International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.59, November 2025

Text-Independent Speaker Identification using Audio Looping with Margin based Loss Functions

Elliot Q.C. Garcia
Universidade Federal Rural de Pernambuco
Recife, Pernambuco, Brasil
Futuro Tech
Recife, Pernambuco, Brasil

Nicéias Silva Vilela
Universidade Federal Rural de Pernambuco
Recife, Pernambuco, Brasil

Kátia Pires Nascimento do Sacramento
Universidade Regional do Cariri
Juazeiro do Norte Ceará, Brasil

Tiago A.E. Ferreira
Universidade Federal Rural de Pernambuco
Recife, Pernambuco, Brasil

ABSTRACT

Speaker identification has become a crucial component in various applications, including security systems, virtual assistants, and personalized user experiences. This paper investigates the effectiveness of CosFace Loss and ArcFace Loss for text-independent speaker identification using a Convolutional Neural Network architecture based on the VGG16 model, modified to accommodate mel spectrogram inputs of variable sizes generated from the Voxceleb1 dataset. The approach involves implementing both loss functions to analyze their effects on model accuracy and robustness, where the Softmax loss function served as a comparative baseline. Additionally, the study examines how the sizes of mel spectrograms and their varying time lengths influence model performance using 3 seconds as the baseline, with 10 seconds being the maximum time length. The experimental results demonstrate superior identification accuracy compared to traditional Softmax loss in the model that was used. Furthermore, the paper discusses the implications of these findings for future research.

General Terms

Convolutional Neural Networks, Deep Learning

Keywords

Speaker Identification, Loss Functions, Data Augmentation, Mel Spectrograms

1. INTRODUCTION

Speaker Recognition (SR) has become a widely used application for security systems, virtual assistants, and personalized user experiences [1, 15]. Traditional methods, such as Mel-Frequency Cepstral Coefficients (MFCCs) [11, 23], have been widely used in SR but often struggle with noise sensitivity and limited discriminative power. With the advent of deep learning, spectrograms have emerged as an alternative for speaker identification [27, 3, 2]. Spectrograms provide a time-frequency representation of speech signals, capturing both temporal and spectral information. These spectrograms are often combined with Convolutional Neural Networks (CNNs) where the CNNs are trained in a supervised process guided by classification loss functions. Current prevailing classification loss functions for SR systems are mostly based on the Softmax Loss function. However, Softmax Loss does not explicitly optimize feature embedding to enforce higher similarity for intra-class samples and diversity for inter-class samples. To remedy this, this study approaches the problem using two different loss functions and compares them to the traditional Softmax approach, namely Arcface and CosFace Loss.

SR systems can be modeled either as Text-Dependent or Text-Independent systems. Text-Dependent SR relies on the user uttering whatever is being prompted, while text-independent SR provides a more flexible approach where there is no constraint on the user of what can be said, and as such, has broader applications. SR systems are generally categorized into three types:

—Speaker Verification (SV). Verifies whether a speaker's claimed identity is true. The system compares the speaker's voice to a stored voiceprint of the claimed identity to confirm their identity [10]. Speaker verification is commonly used in authentication systems, such as secure access to privileged information, devices, or accounts;

—Speaker Identification (SI). Determines the identity of an unknown speaker from a group of known speakers. The system compares the input voice to a database of voiceprints and identifies the closest match. SI can then also be split into two approaches, the closed-set approach and the open-set approach. In the case of closed-set identification, the speakers are initially already enrolled in the database, and the system assumes that the current unknown speaker is already enrolled in the database. In the case of open-set identification, the speaker is not always

enrolled in the database, as such the system needs to be capable of rejecting a speaker [5]. SI is widely used in applications such as law enforcement, voice assistants, and call center analytics [7];

—Speaker Classification (SC). Distinguishes and classifies speakers based on specific characteristics such as age, gender and health. It is commonly used in scenarios such as demographic analysis and targeted marketing [19].

Each category faces unique challenges and applications. For instance, speaker verification must handle variability in a person's voice due to emotional states or background noise, while SI requires robust matching algorithms to distinguish between similar voices, often necessitating large and diverse databases to improve accuracy. SC, on the other hand, requires sophisticated feature extraction techniques and machine learning models to accurately capture and analyze subtle vocal traits.

The underlying system for SI works similarly to a fingerprint matching process. Looking at the spectral content, the system can analyze the unique features of a person's voice and match it against a database. These features are highly distinctive, capturing the physiological and behavioral characteristics of an individual's voice [16].

The present paper makes three main contributions. First, the study evaluates the comparative effectiveness of the advanced loss functions ArcFace and CosFace against the traditional Softmax loss for SI using modified VGG16 architectures with mel-spectrogram inputs. It is worth noting that the time duration of the voice sample can influence the mel-spectrograms' capacity to represent the speaker. Therefore, the study also investigates two different sample durations: 3 and 10 seconds. To extend the sample duration, the paper proposes a new method based on time-loop repetition, where the original sample is repeated until the desired duration is reached. Accordingly, the second contribution focuses on the length of the voice samples. The research demonstrates that audio looping techniques significantly improve identification accuracy by extending shorter recordings to optimal lengths for feature extraction. Finally, the study systematically analyzes how varying mel-spectrogram image dimensions ($224 \times 224 \times 3$, $448 \times 448 \times 3$, and $432 \times 288 \times 3$) affect model performance, identifying optimal input configurations that balance computational efficiency with identification accuracy.

