Deep Information Maximisation to Mitigate Information Loss in Text Independent Speaker Verification

Nipun Fonseka, Nirmal Sankalana, Buddhika Karunarathne, Uthayasanker Thayasivam
Dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
{nipunf.19, nirmalsankalana.19, buddhika, rtuthaya}@cse.mrt.ac.lk

Abstract

This paper presents a novel approach to mitigate information loss in text-independent speaker verification by leveraging Deep Information Maximisation (DIM). The proposed method aims to enhance the retention of speaker-specific information during the pooling process, which is crucial for creating accurate and high-level speech signal representations. By incorporating mutual information maximisation techniques, the DIM method optimises the statistical dependency between frame-level features and their corresponding high-level embeddings. Experiments conducted on the VoxCeleb1 dataset demonstrate the efficacy of the approach, showing a significant reduction in the Equal Error Rate (EER). Our best configuration achieved an EER of 1.5376, an improvement over the baseline model's EER of 1.6119. These findings indicate that the integration of DIM can effectively enhance the performance and accuracy of speaker verification systems.

1 Introduction

Speaker verification is the task of determining whether a speaker's claimed identity is true. This process involves two main phases: the first is converting a speech signal into a fixed-dimensional, high-level representation called an embedding; the second is measuring the similarity between such embedding to verify identity.

In text-independent speaker verification, pooling is essential for combining frame-level features into a single, higher-level representation. However, this process can lead to the loss of crucial speaker information, vital for accurate speaker verification. Various techniques have been proposed to address this issue, including attention-based pooling (Okabe et al., 2018), multi-level pooling (Tang et al., 2019), and vector-based attentive pooling (Gao et al., 2020). Despite these advancements, significant information can still be lost due to the inherent compression involved in pooling.

With the rise of deep learning, deep neural networks have become widely used in speaker verification for producing consistent, high-level representations of speech signals. Starting with x-vector systems (Snyder et al., 2018), various methods have been developed over time, including Time Delay Neural Networks (TDNNs) (Liu et al., 2022), Long Short-Term Memory networks (LSTMs) (Mobiny and Najarian, 2018), Extended Context-Aware Permutation-Invariant TDNNs (ECAPA-TDNNs) (Desplanques et al., 2020), Convolutional Neural Networks (CNNs) (Zhou et al., 2019; Zhao et al., 2020; Kim et al., 2022), and more recently, transformers (Peng et al., 2023). While these neural networks are effective at feature extraction, the pooling process is crucial for creating fixed-dimensional high-level representations. Researchers have developed several pooling methods, evolving from statistical (Variani et al., 2014) techniques to advanced methods like multi-headed attentive pooling (Zhu et al., 2018), aiming to optimise the pooling process.

Mutual information is a measure that quantifies the statistical dependence between two random variables. Belghazi et al. introduced a neural network-based method called the Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018) for estimating the mutual information between two variables. This method was adapted by Hjelm et al. to create deep representations of images by maximising the mutual information between an image and its high-level representation (embedding) (Hjelm et al., 2019).

In this research, we apply the Deep Information Maximisation technique proposed by Hjelm et al. to mitigate information loss during pooling in speaker verification. Our approach aims to enhance the retention of speaker-specific information, thereby improving the performance and accuracy of text-independent speaker verification systems.

2 Related Work

Our baseline model for speaker verification builds on the work of Peng et al., which introduces an attention-based backend for fine-tuning pre-trained Automatic Speech Recognition (ASR) Transformer models. This approach leverages the ability of pre-trained Transformers to capture meaningful acoustic and phonetic representations while introducing a lightweight backend to extract speaker-discriminative features effectively (Peng et al., 2023).

The core component of the attention-based backend is the Multi-Head Factorised Attentive Pooling (MHFA) mechanism (Peng et al., 2023). It aims to condition the speaker representations on the phonetic content of the input utterance, enabling the model to capture both speaker and phonetic information simultaneously. The output feature map of each layer of the transformer is utilised here by assigning two types of attention weights.

