Computing, Telecommunication and Control, 2025, Vol. 18, No. 1, Pp. 60–71.
Информатика, телекоммуникации и управление. 2025. Том 18, № 1. С. 60–71.

Circuits and Systems for Receiving, Transmitting, and Signal Processing

Устройства и системы передачи, приема и обработки сигналов

Research article
DOI: https://doi.org/10.18721/JCSTCS.18105
UDC 004.89

RESNET-SV: FAST AND ACCURATE SPEAKER VERIFICATION WITH A MULTI-LAYER CASCADE ATTENTION MECHANISM

A.A. Aliyev ✉ , *S.A. Molodyakov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ aliev.aa@edu.spbstu.ru

Abstract. One of the most challenging issues of voice biometrics rapid development is the need to develop methods that can combine speed and accuracy. Traditional solutions tend to choose a compromise between these two aspects, which either complicates the speaker verification process or reduces accuracy, especially under real-world conditions in which background noise and fluctuation in speech are substantial obstacles. This paper examines modern approaches and their architectural features. The architecture is based on ResNet, originally designed for computer vision tasks, which was modified and adapted for optimal performance in speech processing. The proposed modification method based on a multi-layer cascade attention mechanism for feature extraction from convolutional blocks is described in detail. This modification allows using fewer layers for feature extraction, thereby increasing the speed of the model, and allows to deal more effectively with the noise in the audio signal. The paper concludes with the model parameters used in the training process, as well as key metrics such as EER and minDCF computed on the VoxCeleb1 dataset. The results are compared with solutions built on other architectures. Through experimentation, the authors were able to achieve a high level of accuracy, with a smaller number of the neural network model parameters. This work brings us closer to a wider application of voice biometric systems in various scenarios.

Keywords: speaker verification, speaker identification, voice biometrics, convolutional neural networks, attention mechanism, speech processing

Citation: Aliyev A.A., Molodyakov S.A. ResNet-SV: Fast and accurate speaker verification with a multi-layer cascade attention mechanism. Computing, Telecommunications and Control, 2025, Vol. 18, No. 1, Pp. 60–71. DOI: 10.18721/JCSTCS.18105

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18105
УДК 004.89

RESNET-SV: БЫСТРАЯ И ТОЧНАЯ ВЕРИФИКАЦИЯ СПИКЕРА С ИСПОЛЬЗОВАНИЕМ МНОГОУРОВНЕВОГО КАСКАДНОГО МЕХАНИЗМА ВНИМАНИЯ

А.А. Алиев ✉ , *С.А. Молодяков*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ aliev.aa@edu.spbstu.ru

Аннотация. Одной из самых сложных проблем быстрого развития голосовой биометрии является необходимость разработки методов, способных сочетать скорость и точность. Традиционные решения, как правило, выбирают компромисс между этими двумя аспектами, что приводит либо к усложнению процесса верификации спикеров, либо к снижению точности, особенно в реальных условиях, когда фоновый шум и колебания речи являются существенными препятствиями. В данной статье рассматриваются современные подходы и их архитектурные особенности. Основой для разработки архитектуры послужила ResNet, изначально предназначенная для задач компьютерного зрения, которая была модифицирована и адаптирована для оптимальной работы в области обработки речи. Подробно описывается предложенный метод модификации на основе многослойного каскадного механизма внимания для извлечения признаков из сверточных блоков. Такая модификация позволяет использовать меньшее количество слоев для извлечения признаков, тем самым увеличивая скорость работы модели, а также позволяет более эффективно бороться с возникшими шумами в аудиосигнале. В заключении статьи представлены параметры модели, использованные в процессе обучения, а также ключевые метрики, такие как EER и minDCF, рассчитанные на выборке данных VoxCeleb1. Результаты сравниваются с решениями, построенными на других архитектурах. В ходе экспериментов авторам удалось достичь высокого уровня точности при меньшем количестве параметров модели нейронной сети. Эта работа приближает нас к более широкому применению систем голосовой биометрии в различных сценариях.

