ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2025.0515

FEATURE EXTRACTION USING I-VECTOR AND X-VECTOR METHODS FOR SPEAKER DIARIZATION

Vinod K. Pande[1], Vijay K. Kale[2] and Sangramsing N. Kayte[3]

[1,2]Department of Basic and Applied Science, Dr G.Y. Pathrikar College of Computer Science and Information Technology, India
[3]Centre for Digital and Computational Humanities, University of Copenhagen, Denmark

Abstract

Speaker diarization is the process of identifying who is speaking at different times in audio recordings. This is important in various situations, such as recording meetings, monitoring calls in call centers, or analyzing media. In this paper, examine how well different methods for speaker diarization perform in real-life scenarios. focus on two modern techniques: I-vectors and X-vectors. I-vectors are effective for automatic speaker recognition because they create compact and efficient representations of speakers using statistical models. However, they struggle in situations involving overlapping voices or background noise. On the other hand, X-vectors overcome these limitations. They use deep neural networks to create more complex and reliable representations, making them better suited for challenging conditions. To evaluate these two approaches, used standard datasets, specifically the AMI Meeting Corpus and VoxCeleb. measured their performance using two indicators: Diarization Error Rate (DER) and Jaccard Error Rate (JER). Results show that while I-vectors are less resource-intensive and work well in ideal conditions, X-vectors perform better in real-world settings where noise and overlapping speech are present. This study provides guidance for practitioners in choosing the right approach based on their needs, considering factors such as accuracy, computational costs, and reliability.

Keywords:

Speaker Diarization, I-Vector, X-Vector, MFCC, Speech Recognition

1. INTRODUCTION

1.1 SPEAKER DIARIZATION

Speaker diarization is the method of tagging sections of an audio recording to identify who is speaking and when. In essence, it addresses the question: "Who spoke at what time?" This is especially crucial for automatic transcription systems, teleconferencing, call centers, forensic audio analysis, and other applications. By distinctly recognizing various speakers, speaker diarization enhances several functions, such as improving automatic speech recognition (ASR) by minimizing confusion caused by overlapping voices [1].

1.2 HISTORICAL CONTEXT AND CHALLENGES

In the past two decades, the way approach speaker diarization, which involves identifying the speakers in a conversation, has changed dramatically. Recent training techniques that integrate speaker and language learning have produced favorable outcomes in recognizing speakers and enabling accurate transcription of speech. To enhance the effectiveness of certain estimation methods, it was crucial to separate the training data [2]. However, these techniques faced challenges such as overlapping speech, background noise, and many speakers. Furthermore, many conventional methods encountered difficulties due to variations in how speakers communicate, including their styles, emotions, and the context of the discussion [3]. The iVector suppression added to the relative error the excitation of a linear system formed by structures of transition networks and splines. The application of iVectors resulted in a considerable improvement in speaker verification and diarization [4]. How- ever, Geoffrey Hinton pointed out, an iVector has some disadvantages when measuring axon projections, for instance, in chopped talks. In recent years, the primary substitute for I-vectors has been X-vectors, which are derived from deep neural networks, or DNNs. Unlike I-vectors, X-vectors capture various characteristics of a speaker using more advanced DNN embeddings [5]. This paper examines the performance of I-vectors and X-vectors in speaker diarization, with an emphasis on feature extraction and segmentation.

2. BACKGROUND AND RELATED WORK

2.1 ADVANCEMENTS IN SPEAKER DIARIZATION TECHNIQUES

Diarization of speakers has undergone tremendous change in the past 2 decades. The first speaker diarization systems used a combination of clustering and segmentation methods based on Gaussian Mixture Models and Hidden Markov Models. In order to distinguish speech signals, one of the methods used low-level acoustic features, for example MFCCs (Reynolds, 2000). Even though these processes proved to be rather effective when used in controlled environments, the versatility of their application was limited as they struggled with various speaker styles, overlap in speech as well as background noise [6]. During the late 2000, I-vectors became a game changer in that it offered a simpler way of speech that indeed required lower dimensional space in that high dimensional feature vectors were projected to a lower dimension. Based on this, Total Variability Modeling which is concerned with both speaker and channel variability was always useful in speaker verification as well as speaker diarization tasks[6]. Due to their effectiveness and low dimensional cost I-vectors were adopted industrywide. I-vectors are however not applicable in practice and complex scenarios, for instance teleconferencing and working with broadcast media for they are unreliable under conditions with excessive noise or use short speech segments

