How to merge your embeddings: statistical vs attention-based speaker embedding aggregation for speaker verification with multiple enrollments

Justyna Krzywdziak*, Piotr Masztalski*†, Michal Romaniuk*, Milosz Dudek*†,
Joanna Stepien*†, Mateusz Matuszewski*, Daria Hemmerling*†
*Samsung R&D Institute, Poland
†AGH University of Krakow, Poland
miloszdudek@agh.edu.pl

***Abstract*—In this paper, we evaluate existing state-of-the-art approaches to enrollment utterance aggregation, while proposing alternative methods to improve the effectiveness of speaker verification (SV) systems, when dealing with multiple utterances of the target speaker during the enrollment phase. We investigate multiple facets of the problem, including text-dependent and text-independent scenarios, as well as near-field and far-field speech. Additionally, we assess the impact of several enrollment utterance augmentation methods on aggregation quality. Our research evaluates embedding aggregation approaches, ranging from straightforward techniques such as calculating the average, max and median, to more advanced attention-based models. We propose a modified attention-based architecture that outperforms other techniques by 5 percentage points of Equal Error Rate (EER) in the performance of the verification system. Moreover we suggest a data augmentation method that can improve presented aggregation methods by almost 4 percentage points EER.**

***Index Terms*—speaker verification, enrollment, speaker embedding aggregation, attention, augmentation**

I. Introduction

The swift development of voice assistants confirms the trend towards hands-free human-computer interaction solutions based on speech, applied in smart homes, smartphones, televisions, and other contemporary technologies integrated into our daily lives. Very often, such systems employ speaker verification modules to provide a personalized user experience and advanced security features. In recent years, a large amount of research in SV has been conducted. The current state-of-the-art solutions are almost exclusively based on deep neural networks (DNN) [1]–[4] and outperform the formerly popular i-vector-based systems. In a typical SV system, new users are required to provide a voice sample during the enrollment phase, which serves as a model voiceprint for verification. Most current speaker verification systems require more than one utterance during enrollment, often three or five short utterances. This study focuses specifically on situations where there are five enrollment embeddings of a single speaker. We aim to find an effective way of combining these five utterances to produce the most exemplary representation of the speaker using embeddings generated by ECAPA-TDNN [5] a state-of-the-art speaker verification model. Moreover, we investigate how augmentation of the enrollment data will impact the effectiveness of embedding aggregation methods. We address this in the context of two types of enrollment utterances:

1) text-dependent scenarios, limited to a small set of words or phrases [6],
2) text-independent scenarios, where arbitrary phrases are spoken in each utterance [6].

In addition, we recognize that speaker verification systems should not be limited to near-field scenarios, but must also work effectively at greater distances, such as in a smart home environment. For this reason, we have distinguished two additional data categories:

1) near-field data, recorded by close talking microphone [7],
2) far-field data, the microphone is more than 1m away from the speaker [7].

Contribution: We present a comprehensive evaluation of techniques for merging embeddings in text-dependent and text-independent scenarios, across both near and far-field environments. What is more, we compare various audio augmentation techniques and evaluate the impact of enrollment augmentation methods during train and test time on the effectiveness of the aggregation system. Additionally, we propose a simple modification to a state-of-the-art attention back-end model [8], which is to replace the second multi-attention block with max or average pooling. This improvement not only enhances performance but also simplifies the architecture and decreases model size by 10% and parameter count from 160k to 148k.

