Interspeech 2024
1-5 September 2024, Kos, Greece

Getting More for Less: Using Weak Labels and AV-Mixup for Robust Audio-Visual Speaker Verification

Anith Selvakumar, Homa Fashandi

LG Electronics, Toronto AI Lab, Canada
a.selvakumarasingam@lge.com, homa.fashandi@lge.com

Abstract

Distance Metric Learning (DML) has typically dominated the audio-visual speaker verification problem space, owing to strong performance in new and unseen classes. In our work, we explored multitask learning techniques to further enhance DML, and show that an auxiliary task with even weak labels can increase the quality of the learned speaker representation without increasing model complexity during inference. We also extend the Generalized End-to-End Loss (GE2E) to multimodal inputs and demonstrate that it can achieve competitive performance in an audio-visual space. Finally, we introduce AV-Mixup, a multimodal augmentation technique during training time that has shown to reduce speaker overfit. Our network achieves state of the art performance for speaker verification, reporting **0.244%**, **0.252%**, **0.441%** Equal Error Rate (EER) on the VoxCeleb1-O/E/H test sets, which is to our knowledge, the best published results on VoxCeleb1-E and VoxCeleb1-H.

Index Terms: speaker verification, multimodal, audio-visual, person representation, multi-task learning

1. Introduction

Human interactivity with artificially intelligent systems continue to gain popularity, especially as devices become capable of advanced tasks and are seamlessly integrated with our daily lives (e.g., digital assistants). A critical component to enabling personalized interactions with such systems is speaker verification, the process of identifying whether a speaker matches with a pre-enrolled speaker profile. Applications can range from user authentication to personalized experiences, however, environmental noise and visual occlusion are only a few of the challenges associated with performing reliable speaker verification in real-world settings [1]. To this end, multimodal systems have been increasing in popularity due to the potential of added robustness and improved performance [2]. For the case of speaker verification, leveraging both audio and visual modalities, in particular, have shown improvement in false-reject rates when compared to audio-only and visual-only systems [3] [4] [5] [6] [7].

This work explores to improve upon the existing benchmarks in audio-visual speaker verification without expending on model complexity, by introducing data-efficient training and augmentation strategies. Our main contributions are as follows:

1. We demonstrate that multi-task learning with inexpensively-obtained, weak labels can be used to enhance the representations learned by DML.
2. We extend the Generalized End to End Loss (GE2E) [8] from a unimodal to a multimodal input space, and for the first time, validate its efficacy beyond an audio-only task.
3. We introduce AV-Mixup, a multimodal augmentation strategy to reduce speaker overfit and improve generalization.

The collection of these contributions yield SOTA performance with EER of 0.244%, 0.252%, and 0.441% on the VoxCeleb1-O/E/H test splits.

2. Background

2.1. Related Works

Prior studies reveal interesting developments in audio-visual speaker verification. However, a majority rely on traditional DML training approaches, and in general, devise complex networks or rely on expensive data collection to realize modest performance gains. For example, Qian et al. [4] introduced a joint learned embedding-level network architecture, trained with their contrastive loss sampling and data augmentation strategy originally presented in [7]. Sun et al. [9] implemented joint-attention pooling on the audio-visual inputs that enhance the weights of impactful time frames. Tao et al. [10] cited noisy labels in large-scale datasets as a significant limitation, and proposed a two-step multimodal deep cleansing network to identify and remove noisy training samples. Finally, Lin et al. [11] introduced a large-scale dataset that showed improved performance when used as a supplementary training set, with best results on their ResNet with frequency-wise Squeeze-Excitation model (denoted M3).

Our work differs by hypothesizing that DML can be enhanced by introducing even a weakly supervised multi-task component to the objective function. This is on the basis of [12], where it was shown that the features learned through classification and contrastive approaches can differ. Further, rather than extensive data collection or dataset cleansing, we hypothesize that noisy labels can be beneficial to achieve more robust speaker representations, by extending training methods and proposing novel augmentation techniques that collectively serve as regularization during training for open-set tasks.

