Elastic Additive Angular Margin Loss Integrated with Mixup for Anomalous Sound Detection

Yihao Zhao*, Yichen Yang*, Xiao Zhang*, and Shoji Makino*
* Waseda University, Japan
E-mail: yihaozhao@akane.waseda.jp, yang_yichen@mail.nwpu.edu.cn, zhang_x07@toki.waseda.jp, s.makino@waseda.jp

***Abstract*—Unsupervised anomalous sound detection (ASD) aims to detect anomalous sounds by modeling only normal sounds. Conventional methods use classification-based self-supervised neural network models to identify normal and abnormal sounds. Building on this idea, Additive Angular Margin Loss (ArcFace) is utilized to enhance intra-class compactness and inter-class discrepancy of normal classes. Furthermore, previous methods introduce mixup to ensure intra-class compactness and alleviate the overlap of normal and anomalous data distributions. However, the additive angular margin applies only to a single target logit but not to learning from interpolated labels and samples generated by mixup. Additionally, its fixed margin is unsuitable for data with inconsistent intra-class and inter-class variations. To solve these issues, we propose a training method applying an elastic angular margin loss to mixup data augmentation, namely Elastic-Margin-Mixup (EMM). The proposed EMM weights the margin penalty by the interpolation factor, enabling the ArcFace loss to adapt appropriately to interpolated samples. Moreover, it introduces proportional randomization across differently weighted margins to further help the model handle inconsistent variations both between and within classes. On the DCASE 2020 Challenge Task 2 dataset, experimental results demonstrate that our proposed EMM achieved an AUC of 95.03% and a pAUC of 89.95%, surpassing previous approaches integrating ArcFace and mixup for anomalous sound detection.**

I. Introduction

Anomalous sound detection (ASD) is designed to determine whether a sound reflects normal conditions or undetected faults. Because anomalous sound data are insufficient, unsupervised ASD is an important method for achieving this detection task by training solely on normal sound data [1].

Interpolation deep neural network (IDNN) [2] and Group Masked Autoencoder for Density Estimation (Group MADE) [3] utilize generative techniques for unsupervised ASD. These methods use the autoencoder to reconstruct the spectrogram and minimize the reconstruction error to learn the distribution of normal sounds. The reconstruction error is used as an anomaly score to identify normal sounds and anomalous sounds. However, these methods cannot give significantly different anomaly scores if the normal and anomalous sounds have similar features [4], [5].

Self-supervised classification approaches leverage the metadata to divide sounds into distinct classes for feature learning and demonstrate superior performance compared to autoencoder-based methods [6]–[10]. The log-Mel spectrogram is widely used in classification-based ASD, but it may filter out high-frequency information. To compensate for this drawback, STgram-MFN [7] proposes fusing the spectral feature Sgram with a complementary temporal feature Tgram to enhance self-supervised classification. Furthermore, it employs ArcFace [11] as its training loss to improve intra-class compactness and inter-class discrepancy. ArcFace realizes this by adding an additive angular margin between each feature embedding and its corresponding class center. However, while ArcFace tightens intra-class distances for normal samples, it may also cause the ASD model to incorrectly map the feature embeddings of anomalous sounds close to their class centers defined by the metadata, leading to the distribution overlap between normal and anomalous sound samples.

To alleviate this issue, mixup [12] has been applied to improve the separation between normal and anomalous sounds [9], [10]. Mixup linearly interpolates random pairs of input samples and their one-hot labels using a random interpolation factor for each pair, thus generating more diverse training data. However, the margin of ArcFace is added to the target logit, whereas mixup trains a model to predict interpolated labels of interpolated samples. To address this problem, ArcMix [9] weights two ArcFace losses by interpolation factors. The refinement of ArcMix, Noisy-ArcMix [9], applies the margin penalty to the class of only one sample in each sample pair, which improves intra-class compactness and inter-class discrimination. Moreover, another work [10] utilizes the interpolation factor to weight the logits with a margin penalty and without a margin penalty for interpolated samples. While these methods enhance intra-class compactness and inter-class separation, their effects on reducing the distribution overlap of normal and anomalous sounds are not significant.

