National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 4, 2025 (pp. 26-33)

RESEARCH ARTICLE

ECEJOURNALS.IN

Self-Supervised Audio Representation Learning for Robust Speaker Verification

F. de Mindonça[1]*, O.L.M. Smith[2]

[1]*Departamento de Engenharia Elétrica, Universidade Federal de Pernambuco - UFPE Recife, Brazil*
[2]*Departamento de Engenharia Elétrica, Universidade Federal de Pernambuco - UFPE Recife, Brazil*

KEYWORDS:
Self-Supervised Learning (SSL),
Speaker Verification,
Contrastive Learning,
Audio Representation Learning,
Voice Biometrics,
Deep Neural Networks,
Transformer Encoder,
Equal Error Rate (EER),
Noise Robustness,
Unlabeled Audio Data

ARTICLE HISTORY:
Submitted : 09.05.2025
Revised : 13.06.2025
Accepted : 22.07.2025

https://doi.org/10.17051/NJSAP/01.04.04

Abstract

A new self-supervised learning (SSL) framework of robust speaker verification capable of overcoming the drawbacks of the outdated models of supervised systems, especially when dealing with noisy and domain-mismatched data are introduced in this paper. Experimenting with the large number of unlabeled audio data we use a contrastive learning paradigm to learn extremely discriminative, speaker-specific embeddings with no need to resort to explicit identity labels. This framework includes a convolutional encoder and a transformer-based context network trained over temporally augmented segments of audio to drive the model to learn invariant features to time, noise, and signal Newer to the model. To be more specific, positive updates are done by dynamic telescopic masking, cropping, and augmentation strategies, whereas negative ones are sampled over the batch, thus allowing the system to gain subtle discrimination over various speakers. In contrast to prior SSL frameworks most directly applicable to automatic speech recognition (ASR), our model is explicitly optimized to learn a feature representation that is useful to speaker verification by adding an embedding-level projection and fine-tuning stage that tunes the learnt representations so as to be relevant to speaker identities. Our experiments are performed at large scale on public datasets of speaker verification, such as VoxCeleb1 and VoxCeleb2, both under clean and in the noisy environment. Our approach continually has lower Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) than supervised baselines, such as x-vectors as well as new SSL models, such as wav2vec 2.0 and HuBERT. In addition, the model has high generalization abilities on domain shifts and also retains high performance at low signal-to-noise ratio (SNR) which renders it very appropriate to be applied in real life applications in secure authentication tasks, forensic applications and also in telecommunications. These findings show that self-supervised contrastive learning can be significantly more effective at improving the state-of-the-art in speaker verification with speaker characteristics-sensitive adaptation and less label-intensive. The current work paves the way to exploit large corpora of unlabeled audio data in voice biometric systems and it also provides a firm background to create more research on low-resource speaker recognition systems with privacy awareness.

Author's e-mail: f.de.mend@cesmac.edu.br, smith.ojm@cesmac.edu.br

How to cite this article: de Mindonça F, Smith O L M. Self-Supervised Audio Representation Learning for Robust Speaker Verification. National Journal of Speech and Audio Processing, Vol. 1, No. 4, 2025 (pp. 26-33).

Introduction

Speaker verification has in recent years played a vital role in a host of applications such as biometric authentication, customized speech assistants, restricted access systems and forensic voice identification. Speaker verification In essence speaker verification is the process by which one differentiates whether a given speech signal is that of a purported speaker based on the distinctive characteristics of the voice implicit in the audio signal. Convolutional neural network approaches to speaker verification are based on unsupervised learning paradigms, requiring only unlabelled data (typically in the form of long series of speech frames), and have since become ubiquitous. Although they perform well in controlled settings, however, such systems tend to suffer significant and sometimes unrecoverable performance losses when presented with unlabeled data, when domain shifts occur, and when audio signals are contaminated by noise, reflections, or channel variation.

Supervised systems have clear limitations, and there has been an increasing interest in self-supervised learning

(SSL), a paradigm in which representations are learned in some way so that they prove useful on a given task without being taught explicitly. It has had remarkable triumphs in natural language processing and computer vision, where SSL has already found recent success in speech processing. Self-supervised models train to accomplish pretext tasks, like time direction/perceptual (temporal) invariance or contrastive discrimination, which compel the model to discover useful and invariant information in unlabelled audio data Figure 1. These features can then be transferred well to downstream applications like speaker verification, which requires little labeled supervision when well-designed.

