CAARMA: Class Augmentation with Adversarial Mixup Regularization

Anonymous ACL submission

Abstract

Speaker verification is a typical zero-shot learning task, where inference of unseen classes is performed by comparing embeddings of test instances to known examples. The models performing inference must hence naturally generate embeddings that cluster same-class instances compactly, while maintaining separation across classes. In order to learn to do so, they are typically trained on a large number of classes (speakers), often using specialized losses. However real-world speaker datasets often lack the class diversity needed to effectively learn this in a generalizable manner. We introduce CAARMA, a class augmentation framework that addresses this problem by generating synthetic classes through data mixing in the embedding space, expanding the number of training classes. To ensure the authenticity of the synthetic classes we adopt a novel adversarial refinement mechanism that minimizes categorical distinctions between synthetic and real classes. We evaluate CAARMA on multiple speaker verification tasks, as well as other representative zero-shot comparison-based speech analysis tasks and obtain consistent improvements: our framework demonstrates a significant improvement of 8% over all baseline models. Code for CAARMA will be released.

1 Introduction

Speaker verification is fundamentally a zero-shot learning (ZSL) task, where verification is accomplished by comparing embeddings from enrollment and verification samples without the need for further training (Wan et al., 2018). This process aligns with the principles of ZSL, where models are expected to operate effectively on unseen data. Therefore, while the following discussion is framed within the broader context of ZSL, it is specifically tailored to address the challenges in speaker verification.

To address the challenge of limited class diversity in training datasets a common issue in speaker verification we propose a novel augmentation-based training paradigm. This approach leverages synthetic data augmentation to enhance the robustness and generalization capabilities of speaker verification systems, particularly in low-diversity environments. Our method not only stays true to the essence of ZSL by facilitating effective generalization to new speakers but also introduces a practical solution to overcome the inherent limitations of traditional training datasets.

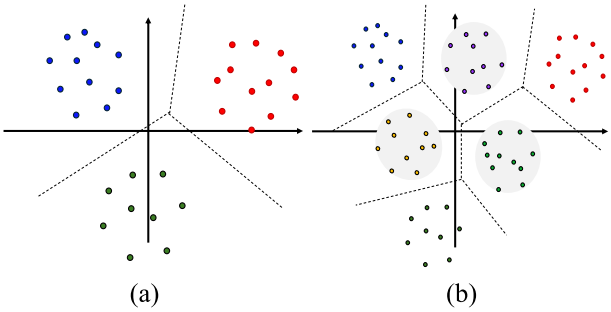

Figure 1: (a) When trained with fewer classes the model can spread the embeddings of individual classes out while still learning to classify the training data accurately and with large margins. This will not, however translate to compact representations for newer unseen classes. (b) With additional synthetic classes (shaded grey), the model must now learn to compact classes more. This will translate to more compact unseen classes as well.

Effective zero-shot learning lies in generating embeddings that cluster same-class (in our case, same-speaker) instances closely while maintaining separation between different classes (Zhu et al., 2019). Traditional training approaches for ZSL models rely on two key components: exposure to a large number of diverse classes and the use of specialized loss functions that promote both inter-class separation and intra-class compactness (Min et al., 2020). The underlying principle here is that

by learning from a sufficiently large number of classes, and through proper encouragement embodied in the losses, the model learns not merely to separate the classes it has seen, but *the more general principle* that instances from a class must be clustered closely together while begin separated from those from other classes (Xian et al., 2018).

However, when the training datasets lack the necessary variety of classes (speakers), this can severely limit the model's ability to develop robust and transferable representations (Xie et al., 2022). Indeed, it may be argued that in the high-dimensional space of the embeddings, *any* finite set of training classes is insufficient to cover the space adequately. This limitation leads to models that fail to generalize effectively to unseen categories, resulting in suboptimal zero-shot inference performance (Gupta et al., 2021).

Popular approaches to address training data limitations often rely on data augmentation techniques. Methods such as AutoAugment (Cubuk et al., 2019) and SpecAugment (Park et al., 2019) generate new samples by modifying existing ones through transformations like geometric distortions, time warping, and frequency masking. However, while these techniques increase intra-class diversity, they do not introduce new classes, limiting their effectiveness in zero-shot learning scenarios. Of most relevance to our paper are *mixup*-based data augmentation techniques, *e.g.* (Verma et al., 2019; Yun et al., 2019), that aim to enhance training by generating new samples through interpolation (Han et al., 2021) of both the features and their labels. Regardless of the interpolation, yet, these methods too do not generate *new* classes; they merely improve the generalization of the model by mapping mixed data to mixed class labels. Still other methods use generative models such as VAEs, GANs, diffusion models *etc.* to generate authentically novel data to enhance the training (Min et al., 2019); however these too are generally restricted to generating novel instances for *known* classes, limiting their effectiveness in zero-shot scenarios (Pourpanah et al., 2022). Thus, while these approaches are generally very successful in improving generalization in classification problems, they fail at addressing the problem ZSL learning faces, that of increasing the number of classes themselves, leading to inconsistent generalization to unseen classes (Xie et al., 2022).