2.2. Multimodal Fusion

Multimodal fusion has been achieved through many different techniques [1] [13] [14]. For robustness in non-ideal scenarios in the audio-visual domain, the attention-based fusion network (AFN) proposed in [1] is of particular interest. AFN learns to adapt to corrupt or missing modalities by re-weighing the contribution of either modality at time of fusion. This is through an attention mechanism that extended across the modality axis to obtain modality attention weights:

$$\hat{a}_{\{a,v\}} = f_{att}([\mathbf{e}_a, \mathbf{e}_v]) = \mathbf{W}^T[\mathbf{e}_a, \mathbf{e}_v] + \mathbf{b}, \tag{1}$$

where e_a and e_v are transformed audio and visual represen-

10.21437/Interspeech.2024-531

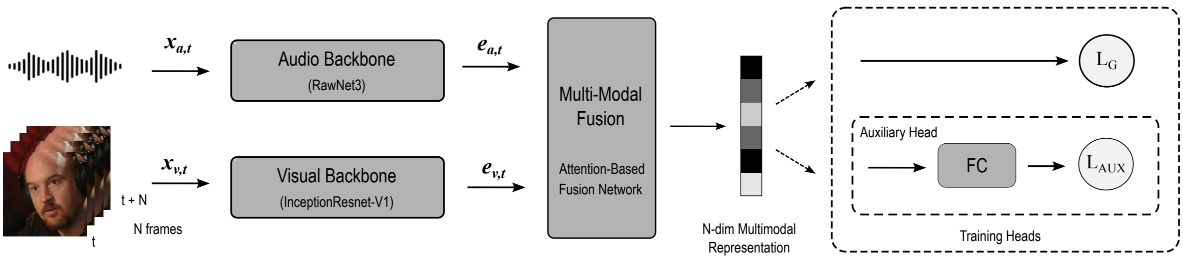


Figure 1: *System level diagram of the REPTAR network. The multimodal representation feeds into the training heads, which are removed during inference time. For inference, the multimodal representation can be used directly for speaker verification*

tations, respectively. Learnable parameters $\mathbf{W}^T$ and $\mathbf{b}$ are optimized during the training process. Modality attention weights are then re-scaled via Softmax to obtain scores between [0, 1] and applied across the embedding axis prior to aggregation to form the fused multimodal representation.

2.3. Generalized End-to-End (GE2E) Loss

The GE2E loss, originally proposed in [8] for audio-based speaker verification, adopts an approach that uses class centroid distances during optimization. Specifically, the loss is calculated from a similarity matrix between each utterance representation embedding to all speaker utterance centroids. From this matrix, a total contrastive loss is calculated based upon positive components and a hard negative component.

3. Robust Audio-Visual Person Encoder

3.1. Generalized End-to End Multimodal Loss

Large scale datasets often contain noisy labels that can confuse networks during training and limit performance. We, however, propose to leverage these noisy samples to improve generalizability. We hypothesize that using a centroid-based optimization approach, outliers and noisy labels will act as a regularizer during training to lead to better generalization. We achieve this through extending the GE2E loss to a multimodal input space, and refer to this new loss function as GE2E-MM.

The GE2E-MM architecture relies on batching $N \times M$ audio and visual inputs, $x_{\{a,v\},ji}$ $(1 \leq j \leq N, 1 \leq i \leq M)$, where N and M are unique speakers and speaker utterances, respectively.