The paper is organized as follows. Section 2 presents related works on SI systems utilizing CNNs with spectrogram inputs and establishes the context for this paper's contributions. Section 3 describes the methodology, including preprocessing techniques applied to the VoxCeleb1 dataset[1] [17], the advanced loss functions employed and their theoretical advantages, and the architectural modifications made to the VGG16 model. Section 4 presents experimental results and discusses the impact of audio length and looping techniques, the comparative effectiveness of different loss functions, and the effects of varying mel-spectrogram dimensions on identification accuracy. Section 5 concludes the paper with a summary of findings and implications for future research.

2. RELATED WORK

This section is to briefly review several related or similar works in line with this paper. The technique of generating speech-based spectrograms is implemented during the feature extraction stage, while CNN are used during the classification stage.

[1]https://www.robots.ox.ac.uk/˜vgg/data/voxceleb/vox1.html

Nagrani *et al.* [17] originally introduced and tested the VoxCeleb1 dataset. They randomly selected three second segments from each speech and generated mel-spectrogram images of size $512 \times 300 \times 3$. For classification, they used a modified VGG CNN, having changed the final maxpool layer to avgpool. They achieved a top-1 accuracy of 80.5% with the traditional Softmax loss function.

Anand *et al.* [3] propose a 3-second random selection of each speech and then convert them to spectrogram images of size $128 \times 300 \times 1$. In the classification stage, the spectrograms are classified by different CNNs, such as VGG, ResNet, and CapsuleNet. They employed a combination of Margin Loss for training the capsule networks and Prototypical Loss for generalizing under unseen speakers. The ResNet classifier for 200 classes from the Voxceleb1 dataset has the best result of top-1 = 71.8%.

An *et al.* [2] improved upon the original SI system's accuracy [17] by altering the VGG network, specifically replacing the final max pooling layer with an average pooling layer and adding a self-attention layer before this pooling layer. This modification enabled the system to manage variable-length segments effectively. The study employed a combination of a penalization term alongside the traditional cross-entropy loss. For feature extraction, the researchers utilized mel-spectrograms, extracted from three-second audio segments, with dimensions of $512 \times 300 \times 3$. The resulting accuracy on the VoxCeleb1 dataset was a top-1 of $= 90.8\%$.

Similarly, Yadav and Rai [27] extracted mel-spectrograms from three-second utterances. However, they were generated with size $301 \times 161 \times 1$. One of the methods used in the classification stage was a VGG13-based CNN by adding a batch normalization layer after every convolutional layer. The system employs joint supervision of Softmax and Center Loss, achieving a top-1 accuracy of 89.5% on the VoxCeleb1 dataset.

Sharif *et al.* [21] initially extracted mel-spectrogram images of size $535 \times 678 \times 3$, resizing them afterwards to $100 \times 128 \times 3$. The study optimized the VGG-13 architecture for speaker recognition by reducing the number of convolutional layers from 10 to 5 and changing the pooling layers from max pooling to average pooling. They also added batch normalization layers after each average pooling layer and decreased the dropout layers from 10 to just 1. Furthermore, the fully connected layers were modified from three layers to a single layer with 1251 hidden neurons, matching the number of speakers in the VoxCeleb1 dataset. The study primarily utilized triple loss, n-pair loss, and angular loss, achieving a top-1 accuracy of 91.17%.

The works discussed in this section demonstrate several trends. Researchers have consistently modified VGG network architectures, achieving positive results in SI tasks. They have employed advanced loss functions, resulting in improved model generalization capabilities. The reported accuracy ranges from 71.8% to 91.17% across different studies [17, 3, 2, 27, 18]. All studies utilized VoxCeleb1 as their benchmark dataset while implementing various data augmentation techniques and experimenting with different spectrogram dimensions. Building upon these established approaches, this work contributes to this research landscape by implementing modifications to the VGG16 architecture, systematically comparing two advanced loss functions, namely ArcFace and CosFace, and experimenting with multiple mel-spectrogram dimensions. Additionally, the paper introduces an audio looping technique as a data augmentation method for the VoxCeleb1 dataset, achieving a competitive top-1 accuracy of 83.15% that falls within the established performance range while providing new insights into optimal input configurations.

Table 1. Information about the database used for experimentation.