Ключевые слова: верификация спикеров, идентификация спикеров, голосовая биометрия, сверточные нейронные сети, механизм внимания, обработка речи

Для цитирования: Aliyev A.A., Molodyakov S.A. ResNet-SV: Fast and accurate speaker verification with a multi-layer cascade attention mechanism // Computing, Telecommunications and Control. 2025. Т. 18, № 1. С. 60–71. DOI: 10.18721/JCSTCS.18105

Introduction

Speaker verification is the core of the authentication process for many applications: security systems, access control, financial transactions, virtual assistants, etc. It is a way to identify a person using his or her voice, which is unobtrusive and natural. Its importance has increased with the digital revolution, with the increased need for highly secure, reliable and user-friendly mechanisms of authentication. These speaker verification systems use the personal characteristics of an individual's voice. They happen to be much more effective in authenticating access than the traditional passwords or PIN codes, which can easily be stolen or forged.

Although, despite the numerous advances in speaker verification technology, there still are quite many difficulties [1] that affect their effectiveness and wider application. The main issue is the variability of the human voice, which is affected by various conditions – from illness and emotional state to aging. The human voice is also affected by environmental conditions – background noise and acoustic environment of the room. Such variabilities can greatly affect the ability of the system to accurately

recognize the speaker's voice; this again leads to an increase on error rate. It also depends on trade-off between accuracy and speed, which is an extremely important issue: faster systems decrease accuracy, while ultra-precise systems may not work in real-time conditions for many applications.

The purpose of this work, therefore, is to address the above-stated challenges by exploring the potential of different convolutional neural network (CNN) architectures [2] to improve the speed and accuracy of the speaker verification systems. CNNs, therefore, provide state-of-the-art performance over a wide range of deep learning architectures, since they allow gradients to propagate efficiently through a deep network. With all this in view, we tend to develop a speaker verification system based on these architectures that can achieve high accuracy in the face of voice variability and environmental noise, and operate at a speed sufficient for the system to be used in real-time applications.

Therefore, this paper makes two main contributions to the field of speaker verification. The first one is introducing a new architecture based on the best practices from different CNN architectures, designed to fit speaker verification problems. The second one is bringing state-of-the-art advances in deep learning techniques, such as feature aggregation techniques with attention mechanisms, to better extract and process voice characteristics. We also conduct rigorous testing of the system to demonstrate that it outperforms the existing benchmarks in both accuracy and speed. Finally, the study provides valuable insights into the application of CNNs in speech processing tasks, hence opening a way for further research and development in this area.

Related Works

Survey of existing methods

Speaker verification technologies have come a long way from traditional to very advanced deep learning-based approaches. Firstly, the traditional methodologies, Gaussian Mixture Models (GMM) [3] and Hidden Markov Models (HMM) [4] approaches led to statistical modeling of voice characteristics. These methods mostly used hand-crafted features, such as Mel-Frequency Cepstral Coefficients (MFCCs) [5], designed to characterize the speaker's voice. With the enormous success of deep learning, neural network-driven methods have managed to revolutionize the speaker verification landscape. Most of them use three types of models: Deep Neural Networks (DNN), CNNs, and Recurrent Neural Networks (RNN), which have proven their ability to automatically learn hierarchical representations of raw audio data. Later Long Short-Term Memory (LSTM) [6] networks emerged, and more recently – attention mechanisms to better capture the temporal dynamics and dependencies in speech data.

Reference architectures in speaker verification

To benchmark the performance of our ResNet-SV architecture, we compared it with several established models in speaker verification:

• **Wav-LM** [7]: it is the state-of-the-art speech model developed by Microsoft, generally used for various types of speech recognition and understanding. It leverages the transformer architecture, proven quite effective for a wide range of applications demanding natural language processing. Wav-LM operates directly on raw audio waveforms with self-supervised training on massive amounts of unlabeled audio data. This way, the model gets to learn powerful speech representations useful for multiple downstream tasks, such as speech transcription, speaker identification, and emotion detection. The model demonstrates robustness, showing good performance over different accents and in varied speech scenarios.

• **ECAPA-TDNN** (Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network-Based Speaker Verification) [8]: it advances the TDNN architecture, including channel-wise attention mechanisms and methods for improving feature aggregation. The model delivers an exceptionally high performance in the extraction of the fine-grained speaker characteristics from complex audio input and thus provides an extremely useful application for such biometric authentication and forensics areas. ECAPA-TDNN is particularly noteworthy in low-resource and challenging acoustic settings.

• **Deep Speaker Recognition (ResNet-50)** [9]: a ResNet-based model that extracts speaker embedding directly from spectrograms for verification. This model, like ours, is based on the original ResNet model, but without the improvements, we made in this study, and is a direct competitor to our model.

Thus, these are baseline models, both traditional and deep learning-based, and as such they represent a relatively comprehensive overview of current methods and best practices in speaker verification.