To overcome the above limitations, we apply an appropriate additive angular margin to the prediction logits of interpolated normal sounds. Inspired by margin-mixup [13], we weight the margin by the interpolation factor to provide a proportional margin penalty to the corresponding logit. This approach integrates the ArcFace loss with interpolated samples and offers direct and appropriate margin penalties to their prediction logits to enhance the discrimination between normal and anomalous sounds. To accommodate inconsistent intra-class and inter-class variations in real-world normal data, we incorporate an elastic margin concept from ElasticFace [14], which uses a random margin drawn from a normal distribution. Building on this concept, we generate each random margin by sampling from a normal distribution centered at the weighted margin, with a standard deviation weighted by the same interpolation factor. This imparts a proportional randomness for each weighted margin, providing the model with flexibility in class separation. We employ STgram-MFN as the backbone model for our proposed Elastic-Margin-Mixup (EMM). The method is evaluated on the DCASE 2020 Challenge Task 2 dataset, where the results demonstrate its effectiveness and superiority over existing ArcFace-based mixup methods.

II. Related Work

A. Spectral-Temporal Feature Fusion for Self-supervised Anomalous Sound Detection

The widely used log-Mel spectrogram in ASD is tuned to the human auditory system's frequency sensitivity but may suppress high-frequency anomalous features [7], [15]. To address this limitation, a temporal feature, Tgram, is extracted from the raw sound wave by a newly designed CNN-based network, TgramNet [7]. In addition, the authors propose a fusion feature STgram that combines the log-Mel spectrogram, namely Sgram, and the Tgram. This fusion feature is input to a lightweight convolutional network, MobileFaceNet (MFN) [16], to learn feature representations of normal sounds. Furthermore, the complete architecture called STgram-MFN applies ArcFace [11] to enhance both intra-class compactness and inter-class discrepancy.

B. ArcFace-based Mixup Methods for Self-supervised Anomalous Sound Detection

Mixup has been applied to increase data diversity in self-supervised anomalous sound detection. However, the interpolated labels of interpolated samples synthesized by mixup do not fit with the ArcFace loss that applies an additive angular margin to the target logit. To overcome this issue, ArcMix (AMix) [9] weights two standard ArcFace losses by the interpolation factor, and Noisy-ArcMix (NAMix) [9] applies an asymmetric margin penalty to the classes of each sample pair. Based on STgram, a feature TASTgram [9] containing temporal-attention information is combined with the above two methods to improve ASD performance. Another approach [10] weights the logits with and without the margin penalty by the interpolation factor during training on interpolated samples. This method also separates the pooling algorithm of MobileFaceNet [16] for the time and frequency axes to better capture the distinct physical meanings in log-Mel spectrograms [17], [18], namely Modified MFN [10]. These ArcFace-based mixup methods tighten intra-class distributions and enhance inter-class separation, but they do not remarkably alleviate the distribution overlap between normal and anomalous sounds. Additionally, their fixed margin penalties for prediction logits are not suitable for interpolation samples.

C. ArcFace-based Mixup Method for Speaker Verification

Different from the above ArcFace-based mixup methods using a fixed additive angular margin for the prediction logits, margin-mixup [13] directly weights the additive angular margin by the interpolation factor for the interpolated samples.

