Enhancing Matryoshka Representation Learning with Structured Supervision

Sunchan Park
sunchanpark@pusan.ac.kr

Contents

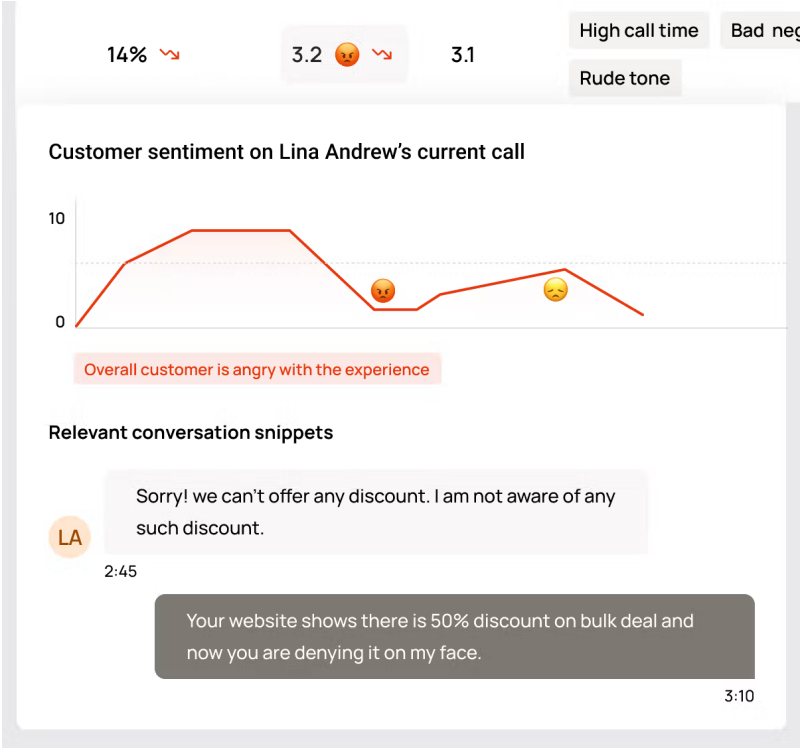
- Introduction
- Background
- Proposed Methods
- Experiments
- Conclusions

Introduction: Presenter

- Research Background
  - Automatic Speech Recognition
  - Speaker Recognition
  - Speech Emotion Recognition

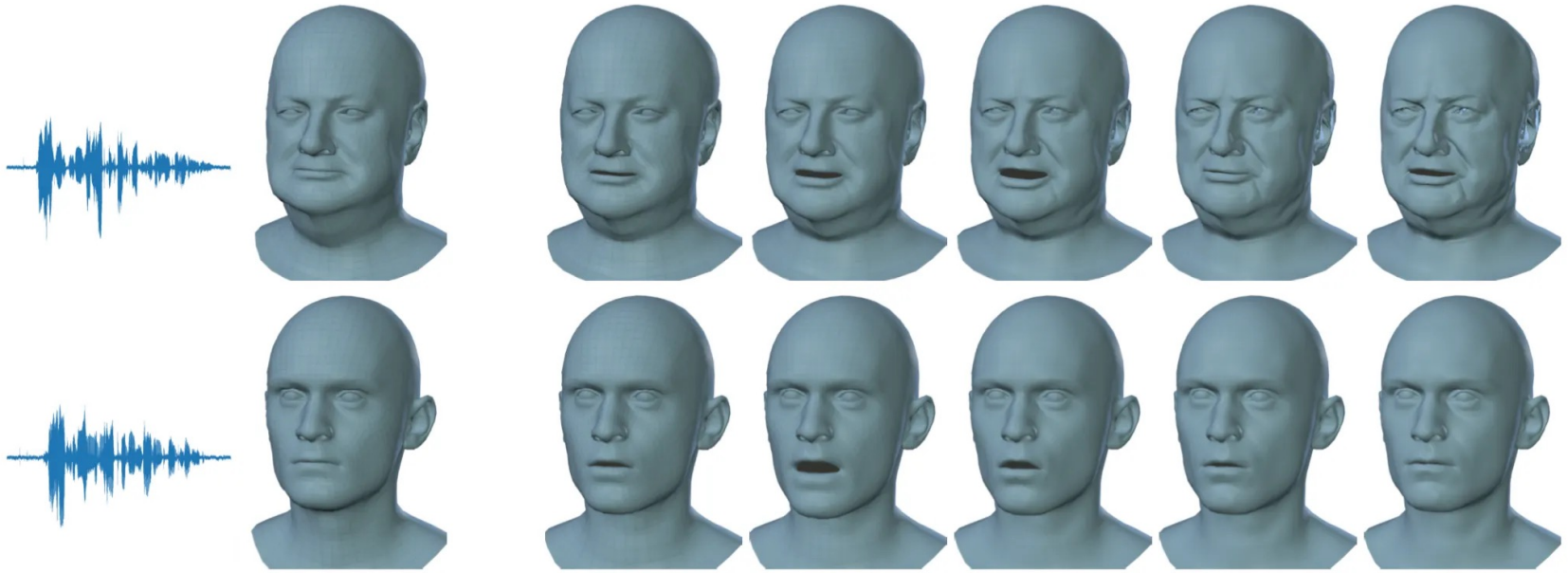






Introduction: Presenter

- Future Research Directions
  - Speech-Driven 3D Facial Animation
  - Audio-Visual Modeling
  - Multimodal LLM

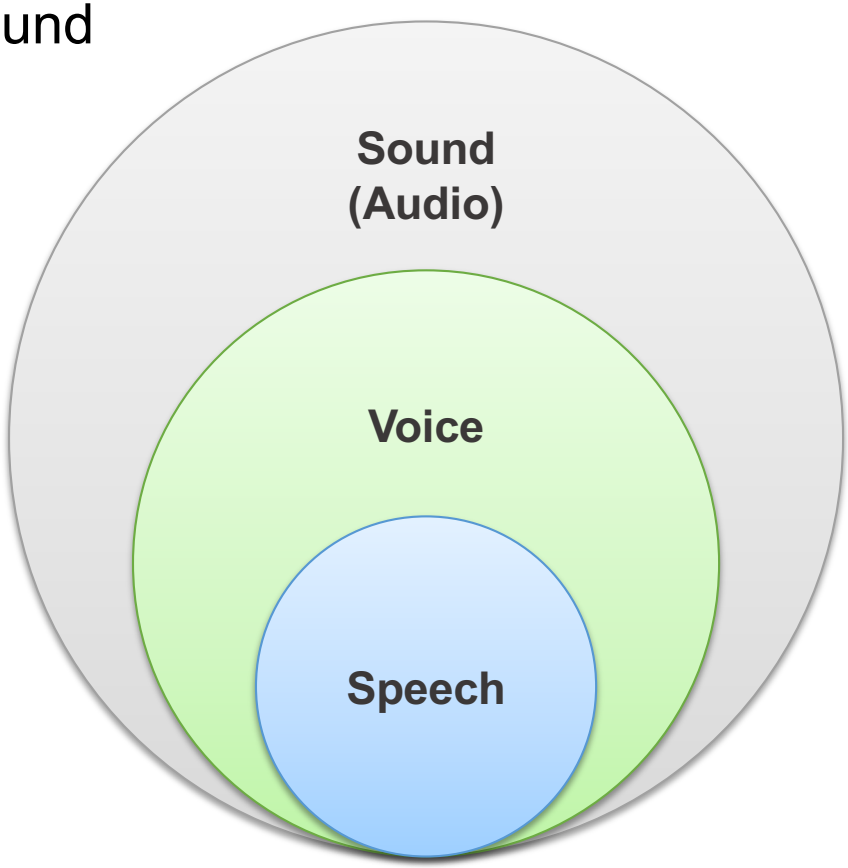


Input: speech signal and 3D template

Output: 3D character animation

Introduction: Presenter

- Hierarchy of Sound
  - Sound: Physical Phenomenon
  - Audio: Electrical Representation of Sound
  - Voice: Human-produced Sound
  - Speech: Meaningful Voice



Introduction: Overview

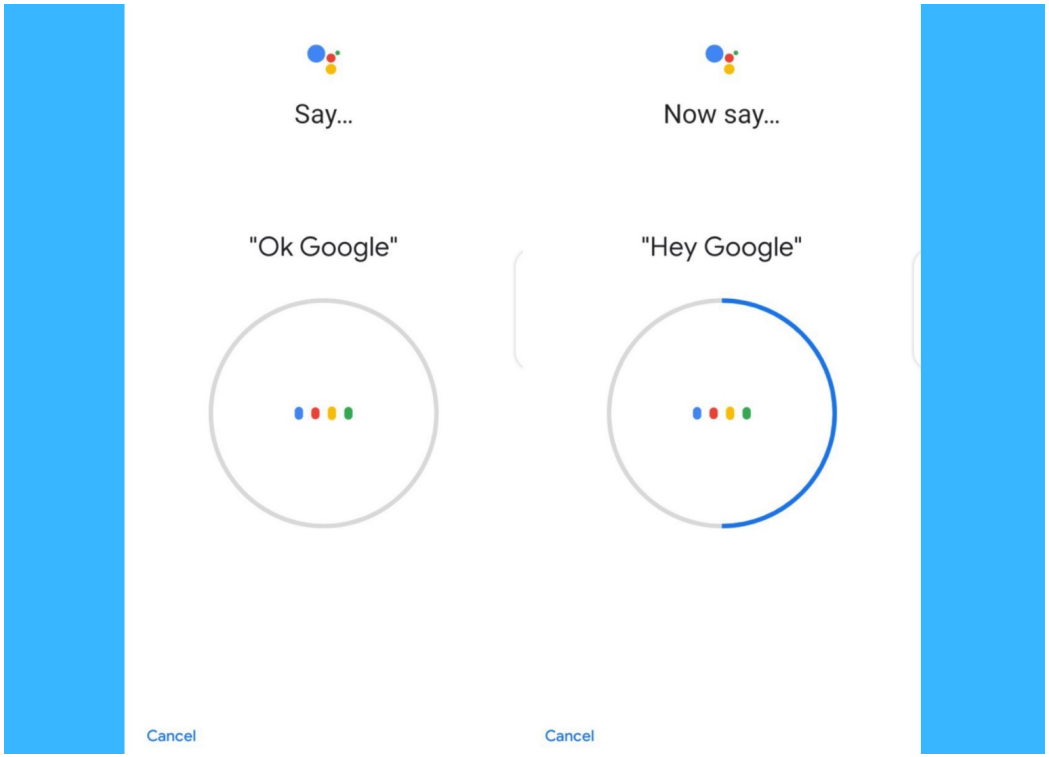
- Applying Matryoshka Representation Learning to speaker embedding improves low-dimensional performance but degrades high-dimensional performance
- To mitigate this issue, we propose following methods:
  - Dimension-Specific Margins
  - Element-wise Gradient Scaling
- Experimental results demonstrate the effectiveness of each method and their combination in the following tasks:
  - Speaker Verification
  - Open-Set Speaker Identification

Contents

- Introduction
- Background
  - Speaker Recognition
  - Deep Speaker Embedding
  - Margin-Based Softmax Loss
  - Matryoshka Representation Learning
- Proposed Methods
- Experiments
- Conclusions

Background: Speaker Recognition

- Speaker Recognition
  - Automatically recognize who is speaking
  - Use speaker-specific information included in speech
  - Text-dependent vs. Text-independent



Background: Speaker Recognition

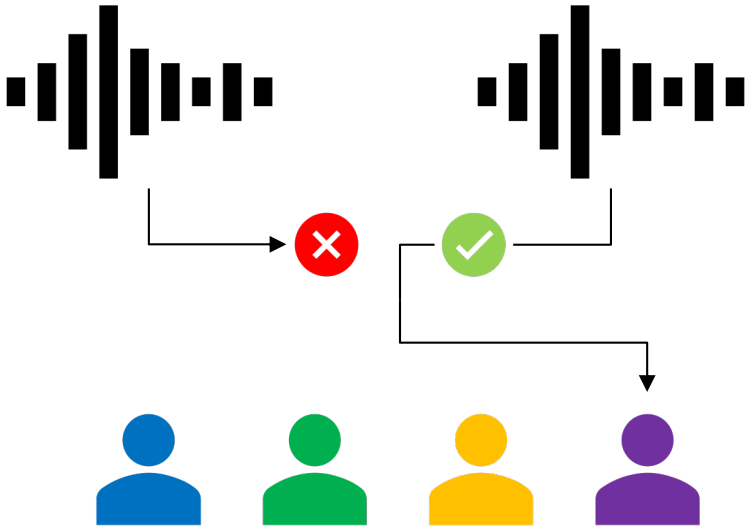
- Speaker Verification
  - Verify whether the speaker matches the claimed identity
- Speaker Identification
  - Identify the speaker that best matches from a known group



Speaker Verification **Speaker Identification**

Background: Speaker Recognition

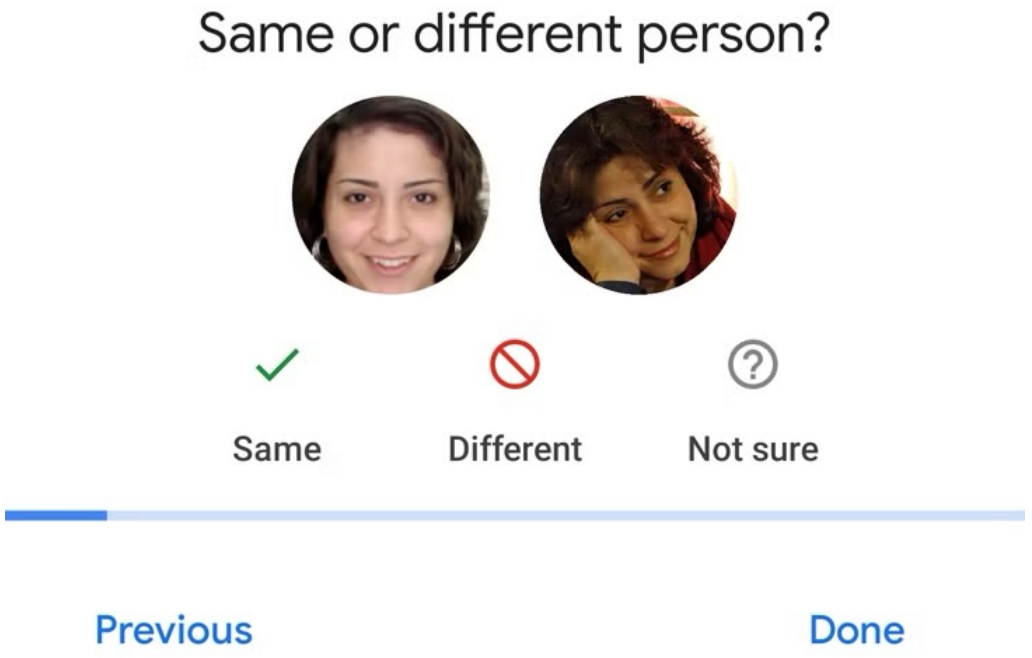
- Closed-Set Speaker Identification
  - Classify speech among enrolled speakers without rejection
- Open-Set Speaker Identification
  - Identify enrolled speakers or reject unknown speakers



Open-Set Speaker Identification

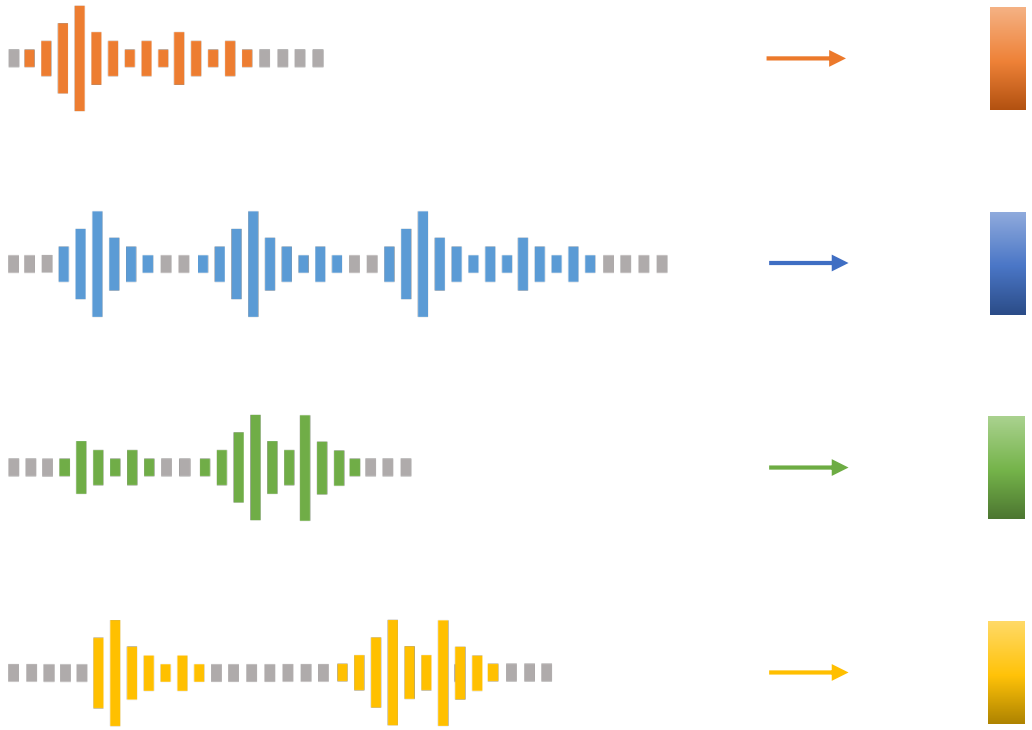
Background: Speaker Recognition

- Face Recognition
  - Similar to speaker recognition: verify/identify using biometric data
  - Similar methods show robustness across both tasks