Parameters	Voxceleb1
Total number of speakers	1251(688 male, 563 female)
Total utterances	153,516
Train utterances	108,018
Validation utterances	23,066
Test utterances	22,432

3. METHODOLOGY

3.1 Dataset

For this study, the VoxCeleb1 dataset [17] was used, a widely used benchmark for SR and verification tasks. VoxCeleb1 contains over $100,000$ utterances from $1,251$ speakers, collected from publicly available interview videos on platforms such as YouTube [17].
The audio samples in VoxCeleb1 are recorded at a sample rate of 16kHz, which is standard for speech processing tasks. The dataset includes recordings in multiple languages, with speakers of diverse accents, ages, and genders, being made up of 688 males and 563 females.
The utterances are recorded in various environments, from outdoor stadiums to quiet indoor studios, with almost all of them having some form of background noise, such as laughter, background chatter, and overlapping speaking.
For the experiment, a train-validation-test split of 70%, 15%, 15% was used, as illustrated in Table 1.

3.2 Preprocessing

A lot of the audios start or end with silence. Figure 1 shows an example where there is a silent situation at the beginning (region with only purple color). The preprocessing begins by reducing, ensuring that the system does not waste resources by processing empty data.

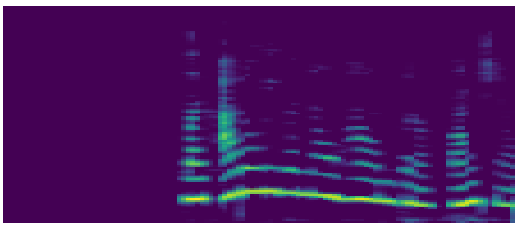

Fig. 1. Spectrogram image of audio before silence removal.

3.2.1 Data Augmentation. After silence removal, recordings that did not reach the desired duration were extended using an audio looping technique, as illustrated in Figure 2. This created two dataset versions: (a) three-second clips and (b) ten-second clips. Three seconds was chosen as a standard in SI research, while ten seconds was selected based on observations of diminishing returns beyond this duration.
Audio looping was applied only to segments shorter than the target duration. Samples meeting or exceeding the target length were either used as-is or truncated to the desired duration. Further investigation is needed to evaluate this method's effectiveness in different environments or with alternative looping strategies.

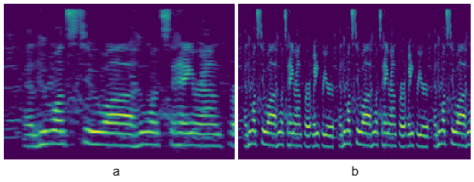

Fig. 2. Visual representation of mel-spectrograms (a) before and (b) after looping. The looping process extends shorter audio segments to meet the required duration.

3.2.2 Mel-spectrogram. Spectrograms are a graphical representation of a signal's frequencies. When applied to an audio or voice signal, it generates a visual representation of signals, depicting the distribution of energy across time and frequency. They are generated using the Short-Term Fourier Transform (STFT) [4], which converts raw audio signals into a two-dimensional representation, where the x-axis represents time, the y-axis represents frequency, and the color intensity represents the energy amplitude, as can be seen in Figure 3.
Spectrograms, particularly mel-spectrograms, are robust to noise and variations in recording conditions, as shown by Nagrani *et al.* [17], Lambamo *et al.* [12], and Saritha *et al.* [20]. In these works, the authors showed that spectrograms can mitigate the effects of background noise by focusing on the frequency patterns that are most relevant to the characteristics of the speaker. The conversion of spectrograms into the Mel scale is done because the Mel scale is designed to mimic the way humans perceive sound, with a non-linear frequency resolution that assigns more weight to lower frequencies. Lower frequency components have been shown to be more distinguishing between speakers in SI systems [14, 28].

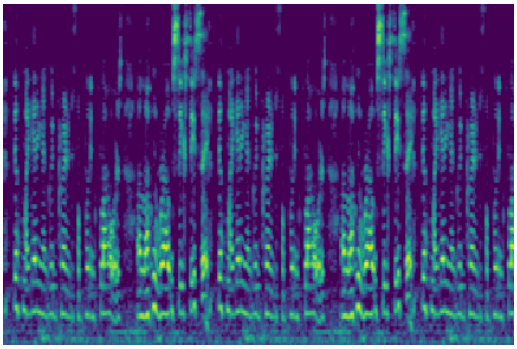

Fig. 3. A typical Mel-spectrogram image used to represent a speech signal.

Using the audio data previously generated, the recordings are converted into ten-second Mel-spectrograms. These are then resized into three specific dimensions: $224 \times 224 \times 3$, $448 \times 448 \times 3$, and $432 \times 288 \times 3$. The same process is applied to the dataset of three-second audio recordings. Additionally, mean and variance normalization is applied on a per-speaker basis to ensure consistent processing across the dataset [17].