II. Related work

The existing research on speaker verification systems most often does not address scenarios involving multiple enrollment utterances [5], [9]–[11]. In [12], the authors only explore the influence of averaged enrolled embeddings on verification system results. They carry out experiments with text-dependent and text-independent datasets, although they focus solely on the multiple enrollment scenario in relation to text-independent data. Averaging enrollment embeddings has been used in SV for more than a decade now even before the emergence of DNN-based systems in the SV field [13], [14]. In [8],

an attention-based back-end is introduced for SV systems with multiple utterances. However, the study solely reports findings for the text-independent scenario, and no details are provided regarding dataset construction for the multiple enrollment scenario. Speaker verification is related to the face verification problem - in both situations, we need to extract a representative feature vector, that will later serve as reference for comparisons. Additionally, the methods for extracting the representative embedding in both domains are very similar [15] [16]. In [17] authors state that the most common aggregation techniques are average pooling and max pooling. In terms of data augmentation, most of the current work focuses on augmenting data to improve SV embedding extraction model as in [18] [19] [20]. This approach does not address the main objectives of our work as we want to concentrate on enhancing embedding aggregation methods. In [21] authors use augmentation to generate extra data to empower speaker enrollment. This research states that by augmenting only enrol data, one can improve verification results by 0.5 Equal Error Rate (EER).

III. Proposed Methods

The primary goal of this work is to find an improved approach for creating a representative speaker embedding from five enrollment utterances. As a baseline, we first evaluated straightforward techniques: average, max, and median pooling. These methods have appeared in studies focusing on multi-utterance enrollment [12], [17]. We then extended our investigation to attention-based approaches for speaker feature aggregation.

A. Statistical Aggregation

We explored three basic techniques to produce a single representative embedding from five enrollment embeddings: averaging, max pooling, and median pooling. Each of the five embeddings has the same dimension, denoted $\mathbb{R}^d$. In all three methods, a new vector is generated by combining the corresponding elements from each of the five enrollment embeddings:

- **Averaging** takes the mean value across the five embeddings at each dimension.
- **Max pooling** selects the highest value at each dimension.
- **Median pooling** selects the median value at each dimension.

B. Attention-Based Method

Attention-based methods have been shown to improve aggregation in multi-utterance enrollment [8]. They leverage scaled dot-product self-attention and feed-forward layers to capture relationships among multiple enrollment embeddings. Figure 1 shows the overall architecture, including our proposed simplification.

First, we concatenate the five enrollment embeddings (each in $\mathbb{R}^d$) from the same speaker into a single matrix E of size $5 \times d$. Next, E is passed through a multi-head scaled dot-product self-attention mechanism [22]. This operation is repeated M times (once per attention head), and the outputs are merged back into a single hidden matrix $H \in \mathbb{R}^{5\times d}$. Residual connections, layer normalization, and a feed-forward sub-layer are included, as described in [22].

Following this initial attention block, the original architecture [8] applies a second multi-head attention layer. In our simplified approach, we replace that second attention module with either average or max pooling across the five rows of H. This pooling produces a single representative embedding $\mathbf{h} \in \mathbb{R}^d$.

Finally, we compute a cosine similarity score between the pooled enrollment embedding and a test embedding $\mathbf{s} \in \mathbb{R}^d$. While [8] reported results only when using both attention blocks, our experiments (see Section VI) show that replacing the second multi-head block with pooling not only streamlines the architecture but also yields superior results in all tested conditions.

IV. Datasets

The availability of datasets tailored explicitly for speaker verification tasks is limited, especially when addressing scenarios that involve the enrollment phase in text-dependent contexts. As a result, the selected datasets had to undergo essential preprocessing steps, the details of which will be elaborated upon in this subsection.

To conduct experiments in text-dependent scenarios, we used the Hi-Mia dataset, as described in [7]. This dataset includes audio recordings of 340 individuals. The specified keyword in this dataset is 'Hi, Mia' in English, and 'ni hao, mi ya' in Chinese. In dataset preparation, any utterance exceeding a 2-second duration was excluded, and the remaining ones were extended with silence to ensure a consistent length of 2 seconds. The recordings, following the dataset authors' guidelines, were divided into near-field and far-field segments. Dataset statistics are shown in Table I

TABLE I
Hi Mia dataset post-processing statistics.

	near-field	far-field
Number of training speakers	254	254
Number of test speakers	44	44
Number of train utterances	19,586	400,000
Number of test utterances	3,263	159,652

The near-field audio recordings were captured using a close-talking microphone positioned 25 cm from the speaker. In contrast, the far-field utterances were captured using six 16-channel circular microphone arrays, positioned around the individual at distances of 1m, 3m, and 5m. Although the signal is multi-channeled we extracted and used mono-channel audio. Distinct enrollment sets were established for both the near and far-field test sets, using test audio files. For each speaker in the test set, five randomly selected files were chosen to create their respective enrollment sets.