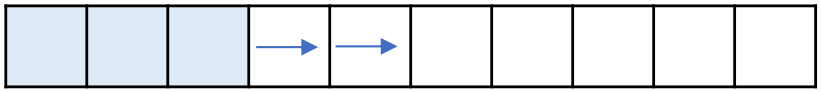
Background: Deep Speaker Embedding

- Speaker Embedding
  - Extract fixed-size vector from variable-length speech signal
  - Used for speaker similarity computation (e.g., cosine similarity)



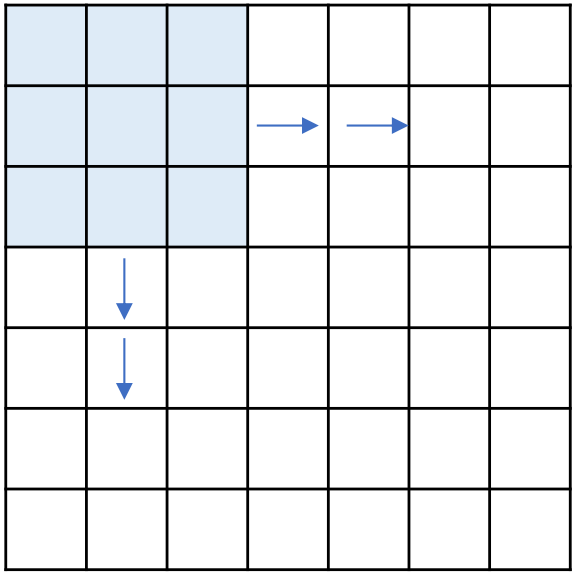
Background: Deep Speaker Embedding

- Speech Signals as Neural Network Input
  - 1D CNN over time axis



[T]

1D CNN

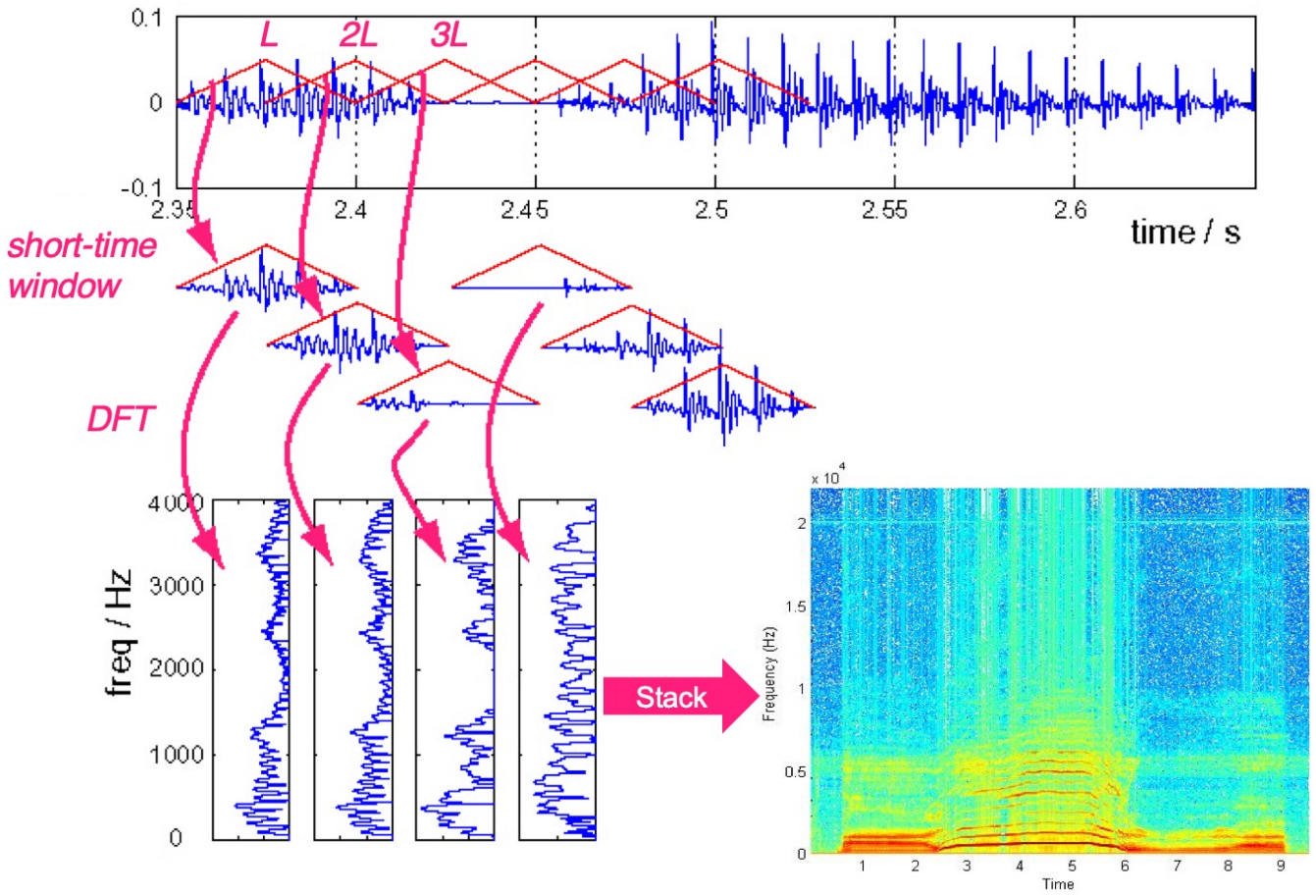


[H, W]

2D CNN

Background: Deep Speaker Embedding

- Speech Signals as Neural Network Input
  - Spectrogram (short-time Fourier transform)



Background: Deep Speaker Embedding

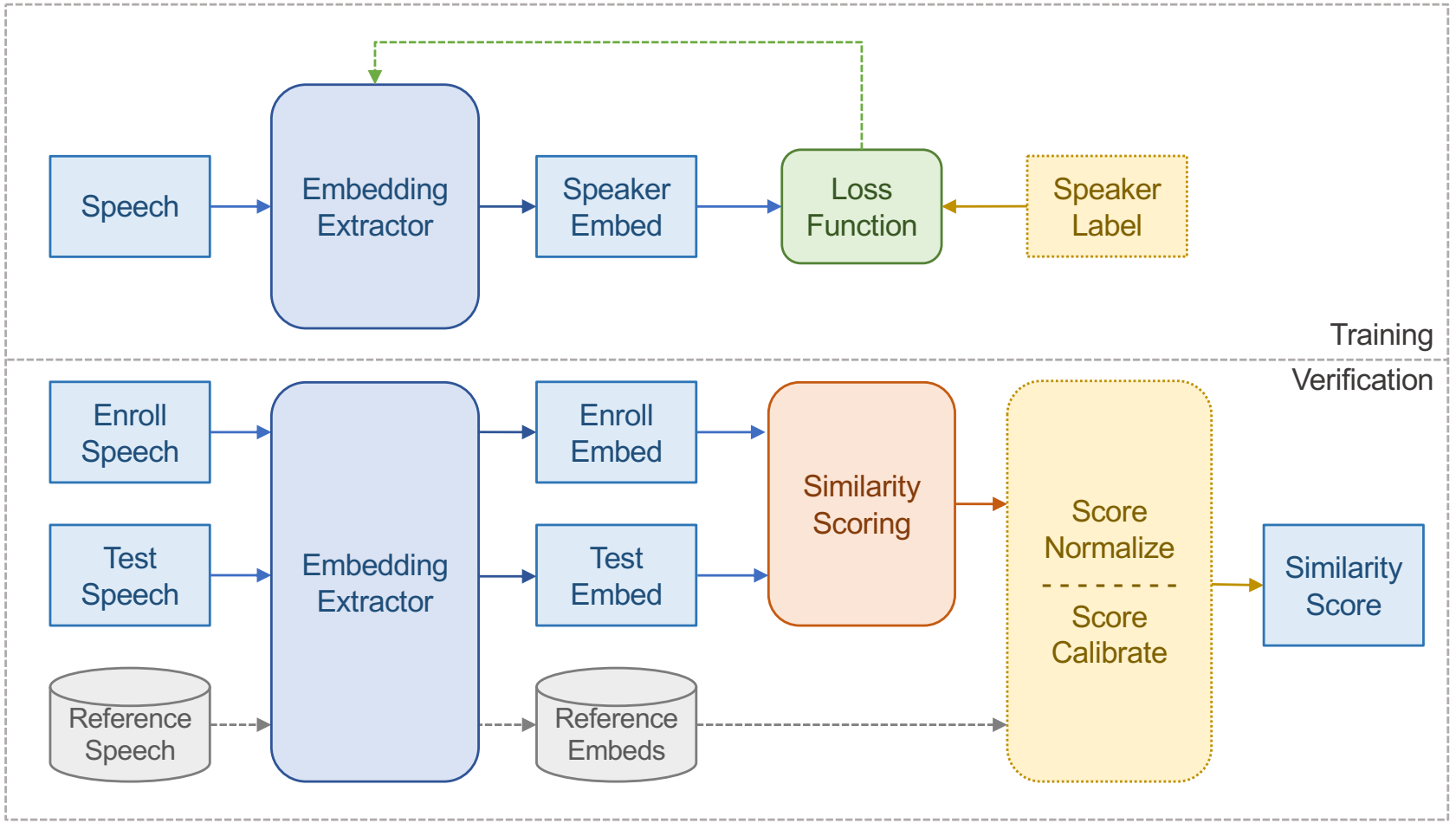
- Network Architectures for Speaker Embedding
  - More complex and better-performing models are being proposed
  - Select an appropriate model for efficient training and evaluation

Architecture	Params	FLOPs	Vox1-H (EER, %)
x-vector [1]	4.6M	0.5G	3.12
ECAPA-TDNN ($C = 1024$) [3]	14.7M	2.7G	1.87
ResNet34 based [5]	6.6M	4.6G	1.73
ResNet293 based [5]	28.6M	28.1G	1.28
ECAPA2 [7]	27.1M	187.0G	1.15
ReDimNet-B6 [8]	15.0M	20.3G	1.00

[1] D. Snyder *et al.*, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5329–5333.
[3] B. Desplanques *et al.*, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
[5] H. Wang *et al.*, "Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.
[7] J. Thienpondt *et al.*, "ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings," in *Proc. ASRU*, 2023, pp. 1–8.
[8] I. Yakovlev *et al.*, "Reshape Dimensions Network for Speaker Recognition," in *Proc. Interspeech*, 2024, pp. 3235-3239.

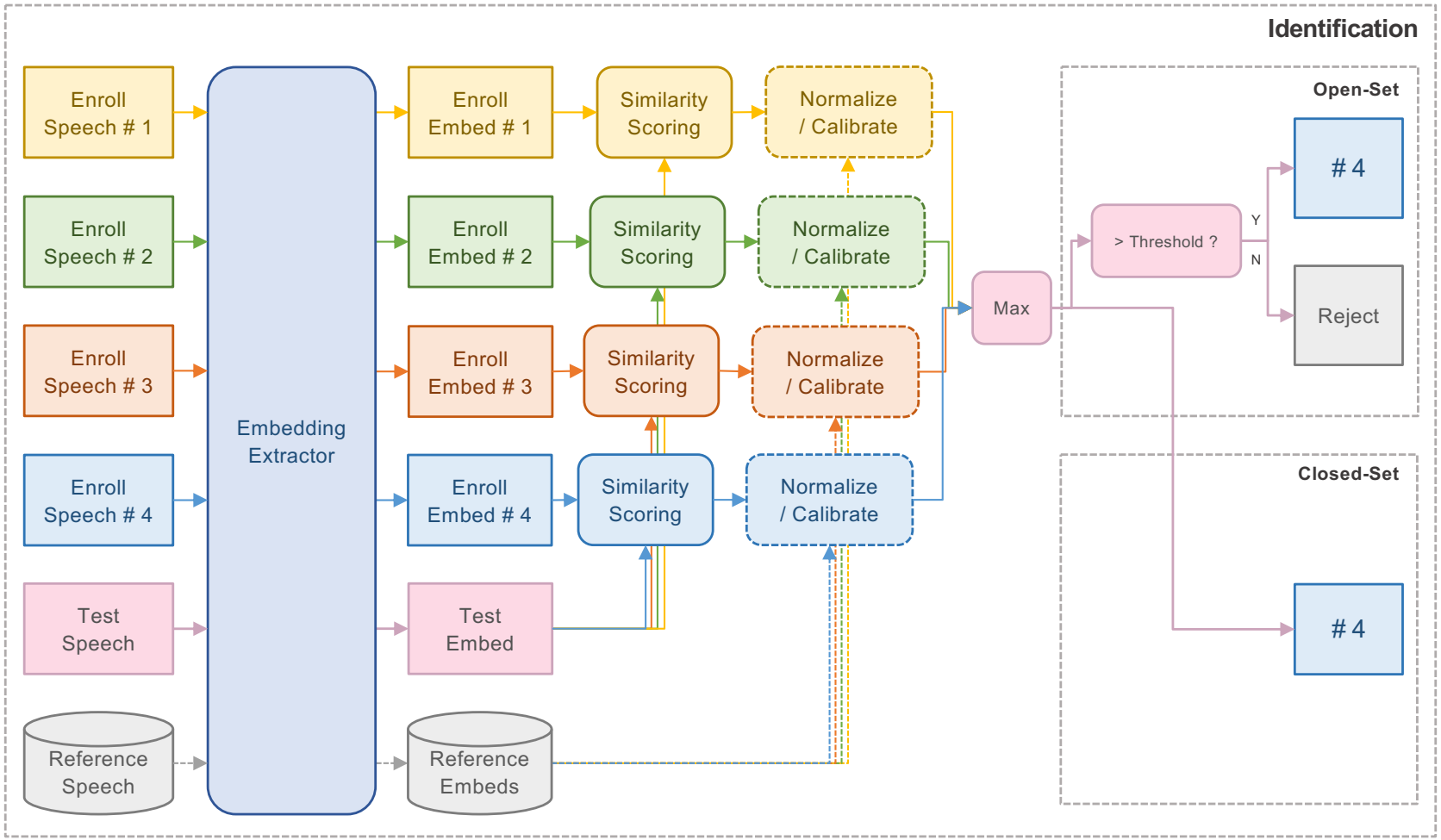
Background: Deep Speaker Embedding

- Deep Speaker Embedding Framework
  - Train embedding extraction and utilize it for speaker verification



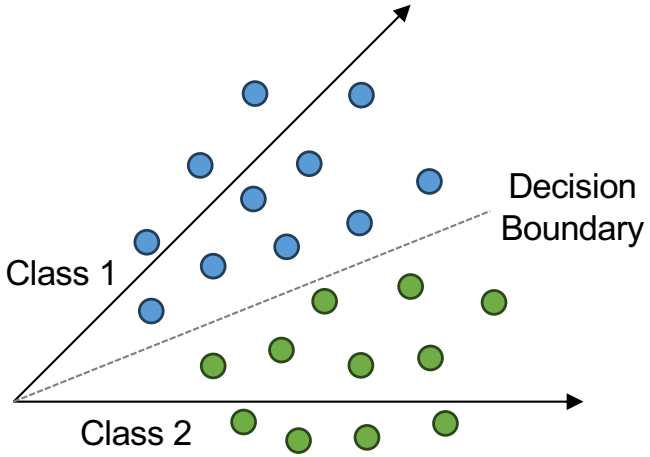
Background: Deep Speaker Embedding

- Deep Speaker Embedding Framework
  - Utilize embeddings for speaker identification

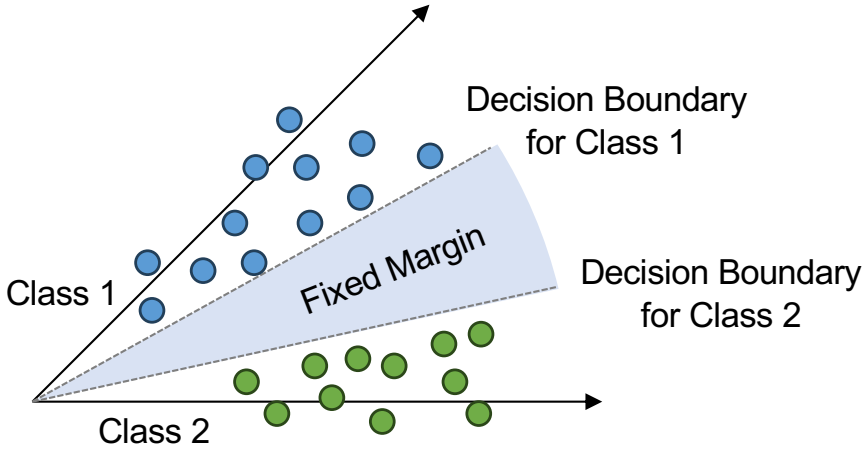


Background: Margin-Based Softmax Loss

- Margin-Based Softmax Loss
  - Apply a fixed margin to the original Softmax
  - Improve inter-class separability and intra-class compactness