Limitations

Despite these remarkable achievements, current speaker verification techniques have some limitations. Classical techniques perform quite well, but fail to cope with high dimensionality of the speech data and usually provide much lower accuracy than deep learning models. Another challenge with deep learning-based methods is that they are computationally inefficient. This is especially true for deeper models, which require enormous resources for training and computation. In addition, the accuracy of many other models is reduced by issues, such as background noise, variations in emotional state, and speech irregularities associated with illness. Some of them suffer from overtraining due to the peculiarities of training pipelines and architectures. Therefore, there is an obvious interest in building more robust and efficient models that can maintain high accuracy even in the most challenging situations.

Rationale for CNNs

CNNs have shown the potential to address such limitations observed in current speaker verification approaches. First developed for the task of image recognition, very deep CNNs like ResNet have shown success in varied domains of deep learning due to an innovative architectural feature that allowed training networks hundreds of layers deep. The main idea of ResNet is the learning through shortcut connections, allowing the gradients to flow across the network without any obstructions, hence avoiding the vanishing gradient problem, that most deeper architectures face. This increases not only the learning capacity of the model, but also its computational efficiency, since this allows learning more representations without significant growth in the usage of computational resources. ResNet shows great promise for speaker verification, since it is able to capture the temporal dynamics and features of normal to complex voice signals, even under adverse conditions, with greater accuracy than other architectures. Besides, their efficiency and scalability would make them to be used in the cryptographic scheme for real-time verification applications, an area where there is a pressing need. The proven success in various domains and unique strengths of the explored application of CNNs in the field of speaker verification make it a quite logical and very promising way to overcome the existing problems.

Materials and methods

Architecture overview

The strategic architectural decisions in the proposed convolutional network architecture for speaker verification should be aimed at achieving the best trade-off between speed and accuracy. Thus, the architecture developed in this work is supposed to be a ResNet structure with modified ResNet blocks, each containing two main paths: the main learning path and the shortcut connection. At the convolutional layers, batch normalization is used, and then ReLU is implemented as an activation function to enable the learning of non-linearity. Both convolutional layers are set to learn and extract features of the input signal at various abstraction levels, and the shortcut connections help the gradients flow directly without facing the problem of vanishing gradients.

As you can see from Fig. 1, the architecture consists of the following parts:

1. **Input**: Starts with a raw audio waveform, which is transformed into a Mel spectrogram using F-Banks (Filter Banks) [10] features.

2. **ResNet-SV Blocks with Convolutional Layers**: A series of 3x3 and 1x1 convolutional layers with increasing feature maps are applied, capturing hierarchical audio features. We have reduced the number of channels in convolutional layers, unlike the original ResNet or Deep Speaker Recognition ResNet.

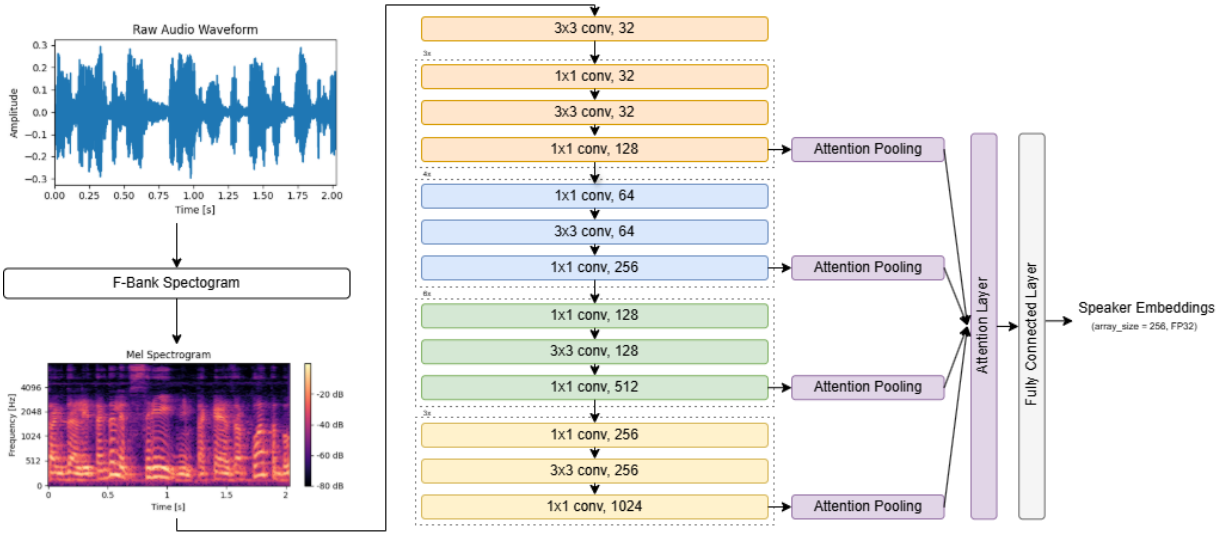

Fig. 1. ResNet-SV architecture

3. **Attention Pooling**: Each convolutional block is followed by an attention [11] pooling layer to focus on important parts of the feature maps.

