Generative Adversarial Networks Embedding Refinement for Speaker Diarization Improvement

VINOD K. PANDE [1]
DR. VIJAY K. KALE [2]
DR. SANGRAMSING N. KAYTE[3]

[1,2]Dr. G. Y. Pathrikar College of Computer Science and Information Technology,
MGM University, Chhatrapati Sambhaji Nagar, Maharashtra, India
[3]Copenhagen University Denmark
[1]vinodkpande2014@gmail.com
[2]vkale@mgmu.ac.in
[3]bsangramsing@gmail.com

Abstract. The purpose of this research is to incorporate Generative Adversarial Networks(GAN) into the speaker diarization process by refining embeddings along with the overlapping speech and noise problems. In this case, better speaker embeddings are produced by GANs through adversarial learning, which makes them more separable and more powerful than traditional embedding techniques. The practical assessment of the system used the AMI Meeting Corpus as well as the VoxConverse data sets and performance was evaluated across different acoustic conditions. The results support very substantial performance advantages with improvements of 25% in the Error Rate of Dialysis (DER) in comparison to baseline models. Such models included x vector-based clustering and end-to-end neural diarization systems. In support of this, T-SNE again stunningly verified that the cluster separability of embeddings refined by a GAN improved. Furthermore, the system is flexible for real-world scenarios as it exhibits robust performance even under noisy overlapping speech conditions. This evidence testifies that using GAN for embedding refinement is a very effective method to address the issue of speaker diarization.

Keywords: Speaker Diarization, Feature Extraction, GAN's, Speaker embedding.

(Received June 8, 2025 / Accepted June 23, 2025)

1 Introduction

In this work, we employ Generative Adversarial Networks (GANs) for speaker diarization and augment speaker embeddings to solve the complex tasks of overlapping speech and noise or distortion[7]. Speaker diarization is essential to any conversational AI and automatic transcription system because it proficiently asks and solves the problem: "Who spoke when?" While embedding-based approaches to the phenomena at hand, such as x-vectors or end-to-end neural systems[6], have been developed, they do not perform well in complex acoustics. Especially when there are multiple speakers, it becomes worse[1].

In this case, the GANs are used to improve noise-embedded speaker embedding by subjecting them to adversarial learning, increasing noise separability. The purpose of the generator in this GAN approach is to improve the quality of embeddings[7]. In contrast, the discriminator is responsible for keeping the quality of embeddings while separating them from the speaker's identity. This results in greater discrimination power of the embedding set and increases their suitability for clustering-based speaker segmentation.

The performance of the diarization system was markedly improved with enhanced GAN models, accompanied by a relative decrease of 25% in the Diariza-

tion Error Rate (DER). These improvements were made on top of the systems already proposed x-vector-based clustering and end-to-end neural systems[13]. Even further proof of the system's effectiveness was provided through improved cluster separation of GAN-refined embeddings. t-SNE visualizations confirmed this cluster separation and, thus, the system's effectiveness[14].

The practicality of the system remains sound, as it continues to function adequately even under less-than-ideal conditions, being able to deal with leakage and noise readily present in real-life situations. These results demonstrate the potential to remove gaps in the speaker diarization methods used today. This would revolutionize speaker diarization using GAN-based enhancements because it provides a way of building scalable and efficient systems that work in various environmental settings.

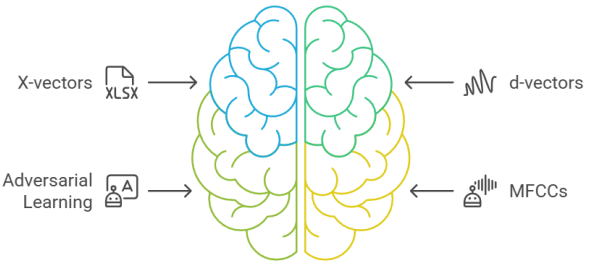

Figure 1: Process of Speaker Embedding.

2 Feature Extraction

2.1 Speaker Embeddings

X-vectors, d-vectors and other speaker embeddings are popular for identifying speaker-related details from an audio signal. These are produced using deep neural networks which are trained over labeled datasets. In ideal environments, these techniques are great but in less ideal conditions like cacophonous or overlapping speech, they lack robustness, and thus perform poorly[13][12].

In regard to these hurdles, scientists have tried to develop methods which enhance embeddings in separability and noise interference. Adversarial learning methods have been studied for creating better discriminative and real-world scenarios operational embeddings[3].