We define the audio-visual latent representation as:

$$\mathbf{e}_{ji} = \frac{f(\mathbf{x}_{a,ji}; \mathbf{x}_{v,ji}; \mathbf{w})}{||f(\mathbf{x}_{a,ji}; \mathbf{x}_{v,ji}; \mathbf{w})||_2} \tag{2}$$

where $f(\mathbf{x}_{a,ji}; \mathbf{x}_{v,ji}; \mathbf{w})$ represents the transfer function of the neural network, with $\mathbf{x}_{a,ji}$ and $\mathbf{x}_{v,ji}$ representing raw audio and visual inputs; and $\mathbf{w}$ representing the network weights. Using this, a similarity matrix, $S_{ji,k}$, of scaled cosine similarities is computed, representing a similarity metric between each multimodal embedding e_{ji} and each speaker centroid, c_k, from the $N \times M$ batch:

$$\mathbf{c}_k = \frac{1}{M} \sum_{m=1}^{M} \mathbf{e}_{km} \tag{3}$$

$$\mathbf{S}_{ji,k} = w \cdot cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b \tag{4}$$

where w, b are learnable parameters. Using this similarity matrix, a contrastive loss is calculated for each multimodal representation, e_{ji}, focusing primarily on all positive pairs and a hard negative pair. The GE2E-MM loss, $\mathcal{L}_G$, is then defined as:

$$\mathcal{L}_G(\mathbf{S}) = \sum_{j,i} \mathcal{L}(\mathbf{e}_{ji}) \tag{5}$$

where,

$$\mathcal{L}(e_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \leq k \leq N \\ k \neq j}} \sigma(\mathbf{S}_{ji,k}) \tag{6}$$

where σ represents the sigmoid function. Optimization of Equation 6 has the effect of pushing embeddings from identical speakers towards its respective centroid, and away from its closest dissimilar speaker centroid.

3.2. Multi-Task Objective Function

We hypothesize that adding an age classification task will force more subtle characteristics to be extracted from the inputs and embedded in the multimodal representation. With this auxiliary task, we can define a multi-task loss function:

$$\mathcal{L}_{MTL} = \gamma \cdot \mathcal{L}_G(S) + (1 - \gamma) \cdot \mathcal{L}_{AUX}, \tag{7}$$

where γ is a scalar weight that is applied in order to balance the losses and prevent one task from dominating. The parameter is obtained through hyper-parameter tuning. $\mathcal{L}_G$ and $\mathcal{L}_{AUX}$ are GE2E-MM and auxiliary task losses, respectively.

3.3. AV-Mixup for Multimodal Augmentation

Audio and visual samples sourced from the same utterance can be highly correlated on speaker-irrelevant features [15] [16]. This can limit the learning of distinctive features and instead cause the training to focus on peripheral attributes such as noise or environmental factors. To prevent this, we propose AV-Mixup, an augmentation technique whereby unique audio-visual pairs are recreated using audio and visual samples that are extracted from disjoint utterances from the same speaker.

Using the AFN for audio-visual fusion, we implemented the above-mentioned developments to form the end-to-end encoder network: '**R**obust **E**ncoder for **P**ersons through Learned Multi-**TA**sk **R**epresentations' (REPTAR), shown in Figure 1.

4. Experimentation

4.1. Setup

4.1.1. Dataset and Preprocessing

VoxCeleb is a large-scale audio-visual dataset for speaker recognition and is widely used in literature. For our experimentation, VoxCeleb2 [17] was used for training while pre-defined test splits from VoxCeleb1 [18] were used for evaluation.

During pre-processing, utterance video files were decomposed into face tracks at a 1 frame per second (FPS) rate and re-scaled to 160 × 160 pixels, and a voice track cropped to a random window size between 4-8 seconds. For evaluation, utterances were sourced from the predefined test splits. Voice clips were limited to a fixed 4 second window with a random onset time and faces were extracted from the same utterance.

4.1.2. Voice and Face Representation

Pre-trained encoders were used to obtain voice and face representations. For voice profiles, RawNet3, proposed in [19] was used. RawNet3 was trained on the VoxCeleb2 dataset and evaluated on the VoxCeleb1-O test split and demonstrates competitive EER performance. For face representation, the InceptionResnet-V1 architecture [20] was used, which was pre-trained on the VGGFace2 [21] dataset. VGGFace2 was confirmed to have a negligible overlap between the test set (5 out of 1251 speakers), which was verified to play no impact to the model performance when rounded to three significant figures.