In this paper, we introduce *Class Augmentation with AdversaRial Mixup regulariAztion* (CAARMA), a data augmentation framework to introduce *synthetic* classes (speakers) to enhance ZSL training for speaker verification. CAARMA utilizes a mixup-like strategy to generate data from fictitious speakers. However, unlike conventional mixup which mixes data in the input space, which would arguably be meaningless in our setting (a straight-forward mix of two speech recordings will merely result in a mixed signal, and not a new speaker), the mixup is performed in the embedding space in a manner that permits assignment of new class identities to the mixed-up data. Critically, we must now ensure that the mixed-up embeddings resemble those from actual speakers. We do so through a discriminator that is used to minimize categorical distinctions between synthetic and authentic data through adversarial training.

We demonstrate CAARMA's effectiveness through extensive evaluation on speaker verification, where it achieves substantial improvements in generalizing to diverse speaker distributions. Additional experiments on other ZSL speech tasks further validate our approach's broad applicability.

Our main contributions are as follows:

- We introduce CAARMA, a novel class augmentation framework that addresses the fundamental limitation of class diversity in zero-shot learning by generating synthetic classes termed as Sythetic Label Mixup (SL-Mixup) through embedding-space mixing, rather than conventional input-space augmentation.

- We develop an adversarial training mechanism that ensures the synthetic classes generated through our mixing strategy maintain statistical authenticity by minimizing categorical distinctions between real and synthetic embeddings.

- We achieve significant performance improvements in zero-shot inference, demonstrated through an 8% improvement over baseline models in speaker verification tasks, with enhanced generalization to diverse speaker distributions and verified applicability across various zero-shot learning tasks.

2 Related-Work

Mixup. The development of mixup strategies has evolved substantially since Mixup's original introduction by (Zhang et al., 2018), which generated

virtual samples and mixed labels by linearly combining two input samples and their corresponding labels. This pioneering method proved particularly successful in enhancing data diversity and improving generalization in visual classification tasks. Extensions such as ManifoldMix (Verma et al., 2019) applied this concept to hidden layers, while CutMix (Yun et al., 2019) introduced a patch-based approach by blending rectangular sections of images, offering a novel alternative for augmenting training data. Subsequent mixup strategies focused on tailoring data mixing to specific contexts or improving the precision of mixing. Static policies like SmoothMix (Lee et al., 2020), GridMix (Baek et al., 2021), and ResizeMix (Qin et al., 2023) used hand-crafted cutting techniques, while dynamic approaches such as PuzzleMix (Kim et al., 2020) and AlignMix (Venkataramanan et al., 2022) incorporated optimal-transport methods to determine mix regions with greater flexibility. For Vision Transformers, strategies such as TransMix (Chen et al., 2022) and TokenMix (Liu et al., 2022b) focused on leveraging attention mechanisms to refine mixing operations, particularly for transformer architectures (Dosovitskiy et al., 2021). Recent developments have adapted mixup techniques to tasks beyond classification, such as regression. C-Mixup, for instance, applies sample mixing based on label distances, using a symmetric Gaussian kernel to select samples that improve regression performance (Cai et al., 2021). Further enhancing robustness, RC-Mixup integrates C-Mixup with multi-round robust training, creating a feedback loop where C-Mixup helps identify cleaner data, and robust training improves the quality of data for mixing (Liu et al., 2022a). These specialized approaches reveal mixup's adaptability across various machine learning tasks, enhancing model performance and data resilience. However, it's important to note that none of these methods involve mixing in the embedding space, which could allow for the creation of entirely new and synthetic class identities,