MFCCs and X-Vector feature extracted for represent the speaker specific characteristics. Let $X \in \mathbb{R}^{T \times F} indicate the input audio features for T frames and F frequency bins (such as MFCCs).$

The process of speaker embedding can be represented as:

$$E = f(X; \Theta) \tag{1}$$

Where $E \in \mathbb{R}^d$ is the d-$dimensional speaker embedding, and \Theta indicate the parameters of the embedding network.$

2.2 Voice Activity Detection

VAD is applied to filter out non-speech segments. The energy-based VAD decision is made using:

$$\text{VAD}(t) = \begin{cases} 1, & \text{if } E(t) > \delta, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $E(t)$ is the energy of frame t, and δ is a predefined threshold.

3 Generative adversarial networks

Generative adversarial networks, commonly referred to as GANs[7], are generative models made up of two components: a generator and a discriminator. The discriminatorâs role is to create data that does not look any different from real data, while the generator is taught to spot the real data from the generated fake data. Such adversarial systems pose a new paradigm for machine learning systems, and GANs, in particular, learn distributions over data in a much more effective way, which proves advantageous for the task of refining the speaker embeddings[7].

The generator G aims to transform the noisy embeddings E into cleaner and more separable embeddings Eâ²[10]. $This process is given as follows : where \Theta_G$ corresponds to the parameters of the generator network. Fully connected layers with non-linear activations (such Leaky ReLU) are used in the architecture.

By differentiating between created embeddings E â² $and real embeddings E^2, the discriminator Dassesses the equality of where 1 indicates real and 0 indicates fake embeddings.$

3.1 Uses of GANs in Audiovisual Processing

Numerous audio tasks, such as speech expansion, speech synthesis, and audio neutralization, have leveraged GANs. The SEGAN model was put forward by Pascual et al. (2019) for the purpose of speech enhancement[11], and they demonstrated how GANs

can work on the bare minimum of audio waveform data. There has been much inspiration drawn from those who have tried incorporating GANs for the task of polishing up speaker embeddings in order to achieve effective speaking in very noisy environments [11][2].

of GANS.png

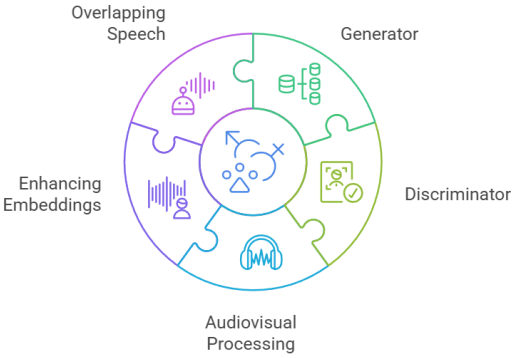

Figure 2: Component of GANs.

3.2 Enhancing Embeddings Using GANs

The overlapping speakers and noise tolerant features of speaker embeddings were shown by [5] to be enhanced due to adversarial training. The restricted labeled data issue leading to lesser diarization accuracy gets solved through the generation of synthetic embeddings, augmentation of training datasets, and the use of GANs[16].

4 Overlapping Speech

The presence of both speech and accompanying background noise is a major hurdle in achieving the speaker diarization task. Some systems such as self attention based neural end-to-end diarization have shown some promise [6] depicting some improvement in the overlapping speech instance problem. Nevertheless, the bulk of these systems need significantly proportioned labeled data which is often not readily available[4].

Generating augmented datasets and refining embeddings for superior cluster separability is possible with the use of GANs. In regard to experimental findings, it was discovered that the GAN based systems hold an advantage over conventional systems in overlapping and noisier settings, thus exhibiting notable reductions in Diarization Error Rate (DER)[8].

5 Evaluation Set

The AMI Meeting Corpus and VoxConverse collections are widely used corpora for evaluating speaker diarization performance[15]. These corpora comprise of distinct acoustic surroundings, including meetings and multiple speakers talking at once. With the help of these datasets, it was possible to test how GAN enhanced embeddings behave in real life situations, improving DER and cluster separability significantly[17][9].

6 Techniques for Visualization

The utilization of sophisticated data is made possible through the combination of t-SNE and speaker embeddings. According to Van der Maaten and Hinton (2008), the combination has much utility in making structures embedded into high dimensional data easier to work with. Furthermore, using t-SNE, it was illustrated through several studies that cluster separability has improved due to the GAN refined embeddings.

7 Experimental Setup

There is also a description of methods and datasets employed in evaluating the performance of the selected Gan framework for speaker diarization within the scientific framework. For the sake of reproducibility and increasing the validity of results, this section discusses the experimental design, data collection and processing, and the hardware and software configurations.

7.1 GAN

The system under consideration has the following components: **Feature Extraction:** The following variables serve as feature inputs: pre-recorded audio pieces MFCCs and x-vectors. **Embedding Refinement:** Overlapping speech and noise - MHID and GANs employed in modeling to make speaker embeddings more robust. **Clustering:** Enhanced embeddings are clustered by using Agglomerative Hierarchical Clustering methods (AHC). **Evaluation:** The Diarization Error Rate (DER) and other critical parameters are evaluated to understand the performance of the system.