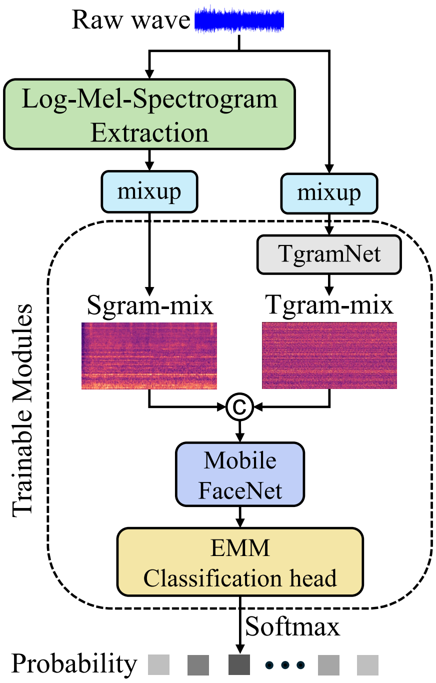


Fig. 1. The training architecture of our proposed Elastic-Margin-Mixup. TgramNet is a CNN-based network proposed in STgram-MFN to extract temporal features. Sgram-mix denotes the interpolated spectrogram generated by mixup, and Tgram-mix denotes the temporal feature extracted from the interpolated raw wave by TgramNet. The two features are concatenated and passed to MobileFaceNet to produce a feature embedding. The classification head transforms the embedding into prediction logits with our proposed EMM training strategy. The softmax function then converts the logits into class probabilities.

During training, each input is an interpolation of two single-speaker waveforms, which enables the embedding space to represent overlapping speakers. Therefore, the margin-mixup training strategy improves the robustness of speaker embeddings for multi-speaker verification.

D. Random Margin for Flexible Class Separation

Beyond the challenge of integrating with mixup, another consideration for ArcFace is its assumption that the geodesic distances within and between different classes can be equally learned using the fixed margin. However, in real-world datasets, intra-class and inter-class variations are inconsistent, making this assumption suboptimal [14]. To alleviate this issue, ElasticFace [14] proposes an enhancement of margin loss for face recognition that samples a random margin from a normal distribution with its mean set to the margin penalty and a fixed standard deviation. This method provides the model with the capacity to learn flexible class separability, improving intra-class compactness and inter-class discrepancy.

III. Proposed Method

In our self-supervised classification method for unsupervised ASD, we propose employing the ArcFace loss [11] during training to enhance intra-class compactness and inter-class discrepancy of normal classes. The ArcFace loss is designed

TABLE I
AUC (%) AND pAUC (%) RESULTS OF DIFFERENT METHODS ON THE TEST DATA OF THE DEVELOPMENT DATASET.

Methods	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
STgram-MFN(ArcFace) [7]	94.04	88.97	91.94	81.75	99.55	97.61	99.64	98.44	94.44	87.68	74.57	63.60	92.36	86.34
TASTgram-MFN(AMix) [9]	96.43	94.52	93.12	85.19	99.08	95.23	99.22	98.83	95.54	87.55	70.89	61.30	92.38	87.10
TASTgram-MFN(NAMix) [9]	**98.32**	**95.34**	95.44	85.99	99.53	97.50	**99.95**	**99.74**	96.76	90.11	77.90	67.15	94.65	89.31
Modified MFN(mixup) [10]	96.75	92.36	94.79	88.00	**99.58**	**97.81**	99.85	99.20	**97.17**	**91.37**	79.88	**68.38**	94.67	89.52
STgram-MFN(EMM)	98.00	94.35	**95.85**	**90.22**	99.52	97.49	99.87	99.30	96.57	90.96	**80.38**	67.38	**95.03**	**89.95**

to incorporate an additive angular margin directly in the angle space, and its loss function is formulated as

$$L_{\text{ArcFace}} = -\log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{k=1,k\neq y_i}^{C} e^{s\cos(\theta_k)}}, \quad (1)$$