Synthetic speech. Recent advancements in synthetic audio generation have emphasized the creation of diverse and high-quality datasets, crucial for training and evaluating audio-based AI models. A notable innovation is ConversaSynth, a framework utilizing large language models (LLMs) to generate synthetic conversational audio across varied persona settings (Gao et al., 2022). This method begins with generating text-based dialogues, which are then rendered into audio using text-to-speech (TTS) systems. The synthetic datasets produced are noted for their realism and topic variety, proving beneficial for tasks such as audio tagging, classification, and multi-speaker speech recognition. These capabilities make ConversaSynth a valuable tool for developing robust, adaptable AI models that can handle diverse audio data and complex conversational contexts. In the domain of speaker verification, SpeechMix introduces a novel method by mixing speech at the waveform level, carefully adjusting ratios to preserve the distinct characteristics of speaker identity (Jindal et al., 2020). However, like many other generative and augmentation techniques, SpeechMix primarily focuses on manipulating known speaker voices rather than generating new identities, thereby limiting its utility for enhancing speaker diversity critical for effective zero-shot learning. Synthio employs a unique approach by using text-to-audio (T2A) diffusion models to augment small-scale audio classification datasets (Joassin and Alvarez-Melis, 2022). It enhances compositional diversity and maintains acoustic consistency by aligning T2A-generated synthetic samples with the original dataset using preference optimization. Additionally, exploring style transfer in synthetic audio, recent work by Ueda et al. employs a VITS-based voice conversion model, conditioned on the fundamental frequency (F0), to produce expressive variations from neutral speaker voices (Ueda et al., 2024). This method achieves cross-speaker style transfer in a FastPitch-based TTS system, incorporating a style encoder pre-trained on timbre-perturbed data to prevent speaker leakage. This technique enhances the utility of synthetic data in applications requiring rich stylistic diversity. These developments underline the increasing sophistication of synthetic audio generation techniques, from multi-speaker conversations in ConversaSynth to Synthio's optimized T2A augmentation for classification, and cross-speaker style transfers with VITS-based models. They collectively demonstrate the power of synthetic data to enrich audio model training by adding diversity and realism. However, these methods still face limitations in generating entirely new speaker identities, which is critical for expanding the range of recognizable voices in speaker verification systems. CAARMA addresses this gap by directly mixing in the embedding space, creating synthetic speakers that enhance zero-shot learning capabilities. CAARMA not only preserves speaker

characteristics but also significantly expands the diversity of speaker identities, offering a superior solution for training more robust and adaptable speaker verification systems.

3 Class Augmentation with Adversarial Mixup Regularization

3.1 Overview

As mentioned in Section 1, ZSL models learn their ability to compactly cluster same-class embeddings while maintaining separation between classes primarily through exposure to a large number of classes during training; the more classes they are exposed to in training, the better they are able to generalize to unseen classes. In the speaker verification setting, this translates to training the model with recordings from a large number of speakers; the more the number of training speakers the better the model generalizes. To improve this generalization CAARMA attempts to increase the number of speakers by creating synthetic speakers while training. Synthetic speakers may be created through generative methods such as (Cornell et al., 2024); however this approach does not scale. Instead, CAARMA creates them through a simple mixup strategy, as convex combinations of real speakers, with a key distinction: the mixup is performed in the *embedding* space, where the classes are expected to form compact (and generally convex) clusters. In order to ensure that these synthetic speakers are indeed representative of actual speakers, CAARMA utilizes a *semantic discriminator*, a discriminator which attempts to distinguish between synthetic and real speakers: if this discriminator is fooled, the synthetic speakers are statistically indistinguishable from real ones.

When training the model, a conventional loss such as the Additive Margin Softmax (AM-Softmax) (Wang et al., 2018) is used. The synthetic classes, which are created dynamically during training, are included by dynamically also expanding the class labels in the loss. In addition, the model also attempts to adversarially fool the discriminator. Once the model is trained, the discriminator is no longer needed and is discarded.

3.2 Framework

Our framework consists of three main components: an encoder for embedding generation, a synthetic label mixup mechanism for class augmentation, and an adversarial training scheme with a semantic discriminator. Figure 2 illustrates the complete pipeline of our approach. The process begins with a waveform input that is transformed into a Mel-spectrogram. This spectrogram serves as input to the encoder $\mathcal{E}$, which generates embeddings e that capture discriminative speaker characteristics. These embeddings undergo our SL-Mixup strategy, which generates synthetic embeddings e_{syn} by mixing embeddings e based on their closest neighbor weights W. Each synthetic embedding receives a corresponding synthetic label ID_{syn} within the mini-batch. The framework employs two primary loss functions: the encoder loss $\mathcal{L}_{\text{real}}$ for original embeddings and the synthetic loss $\mathcal{L}_{\text{syn}}$ for synthetic embeddings. A Self-Supervised Learning (SSL) model serves as the semantic discriminator to distinguish between real (R) and synthetic (S) embeddings. A discriminator loss $\mathcal{L}_D$ guides the discriminator to maximally distinguish between R and S. On the other hand, a generator loss $\mathcal{L}_{\text{gen}}$ guides the encoder to "fool" the discriminator, so that it perceives no distinction between real and synthetic embeddings.