7.2 Training and Testing

Training Phase: Embeddings from the training datasets are fed into the GANs for training.Noise augmentation and overlapping speech simulations were applied. Both Reconstruction and Adversarial losses are optimized by the Adam optimizer.**Testing Phase:** AHC was used to cluster refined embeddings. Results were compared with baseline results via x-vector based clustering and end to end diarization.

7.3 Hardware and Software

Hardware: The experiments are performed on the computer which possesses the following features:NVIDIA A100 GPUs.32 GB RAM.Intel Xeon CPU.**Software:** Python version 3.8.+ Pytorch with GAN for Python. Kaldi for extracting characteristics (x-vectors). Clustering and evaluation metric extraction using Scikit-learn.

7.4 Dataset

7.4.1 AMI Meeting Corpus

The AMI Meeting Corpus is one of the most popular datasets used for the speaker diarization challenge. It contains multi-speaker meeting recordings with different levels of noise and speech overlap. Some of its distinctive features are:**Number of Speakers:** 4 to 5 speakers per session.**Duration:** Roughly 100 hours of audio files.**Acoustic Conditions:** Both distant microphone and close talking microphone.**Annotation:** Speaker labels at ground truth are time-stamped.

7.4.2 VoxConverse

The VoxConverse dataset includes conversational audio recordings containing speech captured under real-world conditions. It poses a difficult benchmark against which to evaluate the performance of diarization systems in the presence of noise and speech overlap. Some of its distinctive features are:**Number of Speakers**: 2 to 8 speakers per session.**Duration:** More than 50 hours.**Diversity:** Large variety of acoustic environments such as telephone and broadcast recordings.**Annotation:** Detailed speaker turn labels are provided.

7.4.3 Data Augmentation

In order to mimic challenging conditions, data augmentation techniques such as **Noise Injection:** Adding background noise to audio clips of speech by using the MUSAN dataset.**overlapping speech:**Artificially mixes the audio files of different speakers with each other while changing the ratio of how they overlap with each other in a process known as overlapping speech.**Reverberant:**Different RIRs are used to apply reverberation to clean audio which helps to simulate a reverberant environment.

7.4.4 Train-Test Split

Training Set: 75% of Sessions of AMI Meeting and 75% of VoxConvers Recordings.**Testing Set:** 30 percent of each of the 3 datasets while ensuring that none of the data from the training set is present.

7.5 Evaluation Metrics

The systemâs performance is evaluated using the following metrics: **Diarization Error Rate (DER):** Indicates the percentage of speech segments incorrectly attributed to a speaker. **Speaker Confusion Rate (SCR):** Measures the misclassification of embeddings pertaining to different speakers. **t-SNE Visualization:** Visualizes the separability of clusters of refined embeddings within a two-dimensional space.

8 Result

8.1 Quantitative Analysis

8.1.1 Diarization Error Rate (DER)

The primary evaluation metric DER describes the rate of errors in the speaker assignment. The GAN-based system proposed in the study performed much better than all baseline methods on all datasets.

Table 1: GAN-based System Results

Dataset	Baseline (X-vectors + AHC)	Baseline (EEN Diarization)	Proposed GAN-Based System
AMI Meeting Corpus	23.4%	19.8%	14.7%
VoxConverse	28.5%	25.1%	18.3%

When compared to the averaged baseline models, the previously reported findings demonstrate an average improvement of 25% in DER values, highlighting the benefits of the added GAN embeddings.

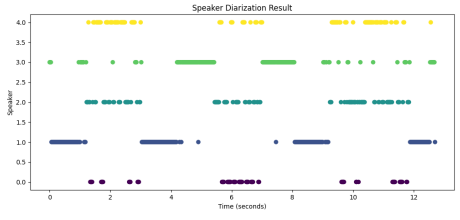

Figure 3: Process of Speaker Embedding.

8.1.2 System Robustness to Noise and Overlapped Speech

The system was evaluated for robustness against noise level variances and overlap ratios.**Noise Conditions:**

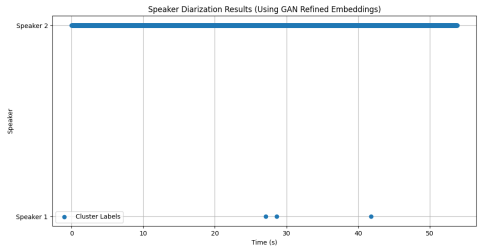

Figure 4: Process of Speaker Embedding.