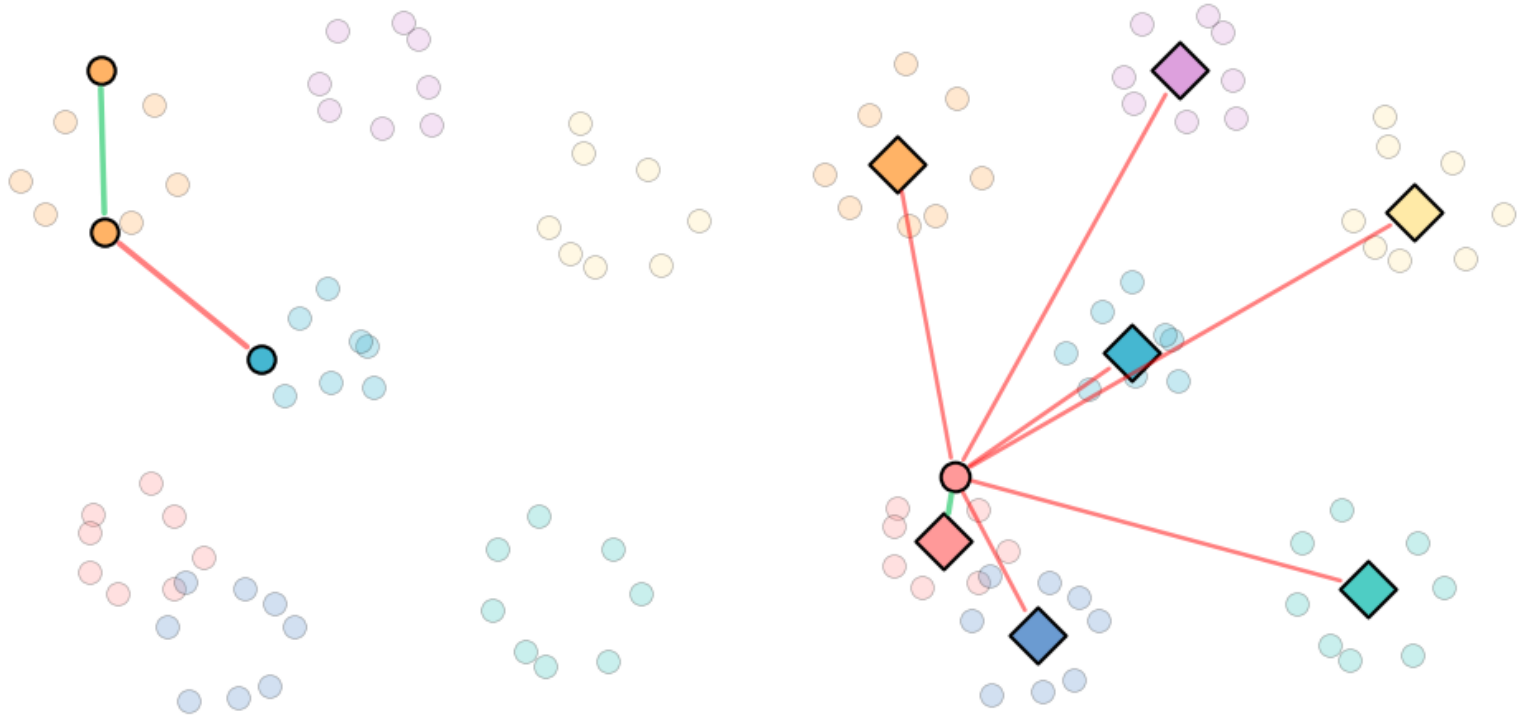
Original Softmax



Margin-based Softmax

Background: Margin-Based Softmax Loss

- Metric Learning vs. Margin-based Softmax
  - Metric learning: sees only mini-batch relationships
  - Margin-based: global weight matrix updates with all classes



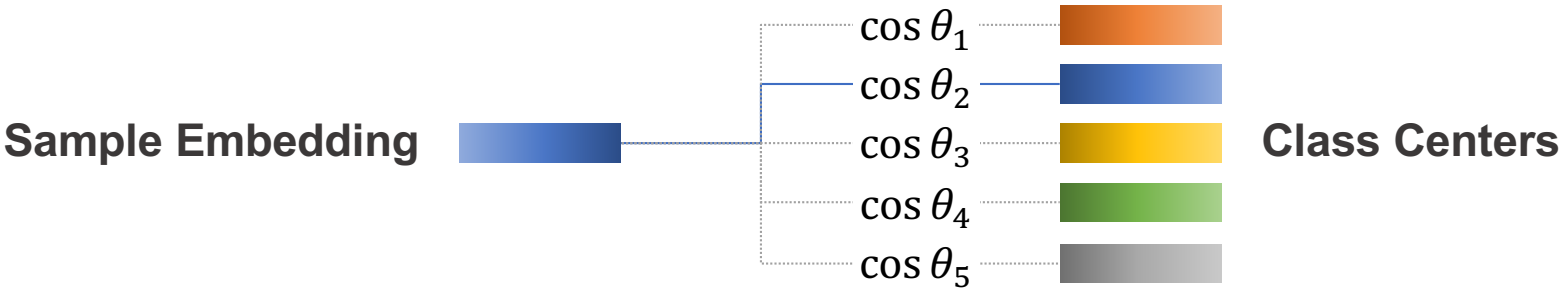
Metric Learning (Triplet)

Margin-based Softmax

Background: Margin-Based Softmax Loss

- Modified Softmax for Speaker Embedding
  - Remove bias term
  - Replace logit with cosine similarity
  - Apply scaling factor (s) → prevent vanishing gradient

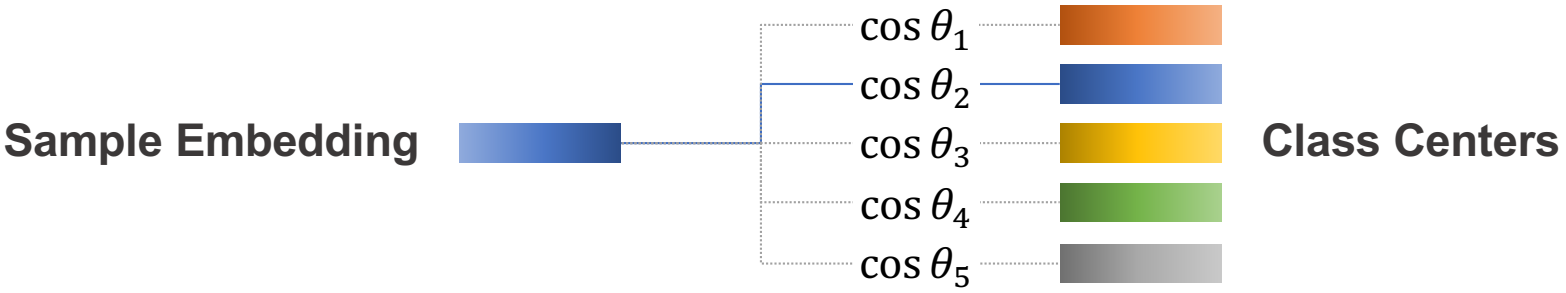
$$\ell_{\text{modified}} = -\log \frac{e^{s\cos\theta_{y_i}}}{\sum_{j=1}^{c} e^{s\cos\theta_j}}$$



Background: Margin-Based Softmax Loss

- Additive Angular Margin Softmax (AAM-Softmax)
  - Commonly used in speaker embedding training
  - Apply additive angular margin (m) to the angle θ_{y_i}
  - Hyperparameters: margin m, scaling factor s

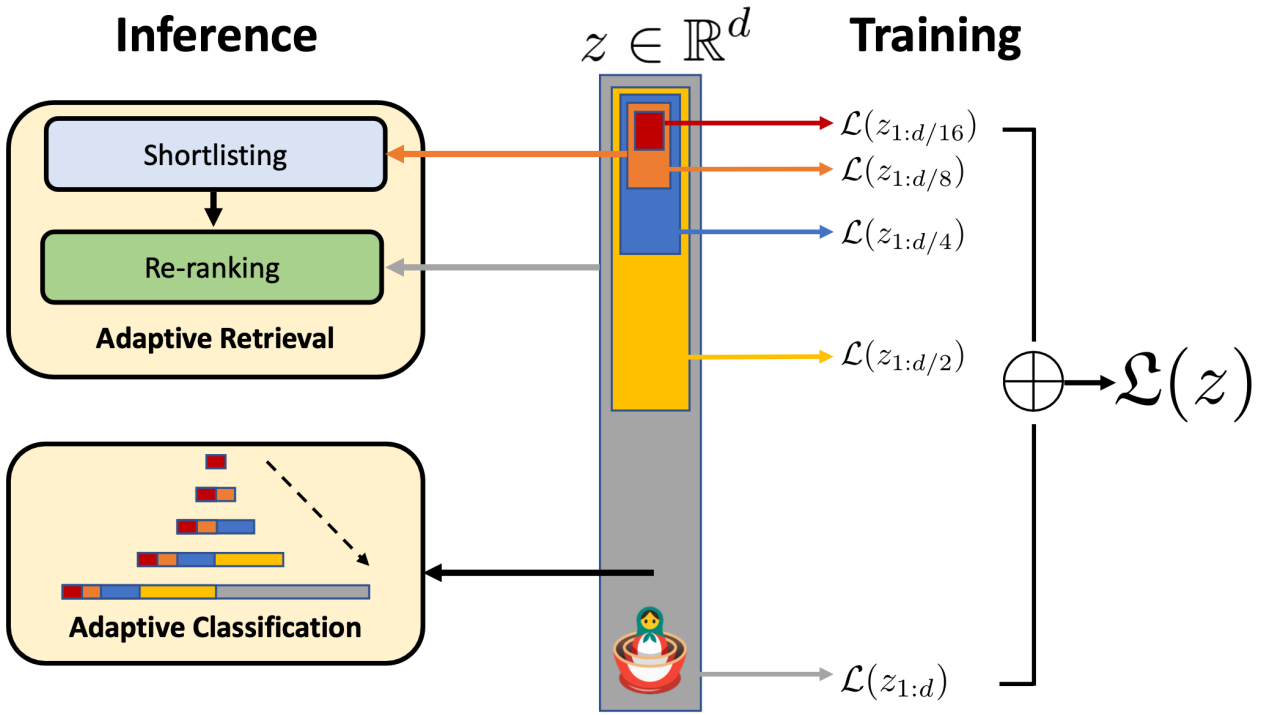
$$\ell_{\text{AAM}} = -\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{c} e^{s\cos\theta_j}}$$



[18] J. Deng *et al.*, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4685–4694.

Background: Matryoshka Representation Learning

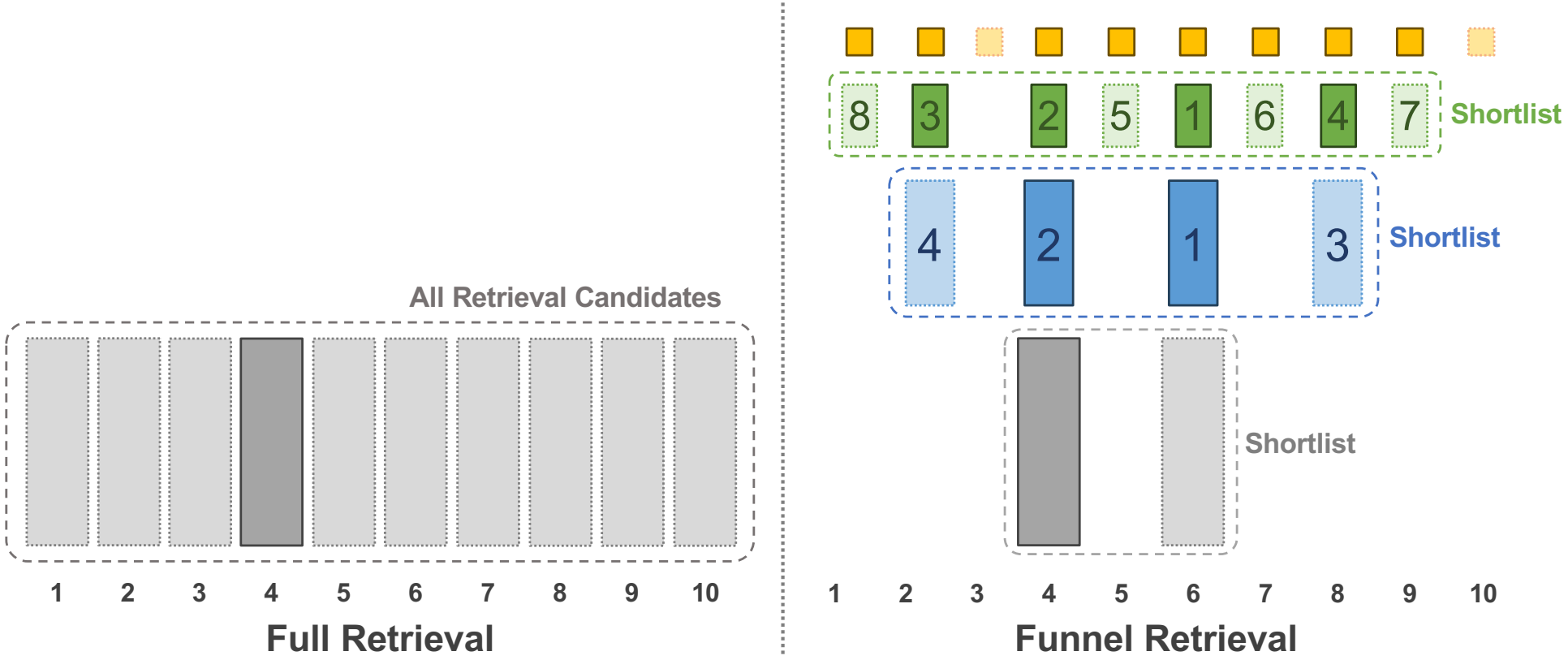
- Matryoshka Representation Learning (MRL)
  - Simultaneously train nested lower-dimensional vectors
  - Enable flexible adaptation to task-specific requirements
  - Improved performance in lower-dimensional vectors



[38] A. Kusupati *et al.*, "Matryoshka Representation Learning," in *Proc. Adv. in Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 30233–30249.

Background: Matryoshka Representation Learning

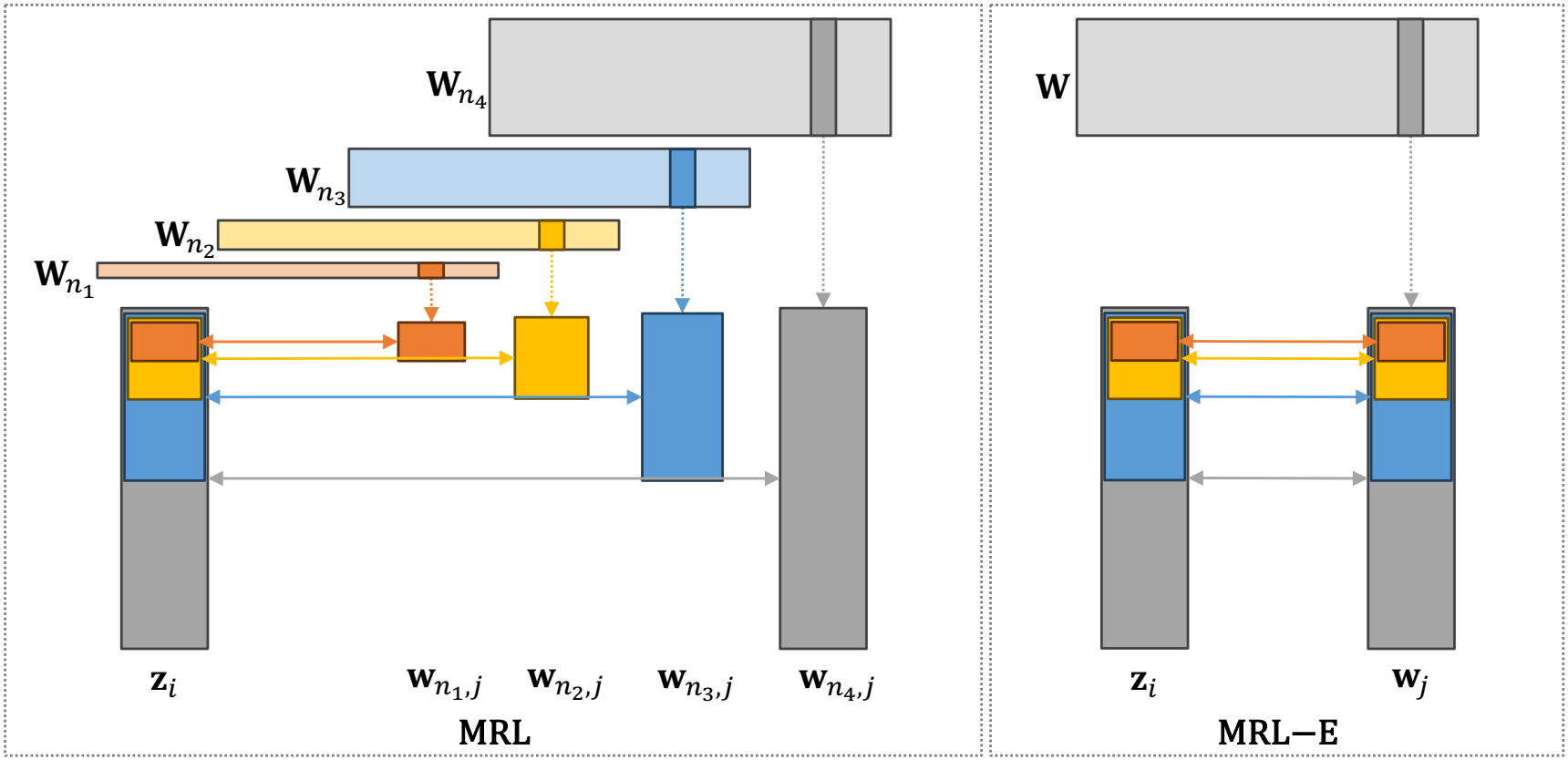
- Adaptive Retrieval
  - Reduce computational cost through shortlisting and re-ranking
  - Funnel Retrieval: iteratively halve shortlist and double vector size



[38] A. Kusupati *et al.*, "Matryoshka Representation Learning," in *Proc. Adv. in Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 30233–30249.