The VoxCeleb dataset [23] is a collection of audiovisual content, carefully curated from interview videos available on

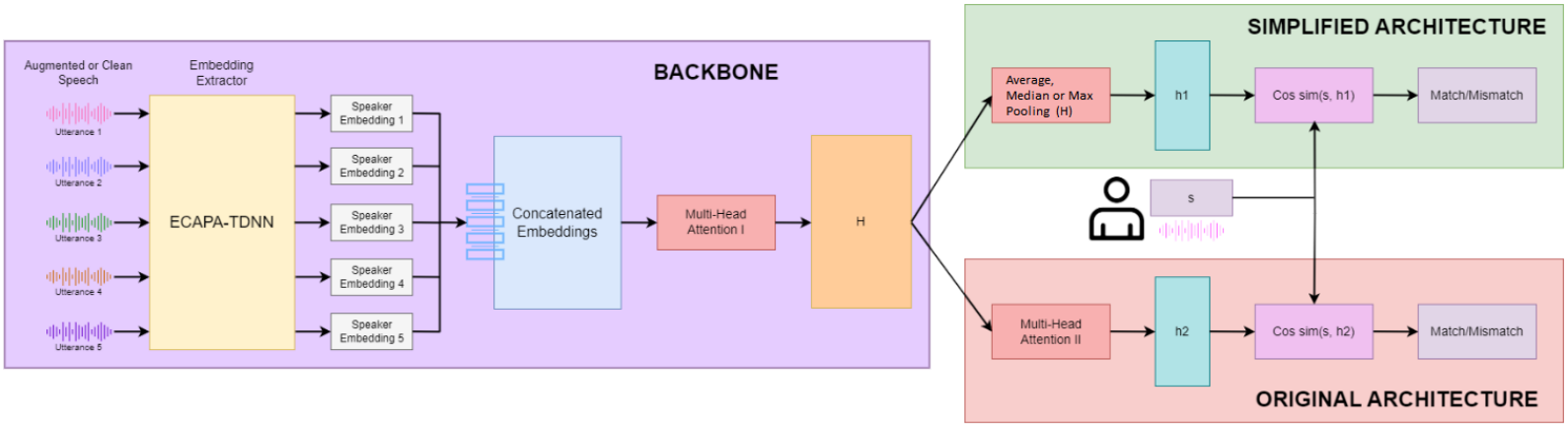

Fig. 1. Attention-based architecture. H is the hidden matrix output of the first multi-head attention layer, h_1 is derived from an average or max pooling operation on H, h_2 is obtained from a second multi-head attention layer, and $\mathbf{s}$ is a test speaker embedding.

the YouTube platform. It is gender balanced and covers a wide range of backgrounds, including different ethnicities, regional accents, professional roles, and age groups. For the speaker verification task, the dataset underwent a series of preparatory steps. First, all audio files were split into 2-second segments, with any remaining fragments from the last split that fell below the 2-second mark excluded. The enrollment set was then constructed by selecting five different utterances from one of the interviews for each speaker test data. Any excess audio files from the interview chosen for enrollment were intentionally omitted to avoid their inadvertent inclusion in subsequent training and testing phases. Dataset statistics are shown in Table II.

TABLE II
VOXCELEB 1&2 DATASET POST-PROCESSING STATISTICS.

