MARGIN-MIXUP: A METHOD FOR ROBUST SPEAKER VERIFICATION IN MULTI-SPEAKER AUDIO

Jenthe Thienpondt, Nilesh Madhu, Kris Demuynck

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
jenthe.thienpondt@ugent.be, nilesh.madhu@ugent.be, kris.demuynck@ugent.be

ABSTRACT

This paper is concerned with the task of speaker verification on audio with multiple overlapping speakers. Most speaker verification systems are designed with the assumption of a single speaker being present in a given audio segment. However, in a real-world setting this assumption does not always hold. In this paper, we demonstrate that current speaker verification systems are not robust against audio with noticeable speaker overlap. To alleviate this issue, we propose margin-mixup, a simple training strategy that can easily be adopted by existing speaker verification pipelines to make the resulting speaker embeddings robust against multi-speaker audio. In contrast to other methods, margin-mixup requires no alterations to regular speaker verification architectures, while attaining better results. On our multi-speaker test set based on VoxCeleb1, the proposed margin-mixup strategy improves the EER on average with 44.4% relative to our state-of-the-art speaker verification baseline systems.

***Index Terms*—** speaker verification, multiple speakers, mixup, ECAPA-TDNN, ResNet

1. INTRODUCTION

Speaker verification tries to answer the question if two utterances originate from the same person. Speaker verification systems have recently gained impressive results due to the abundance of labelled training data [1, 2] and the advent of deep learning based speaker recognition architectures [3, 4].

Most speaker recognition pipelines are based on the assumption of a single speaker being present in training and test utterances. Although augmentation techniques such as babble noise [5] are common to provide a certain robustness towards background speakers, we argue that these systems fail to distinguish speakers in a multi-speaker audio scenario.

In prior work done on speaker verification in multi-speaker audio, a few approaches have been suggested. The most straightforward method is the usage of speaker diarization prior to speaker verification [6]. These systems perform well in the case of limited temporal overlap between speaker turns but are not designed to handle mixtures with significant interference between speakers. In [7], target-speaker extraction [8] is applied in a joint training scheme with speaker verification to handle the significant speaker overlap. Another approach is used in [9], with an architecture injecting the target speaker embedding into the hidden frame-level features before the statistics pooling layer. However, these approaches use task-specific architectures or require significant adaptations of existing speaker verification pipelines.

To alleviate this issue, we propose margin-mixup, a training strategy based on mixup [10] to produce speaker embeddings significantly more robust in the multi-speaker scenario. Margin-mixup mixes input waveforms on the batch-level in combination with the usage of an adapted version of the Additive Angular Margin (AAM) [11] softmax loss function, used in current state-of-the-art speaker verification systems [3, 12]. The margin-mixup training strategy can be employed by existing speaker verification models without any computational overhead and requires no architectural alterations.

The paper is organized as follows: Section 2 gives an overview of current state-of-the-art speaker verification systems and our baseline models. Section 3 describes the proposed margin-mixup training strategy. Subsequently, Section 4 explains the experimental setup to verify the proposed margin-mixup strategy while Section 5 analyzes the results. Finally, Section 6 gives some concluding remarks.

2. BASELINE SYSTEMS

We verify the multi-speaker verification performance of three state-of-the-art models and use those as our baseline for the proposed margin-mixup training strategy. Our first baseline system is ECAPA-TDNN [3], which improves upon the x-vector [13] architecture with an adapted version of Squeeze-Excitation (SE) [14], channel- and context dependent statistics pooling and the aggregation of multi-layer features. Secondly, we employ the ECAPA-CNN-TDNN model [4], a hybrid architecture which adds a 2D-convolutional stem to the ECAPA model to enhance its capability to model frequency independent features. Lastly, we consider fwSE-ResNet34 [12], a ResNet based architecture with an adapted frequency-wise SE layer and usage of positional encodings.

This work is supported by the Research Foundation - Flanders (FWO) under grant numbers G081420N and S004923N.