4. **Attention Layer**: Outputs from all attention-pooled layers are merged.

5. **Fully Connected Layer**: Finally, a fully connected layer generates 256-dimensional speaker embedding as the output.

We use features from different blocks of our convolutional architecture to be able to capture different levels of abstraction of audio features, the upper layers capture more complex audio data, while the lower layers often display more basic information. A more detailed architecture of ResNet-SV blocks is shown in Fig. 2.

As can be seen from Table 1, unlike the original ResNet paper [9], we have reduced the number of channels in the convolutional layers by a factor of two, which naturally had a strong impact on the learning and inference speed, while not significantly affecting the quality of the neural network due to the way the architecture is structured.

As in other similar ResNet architectures, we can increase the size of our network by increasing the number of ResNet-SV blocks, some of them are shown in Table 1.

One of the significant improvements of ResNet-SV is the inclusion of an attention mechanism [12] between the convolutional blocks, designed specifically to make the model better focus on those salient features that are critical in between-speaker discrimination. The context-aware attention module is used here to dynamically weight between diverse region importance of the input feature map. The mechanism of our attention pooling layers is as follows:

1. First, we calculate α (attention) for the original input feature map (x_i). We use two Conv1D layers with tanh activation function. Here we take tanh function activation instead of ReLU, because tanh function converges better and faster. After this, we calculate SoftMax function values for output.

$$\alpha = \sigma\left(\mathrm{Conv1D}\left(\tanh\left(\mathrm{Conv1D}\left(x_i\right)\right)\right)\right);$$

$$\sigma\left(x\right)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}.$$

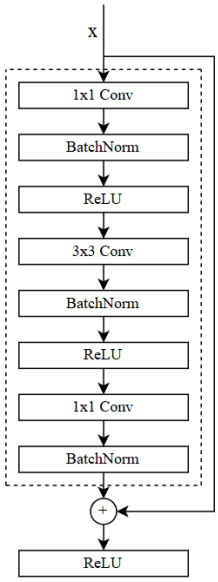

Fig. 2. ResNet-SV block structure

Table 1

ResnetSV architecture

Layer name	Output size	ResNet-SV-50	ResNet-SV-101	ResNet-SV-152
conv1	T × 80 × 32	3 × 3.32		
ResNet-SV-Block (resnet-sv_conv2_x)	T × 80 × 128	$\begin{bmatrix} 1\times1.32 \\ 3\times3.32 \\ 1\times1.128 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1.32 \\ 3\times3.32 \\ 1\times1.128 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1.32 \\ 3\times3.32 \\ 1\times1.128 \end{bmatrix}\times3$
ResNet-SV-Block (resnet-sv_conv3_x)	T/2 × 40 × 256	$\begin{bmatrix} 1\times1.64 \\ 3\times3.64 \\ 1\times1.256 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1.64 \\ 3\times3.64 \\ 1\times1.256 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1.64 \\ 3\times3.64 \\ 1\times1.256 \end{bmatrix}\times4$
ResNet-SV-Block (resnet-sv_conv4_x)	T/4 × 20 × 512	$\begin{bmatrix} 1\times1.128 \\ 3\times3.128 \\ 1\times1.512 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1.128 \\ 3\times3.128 \\ 1\times1.512 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1.128 \\ 3\times3.128 \\ 1\times1.512 \end{bmatrix}\times6$
ResNet-SV-Block (resnet-sv_conv5_x)	T/8 × 10 × 1024	$\begin{bmatrix} 1\times1.256 \\ 3\times3.256 \\ 1\times1.1024 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1.256 \\ 3\times3.256 \\ 1\times1.1024 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1.256 \\ 3\times3.256 \\ 1\times1.1024 \end{bmatrix}\times3$
Attention Pooling (atn_conv2_x)	20480	–		
Attention Pooling (atn_conv3_x)	20480	–		
Attention Pooling (atn_conv4_x)	20480	–		
Attention Pooling (atn_conv5_x)	20480	–		
Attention	81920	–		
Fully Connected	256	–		