	VoxCeleb 1&2
Number of training speakers	7,205
Number of test speakers	40
Number of train utterances	4,299,065
Number of test utterances	17,097

A. Data augmentation

Inspired by [21] we decided to apply augmentation techniques directly on enrollment and test samples to enrich the speaker representation for more robust aggregation. For simple aggregation techniques only enrollment and test samples were augmented and for attention-based methods we also augmented the aggregation module training data. The augmentation methods we focused on were Gaussian Noise, Air Absorption (lowpass-like filterbank with variable octave), Gain (multiply the audio by a random amplitude factor to reduce or increase the volume), Bitcrush (reducing the signal to a given bit depth), Low Pass Filter, High Pass Filter, and Band Pass Filter. All augmentation techniques were implemented using python libraries Pedalboard [24] and Audiomentations [25]. To perform more experiments on data augmentation we decided to use a 5% subset of the Voxceleb training dataset therefore there is a difference in the EER value between Voxceleb results. This subset was used only for augmentation experiments, other experiments are carried out using the whole dataset.

V. EXPERIMENTAL SETUP

Each end-to-end speaker verification system consists of two primary components: a model responsible for extracting speaker embeddings from audio files, and a scoring mechanism that calculates the similarity between two vectors - the target speaker embedding and the test embedding. As the speaker encoder, we employed the pre-trained ECAPA-TDNN model [1] from the Speechbrain library [26]. To ensure more robust results in text-dependent scenarios, we fine-tuned ECAPA-TDNN, pre-trained on the full Voxceleb dataset, on both the near-field and far-field subsets of the Hi Mia dataset with cosine similarity was used for scoring. Our evaluation metrics included the widely accepted Equal Error Rate (EER) and the Minimum Detection Cost Function (minDCF), both measuring the effectiveness of the speaker verification systems. We report minDCF at priors of 0.01 and 0.001, reflecting scenarios in which the genuine speaker is assumed to appear with a 1% or 0.1% chance, respectively. The attention back-end architecture was trained separately for each scenario using the repository[1]. For the aggregation module we employ the same contrastive training strategy as in [8]. Each training was conducted using stochastic gradient descent with momentum, with a learning rate set to 0.01, binary cross-entropy loss, and employing two attention heads. Text-independent scenario experiments were conducted using the VoxCeleb dataset and the best results were obtained after 10 epochs, with a batch size of 128 speakers and 5 enrollment utterances per speaker. We examined the text-dependent scenario using the Hi Mia dataset, considering both near and far-field audio types. Training sessions with the best results were conducted for 20 epochs, with a batch size of 20 speakers and 5 enrollment utterances per speaker. In the

[1]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

evaluation process, we used a predefined list of enrollment and test embeddings pairs, with negative pairs denoted by 0 and positive denoted by 1, depending on whether they originated from a different or the same speaker.

VI. Results

To validate our findings, we conducted a series of experiments. The most promising results are summarised in the Tables III, V, and VII. Results for the augmentation experiments are in the Tables IV, VI and VIII. Table III illustrates the performance of each method on the Hi Mia near-field subset. Notably, using attention-based techniques substantially enhances EER, with a reduction of almost 4 p.p. (percentage points) when compared to the simple average. Furthermore, slight modifications to the attention structure led to a further decrease in EER by 0.5 p.p relative to the attention technique presented in [8]. Table IV demonstrates how each augmentation method impacts the EER for all aggregation methods on the Hi Mia near-field subset. The best results are obtained using Gaussian noise augmentation combined with attention with max pooling aggregation outperforms the baseline average by more than 5 p.p. Multiple augmentation did not improve results.

TABLE III
Results on near-field Hi Mia subset.

Method	EER (%)	minDCF at 0.01	minDCF at 0.001
Average	9.70	0.51	0.60
Median	9.83	0.52	0.61
Max	10.77	0.70	0.79
Attention I & II	6.55	0.52	0.75
Attention + avg	**5.97**	**0.46**	**0.67**
Attention + max	6.57	0.52	0.74

TABLE IV
Augmentation EER (%) results on near-field Hi Mia subset. Column headers: AIR = Air Absorption, GAIN = Gain, GAUS = Gaussian Noise, HP = High Pass Filter, LP = Low Pass Filter, BP = Band Pass Filter.