3.3 Loss functions

Much research has been conducted using the cross-entropy loss function, commonly known as Softmax loss [17, 2]. The Softmax loss function for a batch of n samples is mathematically defined as:

$$L_{\text{AMS}} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(P(y=y_i|x_i;W)\right) \tag{1}$$

where $(P(y=y_i|x_i;W))$ is the probability of y is the class y_i, given the sample x_i and the weight vector W.
While these approaches have yielded positive results, Softmax presents several limitations for SI. A primary drawback is its failure to explicitly enforce discriminative margins between different speaker classes. This can result in learned embeddings for distinct speakers being too close in the feature space, hindering the model's ability to differentiate them, particularly under noisy conditions or with limited training data. Furthermore, Softmax does not directly optimize for intra-class compactness; while it pushes embeddings away from decision boundaries, it doesn't strongly encourage embeddings from the same speaker to cluster tightly together as seen in Figure 4. This can lead to more spread-out representations for a single speaker, potentially reducing the robustness of the SI system. An additional concern is that Softmax treats all classes equally, which can be problematic in SI, where the number of speakers (classes) can be very large, potentially leading to imbalanced class distributions and a bias towards more frequent speakers in the training data.

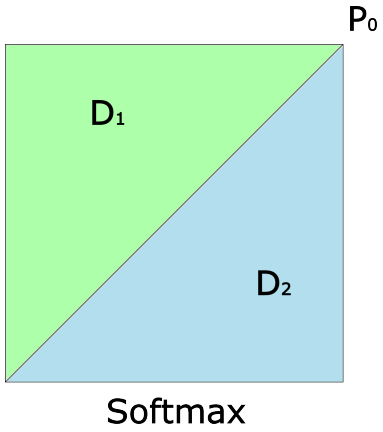

Fig. 4. Traditional Softmax's decision boundary meets at P_0. D_1 and D_2 represent class 1 and class 2, respectively.

To address these limitations, this study investigates the effectiveness of alternative loss functions: CosFace and ArcFace [26, 6]. These functions are designed to directly improve class separation and intra-class compactness, which are crucial for robust SI. The study focuses on enhancing intra-class similarities by employing these functions, which, although originally developed for facial recognition, have demonstrated significant potential for SI tasks. Both CosFace and ArcFace work by enhancing intra-class similarities through the imposition of angular or margin-based constraints [26, 25, 6].

3.4 CosFace

CosFace Loss is an enhanced version of the traditional Softmax loss, designed to improve the discriminative power of deep neural networks by imposing a fixed angular margin between classes [25]. This is achieved by modifying the target logit through the addition of a margin to the cosine similarity score of the target class.
The CosFace loss introduces an additive margin to the cosine similarity score of the target class, formulated as:

$$\psi(x) = x - m \tag{2}$$

where $x = \cos\theta_{yi}$ is the cosine similarity between the feature vector of the target class, and m is the additive margin. This formulation ensures that the decision boundary is pushed away from the target class by a fixed margin, thereby improving class separation.
The geometric interpretation of this concept is illustrated in Figure 5. While traditional Softmax loss places the decision boundary at a single hyperplane P_o (two-class problem), CosFace introduces a marginal region, shifting the boundary away from the target class. This shift is calculated as $m = (W_1 - W_2)^T P_1$. Where W_1 and W_2 are the weight vectors of the two classes, and P_1 is the decision boundary for class 1. This fixed angular margin leads to more discriminative and compact features.

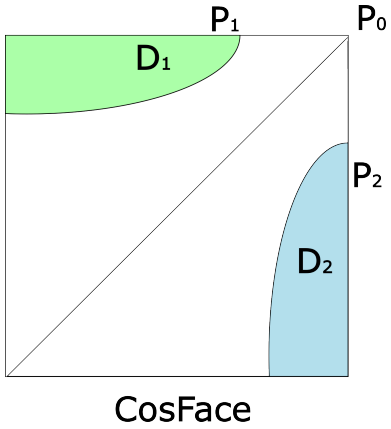

Fig. 5. A comparison between traditional Softmax's decision boundary and CosFace's decision boundary. Softmax's decision boundary meets at P_0 whereas CosFace's decision boundary for class 1 is at P_1 and for class 2 it is at P_2. D_1 and D_2 represent class 1 and class 2, respectively.

The CosFace loss function for sample x_i is mathematically defined as:

$$L_{\text{CF}} = -\log\left(\frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}\right) \tag{3}$$

where:

—s is the scaling factor, which amplifies the separation between classes;

—m is the additive margin, which enforces a fixed angular separation;

—$\cos(\theta_{y_i})$ is the cosine similarity between the features and the target weights.

For this study, $s = 22$ and $m = 0.2$, which were chosen to optimize the trade-off between class separation and model convergence.

3.5 ArcFace

The ArcFace loss function, introduced by Deng *et al.* [6], was originally used for face verification, but has seen some recent use in speaker recognition technology [8]. This function is specifically designed to optimize the geodesic distance margin on a hypersphere by introducing an additive angular margin that pushes the decision boundary further from the target class, thereby enhancing class separation.

ArcFace optimizes speaker embeddings through a dual approach: maximizing intra-class compactness while simultaneously increasing inter-class separation as seen in Figure 6. This optimization strategy has seen some success in improving speaker distinction, particularly when paired with large-scale datasets where subtle differences between speakers must be accurately captured.