Background: Matryoshka Representation Learning

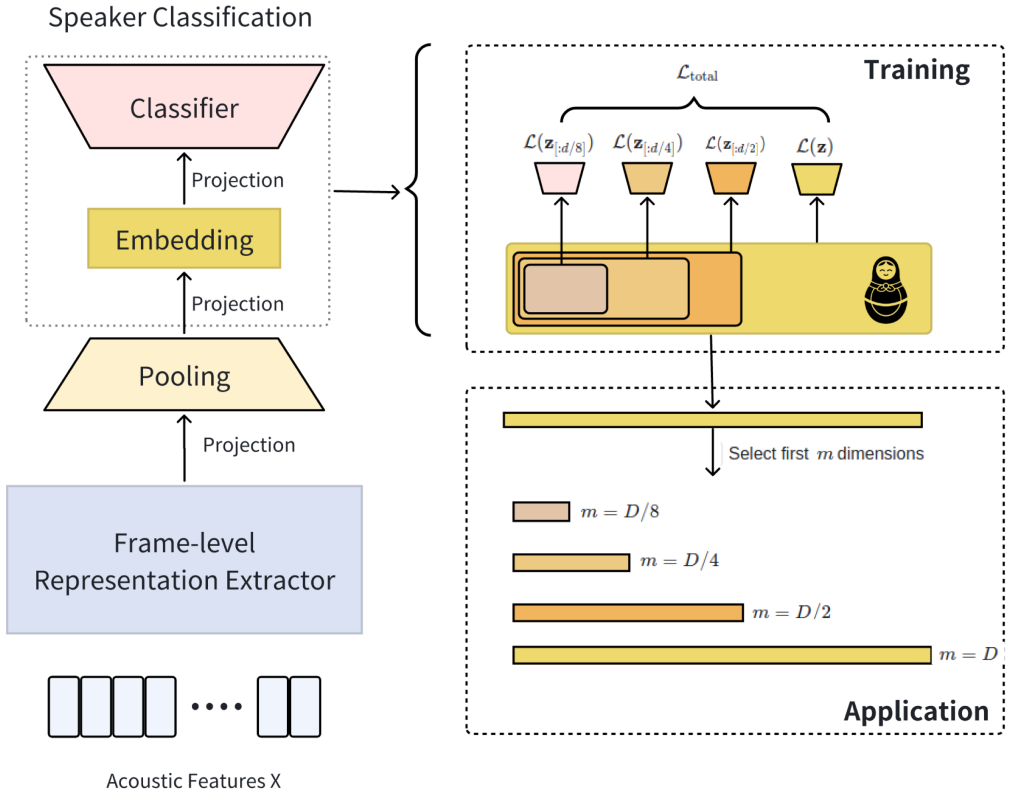
- Efficient Matryoshka Representation Learning (MRL-E)
  - Use weight-tying to reduce memory consumption
  - Improve memory efficiency and enable fast convergence



[38] A. Kusupati *et al.*, "Matryoshka Representation Learning," in *Proc. Adv. in Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 30233–30249.

Background: Matryoshka Representation Learning

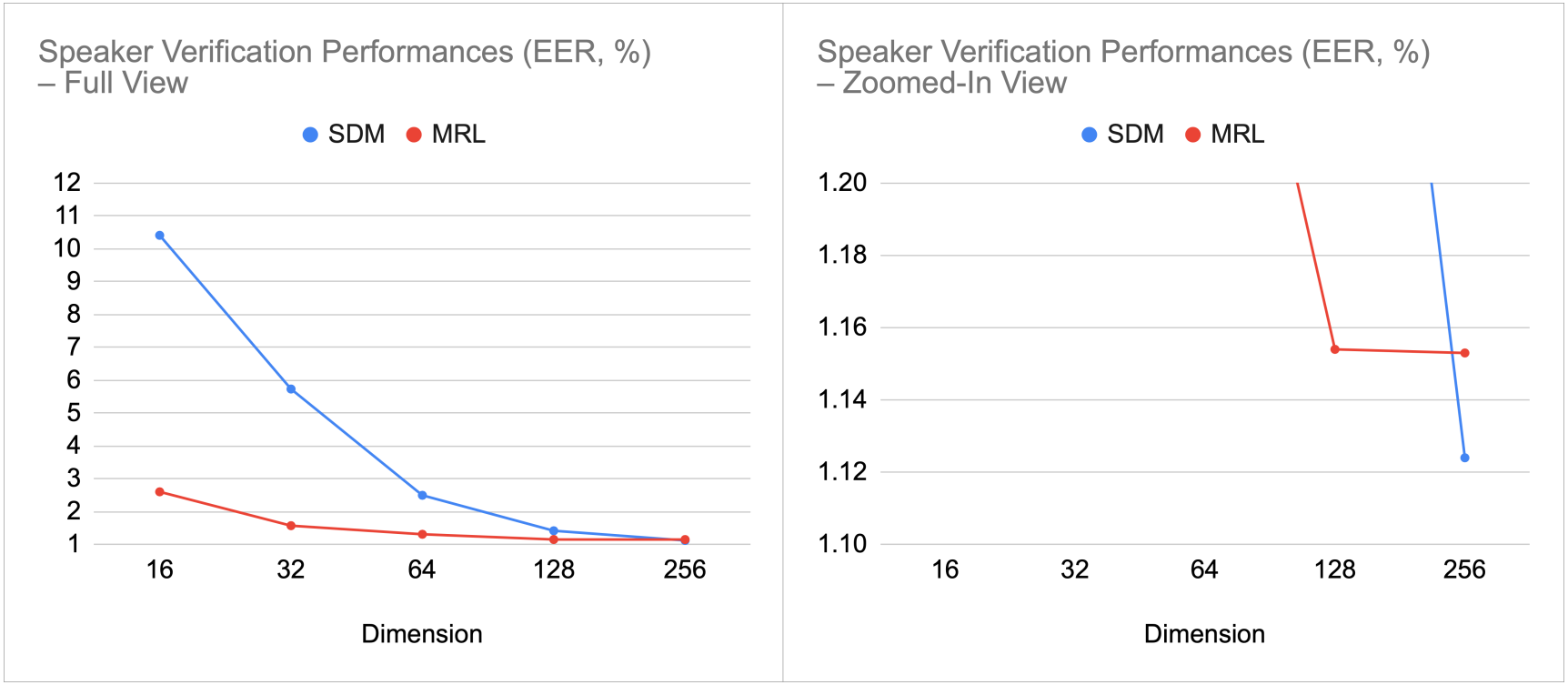
- Matryoshka Speaker Embedding
  - Apply MRL to train speaker embeddings
  - Train each dimension of the embedding with AAM-Softmax



[39] S. Wang, *et al.*, "M-Vec: Matryoshka Speaker Embeddings with Flexible Dimensions," 2024, *arXiv:2409.15782*.

Background: Matryoshka Representation Learning

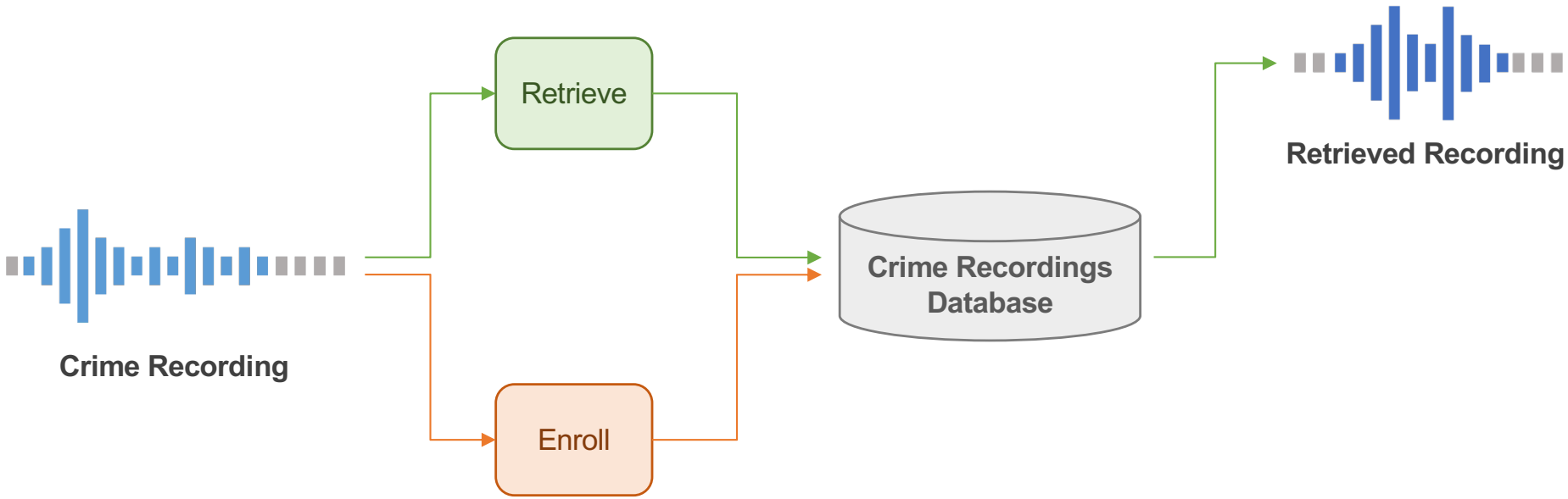
- Evaluation of Matryoshka Speaker Embedding
  - Compare MRL model with Single-Dimensional Models (SDM)
  - MRL improves performance in low dim. but degrades it in high dim.
  - Limits the model's maximum performance



[39] S. Wang, *et al.*, "M-Vec: Matryoshka Speaker Embeddings with Flexible Dimensions," 2024, *arXiv:2409.15782*.

Background: Matryoshka Representation Learning

- Application: Speech-based Criminal Investigation
  - Find the same speaker from past crime recordings
  - Adaptive retrieval improves efficiency as more recordings enroll
  - Maximum performance drop in MRL can reduce retrieval accuracy

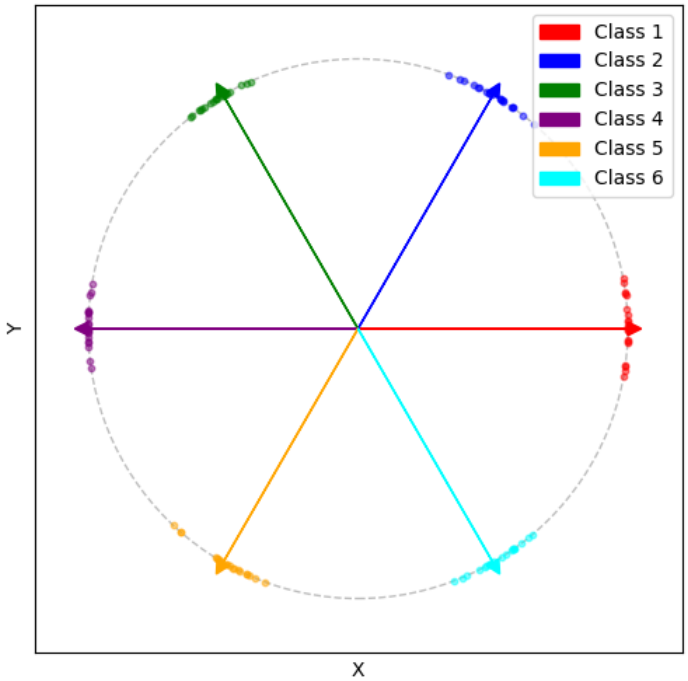


Contents

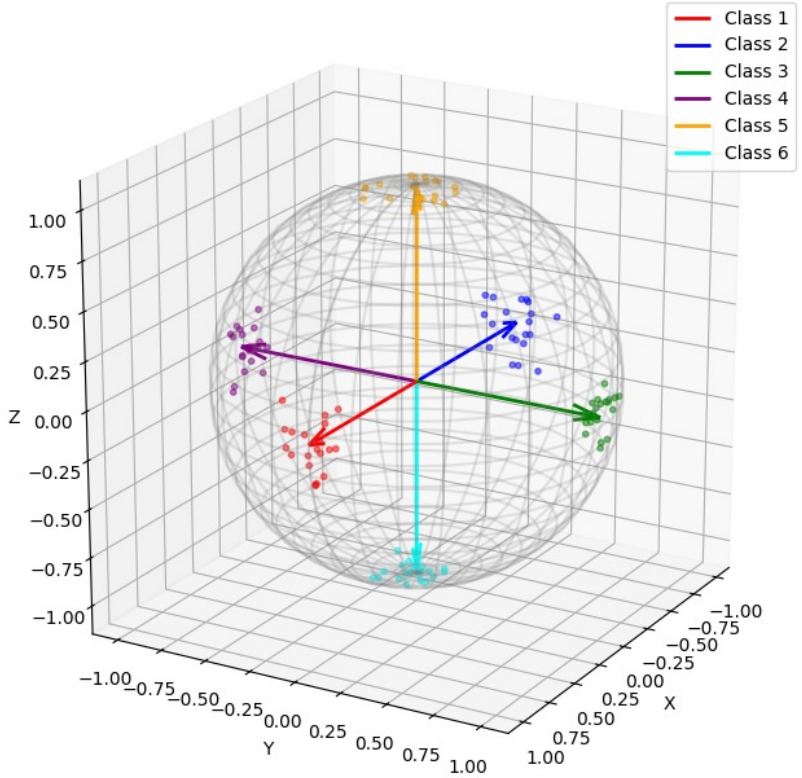
- Introduction
- Background
- Proposed Methods
  - Dimension-Specific Margins
  - Element-Wise Gradient Scaling
- Experiments
- Conclusions

Proposed Method 1: Dimension-Specific Margins

- Motivation
  - Low dimension → Smaller angles between class centers
  - High dimension → Larger angles between class centers



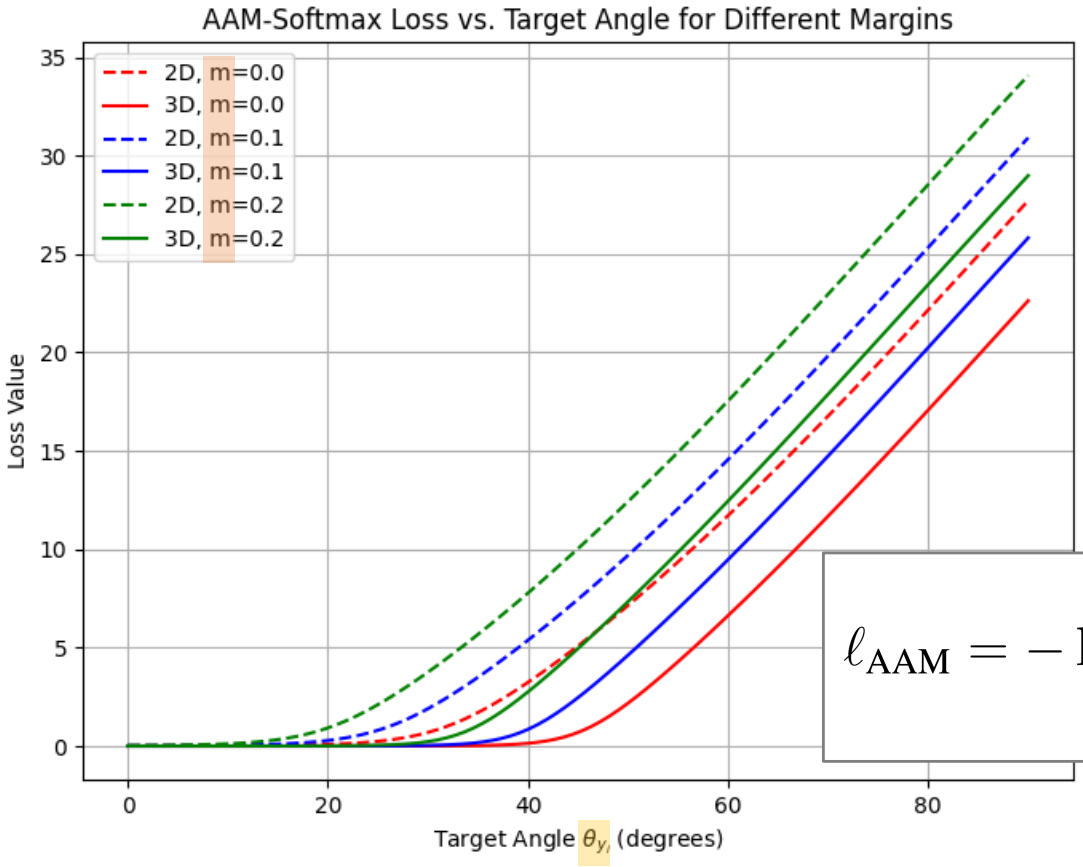
2-Dimensional Embeddings



3-Dimensional Embeddings

Proposed Method 1: Dimension-Specific Margins

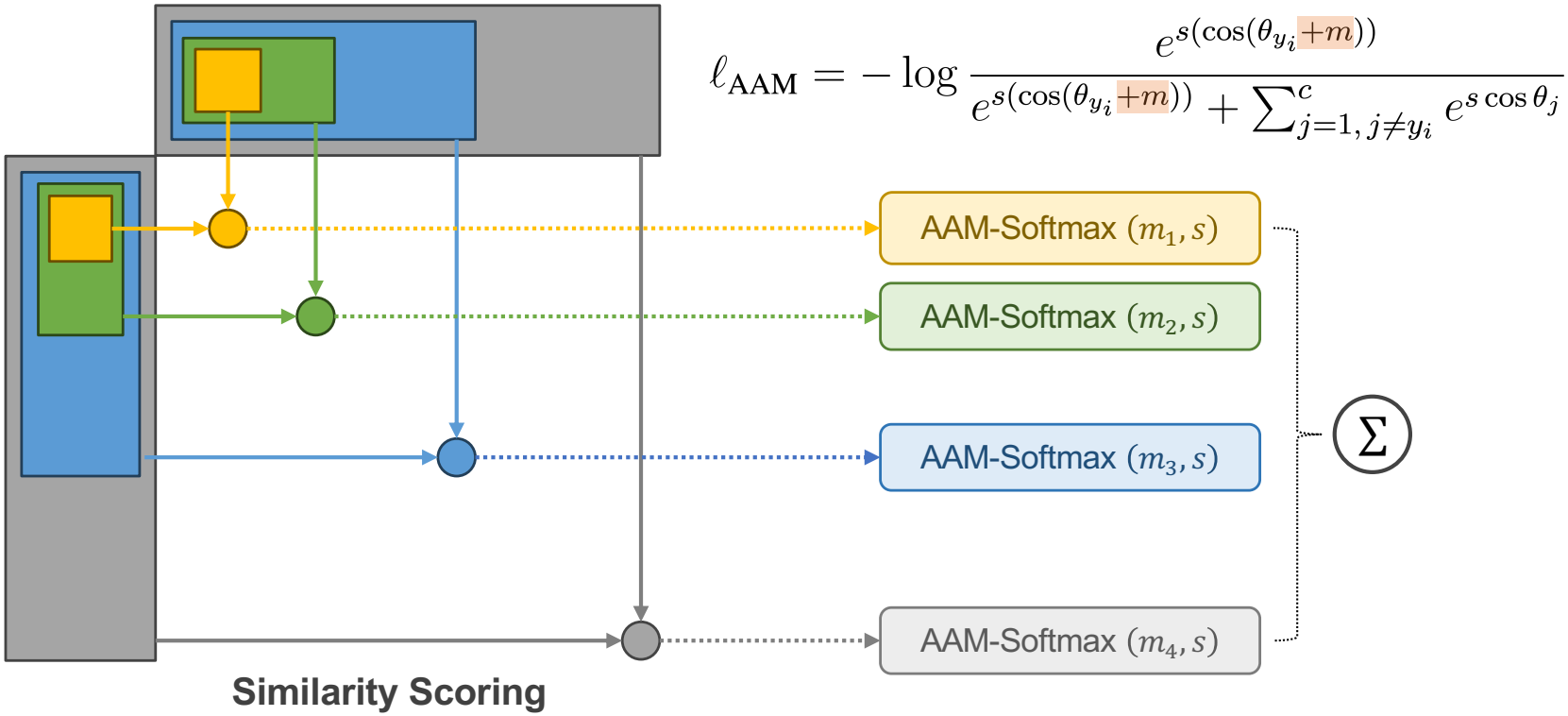
- Motivation
  - Larger margin leads to higher loss values
  - In low dimension, large margin can cause excessively high loss



$$\ell_{\text{AAM}} = -\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{c} e^{s\cos\theta_j}}$$