3.3 Encoder

The encoder $\mathcal{E}$ transforms Mel-spectrograms into discriminative embeddings e that capture speaker-specific acoustic features. We employ an MFA-Conformer model (Zhang et al., 2022) as our encoder architecture, which combines feed-forward networks (FFNs), multihead self-attention (MHSA), and convolution modules. The model incorporates positional embeddings to handle variable-length input sequences effectively. For training, we utilize the AM-Softmax function as our encoder loss $\mathcal{L}_{\text{real}}$.

$$\mathcal{L}_{\text{real}} = -\log \frac{e^{s\cdot(\cos(\theta_y)-m)}}{\sum_{j=1}^{C} e^{s\cdot\cos(\theta_j)}}$$

where s is a scaling factor used to stabilize gradients, C represents the number of classes, $\cos(\theta_y)$ denotes the cosine similarity for the true class, and m is an additive margin that enhances class separation by increasing inter-class distances.

3.4 Synthetic Label Mixup

Our SL-Mixup strategy generates synthetic embeddings e_{syn} within each mini-batch by mixing embeddings e according to their closest neighbor weights W, as detailed in Algorithm 1. This approach ensures synthetic embeddings remain

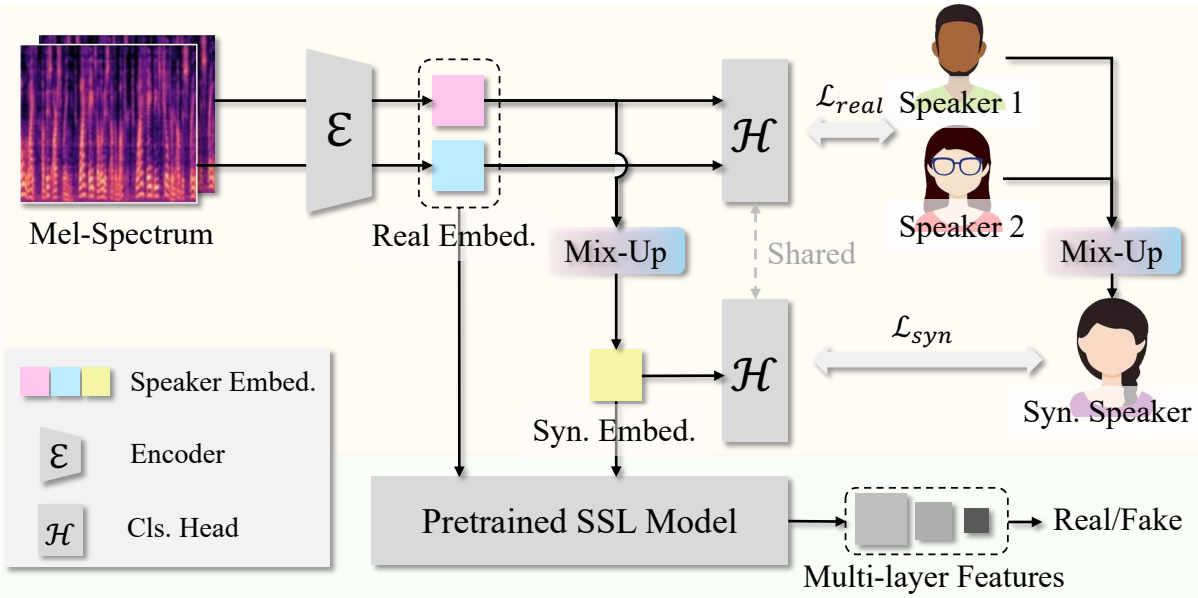

Figure 2: Illustration of CAARMA framework. (a) The encoder ($\mathcal{E}$) extracts embeddings from Mel-spectrograms, which are processed by a classification head ($\mathcal{H}$) for speaker identification and through Mix-Up for synthetic embedding generation. (b) Both real and synthetic embeddings are fed into a pretrained SSL model that acts as a discriminator, distinguishing between real and synthetic samples.

within the same manifold as real embeddings, avoiding arbitrary generation. The strategy creates synthetic labels ID_{syn} and embeddings dynamically during training, enabling effective representation learning and facilitating the potential use of unlabeled data. To ensure that synthetic speakers are minimally confusable with their component speakers, we only combine pairs of speakers with a fixed weight of 0.5. This approach maintains a balanced contribution from each component speaker, preventing synthetic embeddings from collapsing into a single identity while maintaining inter-class separation. The synthetic loss $\mathcal{L}_{\text{syn}}$ is computed using the AM-Softmax loss function applied to synthetic embeddings e_{syn}. This loss is integrated into the main encoder loss $\mathcal{L}_{\text{real}}$, scaled by $1/\lambda$, where λ represents the number of speakers. This integration ensures proper alignment of synthetic embeddings within the embedding manifold.