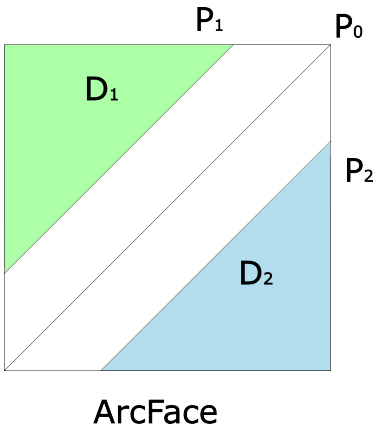

Fig. 6. A comparison between traditional Softmax's decision boundary and ArcFace's decision boundary. Softmax's decision boundary meets at P_0, whereas ArcFace's decision boundary for class 1 is at P_1 and for class 2 it is at P_2. D_1 and D_2 represent class 1 and class 2, respectively.

The integration of ArcFace with SincNet architectures [8] has enabled direct feature extraction from raw speech signals. When combined with dual attention mechanisms, this approach delivers performance improvements, especially in short-utterance scenarios where limited audio data is available for SI.

ArcFace demonstrates the capability to effectively handle difficult operational conditions, including short utterances and cross-domain mismatches. In far-field speaker verification applications, the loss function has been successfully adapted to address domain mismatches [13], with optimized margin penalties during training significantly enhancing overall system performance.

The practical advantages of ArcFace extend to various challenging environments. Models incorporating ArcFace have shown reduced error rates in short speech scenarios, making them effective for real-world applications where audio samples are typically brief. Additionally, the enhanced feature extraction and classification capabilities contribute to superior performance in noisy environments, addressing one of the most common challenges in SI systems.

Comparative studies have demonstrated that ArcFace-based models consistently achieve lower error rates than traditional methods [24], underscoring their effectiveness in SI tasks. Furthermore, the loss function's inherent ability to stabilize training processes and enhance feature learning establishes it as an invaluable tool in deep learning-based speaker recognition systems [24].

Table 2. Original VGG16 Architecture. This table details the sequential layers, kernel sizes, strides, and activation functions characteristic of the VGG16 network as described by Simonyan and Zisserman[22].

Original VGG16					
Layer	Feature Map	Size	Kernel Size	Stride	Activation
Image	1	224 x 224 x3	-	-	-
2 X Convolution	64	224 x 224 x 64	3x3	1	relu
Max Pooling	64	112 x 112 x 64	3x3	2	relu
2 X Convolution	128	112 x 112 x 128	3x3	1	relu
Max Pooling	128	56 x 56 x 128	3x3	2	relu
2 X Convolution	256	56 x 56 x 256	3x3	1	relu
Max Pooling	256	28 x 28 x 256	3x3	2	relu
3 X Convolution	512	28 x 28 x 512	3x3	1	relu
Max Pooling	512	14 x 14 x 512	3x3	2	relu
3 X Convolution	512	14 x 14 x 512	3x3	1	relu
Max Pooling	512	7 x 7 x 512	3x3	2	relu
Dense	-	25088	-	-	relu
Dense	-	4096	-	-	relu
Dense	-	4096	-	-	relu
Dense	-	1000	-	-	Softmax

For a sample (x_i, y_i), where x_i is the input sample and y_i is the target label, the mathematical formulation of the ArcFace loss function is as follows:

$$L_{\text{AF}}(x_i) = -log\left(\frac{e^{s(\cos(\theta_i+m))}}{e^{s(\cos(\theta_i+m))} + \sum_{j\neq y_i} e^{s(\cos(\theta_j))}}\right) \quad (4)$$

where:

—s is the scaling factor;

—m is the additive angular margin;

—$\cos(\theta_{y_i})$ is the angle between the feature vector and the target weights.

In this implementation, the initial scaling factor s was set to $s = 22$. The additive angular margin m was set to 0.2. This configuration was chosen to ensure more compact and discriminative features.

3.6 Recognition Model

3.6.1 VGG16. The VGG16 model is a CNN comprising 16 weight layers: 13 convolutional layers and 3 fully connected layers. These layers are organized into five blocks, each followed by a max-pooling layer to progressively reduce the spatial dimensions of the feature maps [22] as can be seen in Table 2. The core of the architecture can be broken down as follows:

—Convolutional Layers: The first two blocks contain two convolutional layers each, while the last three blocks contain three convolutional layers each. Each convolutional layer applies a 3×3 filter with stride 1 and padding to preserve spatial resolution, followed by a ReLU activation function;

—Pooling Layers: A max-pooling layer with a 2×2 filter and stride 2 is applied after each block to downsample the feature maps.

3.6.2 Modified VGG16. To enhance the model's performance and adaptability for SI, we proposed the following modifications using PyTorch, as seen in Table 3:

(1) First, the fixed size average pooling layer was replaced with a global average pooling layer to handle inputs of varying spatial dimensions.

Table 3. Proposed Modified VGG16 Architecture.

