Multimedia Tools and Applications (2024) 83:54433–54448
https://doi.org/10.1007/s11042-023-17017-x

Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization

Aishwarya Gupta[1] · **Archana Purwar**[1]

Received: 11 November 2022 / Revised: 13 August 2023 / Accepted: 8 September 2023 /
Published online: 5 December 2023

Abstract
In this digitally-driven culture, the need and demand for diarizing online meetings, classes, conferences, and medical diagnoses have increased a lot. Speaker Diarization, a sub-domain of Speaker Recognition has grown with the advent of neural networks in the last decade. Diarize generally refers to obtaining the duration of individual speakers in any event. Researchers have suggested various approaches for multiple-speaker diarization. However, it still suffers from a problem of various environmental noises, and non-speech sounds like laughter, murmuring, clapping, etc. in the datasets. Hence, this paper proposes an improved speaker diarization pipeline to deal with the noise present in a dataset having multiple speakers. This improved diarization pipeline uses Bi-directional Long Short-Term Memory (Bi-LSTM), based speech refinement pre-processing module, and Modified Spectral Clustering with Symmetrized Singular Value Decomposition (MSC-SSVD). MSC-SSVD is used to cater to the problem of spectral clustering in large datasets. The proposed diarization pipeline is evaluated using the publicly available VoxConverse dataset. The Diarization Error Rate (DER) obtained after experimentation are 37.2%, 37.1%, and 43.3% respectively for three batches of dataset under study. The results are also compared with the baseline system and significant change in DER by 6.1%, 4.7%, and 7% respectively for three batches is observed.

Keywords Speaker Diarization · Speech Refinement · Bi-directional Long Short-Term Memory (Bi-LSTM) · Skip U-Net Connections · Singular Value Decomposition · Spectral clustering

✉ Aishwarya Gupta
aishgupta.93@gmail.com

Archana Purwar
archana.purwar@jiit.ac.in

[1] Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

1 Introduction

Speaker Recognition and Diarization are the domains that deal in the recognition of human voices but the former lets you distinguish between the speakers only whereas the latter helps in finding speaker count along with their duration of speech. Diarization provides us with the specific time of each speaker during any multi-speaker conversation. It mainly deals with the problem of "Who spoke when?" [2, 4]. But finding a few speakers in any conversation is an easily achievable task; a challenge arises when it comes to a multiple-speaker scenario. Extract multiple speakers' duration from meetings, seminars, conferences, discussions, and telephonic conversations. Another complex task is to distinguish various speakers' speech from various background noises such as clapping, murmuring, laughing, reverberation, and overlapped speech as well [3]. These noises create disturbances and affect the performance of the complete diarization pipeline.

Generally, the diarization process can be termed as unsupervised or supervised in terms of segmentation, embedding, and clustering behavior. Initially, unsupervised diarization system [25] with various types of embeddings (i-vector [3, 13, 26], x-vector [11, 12, 27], d-vector [14, 15, 28, 30] embeddings) and different types of clustering (Mean Shift [16, 17], Agglomerative Hierarchical Clustering [3, 18, 22], Spectral Clustering [19–22, 30]) were known. Recently, a shift from unsupervised to supervised and now to fully supervised diarization system like UIS-RNN (Unbounded Interleaved State Recurrent Neural Network) [29] has scaled up the research in online and offline processes for diarization. All this is introduced after the rise in the development of deep neural network architectures. Likewise, in [30] Wang Q. et. al. introduced a diarization system using LSTM (Long Short-Term Memory) neural network using various offline and online clustering algorithms with d-vector embeddings. Speaker diarization system (pipeline) present in literature comprised of the following sub-modules [2];

1. **Speech Activity Detection**: It divides speech segments from non-speech segments such as noise, reverberation, etc. In this module important features are being extracted like Mel Frequency Cepstral Coefficient (MFCC), zero crossing rate, spectral features, etc. [1]. Thereafter, speech separation on the speech part is executed. A classifier model is used to predict whether the input frame is speech or not. Earlier Gaussian Mixture Models (GMM) and Hidden Markov Models [5] were used for this detection but now after the introduction of Deep Neural Networks (DNN) has performed well in it.
2. **Speech Segmentation**: Small segments from larger chunks of data are being formed to make the speaker assignment processing easy. These segments obtained from speaker change point detection works better than any other method. Also, these are used to specify speaker labels. Various techniques like Generalised Likelihood Ratio [6], and Bayesian Information Criterion were used for segmentation [7, 8] previously. Before i-vector existed Kullback–Leibler and information change rate were popular methods to calculate distance between speech segments in the last decade. With the introduction of i-vectors and d-vectors uniform segmentation prevailed to date. Fixed window length and overlap length are considered for uniform segmentation [9].
3. **Embedding Extraction:** Previously GMM and GMM-UBM (Universal Background Model) models were preferred for the speaker representation until Joint Factor Analysis and i-vector were used. JFA overcame problems faced by MAP (Maximum-A-Posterior) like channel and background noise [1]. Then, neural representation techniques took all

over and reformed the process [10]. Various combinations of x-vectors [11, 12] and i-vectors [13], exist but d-vectors [14, 15, 30] are still ruling.