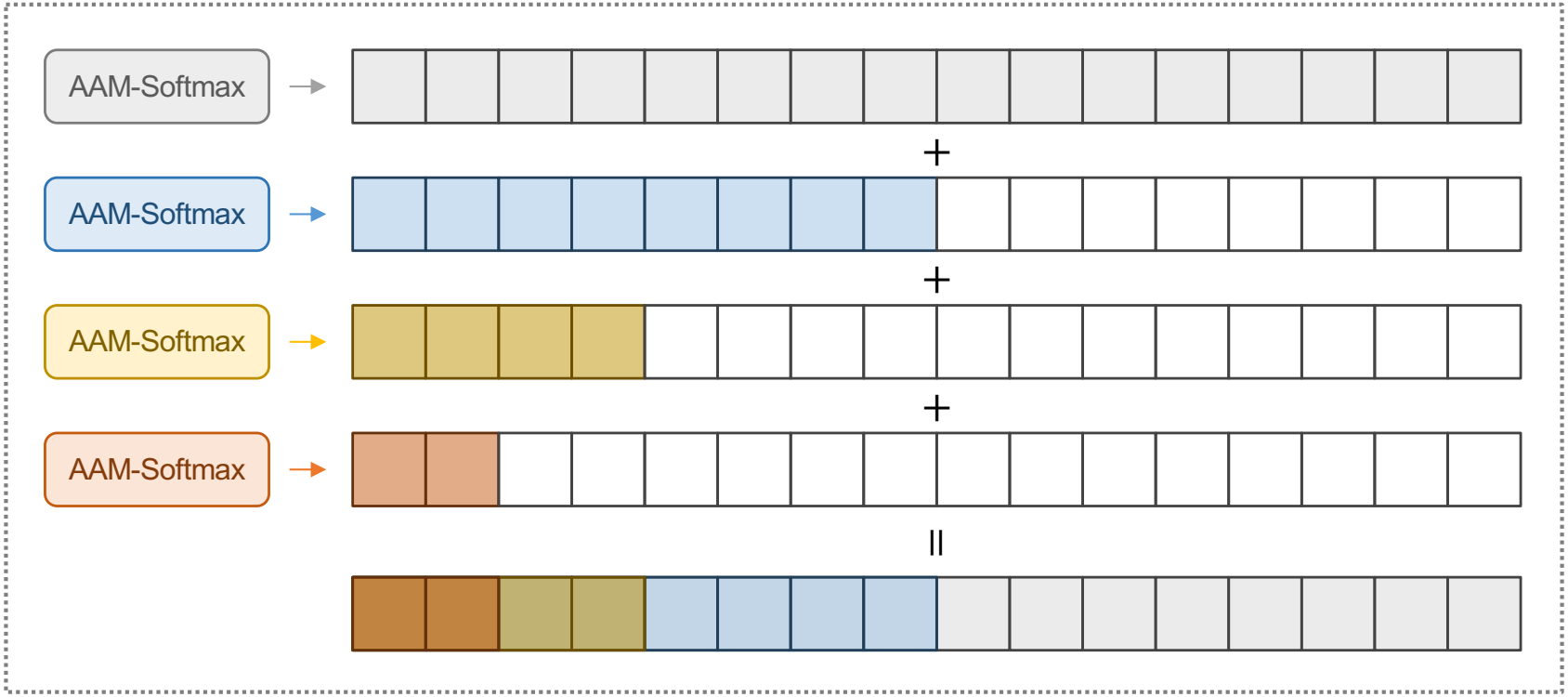
Proposed Method 1: Dimension-Specific Margins

- Dimension-Specific Margins (DSM)
  - Instead of fixed margin, adjust margins for each dimension
  - Stabilize training by preventing high loss in low dimensions
  - Use single-dimensional models to determine optimal margins



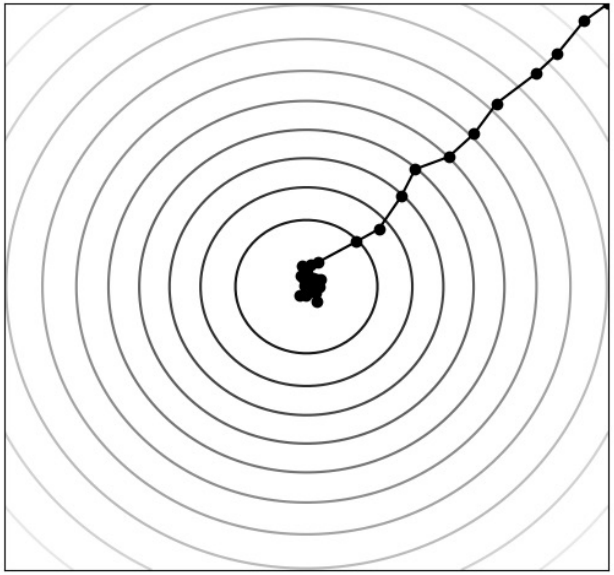
Proposed Method 2: Element-Wise Gradient Scaling

- Motivation
  - Gradient imbalance occurs across elements in MRL
  - Element-wise gradients decrease with increasing dimension
  - Each AAM-Softmax loss affects different elements

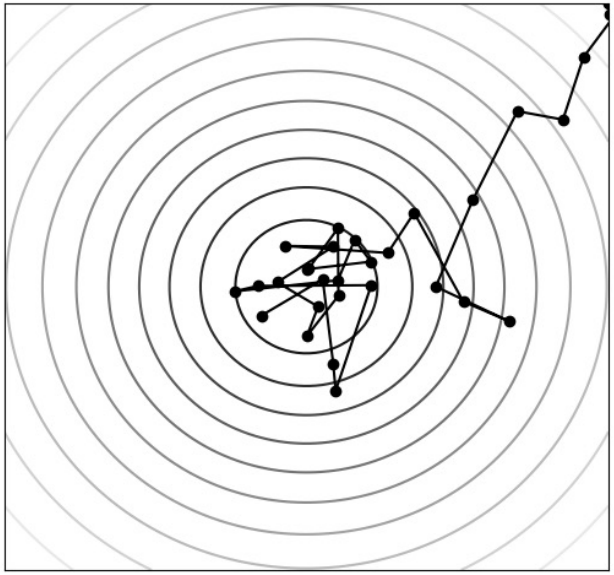


Proposed Method 2: Element-Wise Gradient Scaling

- Motivation
  - L2 normalization is applied to embeddings in AAM-Softmax
  - L2 normalization tends to slow down convergence
  - Accumulated gradients lead to larger embedding magnitudes



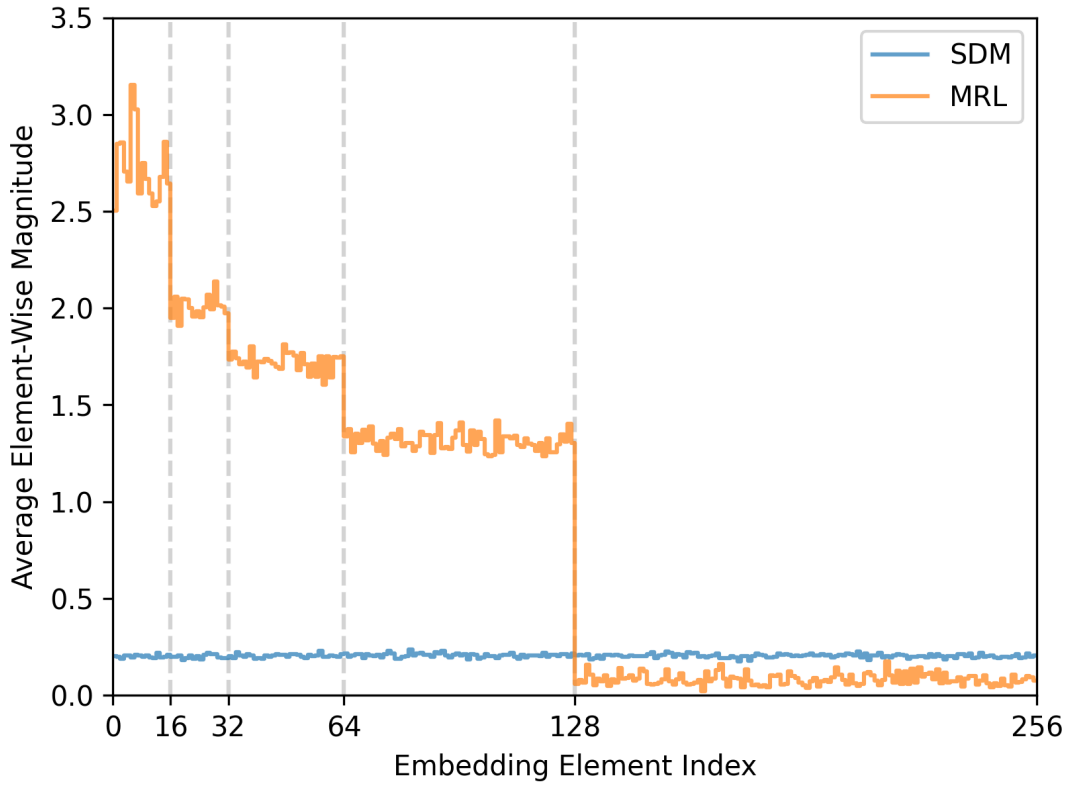
Gradient Descent without L2 Normalization



Gradient Descent with L2 Normalization

Proposed Method 2: Element-Wise Gradient Scaling

- Motivation
  - Overall magnitude shift does not affect performance
  - Magnitude imbalance can limit performance in specific dimensions
  - Approximate element-wise gradient to mitigate imbalance



Proposed Method 2: Element-Wise Gradient Scaling

- Approximation of Element-Wise Gradient
  - Assume cosine similarity is invariant across nested dimensions
  - Assume uniform contribution from all elements
  - Element-wise gradient in n dimension is approximately:

$$\frac{\partial \ell_{\text{AAM}}}{\partial \mathbf{z}_i} = \sum_{j=1}^{c} (p_j - \delta_{j,y_i})\, s \left(\frac{\mathbf{u}_j}{\|\mathbf{z}_i\|} - \cos\theta_j \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|^2} \right)$$

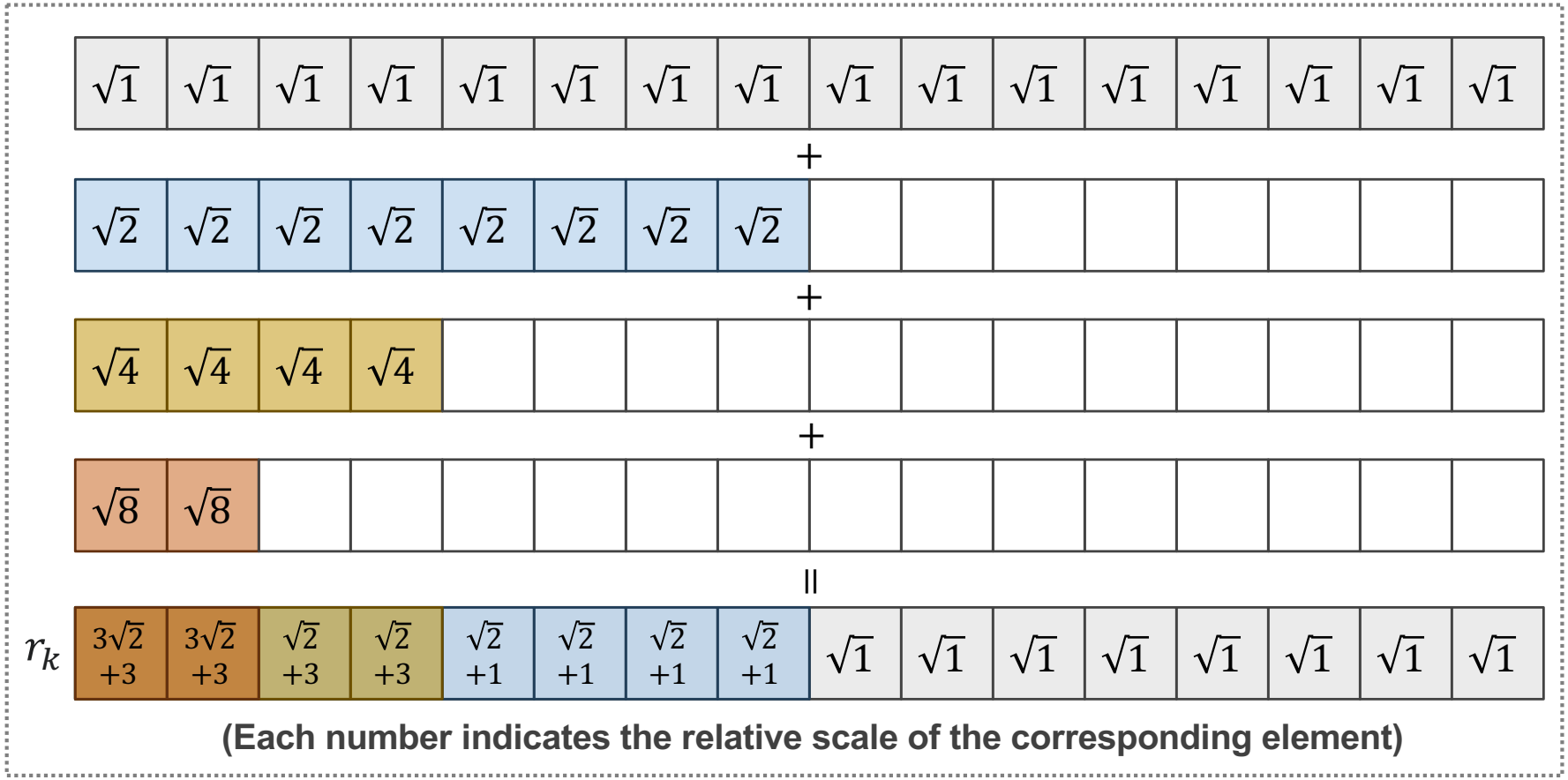
$$p_j = \frac{e^{l_j}}{\sum_{t=1}^{c} e^{l_t}} \qquad l_j = \begin{cases} s\cos(\theta_{y_i} + m), & j = y_i, \\ s\cos\theta_j, & \text{otherwise} \end{cases} \qquad \delta_{j,y_i} = \begin{cases} 1, & j = y_i, \\ 0, & \text{otherwise} \end{cases} \qquad \mathbf{u}_j = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$$

$$\|\mathbf{z}_i\| = \sqrt{\sum_{k=1}^{d} (z_{i,k})^2} \qquad \|\mathbf{z}_i^{(1:n)}\| = \sqrt{\sum_{k=1}^{n} (z_{i,k})^2} \qquad \|\mathbf{z}_i^{(1:n)}\| \approx \sqrt{\frac{n}{d}} \cdot \|\mathbf{z}_i\|$$

$$\frac{\partial \ell_{\text{AAM}}^{(1:n)}}{\partial \mathbf{z}_i^{(1:n)}} \approx \sqrt{\frac{d}{n}} \cdot \frac{\partial \ell_{\text{AAM}}}{\partial \mathbf{z}_i}$$

Proposed Method 2: Element-Wise Gradient Scaling

- Element-Wise Gradient Scaling (EGS)
  - Scaling factor r_k is defined by accumulated relative gradient
  - Rescale each element's magnitude by $1/r_k$