where C represents the number of classes. Let $\mathbf{z}_i \in \mathbb{R}^{D\times 1}$ and $\mathbf{w}_k \in \mathbb{R}^{D\times 1}$ represent the feature embedding vector of the sample $\mathbf{x}_i$ and the weight vector of the k^{th} class center, respectively. Here, D is the embedding dimension, and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_C] \in \mathbb{R}^{D\times C}$ is the fully connected layer's projection matrix in the classification head. After $\mathbf{z}_i$ and $\mathbf{w}_k$ are ℓ_2-normalized, the angle between them can be denoted by $\theta_k = \arccos(\mathbf{w}_k^{\mathsf{T}}\mathbf{z}_i)$. A fixed margin m is added only to the true-class angle θ_{y_i}, turning the target logit $\cos(\theta_{y_i})$ to $\cos(\theta_{y_i}+m)$. Every logit is scaled by a scale parameter s to enlarge the difference between logits. By adding the additive angular margin m to the target logit during training on normal sound classes, ArcFace strengthens intra-class compactness and inter-class discrepancy. However, applying this margin penalty may lead the ASD model to incorrectly map the feature embeddings of anomalous samples close to their class centers defined by the metadata, thereby increasing the overlap between the distributions of normal and anomalous sounds.

To achieve a better distinction between normal and anomalous sounds, we introduce mixup [12] to synthesize interpolated random normal samples, which extends the normal data distribution. Given a mini-batch of N uninterpolated sound samples, we randomly draw N sample pairs $(\mathbf{x}_i, \mathbf{x}_j)$ with corresponding one-hot labels $\mathbf{y}_i$ and $\mathbf{y}_j$. The interpolation factor λ is sampled from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$ with $\alpha = 0.5$. For each pair, the linearly interpolated input sample $\mathbf{x}_{ij}$ is computed as

$$\mathbf{x}_{ij} = \lambda\mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \quad (2)$$

and its interpolated one-hot label $\mathbf{y}_{ij}$ is given by

$$\mathbf{y}_{ij} = \lambda\mathbf{y}_i + (1-\lambda)\mathbf{y}_j. \quad (3)$$

However, the additive angular margin of ArcFace is defined for the single target logit of each uninterpolated sample, making it unsuitable for interpolated samples that have interpolated labels. To solve this problem, we follow margin-mixup [13] to directly weight the additive angular margin by λ. Specifically, for an interpolated sample $\mathbf{x}_{ij}$, the margin applied for the logit of class y_i is weighted by λ, and for class y_j by $1-\lambda$. This allows the margin penalty to adjust dynamically according to the interpolation factor. Consequently, the direct appropriate margin penalties ensure adequate intra-class compactness and prompt the ASD model to keep discrimination on subtle differences among normal samples, thereby improving the detection of unseen anomalies and alleviating the distribution overlap between normal and anomalous sounds.

Furthermore, regarding inconsistent variations between and within normal sound classes, the flexible separability of normal classes is important for the ASD model. To address this issue, we introduce the elastic margin in ElasticFace [14], which samples a random margin from a normal distribution. Formally, the probability density function of a normal distribution with mean μ and standard deviation σ can be expressed as

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (4)$$

Considering that our method uses a weighted margin that varies with λ across different interpolated samples, applying a fixed standard deviation for the randomization is suboptimal. We propose that the degree of randomness should correlate with the magnitude of the weighted margin, since larger margins have a more significant effect on class separation, whereas the effect of smaller margins is less pronounced. To meet this requirement, we set the mean of the normal distribution as the weighted margin and weight its standard deviation by the same interpolation factor. Thus, the random margin is drawn from a weighted normal distribution $\mathcal{N}(\lambda m, (\lambda\sigma)^2)$.