4. **Clustering:** This module provides speaker count by labeling them with clusters separately. As soon as the embeddings are fed into the pipeline; the clustering algorithm is implemented. Initially mean shift clustering algorithm [16, 17] and agglomerative hierarchical clustering [3, 18] were used in general with most representation techniques. In diarization recently used clustering is a spectral clustering algorithm [19–21]. Many of its improved variants [23, 24, 58–60] are working well with deep learning models.

But besides these 4 modules of a diarization pipeline, there are pre and post-processing modules [1]. They both help in decreasing the complexity within the system by refining and smoothening the input and output of the diarization pipeline respectively. With the emergence of multi-disciplinary large datasets like VoxCeleb, VoxConverse and other Meeting corpuses various challenges like unwanted background noise, multiple speakers (more than 10 or 15), lengthy audio files, overlapping segments, processing computation, low resource compatibility, etc. came into existence. Sun L. et. al., in [3] addressed the noise issue by using an LSTM-based speech denoising model with i-vector embedding using AHC was introduced by. In [30] a d-vector-based LSTM model was suggested by Wang Q.et. al., and compared to i-vectors it proved to be more robust and effective for the diarization system. In 2020, authors in [40] proposed LSTM based speech enhancement block for reducing the noises using densely connected progressive learning.

However, the pipeline suffers any discrepancies related to the removal of background noise with laughter, clapping and murmuring, etc., and after that dealing with large datasets like VoxConverse heavy computation is another challenge. To overcome both of these challenges of working with large datasets this paper has proposed a modified speaker diarization pipeline. It contains a modified pre-processing module in which a speech refinement model using Bi-LSTM Skip U-Net connection network for noise reduction is added. Along with that paper has also suggested a solution to overcome the slow computation problem of spectral clustering on large datasets by symmetrizing the affinity for smooth calculation and evaluating eigenvector using singular value decomposition.

The rest of the paper is organized as follows. In section 2 background study concerning both modified modules is presented. In sub-section 2.1 a brief review of speech enhancement models is discussed whereas in sub-section 2.2 various modifications done in spectral clustering from traditional to latest are mentioned. Section 3 will give highlights about the material and methods used in the proposed system pipeline. Then, in section 4 proposed speaker diarization pipeline is introduced with a modified version of both pre-processing and clustering modules in detail. The metric used, datasets and the experimental setup are specified in section 5. The implemented results are shared in section 6. Finally, conclusions and future scope is discussed in section. 7.

2 Background study

2.1 Review of speech enhancement methods

Enhancement in the speech domain is one of the most essential pre-processing steps. It helps in the reduction or removal of various noises such as background, applauding,

laughter, reverberation, etc. to obtain improved speech signals. It is segregated into 2 domains i.e., *Frequency Domain and Time Domain.*

- **In the frequency domain**, single-channel speech enhancement [31], frequency domain linear predicting (FDLP) [32], and conventional Fast Fourier Transform magnitude spectra [33] have been replaced by wavelet threshold multi-taper spectra [34, 35] which employ significant improvement in SNR (signal-to-noise ratio) metric of evaluation.
- **In time domain** features earliest speech enhancement technique used was spectral subtraction [36], in which noise spectrum is extracted from noisy spectrum to obtain the clean speech spectrum. It helps in improving quality and intelligibility. Thereafter many variants of spectral subtraction approached like spectral over subtraction, multi-band spectral subtraction, iterative spectral subtraction, and finally Wiener filtering which replaced all. Mainly, there are the following time domain techniques employed for speech enhancement.
  - **Spectral Techniques** [37]. In this technique, authors have discussed the addition of 2 factors (spectral over subtraction and spectral floor) in spectral over subtraction which improved the basic algorithm, but remnant noise could not be removed with this. So, multiband spectral subtraction was introduced for controlling real-world noise. This type of spectral subtraction is implemented over 4 equally spaced frequency bands. To customize the noise removal, process an additional band subtraction element had been added. It also provided control over noise subtraction at each band. Another one is **Wiener Filtering Technique.** It improved spectral subtraction by minimizing the mean square error (MSE) between the original and assessed signal. It uses a fixed gain function for the calculation of accuracy which degrades the quality of clean speech signal. Finally, iterative spectral subtraction came as an improvement of wiener filtering in which an output of the enhanced speech signal is used as input for the next signal in the process. In 2013, Abd-El Fattah, et.al. [38] improvised normal wiener filtering to an adaptive wiener filtering by targeting the drawback of the former one i.e., spectral subtraction is being applied over stationary signals but this adaptive one learns sample-to-sample filter response. Finally, **Gating Technique is** used to reduce **noise** generally in the music industry [57] for many years. It uses a gate that monitors the audio level. In [50] spectral gating was applied on the pre-processing module of the speaker diarization pipeline and hence received an improved performance. Table 1 shows how gating techniques help in reducing noise present in raw audio files on the VoxConverse dev set.
  - **Deep Learning-based Techniques**: Recently various neural network-based filtering algorithms such as Convolution Neural Network (CNN) based Speech Enhancement method [39] and Kalman Filter based Deep Neural Network (DNN) [40] have been implemented. Neural network models performed well in noisy conditions on