3. MARGIN-MIXUP

Current state-of-the-art speaker verification pipelines attain impressive performance on audio containing a single speaker. While those systems are generally robust against babble-like background noise, we argue they are not able to handle significant interference of multiple speakers. This is mainly due to the emphasis of single-speaker audio in most speaker recognition training datasets and the subsequent absence of a suitable loss function to model speaker overlap. This is further accentuated by the usage of margin based loss functions, such as the commonly used AAM-softmax. Margin based loss functions increase discriminative performance of speaker recognition systems by inducing a margin penalty on a single target speaker during training. However, this makes margin-based loss functions unsuitable for modelling multiple overlapping speakers.

To alleviate this issue, we introduce margin-mixup, a training strategy which enables the speaker embedding space to model overlapping speakers. As we demonstrate, margin-mixup can easily be adopted by regular speaker verification models without any alterations to the speaker embedding extractor architecture, making it a more flexible method compared to task-specific multi-speaker verification approaches.

3.1. Proposed margin-mixup training strategy

Margin-mixup is based on the mixup training strategy introduced in [10], which proposes to sample input features during training from a vicinal distribution by the linear interpolation of input features. With the consequent interpolation of target labels, the model extends the embedding space with a notion of in-between classes.

In this paper, we keep the interpolation limited to two speakers. To adapt the mixup training strategy for speaker recognition systems, an interpolated input waveform $\hat{\mathbf{x}}$ is constructed as following:

$$\hat{\mathbf{x}} = \lambda \frac{\mathbf{x}_a}{\|\mathbf{x}_a\|} + (1-\lambda)\frac{\mathbf{x}_b}{\|\mathbf{x}_b\|} \tag{1}$$

with λ being the interpolation strength between two single-speaker waveforms $\mathbf{x}_a$ and $\mathbf{x}_b$. The input waveforms are energy normalized before the weighted interpolation. Subsequently, the corresponding interpolated target label $\hat{\mathbf{y}}$ is defined as:

$$\hat{\mathbf{y}} = \lambda \mathbf{y}_a + (1-\lambda)\mathbf{y}_b \tag{2}$$

with $\mathbf{y}_a$ and $\mathbf{y}_b$ being one-hot label vectors. We note that the initial intended goal of mixup training is to make systems generalize better on the unmixed task and making them less prone to corrupt labels [10]. However, our main goal during mixup training is to learn an embedding space which can cope with overlapping speakers.

3.2. Margin penalty mixing

The interpolated target label $\hat{\mathbf{y}}$ poses a problem for margin based loss functions as they are designed to impose a margin penalty on a single target label. Since current state-of-the-art speaker verification systems are based on such loss functions, we use an adapted version of AAM-softmax in our proposed margin-mixup training strategy.

AAM-softmax is based on the cosine distance between a speaker embedding $\mathbf{e}_i$ with speaker label i and the corresponding class center $\boldsymbol{W}_i$ with $\boldsymbol{W} \in \mathbb{R}^{D\times N}$. D and N indicate the embedding size and number of training speakers, respectively. In addition, a margin penalty is applied on the angle between the speaker embedding and the target class to enforce a tighter boundary around the speaker classes and subsequently improve speaker verification performance.

During margin-mixup training, the margin penalty m is applied on the angle θ_i between the embedding of the mixed input utterance $\mathbf{e}_{a,b}$ containing both speakers a and b and their corresponding class centers $\boldsymbol{W}_i$ when $i \in [a, b]$ weighted according to the interpolation value λ:

$$\hat{\theta}_i = \begin{cases} \theta_i + \lambda m, & \text{if } i = a \\ \theta_i + (1-\lambda)m, & \text{if } i = b \\ \theta_i, & \text{else} \end{cases} \tag{3}$$

With the applied margin penalty given in Equation 3, the margin-mixup loss function L is subsequently given by:

$$L = \lambda \log \frac{e^{s\,cos(\hat{\theta}_a)}}{\sum_{j=1}^{N} e^{s\,cos(\hat{\theta}_j)}} + (1-\lambda)\log \frac{e^{s\,cos(\hat{\theta}_b)}}{\sum_{j=1}^{N} e^{s\,cos(\hat{\theta}_j)}} \tag{4}$$

with $\hat{\theta}_a$ and $\hat{\theta}_b$ indicating the angle with the applied margin penalty between the multi-speaker input embedding $\mathbf{e}_{a,b}$ and the class centers of a and b, respectively. The hyperparameter s is a scale factor to optimize the gradient flow during training.