Therefore, for each interpolated sample $\mathbf{x}_{ij}$, we add a random margin weighted by λ to the angle between its feature embedding $\mathbf{z}_{ij}$ and the center of class y_i, and add another random margin weighted by $1-\lambda$ to the angle for class y_j. Let $G(\lambda m, \lambda\sigma)$ denotes the weighted random margin sampled from $\mathcal{N}(\lambda m, (\lambda\sigma)^2)$. The modified angle $\widetilde{\theta}_k$ for each class k can be expressed as

$$\widetilde{\theta}_k = \begin{cases} \theta_k + G(\lambda m, \lambda\sigma), & k = y_i, \\ \theta_k + G((1-\lambda)m, (1-\lambda)\sigma), & k = y_j, \\ \theta_k, & \text{otherwise.} \end{cases} \quad (5)$$

These elastic angles $\{\widetilde{\theta}_k\}_{k=1}^{C}$ are subsequently incorporated into the ArcFace loss. Thus, the Elastic-Margin-Mixup loss

TABLE II
MIXUP AND MARGIN RANDOMIZATION ABLATION ON THE TEST DATA OF THE DEVELOPMENT DATASET IN TERMS OF AUC (%) AND PAUC (%).

Methods	mixup	σ	$\lambda\sigma$	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
				AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
ArcFace [11]	✗	0.0	✗	92.57	85.02	95.34	86.79	**99.71**	**98.47**	99.30	96.80	95.35	89.10	76.24	63.50	93.08	86.61
ElasticFace [14]	✗	0.0125	✗	92.75	87.96	94.26	85.55	99.68	98.34	99.62	98.07	96.77	89.66	73.68	60.86	92.79	86.74
	✗	0.0175	✗	94.38	88.27	93.35	84.48	99.67	98.26	99.49	97.52	96.07	87.93	77.98	64.32	93.49	86.80
Margin-mixup [13]	✓	0.0	✗	98.25	93.76	**95.94**	89.57	99.57	97.75	99.87	99.36	**96.80**	**91.64**	78.60	66.29	94.84	89.73
Elastic-Margin-Mixup	✓	0.0125	✗	97.19	93.41	95.37	89.10	99.51	97.44	99.79	98.91	96.79	91.18	79.55	67.62	94.70	89.61
	✓	0.0125	✓	98.00	94.35	95.85	90.22	99.52	97.49	99.87	99.30	96.57	90.96	80.38	67.38	**95.03**	**89.95**
	✓	0.0175	✗	**98.56**	**95.01**	94.73	88.79	99.58	97.77	99.70	98.43	96.57	90.32	80.32	**68.23**	94.91	89.76
	✓	0.0175	✓	98.36	94.25	95.25	**90.35**	99.59	97.87	**99.91**	**99.64**	96.11	90.06	**80.66**	67.03	94.98	89.87

TABLE III
MAUC (%) RESULTS OF DIFFERENT METHODS ON THE TEST DATA OF THE DEVELOPMENT DATASET.

Method	STgram-MFN (ArcFace)	TASTgram-MFN (Noisy-ArcMix)	**STgram-MFN (EMM)**
Fan	81.39	**92.67**	92.19
Pump	83.48	**91.17**	89.07
Slider	**98.22**	97.96	98.09
Valve	98.83	**99.89**	99.53
ToyCar	83.07	**88.81**	88.32
ToyConveyor	64.16	68.18	**70.12**
Average	84.86	**89.78**	89.55

function L for the interpolated sample $\mathbf{x}_{ij}$ is given by

$$L = -\lambda \log \frac{e^{s\cos(\widetilde{\theta}_{y_i})}}{\sum_{k=1}^{C} e^{s\cos(\widetilde{\theta}_k)}} - (1-\lambda)\log \frac{e^{s\cos(\widetilde{\theta}_{y_j})}}{\sum_{k=1}^{C} e^{s\cos(\widetilde{\theta}_k)}}. \tag{6}$$

By weighting both the margin and the standard deviation with the mixup interpolation factor, the proposed method applies an adaptive random margin penalty to each interpolated sample. Consequently, the proposed method ensures intra-class compactness, enables flexible class separability, and reduces the distribution overlap between normal and anomalous sounds. We apply our method to the backbone model STgram-MFN [7] and show the training architecture in Fig. 1. Following Noisy-ArcMix [9], we perform mixup on both log-Mel spectrograms and raw waves, which serve as inputs to the trainable modules. Lastly, we replace the original ArcFace loss with the proposed Elastic-Margin-Mixup loss during training on these interpolated inputs.