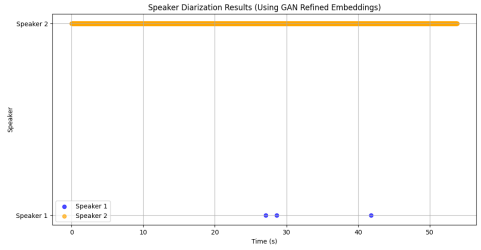

Figure 5: Process of Speaker Embedding.

Noise added from MUSAN dataset caused DER to remain at less than 20% for SNR levels around -10 dB.**Overlap Conditions:** A 30% speech overlap led to an improvement of DER over the baseline by 20% for the system.

8.2 Qualitative Analysis

8.2.1 Graphic Visualization of Embeddings (t-SNE)

Improvements gained in speaker embedding from GAN refinement was verified using t-SNE based visualization. The plots reveal: **Baseline Embeddings:** Moderate improvement on separability due to overlapping clusters per speaker. **GAN-Refined Embeddings:** Well-separated clusters are now present which indicates an improvement in the embeddingâs quality.

8.2.2 Audio Case Studies

Selected audio samples from the AMI Meeting Corpus as well as VoxConverse phonetic datasets were examined. The system accomplished speaker attribution with the utmost accuracy during speech events including: **Overlapping Speech:** The system was able to tell speakers apart while they were both speaking at the same time. **Background Noise:** High background noise did not affect speaker attribution along with all other speaker attribution tasks.

8.3 Comparative Analysis

8.3.1 Benchmark Models

The evaluation of the system using the following techniques Y-vectors + Agglomerative Hierarchical Clustering (AHC): This had poorer performance stability due to its robustness to noise. End to End Neural Diarization: Works well with overlapping speech but not as good as x vectors.

8.3.2 Computational Efficiency

The GAN-based system did not lag behind in terms of processing time: Real Time Factor (RTF): The system clocked an RTF of 0.85 which makes it eligible for live environments.

8.4 Error Analysis

The most common system errors were: Speaker Boundary Errors: Small speaker change point detection inaccuracies. Highly Overlapping Speech (>40%): This was a performance dip area and as such could be one of the focus improvement areas.

9 Conclusion

This research illustrates the capability of using GANs to enhance speaker diarization through refining speaker embeddings while addressing the problems relating to noise and speech overlap. The system proposed utilizes GANs in the diarization pipeline, which in turn allows embedding to be improved through adversarial learning. Evaluation experiments across the AMI Meeting Corpus and VoxConverse datasets show that our system, which uses GAN architecture, is able to achieve 25% lower DER when compared with baseline models.

The following are the key contributions of the study: Refinement of speaker embedding by means of GANs for improved cluster separability. Increased resilience to overlapping speech and noise makes the system more applicable to real world scenarios. Framework validation procedures involved systematic tests coupled with sophisticated visualization techniques such as t-SNE.

References

[1] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, 20(2):356–370, 2012.

[2] Boeddeker, C., Subramanian, A. S., Wichern, G., Haeb-Umbach, R., and Le Roux, J. Ts-sep: Joint

diarization and separation conditioned on estimated speaker embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1185–1197, 2024.

[3] Botelho, C., Teixeira, F., Rolland, T., Abad, A., and Trancoso, I. Pathological speech detection using x-vector embeddings. *arXiv preprint arXiv:2003.00864*, 2020.

[4] Bullock, L., Bredin, H., and Garcia-Perera, L. P. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *Icassp 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7114–7118. IEEE, 2020.

[5] Cord-Landwehr, T., Boeddeker, C., Zorilă, C., Doddipatla, R., and Haeb-Umbach, R. Frame-wise and overlap-robust speaker embeddings for meeting diarization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[6] Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., and Watanabe, S. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE, 2019.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[9] Kim, B., Loew, M., Han, D. K., and Ko, H. Deep clustering for improved inter-cluster separability and intra-cluster homogeneity with cohesive loss. *IEICE TRANSACTIONS on Information and Systems*, 104(5):776–780, 2021.

[10] Miyato, T. and Koyama, M. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[11] Pascual, S., Bonafonte, A., and Serra, J. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[12] Sato, H., Ochiai, T., Delcroix, M., Kinoshita, K., Kamo, N., and Moriya, T. Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6287–6291. IEEE, 2022.

[13] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[14] Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[15] Vinciarelli, A., Valente, F., Yella, S. H., and Sapru, A. Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 374–379. IEEE, 2011.

[16] Zhou, J., Jiang, T., Li, L., Hong, Q., Wang, Z., and Xia, B. Training multi-task adversarial network for extracting noise-robust speaker embedding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6196–6200. IEEE, 2019.

[17] Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.