We follow the original paper on mixup [10] to determine the interpolation weight λ by sampling the value from a beta distribution with α and β equal to 0.2. This will ensure the interpolation of features is mostly focused on one of the utterances, as we do not want to impact the performance on regular speaker verification with single-speaker test utterances.

4. EXPERIMENTAL SETUP

4.1. Multi-speaker test dataset

To analyze the proposed margin-mixup training strategy for multi-speaker audio settings, we construct a custom test set Vox1-M based on the original VoxCeleb1 [1] (Vox1-O) test set. The overlapping speaker set Vox1-M is constructed by the addition of an interfering utterance from the unused training partition of VoxCeleb1 to the Vox1-O utterances. The corresponding signal-to-noise ratio (SNR) is uniformly sampled

between 0-5dB. If necessary, we repeat the utterance of the interfering speaker to fully overlap the target speaker utterance. The resulting multi-speaker test set is challenging due to the in-domain utterance mixing combined with low SNR levels. All models are trained using the training partition of VoxCeleb2 [2] and will be evaluated by the EER metric.

4.2. Training configuration

Our ECAPA-TDNN baseline model follows the same structure as defined in [3] with a hidden feature dimension of 1024 and an additional SE-Res2Block with kernel size 3 and dilation factor 5 at the end of the frame-level feature extractor. The fwSE-ResNet34 and ECAPA-CNN-TDNN models correspond to the architectures presented in [4]. More details about these models can be found in the accompanying papers.

During training, we take random crops of 2 seconds of audio to prevent overfitting. In addition, we apply one random augmentation based on the MUSAN library [5] (additive noise, music and babble) or the RIR corpus [15] (reverb). Our input features consists of 80-dimensional log Mel-filterbanks with a window length and hop length of 25ms and 10ms, respectively. Subsequently, we apply SpecAugment [16] which randomly masks 0-10 contiguous frequency bins and 0-5 time frames. Finally, we mean normalize the log filterbank energies across the temporal dimension. The margin penalty m and scale factor s of the AAM-softmax loss function are set to 0.2 and 30, respectively.

To optimize the speaker embedding extractors, we use the Adam [17] optimizer with a weight decay of 2e-4. A Cyclical Learning Rate (CLR) [18] strategy is used with the *triangular2* decaying strategy and a minimum and maximum learning rate of 1e-8 and 1e-3, respectively. The cycle length is set to 130K iterations with a batch size of 128.

We apply large margin fine-tuning (LM-FT) [19] after the initial training phase to encourage intra-speaker compactness and increase inter-speaker distances. The margin penalty and crop length are raised to 0.5 and 5s, respectively. All augmentations are disabled during fine-tuning. The CLR maximum learning rate is decreased to 1e-5 with the cycle length lowered to 60K.

After training the speaker embedding extractor, we score the test trials by computing the cosine distance between the enrollment and target speaker embeddings. Subsequently, we apply top-1000 adaptive s-normalization [20, 21] with an imposter cohort consisting of the average speaker embedding of the training utterances in VoxCeleb2.

4.3. Margin-mixup configuration

We apply the proposed margin-mixup training strategy described in Section 3 during both the initial and fine-tuning training phase with the interpolation weight λ sampled from a beta distribution with both α and β set to 0.2.

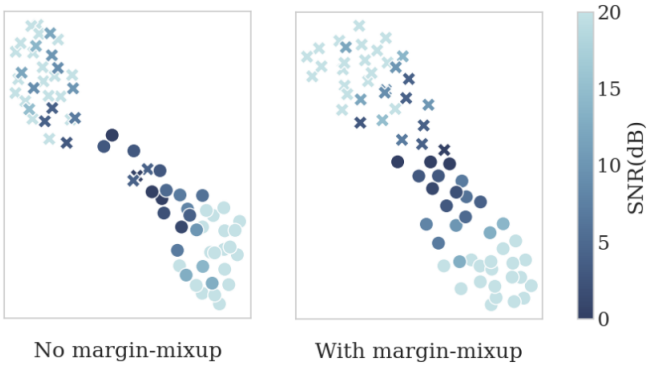

Fig. 1. UMAP-reduced [22] embeddings of utterances from two speakers, indicated by the circle and cross marker, mixed together with various SNR levels from a model trained without (left) and with (right) margin-mixup. Notice that the embedding space in the model trained with margin-mixup has a better notion of in-between speakers.