2.2 PROCESS OF X-VECTORS FORMATION

Deep learning techniques have indeed vastly contributed to the effective resolution of the speaker diarization task, particularly using X-vectors. In contrast to I-vectors, X-vectors owe their genesis to Deep Neural Networks (DNNs), which can learn complex and more abstract relationships that characterize individual speakers [6]. The usual X-vector systems consist of a time-delay neural network (TDNN) trained on large corpora to reliably classify speakers. These embeddings have outperformed traditional approaches in scenarios involving overlapping and concurrent speech, background noise, and short phrases [6] [7]. In practical use, X-vectors are more widely integrated into real-time conference systems and automatic voice recognition systems due to their agnostic nature regarding the speaker and background sounds [7]. Furthermore, X-vectors have demonstrated strong performance in tasks involving multiple speakers and in cross-linguistic tasks. This makes such technology highly useful in the current context of speaker recognition systems [7].

2.3 COMPARATIVE STUDIES

There have been some comparative studies evaluating the effectiveness of I-vectors and X-vectors in speaker diarization tasks. One such study, Sell et al. [8], tested these two approaches using a DIHARD Challenge dataset, which contains audio data from various recordings conducted under the presence of noise and overlapping speech. The study found that X-vectors demonstrated a reduced bias in terms of the Diarization Error Rate (DER) and exhibited greater robustness against noise compared to I-vectors. However, the high computational requirements for processing X-vectors remain an issue, particularly for real-time streaming. Another practical comparison, conducted by Wang et al. [9], noted that while I-vectors are still suitable in clean and controlled settings, X-vectors perform better in real-world applications, such as conference call transcription and media analysis. These findings underscore the importance of choosing the best feature extraction method based on the specific use case.

2.4 REAL-WORLD APPLICATIONS

Meeting Transcription: Diarization systems offered by companies like Zoom and Microsoft Teams have integrated the possibility of prescriptive diagnosis. With the ongoing improvement of these systems, the ability to record individual speakers has also been incorporated. X-vectors play an especially important role in virtual meetings where speech and background noise overlap [1]-[7].

- **Call Centers**: According to the study, transcription is useful for overseeing exchanges, monitoring compliance, and enhancing first-call resolution rates. The use of i-vectors is common because they are computationally inexpensive, yet x-vectors are gaining popularity due to their superior performance in noisy environments [1]-[9].
- **Broadcast Media**: In news and talk shows, which also serve as the Talk and News segments of the program, several speakers talk and are subjected to noise. X-vectors provide better segmentation, and the clustering process improves the quality of transcription and analysis [1]-[9].
- **Forensic Analysis**: Diarization is used here to understand audio recordings and to identify who is watching whom in a surveillance exercise. X-vectors' robustness to noise and variability makes them useful for forensic purposes [10].

3. METHODOLOGY

3.1 DATA COLLECTION AND PREPROCESSING

The study employed two datasets that are publicly accessible:

3.1.1 AMI Meeting Corpus:

The AMI Meeting Corpus is an excellent benchmark dataset in the domains of speaker diarization and automatic meeting transcription. It comprises multi-speaker audio recordings that were collected in artificial meeting settings. Salient points are as follows [12]:

- *Number of speakers*: Meetings typically have about 3 to 5 members, which approximates actual conversational circumstances.
- *Recording Conditions*: Individual headset microphones and far-field microphone arrays are used to record audio clips, allowing for the testing of both close-talk and far-talk situations.
- *Speech variability*: The database contains spontaneous speech, overlapping dialogue, speech interruptions, and noise. Therefore, it is an excellent dataset for meeting scenarios.
- *Annotations*: The reliability evaluation of diarization systems is achieved through detailed speaker labels, timestamps, and transcription. This dataset was introduced in the AMI project, whose objective was to facilitate research and analysis of multimodal interactions. Its realistic representation of women and their rich diversity makes it a primary candidate for speaker diarization models.

3.1.2 VoxCeleb Dataset:

The VoxCeleb dataset is a collection of a large-scale speaker identification database compiled from YouTube recordings of famous celebrities during interviews and informal conversations. The dataset has some salient features, such as [13]:

- *Size and Diversity*: The VoxCeleb dataset is constituted of a wide variety of speech samples with more than thousands of speakers and audio exceeding two thousand hours.
- *Recording Conditions*: The dataset also features audio recordings with higher noise levels including music, background talk, and microphone noise making it best for evaluation of diarization systems in realistic conditions.
- *Utterance Variability*: The dataset also consists of shorter and longer utterances aiding in improved testing of diarization models of different segment lengths. The data set was created by Visual Geometry Group (VGG) operating at University of Oxford with the intention of aiding researchers in the domain of speaker recognition and verification problems [9] [13]. VoxCeleb provides more contractually difficult realistic audio scenarios than AMI Corpus.