IV. EXPERIMENT

A. Dataset

The proposed EMM was evaluated on the DCASE 2020 Challenge Task 2 development dataset together with an additional dataset [1]. Both datasets include parts of audio recordings from the MIMII [19] and ToyADMOS datasets [20]. The MIMII dataset contains four machine types, namely Fan, Pump, Slider, and Valve, with each type comprising seven distinct machines. The ToyADMOS dataset includes two more types, ToyCar and ToyConveyor, with seven and six different machines, respectively. Within each machine type, the individual machines have distinct IDs. We used machine types and machine IDs to divide the datasets into a total of 41 unique machine classes. For training, we employed only the normal samples from both the development dataset and the additional dataset. The evaluation was conducted on the development test dataset, which provides both normal and anomalous sounds.

B. Evaluation Metrics

AUC, partial-AUC (pAUC) [1], and minimum AUC (mAUC) [7] served as the evaluation metrics in the experiments. AUC measures the area under the ROC curve and ranges from 0 to 1. A higher AUC indicates better model performance. pAUC refers to the partial area under the ROC curve where the false positive rate is within the range $[0, p]$. In this experiment, p was set to 0.1, following the setting in [7]. mAUC represents the minimum AUC achieved across individual machine IDs for each machine type. This metric reflects the worst case of the evaluated model within each machine type.

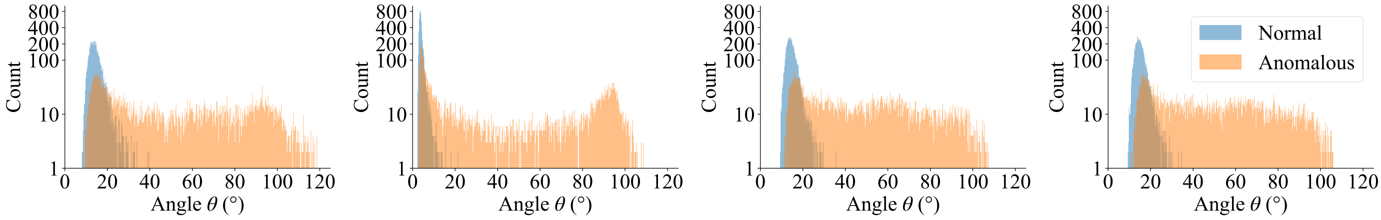
C. Implementation Details

We used STgram-MFN [7] as the backbone model to evaluate our proposed training method EMM. Self-supervised classification was performed on 41 normal classes. The hyperparameter α of the Beta distribution was set to 0.5, following previous works [9], [10]. The proposed margin loss was based on the ArcFace loss [11] and employed an additive angular margin m of 1.1 and a scale factor s of 30. Additionally, the standard deviation σ was set to 0.0125 and 0.0175, following [14]. Training parameters include a batch size of 64 and a learning rate of 0.0001. The model was trained for 200 epochs using the Adam optimizer [21] and the cosine annealing learning rate decay schedule [22].

D. Performance Comparison

The performance of our proposed method, STgram-MFN with EMM, was evaluated against several self-supervised classification approaches for ASD, including STgram-MFN with ArcFace [7], TASTgram-MFN with ArcMix (AMix) and Noisy-ArcMix (NAMix) [9], and Modified MFN with an ArcFace-based mixup method [10].