4.1.3. Multimodal Fusion

For our implementation of the AFN, the 512 dimension face embedding and 256 dimension voice embedding obtained from their pre-trained models are L2 normalized and transformed independently into a 512 dimension space for equal representation prior to fusion. This transformation consists of two linear layers, with a ReLU activation and batch normalization layer in between. The attention layer is implemented as a linear layer with input size 1024 and output of 2 to represent the modality scores. These softmaxed scores are applied as multiplicative factors to the transformed representations, which are then concatenated to form the multimodal representation.

4.1.4. Weakly-Supervised Auxiliary Task

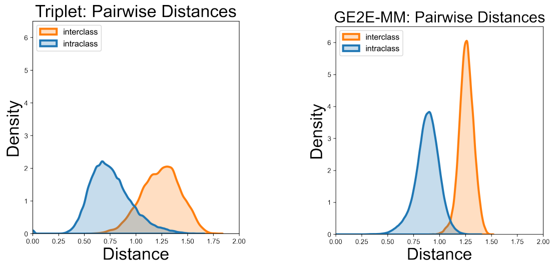
An age prediction auxiliary task was implemented to enhance feature learning during training. The task head consisted of two linear layers with a ReLU activation and batch normalization layer in between. A sigmoid layer was used at the end of the network to represent normalized age predictions. Mean-squared error loss was used as the objective function. The optimal γ value in the compound loss function of Equation 7 was determined to be 0.015 through hyperparameter tuning.

Age labels were obtained from the AgeVoxCeleb dataset [22], which contain estimated ages for approximately 5000 of 6112 speakers of the VoxCeleb2 dataset. These labels can be considered weak due to label inaccuracies and incompleteness.

4.1.5. Training and Evaluation

Our proposed REPTAR model was trained on a Tesla V100 GPU. Batch size was set to 64 with 10 utterances per speaker. The Adam optimizer [23] was used with an exponentially decaying learning rate, initially set to 0.05 and decaying at a factor of 0.9 per epoch. Early stopping with a patience of 5 epochs was used to prevent overfitting. Multiple seeds were tested to demonstrate reproducibility of results.

The Equal Error Rate (EER) metric was calculated for each of the VoxCeleb1 test splits to evaluate the performance of the speaker verification system. EER corresponds to the error rate at which the False Positive Rate (FPR) and False Negative Rate (FNR) is equal, and is a standard metric used for speaker verification [18].



(a) *Triplet Loss* (b) *GE2E-MM Loss*

Figure 2: *Interclass and Intraclass Pairwise Distance Distribution of Triplet and GE2E-MM models for a random speaker.*

Table 1: *Quality of encoded person representation clusters against triplet and GE2E-MM loss (with, without auxiliary task)*

Clustering Metric	$\mathcal{L}_{triplet}$	$\mathcal{L}_{GE2E-MM}$	
		$NoAux$	Aux
Silhouette Score [24] (↑)	0.233	0.501	0.504
Calinski-Harabasz [25] (↑)	4541	6264	6454
Davies-Bouldin [26] (↓)	1.4657	0.8613	0.8599

4.2. Results and Ablation Studies

We performed extensive experiments to study the effect of each of the proposed strategies. This section describes the results obtained along with a comparison to other relevant works.

4.2.1. Effect of the GE2E-MM Loss:

The impact of the GE2E-MM loss was analyzed by studying the encoded audio-visual representations of speaker utterances from identical networks that differed only on the objective function they were trained on (GE2E-MM or Triplet [27]). Intraclass pairwise Euclidean distance was measured between all combinations in the set of utterance encodings obtained from the same speaker. Similarly, interclass pairwise distance was measured as the Euclidean distance between the reference speaker's encoded utterance *centroid* with that of the utterance centroid for every other speaker in the set.