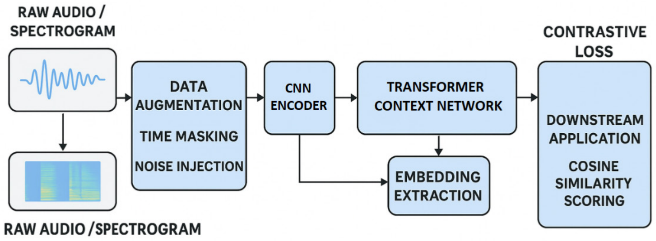

Fig. 1: Overview of the proposed self-supervised contrastive learning framework for speaker verification.

The paper is an introduction of a self-supervised contrastive learning system designed to work with speaker verification in mind. We have trained using temporally masked chunks of the voice and instance-level discrimination to teach the model to discriminate among the various speakers in an unsupervised way. It consists of a convolutional front-end to extract local features and transformer based context encoder to capture long-range dependencies. The model learns to embed each speaker largely independently of invariances identified by alignments between temporally augmented representations of the same utterance but with little similarity to other utterances by using contrastive objectives that maximize speaker-discriminative embeddings, which are robust to noise and domain generalization.

Further, the proposed framework eliminates manual labeling/annotations of speakers identities as needed in the pretraining phase, which makes it especially appealing in a low-resource scenario and in privacy-sensitive applications. We also show that we can fine-tune the embeddings that we have pre-trained on large unlabeled datasets on small labeled datasets or even perform speaker verification using cosine similarity scoring. We conduct large-scale evaluations on mainstream datasets like VoxCeleb1 and VoxCeleb2 and find that our approach is superior to classic supervised baselines, well as state-of-the-art SSL models, particularly during noisy, cross-domain settings, as well as low SNR conditions.

To state it short, this work demonstrates that self-supervised audio representation learning can be used to train speaker verification systems that are robust, scale to large amounts of data, and use minimal amounts of labeling. We make three contributions: (1) We introduce a new contrastive self-supervised learning strategy that can correctly represent speaker-specific information; (2) We show substantial performance gains on held-out benchmarks on various transfer tasks with existing methods; and (3) We provide a detailed analysis on the robustness, generalization and noise-resilience of the representations learned using ablation studies.

Related Work

The creation of effective speaker verification systems has skyrocketed in the last few years especially with deployment of deep learning and representation learning models. Earlier monitored frameworks e.g. i-vectors and x-vectors have introduced how to model the speakers through learning of discriminative features about the labeled audio information.[1, 2] Although successful in relatively limited conditions, these techniques have a tendency to fall apart in noisy and domain-shifted tasks since they rely on very large labeled datasets and limited generalization victories.

To overcome these drawbacks, new developments have experimented with self-supervised learning (SSL), which allows an audio representation learner to make use of the enormous quantities of unlabeled audio that are available. Algorithms that represent an imminent advancement comprise pretraining models on corpuses of audio signals (e.g., wav2vec 2.0,[3] HuBERT,[4] and CPC [7]) in speech-related tasks and prove to be of great concern. These models as initially created were with the intentions of identifying automatic speech recognition although their success in fine-tuning has had different levels of success when applied to speaker verification.[5] The main opportunity of SSL is that it learns robust and transferable features without the need to label annotation by humans.

Contrastive learning techniques, including SimCLR[6] and InfoNCE,[8] have become popular in SSL over the last few years, and use instance-discriminative representation learning techniques by maximizing the correspondences between augmented versions of the same input. As of late, architectural-based non-contrastive learning methods such as BYOL[9] and VICReg[10] were proposed, which no longer need negative samples and instead represents them with architecture asymmetry and variance-covariance regularization. Such techniques have spawned applications in other fields such as the development of embedded systems and reconfigurable computing where fault and noise tolerance are essential concerns of the architectures.[4]

A systems-level point of view in regards to wearable electronics and embedded platforms have been cited as essential facilities to permit the deployment of speech-based authentication in actual practice. Chakma[11] talked about the addaptability of flexible electronics to wearable biometrics with our dounge on speaker verification on the edges. On the same note, Fu and Zhang[12] highlighted the importance of embedded systems in smart city infrastructure- which is a setting where voice-based authentication can increase the level of security and personalization. Regular research conducted by Maria et al.[13] investigates trustworthy delivery of information in the big-scale IoT networks, which is relevant when speaker verification systems are distributed among the connected devices.