As shown in Table I, the STgram-MFN model with our proposed EMM achieved the highest average AUC and pAUC



(a) ArcFace $(m = 0.7)$ (b) ArcFace $(m = 1.1)$ (c) Margin-mixup $(m = 1.1)$ (d) Elastic-Margin-Mixup $(m = 1.1)$

Fig. 2. Distribution of the angles θ between each feature embedding and its class center. The plots show the results of the model STgram-MFN trained with (a) ArcFace $(m = 0.7)$, (b) ArcFace $(m = 1.1)$, (c) Margin-mixup $(m = 1.1)$, and (d) Elastic-Margin-Mixup $(m = 1.1)$. The y-axis represents the sample count on a base-10 logarithmic scale. Results aggregate all machine types included in the test data of the DCASE 2020 Challenge Task 2 development dataset.

TABLE IV
AUC (%) AND PAUC (%) RESULTS UNDER DIFFERENT MARGIN SETTINGS FOR THE WEIGHTED MARGIN ON THE TEST DATA OF THE DEVELOPMENT DATASET.

m	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
0.7	95.70	90.34	94.69	88.34	99.52	97.51	99.57	98.13	**97.16**	**91.82**	78.79	65.54	94.24	88.62
0.8	96.27	90.62	94.92	89.40	99.59	97.92	99.72	98.96	96.97	91.06	80.32	66.90	94.63	89.14
0.9	96.70	92.36	95.59	**90.12**	**99.63**	**98.05**	99.78	99.00	96.48	91.70	78.20	66.17	94.40	89.57
1.0	97.97	93.51	94.54	88.50	99.61	97.96	99.55	97.97	96.49	91.49	**81.56**	**68.22**	**94.96**	89.61
1.1	**98.25**	93.76	**95.94**	89.57	99.57	97.75	**99.87**	99.36	96.80	91.64	78.60	66.29	94.84	**89.73**
1.2	97.97	**93.93**	95.28	88.61	99.52	97.46	99.79	**99.38**	96.42	90.54	78.47	66.67	94.57	89.43

among all methods. Table III presents a comparison of mAUC, where the proposed EMM outperformed STgram-MFN with ArcFace, but its mAUC was slightly lower than TASTgram-MFN with Noisy-ArcMix [9].

E. Ablation Study

In Table II, we demonstrated the contributions of mixup, margin randomness, and weighted standard deviation to model performance under the condition that margin m equals 1.1. A value of zero σ indicates an absence of randomness in the margin. A checkmark in the column labeled $\lambda\sigma$ indicates the application of the weighted standard deviation, whereas a cross indicates that the standard deviation is unweighted. The results show that applying a random margin without mixup, as in the ElasticFace [14], yields a performance improvement, suggesting the possible existence of inconsistent intra-class and inter-class variations of normal sound classes. Employing mixup with a weighted margin, as in the margin-mixup [13], produced a substantial gain in average AUC and pAUC. This indicates that using the weighted margin as an integration of mixup and the ArcFace loss enhances the discrimination between normal and anomalous sounds, which makes this method suitable for ASD. Furthermore, employing the weighted standard deviation to generate the random margin based on the weighted margin outperformed both margin-mixup and the approach using the fixed standard deviation. The result means that the proportional randomness applied to the weighted margin helps to better address the inconsistent intra-class and inter-class variations.

Fig. 2 presents the angles between each feature embedding and its corresponding class center. We used a standard deviation of 0.0125 for our proposed EMM in this case. ArcFace $(m = 1.1)$ in Fig. 2b yielded tighter intra-class distributions for normal classes compared with ArcFace $(m = 0.7)$ in Fig. 2a, but it resulted in a higher overlap of normal and anomalous sound distributions. The large margin penalty greatly enhances intra-class compactness of normal samples but reduces the ability to distinguish intra-class diversity in normal classes, thereby decreasing the discrimination of normal and anomalous sounds. Incorporating mixup into ArcFace by weighted margin penalties preserves intra-class tightness and reduces the overlap between normal and anomalous sound distributions, as shown in Fig. 2c and Fig. 2d. Specifically, applying weighted margin penalties to interpolated samples helps the ASD model detect subtle discrepancies among normal samples and better identify unseen anomalies, thereby enhancing the detection of anomalous sounds. Although margin-mixup [13] achieved slightly higher intra-class compactness than our proposed EMM, EMM achieved higher ASD performance, as shown in Table II. This indicates that the proportional randomness realized by weighted standard deviation brings flexible class separation, which helps alleviate the intermingling of normal and anomalous samples. Consequently, EMM achieved a better balance between promoting intra-class compactness and reducing the distribution overlap between normal and anomalous sounds.