The intraclass and interclass sample distributions for the triplet loss and GE2E-MM loss models are shown in Figure 2. Results show significant reduction in the overlap between intraclass and interclass distributions on the GE2E-MM loss trained network, implying more compact person representations. To validate this further, we employ the Silhouette coefficient [24], Calinski-Harabasz score [25], and the Davies-Bouldin score [26] on the test set. As shown in Table 1, the GE2E-MM loss show improved cluster quality for all three metrics.

4.2.2. Effect of the Weakly-Supervised Auxiliary Task and AV-Mixup Multimodal Augmentation:

The results in Table 1 show the improvement of cluster quality when comparing to the same network trained without the additional task loss. This sentiment is echoed in the results of Ta-

Table 2: *Ablation study showing the effect of proposed loss function, auxiliary task training, and AV-Mixup on EER.*

Loss	Training Config		Evaluation (VC1)		
	Aux	AV-Mix	O	E	H
Triplet [27]	N	Y	6.85	9.57	5.16
GE2E-MM	N	N	0.427	0.321	0.568
GE2E-MM	N	Y	0.323	0.292	0.507
GE2E-MM	**Y**	**Y**	**0.244**	**0.252**	**0.441**

Table 3: *EER of REPTAR in the presence of corrupted and missing modalities on the VoxCeleb1-O test split*

Architecture	Audio Input	Visual Input	**EER**
REPTAR	*clean*	*clean*	0.24
	corrupt	*clean*	1.98
	missing	*clean*	1.12
	clean	*corrupt*	6.12
	clean	*missing*	1.64

Table 4: *EER measurements on various training dataset configurations. Best results per evaluation split are in **bold***

Model	Train Set		Evaluation (VC1)		
	VC2	VB	O	E	H
Lin et al (M3)[11]	Y	N	0.622	0.761	1.391
REPTAR	Y	N	0.244	**0.252**	**0.441**
Lin et al (M3)[11]	Y	Y	0.441	0.681	1.268
REPTAR	Y	Y	**0.196**	0.316	0.537

ble 2, showing an average 17% EER improvement on the VC1-O/E/H test splits. We believe that the age classification auxiliary task ensured that distinctive markers from both modalities are preserved in the multimodal representation to help generalization and improve overall performance.

Through randomization of visual and audio speaker inputs by AV-Mixup, we were able to see an average 15% improvement in performance compared to a model trained using audio and visual inputs from the same utterance. Similar to the findings of Nagrani et al. [16] on disentangled linguistic content and speaker identity yielding better generalization, we believe our improvement is also as result of minimizing mutual information, but within the speaker audio-visual input space.

4.2.3. Effect of Corrupted and Missing Modality:

Measuring robustness to non-ideal situations was performed by recreating absent or corrupt modalities. An absent modality was emulated by setting the input to zero. A corrupt input was emulated using additive white Gaussian noise (AWGN), with $\mu = 0$ and $\sigma_v = [0, 255]$ or $\sigma_a = [-1, 1]$, where μ and σ are mean and standard deviation, respectively. This methodology is consistent with existing literary works [1].

The results in Table 3 show that the multimodal network is robust to missing or corrupt inputs without significantly compromising performance. This can be compared to other architectures that rely on both modalities to be present, a constraint that is not always feasible in real-world scenarios.

Table 5: *Proposed model EER performance compared to SOTA. Lower value signifies a better result. Best results are in **bold***

Model	Modality	VoxCeleb1		
		O	E	H
Chen et al [7]	A	2.31	2.23	3.78
	V	2.26	1.54	2.37
	AV	0.585	0.427	0.735
Qian et al [4]	A	1.62	1.75	3.16
	V	3.04	2.18	3.23
	AV	0.558	0.441	0.793
Sun et al [9]	A	**0.99**	1.24	2.27
	V	1.44	1.28	2.14
	AV	**0.18**	0.26	0.49
Lin et al (M3) [11]	AV	0.622	0.761	1.39
Lin et al (M4) [11]	AV	0.580	0.775	1.44
REPTAR	A	1.64	**1.12**	**1.85**
	V	**1.12**	**1.19**	**1.82**
	AV	0.244	**0.252**	**0.441**