Given the output representations $\mathbf{Z}_l \in R^{T\times F}$ from the l-th Transformer layer of the pre-trained model, where T is the number of frames and F is the feature dimension, the MHFA method computes two factorised representations: keys $\mathbf{K}$ and values $\mathbf{V}$, as follows:

$$\mathbf{K} = \sum_{l=1}^{L} w_k^l \mathbf{Z}_l \mathbf{S}_k, \quad \mathbf{V} = \sum_{l=1}^{L} w_v^l \mathbf{Z}_l \mathbf{S}_v \qquad (1)$$

Here, $\mathbf{w}_k^l$ and $\mathbf{w}_v^l$ are learnable weights that aggregate the layer-wise outputs, and $\mathbf{S}_k \in R^{F\times D}$ and $\mathbf{S}_v \in R^{F\times D}$ are linear projections that reduce the dimensionality of keys and values, respectively, to $\mathbf{D}$.

The multi-head attention mechanism is then applied to aggregate the values $\mathbf{V}$ over frames, conditioned on the keys $\mathbf{K}$:

$$\mathbf{A} = \text{softmax}(\mathbf{K}\mathbf{Q}^\top) \qquad (2)$$

$$\mathbf{c}_h = \sum_{t=1}^{T} \mathbf{A}_{ht} \mathbf{V}_t \qquad (3)$$

$$\mathbf{c} = \text{concat}(\mathbf{c}_1, \ldots, \mathbf{c}_H) \qquad (4)$$

Here, $\mathbf{Q} \in R^{D\times H}$ contains the learnable query vectors for each of the H attention heads, $\mathbf{A} \in R^{T\times H}$ is the attention matrix, and $\mathbf{c}_h \in R^{1\times D}$ and $\mathbf{c} \in R^{1\times HD}$ are the sub-representations and the final utterance-level speaker representation, respectively.

The key idea behind MHFA is that the keys $\mathbf{K}$ capture phonetic information, allowing each attention head to focus on a specific set of phonetic units. Simultaneously, the values $\mathbf{V}$ encode speaker discriminative information, ensuring that the final representation $\mathbf{c}$ is conditioned on both speaker and phonetic characteristics.

To stabilise the fine-tuning process and improve performance, propose two strategies (Peng et al., 2023):

1. Fine-Tuning Regularisation: An $\mathbf{L}_2$ regularization term is added to the overall loss function, encouraging the fine-tuned model's weights to remain close to the initial pre-trained weights:

$$\mathcal{L} = \mathcal{L}_{spk} + \lambda \sum_{j=1}^{|\Theta|} \|\theta_j - \theta_j^p\|_2^2 \qquad (5)$$

Here, $\mathcal{L}_{spk}$ is the speaker classification loss, Θ denotes the model parameters, θ_j^p are the corresponding parameters from the initial pre-trained model, and λ is a hyperparameter controlling the strength of the regularisation.

2. Layer-wise Learning Rate Decay (LLRD): Instead of using the same learning rate for all Transformer layers during fine-tuning, LLRD assigns lower learning rates to the bottom layers and higher rates to the top layers, as follows:

$$\text{LR}_l = \text{LR}_1 \cdot \xi^{l-1} \qquad (6)$$

Here, LR_l is the learning rate for the l-th Transformer layer, LR_1 is the base learning rate for the bottom layer, and ξ is a weight decay factor controlling the rate of increase in learning rates across layers.

The authors demonstrate that these fine-tuning strategies, combined with the MHFA backend, achieve state-of-the-art performance in speaker verification while significantly reducing training time compared to previous approaches.

3 Proposed Method

The proposed method is inspired by the research presented in Deep Information Maximiser (Hjelm et al., 2019). Here we introduce a regularisation mechanism aimed at increasing the mutual information between the high-level final embedding and the frame-level features. This enhancement seeks to retain valuable information from the frame-level features.