Method	AIR	BIT	GAIN	GAUS	HP	LP	BP
Average	10.40	9.88	9.74	5.95	9.78	11.30	11.04
Median	10.50	9.88	9.87	6.24	10.04	11.53	11.32
Max	11.48	10.74	10.67	6.62	10.74	12.51	12.26
Attention I & II	5.86	6.01	5.97	4.41	6.56	5.50	6.55
Attention + avg	5.76	5.89	6.14	4.35	6.62	5.28	5.91
Attention + max	6.20	5.91	6.01	**4.01**	6.67	5.72	5.89

Similarly, the subset of the far-field Hi Mia produced results consistent with the near-field environment, highlighting the robustness of the attention-based approach for both near and far-field scenarios, as shown in Table V. The improvement achieved in this scenario is almost 5 p.p. in EER values when compared to the average.

Results for the augmented Hi Mia far-field subset are not as optimistic as in the near-field scenario. In far field scenario speech is quieter than in the near-field scenario and already noised by music/TV, so even subtle augmentation can drown out speech. No single best augmentation can be chosen. Although for simple aggregation methods Gaussian Noise is noticeably reducing the EER, for attention-based methods it achieves similar results to other techniques. The changes we carried out resulted in enhancements also on the VoxCeleb dataset. The attention structures are less effective in this case, with an EER increase of more than 0.1 p.p. compared to the average and median techniques as shown in Table VII. The best method in this case is the attention with average pooling. Table VIII shows the results obtained by augmenting enrollment and test data while evaluating the VoxCeleb dataset. The difference between augmented and non-augmented data is negligible. We attribute this to the fact that the VoxCeleb training dataset was already heavily augmented.

TABLE V
Results on far-field Hi Mia subset.

Method	EER (%)	minDCF at 0.01	minDCF at 0.001
Average	11.83	0.85	0.95
Median	11.96	0.86	0.96
Max	13.58	0.95	0.99
Attention I & II	7.03	**0.66**	0.85
Attention + avg	6.95	**0.66**	0.86
Attention + max	**6.84**	0.67	**0.85**

TABLE VI
Augmentation EER (%) results on far-field Hi Mia subset.

Method	AIR	BIT	GAIN	GAUS	HP	LP	BP
Average	11.35	11.36	11.21	8.69	9.23	12.13	10.22
Median	11.33	11.53	11.35	9.05	9.58	12.14	10.35
Max	13.09	13.26	13.21	11.47	12.14	13.96	12.77
Attention I & II	6.76	6.99	6.97	6.74	**6.50**	6.91	6.59
Attention + avg	6.64	6.91	6.96	6.72	6.59	6.98	6.57
Attention + max	6.62	7.03	6.97	6.59	6.59	6.97	6.52

TABLE VII
Results on VoxCeleb dataset.

Method	EER (%)	minDCF at 0.01	minDCF at 0.001
Average	1.71	**0.21**	0.39
Median	1.83	0.23	0.40
Max	4.62	0.41	0.58
Attention I & II	1.62	0.23	0.44
Attention + avg	**1.61**	0.22	**0.40**
Attention + max	1.66	0.23	0.43

TABLE VIII
Augmentation EER (%) results on VoxCeleb subset. CLEAN column: subset without augmentation.

Method	CLEAN	AIR	BIT	GAIN	GAUS	HP	LP	BP
Average	1.71	1.70	**1.62**	1.75	1.91	1.91	1.94	2.11
Median	1.83	1.78	1.79	1.82	2.00	2.01	2.05	2.05
Max	4.62	4.53	4.62	4.68	5.90	5.24	5.06	5.19
Attention I & II	3.44	3.45	3.36	3.46	3.37	3.68	3.57	3.86
Attention + avg	3.46	3.46	3.35	3.52	3.42	3.70	3.58	3.94
Attention + max	3.45	3.46	3.35	3.50	3.37	3.64	3.56	3.90