3.1.3 Preprocessing Steps:

Noise Reduction: Used spectral gating to lessen interference from ambient noise. This is essential in increasing the quality of feature extraction where noise in the environment is high [14].

- **Voice Activity Detection (VAD)**: For cleaning up the lack of sound parts, used modules of WebRTC VAD. VAD is important in filtering out noise and silence and retaining meaningful speech parts.
- **Segmentation**: The audio files were segmented into audio files of 2 seconds to enable uniformity in feature extraction. Shorter segments help reduce the workload needed for

computation, while increasing the accuracy of the I-vector and X-vector models [7].

3.2 FEATURE EXTRACTION

3.2.1 I-vector Extraction:

An I-vector is derived using the Total Variability Model (TVM), a model aimed at capturing both speaker and channel variation in a unified low dimensional representation. The procedure consists of:

- *Frame-Level Feature Extraction*: took MFCCs from every audio piece, which are the coefficients and parameters that encode different features, that characterize the auditory signal [11] [13] [14].
- *GMM-UBM Training*: A GMM with 512 components was trained to be UBM model as a universal background model, which is a type of model designed to train a model that can distinguish between some speaker dependent features.
- *Total Variability Space*: These coefficients are projected into a unified low dimensional space, referred to as I-vector space. Each I-vector contains information related to a particular speaker and indicates an entire audio segment thereby compressing the size of the representation.

3.2.2 X-vector Extraction:

Embeddings from deep neural networks lead to the vectors known as x-vectors. The embedding involves:

- *Feature Representation*: The time delay neural network (TDNN) is inputted with MFCCs. Large datasets like VoxCeleb are used for pre-training the network so that it learns speaker discriminative features [14].
- *Layer-Wise Aggregation*: The TDNN has numerous layers, hence giving it the ability to model multiple characteristics of a speaker. The output from the last layer is averaged and then normalized to give the embedded x-vector [15].
- *Speaker Embedding*: The x-vectors are speaker characteristics for each segment. These embeddings are less prone to noise and speaker variability than the I-vectors. X-vectors are extracted as fixed-length embeddings, which reflects the speaker characteristics of each segment.

3.3 SPEAKER DIARIZATION PROCESS

3.3.1 Segmentation:

A combination of sliding Window Approach with a window size of 1.5 seconds and a shift of 0.5 seconds is employed to cut the preprocessed audio into self-contained sections. Individual segments are contained by all the features extracted.

3.3.2 Clustering Algorithms:

After feature extraction, the segments are grouped based on the speaker identity using clustering techniques:

- Agglomerative Hierarchical Clustering (AHC) method: This method takes each segment as an individual cluster and combines any segments that appear to be similar, until the desired number of speakers is obtained [16].
- K-means Clustering: The K-means serves as a comparison baseline. While it is notable, it is easier as a computing resource, K-means has difficulties dealing with segments of speech of uneven length. Evaluates how similar speaker embeddings are. This process increases the probability of correctly classifying the speaker before and after clustering by increasing the differences between various speakers and therefore helps to enhance the cluster outputs [4].

3.4 EVALUATION METRICS

The performance of the diarization system is evaluated using: Diarization Error Rate, (DER), DER is claimed to be a proportion of time that each speaker is flagged erroneously. The greater the DAS, the better performance is perceived. The ratio of overlap instead of ratio of union for the purpose of evaluation is the Jaccard Error Rate (JER) metric.

Table.1. Comparative Analysis

Aspect	I-vectors	X-vectors
Feature Extraction	Based on GMM-UBM and Total Variability Model	Uses DNNs (TDNN architecture) for high-level embeddings
Robustness to Noise	Limited, especially in noisy environments	High, can handle noisy and overlapping speech
Short Utterance Handling	Less effective due to data requirements	Effective, captures speaker characteristics quickly
Computational Cost	Low, efficient for real-time processing	High, requires significant computational resources
Best Use Case	Controlled environments, call centers	Noisy environments, meetings, forensic analysis.

4. EXPERIMENTAL SETUP

This subsection provides an account of the practical aspect of the speaker diarization using i-vectors and x-vectors with particular emphasis on the test environment, tools and evaluation parameters.

4.1 SYSTEM CONFIGURATION

The following testing conditions and parameters were used in the computer.