Table 1 Average value of components of DER (Diarization Error Rate) & DER % before and after applying spectral gating on VoxConverse dev set [50]

Components	Average value with noise	Average value without noise
False Alarm	11.25	8.81
Confusion	9.41	7.78
Missed Detection	2.41	4.86
DER%	74	54.6

both metrics of quality and intelligibility i.e., Perceptual Evaluation of Speech Quality (PESQ) and short-time objective intelligibility (STOI) respectively as noted by authors in [39, 40]. Hence, deep learning paved the way for some more improvement in reducing these background disturbances and unnecessary noises. Also, Authors in [66, 67] discussed U-Net architecture which has shown improved results in terms of speech enhancement.

The enhancement techniques discussed above are based on filtering, gating, and neural network methods. Gating techniques have been earlier used by music enthusiasts researching the domain. But from recent research, it can be observed that neural network has performed better than both other techniques [37, 38, 40]. Still, there are some distracting noises like laughter, murmuring, clapping, etc. which are sustained and could not be reduced or removed from audio files in the case of multidisciplinary datasets.

2.2 Review on spectral clustering

Clustering is an important module of the diarization pipeline in which speakers are labeled through clusters. Spectral clustering has been one of the finest clustering algorithms that overcame the drawback of k-means clustering of not being compatible with anisotropic data (spherical or round cluster formation), it is sensitive to initializing centroids prior and converging easily locally. The basic steps involved in the traditional spectral clustering algorithm by Ng A., et. al., [42] are as follows:

- Form an affinity matrix $A_{ij} = \exp(\frac{-\|s_i - s_j\|^2}{2\sigma^2})$, considering $i \neq j$ and $A_{ij} = 0$.
- Let D be the diagonal matrix whose elements (i, j) be the sum of $A's$ i-th row and construct the Laplacian matrix $L = D^{-1/2}AD^{-1/2}$.
- Find the largest k eigenvectors of L from $x_1, x_2, \ldots \ldots \ldots, x_k$ and form a column.
- Form matrix Y from X by renormalizing each of $X's$ row to have unit length.
- Apply k-means clustering and obtain cluster labels separately for each cluster.

Some variants of spectral clustering algorithms came with different modifications and implemented different machine learning algorithms and neural network systems. Some of them are discussed below:

A. Unnormalized Spectral Clustering

This was the very first change made by Andrew Ng and discussed by Luxumberg in [19]. As the name suggests Unnormalized Spectral clustering computed its unnormalized Laplacian matrix (which has satisfied some properties of being symmetric, semi-definite, and has non-negative eigenvalues). The basic steps of unnormalized spectral clustering are mentioned as follows:

- With W as a weighted adjacency matrix, D as a diagonal matrix, and L as unnormalized Laplacian matrix, it is calculated as $L = D - W$
- First k eigenvectors are calculated $u_1, u_2, \ldots \ldots, u_k$ of$L$.

- A column matrix U is formed from these eigenvectors.
- Clusters labels are obtained using k-means algorithm.

B. Normalized Spectral Clustering

Unlike the above normalized spectral uses normalized Laplacian matrix (L_{sym}(symmetric) or L_{rw}(random walk)) and proceed further with similar steps as in the case of unnormalized clustering [19].

- **For random walk Laplacian matrix:** When Eigenvectors(u) are calculated from generalized eigenproblem with D as diagonal matrix and λ is an eigenvalue of L_{rw} from $Lu = \lambda Du$
- **For symmetric matrix:** Here, λ is an eigenvalue of L_{sym} with Eigenvector $w = D^{-1/2}u$ First k eigenvectors are calculated from (L_{sym} and then normalized by 1 to form another matrix T.

C. Self-tune Spectral Clustering

In self-tune authors tried to automate a complete spectral clustering algorithm [24] by modifying a few steps in traditional spectral clustering. The self-tune spectral clustering steps are given below:

- Compute local scale σ_i for all the points to be clustered.
- Form local scaled affinity matrix $\hat{A}$ and then construct normalized affinity matrix $L = D^{-1/2}AD^{-1/2}$.using diagonal matrix D.
- Form matrix $X = [x_1, \ldots .., x_c]$ with C largest eigenvectors of L.
- Rotate the eigenvectors for maximal sparse representation.

D. Auto-tune Spectral Clustering

This is the latest automated version of spectral clustering developed by Park T. J. et. al., [23] in 2019. Here, hyperparameters used in clustering such as p (the threshold used for row-wise binarization) and g_p(normalized maximum eigengap value) are being tuned automatically. These both share a linear relationship and thus play an important role in the calculation of good DER as p/g_p can suggest a value of p from which presumably DER can be the lowest. As a result, it showed a better performance of SC with cosine similarity than AHC when coupled with PLDA (Probabilistic Linear Discriminant Analysis) model. Other variations were also introduced like Constrained Spectral Clustering [58], Scalable Constrained spectral clustering [59], Multi-view Spectral clustering [61], Multiclass spectral clustering [60], and many others.