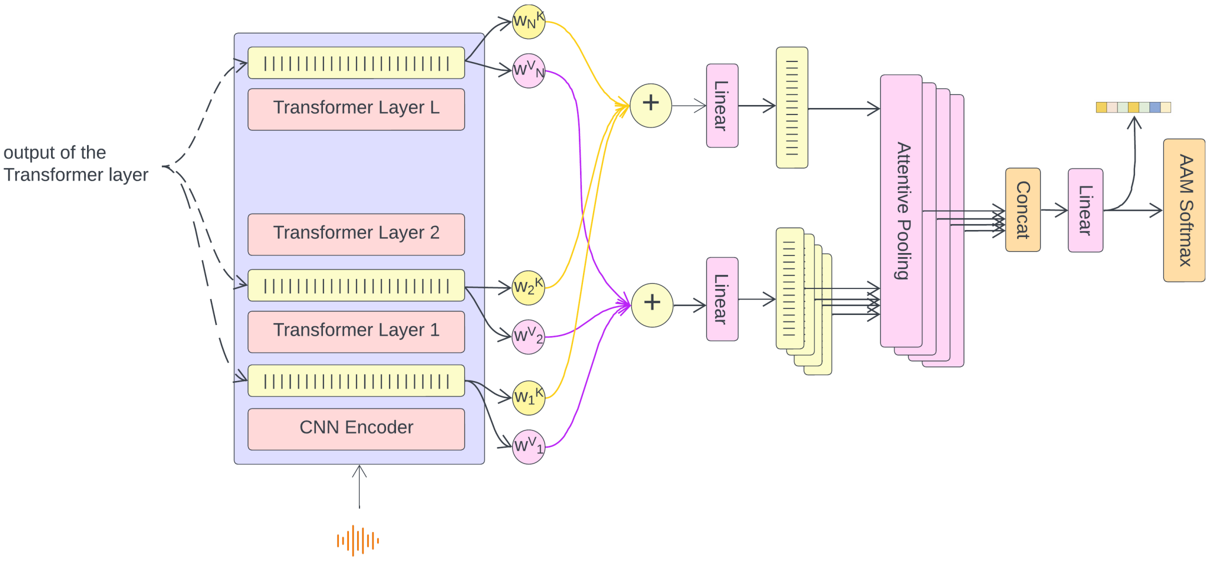

Figure 1: baseline model (Peng et al., 2023)