(Each number indicates the relative scale of the corresponding element)

Contents

- Introduction
- Background
- Proposed Methods
- Experiments
  - Speaker Verification
  - Open-Set Speaker Identification
- Conclusions

Experiments: Speaker Verification

- Evaluation Protocol
  - Positive and negative evaluation pairs are equally sampled
  - Errors from positive and negative pairs vary with the threshold

Spk1-Utt1	Spk1-Utt2	0.6	O	
Spk2-Utt1	Spk2-Utt2	0.5	X	False Rejection
Spk3-Utt1	Spk3-Utt2	0.4	X	False Rejection
Spk4-Utt1	Spk4-Utt2	0.3	X	False Rejection

Positive Pairs

Spk1-Utt1	Spk4-Utt2	0.6	O	False Acceptance
Spk2-Utt1	Spk3-Utt2	0.5	X	
Spk3-Utt1	Spk2-Utt2	0.4	X	
Spk4-Utt1	Spk1-Utt2	0.3	X	

Negative Pairs

$\tau = 0.55$

Spk1-Utt1	Spk1-Utt2	0.6	O	
Spk2-Utt1	Spk2-Utt2	0.5	O	
Spk3-Utt1	Spk3-Utt2	0.4	X	False Rejection
Spk4-Utt1	Spk4-Utt2	0.3	X	False Rejection

Positive Pairs

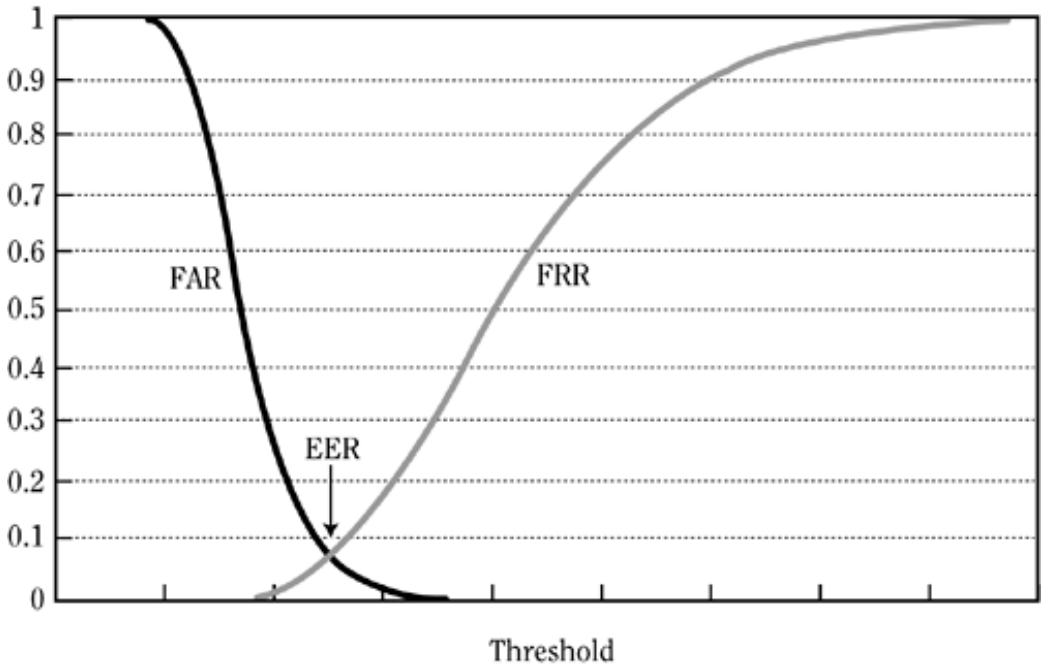
Spk1-Utt1	Spk4-Utt2	0.6	O	False Acceptance
Spk2-Utt1	Spk3-Utt2	0.5	O	False Acceptance
Spk3-Utt1	Spk2-Utt2	0.4	X	
Spk4-Utt1	Spk1-Utt2	0.3	X	

Negative Pairs

$\tau = 0.45$

Experiments: Speaker Verification

- Evaluation Metric: Equal Error Rate (EER)
  - False Acceptance Rate (FAR)
  - False Rejection Rate (FRR)
  - Error at the threshold where FAR and FRR are equal
  - Lower is better



Experiments: Speaker Verification

- Dataset: VoxCeleb
  - Recordings extracted from interview videos on YouTube
  - Training: VoxCeleb2-dev (Network Architecture: ResNet34 based)
  - Evaluation: VoxCeleb1-O/E/H
  - Focus on VoxCeleb1-E/H as they provide more reliable results

	VoxCeleb2-dev	VoxCeleb1-O	VoxCeleb1-E	VoxCeleb1-H
Num. Speakers	5,994	40	1,251	1,190
Num. Utterances	1,092,009	4,708	145,160	137,924
Num. Evaluation Pairs	-	37,611	579,818	550,894

[54] J. S. Chung *et al.*, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

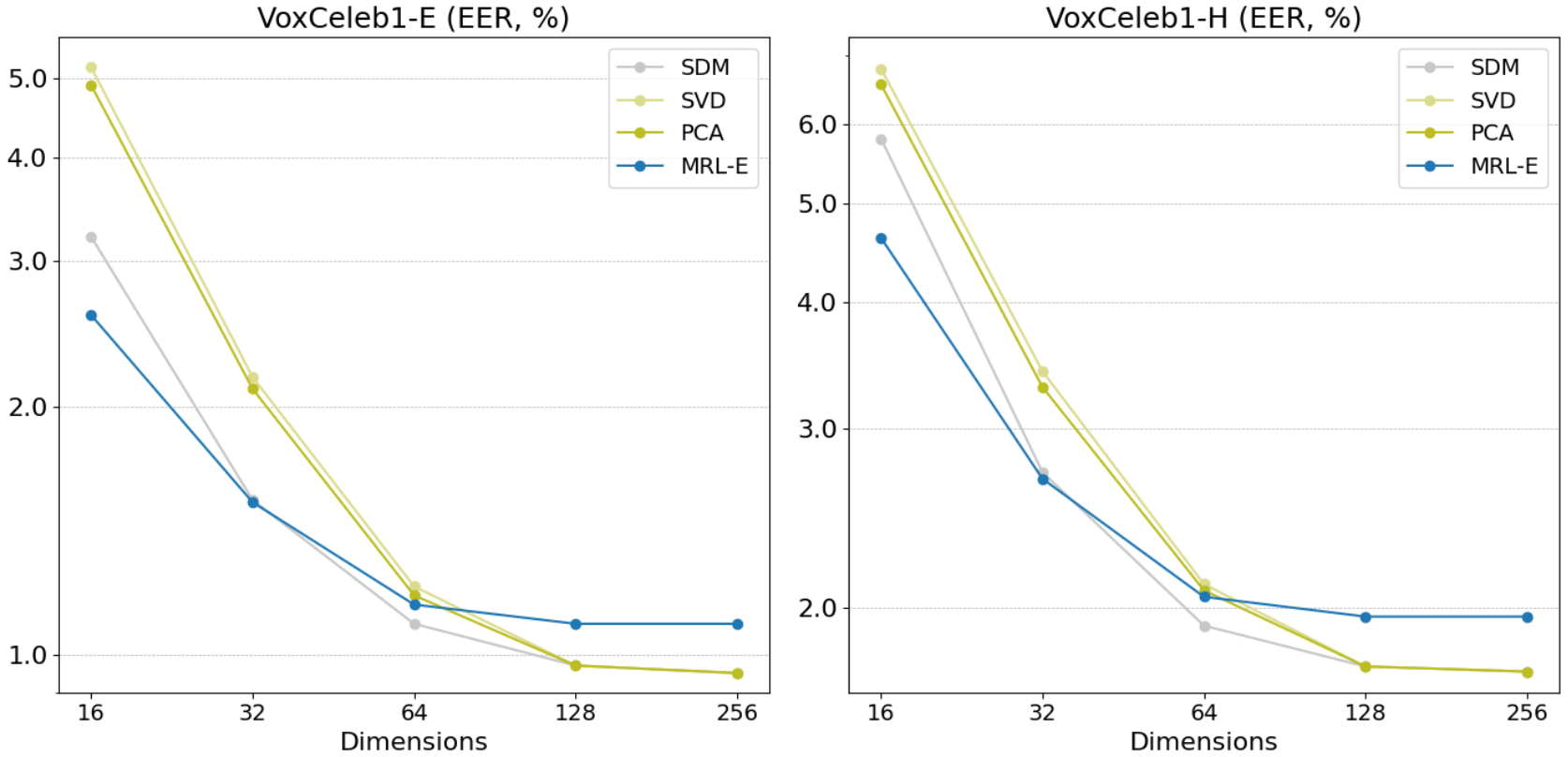
Experiments: Speaker Verification

- Results: DSM and EGS (EER)
  - Evaluate impact of DSM and EGS, individually or together
  - Compare with Single-Dimensional Models (SDM)
  - Compare with SVD/PCA-based dimension reduction methods

	Method		VoxCeleb1-O					VoxCeleb1-E					VoxCeleb1-H				
	DSM	EGS	16	32	64	128	256	16	32	64	128	256	16	32	64	128	256
SDM	✗	✗	3.18	1.53	0.94	0.85	0.80	3.21	1.54	1.09	0.97	0.95	5.80	2.72	1.92	1.75	1.73
	✓	✗	2.53	1.45	0.94	0.85	0.80	2.81	1.54	1.09	0.97	0.95	5.11	2.72	1.92	1.75	1.73
SVD	✗	✗	4.35	1.95	1.04	0.81	0.80	5.15	2.17	1.21	0.97	0.95	6.80	3.42	2.11	1.75	1.73
PCA	✗	✗	4.08	1.88	1.05	0.80	0.80	4.90	2.10	1.18	0.97	0.95	6.56	3.30	2.08	1.75	1.73
MRL	✗	✗	2.52	1.47	1.18	0.96	0.96	2.62	1.52	1.20	1.07	1.07	4.61	2.69	2.12	1.91	1.91
	✓	✗	2.45	1.32	0.95	0.86	0.85	2.65	1.50	1.14	1.04	1.04	4.68	2.68	2.00	1.84	1.84
	✗	✓	2.38	1.35	1.07	0.82	0.77	2.58	1.52	1.19	1.04	0.97	4.61	2.63	2.09	1.86	1.75
	✓	✓	2.55	1.37	1.04	0.83	0.84	2.63	1.52	1.15	1.03	0.97	4.59	2.68	2.04	1.84	1.75
MRL-E	✗	✗	2.40	1.39	0.99	0.94	0.94	2.58	1.53	1.15	1.09	1.09	4.63	2.68	2.05	1.96	1.96
	✓	✗	2.40	1.47	1.01	1.00	1.00	2.60	1.49	1.12	1.07	1.07	4.63	2.60	1.97	1.86	1.86
	✗	✓	2.56	1.26	0.89	0.78	0.76	2.57	1.45	1.13	1.00	0.96	4.57	2.53	1.97	1.79	1.72
	✓	✓	2.58	1.33	0.98	0.81	0.80	2.59	1.47	1.10	0.99	0.96	4.54	2.55	1.94	1.76	1.71

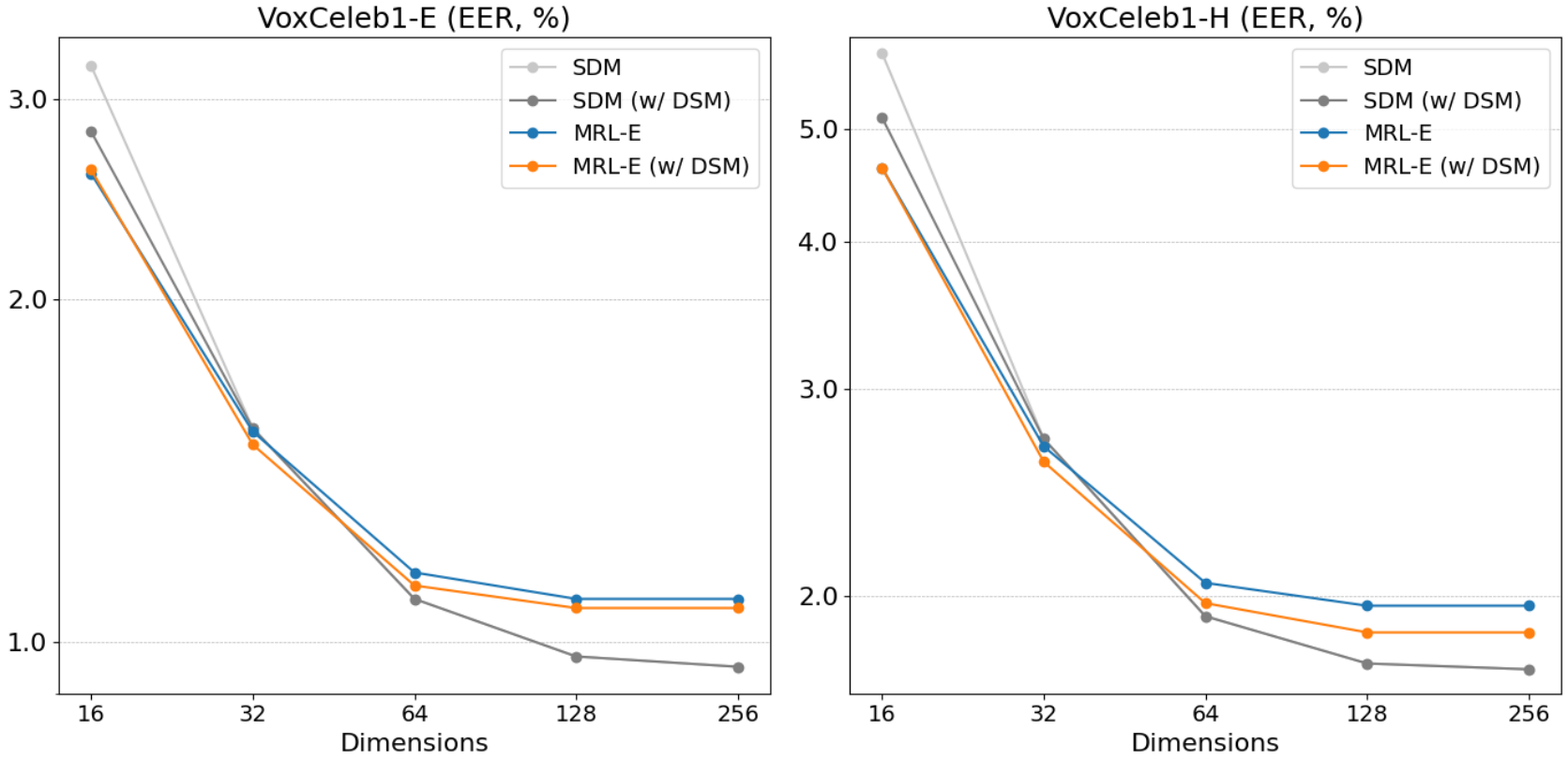
Experiments: Speaker Verification

- Results: DSM and EGS (EER)
  - MRL-E outperforms SVD, PCA and SDM in lower dimensions
  - MRL-E underperforms SVD, PCA and SDM in higher dimensions



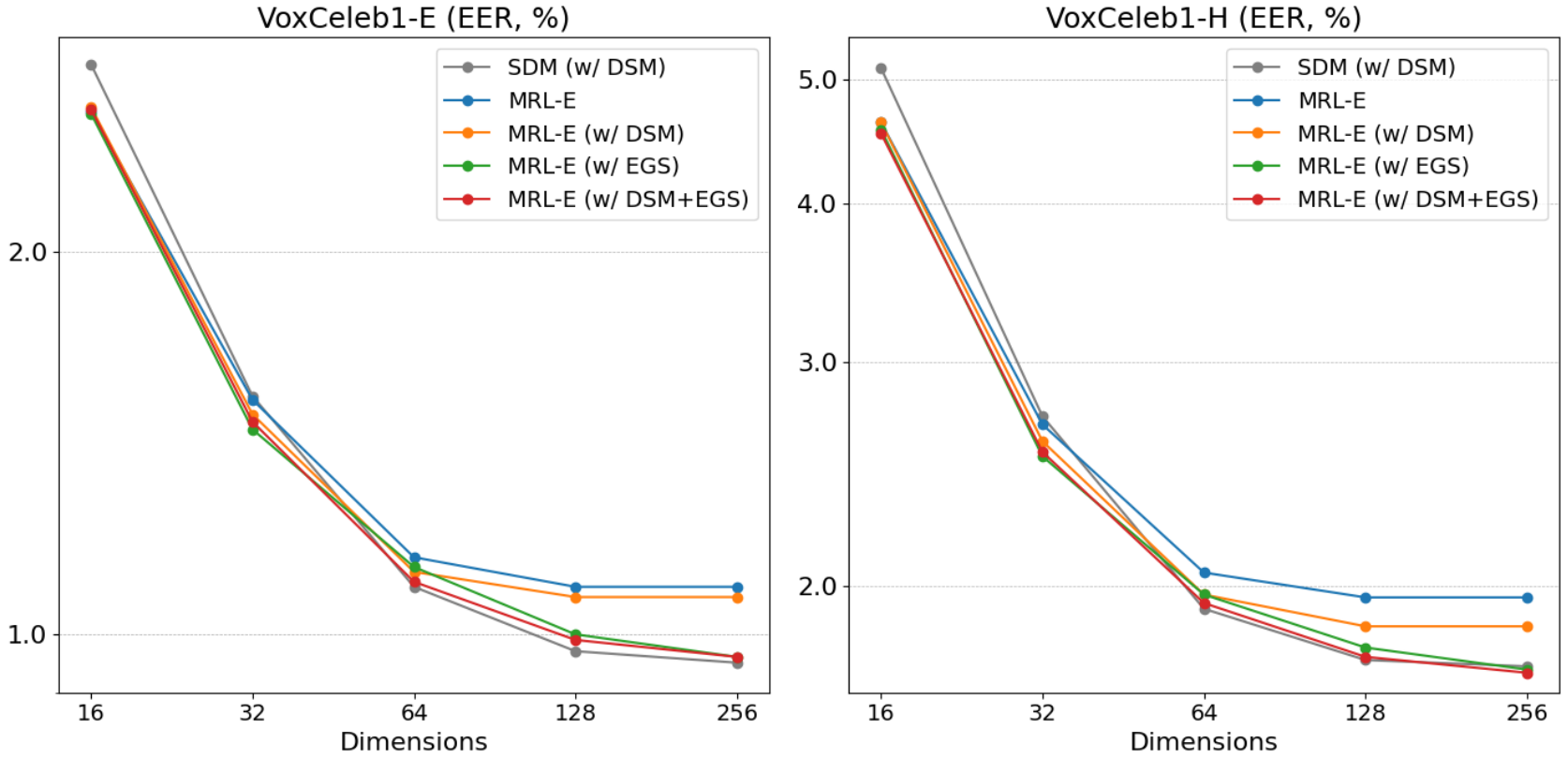
Experiments: Speaker Verification

- Results: DSM and EGS (EER)
  - DSM improves performance of SDM in lower dimensions
  - DSM improves performance of MRL-E in higher dimensions