However, even after so many modifications in spectral clustering it still suffers some drawbacks for large and noisy datasets such as VoxCeleb [63] and VoxCeleb2 [64] and VoxConverse [68]. It faces performance degradation in terms of speed, cost computation, and heavy calculation of informative eigenvector selection within the algorithm. To overcome these setbacks spectral clustering is facing, another better technique of eigenvector decomposition is suggested in this paper i.e., using singular value decomposition, the eigenvector selection gets easy and the computational load is balanced for the complete pipeline which affects the overall performance as well. The complete detailed modified clustering is discussed in module M5 of Sect. 3.

3 Methodology

3.1 Bi-LSTM model with skip U-Net connections

Speech denoising became a crucial pre-processing step in the field of diarization after the advent of large and multidisciplinary datasets which contains multiple speaker conversations in the field of news broadcast, interviews, conferences, meeting, and discussions [1]. Various types of noises disrupt the clean and effective diarization.

These noisy disturbances are present everywhere. Thus, proper reduction or removal technique by pre-processing audio files is required to achieve refined audio for further evaluation.

Demucs architecture developed by Défossez A. et. al., [54] has shown great results for music source separation in the waveform domain [53]. It helped in enhancing the quality of speech by suppressing environmental noise, reverberations, background noises, etc. up to a great extent. It has inherited its structure from Wave-U-Net [55] which was introduced for audio source separation from music datasets.

The paper has adapted Demucs architecture for a speech refinement module that consists of a Convolutional Encoder, Bi-LSTM, and Convolutional Decoder in which the encoder-decoder is linked with skip U-Net connections. This architecture can reduce stationary as well as non-stationary noises. It has also improved the naturalness of the audio. A brief description of the architecture used is given below.

- It consists of L encoder layers, numbered from 1 to L whereas decoder layers are in reverse order from L to 1.
- Firstly, an encoder network has an internal structure has a convolution layer with 2^{i-1}H output channels with Recurrent Linear Units (ReLU) activation, then 1×1 convolution with 2^{i}H output channels, and finally Gated Linear Units (GLU) activation that converts the number of channels to 2^{i-1}H again.
- Secondly, the Bi-LSTM network consists of 2 layers and 2^{L-1} H hidden layers. L1 loss function has been used over the waveform and STFT loss spectrogram magnitude for the architecture.
- Lastly, a decoder network takes input from a neural network and gives an output as a clean signal. It takes 2^{i-1} H channels and applies a 1*1 convolution with 2^{i} H channels which are followed by a GLU activation of 2^{i-1}H channels and transposed convolution with K = 8 (kernel size), S = 4 (stride) and output channels = 2^{i-2}H channels.
- Finally, at last, a ReLU function is applied for all layers except the last one where no ReLU is applied and only a single channel is at the output. The role of skip-u net connections is that it connects $i-th$ layer of an encoder to a $i-th$ layer of a decoder.

A combination of Bi-LSTM network with Skip U-Net Connections is employed in the proposed diarization pipeline described in Sect. 4.

3.2 Singular Value Decomposition (SVD)

SVD is a matrix factorization technique that decomposes a single matrix A into 3 matrices as mentioned below

$$A = U \bullet \Sigma \bullet \mathrm{V} \tag{1}$$

$$A = AA^T \bullet \Sigma \bullet A^T A \tag{2}$$

where A is an $m \times n$ matrix, U is an $m \times m$ left singular matrix, V is an $n \times n$ right singular matrix, and Σ is diagonal matrix with $m \times n$ size. A^T is the transpose of matrix A, m denotes the number of rows and n denotes the number of columns.

4 Proposed system description

This paper proposes an improved speaker diarization pipeline as shown in Fig. 1 for an audio system by employing BILSTM skip U-Net Model in pre-processing module for speech refinement and developing Modified Spectral Clustering (MSC) with SSVD (Symmetricity and Singular value decomposition for clustering module. All the modules of the improved speaker diarization pipeline are explained in detail thereafter in sub-sections.