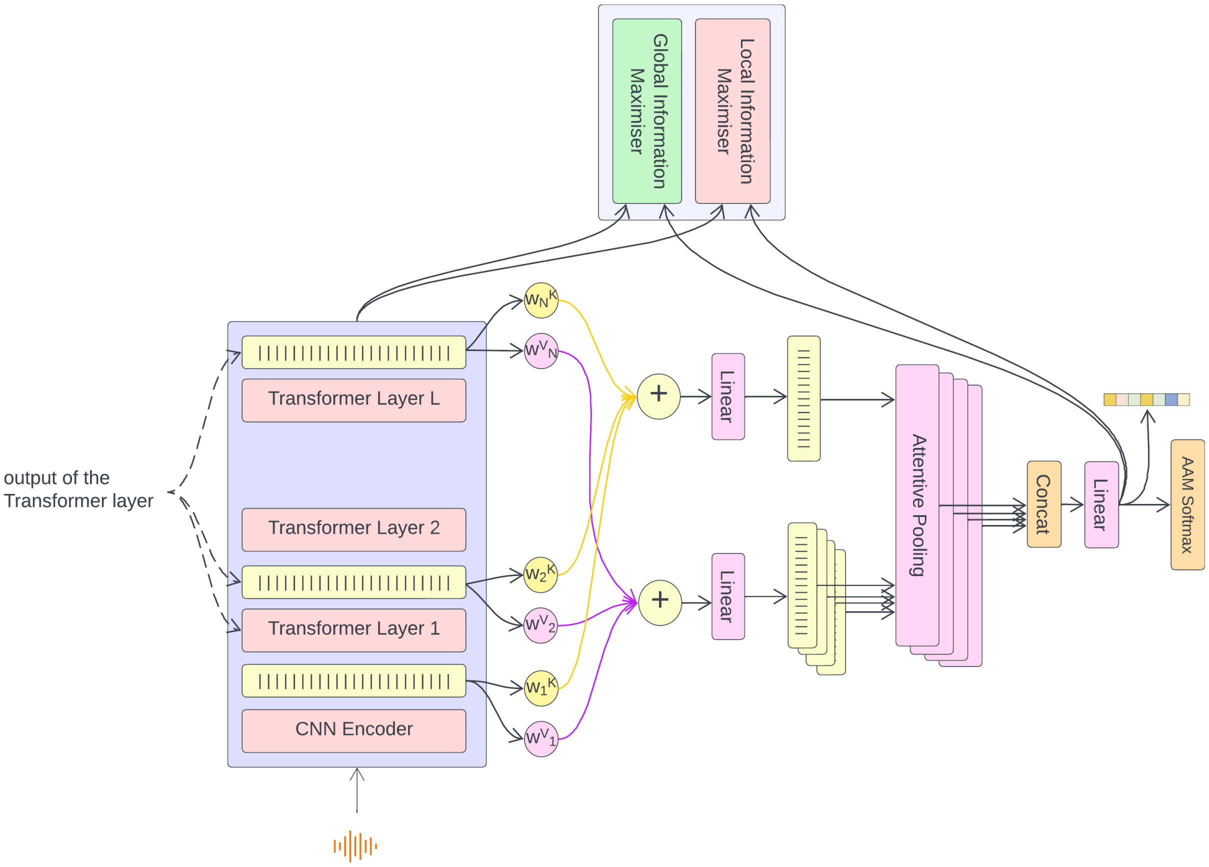

Figure 2: Proposed Model

An additional discriminator, taking a pair of frame-level and high-level embedding as input, is introduced to estimate and maximise the mutual information between these two sets of features. This process effectively functions as a regularising term for the pooling layer, encouraging the embedding vector to capture as much meaningful information from the frame-level features as possible. The discriminator, which functions as a neural network, assesses whether a given concatenated pair of frame-level (low-level) feature maps and high-level embedding corresponds, thereby estimating the common information shared between them.

Two specific discriminators, Global InfoMax (GIM) and Local InfoMax (LIM), which are tailored to capture local and global structures, are employed to estimate and maximise local and global mutual information, respectively.

3.1 Global Information Maximisation (GIM)

Global Information Maximisation (GIM) seeks to increase the mutual information between the output feature map from the ASR encoder and the final speaker embedding. This approach is designed to enhance consistency and contextual understanding within the speaker verification process. GIM optimises $E_\psi : X \to Y$ with neural network parameters ψ, by maximising the mutual information $\mathcal{I}(X, E_\psi(X))$ between X and $E_\psi(X)$. Here, X is the intermediate feature map, and $E_\psi(X)$ is the final embedding created after pooling.

$$(\hat{\omega}, \hat{\psi})_G \in \arg\max_{\omega,\psi} \hat{\mathcal{I}}_\omega(X; E_\psi(X)), \qquad (7)$$

To achieve this, GIM flattens the ASR transformer's feature maps along the feature axis and then concatenates them with the final speaker embedding. Based on this concatenation, GIM assigns a score to measure the mutual information, thereby providing a more accurate representation of the speaker's unique characteristics.

3.2 Local Information Maximiser (LIM)

While GIM can introduce irrelevant dependencies, such as noise, that are not useful for classification, the Local Information Maximiser (LIM) addresses this by focusing on maximising the average mutual information between the high-level embedding and all local frames of the feature map. This approach encourages high-level representation to maintain high mutual information with all frames, promoting the encoding of aspects of data that are shared across frames.

LIM optimises E_ψ with neural network parameters ψ, by maximising the average mutual information $\mathcal{I}(X, E_\psi(X))$ between all the frames F and $E_\psi(X)$. Here, X represents the intermediate feature map, and $E_\psi(X)$ is the final embedding created after pooling.

$$(\hat{\omega}, \hat{\psi})_L = \arg\max_{\omega,\psi} \frac{1}{F} \sum_{i=1}^{F} \mathcal{I}_{\omega,\psi}(x_i; E_\psi(X)) \quad (8)$$

In this formulation, the final embedding is concatenated with each frame of the ASR transformer feature map (intermediate representation). By maximising mutual information between each local frame and the high-level embedding, LIM ensures that the high-level embedding captures the most relevant and shared information across all frames, enhancing the robustness and accuracy of the classification task.

4 Loss Function

Both LIM and GIM are applied together to train the model, optimising the classification loss during training. The overall loss function can be described as follows:

$$\begin{aligned} L_{Total} &= L_{Classification} \\ &+ \alpha \hat{I}_{\omega_G,\psi}(X; E_\psi(X)) \\ &+ \beta \hat{I}_{\omega_L,\psi}(x_i; E_\psi(X)) \end{aligned} \qquad (9)$$