3.5 Adversarial Training

Our adversarial training process alternates between optimizing the encoder and discriminator, as described in Algorithm 2. This optimization scheme continuously refines the embedding manifold through the interaction between real and synthetic embeddings:

Algorithm 1 SL-Mixup

Input: Feature matrix X, Label vector Y, Weight matrix W
Initialize $W_{\text{syn}} \leftarrow \mathbf{0}$, $Y_{\text{syn}} \leftarrow \mathbf{0}$, $X_{\text{syn}} \leftarrow \mathbf{0}$
for $y_i \in Y$ **do**
  distances $\leftarrow \|\mathbf{W}[:,i] - \mathbf{W}[:,j]\|_2 \quad \forall j \in$ label_set $\setminus \{i\}$
  neighbor$(y_i) \leftarrow \arg\min(\text{distances})$
end for
for $i \in$ Batch **do**
  $l_1 \leftarrow Y[i],\ l_2 \leftarrow \text{neighbor}(l_1)$
  $W_{\text{syn}}[:,i] \leftarrow 0.5 \times (W[:,l_1] + W[:,l_2])$
  $Y_{\text{syn}}[i] \leftarrow \text{new_label}(l_1, l_2)$
  $\text{index}[i] \leftarrow \text{find}(Y = l_2)$
  $X_{\text{syn}}[i] \leftarrow 0.5 \times (X[i] + X[\text{index}[i],:])$
end for
Return: $X_{\text{syn}}, Y_{\text{syn}}, W_{\text{syn}}$

- **Discriminator Training:** The discriminator D learns to differentiate between real embeddings e and synthetic embeddings e_{syn} using features extracted from multiple model layers. The discriminator loss is defined as:

$$\mathcal{L}_D = \text{BCE}(D(e), 1) + \text{BCE}(D(e_{\text{syn}}), 0) \tag{1}$$

Algorithm 2 Adversarial Training with Synthetic Embeddings

Input: Feature extractor $f(X)$, Model M, Discriminator D, Dataset $\mathcal{D}$ (waveforms X and labels Y), Adversarial weight λ_{adv}
for each epoch $e \in [1, N_{\text{epochs}}]$ **do**
  for each batch $(X, Y) \in \mathcal{D}$ **do**
    Extract features $F = \text{Mel}(X)$
    Compute embeddings $e = \mathcal{E}(F)$
    Compute AM-Softmax loss $\mathcal{L}_{\text{real}}$
    Generate synthetic embeddings e_{syn} via mixup
    Compute real predictions $D(e)$ and fake predictions $D(e_{\text{syn}})$
    Compute discriminator loss:
    $\mathcal{L}_D = \text{BCE}(D(e), 1) + \text{BCE}(D(e_{syn}), 0)$
    Update D using $\nabla \mathcal{L}_D$
    Compute generator loss
    $\mathcal{L}_G = \text{BCE}(D(e_{\text{syn}}), 1) + \text{BCE}(D(e), 0)$
    Adjust λ_{adv} based on $\mathcal{L}_{\text{real}}/\mathcal{L}_G$
    Compute total loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{real}} + \lambda_{\text{adv}}\mathcal{L}_G$

    Update M using $\nabla \mathcal{L}_{\text{total}}$
  end for
end for
Return: Trained M and D

where BCE represents binary cross-entropy loss.

- **Generator Loss:** The encoder incorporates a generator loss $\mathcal{L}_G$ that guides embedding alignment with the manifold structure:

$$\mathcal{L}_G = \text{BCE}(D(e_{\text{syn}}), 1) + \text{BCE}(D(e), 0) \tag{2}$$

3.6 Semantic Discriminator

To enhance the discriminative power of our framework, we incorporate a self-supervised model (HuBERT) (Hsu et al., 2021) as a semantic discriminator. This discriminator leverages the pre-trained representations to provide richer gradients during adversarial training, improving the stability and quality of the learned embeddings. The semantic discriminator processes embeddings through an adapter module that projects them into a compatible feature space. The adapter consists of a down-projection layer with spectral normalization, followed by fully connected layers with GELU activation (Hendrycks and Gimpel, 2016). We employ skip connections and layer normalization to ensure stable training. The discriminator extracts features from multiple HuBERT (Hsu et al., 2021) layers (7, 9, 11, and 12) to capture diverse speaker characteristics. These features are combined using learnable weights and processed through a residual classification block with spectral normalization and LeakyReLU activation (Xu, 2015) to determine whether an embedding is real or synthetic.