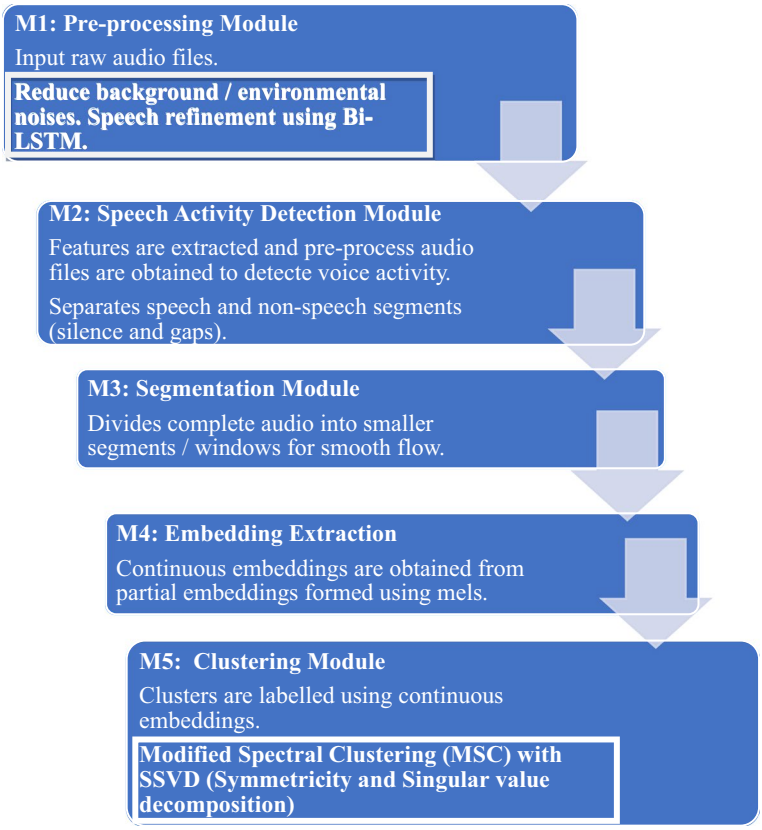

Fig. 1 Proposed speaker diarization pipeline

4.1 M1: Pre-processing module: Speech refinement model using Bi-LSTM

Pre-processing is the first module of our proposed diarization pipeline. It refines and enhances the raw and noisy audio files. The paper proposed a speech refinement model using Bi-LSTM with Skip U-Net connections in this module. The brief network architecture of the refinement module is depicted in Fig. 2.

The Bi-LSTM model is capable of handling sequential and time series data very well. As audio data captures long-range dependencies and finds temporal patterns in both directions. It will be easy to identify noisy patterns within files in a short period using Bi-LSTM rather than an LSTM network. Likewise, Skip U-Net connections extract both high-level and low-level features from an audio signal. It helps in retaining the same structure and characteristics of clean speech or sound which is beneficial for denoising purposes. The combination of Bi-LSTM with Skip U-Net connections works better in understanding the structure of audio and distinguishing between noise and actual signal components. They both combinedly creates a robust network for speech refinement purpose. The pseudocode for the proposed speech refinement module is given below:

a) Input a raw audio waveform to a pre-processing module where, using the pre-trained models a state dictionary is loaded.
b) Then, set all the parameters according to our model's requirement. Pass the waveform to the speech refinement model.

   i. Firstly, it will enter an encoder network, where gated linear units (GLU) are used to boost performance and at the output, both recurrent linear units (ReLU) and GLU activation is being applied for enhancement.
   ii. Then, the processed signal passes through the Bi-LSTM network with 48 hidden layers.

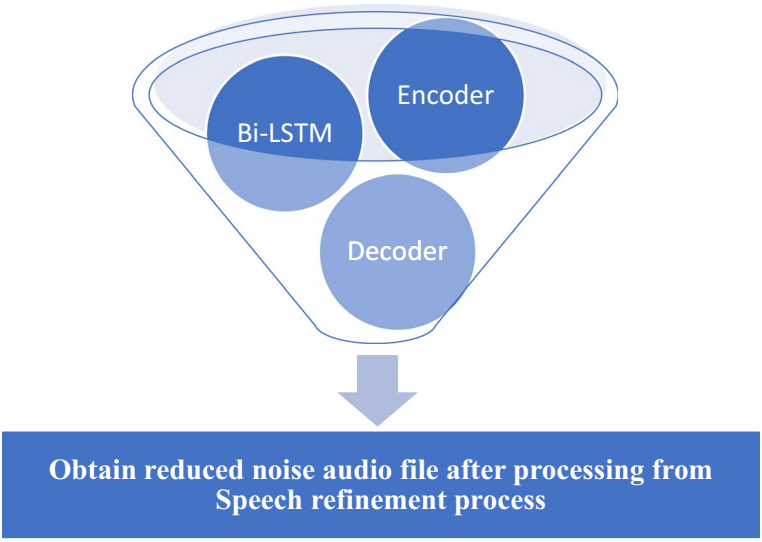

Fig. 2 Speech refinement model using Bi-LSTM

iii. Lastly, the decoder network takes the channels in and provides a clean signal after performing convolution, gated linear activation, and transposed convolution on it and passes through a recurrent linear activation at the end to provide an enhanced signal for the further processing through a speaker diarization pipeline.

Now, refined audio files after the reduction of background noises are received and passed for speech and non-speech detection to the next module.

4.2 M2: Speech activity detection module

In this stage files obtained from the above module after pre-processing are converted to an array, and finally, L1 normalization is applied. The overlap detection is also a part of SAD but here in this framework, we have not considered the overlapping speech segments.