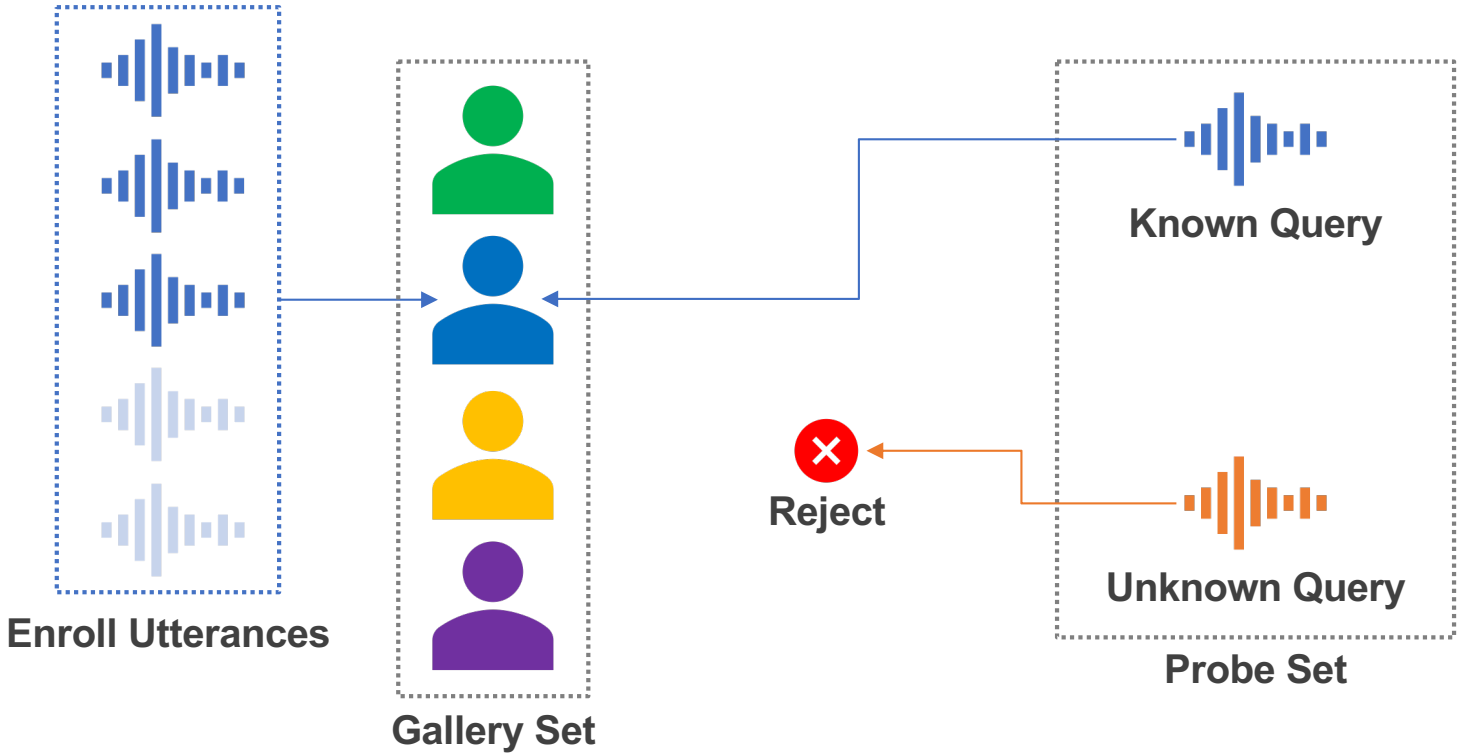
Experiments: Speaker Verification

- Results: DSM and EGS (EER)
  - EGS improves performance of MRL-E in higher dimensions
  - DSM+EGS reduces error by 12.8% in MRL-E (256-D, VoxCeleb1-H)



Experiments: Open-Set Speaker Identification

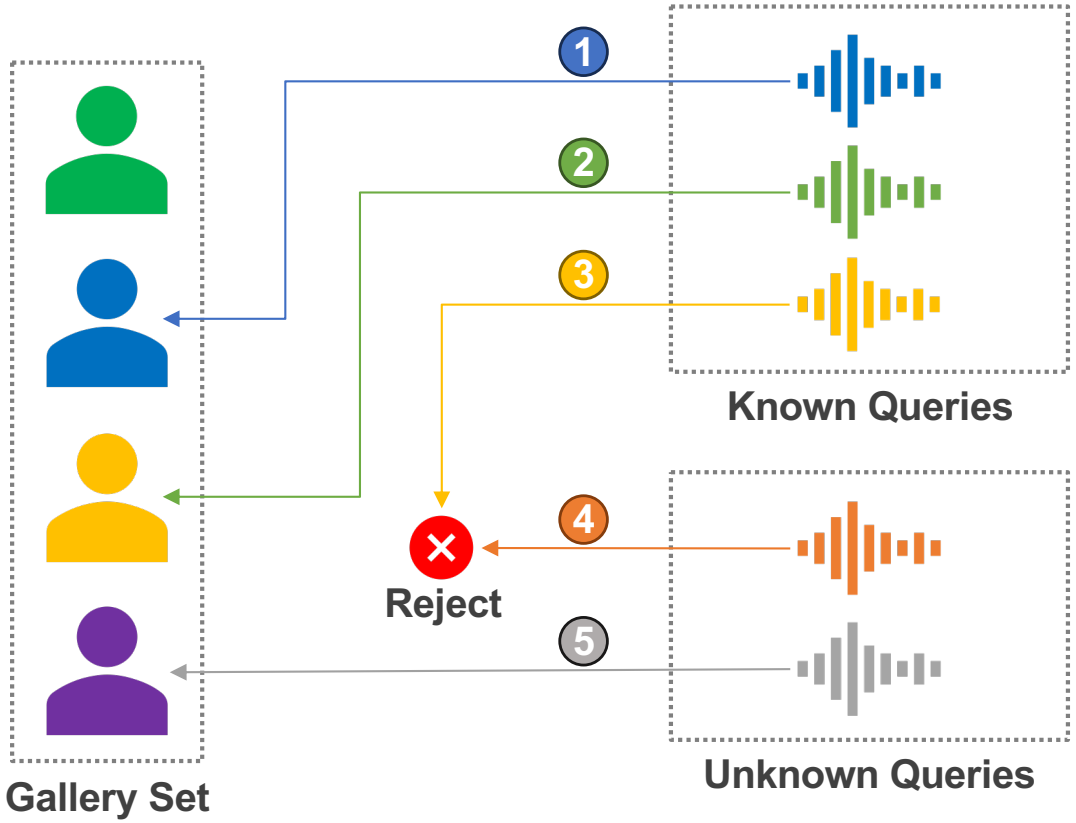
- Evaluation Protocol
  - Gallery Set: Enroll utterances for known speakers
  - Probe Set: Test samples compared against the gallery set



[55] Y. Lin *et al.*, "VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark," in *Proc. Interspeech*, 2024, pp. 4263-4267.

Experiments: Open-Set Speaker Identification

- Evaluation Metric: DIR@FAR
  - Detection and Identification Rate (DIR) = $\frac{①}{①+②+③}$
  - False Alarm Rate (FAR) = $\frac{⑤}{④+⑤}$
  - Higher is better



Experiments: Open-Set Speaker Identification

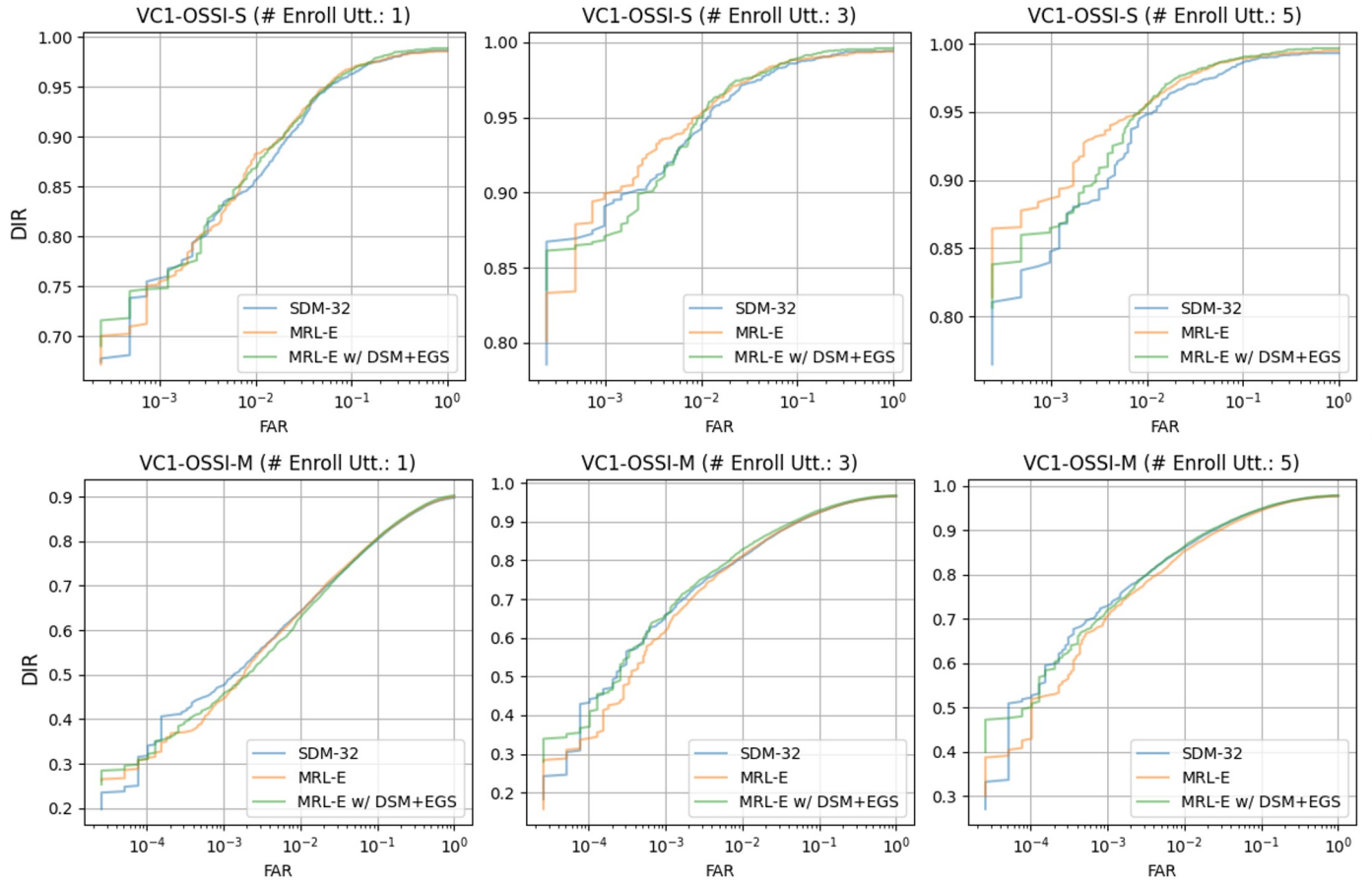
- Dataset: VoxCeleb
  - VoxBlink dataset is not publicly accessible
  - Construct evaluation sets using VoxCeleb instead
  - Following VoxBlink's evaluation subset settings

VoxBlink	VoxCeleb	Gallery Set	Probe Set (Known Queries)	Probe Set (Unknown Queries)
VB-Eval-S	VC1-OSSI-S	60	30	30
VB-Eval-M	VC1-OSSI-M	600	300	300
VB-Eval-L		6,000	3,000	3,000

[55] Y. Lin *et al.*, "VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark," in *Proc. Interspeech*, 2024, pp. 4263-4267.

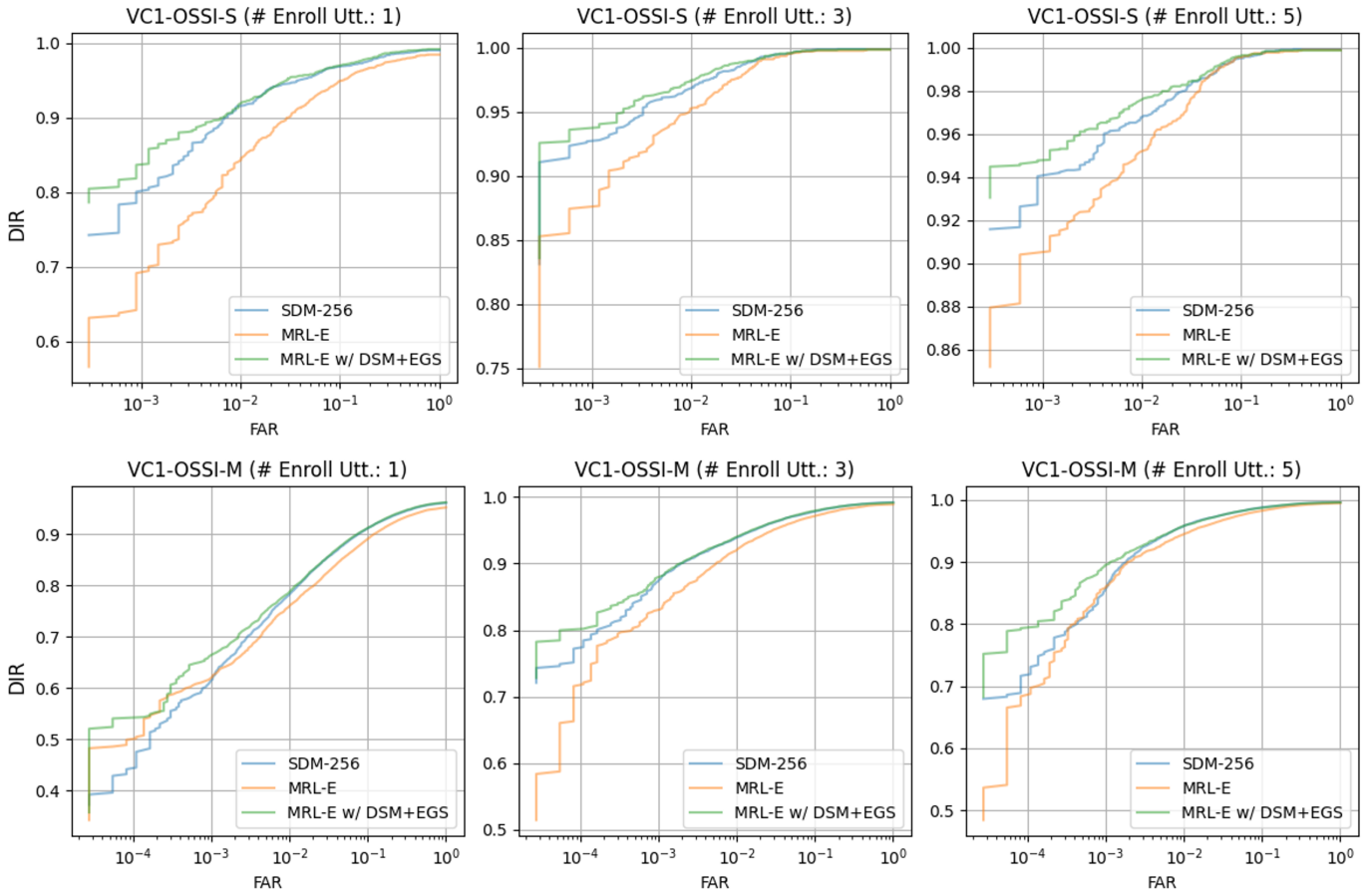
Experiments: Open-Set Speaker Identification

- Results: Model Comparison (32-D)
  - No noticeable performance difference in low dimension
  - Consistent with speaker verification results