Modified VGG16					
Layer	Feature Map	Size	Kernel Size	Stride	Activation
Image	1	Any Height x Any Width x 3	-	-	-
2 X Convolution	64	224 x 224 x 64	3x3	1	relu
Max Pooling	64	112 x 112 x 64	3x3	2	relu
2 X Convolution	128	112 x 112 x 128	3x3	1	relu
Max Pooling	128	56 x 56 x 128	3x3	2	relu
2 X Convolution	256	56 x 56 x 256	3x3	1	relu
Max Pooling	256	28 x 28 x 256	3x3	2	relu
3 X Convolution	512	28 x 28 x 512	3x3	1	relu
Max Pooling	512	14 x 14 x 512	3x3	2	relu
3 X Convolution	512	14 x 14 x 512	3x3	1	relu
Global Avg Pooling 2D	-	512	N/A	N/A	-
Dense	-	1024	-	-	relu
Dropout (0.3)	-	1024	-	-	-
Dense	-	256, followed by L2 normalization	-	-	relu
Dense	-	1251 (num speakers)	-	-	Softmax

(2) The fully connected classifier was replaced with a custom classifier designed to output embeddings of a specified dimension. The output of the convolutional layers is first flattened and passed through a fully connected layer with 1024 units, followed by a ReLU activation and dropout of 0.3 for regularization. Finally, it outputs the optimal 256-dimensional embeddings through another fully connected layer [9];

(3) A classification head was added. This head is a fully connected layer that maps the embeddings to the number of classes, in this case, 1251;

(4) To ensure the embeddings are normalized to unit vectors, the model is wrapped in a custom class. During the forward pass, the embeddings are normalized using L2 normalization ($||x||^2 = 1$), because cosine similarity is used to compare embeddings.

The model parameters were optimized using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 1×10^{-4}. To adapt the learning rate during training, a ReduceLROnPlateau scheduler was employed that monitored validation loss. The scheduler reduced the learning rate by a factor of 0.1 when no improvement in validation loss was observed for 5 consecutive epochs, with a minimum improvement threshold of 1×10^{-4} to qualify as meaningful progress. An early stopping mechanism was also incorporated with a patience of 15 epochs of no improvement, while also enforcing a minimum of 30 epochs to ensure that the model had sufficient opportunity to converge, even if progress was initially slow.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Impact of audio length and looping technique

The experimental results presented in Table 4 indicate that ten-second audio clips, which utilized the looping technique, consistently achieved higher top-1 accuracy compared to three-second clips, with improvements ranging from 8.65% to 19.64% depending on the specific configuration. For example, ten-second clips achieved up to a top-1 accuracy of 80.79% when using the ArcFace loss function and 83.15% with the CosFace loss function, the traditional Softmax achieved the lowest results of the three, obtaining a top-1 accuracy of 76.41%. In contrast, the three-second clips achieved lower accuracies, ranging from 66.28% to 69.82%. This significant difference underscores the impact of audio length in capturing speaker-specific features effectively.

The superior performance aligns with previous studies that emphasize the importance of sufficient audio duration for accurate SI [17, 9]. However, in this case, the duration of the clips can be attributed to the looping technique, which allowed the model to process a more extended segment of the audio signal.

Table 4. Results of speaker identification accuracy for different loss functions, mel-spectrogram dimensions, and audio lengths. The best accuracy result is highlighted in boldface.

Loss function	Mel-Spectrogram Image Dimensions	Audio length	Top-1 accuracy
Softmax	$224 \times 224 \times 3$	3 seconds	65.76%
		10 seconds	75.41%
	$432 \times 288 \times 3$	3 seconds	67.58%
		10 seconds	**76.41%**
	$448 \times 448 \times 3$	3 seconds	67.62%
		10 seconds	75.37%
CosFace	$224 \times 224 \times 3$	3 seconds	$66,51\%$
		10 seconds	$79,42\%$
	$432 \times 288 \times 3$	3 seconds	$69,82\%$
		10 seconds	**$83,15\%$**
	$448 \times 448 \times 3$	3 seconds	$64,93\%$
		10 seconds	$79,56\%$
ArcFace	$224 \times 224 \times 3$	3 seconds	$66,28\%$
		10 seconds	**$81,33\%$**
	$432 \times 288 \times 3$	3 seconds	$68,05\%$
		10 seconds	$80,79\%$
	$448 \times 448 \times 3$	3 seconds	$64,15\%$
		10 seconds	$79,32\%$

4.2 Effect of Mel-Spectrogram dimensions

The experimental results demonstrate that mel-spectrogram dimensions significantly influence top-1 accuracy in SI. The $432 \times 288 \times 3$ configuration consistently achieved the highest accuracy at 83.15%, followed by $224 \times 224 \times 3$ and $448 \times 448 \times 3$.

This performance variation suggests that both resolution and aspect ratio are critical for capturing spectral and temporal characteristics that distinguish speakers. The $432 \times 288 \times 3$ configuration likely provides an optimal balance for the CNN to learn speaker-specific patterns. Notably, larger dimensions do not guarantee better performance; rather, dimensions must be optimized to suit the data structure for maximum accuracy.