4 Experiments

4.1 Datasets

We utilize four datasets in our proposed approach: VoxCeleb1 (Nagrani et al., 2017), VoxCeleb2 (Chung et al., 2018), and two datasets from the Dynamic-SUPERB (Huang et al., 2024) benchmark such as HowFarAreYou and DailyTalk datasets. The dataset statistics are summarized in Table 1. These datasets were employed across different tasks to evaluate the adaptability and generalizability of our pipeline:

- **Speaker Identification:** The primary task of our study involves speaker identification using VoxCeleb1 and VoxCeleb2. These large-scale datasets contain speech recordings from thousands of speakers.
- **Speaker Distance Estimation** The HowFarAreYou dataset originates from the 3DSpeaker dataset, designed to assess the distance of a speaker from the recording device. The task involves predicting distance labels (e.g., 0.4m, 2.0m) based on speech recordings.
- **Emotion Recognition:** We use the DailyTalk dataset to classify the emotional state of a speaker based on speech utterances. This dataset contains speech samples labeled with seven distinct emotion categories. To maintain consistency with other datasets, we resample all recordings to 16 kHz before processing.

Table 1 provides an overview of the datasets used in our experiments, including the number of classes and total utterances per dataset.

4.2 Experimental Setup

4.2.1 Model Architecture

We train two baseline architectures for speaker verification:

ECAPA-TDNN (Desplanques et al., 2020): Contains three SE-Res2Blocks with 1024 channels (20.8M parameters).

Table 1: Dataset statistics used in our experiments.

ID	DATASET	CLASSES	UTTERANCES
1	VOXCELEB1	1211	1,087135
2	VOXCELEB2	5994	153,516
3	HOWFARAREYOU	3	30,00
4	DAILYTALK	7	16,600

Table 2: EER Results for MFA Conformer baseline, Adversarial Training (AT), Semantic Discriminator (SD), and Synthetic Loss L_{syn} using VoxCeleb1 for the SV tasks.

ID	L_{syn}	AT	SD	RESULTS
1				3.33
2	✓			3.28
3		✓		3.15
4		✓	✓	3.18
5	✓	✓		3.17
6	✓	✓	✓	**3.09**

MFA-Conformer (Zhang et al., 2022): Employs 6 Conformer blocks with 256-dimensional encoders, 4 attention heads, and convolution kernel size of 15 (19.7M-20.5M parameters).

Both architectures generate 192-dimensional embeddings for fair comparison. For emotion and distance tasks, we utilize HuBERT-Large (pretrained on LibriSpeech) with 1024-dimensional embeddings and 768 hidden units.

4.2.2 Adversarial Training

We incorporate adversarial training into our baseline experiments. In this approach, each model is retrained from scratch within our adversarial framework. The discriminator is trained concurrently with the encoder. The discriminator's role is to effectively distinguish between real and synthetic embeddings, enforcing a well-structured representation.

Semantic discriminator. To determine the most informative HuBERT hidden layers for speaker representation, we conduct an ablation study. We experiment with different layer configurations, including $\{h_3, h_6, h_9, h_{12}\}$, $\{h_7, h_9, h_{11}, h_{12}\}$, $\{h_6, h_7, h_8, h_9\}$ and evaluate their impact on model performance.

4.2.3 Implementation Details

We implement all baseline systems and discriminators using the PyTorch framework (Yun et al., 2019). Each utterance is randomly segmented into fixed 3-second chunks, with 80-dimensional Fbanks as input features, computed using a 25 ms window length and a 10 ms frame shift, without applying voice activity detection. All models are trained using AM-Softmax loss with a margin of 0.2 and a scaling factor of 30. We use the AdamW optimizer with an initial learning rate of 0.001 for model training, while the discriminator is optimized separately with AdamW at an initial learning rate of 2e-4. To prevent overfitting, we apply a weight decay of 1e-7 and use a linear warmup for the first 2k steps, though no warmup is applied to the discriminator. Training is conducted on NVIDIA V100 GPUs with a batch size of 50, and all models are trained for 30 epochs.

Table 3: EER Results for two different encoders ECAPA-TDNN and MFA Conformer showing performance in baseline, Adeversarial Training (AT), and Synthetic Loss (L_{syn}) on VoxCeleb1-O.