2. Then we calculate the mean and variance [13], taking into account the weights from the attention layer.

$$\text{mean} = \frac{\sum_{i=1}^{T}(\alpha_i \cdot x_i)}{T};$$

$$\text{var} = \frac{\sum_{i=1}^{T}(\alpha_i \cdot x_i - \text{mean})}{T}.$$

3. Then we calculate the standard deviation based on the variance.

$$\text{std} = \sqrt{\text{var}}.$$

We perform these calculations for each feature, in our case $F_{num} = 10240$. Since we count mean and std for each feature, we get $F_{num} = 2 \times 10240 = 20480$ features. We use the attention layers in this way after each of our ResNet blocks. At the end we use the obtained arrays of mean and std values, combine them and feed them further to the next layer of attention.

In order to improve robustness, especially in noisy conditions, a noise-aware training strategy is implemented into the model. The general approach for improved invariance is to augment training data with different types of background noise. This, in turn, increases invariance and helps the network learn more discriminative features in noisy environments.

Data preprocessing

Below are the pre-processing stages that the pipeline undergoes in preparing the raw audio data for effective learning.

In the first stage, we augment the data with a speed perturbation and a volume change. In addition, we use various noise and other signals from datasets such as RIRS [14] and MUSAN [15]. Each of these types of augmentation is applied to each example with a probability of 0.6. This approach allows us to secure our neural network from overfitting and improves performance under challenging conditions.

Since raw audio contains too much data that we often do not need, we need to convert raw audio into a more meaningful data type. In speech processing tasks, spectrograms are often used. There are several ways to create a spectrogram. We decided to use filter banks (F-banks). In the case of audio signal processing, F-banks serve to segment the signals into shorter, overlapping frames to capture the dynamic nature of speech more effectively. F-banks apply a set of band-pass filters corresponding to the critical bands of human hearing to each frame to obtain a representation with bands representative of specific frequency ranges. This follows close to the auditory scale of human perception in every frame. These F-bank energies have a normal zero-mean and unit variance across all frames. Normalization acts as a standardization method that reduces influence due to varying amplitudes of signals on the performance of the model, thus assuring that the focus is on the spectral characteristics of the audio rather than on loudness.

We used two second segments, the frame size of which was 25 ms with a 10 ms step. Taking into account the input segment size and step size, we get the final input tensor as (B, 80, 200), where B is the batch size.

Training procedure

ResNet-SV is trained on curated data that collects several publicly available corpora, providing variations in accent, dialect, and recording conditions. The training, validation, and test splits of the dataset are kept independent of each other, preserving speaker independence from one split to another, except that data values do not creep through splits.

The AAM-Softmax [16] loss is applied to the model, which helps keep high classification accuracy value. We use the Adam [17] optimizer with adaptive learning rate features and fast convergence. The

parameters are trained using a learning rate initialized at 0.001 and have a schedule decay. The batch size is 64 and the epochs are set to a maximum of 100, with early stopping based on the validation loss to curb overfitting.

Evaluation metrics

We evaluate ResNet-SV using two core metrics that have seen wide use in the speaker verification community – **Equal Error Rate** (**EER**) [18] and the **minimum Detection Cost Function** (**minDCF**) [19]. The EER point is the point, at which the false acceptance rate is exactly equal to the false rejection rate and points out equal errors for both, which can be taken as the overall error rate of the system. The minDCF is one more factor that has been implemented to correct the decision threshold in the verification cost evaluation. The minDCF is found by searching for the particular system configuration's cost that minimizes the detection cost function. These metrics provide holistic views on the effectiveness and practical utility of the model used for speaker verification tasks.

The EER itself cannot be represented as a mathematical formula because it refers to the point of intersection of two errors:

$$FAR(t) = FRR(t),$$

where $FAR(t)$ is the rate, at which impostor attempts are incorrectly accepted above the threshold t; $FRR(t)$ is the rate, at which genuine attempts are incorrectly rejected below the threshold t.