Single attention layer combined with mean or average

pooling achieves the best results in all scenarios because as we suspect double attention is limiting the significant voice characteristic information from each speaker and adding additional complexity, whereas single attention allowed to keep more significant voice features, which then were average or max pooled. The lowest EER was obtained in the VoxCeleb experiments, as expected due to the difference in the amount of data between the VoxCeleb and Hi-Mia datasets. This difference affected not only the training of attention-based aggregation techniques but also the ability of ECAPA-TDNN to extract representative embeddings from audio data. Although we fine-tuned ECAPA with the Hi-Mia dataset, there was not enough data to fine-tune it as well as it works for the VoxCeleb dataset. We hypothesize the reason for the poor results of augmentation on the Voxceleb and Hi-Mia far-field subset is that these data are already noisy. VoxCeleb audio is extracted from interviews - a real-life scenario, and Hi Mia far-field was collected in a smart home environment - with TV noise or music. The Hi Mia near-field subset was collected with a high-quality close-talking microphone, so the recorded sound was free of noise and distortion. We believe this is why the enhancement techniques performed best in this scenario.

VII. Conclusions

This study addresses the rarely explored scenario of multi-utterance enrollment in speaker verification systems, typical of mobile or smart home devices. We conducted a comprehensive investigation of text-dependent and text-independent settings under both near-field and far-field conditions, and proposed a simplified attention-based architecture. This model outperforms prior aggregation methods by up to 5 p.p. EER (relative to [12]) while reducing the complexity of the network, making it more suitable for resource-limited devices. Furthermore, we show that applying augmentation to both enrollment and test samples allows statistical and attention-based approaches to achieve competitive results. As speech-based technologies become increasingly commonplace, these findings underscore the importance of efficient and accurate speaker verification solutions.

References

[1] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[2] I. Shahin, A. B. Nassif, N. Nemmour, A. Elnagar, A. Alhudhaif, and K. Polat, "Novel hybrid dnn approaches for speaker verification in emotional and stressful talking environments," *Neural Computing and Applications*, vol. 33, no. 23, pp. 16 033–16 055, 2021.

[3] N. Li, M.-W. Mak, and J.-T. Chien, "Dnn-driven mixture of plda for robust speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1371–1383, 2017.

[4] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[6] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, 2022.

[7] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7613.

[8] C. Zeng, X. Wang, E. Cooper, X. Miao, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6717–6721.

[9] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[10] F. Tong, M. Zhao, J. Zhou, H. Lu, Z. Li, L. Li, and Q. Hong, "Asv-subtools: Open source toolkit for automatic speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6184–6188.

[11] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[12] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[13] G. Liu, T. Hasan, H. Boril, and J. H. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7755–7759.

[14] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. Hansen, "Crss systems for 2012 nist speaker recognition evaluation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6783–6787.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[16] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[17] Z. Liu, H. Hu, J. Bai, S. Li, and S. Lian, "Feature aggregation network for video face recognition," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[18] W. Lin and M.-W. Mak, "Robust speaker verification using population-based data augmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7642–7646.

[19] F. Tong, M. Zhao, J. Zhou, H. Lu, Z. Li, L. Li, and Q. Hong, "Asv-subtools: Open source toolkit for automatic speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6184–6188.

[20] M. Dua, S. Joshi, and S. Dua, "Data augmentation based novel approach to automatic speaker verification system," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 6, p. 100346, 2023.

[21] A. K. Sarkar, H. Sarma, P. Dwivedi, and Z.-H. Tan, "Data augmentation enhanced speaker enrollment for text-dependent speaker verification," in *2020 3rd International Conference on Energy, Power and Environment: Towards Clean Energy Technologies*, 2021, pp. 1–6.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[23] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[24] P. Sobot, "Pedalboard," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.7817838

[25] I. Jordal, "'iver56/audiomentations: v0.34.1'," Nov. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10202830

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.