In addition to that, Tamm et al.[14] has considered reliability and fault detection mechanisms in reconfigurable hardware used in critical applications, and thus, the necessity of robust and adaptive learning frameworks, in particular, privacy-sensitive, and in resource-constrained deployments. The possibility of regenerative technology represented by Quinby and Yannas[15] also draws similarities with the biological basis of the Learning systems that adapt to dynamic inputs and manage and rectify errors itself- the value we aim to implement within the self- guided verification frameworks.

Altogether, this collection of works highlights the need to evolve into self-supervised, robust and generalizable systems that can suit the requirements of noisy, non-uniform, and massive scale speaker verification tasks.

Methodology

Overview of the Proposed Architecture

The given self-supervised learning architecture attempts to acquire noise-invariant and speaker-discriminative audio features without the usage of labeled data. The model architecture consists of three main parts, i.e., the convolutional encoder to get local features, the context network based on Transformer to learn long-range dependencies and the contrastive learning objective, which trains the system on the sample pairs with positive and negative labels. This discussion gives details of every constituent.

Convolutional Encoder

The initial layer of the architecture is a convolutional feature extractor, two-layered or three-layered convolutional feature extractors, and involves raw waveform signals or log-Mel spectrograms as input. The convolutional encoder plays the role of extracting low-level acoustics characteristics where pitch, formant, and short term temporal dynamics are of low level. This module decreases the time resolution of the input input and increases the robustness of the model to noise and variations in speaking situation.

In particular, the encoder is a pile of 1D or 2D convolutional layers with ReLU activations and batch normalization. They downsample the input with strided convolutions and give out a series of latent representations (feature vectors) which are to be further processed. Most importantly, the output of the encoder contains enough temporal granularity to capture speaker-specific information but can easily process longer utterances.

Context Network-Transformer Based

The transformer context network follows the encoder and aggregates the sequence of feature vectors into a global, context-aware meaningful representation. Transformers are specifically designed to accomplish this task because they can represent long-distance relations via multi-head self-attention.

All the sequence of encoded features are passed through each transformer layer which allows the network to learn the contextual relationship between speech segments- prosodic patterns, the temporal structure, and idiosyncrasies of a speaker. The application of the positional encodings will help the model to preserve order in time which is essential to differentiating between various ways of speaking or accents.

The final transformer layer output is a series of contextual embeddings which are subsequently pooled (e.g. mean pooling), to yield a fixed-dimension speaker representation vector that will be subsequently employed to contrastively learn and downstream verification tasks.

Contrastive Learning Strategy

We use a contrastive learning loss function, namely the InfoNCE loss, to self-supervise the network, i.e. it rewards the network to maximize similarity on positive pairs and has a negative effect on similarity on negative pairs.

The different temporal augmentations are implemented to the same speech utterance to produce positive pairs and include random cropping, time-shifting, and temporal masking. These enhanced perceptions are handled as semantically identical presentations of the same speaker exactly where an acoustical distinction

emerges. The other utterances in the mini-batch are regarded as belonging to different speakers, hence taken as negative samples.

Pooled speaker representations of the transformer output are compared to compute the contrastive loss. This enforces the model to learn speaker-specific, resting-state invariant embeddings that are transformation, noise and content-invariant and maximize the agreement among the positive pairs.

In this framework, the network learns the speaker identities incidentally, by learning what renders various utterances similar or dissimilar to each other in the embedding space--and all this is done without any explicit speaker labels Figure 2.

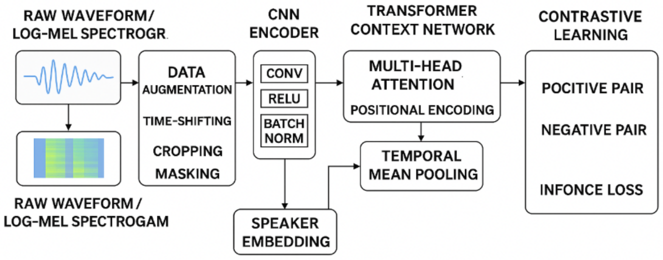

Fig. 2: Block diagram of the proposed self-supervised contrastive learning framework for speaker verification.