4.3 M3: Segmentation Module

During this phase, speech and non-speech segments are separated to further reduce the complexity by removing silences, gaps, and non-speech segments from the speech ones. These smaller segments help in smoothening the process by gathering the vocal segments in one place.

4.4 M4: Embedding extraction module

Using those speech segments, we receive embeddings from embed utterances, and then MFCC features are extracted. Obtain continuous embedding from partial embedding. In this paper, d-vector embeddings are extracted from the pre-trained PyTorch model Resemblyzer.

4.5 M5: Clustering module: Spectral clustering modification using MSC-SSVD

Now, continuous embeddings are received from the previous module and firstly number of clusters should be predicted to be given as an input to spectral clustering. Another modification of the proposed pipeline is implemented here in the spectral clustering algorithm. The need of modifying the traditional algorithm is just to use the basic dimensionality reduction technique and fasten the process of eigenvector calculation and decomposition. Firstly, the affinity matrix is symmetrized to obtain a transformed adjacency matrix and then singular value decomposition is applied for eigenvector calculation.

The steps of the modified spectral clustering- symmetrized singular value decomposition (MSC-SSVD) are as follows.

a) Form an affinity matrix A using cosine similarity measure ($A_{ij} = d(w_i, w_j)$); which is the distance between the 2 speakers embedding from 2 speech segments.
b) Then, a symmetric operation on the affinity matrix is applied and a transformed matrix is obtained. The affinity matrix will be transformed into an undirected adjacency matrix by taking an average of an original and transposed versions of the affinity matrix.

$$A_s = \frac{1}{2}(A + A^T) \tag{3}$$

c) Compute the eigenvalues and eigenvectors using Singular Value Decomposition and arrange the eigenvectors in descending order.
d) Apply the k-means clustering algorithm on the obtained spectral embeddings to get the number of cluster labels.
e) The obtained cluster labels signify the number of speaker counts.

Lastly, individual speaker labels are obtained after the final clustering module. The labels signify the speech duration of each speaker separately.

5 Dataset, experimental setup, and evaluation metrics

5.1 Dataset

The proposed speaker diarization pipeline is implemented on the VoxConverse dataset [64] which is made publicly available in 2020. It consists of 50 h of multi-speaker clips of human conversations in various forms of telephonic calls, broadcasting news and interviews, and other conversations.

The implementation using this VoxConverse dataset has been done in 3 batches of these files with different durations for adding up the variations and to get a complete understanding of different timing durations.

The reason for implementing the pipeline on this VoxConverse dataset is that it contains varieties of multi-speaker clips in many different scenarios which will be very helpful in gathering most of the variants in a single place. This helped us in analyzing multidisciplinary domains in a single audio dataset.

5.2 Experimental setup

The implementation is done on "NVIDIA-SMI 471.41, Driver Version: 471.41, CUDA Version: 11.4 with 8 GB RAM and 512 SSD NVIDIA GeForce". It was another challenge to run this heavy dataset in a low-resource environment. Pyannote. audio 2.0.1 has also been taken into use for the SAD task. Metric evaluation has been done using pyannote. metrics. Resemblyzer pre-trained embedding extractor is used for extracting d-vector embeddings.

5.3 Evaluation metric

To analyze our speaker diarization pipeline we used Diarization Error Rate (DER) [65]. It calculates the total percent of reference speaker duration that is not assigned correctly to a speaker. Here, the correctly assigned is nothing but a one-to-one mapping of speaker labels between their ground truth and hypothesis. Overlapping segments are ignored.

Diarization error Rate namely consists of 3 sub-components as mentioned in Equation 4.

- **False Alarm:** It refers to the percentage of scored time that a hypothesized speaker is labeled as a non-speech in the reference.
- **Missed Detection:** It refers to the percentage of scored time that a hypothesized non-speech segment corresponds to a reference speaker segment.
- **Confusion (Speaker error):** It refers to the percentage of scored time that a speaker ID is assigned to the wrong speaker

$$DiarizationErrorRate = \frac{\text{FalseAlarm} + \text{MissedDetection} + \text{Confusion}}{TotalReference} \tag{4}$$

6 Results and discussions

The proposed speaker diarization pipeline is evaluated on the VoxConverse dataset and results obtained after employing Bi-LSTM with spectral clustering are tabulated in Table 2 and compared with state of the art system which used an LSTM neural network with spectral clustering for the diarization pipeline [30]. According to the result, there is a significant decrease in DER after the addition of the Bi-LSTM model during pre-processing. An average of 3.8% decrease in DER can be seen.

Table 3 also shows a DER% after applying the proposed modified spectral clustering with symmetrized singular value decomposition. It also shows with state of art clustering algorithms like Particle Swarm Optimization k-means (PSO k-means), Agglomerative Hierarchical Clustering (AHC), and Spectral Clustering for the diarization pipeline. During experimentation, the LSTM network is combined with all clustering and without any refinement procedure. DER% has been reduced by a noticeable margin by our proposed MSC – SSVD technique from existing spectral clustering [30].