- Hardware: Intel core i7-10750H CPU, 32GB RAM, NVIDIA GeForce GTX 1660 GPU
- Software: Operating system: Windows 10; Programming Language: Python version 3.8; Libraries: Pytorch, Kaldi and Scikit learn; Toolkit/Resources: The Kaldi Speech Recognition Toolkit was applicable in feature extraction and model training.

This specification is useful so that both the I-vector and X-vector models are trained and tested in uniform conditions which will reduce the variability in metrics used to measure performance.

4.2 IMPLEMENTATION PIPELINE

4.2.1 *Data Loading and Preprocessing:*

- ***Data Set Preparation***: AMI Meeting Corpus and VoxCeleb datasets were loaded and preprocessed. Each audio file was segmented using a Voice Activity Detector (VAD) to delete non-speech segments.[13],[14]. In this case, files that were conforming to 16 kHz mono WAV files were used as standard input. Snyder et al (2018).
- ***Noise filtering*****:** To reduce background noise, spectral gating was employed, using the Weiner filter algorithm [12]-[16]. Feature Extraction
- **I-vector Extraction:** For this purpose, MFCCs were obtained from each segment and introduced in GMM-UBM model [15]. The system was set up to produce a hundred-dimensional I-vectors to present each of the speech segments in a compressed style.
- **X-vector Extraction:** A neural network based on TDNN architecture was trained with the Kaldi toolkit as per Snyder et al [7]. X-vectors of an embedding size of five hundred and twelve were extracted as the speaker related information at a greater level.
- ***Clustering***: Agglomerative Hierarchical Clustering (AHC): It was used to cluster different segments based on the speakers' voice characteristics using PLDA scoring [14]-[16].
- ***K-means Clustering***: Used in this study for the justification of the baseline, embedding techniques were adopted on I-vector/x-vector where embeddings were traced using the Euclidean distance method [6].

4.3 EVALUATION METRICS

- **Diarization Error Rate (DER):** It is used to express the proportion of attributable speaker time which has been attributed incorrectly, its expression is defined as [1]:

DER=(Speaker Error+False Alarm+Missed speech)/(Total Time) (1)

- **Jaccard Error Rate (JER):** Focusing on the accuracy of segmentation where there is a likelihood of nodal speech segments which probably overlap ensures that wider evaluation metrics are taken into account [7].

4.4 BENCHMARK TESTING AND CASE STUDIES

- *Scenario 1: Multi-Speaker Meeting Transcription*: The first set of tests were conducted using AMI Corpus with different noise and overlapping speakers' conditions. The results showed much better performance of X-vectors in the noisy segments [12].
- *Scenario 2: Call Center Interaction Analysis*: These models have been tested with artificial call center datasets. Clustering the I-vectors proved to be efficient, while the overlapping dialogue posed challenges to Kaldi models [10].

5. RESULTS AND DISCUSSION

In this section, the results collected from the speaker recognition algorithm implementations using I-vector and X-vector are presented. The evaluation of the results is performed in terms of Diarization Error Rate (DER) and Jaccard Error Rate (JER), followed by a qualitative analysis.

5.1 QUANTITATIVE ANALYSIS

Comparison of the derivation of DER metrics, specifically DER and JER. Doubled DER values for both methods were computed on the AMI Meeting Corpus and VoxCeleb datasets. The findings are summarized in the table below:

Table.2. Quantitative Analysis

Method	AMI Corpus DER (%)	VoxCeleb DER (%)
I-Vector	16	22
X-Vector	9	11

Method AMI Corpus DER in percentage scale VoxCeleb DER in percentage scale. X- vector's consistency demonstrates consistent outperformance of I-vectors across both datasets. Lower DER confirms that X-vectors provide greater protection to speaker variance and use cases in a low signal-to-noise ratio environment. This variant of performance was accentuated in the dataset VoxCeleb which includes a wider range of audio conditions. This indicates the effectiveness of X-Vectors in practical challenges.

5.2 COMPARISON OF JER

The table showing JER related to the accuracy of segmentation are as follows:

Table.3. Comparison of Jaccard Error rate (JER)

Method	AMI Corpus JER (%)	VoxCeleb JER (%)
I-Vector	15	21
X-Vector	8	8

5.3 KEY OBSERVATIONS

- I-vectors techniques were good but applicable in controlled environments, however their use on short utterances background noise were problematic because that is how the real-world record.
- X-vectors recorded a lower JER overall, even in scenarios where overlapping speeches were present. This evidence is consistent with existing research that X-vectors can capture complex speaker impersonation characteristics better than others.