The minDCF formula is as follows:

$$C_{det}(t) = C_{miss} \times P_{miss}(t) \times P_{target} + C_{fa} \times P_{fa}(t) \times (1 - P_{target}),$$

where C_{miss} is the cost of a miss (false rejection); C_{fa} is the cost of a false alarm (false acceptance); $P_{miss}(t)$ is the probability of a miss (false rejection rate) at threshold t; $P_{fa}(t)$ is the probability of a false alarm (false acceptance rate) at threshold t; P_{target} is the a priori probability of the speaker being the target (i.e., a genuine user).

The minDCF is then calculated by finding the value of t that minimizes $C_{det}(t)$.

Experiments

Datasets

In this study, the VoxCeleb [20] datasets have been used for training and testing the proposed ResNet-SV model. This dataset is one of the best in benchmarking for speaker verification.

Training dataset VoxCeleb2 contains over a million utterances from 6112 celebrities collated from YouTube video sources. It represents a full range of accents, languages, and acoustic conditions, making it perfect for training very vibrant models in speaker verification. The quantity and diversity of VoxCeleb2 makes it possible to train deep neural networks, which means that models trained on it generalize well to populations of different speakers and conditions.

Testing Dataset VoxCeleb1 includes more than 100000 utterances of 1251 celebrities. This dataset shares the same characteristics with VoxCeleb2, except that the speakers are different, so it provides a rigorous test to the model's generalization ability on unseen data. This is a standard dataset type in speaker-verification research tasks, given that VoxCeleb1 is intended for testing and hence for direct comparison with earlier works.

These VoxCeleb datasets, containing real-world complexity, diversity, natural noise, and other variability, are a very suitable domain for evaluating the efficacy of the proposed noise-robust convolutional network architecture. We also chose the VoxCeleb dataset, as it is a global benchmark standard for speaker verification tasks and allows us to more easily compare our work with other works and simplifies the replication of our work.

Experimental setup

Experiments are carried out on an NVIDIA SuperPOD A100 high-performance computing cluster node, which was equipped with 8x NVIDIA A100-SXM4-40GB graphics cards and with 2x AMD EPYC 7742 processors with 64 cores, 128 threads with 1TB of RAM.

Implementation details

For replicability, the following implementation details are provided:

Preprocessing: The audio was resampled at 16 kHz, and the feature extracted was the Mel spectrogram using F-bank features. The short-time Fourier transform (STFT) of the audio signal was used to get the window size at 25 ms with a 10 ms step.

Model Training Parameters: The model was trained using an Adam optimizer. For training and evaluation, the batch size used was 64 samples.

Evaluation Protocol: The speaker verification performance was evaluated on a standard split of the VoxCeleb1 test set, which ensures a fair comparison of the performance presented in this experiment with the baseline and other architectures.

This will give an even more robust, and at the same time, reproducible setup for other researchers.

Results and discussion

Performance evaluation

In summarizing the evaluation results of the proposed ResNet-SV architecture and comparing it with baseline models, we focus mainly on EER and minDCF.

Table 2

Speaker verification models comparisons

Model	Params	EER			minDCF		
		Vox1-O	Vox1-E	Vox1-H	Vox1-O	Vox1-E	Vox1-H
WavLM Large	316.62M	0.62	0.66	1.32	–	–	–
WavLM Base+	94.70M	0.84	0.93	1.76	–	–	–
ECAPA-TDNN (C=512)	6.2M	1.01	1.24	2.32	0.1274	0.1418	0.2181
ECAPA-TDNN (C=1024)	14.7M	0.87	1.12	2.12	0.1066	0.1318	0.2101
ResNet-50	25.6M	3.95	4.42	7.33	0.4290	0.5240	0.6730
ResNet-SV-50 (Ours)	**13.76M**	0.68	0.82	1.48	0.060	0.089	0.135
ResNet-SV-101 (Ours)	18.52M	0.62	0.73	1.36	0.055	0.078	0.127
ResNet-SV-152 (Ours)	22.45M	**0.54**	**0.68**	**1.29**	**0.049**	**0.071**	0.119

Table 2 presents performance characteristics for various speaker verification models with three test conditions of the VoxCeleb dataset: Vox1-O (original), Vox1-E (extended), and Vox1-H (hard). Each model is evaluated using EER and, where available, minDCF. The models vary in complexity, from the large-scale WavLM Large with over 300 million parameters to smaller models like the ECAPA-TDNN. Notably, the WavLM models do not report minDCF values. The data demonstrates a range of performances with some models, particularly the ResNet-SV series developed by us, showing notably lower EER and minDCF, suggesting better verification accuracy.