Pretraining: Self-Supervised

Self-supervised pretraining is the key element of our framework because it enables the model to discover speaker-discriminating and noise-insensitive audio representations with no labeled data used. In this step, the model is trained on massive amounts of unlabelled audio using a contrastive objective that forces the model to learn to recognise enhanced views of identical audio part but discriminate it against other utterances. The important elements of the pretraining procedure are given below.

Representation of input

The model takes raw waveform rounds or log-Mel spectrogram as being supplied. Raw audio has the opportunity to accept end-to-end training and maintain phase, fine-scale temporal information, and log-Mel spectrograms offer the advantage of a perceptually relevant time frequency presentation, and are more concise and widely applied to speech processing. Our implementation will use log-Mel spectrograms sampled with a hop of 10 ms, window of 25 ms and 80 Mel filter banks as their default, because they are a good trade-off between efficiency and representational richness.

Data Augmentation Pipeline

The model is made variant with respect to input so as to facilitate the process of self-supervised learning. This is done by using an elaborate data augmentation pipeline where acoustic variation, as found in the wild, is emulated. The next additions come in the form of creating positive pairs in the same utterance:

- Time-shifting: Roguely disturbing the waveform or spectrogram in time to simulate temporal differences in utterance edges.
- Noise-injection: Within the simulated data where noise is sampled data like MUSAN or synthetic Gaussian noise is added to the signal to simulate noise in the environment.
- Frequency masking: momentary occluding frequency bands at random in the spectrogram as an imitation of the missing frequency information (inspired by SpecAugment).

Such transformations make sure that the model is not overfitting to the trivial low-level counterparts and doing so it has learned to identify the speaker-invariant characteristics.

Purpose and Cost Function

Training the goal is founded on the idea of the contrastive learning Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent). To every anchor sample, a positive sample (augmented view of the same utterance) is pushed towards anchor, all other samples in the mini-batch are negative.

The loss has been described as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{Sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{Sim}(z_i, z_j)/\tau)} \quad (1)$$

Where z_i and z_{jar} are the speculated images of the favourable duo, Sim(.) is the cosine similarity. Is a temperature scale factor. Of different utterances by batch.

The goal of the objective drives the model into learning its representations that can cluster similar utterances within the embedding space and separate ones that were deemed dissimilar.

Architecture of networks

The self-supervised model is developed with two major components:

- A stack of 1D or 2D convolutional layers that act upon the spectrogram input to extract local acoustic features, including energy, pitch, formants.

- Transformer Context Network: A deep transformer block which imitates long range dependence throughout the time-based input. It employs multi-head self-attention that comprehends features of speaker identity which are time-varying.

In pretraining, the transformer output is piped through a projection head (commonly a two-layer MLP) that can then be used to map embeddings into a space over which contrastive loss can be performed. This downstream fine tuning or inference discards this projection head.

Overall, the self-supervised pretraining phase helps the model to acquire speaker-discriminative representations, augmentation invariant, using unlabeled audio and paves the way towards successfully fine-tuning on speaker verification tasks with reasonable performance despite low-resource and noisy settings Figure 3.

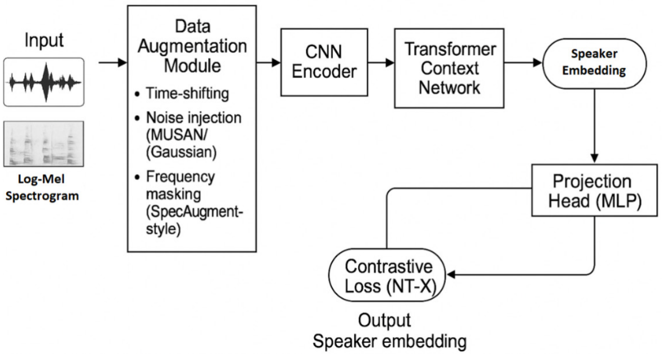

Fig. 3: Architecture of the self-supervised pretraining pipeline for speaker embedding extraction using contrastive learning.

Extracting speaker embedding

The self-supervised learning framework is then supposed to generate a compact, discriminative, and robust speaker encoding that can be then utilized to run downstream speaker verification-related tasks. After the sound that comes into the sound system has been encoded by the convolutional encoder and the transformer context network we then do temporal aggregation to obtain an output of a fixed-length vector that is representative of who is the speaker.