The first term, $L_{Classification}$, is the speaker classification loss. The Additive Angular Margin (AAM) Softmax loss function is used as the classification loss. The second and third terms are the global and local MINE objectives ω_G and ω_L are the parameters for the global and local discriminators, respectively). These MINE objectives act as regularisation terms with weights α and β during the training of the entire system. The total loss function is calculated as follows:

$$\begin{aligned} &\hat{T}^{(JSD)}_{\omega,\psi}(X; E_\psi(X)) = \\ &E_P[-\text{sp}(-T_{\psi,\omega}(x, E_\psi(x)))] \\ &-E_{\tilde{P}}[\text{sp}(T_{\psi,\omega}(x', E_\psi(x)))] \end{aligned} \qquad (10)$$

The Jensen-Shannon Divergence (JSD) is used as the objective function for MINE. It returns the

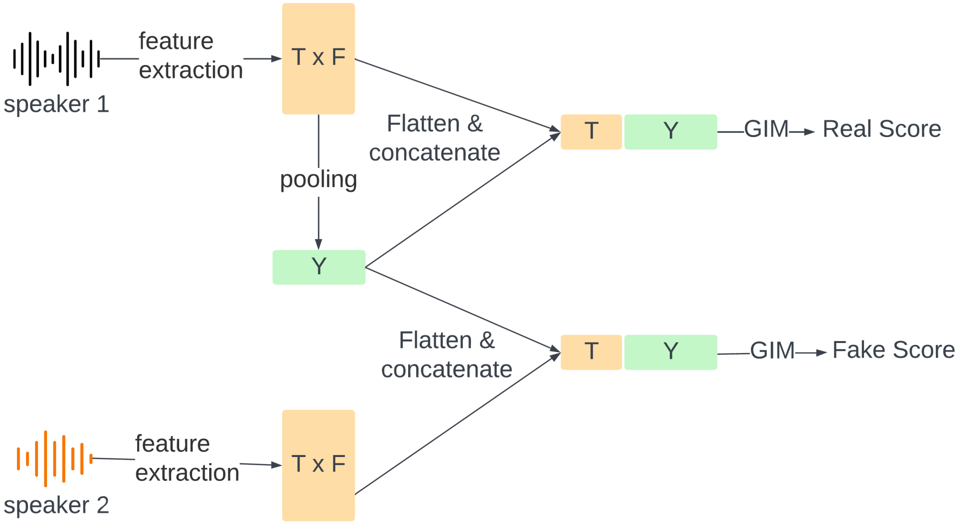

Figure 3: Global Information Maximiser

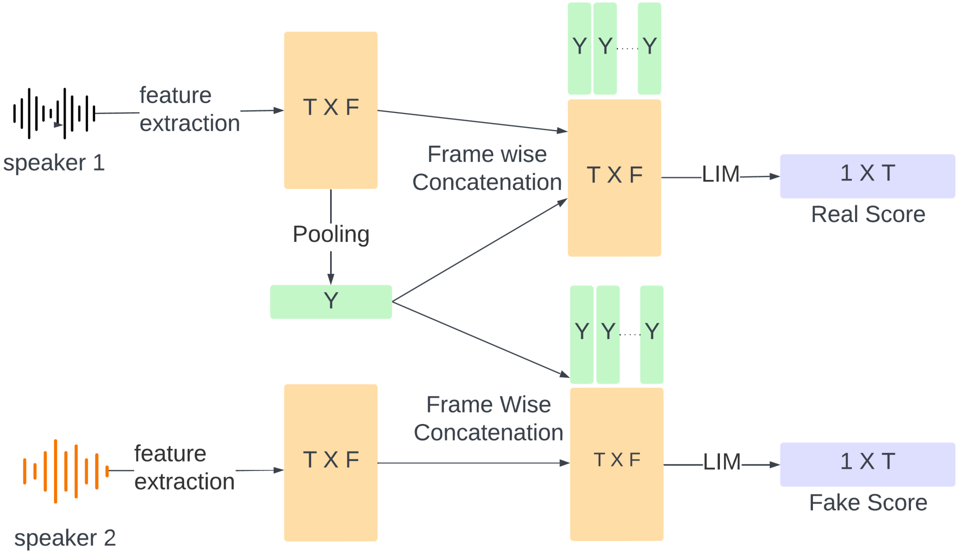

Figure 4: Local Information Maximiser

difference between the softmaxed estimated mutual information of positive pairs (marginal distribution) and the softmaxed estimated mutual information of negative pairs. The JSD provides better and more stable results (Ravanelli and Bengio, 2019) compared to the Kullback-Leibler (KL) divergence used by Belghazi et al.