Systems	M-M	EER(%)	
		Vox1-O	Vox1-M
ECAPA-TDNN	-	0.76	22.31
	✓	**0.74**	**12.42**
ECAPA-CNN-TDNN	-	**0.66**	22.15
	✓	0.68	**12.13**
fwSE-ResNet34	-	**0.63**	21.11
	✓	0.66	**11.85**

Table 1. Performance analysis of the margin-mixup (M-M) strategy on the Vox1-O and multi-speaker Vox1-M test sets.

5. RESULTS

Table 1 gives an overview of the impact of applying the proposed margin-mixup strategy on the baseline models described in Section 2. We notice a significant performance degradation across all baseline models when mixing another speaker in the test trials as done in the Vox1-M test set. This supports our hypothesis that the single-speaker assumption held in most speaker verification setups results in embeddings unable to model significant speaker overlap. However, while margin-mixup has a negligible performance impact on the single-speaker test set Vox1-O, the EER on Vox1-M improves on average with 44.4% relative over the baseline systems. This shows that the proposed margin-mixup training strategy significantly helps to create an embedding space which successfully models in-between speakers, as illustrated in Figure 1, without impacting performance on the single-speaker scenario.

To get a better understanding of the impact of the various components of the margin-mixup training strategy, an ab-

	Method	Vox1-O	Vox1-M
	fwSE-ResNet34	0.66	11.85
A	no mixed margin	0.77	13.15
B	no mixup loss	0.74	17.23
C	only input mixup	0.78	17.65

Table 2. Ablation study on the components of margin-mixup.

lation study is done in Table 2. The baseline system is the fwSE-ResNet34 model trained with margin-mixup. Note that all ablation experiments in this table are still using the input mixup as described in Section 3.1. In experiment *A*, we trained the model without mixing the AAM-softmax margins and applied the margin penalty only to the original speaker by setting $\lambda = 1$ in Equation 3. Without mixing the margins, a performance degradation on both Vox1-O and Vox1-M are observed, indicating the importance of properly mixed margin penalties to fully exploit the AAM-softmax loss function in a multi-speaker setup. When not applying the mixup loss by setting $\lambda = 1$ in Equation 4 and consequently not imposing the model to explicitly detect both speakers, we see a large performance degradation on Vox1-M of 45.4% EER relative. Experiment *C* takes this further by setting $\lambda = 1$ in both Equation 3 and 4, with an additional degradation on both test sets. Notably, the performance on Vox1-M is still an improvement over the fwSE-ResNet34 baseline model without margin-mixup in Table 1. We suspect this is due to the input still being mixed with another speaker and can be regarded as a regular augmentation, bringing the training condition closer to the multi-speaker test setup of Vox1-M.

Systems	Vox1-O	Vox1-M		
		0dB	2dB	0-5dB
baseline	**0.63**	33.67	23.11	21.11
$\alpha, \beta = 0.1$	0.64	19.12	13.12	12.82
$\alpha, \beta = 0.2$	0.66	18.64	12.57	11.85
$\alpha, \beta = 0.4$	0.74	17.32	11.63	11.12
$\alpha, \beta = 0.8$	0.94	15.23	10.62	10.31
$\alpha, \beta = 1$	1.12	**14.84**	**9.44**	**8.93**

Table 3. Analysis of the impact of different α and β parameters in the beta distribution used to sample the interpolation weight λ in the margin-mixup strategy.

Subsequently, we analyze the impact of the α and β parameters of the beta distribution described in Section 3.2 used to sample the interpolation strength λ during margin-mixup training in Table 3. We observe a trade-off between the performance on the standard single-speaker test set Vox1-O and the multi-speaker test set Vox1-M at various SNR levels. In the case of a uniform sampling probability of λ when $\alpha, \beta = 1$, a