ENCODER	BASELINE	AT	AT+$\mathcal{L}_{syn}$
ECAPA TDNN	4.22	3.96	3.87
MFA CONFORMER	3.33	3.18	3.09

Table 4: Ablation study of different hidden layers for Semantic Discriminator (SD) reporting EER (%) and minDCF.

ID	HIDDEN LAYERS	EER (%)	MINDCF
1	h_3, h_6, h_9, h_{12}	3.22	0.31
2	h_6, h_7, h_8, h_9	3.12	0.30
3	h_7, h_9, h_{11}, h_{12}	**3.09**	0.28

5 Results & Analysis

To validate the effectiveness of our approach, we conduct comprehensive experiments across speaker verification, emotion classification, and speaker distance estimation tasks. Our analysis demonstrates significant improvements through adversarial refinement on model generalization.

5.1 Speaker Verification Task

We evaluate our models on VoxCeleb1-O, the official test set of VoxCeleb1.

5.1.1 Small Scale

Our initial evaluations focused on models trained on VoxCeleb1 to facilitate thorough experimentation with different model configurations.

Using the MFA Conformer as the Encoder, we conduct several experiments (Table 2) reporting the Equal Error Rate (EER). The addition of synthetic loss alone (Model ID_2) yield slight improvements over the baseline (Model ID_1). More sub-

Table 5: *Performance overview of all systems on VoxCeleb1-O*

Model	**# Parameters**	**EER(%)**	**minDCF**
ResNet34	23.2M	1.03	0.112
ECAPA-TDNN	20.8M	0.82	0.112
MFA-Conformer	19.8M	0.86	0.066
MFA-Adversarial	19.8M	0.81	0.036

Table 6: Classification accuracies for Hubert Encoder baseline and with Adversarial Training on two different tasks.

DATA SET	BASELINE	AT
EMOTION CLASSIFICATION	83%	85.50%
HOWFARSPK	77.91%	79.97%

stantial gains were achieved through adversarial training (Model ID_3), with further improvements observed when incorporating the semantic discriminator (Model ID_4). The best performance was achieved by Model ID_6, which combined all three components: synthetic loss, semantic discriminator, and adversarial training.

To further validate the generalizability of our approach, we implement it with an alternative speaker encoder. As shown in Table 3, the combination of adversarial training and semantic discriminator improved performance by 6.56% compared to the baseline. Adding synthetic loss further enhanced the improvement to 8.29%.

We conduct an ablation study to identify the most informative hidden layers for speaker representation. Table 4 presents the EER and minimum Detection Cost Function (minDCF) across various layer configurations. Our analysis revealed that layers 7, 9, 11, and 12 provide the most effective speaker characteristics representation, suggesting that later layers capture more valuable speaker-specific information.

5.1.2 Large Scale

To demonstrate the scalability of our framework, we conduct experiments on models trained on both VoxCeleb1 and VoxCeleb2, as a large-scale dataset. As shown in Table 5, we incorporate the MFA Conformer as the encoder within our adversarial training framework, leveraging its strong baseline performance on smaller datasets. The results indicate that our adversarial refinement approach scales effectively to larger datasets, consistently improving both EER and minDCF metrics.

5.2 Emotion and Speaker Distance Tasks

To demonstrate the generalizability of our approach across different speech processing domains, we evaluate its effectiveness on emotion classification and speaker distance estimation using the DailyTalk and HowFarAreYou test sets, respectively. As shown in Table 6, our method improved classification accuracy across both tasks: emotion classification accuracy increased from 83% to 85.50%, while speaker distance estimation improved from 77.91% to 79.97%. These results demonstrate that our approach can be effectively integrated with various models across various domains.

6 Conclusion

In this work, we introduce CAARMA, a novel class augmentation framework designed to tackle the challenge of limited class diversity in zero-shot inference tasks. Our approach synthesizes strategic data mixing with an adversarial refinement mechanism to align real and synthetic classes effectively within the embedding space. We validate CAARMA's effectiveness in speaker verification, achieving an 8% improvement over baseline models, and extend its application to emotion classification and speaker distance estimation, where it also shows significant gains. These results underscore CAARMA's capability to enhance embedding structures in various zero-shot inference scenarios. Our framework offers a scalable solution to the class diversity problem, facilitating integration into existing systems without the need for new real-world data collection. With our code released publicly, we anticipate that CAARMA will aid both research and practical applications in zero-shot learning. Future work will focus on expanding CAARMA's utility to larger datasets and other domains, such as computer vision.

Limitations

The CAARMA framework, while showcasing notable enhancements in speaker verification and

zero-shot learning tasks, is subject to several limitations that merit further exploration. Although it performs well in controlled settings, its scalability to extremely large or diverse datasets, as well as its applicability to real-world scenarios with high speaker variability, has yet to be fully established. This also adds complexity to the implementation and increases computational demands, which may restrict accessibility for those with limited resources.