5 Experiments

5.1 Experiment Setup

To train our proposed model, we utilised the VoxCeleb1 development set (Nagrani et al., 2017), a widely recognised large-scale dataset for text-independent speaker verification. We evaluated the model's performance using the VoxCeleb1 test corpus. For the ASR transformer, we employed the WaveLM-Base-Plus (Chen et al., 2022) model due to its strong performance in previous studies (Peng et al., 2023). The transformer's output had dimensions of 150 x 768, with 150 representing the total number of frames and 768 representing the feature dimension for each frame. Model training was conducted on two 16GB NVIDIA Tesla GPUs in a distributed manner, with a batch size of 32. We conducted experiments both with and without the Deep Information Maximisation (DIM) method, adjusting parameters such as α, β.

The learning rate was set to 0.0001, with a decay rate of 0.95. Both the Local Information Maximiser (LIM) and the Global Information Maximiser (GIM) were implemented using one-

Layer	in channels	out channels	feature dimension	kernel size
conv 1	768	256	150	3
conv 2	256	64	148	3
fc 1	9600	512	-	-
fc 2	512	1	-	-

Table 1: Layer configuration for GIM

Layer	in channels	out channels	feature dimension	kernel size
conv 1	256 + 768	512	150	1
conv 2	512	512	150	1

Table 2: Layer configuration for LIM

dimensional Convolutional Neural Networks (1D CNNs) and fully connected layers.

We used one-dimensional Convolutional Neural Networks (1D CNNs) for both Local and Global Information Maximisation (InfoMax) because; firstly, audio data is inherently sequential, with each time step represented by a feature vector, making 1D convolutions ideal for capturing temporal dependencies and local patterns along the time axis. This approach also reduces computational complexity and the number of model parameters compared to 2D convolutions, enhancing efficiency.

5.2 Experiment Results

The experiment evaluated the proposed Deep Information Maximisation (DIM) approach integrated with an attention-based backend for text-independent speaker verification. The primary metric used for performance evaluation was the Equal Error Rate (EER), where a lower EER indicates better performance.

No attention heads	EER
1	1.877
2	1.681
4	1.612
8	1.485
16	1.419
32	1.336

Table 3: Experimental Results for baseline model with different number of attention heads

As the number of attention heads increases, the EER consistently decreases, demonstrating that using more attention heads improves the accuracy of the speaker verification system. The lowest EER of 1.336 is achieved with 32 attention heads.

No of attention heads	Baseline EER	DIM integrated baseline EER
1	1.877	1.845
2	1.681	1.677
4	1.612	1.538
8	1.485	1.441
16	1.419	1.389
32	1.336	1.392

Table 4: Comparison of Baseline and DIM integrated baseline for different attention heads

	$\beta = 0.01$	$\beta = 0.05$	$\beta = 0.1$
$\alpha = 0.01$	1.8664	1.5376	1.7656
$\alpha = 0.05$	1.6278	1.7073	2.0308
$\alpha = 0.1$	1.7709	1.7232	1.9618

Table 5: Experiment results(EER) of the DIM integrated base for different α and β values with four attention heads.

The DIM method was tested with various configurations of the hyperparameters α and β, which control the weights of the global and local mutual information maximisation terms, respectively. The results are presented in the table below, comparing different values of α and β with the baseline model, which had an EER of 1.612.

The experiment results indicate that the integration of the DIM method can improve the performance of the speaker verification system. The configuration with α=0.01 and β=0.05 achieved the best EER of 1.5376, which is an improvement over the baseline EER of 1.612.

In high levels, increasing the number of attention heads generally leads to lower EER, indicating better performance. With up to 16 attention heads, DIM-integrated models outperform the baseline.

Optimal values for α and β significantly impact performance, with lower values generally resulting in better EER, with the optimal combination being α=0.01 and β=0.05. Compared to α, the β has a stronger influence on EER, which emphasizes the importance of Local InfoMax. Overall, the DIM integrated baseline shows consistent improvements over the baseline, confirming the effectiveness of the DIM integration.