4.3 Impact of loss functions

CosFace demonstrated clear superiority over both ArcFace and traditional Softmax across most configurations. CosFace achieved the highest identification accuracy of 83.15%, representing a 2.36 percentage point improvement over ArcFace's best performance (80.79%) and a 6.74 percentage point advantage over traditional Softmax (76.41%) when using optimal conditions.

The superior performance of CosFace can be attributed to its ability to enforce a fixed angular margin combined with a scaling factor, which creates more discriminative decision boundaries and pushes speaker embeddings further apart in the feature space.

Unlike traditional Softmax that fails to explicitly optimize for class separation, CosFace addresses fundamental limitations by directly enhancing intra-class compactness while increasing inter-class diversity.

While ArcFace also incorporates margin-based constraints through geodesic distance optimization, it demonstrated more variability across configurations. Notably, ArcFace achieved its peak performance of 81.33% with $244 \times 244 \times 3$ dimensions and ten-second clips, but consistently underperformed compared to CosFace when using the optimal $432 \times 288 \times 3$ configuration.

5. CONCLUSION

The study comprehensively achieved all stated objectives.

—The study successfully evaluated the comparative effectiveness of the ArcFace and CosFace loss functions against traditional Softmax loss and identified their superior performance when used with the modified VGG16 architecture.

—An audio looping technique was developed and validated that effectively extends shorter audio samples, proving that temporal information enhancement can overcome duration limitations in practical applications.

—Optimal mel-spectrogram dimensions (432×288×3) were systematically identified that balance computational efficiency with identification accuracy.

5.1 Practical Implications and Applications

The enhanced accuracy and robustness achieved through this approach could benefit security systems requiring reliable voice authentication, improve virtual assistant personalization capabilities, and support forensic voice analysis applications. The audio looping technique particularly addresses practical challenges where only short utterances are available, making the system more applicable to real-world scenarios where extended speech samples are not always obtainable.

5.2 Study Limitations and Considerations

It is important to contextualize objectives and findings of the present study within its intended research scope. This investigation was not designed as an attempt to achieve state-of-the-art performance on the VoxCeleb1 dataset, but rather as a systematic exploration of audio looping behavior across different system configurations. The primary research objective centered on understanding how temporal augmentation through audio looping interacts with various loss functions and input dimensions, providing foundational insights for future optimization efforts.

The experimental design prioritized comprehensive configuration space exploration over performance maximization. By systematically evaluating audio looping across three distinct loss functions (Softmax, ArcFace, CosFace), multiple mel-spectrogram dimensions ($224 \times 224 \times 3$, $448 \times 448 \times 3$, $432 \times 288 \times 3$), and two temporal durations (3-second, 10-second), the study established a controlled framework for understanding the behavioral characteristics of the proposed augmentation method. This systematic approach enables reliable conclusions about the relative effectiveness of different configurations and provides guidance for practitioners considering similar approaches.

The choice to maintain consistent architectural and preprocessing approaches across all experiments, while potentially limiting absolute performance, ensures that observed differences can be attributed to the specific variables under investigation rather than confounding factors.

The comparison with existing literature serves as a contextual framework for understanding where audio looping fits within the broader spectrum of augmentation strategies. The competitive performance achieved relative to some established methods (notably matching Yadav and Rai's 83.5% within 0.35% [27]) demonstrates that audio looping represents a viable augmentation approach worthy of further development.

The identification of optimal input dimensions ($432 \times 288 \times 3$) and the superior performance of CosFace over ArcFace in this context represent practical contributions that can inform future system design decisions, even as absolute performance optimization remains a goal for subsequent research phases.

5.3 Future Research Directions

The audio looping technique warrants systematic investigation to establish optimal operational parameters. Questions include determining the minimum viable audio duration for effective looping and the maximum extension length before performance plateaus. Future research should explore alternative looping strategies beyond simple repetition, such as partial overlap looping or intelligent looping that prioritizes high-energy audio segments. Additionally, adaptive algorithms that adjust looping patterns based on spectral content analysis could yield superior results compared to full-segment repetition.

Combining audio looping with established techniques could address multiple system limitations simultaneously. Research should investigate optimal combinations of audio looping with random sampling for class balancing, noise injection for robustness, and other modifications. Adaptive systems that select augmentation strategies based on per-speaker data availability represent a particularly promising direction for addressing both class imbalance and temporal constraints within unified frameworks.

The findings regarding network depth suggest systematic architectural investigation could yield improvements. Future research should explore modern architectures—including transformers and efficient convolutional designs—specifically combined with audio looping. Additionally, developing networks explicitly designed to leverage the temporal patterns created by looping could achieve superior performance compared to standard architectures applied to extended inputs.