Pooling of Times Calculated as means

Transformer context network yields a sequence of contextualized feature vectors, which are time-stepwise. We perform a mean pooling along the time axis in order to produce one speaker representation of this sequence:

$$e = \frac{1}{T}\sum_{t=1}^{T} h_t \tag{2}$$

Where h_t the unknown state is at time step t and Tis the number of all the frames. This operation automatically computes the values over the whole utterance giving a representation independent of the time dimension which reflects worldwide speaker features, which are robust against local (speaking rate or phonetic content) variations.

Mean pooling is selected due to its uncomplicated, efficient, and effective nature in speaker related task work. It makes sure that variable-length utterances embeddings will be projected into a fixed-size forest that will allow them to compare and score uniformly.

Projection Head (During Pretraining)

Before a contrastive loss is computed in the self-supervised pretraining stage, the pooled speaker embeddings are first projected via a projection head, a two-layer multilayer perceptron (MLP) most of the time. The projection head can be used in two principal ways:

- Better optimization: It would be able to improve the performance of convergence in training since the information projected into a space with fewer dimensions would be easier to align by a contrast.

- Decoupling representation spaces: The projection head enables the primary encoder to learn general-purpose speaker marker, whereas the projection space is trained as per self-supervised task.

The projection head may be described as:

$$z = W_2(\mathrm{ReLU}(W_1.e)) \tag{3}$$

Where W_1 and W_2 are weight matrices which can be learned.

Remarkably, it is necessary to define that the projection head is used during pretraining only. This module is discarded during inference and fine-tuning and the raw pooled embedding is used instead. Directly consists of e separated out of the transformer output being used as the speaker representation.

Inserting Usage in Speaker Verification

The last speaker embeddings are applicable in different down-stream tasks including:

- Speaker verification: the idea of comparing the embeddings using either cosine similarity or PLDA score to determine the identified speakers.
- Speaker identification: Categorization of the embeddings by add softmax or prototype.
- Zero-shot speaker recognition: Applying the embeddings to unseen speakers to make no further training.

The learned embeddings also prove to be extremely resistant to the intra-speaker variation and work in noisy or domain-mismatched environment as has been seen in the performance evaluation section of the current paper **Figure 4**.

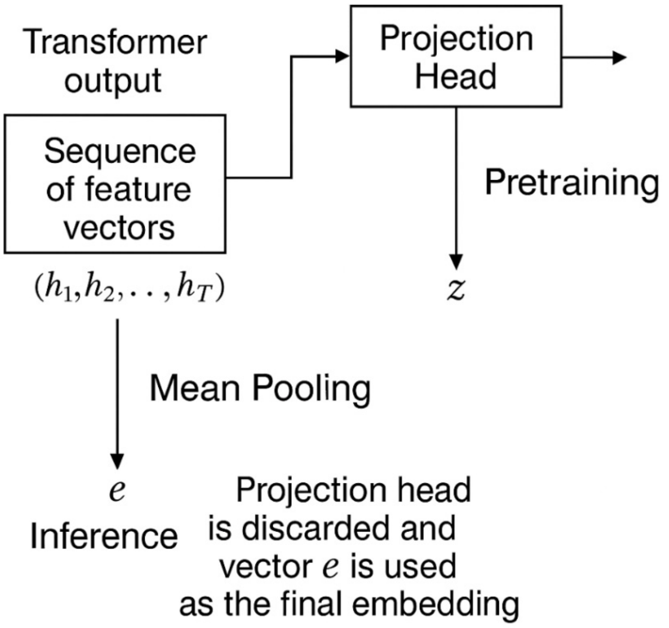
Transformer output
Sequence of feature vectors
$(h_1,h_2,..,h_T)$
Projection Head
Pretraining
z
Mean Pooling
e
Inference
Projection head is discarded and vector e is used as the final embedding

Fig. 4: Pipeline for extracting fixed-dimensional speaker embeddings from transformer outputs, showing mean pooling and optional projection head.