4.2.4. Effect of Additional Training Data:

The VoxBlink-clean (VB) dataset was explored as a potential complementary training dataset [11], and was compared to against the reported benchmarks. Results are shown in Table 4. Despite the additional set of 1.45M utterances across 38K new speakers, a degradation of performance on the VC1-E/H splits was observed. This leads us to believe that the training and optimization strategies proposed (i.e. MTL, AV-Mixup) in REPTAR can be seen as an alternative, more data-efficient way to improve model performance.

4.2.5. Summary of Results

Our proposed model REPTAR achieves competitive performance against the previous state of the art, yielding best published results in 7 of 9 test configurations. Results are shown in Table 5 along with a comparison to related works. The results of Tao et al. [10] were omitted due to train-test contamination.

Our proposed model is able to achieve SOTA performance on all test configurations of the VoxCeleb1-E and VoxCeleb1-H test splits, which are considerably larger and targets a broader demographic compared to the VoxCeleb1-O split. Jointly, VoxCeleb1-E and VoxCeleb1-H can be used to describe the REPTAR's generalizeability and quality of feature extraction. The results show that REPTAR goes beyond encoding basic high-level features such as nationality and gender in the representation space, and is able to extract features that can be used to distinguish between even the most similar of speakers.

5. Conclusion

In this paper we explored data-efficient approaches to improving the robustness of speaker verification systems. Specifically, we demonstrated how DML representation learning can be enhanced, by introducing an auxiliary task trained on inexpensive, weak labels and measuring the quality of the resulting speaker representations. We also show how noise in the training set can be leveraged to improve generalization by introducing the GE2E-MM loss as well as AV-Mixup, a multimodal data augmentation technique. A comprehensive study of MTL task selection and task weighting strategies is left as future work.

6. References

[1] S. Shon, T. H. Oh, and J. Glass, "Noise-tolerant Audio-visual Online Person Verification Using an Attention-based Neural Network Fusion," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[2] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What Makes Multi-modal Learning Better than Single (Provably)," *Advances in Neural Information Processing Systems*, 2021.

[3] B. Shi, A. Mohamed, and W. N. Hsu, "Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022.

[4] Y. Qian, Z. Chen, and S. Wang, "Audio-Visual Deep Neural Network for Robust Person Verification," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2021.

[5] L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[6] R. Tao, R. K. Das, and H. Li, "Audio-visual speaker recognition with a cross-modal discriminative network," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020.

[7] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020.

[8] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[9] P. Sun, S. Zhang, Z. Liu, Y. Yuan, T. Zhang, H. Zhang, and P. Hu, "Learning Audio-Visual embedding for Person Verification in the Wild," *arXiv preprint arXiv:2209.04093*, 2022.

[10] R. Tao, K. A. Lee, Z. Shi, and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[11] Y. Lin, X. Qin, M. Cheng, N. Jiang, G. Zhao, and M. Li, "VoxBlink: X-Large Speaker Verification Dataset on Camera," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. [Online]. Available: http://arxiv.org/abs/2308.07056

[12] K. Kobs, M. Steininger, A. Dulny, and A. Hotho, "Do Different Deep Metric Learning Losses Lead to Similar Learned Features?" *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[13] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[14] J. Arevalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González, "Gated multimodal units for information fusion," *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2019.

[15] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," *arXiv preprint arXiv:2208.08012*, 2022.

[16] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled Speech Embeddings Using Cross-Modal Self-Supervision," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.

[18] A. Nagraniy, J. S. Chungy, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[19] J. W. Jung, Y. J. Kim, H. S. Heo, B. J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[21] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *13th IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[22] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-VOX-Celeb: Multi-Modal Corpus for Facial and Speech Estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[23] D. P. Kingma and J. Lei Ba, "ADAM: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2015.

[24] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, 1987.

[25] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[26] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.