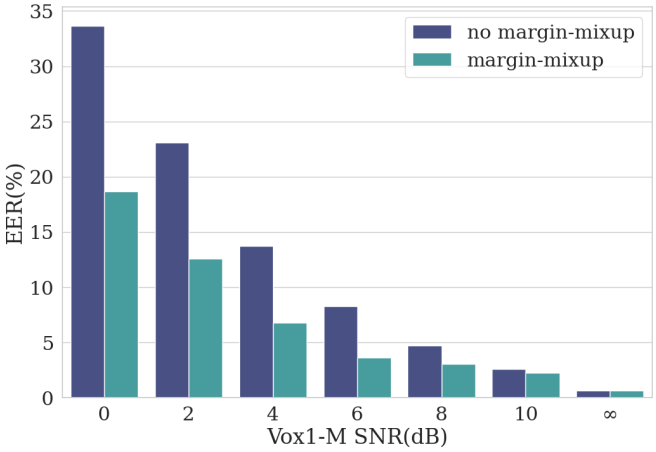

Fig. 2. Bar chart depicting the impact of the margin-mixup training strategy on the fwSE-ResNet34 model evaluated on the multi-speaker Vox1-M test set with different SNR values.

significant performance increase is noted on the *0dB* case with a corresponding performance drop on the regular Vox1-O test set. This indicates that more aggressive mixup conditions during training results in more robust embeddings in severe SNR test scenarios, at the cost of regular speaker verification performance. As seen in the case of $\alpha, \beta = 0.2$, no significant degradation is observed on Vox1-O while still attaining more robust embeddings in the multi-speaker scenario.

Finally, to analyze the impact of the SNR values in overlapping speaker verification, we evaluated multiple versions of the Vox1-M test set with a fixed SNR value using the fwSE-ResNet34 baseline model in Figure 2. As expected, the SNR value has a drastic impact on the speaker verification performance with an inverse correlation between SNR and EER. We note that margin-mixup has the most impact in multi-speaker setups with low SNR values. We suspect this is mainly due to the increasing tendency of the baseline models to see the mixed speaker as background noise in higher SNR scenarios.

6. CONCLUSION

In this paper we presented margin-mixup, a training strategy to make speaker embeddings more robust in a multi-speaker audio setup. In contrast to other approaches, margin-mixup requires no architectural changes to speaker verification pipelines to adapt to the multi-speaker scenario, while attaining significant performance improvements in this challenging condition. Training our baseline models with the proposed margin-mixup strategy improves the EER on average with 44.4% relative over the baseline performance on a multi-speaker version of the original VoxCeleb1 test set. In future work, we aim to extend margin-mixup with a more suitable similarity metric then the cosine distance to increase the multi-speaker modelling capabilities.

7. REFERENCES

[1] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[4] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," in *Proc. Interspeech 2021*, 2021, pp. 2302–2306.

[5] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019*, 2019, pp. 5796–5800.

[7] Wei Rao, Chenglin Xu, Eng Siong Chng, and Haizhou Li, "Target Speaker Extraction for Multi-Talker Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 1273–1277.

[8] Marvin Borsdorf, Chenglin Xu, Haizhou Li, and Tanja Schultz, "Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers," in *Proc. Interspeech 2021*, 2021, pp. 1469–1473.

[9] Ahmad Aloradi, Wolfgang Mack, Mohamed Elminshawi, and E. A. P. Habets, "Speaker verification in multi-speaker environments using temporal feature fusion," in *EUSIPCO 2022*, 2022.

[10] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR 2018, Conference Track Proceedings*, 2018.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4685–4694.

[12] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Tackling the score shift in cross-lingual speaker verification by exploiting language information," in *ICASSP 2022*, 2022, pp. 7187–7191.

[13] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018*, 2018, pp. 5329–5333.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[15] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP 2017*, 2017, pp. 5220–5224.

[16] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[17] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

[18] Leslie N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 464–472.

[19] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *ICASSP 2021*, 2021, pp. 5814–5818.

[20] Pavel Matejka, Ondřej Novotný, Oldřich Plchot, Lukas Burget, Mireia Diez, and Jan Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proc. Interspeech 2017*, 2017, pp. 1567–1571.

[21] Sandro Cumani, Pier Batzu, Daniele Colibro, Claudio Vair, Pietro Laface, and Vasileios Vasilakakis, "Comparison of speaker recognition approaches for real applications.," in *Proc. Interspeech 2011*, 2011, pp. 2365–2368.

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, pp. 861, 2018.