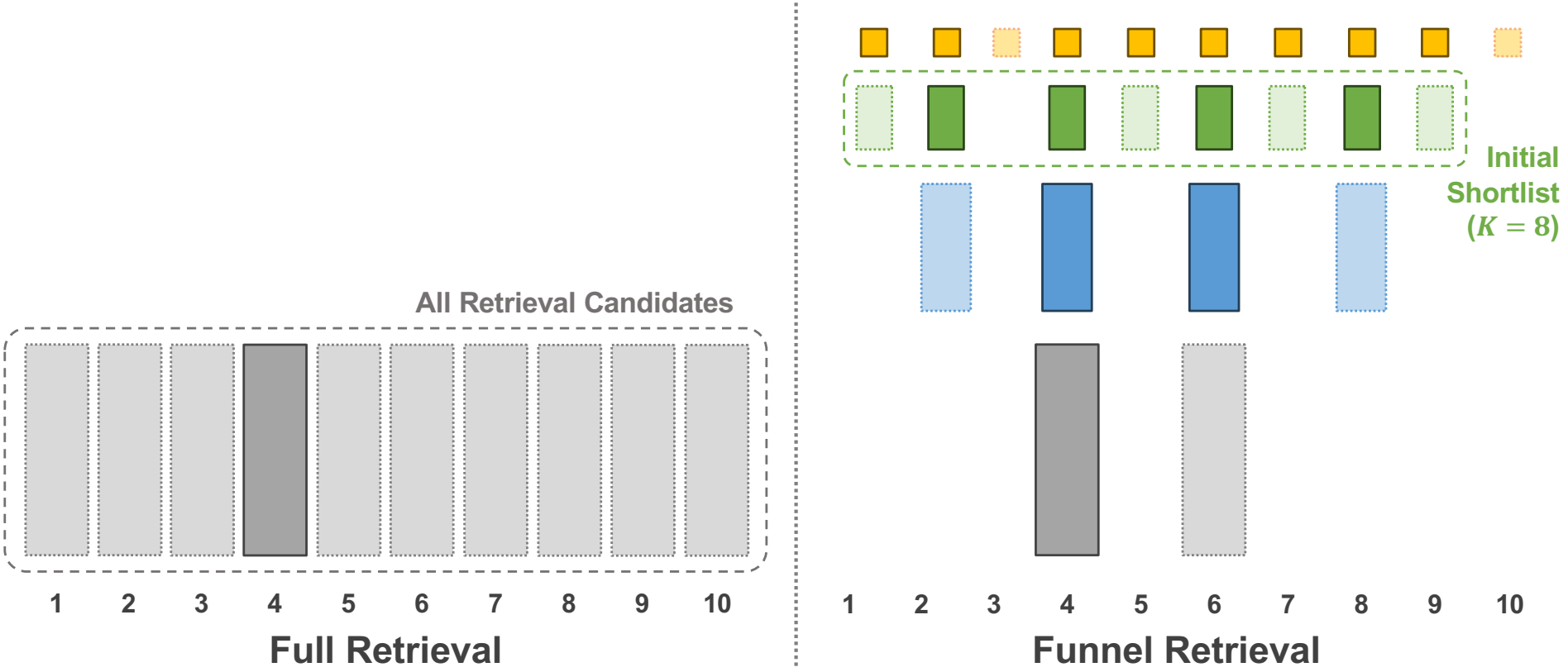
Experiments: Open-Set Speaker Identification

- Results: Model Comparison (256-D)
  - (MRL-E w/ DSM+EGS) > (SDM-256) > (MRL-E)
  - Difference is more pronounced under low FAR conditions



Experiments: Open-Set Speaker Identification

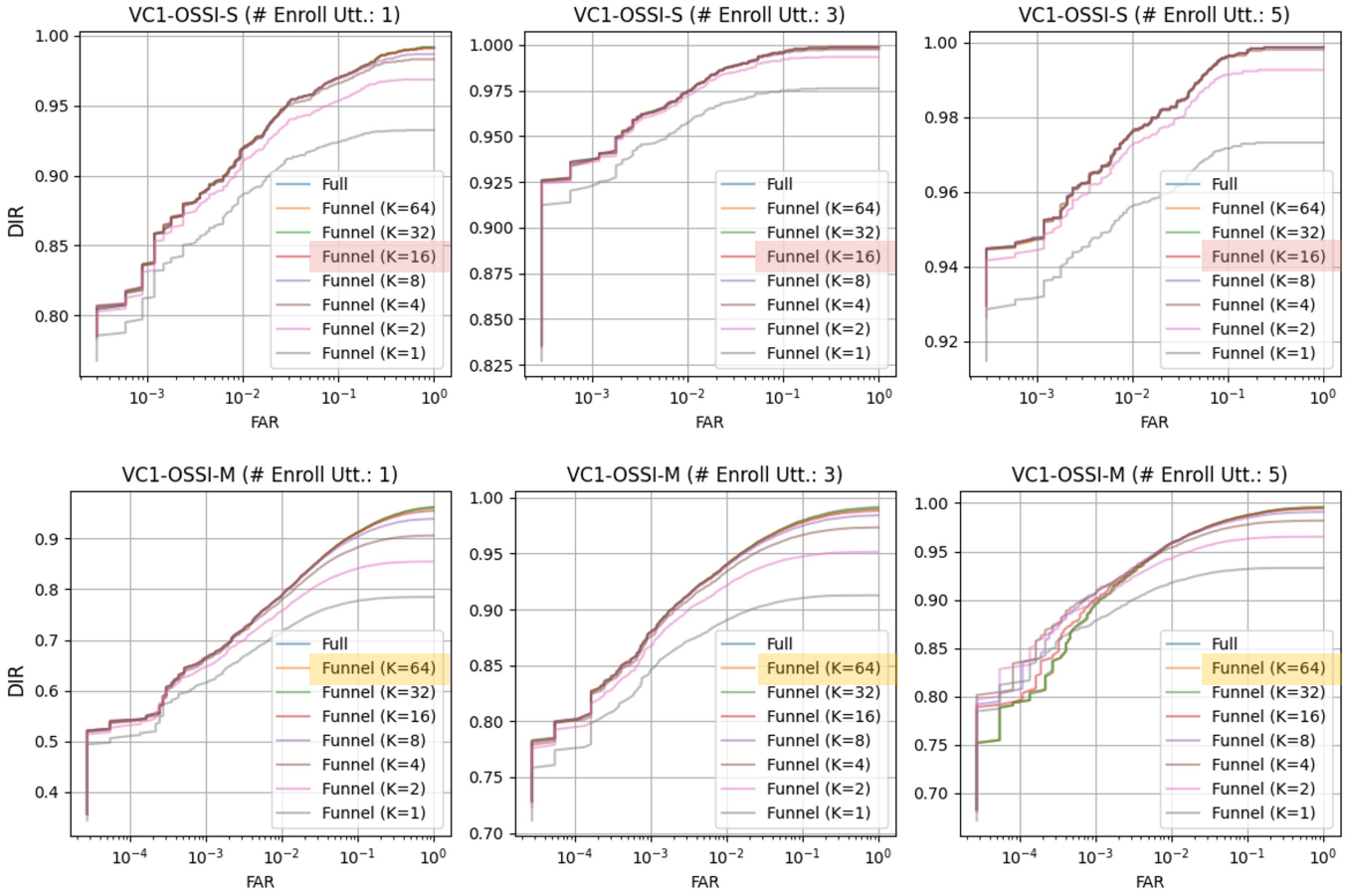
- Comparison of Retrieval Methods
  - Compare Full Retrieval (256-D) with Funnel Retrieval
  - Computation and performance vary with initial shortlist size K



[38] A. Kusupati *et al.*, "Matryoshka Representation Learning," in *Proc. Adv. in Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 30233–30249.

Experiments: Open-Set Speaker Identification

- Results: Retrieval Methods (MRL-E w/ DSM+EGS)
  - Reducing K lowers both performance and computation
  - Comparable performance to full retrieval at $K = 16$ (S), $K = 64$ (M)



Experiments: Open-Set Speaker Identification

- Results: Retrieval Methods (MRL-E w/ DSM+EGS)
  - VC1-OSSI-S: 95% less computation, ≤ 0.1% DIR@FAR drop
  - VC1-OSSI-M: 98% less computation, ≤ 0.1% DIR@FAR drop
  - Higher efficiency is expected with more enrolled speakers

Evaluation Set	# Enroll Spk.	# Enroll Utt.	Retrieval	K	MFLOPs	DIR@FAR [%]			
						0.001	0.01	0.1	1
VC1-OSSI-S	60	1	Full	–	562	83.7	92.0	97.0	99.2
			Funnel	64	83	83.7	92.0	97.0	99.2
			Funnel	32	46	83.7	92.0	97.0	99.2
			Funnel	16	28	83.6	92.0	97.0	99.1
			Funnel	8	18	83.6	91.9	96.9	98.7
			Funnel	4	18	83.6	91.9	96.6	98.3
			Funnel	2	18	83.1	91.0	95.3	96.9
			Funnel	1	18	81.2	88.7	92.4	93.2
		3	Full	–	562	94.1	97.5	99.6	99.9
			Funnel	64	83	93.6	97.5	99.6	99.9
			Funnel	32	46	93.6	97.5	99.6	99.9
			Funnel	16	28	94.1	97.5	99.6	99.9
			Funnel	8	18	93.6	97.5	99.6	99.8
			Funnel	4	18	94.0	97.5	99.5	99.8
			Funnel	2	18	93.9	97.3	99.2	99.3
			Funnel	1	18	92.2	95.7	97.5	97.6
		5	Full	–	562	94.8	97.6	99.6	99.9
			Funnel	64	83	94.8	97.6	99.6	99.9
			Funnel	32	46	94.7	97.6	99.6	99.9
			Funnel	16	28	94.7	97.6	99.6	99.9
			Funnel	8	18	94.7	97.6	99.7	99.9
			Funnel	4	18	94.7	97.6	99.6	99.8
			Funnel	2	18	94.4	97.3	99.1	99.3
			Funnel	1	18	93.1	95.6	97.2	97.3

Evaluation Set	# Enroll Spk.	# Enroll Utt.	Retrieval	K	MFLOPs	DIR@FAR [%]			
						0.001	0.01	0.1	1
VC1-OSSI-M	600	1	Full	–	5538	66.4	78.7	91.3	96.2
			Funnel	64	83	66.5	78.8	91.2	96.2
			Funnel	32	46	66.6	78.7	91.2	96.0
			Funnel	16	28	66.5	78.8	91.1	95.5
			Funnel	8	18	66.6	78.8	90.4	93.9
			Funnel	4	18	66.0	78.1	88.2	90.6
			Funnel	2	18	64.9	75.8	84.1	85.5
			Funnel	1	18	61.6	71.6	77.7	78.5
		3	Full	–	5538	88.0	94.0	97.9	99.2
			Funnel	64	83	88.0	94.0	97.9	99.1
			Funnel	32	46	88.0	94.0	97.9	99.0
			Funnel	16	28	88.0	94.0	97.8	98.8
			Funnel	8	18	87.9	93.8	97.5	98.4
			Funnel	4	18	87.6	93.4	96.7	97.3
			Funnel	2	18	86.7	92.1	94.7	95.2
			Funnel	1	18	84.5	89.0	91.0	91.3
		5	Full	–	5538	89.5	95.8	98.7	99.6
			Funnel	64	83	89.5	95.8	98.7	99.6
			Funnel	32	46	89.6	95.9	98.7	99.5
			Funnel	16	28	90.1	95.9	98.7	99.4
			Funnel	8	18	90.6	95.8	98.5	99.1
			Funnel	4	18	90.7	95.4	97.8	98.2
			Funnel	2	18	90.0	94.3	96.3	96.5
			Funnel	1	18	87.8	91.7	93.2	93.3

Conclusions

- DSM, EGS and their combination effectively enhance model performance in higher-dimensional embeddings
- The proposed methods achieve higher accuracy in open-set speaker identification, while preserving MRL-E's efficiency
- The concept of EGS can be applied to other multi-loss or multi-task learning scenarios

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. I EEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5329–5333.

[2] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Ba sed Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CV PR)*, 2016, pp. 770–778.

[5] H. Wang *et al.*, "Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit," in *Proc. IEEE Int. Conf. Acoust., Spee ch, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[6] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Inform ation in 2D ResNets to Enhance Speaker Verification," in *Proc. Interspeech*, 2021, pp. 2302-2306.

[7] J. Thienpondt and K. Demuynck, "ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, 2023, pp. 1–8.

[8] I. Yakovlev *et al.*, "Reshape Dimensions Network for Speaker Recognition," in *Proc. Interspeech*, 2024, pp. 3235-3239.

[9] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *Proc. IEEE Con f. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 539–546.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proc. IEEE Conf. Comp ut. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[11] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddi ngs," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1633–1644, Sept. 2018.

[12] J. Wang, K. C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-Based Deep Metric Learning for Speaker Recognition," in *Proc. IEE E Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 3652–3656.

[13] J. S. Chung *et al.*, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[14] Y. Sun *et al.*, "Noise-Disentanglement Metric Learning for Robust Speaker Verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc ess. (ICASSP)*, 2023, pp. 1–5.

[15] W. Liu *et al.*, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 212-220.

References

[16] H. Wang *et al.*, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5265–5274.

[17] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, July 2018.

[18] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4685–4694.

[19] X. Xiang *et al.*, "Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2019, pp. 1652–1656.

[20] Y. Liu, L. He, and J. Liu, "Large Margin Softmax Loss for Speaker Verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.

[21] Y. Wei, J. Du, and H. Liu, "Angular Margin Centroid Loss for Text-Independent Speaker Recognition," in *Proc. Interspeech*, 2020, pp. 3820–3824.

[22] B. Han, Z. Chen and Y. Qian, "Exploring Binary Classification Loss for Speaker Verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[23] Y. Zhang, S. Herdade, K. Thadani, E. Dodds, J. Culpepper and Y. -N. Ku, "Unifying Margin-Based Softmax Losses in Face Recognition," in *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis. (WACV)*, 2023, pp. 3537-3546.

[24] D. Zhou *et al.*, "Dynamic Margin Softmax Loss for Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3800–3804.

[25] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality Adaptive Margin for Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18729–18738.

[26] R. Li, S. Fang, C. Ma, and L. Li, "Adaptive Rectangle Loss for Speaker Verification," in *Proc. Interspeech*, 2022, pp. 301–305.

[27] L. Zhang, Z. Chen, and Y. Qian, "Adaptive Large Margin Fine-Tuning for Robust Speaker Verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[28] Li *et al.*, "HAMFace: Hardness Adaptive Margin Loss for Face Recognition with Various Intra-class Variations," *Expert Syst. Appl.*, vol. 240, Art. no. 122384, April 2024.

[29] Xu *et al.*, "X2-Softmax: Margin Adaptive Loss Function for Face Recognition," *Expert Syst. Appl.*, vol. 249, Art. no. 123791, Sept. 2024.

[30] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14220–14229.

References

[31] N. Kuzmin, I. Fedorov, and A. Sholokhov, "Magnitude-Aware Probabilistic Speaker Embeddings," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, 2022, pp. 1–8.

[32] Q. Liu, X. Zhang, X. Liang, Y. Qian, and S. Yao, "AWLloss: Speaker Verification Based on the Quality and Difficulty of Speech," *IEEE Signa l Process. Lett.*, vol. 30, pp. 1337–1341, 2023.

[33] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World G raphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020.

[34] D. Cai, "A Revisit of Hashing Algorithms for Approximate Nearest Neighbor Search," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2337 –2348, Jun. 2021.

[35] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized Product Quantization for Approximate Nearest Neighbor Search," in *Proc. IEEE Conf. Comput. V is. Pattern Recognit. (CVPR)*, 2013, pp. 2946–2953.

[36] Y. Ma and L. Zhu, "A Review on Dimension Reduction," *Int. Stat. Rev.*, vol. 81, no. 1, pp. 134–150, 2013.

[37] G. Chandrashekar and F. Sahin, "A Aurvey on Feature Selection Methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[38] A. Kusupati *et al.*, "Matryoshka Representation Learning," in *Proc. Adv. in Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 30233–30249.

[39] S. Wang, P. Zhu, and H. Li, "M-Vec: Matryoshka Speaker Embeddings with Flexible Dimensions," 2024, *arXiv:2409.15782*.

[40] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[41] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Trans. Au dio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[42] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Trans. A udio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[43] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. Audi o, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[44] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE/ACM Trans. Audi o Speech Lang. Process.*, vol. 19, no. 4, pp. 788-798, May 2011.

[45] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proc. IEEE Conf. Comput. Vis. P attern Recognit. (CVPR)*, 2007, pp. 1-8.

References

[46] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score Normalization for Text-Independent Speaker Verification systems,” *Digital Signal P rocess.*, vol. 10, no. 1, pp. 42–54, 2000.

[47] H. Aronowitz, D. Irony, and D. Burshtein, “Modeling Intra-Speaker Variability for Speaker Recognition,” in *Proc. Interspeech*, 2005, pp. 2177 -2180.

[48] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, 2010, pp. 14.

[49] Z. N. Karam, W. M. Campbell, and N. Dehak, “Towards Reduced False-Alarms Using Cohorts,” in *Proc. IEEE Int. Conf. Acoust., Speech, Sign al Process. (ICASSP)*, 2011, pp. 4512-4515.

[50] N. Brümmer and E. de Villiers, “The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF,” 2013, *arXiv:1304.2865*.

[51] J. Thienpondt, B. Desplanques and K. Demuynck, “The Idlab VoxSRC-20 Submission: Large Margin Fine-Tuning and Quality-Aware Score C alibration in DNN Based Speaker Verification,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 5814-5818.

[52] Y. Zheng, Y. Zhang, C. Niu, Y. Zhan, Y. Long and D. Xu, “Score Calibration Based on Consistency Measure Factor for Speaker Verification,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 12371-12375.

[53] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in *Proc. Interspeech*, 2017, pp. 2616–2 620.

[54] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in *Proc. Interspeech*, 2018, pp. 1086–1090.

[55] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, “VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Ident ification Benchmark,” in *Proc. Interspeech*, 2024, pp. 4263-4267.

[56] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, *arXiv:1510.08484*.

[57] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A Study on Data Augmentation of Reverberant Speech for Robust Speech Rec ognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 5220–5224.

[58] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, “Speaker Augmentation and Bandwidth Extension for Deep Speaker Embedding,” in *P roc. Interspeech*, 2019, pp. 301–305.