Moreover, in terms of performance, ResNet-SV improves remarkably with respect to the baseline models, reporting the lowest EER and minDCF over all the considered methods. This is an indication of the clear existence of a substantially higher level of accuracy and lower cost of errors in the task of speaker verification, providing further evidence of the effectiveness of our devised attention mechanisms and noise-aware training of the convolutional network architecture.

Analysis

The results indicate several key insights.

Robust to Noise: ResNet-SV performs better, hence has stronger feature extraction capabilities, where the model could focus well on salient speaker characteristics while effectively reducing the effect of background noise by a built-in attention mechanism and training strategy aware of the noise. This is especially evident in the Vox1-H dataset, where there are complex examples.

Efficiency and Scalability: Convolutional connections in the deep architecture of ResNet-SV assure efficient training and inference with real-time applications, where modern hardware requires high throughput.

Strong generalization to different speakers and environments is a key attribute in practical speaker verification systems. The improved performance on the VoxCeleb1 test set – unseen during training – provides empirical evidence of the improved generalization of ResNet-SV.

Interpretation of the results

We validate the effectiveness and efficiency of the proposed ResNet-SV architecture for speaker verification through our experimental results. With a significantly lower EER and minDCF compared to baseline models, it proves that it can indeed verify speakers more accurately, even under adverse conditions. In particular, attention mechanisms and noise-aware training strategies have improved the attention to relevant speaker characteristics and the robustness to background noise in the model.

The convolutional connections ensure the efficiency of the model. Thus, ResNet-SV quickly processes the inputs and makes verification decisions even with the much-required deep architecture in real-time applications. ResNet-SV balances high accuracy and performance efficiency, which is an important achievement for speaker verification.

Comparison with previous work

Compared to traditional models such as ResNet and newer deep learning models like Wav-LM and ECAPA-TDNN, ResNet-SV significantly improves the verification performance. Similar attempts have largely failed to achieve this goal in their approach towards restoring the degraded accuracy of noisy conditions, while achieving computational efficiency. The improved performance of ResNet-SV, in this case, is measured by lower EER and minDCF, indicating a significant improvement over these models. Additionally, the attention mechanisms embedded in the architecture of the convolutional network for the nuances of speaker verification clearly separate our work from previous works, which mainly tackled either architectural advances or feature-level improvements, without properly covering both aspects.

Limitations and challenges

Despite the promising results, a number of limitations and challenges must be acknowledged:

Computational Resources: Training and fine-tuning of ResNet-SV requires enormous computations, which might not be available in most of the research or production systems.

Model Complexity: Despite very good performance, model complexity prevents the developed architecture from being used in resource-limited environments or in applications that are sensitive to the extremely low-latency processing of information.

Generalization in Various Conditions: How well the model generalizes to such a wide range of settings – more diverse than those in VoxCeleb datasets and potentially including conditions that are more extreme – remains to be fully tested.

The design choices will inevitably lead to increased complexity in both the system and the training, which will help to expand the frontiers of finding more accurate, efficient, and robust speaker verification systems.

Future studies should aim to address these limitations, as well as discuss the strategies that might reduce the computational demand, make models more portable, and improve the generalization capability to cover a wider range of speakers and conditions.

Conclusion and further research

In this work, we proposed a new convolutional network architecture for speaker verification, named ResNet-SV, having an attention mechanism and trained with noise-awareness to ensure the accuracy and robustness of the model to environmental noise. The rigorous experimental comparison also proved that the tested ResNet-SV model achieved significantly improved performance compared to the already existing baseline models, with significantly lower EER and minDCF. The integrated attention mechanisms in the convolutional blocks played an important role in focusing on only relevant speaker characteristics, while the noise-aware training strategy further improved the resilience of the model to background noise. Those results indicate the potential applicability of deep learning, more precisely the ResNet family of CNNs, to enhance the state-of-the-art in speaker verification systems towards more accurate, efficient, and robust setups.

This is very important in that it allows us to chart the path forward for speaker verification systems that can effectively work in noisy conditions with variables found in the real world, conditions, which is known to be particularly challenging. ResNet-SV not only addresses these issues, but also sets a new standard in terms of performance and makes speaker verification technology generally more applicable to many, if not most, scenarios and environments.