Ethics Statement

The CAARMA framework is developed with a commitment to ethical considerations, especially concerning privacy and the potential for surveillance misuse. It is crucial to ensure that this technology, while advancing the capabilities of speaker verification systems, is employed within the confines of strict ethical guidelines and privacy regulations to prevent any invasion of individual privacy. As this framework facilitates the generation of synthetic data, we also focus on preventing biases that could arise in synthetic datasets, ensuring fair representation across different groups. In adherence to the ACL Ethics Policy, we emphasize transparency in the deployment of CAARMA and advocate for its use in ethically justifiable manners that respect individual rights and data integrity.

References

Jinwoo Baek, Hwanjun Kim, Seunghyeon Jeong, Wonjun Choi, and Chulhee Kim. 2021. Gridmix: A simple grid-based data augmentation method for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3255–3264.

Yang Cai, Zifeng Gao, Bohan Zhou, Kunpeng Li, Qi Huang, and Qi Dai. 2021. C-mixup: Class-aware mixup for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7769–7778.

Ting Chen, Xiaowei Wang, Yunzhi Shen, Weihao Nie, Han Zhang, Binglie Li, and Yu Zhang. 2022. Transmix: Attending to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7367–7377.

J Chung, A Nagrani, and A Zisserman. 2018. Voxceleb2: Deep speaker recognition. *Interspeech 2018*.

Samuele Cornell, Jordan Darefsky, Zhiyao Duan, and Shinji Watanabe. 2024. Generating data with text-to-speech and large-language models for conversational speech recognition. In *Proc. SynData4GenAI 2024*, pages 6–10.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jue Gao, Tri M Dang, Michael L Seltzer, and Rif A Saurous. 2022. Conversasynth: Exploring the landscape of synthetic conversations for audio understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6543–6558.

Nilesh Gupta, Sakina Bohra, Yashoteja Prabhu, Saurabh Purohit, and Manik Varma. 2021. Generalized zero-shot extreme multi-label learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 527–535.

Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2371–2381.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP*, pages 12136–12140. IEEE.

Amit Jindal, Narayanan Elavathur Ranganatha, Aniket Didolkar, Arijit Ghosh Chowdhury, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Speechmix-augmenting deep sound recognition using hidden space interpolations. In *INTERSPEECH*, pages 861–865.

Rémy Joassin and David Alvarez-Melis. 2022. Synthio: Towards unified audio data augmentation via conditional text-to-audio diffusion. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6511–6526.

Youngdong Kim, Serena Seo, Bohyung Cho, and Sung Ju Yun. 2020. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Learning Representations*.

Haekyu Lee, Jaeho Lee, Byeongho Kim, and Sungrae Kim. 2020. Smoothmix: A simple yet effective data augmentation to train robust classifiers. In *International Conference on Learning Representations*.

Yichen Liu, Xiting Song, Qingyao Cao, Weizhi Chen, Zhaowei Wu, Xing Li, and Jieping Huang. 2022a. Rc-mixup: Robust and clean mixup for improving generalization. In *International Conference on Machine Learning*, pages 13525–13538. PMLR.

Zhaoyang Liu, Jingfeng Geng, Zhiyong Zhang, and Guanghui Liu. 2022b. Tokenmix: Re-thinking token mixing in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16012–16021.

Shaobo Min, Hantao Yao, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. 2019. Domain-specific embedding network for zero-shot recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2070–2078.

Shaobo Min, Hantao Yao, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. 2020. Domain-oriented semantic embedding for zero-shot learning. *IEEE Transactions on Multimedia*, 23:3919–3930.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070.

Zhaohui Qin, Dong Wang, Zitong Zhang, and Bin Wu. 2023. Resizemix: A simple data augmentation method for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yusuke Ueda, Tomoki Saito, Kazuhiro Tachibana, Carlos Esteban, and Junichi Yamagishi. 2024. Expressive speech synthesis with style transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Lakshmi Venkataramanan, Aditi Raghunathan, Ananya Kapoor, and Gunjan Joshi. 2022. Alignmix: Improving consistency and robustness in vision transformers via aligned mixup regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10338–10347.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Guo-Sen Xie, Zheng Zhang, Huan Xiong, Ling Shao, and Xuelong Li. 2022. Towards zero-shot learning: A brief review and an attention-based embedding network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1181–1197.

Bing Xu. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. 2022. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification. *Interspeech*.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2019. Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.