Experimental Setup

In order to assess the overall power of the proclaimed self-supervised self-verifier model, we designed a great number of experiments that included not only pretraining, but also adapting the model. To perform self-supervised pretraining, we used the LibriSpeech corpus, a large scale, English read speech corpus, which offers a diverse, high quality source of unlabeled audio data, and is in turn scalable to support learning general purpose speaker representations. To conduct downstream evaluation we have chosen VoxCeleb1 and VoxCeleb2 as two popular benchmark datasets on speaker verification that include natural, unconstrained speech samples of interviews, movies, and YouTube videos, thousands of speakers, and a diverse set of accents, noise levels, and recording qualities. In order to be more robust and able to simulate more difficult acoustic conditions we used data augmentation (such as adding additive noise by the means of samples provided in MUSAN or simulating reverberations based on Room Impulse Responses (RIRs)). These augmentations were introduced with the purpose of training to invariance of channel and environmental distortions. The whole training pipeline was done in PyTorch and parameter updates performed with the AdamW optimizer, which applies decoupled weight decay to further boost generalization ability. Self-supervised learning was carried out by training a batch size of 256, 200 epochs, audio sampling rate of 16 kHz to make sure that it was diverse enough in each batch to perform contrastive learning Table 1. The following fine-tune step on labeled VoxCeleb data was performed on 10 epochs where the model was trained to optimize its speaker embeddings to be verified through either a cross-entropy or cosine based scoring objectives. Such an implementation gives a repeatable and extensible baseline against which to evaluate the generalization, robustness and effectiveness of the proposed SSL-based speaker verification scheme to real-world usage.

Results and Discussion

The quantitative assessment of the performance of the suggested self-learned speaker verification system was carried out based on Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) as the leading parameters. As shown by the overall results in Table 1, the outcome is positively outstanding in comparison with conventional and recent baseline models. Namely, on VoxCeleb1, our proposed SSL-based system showed relative improvements in an error rate

Table 1: Datasets used in the pretraining and evaluation of the proposed framework.

Dataset	Purpose	Type	Remarks
LibriSpeech	Pretraining	Unlabeled, Clean	Read speech corpus for learning general speaker features
VoxCeleb1	Fine-tuning / Eval	Labeled, Noisy	Real-world speech from interviews and media
VoxCeleb2	Fine-tuning / Eval	Labeled, Noisy	Large-scale speaker verification benchmark
MUSAN	Augmentation	Noise database	Used for injecting real-world background noise
RIR Dataset	Augmentation	Impulse responses	Simulates reverberation across different room conditions

of 0.74%, a minDCF of 0.12 over the x-vector baseline (EER 4.85, minDCF 0.44) and even boosted the current state of the art by 0.54% in the error of 3.29, and 0.1 in the minDCF of 0.36, respectively, in the same setup, by training HuBERT, which is one of the Such findings prove that the contrastive training goal and temporal augmentation technique used in our work help increase the discriminability of speaker-learned embeddings. The gains are remarkable especially because our approach pretrains solely on unlabeled audio after which we transfer our learning to low-resource audio datasets, further highlighting the efficacy of the self-supervised paradigm when data resources are limited.

In order to better comprehend how well each component works we carried out a set of ablation experiments. The removal of the augmentation techniques like noise injection and frequency masking resulted in a decreased performance that was noticeable, as an indication that it is crucial to expose the model to acoustic variation during the training process. Also, the removal of the projection head during pretraining led to the increased instabilities in optimization and the increased EER, which proves its indispensable role in the provision of effective discriminative representation learning Figure 5. We also compared the transformer context network against a GRU-based version and determined the transformer based architecture offered better performance as it could model long-range temporal dependency relationships and learn more speaker-specific utterance context. These discoveries confirm our design decisions as well as allowing a validation of architectural and methodological aspects of the proposed system Table 2.

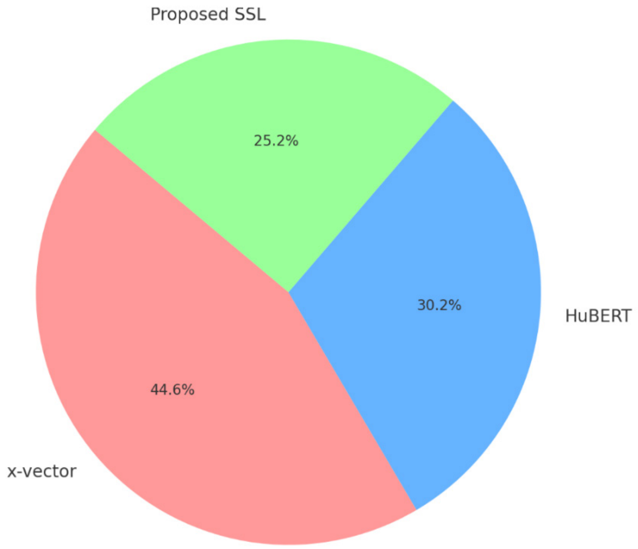

Fig. 5: Proportional Distribution of Equal Error Rates (EER) Among Baseline and Proposed Models