Systematic evaluation across datasets featuring different languages, acoustic conditions, and speaker demographics would establish the method's robustness and identify potential limitations. Cross-linguistic studies examining effectiveness across different phonetic structures and prosodic patterns are particularly important for understanding universal applicability.

Extension to open-set SI scenarios would broaden practical applicability, though this requires investigating how temporal extension affects unknown speaker rejection capabilities. The development of threshold optimization strategies and uncertainty quantification techniques specifically designed for looped audio inputs could address deployment challenges while maintaining identification benefits.

5.4 Concluding Remarks

In conclusion, this research provides a systematic evaluation of margin-based loss functions in conjunction with an audio looping data augmentation strategy, within the framework of CNN-based SI. Achieving a top-1 accuracy of 83.15% on the VoxCeleb1 dataset, the system delivers competitive performance within the existing literature range. The primary value of this work lies in the proposed audio looping technique which, despite its simplicity, offers a practical and easily implementable solution for extending limited audio samples. This method is particularly well-suited for resource-constrained applications, and its integration with advanced loss functions demonstrates a viable pathway for improving SI accuracy in scenarios with short-duration recordings.

6. REFERENCES

[1] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 730–747, 2021.

[2] Nguyen Nang An, Nguyen Quang Thanh, and Yanbing Liu. Deep cnns with self-attention for speaker identification. *IEEE access*, 7:85327–85337, 2019.

[3] Prashant Anand, Ajeet Kumar Singh, Siddharth Srivastava, and Brejesh Lall. Few shot speaker recognition using deep neural networks, 2019.

[4] Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, and Sung Wook Baik. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78:5571–5589, 2019.

[5] Joseph P Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 2002.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[7] Sadaoki Furui. Speaker recognition in smart environments. In *Human-centric interfaces for ambient intelligence*, pages 163–184. Elsevier, 2010.

[8] Miao Guo, Jiaxiong Yang, and Shu Gao. Speaker recognition method for short utterance. In *Journal of physics: conference series*, volume 1827, page 012158. IOP Publishing, 2021.

[9] Mahdi Hajibabaei and Dengxin Dai. Unified hypersphere embedding for speaker recognition. *arXiv preprint arXiv:1807.08312*, 2018.

[10] Biing Hwang Juang, M Mohan Sondhi, and Lawrence R Rabiner. Digital speech processing. 2003.

[11] SM Kamruzzaman, ANM Karim, Md Saiful Islam, and Md Emdadul Haque. Speaker identification using mfcc-domain support vector machine. *arXiv preprint arXiv:1009.4972*, 2010.

[12] Wondimu Lambamo, Ramasamy Srinivasagan, and Worku Jifara. Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition. *applied sciences*, 13(1):569, 2022.

[13] Yuke Lin, Xiaoyi Qin, and Ming Li. Cross-domain arcface: Learnging robust speaker representation under the far-field speaker verification. In *Proc. FFSVC 2022*, pages 6–9, 2022.

[14] Xugang Lu and Jianwu Dang. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech communication*, 50(4):312–322, 2008.

[15] André Filipe da Silva Magalhães et al. Voice recognition of users for virtual assistant in industrial environments. Master's thesis, 2021.

[16] DAMIAN A MORANDI. Effect of pitch modification on the voice identification of the speakers.

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[18] Kristiawan Nugroho, Edi Noersasongko, et al. Enhanced indonesian ethnic speaker recognition using data augmentation deep neural network. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4375–4384, 2022.

[19] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana. Deep neural network framework and transformed mfccs for speaker's age and gender classification. *Knowledge-Based Systems*, 115:5–14, 2017.

[20] Banala Saritha, Mohammad Azharuddin Laskar, Anish Monsley Kirupakaran, Rabul Hussain Laskar, Madhuchhanda Choudhury, and Nirupam Shome. Deep learning-based end-to-end speaker identification using time–frequency representation of speech signal. *Circuits, Systems, and Signal Processing*, 43(3):1839–1861, 2024.

[21] M Sharif-Noughabi, S Razavi, and S Mohamadzadeh. Improving the performance of speaker recognition system using optimized vgg convolutional neural network and data augmentation. *International Journal of Engineering*, 38(10):2414–2425, 2025.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] MS Sinith, Anoop Salim, K Gowri Sankar, KV Sandeep Narayanan, and Vishnu Soman. A novel method for text-independent speaker identification using mfcc and gmm. In *2010 International Conference on Audio, Language and Image Processing*, pages 292–296. IEEE, 2010.

[24] Yuwu Tang, Ying Hu, Liang He, and Hao Huang. A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition. *Speech Communication*, 143:21–32, 2022.

[25] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[26] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[27] Sarthak Yadav and Atul Rai. Learning discriminative features for speaker identification and verification. In *Interspeech*, pages 2237–2241, 2018.

[28] Youssef Zouhir, Mohamed Zarka, Kaïs Ouni, and Lilia El Amraoui. Power wavelet cepstral coefficients (pwcc): An accurate auditory model-based feature extraction method for robust speaker recognition. *IEEE Access*, 2025.