To illustrate the effect of the additive angular margin m, we evaluated the performance using the weighted margin without randomness. The results shown in Table IV represent the relationship between the performance and the magnitude of the margin penalty. We found that the model achieved better results with a larger margin than the baseline [7] setting of 0.7. This behavior may stem from the fact that mixup draws samples from a vicinal distribution extending the original training distribution [12], which shifts the optimal margin magnitude. Furthermore, using a weighted margin reduces the margin penalties on interpolated samples and thus may require a larger value for the margin m to achieve a more effective separation of normal and anomalous sounds.

V. Conclusion

In this paper, we propose an Elastic-Margin-Mixup (EMM) method that integrates the additive angular margin loss with mixup. When training on interpolated samples generated by mixup, our proposed EMM applies additive angular margins weighted by interpolation factors to the prediction logits. Furthermore, EMM draws the final random weighted margin from a normal distribution with a mean of this weighted margin and a standard deviation weighted by the same interpolation factor. This proportional randomness adapts the weighted margin to inconsistent variations within and between normal sound classes, enabling flexible class separability. Consequently, EMM ensures intra-class compactness and reduces the distribution overlap between normal and anomalous sounds. To demonstrate the effectiveness of our proposed EMM, it was evaluated on the DCASE 2020 Challenge Task 2 dataset, achieving superior performance compared to previous ArcFace-based mixup methods. In future work, we will investigate how margin and standard deviation values influence results and develop a practical method for selecting their optimal settings.

References

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[2] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE ICASSP*, 2020, pp. 271–275.

[3] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection," in *Proc. DCASE Workshop*, 2020, pp. 51–55.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, Jan. 2019.

[5] J. Guan, Y. Liu, Q. Kong, *et al.*, "Transformer-based autoencoder with id constraint for unsupervised anomalous sound detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 1, p. 42, Oct. 2023.

[6] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. DCASE Workshop*, 2020, pp. 46–50.

[7] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. IEEE ICASSP*, 2022, pp. 816–820.

[8] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[9] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *Proc. IEEE ICASSP*, 2024, pp. 516–520.

[10] J. Choi and J.-W. Choi, "Integrating mixup and arcface for enhanced anomalous sound detection," in *Proc. INTER-NOISE*, 2024, pp. 6778–6785.

[11] 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.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *Proc. ICLR*, 2018.

[13] J. Thienpondt, N. Madhu, and K. Demuynck, "Margin-mixup: A method for robust speaker verification in multi-speaker audio," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[14] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF CVPR*, 2022, pp. 1578–1587.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[16] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Proc. CCBR*, 2018, pp. 428–438.

[17] H. Xing, S. Zhang, D. Takeuchi, D. Niizumi, N. Harada, and S. Makino, "Enhancing spectrogram for audio classification using time-frequency enhancer," in *Proc. APSIPA ASC*, 2023, pp. 1155–1160.

[18] X. Zhang, H. Xing, M. Song, D. Takeuchi, N. Harada, and S. Makino, "Prediction-error-based adaptive specaugment for fine-tuning the masked model on audio classification tasks," in *Proc. APSIPA ASC*, 2024, pp. 1–6.

[19] H. Purohit, R. Tanabe, K. Ichige, *et al.*, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. DCASE Workshop*, 2019, pp. 209–213.

[20] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proc. IEEE WASPAA*, 2019, pp. 313–317.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[22] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017.