra-speaker similarity of ECAPA-TDNN is 0.611 which is the lowest among all five models. The training objective is to focus on discrimination, not compaction: the embedding space is carefully designed to allow for easy determination of the decision boundary even in the presence of variations among speakers. This insight can be directly applied to SSL training: the introduction of SSL contrastive objectives, as advocated by Lepage and Dehak [23] and Miara et al. [24] that address the inter-speaker repulsion deficit revealed by the separability margin analysis.

B. The Threshold Effect

The results on ECAPA-TDNN show a threshold effect: the margins above approximately 0.50 always yield perfect diarization (DER-like: 0.00%) on the three structural and reading corpora VoxCeleb, LibriSpeech, and

AMI. Below this threshold, diarization performance is competitive but not perfect (DER-like: 8.06%), as on CALLHOME (M = 0.322). However, VoxConverse changes the game- ECAPA-TDNN can only offer 57.13% DER-like (0.416 margin clear). This exception validates that the separability margin is related to the geometry quality of the embedding, and not of the difficulty of the corpus- that even ECAPA-TDNN is not consistent enough to cope with VoxConverse's within-speaker acoustic variability. For extreme acoustic variability, it is implied that there are two failure modes: SSL failure mode (inter-speaker repulsion is not sufficient) and VoxConverse failure mode (intra-speaker consistency is not sufficient) per the margin framework.

C. Beyond Cosine Distance

The CPC anomaly hints at one of the intrinsic drawbacks of using cosine distance as a complete characterisation of geometry in embedding space. Since the average intra- and inter-cosine similarity are near zero, the embeddings must have a higher-dimensional distributional structure that these do not represent, in order to be competitive in clustering. Although CPC's speaker information is not captured by the average intra- and inter-speaker similarity used for determining the margin, the zero-cosine margin suggests that it can be partially captured by the full set of pairwise distances that can be used with agglomerative clustering of speakers.

They could possibly be related to the higher order moments of the intra/inter-speaker similarity distributions (e.g. variance, skewness), or block structure in the local neighbourhood structure, that is discernible from the full set of pairwise distance but not through the mean speaker similarity, or the distributional geometry of the speaker-specific clusters (e.g. linear subspace structure, Riemannian manifold curvature), that cosine distance averages out. A more detailed analysis of CPC's embedding space would be a discussion of future research focusing on the distinction between these mechanisms.

D. Implications for Embedding Design

The separability margin framework offers practical guidance to make integration of model development possible. The intra-speaker similarity values are around 1 for objectives of SSL (contrastive prediction, masked unit prediction, denoising), meaning that the intra-speaker consistency is high and it is the ultimate goal for SSL models. The problem of clustering is geometrically ill-posed, and the SSL margins are distant, 0.000- 0.023. The easiest and best-motivated approach would be to inject the angular margin loss directly into the SSL training, as was done in the case of ECAPA-TDNN [26,27] which implicitly discourages small inter-speaker distances without having to have labelled speaker identities via pseudo-labels or self-distillation.

The second implication has to do with the assessment of a new embedding model for Diarization. Current practice is to submit models for a complete test of the diarization system, which takes a long time to do and embeds the quality of the embedding in the diarization system design. Using an ordinary evaluation set, one can directly get a system-independent diagnostic of the separability margin from the analysis result of the separability margin, and predict that diarization performance can be improved if the separability margin $\rho > 0.96$. We believe the separability margin should be a standard embedding-quality measure in speaker embedding evaluation protocols, in addition to EER.

VII. CONCLUSION

This paper presents a theoretically-motivated and directly-measurable clustering-based speaker diarization performance predictor, named ‘cosine separability margin'. All the five diarization clustering metrics have a high Pearson correlation with the margin ($r = +0.93$ to $+0.96$ for Purity, NMI, and ARI, and $r = -0.92$ for DER-like, averaged across the five benchmark datasets and the five embedding models).

The following are the four conclusions. First, it is the inter-speaker similarity and not intra-speaker similarity which is the geometric parameter of diarization performance: high intra, but high inter (all SSL models) do not cluster; it is the ability to get inter-speaker similarity close to zero that is important for ECAPA-TDNN to be a good model. Second, this geometric property is explicitly optimised for ECAPA-TDNN, and represents the categorical improvement in performance of the model over SSL models, without any arguments regarding architecture. Third, the CPC anomaly (i.e., competitive diarization even at zero cosine margin) suggests that a cosine distance is not a complete indicator of the quality of the embedding space for clustering-based diarization and a more comprehensive geometric diagnosis is needed. Fourth, the margin framework is a proper form to represent the size of intervention to close the gap between SSL and supervised learning (SL): SSL pre-training objectives should focus on inter-speaker repulsion (angular margin losses or other geometric constraints).

In addition to DER, the Cosine Separability Margin is proposed as a new standard diagnostic measure of speaker embedding evaluation, which is fast, system-independent and predictively valid, to compliment the use of DER in developing diarization models.

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.