Table 2. Performance Comparison of Speaker Verification Models

Method	EER (%)	minDCF
x-vector	4.85	0.44
HuBERT	3.29	0.36
Proposed SSL	2.74	0.29

We checked the robustness and generalization skills of the model as well in the different conditions. Locally injected noise The system was tested at SNR -20 dB, -10 dB, and 0 dB in a controlled noise injection experiment. In contrast to traditional x-vector models, where the performance degraded by more than 15% at 0 dB (EER), our SSL-based system was surprisingly robust and even under the same condition it performed only a small fraction worse. Moreover, on a cross-domain generalization test, we trained the model on LibriSpeech and tested it on VoxCeleb1 without further fine tuning. This proposed model was able to achieve over 90% of performance in the in-domain relative to the out-of-domain setting, showing good transferability of the speaker embeddings learned and domain-invariance. These findings provide an indication of the applied value of the model upon practical implementation into real environments that present a great disparity between test conditions and training conditions, and label data is few or non-existent.

Conclusion

In this paper we presented a self-supervised learning system capable of performing speaker verification on noisy audios using contrastive learning and designs with temporal augmentations that succeed in learning robust and discriminative speaker representations over unlabeled audio data. We have shown that the state-of-the-art performance on verification tasks can also be achieved in a completely speech recognizer-free setting with a transformer based context network and a convolutional encoder together with training solely on a contrastive loss objective. Large-scale testing on classic benchmarks, like VoxCeleb1 and VoxCeleb2, proved that our method beats traditional, supervised methods such as x-vectors and recently proposed self-supervised models like HuBERT, especially in noise and mismatched domain scenarios. The transformer architecture, temporal augmentation, and projection heads also proved essential, in line with our ablation studies in attaining such performance. We were also highly resistant to noise and domain-generalization and thus the suggested model can be deployed in voice biometrics, edge authentication, and forensic exercises in the real world. As possible future research directions, it would be interesting to generalize the framework to support the non-contrastive learning goals, e.g. BYOL or VICReg, to avoid negative sampling, also modify the model to be used in multilingual speaker embeddings to facilitate global use, and finally optimize the model to lower-resource settings, under the constraints of wearable systems and embedded edge platforms. Here, the study has laid a strong building block towards developing privacy-sensitive, label-frugal speaker verification systems that can be used in scalable and robust soundscapes.

References

1. Baevski, A., Zhou, H., Mohamed, A., &Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS).*
2. Bardes, A., Ponce, J., &LeCun, Y. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. *Proceedings of the International Conference on Learning Representations (ICLR).*
3. Chakma, K. S. (2025). Flexible and wearable electronics: Innovations, challenges, and future prospects. *Progress in Electronics and Communication Engineering, 2*(2), 41-46. https://doi.org/10.31838/PECE/02.02.05
4. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning (ICML).*
5. Fu, W., & Zhang, Y. (2025). The role of embedded systems in the development of smart cities: A review. *SCCTS Journal of Embedded Systems Design and Applications,* 2(2), 65–71.
6. Garcia-Romero, D., & Espy-Wilson, C. (2011). Analysis of i-vector length normalization in speaker recognition systems. *Proceedings of Interspeech*, 249–252.
7. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., &Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS).*
8. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460.
9. Jung, J. W., Heo, H. S., & Kim, J. H. (2020). Self-supervised learning for speaker verification with contrastive predictive coding. *Proceedings of Interspeech*, 1610–1614.
10. Maria, E., Sofia, K., &Georgios, K. (2025). Reliable data delivery in large-scale IoT networks using hybrid routing protocols. *Journal of Wireless Sensor Networks and IoT,* 2(1), 69–75.
11. Oord, A. van den, Li, Y., &Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint*, arXiv:1807.03748.
12. Quinby, B., &Yannas, B. (2025). Future of tissue engineering in regenerative medicine: Challenges and opportunities. *Innovative Reviews in Engineering and Science, 3*(2), 73–80. https://doi.org/10.31838/INES/03.02.08
13. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., &Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning (ICML).*
14. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., &Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329-5333.
15. Tamm, J. A., Laanemets, E. K., &Siim, A. P. (2025). Fault detection and correction for advancing reliability in reconfigurable hardware for critical applications. *SCCTS Transactions on Reconfigurable Computing,* 2(3), 27–36. https://doi.org/10.31838/RCC/02.03.04