Possible Reasons for Variations in Experimental Results for baseline and DIM Integrated in 32 attention heads. (1) Increasing the number of attention heads enhances the model's ability to capture detailed speaker-specific characteristics by simultaneously focusing on multiple aspects of the input features. While this can reduce the Equal Error Rate (EER), it also increases model complexity, which can lead to overfitting if the training data lacks sufficient diversity. (2) There appears to be a saturation point beyond which adding more attention heads does not significantly improve performance. Beyond this point, additional attention heads provide diminishing returns, potentially leading to inefficiency and a low performance/cost ratio. (3) The DIM method aims to maximise mutual information between frame-level and high-level features, preserving discriminative features essential for effective speaker verification. In models with 32 attention heads, the high-dimensional space created can make it difficult for DIM to preserve and maximise relevant information without introducing noise or redundancy, thus conflicting with the model's complexity. (4) The hyperparameters α and β are crucial for model performance as they control the emphasis on mutual information maximisation. Lower values are generally preferred to prevent the model from overly focusing on mutual information at the expense of classification accuracy. However, if set too low, the model might not fully leverage the benefits of Deep InfoMax.

6 Conclusion

The research investigated the use of Deep Information Maximisation (DIM) to mitigate information loss in text-independent speaker verification systems, focusing on the impact of attention heads and DIM integration on performance. Findings revealed that Local Information Maximisation (LIM) plays a significantly larger role than Global Information Maximisation (GIM) in maximising mutual information, highlighting the importance of preserving local context for accuracy improvement.

Experiments showed that up to 16 attention heads, the DIM-integrated model outperformed the baseline by reducing the Equal Error Rate (EER). Beyond this point, the EER increased, indicating limitations in handling higher complexity and potential noise introduction. This suggests that while DIM is beneficial, its integration with numerous attention heads requires careful balancing to avoid overfitting and diminishing returns. The results emphasize the potential of attention mechanisms in capturing detailed speaker-specific characteristics but also underline the need to manage model complexity for optimal performance.

7 Future Work

Future research should aim to identify the optimal number of attention heads to balance model complexity and performance, involving further experiments and validation across diverse datasets, introducing data augmentation techniques, such as noise addition, which will allow us to evaluate the robustness and generalisation capability of our proposed method in more challenging and realistic conditions.

Advanced techniques to integrate DIM with more attention heads should be explored, including refining mutual information maximisation and incorporating additional regularisation to reduce noise.

Robust hyperparameter tuning for α and β is crucial. Studies should explore a broader range of these parameters to better understand their impact and identify the most effective configurations. Finally, future work should address the computational demands of training models with many attention heads by optimising training stability and efficiency or exploring alternative architectures.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022.

Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

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

Hongcan Gao, Xiaolei Hou, and Jing Xu. 2020. Vector-Based Attentive Pooling for Text-Independent Speaker Verification. In *Interspeech 2020*, pages 936–940. ISCA.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Seong-Hu Kim, Hyeonuk Nam, and Yong-Hwa Park. 2022. Decomposed Temporal Dynamic CNN: Efficient Time-Adaptive Network for Text-Independent Speaker Verification Explained with Speaker Activation Map. ArXiv:2203.15277 [cs, eess].

Tianchi Liu, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. 2022. Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7517–7521.

Aryan Mobiny and Mohammad Najarian. 2018. Text-Independent Speaker Verification Using Long Short-Term Memory Networks. ArXiv:1805.00604 [cs, eess].

A. Nagrani, J. S. Chung, and A. Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive Statistics Pooling for Deep Speaker Embedding. In *Interspeech 2018*, pages 2252–2256. ArXiv:1803.10963 [cs, eess].

Junyi Peng, Oldřich Plchot, Themos Stafylakis, Ladislav Mošner, Lukáš Burget, and Jan Černocký. 2023. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.

Mirco Ravanelli and Yoshua Bengio. 2019. Learning speaker representations with mutual information. pages 1153–1157.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, Calgary, AB. IEEE.

Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Deep speaker embedding learning with multi-level pooling for text-independent speaker verification. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6116–6120.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056, Florence, Italy. IEEE.

Yong Zhao, Tianyan Zhou, Zhuo Chen, and Jian Wu. 2020. Improving Deep CNN Networks with Long Temporal Context for Text-Independent Speaker Verification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6834–6838. ISSN: 2379-190X.

Tianyan Zhou, Yong Zhao, Jinyu Li, Yifan Gong, and Jian Wu. 2019. CNN with Phonetic Attention for Text-Independent Speaker Verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 718–725, SG, Singapore. IEEE.

Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. 2018. Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification. In *Interspeech 2018*, pages 3573–3577. ISCA.