REFERENCES

1. **Singh N., Agrawal A., Khan R.A.** Automatic speaker recognition: Current approaches and progress in last six decades. *Global Journal of Enterprise Information System*, 2017, Vol. 9, No. 3, Pp. 38–45. DOI: 10.18311/gjeis/2017/15973

2. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. DOI: 10.48550/arXiv.1512.03385

3. **Chakroun R., Frikha M.** Robust text-independent speaker recognition with short utterances using Gaussian mixture models. *2020 International Wireless Communications and Mobile Computing* (*IWCMC*), 2020, Pp. 2204–2209. DOI: 10.1109/IWCMC48107.2020.9148102

4. **Wei Y.** Adaptive speaker recognition based on Hidden Markov Model parameter optimization. *IEEE Access*, 2020, Vol. 8, Pp. 34942–34948. DOI: 10.1109/ACCESS.2020.2972511

5. **Ayvaz U., Gürüler H., Khan F., Ahmed N., Whangbo T., Abdusalomov A.B.** Automatic speaker recognition using Mel-Frequency Cepstral Coefficients through machine learning. *Computers, Materials & Continua*, 2022, Vol. 71, No. 3, Pp. 5511–5521. DOI: 10.32604/cmc.2022.023278

6. **Yu Y., Si X., Hu C., Zhang J.** A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 2019, Vol. 31, No. 7, Pp. 1235–1270. DOI: 10.1162/neco_a_01199

7. **Chen S., Wang C., Chen Z., Wu Y., Liu S., Chen Z. et al.** WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022, Vol. 16, No. 6, Pp. 1505–1518. DOI: 10.1109/JSTSP.2022.3188113

8. **Desplanques B., Thienpondt J., Demuynck K.** ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification. *arXiv:2005.07143*, 2020. DOI: 10.48550/arXiv.2005.07143

9. **Chung J.S., Nagrani A., Zisserman A.** VoxCeleb2: Deep speaker recognition. *arXiv:1806.05622*, 2018. DOI: 10.48550/arXiv.1806.05622

10. **Wu Y.-P., Mao J.-M., Li W.-F.** Robust speech recognition by selecting mel-filter banks. *Advances in Engineering Research* (*AER*), 2016, Vol. 117, Pp. 407–416.

11. **Wijayasingha L., Stankovic J.A.** Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health*, 2021, Vol. 19, Art. no. 100165. DOI: 10.1016/j.smhl.2020.100165

12. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.** Attention is all you need. *arXiv:1706.03762*, 2017. DOI: 10.48550/arXiv.1706.03762

13. **Snyder D., Garcia-Romero D., Povey D., Khudanpur S.** Deep neural network embeddings for text-independent speaker verification. *Proceedings of Interspeech*, 2017, Pp. 999–1003. DOI: 10.21437/Interspeech.2017-620

14. **Snyder D., Chen G., Povey D.** MUSAN: A music, speech, and noise corpus. *arXiv:1510.08484*, 2015. DOI: 10.48550/arXiv.1510.08484

15. **Ko T., Peddinti V., Povey D., Seltzer M.L., Khudanpur S.** A study on data augmentation of reverberant speech for robust speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2017, Pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152

16. **Deng J., Guo J., Xue N., Zafeiriou S.** ArcFace: Additive Angular Margin Loss for deep face recognition. *arXiv:1801.07698*, 2018. DOI: 10.48550/arXiv.1801.07698

17. **Kingma D.P., Ba J.** Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014. DOI: 10.48550/arXiv.1412.6980

18. **Thian N.P.H., Bengio S.** *Evidences of equal error rate reduction in biometric authentication fusion*. Switzerland: IDIAP, 2004. 27 p.

19. **Scheffer N., Ferrer L., Graciarena M., Kajarekar S., Shriberg E., Stolcke A.** The SRI NIST 2010 speaker recognition evaluation system. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2011, Pp. 5292–5295. DOI: 10.1109/ICASSP.2011.5947552

20. **Nagrani A., Chung J.S., Zisserman A.** VoxCeleb: a large-scale speaker identification dataset. *arXiv:1706.08612*, 2017. DOI: 10.48550/arXiv.1706.08612

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Aliyev Ali A.
Алиев Али Ахмед оглы
E-mail: aliev.aa@edu.spbstu.ru
ORCID: https://orcid.org/0000-0002-2813-2676

Molodyakov Sergey A.
Молодяков Сергей Александрович
E-mail: samolodyakov@mail.ru
ORCID: https://orcid.org/0000-0003-2191-9449

Submitted: 24.11.2024; Approved: 22.01.2025; Accepted: 27.01.2025.
Поступила: 24.11.2024; Одобрена: 22.01.2025; Принята: 27.01.2025.