5.4 QUALITATIVE ANALYSIS

5.4.1 *Pragmatics of Overlapping Speech:*

Overlapping speech is in our view one of the most common sources of errors in diarization tasks. From our experiments noted that:

- Out of all the models, X-vectors performed the best regarding overlaps because DNN embeddings enabled these models to learn long-term speaker features [13].
- They noted that I-vectors made most errors with overlapping segments and increased error rates. This concurs with

conclusions drawn from the observations of the DIHARD Challenge [14].

5.4.2 Computational Efficiency:

X-vectors as shown in the results of this study stand superior to all other speakers, however the cost of energy on I-vectors are less:

- There are considerable physical loads as well as time delays that are experienced in the process of training the TDNN models associated with X-vector extraction.
- On the other hand, I-vectors are less burdensome in terms of processing power and are thus useful in time-sensitive applications even though there may be some level of accuracy compromise [15].

6. CONCLUSION

Speaker diarization is the task of answering the question "who spoke when" and is crucial in many areas such as meeting transcripts, call center division, and even media centers. In this study, assessed the performance of two new feature extraction techniques, I-vectors and X-vectors, in performing speaker diarization tasks under real-world conditions. The experiment used the AMI Meeting Corpus, VoxCeleb, and a variety of real-world environments to evaluate the pros and cons of these methods.

X-vectors were able to outperform I-vectors in each instance, with X-vectors achieving a Diarization Error Rate of 8% on the AMI Corpus and 8% on VoxCeleb, while I-vectors recorded 16% and 22%, respectively. With the Jaccard Error Rate, X-vectors achieved 9% on AMI and 11% on VoxCeleb, whereas I-vectors received 16% and 22%, respectively. This highlights the strength of X-vectors in coping with overlapping speech, noise, and short speech in real-world devices.

I-vectors remain a viable solution for instant applications or those with restrictions in processing power or real-time requirements. However, the presence of competing speech and noisy backgrounds greatly hindered their capability in diverse or uncontrollable environments, making them ineffective in broader and noisier settings.

According to results, observed improvements in the quality of features extracted for systems built on both I-vectors and X-vectors, thanks to robust preprocessing steps such as noise removal and voice activity detection.

REFERENCES

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 2, pp. 356-370, 2012.

[2] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech communication*, Vol. 17, No. 1, pp. 91-108, 1995.

[3] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 4, pp. 788-798, 2010.

[4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", *Proceedings of International Workshop on Odyssey Speaker and Language Recognition*, pp. 1-6, 2010.

[5] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification", *Proceedings of International Conference on Interspeech*, pp. 999-1003, 2017.

[6] Nakagawa Seiichi, Longbiao Wang and Shinji Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 4, pp. 1085-1095, 2011.

[7] Milner Rosanna and Thomas Hain, "DNN Approach to Speaker Diarisation using Speaker Channels", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-7, 2017.

[8] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey and S. Watanabe, "Diarization is Hard:Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge", *Proceedings of International Conference on Interspeech*, pp. 2808-2812, 2018.

[9] Q. Wang, C. Downey, L. Wan, P.A. Mansfield and I.L. Moreno, "Speaker Diarization with LSTM", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 5239-5243, 2018.

[10] K.V. Khadar, R.K. Aljinu, Sunil Kumar and V.V. Sameer, "Speaker Diarization based on X Vector Extracted from Time-Delay Neural Networks using Agglomerative Hierarchical Clustering in Noisy Environment", *International Journal of Speech Technology*, pp. 1-14, 2024.

[11] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij and M. Kronenthal, "The AMI Meeting Corpus: A Pre-Announcement", *Proceedings of International Workshop on Machine Learning for Multimodal Interaction*, pp. 28-39, 2005.

[12] A. Nagrani, J.S. Chung and A. Zisserman, "Voxceleb: A Large-Scale Speaker Identification Dataset", *Computer Science*, pp. 1-6, 2017.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian and P. Schwarz, "The Kaldi Speech Recognition Toolkit", *Proceedings of International Workshop on Automatic Speech Recognition and Understanding*, pp. 1-7, 2011.

[14] Mateju Lukas and J. Malek, "Overlapped Speech Detection in Broadcast Streams using X-Vectors", *Proceedings of International Conference on Interspeech*, pp. 1-6, 2022.

[15] Snyder David, G. Sell, K. Sanjeev and D. Povey, "X-vectors: Robust DNN Embeddings for Speaker Recognition", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-6, 2018.

[16] Zhang Yingjie and Liu Liu, "Multi-Task Learning for X-Vector based Speaker Recognition", *International Journal of Speech Technology*, Vol. 26, No. 4, pp. 817-823, 2023.