Finally, the comparison between the baseline system [30] and the proposed modified speaker diarization system has been compiled in Table 4. Along with these both proposed modifications separately are also evaluated. It is obvious from the results that the Bi-LSTM

Table 2 Results of DER% by Speech refinement using BiLSTM on VoxConverse dataset

Batches	LSTM + SC [30]	Speech Refinement using BiLSTM (our proposed) + SC
Batch 1	43.3	38.5
Batch 2	41.8	37.4
Batch 3	50.3	48.1

Table 3 Comparison of different clustering algorithms with LSTM network with our proposed Modified SC-SSVD on VoxConverse dataset

Batches	LSTM + PSO-k-means [68]	LSTM + AHC [68]	LSTM + SC [30]	LSTM + Modified SC-SSVD (Our proposed)
Batch 1	83.1	55.3	43.3	40.6
Batch 2	72.6	52.2	41.8	41.7
Batch 3	81.1	67.2	50.3	48.9

Table 4 Comparison of DER of our proposed pipeline with other state-of-art systems on the VoxConverse dataset (All the values are in % and the lower is better)

Batches	LSTM + SC [30]	LSTM + Modified SC-SSVD	BiLSTM + SC	BiLSTM + Modified SC-SSVD (Our Proposed)
Batch 1	43.3	40.6	38.5	37.2
Batch 2	41.8	41.7	37.4	37.1
Batch 3	50.3	48.9	48.1	43.3

model has made remarkable changes in DER when combined with the MSC-SSVD clustering technique. Overall, absolute change in DER comes out to be 6.1%, 4.7%, and 7% respectively for 3 batches. It depicts that background noise plays a significant role in degrading the quality of an audio file.

7 Conclusion

Diarization is the task of identifying and tracking the speaker's speech duration in an audio recording. Nowadays, it has spread its scope to speaker indexing, content structuring, and audio information retrieval. The paper aims to reduce the extraneous noises generated from non-speech sounds like clapping, murmuring, laughing, etc. This paper proposed an improved speaker diarization pipeline with a speech refinement module using Bi-LSTM with skip U-Net connection and an improved spectral clustering algorithm with symmetrized singular value decomposition. The DER obtained after implementing the proposed solution is 37.2%, 37.1%, and 43.3% respectively for 3 batches on the multi-disciplinary VoxConverse dataset. The results are compared with the baseline [30] approach and a significant decrease of 6.1%, 4.7%, and 7% is observed. The modified pipeline paved the way for retrieving audio files with reduced background and unnecessary noises. But the pipeline suffers problems in understanding similar voices at times.

The proposed pipeline can be extended to multimodal functionalities in other modes like videos and emotion recognition with audio. It has still vast scope for improvement in tackling speaker variability, adaptation, and real-time performance.

Data availability DAS: The datasets analyzed during the current study are publicly available from the link: https://www.robots.ox.ac.uk/~vgg/data/voxconverse/ and can be currently downloaded from [https://github.com/joonson/voxconverse] repository at GitHub.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S (2022) A review of speaker diarization: Recent advances with deep learning. Comput Speech Lang 72:101317. https://doi.org/10.1016/j.csl.2021.101317

2. Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O (2012) Speaker diarization: A review of recent research. IEEE Trans Audio Speech Lang Process 20(2):356–370. https://doi.org/10.1109/TASL.2011.2125954
3. Sun L, Du J, Jiang C, Zhang X, He S, Yin B, Lee C (2018) Speaker diarization with enhancing speech for the first DIHARD challenge. Interspeech
4. Sinclair M, King S (2013) Where are the challenges in speaker diarization?. In: 2013 IEEE International conference on acoustics, speech and signal processing. IEEE, pp 7741–7745
5. Sarikaya R, Hansen JH (1998) December). Robust detection of speech activity in the presence of noise. Proc ICSLP 4:1455–1458
6. Meignier S, Moraru D, Fredouille C, Bonastre J-F, Besacier L (2006) Stepby-step and integrated approaches in broadcast news speaker diarization. Comput Speech Lang 20:303–330
7. Chen S, Gopalakrishnan P (1998) Speaker, environment and channel change detection and clustering via the bayesian information criterion. In Proc. DARPA broadcast news transcription and understanding workshop, vol. 8. DARPA, pp 127–132
8. Delacourt P, Wellekens CJ (2000) Distbic: A speaker-based segmentation for audio data indexing. Speech Commun 32:111–126
9. Senoussaoui M, Kenny P, Stafylakis T, Dumouchel P (2013) A study of the cosine distance-based mean shift for telephone speech diarization. IEEE/ACM Trans Audio Speech Lang Process 22:217–227
10. Landini F, Glembek O, Matějka P, Rohdin J, Burget L, Diez M, Silnova A (2021) Analysis of the but diarization system for voxconverse challenge. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5819–5823
11. Snyder, D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5329–5333
12. Landini Federico et al. (2021) Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation, and analysis on standard tasks. Comput Speech Lang. https://doi.org/10.1016/j.csl.2021.101254
13. Sell G, Garcia-Romero D (2014) Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In: 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp 413–417
14. Kang W, Roy BC, Chow W (2020) Multimodal speaker diarization of real-world meetings using d-vectors with spatial features. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6509–6513
15. Novoselov S, Gusev A, Ivanov A, Pekhovsky T, Shulipa A, Avdeeva A et al (2019) Speaker diarization with deep speaker embeddings for DIHARD challenge II. In: Interspeech. pp 1003–1007
16. Comaniciu, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619. https://doi.org/10.1109/34.1000236
17. Stafylakis T, Katsouros V, Carayannis G (2010) Speaker clustering via the mean shift algorithm. ReCALL 2:7
18. Han KJ, Narayanan SS (2007) A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In: Interspeech. pp 1853–1856
19. Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17:395–416. https://doi.org/10.1007/s11222-007-9033-z
20. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. Adv Neural Inf Process Syst 14:849–856. https://dl.acm.org/doi/abs/https://doi.org/10.5555/2980539.2980649
21. Ning H, Liu M, Tang H, Huang TS (2006) A spectral clustering approach to speaker diarization. In: Ninth international conference on spoken language processing
22. Luque J, Hernando J (2012) On the use of agglomerative and spectral clustering in speaker diarization of meetings. In: Odyssey 2012-The speaker and language recognition workshop
23. Park TJ, Han KJ, Kumar M, Narayanan S (2019) Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. IEEE Signal Process Lett 27:381–385. https://arxiv.org/abs/2003.02405
24. Zelnik-Manor L, Perona P (2004) Self-tuning spectral clustering. Adv Neural Inform Proc Syst 17
25. Shum Stephen H et al (2013) Unsupervised methods for speaker diarization: An integrated and iterative approach. IEEE Trans Audio Speech Lang Process 21.10:2015–2028. https://doi.org/10.1109/TASL.2013.2264673
26. Rouvier M, Bousquet PM, Favre B (2015) Speaker diarization through speaker embeddings. In: 2015 23rd European Signal Processing Conference (eusipco). IEEE, pp 2082–2086
27. Toruk M, Bilgin G, Serbes A (2020) Speaker diarization using embedding vectors. In 2020 28th Signal Processing and Communications Applications Conference (SIU). IEEE, pp 1–4

28. Sun G, Liu D, Zhang C, &Woodland PC (2021) Content-aware speaker embeddings for speaker diarisation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 7168–7172
29. Zhang A, Wang Q, Zhu Z, Paisley J, Wang C (2019) Fully supervised speaker diarization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6301–6305
30. Wang Q, Downey C, Wan L, Mansfield PA, Moreno IL (2018) Speaker diarization with LSTM. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5239–5243
31. Nakanishi I, Nagata Y, Itoh Y, Fukui Y (2006) Single-channel speech enhancement based on frequency domain ALE. In: 2006 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, pp 4
32. Li W (2008) Effective post-processing for single-channel frequency-domain speech enhancement. In: 2008 IEEE International conference on multimedia and expo. IEEE, pp 149–152
33. Parchami M, Zhu WP, Champagne B, Plourde E (2016) Recent developments in speech enhancement in the short-timeFourier transform domain. IEEE Circ Syst Mag 16(3):45–77. https://doi.org/10.1109/MCAS.2016.2583681
34. Hu Y, Loizou PC (2004) Incorporating a psycho acoustical model in frequency domain speech enhancement. IEEE Signal Process Lett 11(2):270–273
35. Hu Y, Loizou PC (2004b) Speech enhancement based on wavelet thresholding the multi-taper spectrum. IEEE Trans Speech Audio Process 12(1):59–67. https://doi.org/10.1109/tsa.2003.819949
36. Boll SF (1979) Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans Acoust Speech Signal Process 27:113–120. https://doi.org/10.1109/TASSP.1979.1163209
37. Upadhyay N, Karmakar A (2015) Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. Procedia Comput Sci 54:574–584. https://doi.org/10.1016/j.procs.2015.06.066
38. Abd El-Fattah MA, Dessouky MI, Abbas AM et al (2014) Speech enhancement with an adaptive Wiener filter. Int J Speech Technol 53–64. https://doi.org/10.1007/s10772-013-9205-5
39. Pandey A, Wang DL, Fellow IEEE (2019) A new framework or CNN-based speech enhancement in the time domain. IEEE Trans Audio Speech Lang Process 27(7):1179–1188. https://doi.org/10.1109/taslp.2019.2913512
40. Yu H, Ouyang Z, Zhu WP, Champagne B, Ji Y (2019) A deep neural network based Kalman filter for time domain speech enhancement. In: 2019 IEEE International Symposium on Circuits And Systems (ISCAS). IEEE, pp 1–5
41. Sainburg T (2018) Noise reduction using spectral gating in python. Tim Sain-burg
42. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Adv Neural Inform Proc Syst 14
43. Xia W, Lu H, Wan Q, Tripathi A, Huang Y, Moreno IL, Sak H (2022) Turn-to-diarize: online speaker diarization constrained by transformer transducer speaker turn detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 8077–8081
44. Mihov SG, Ivanov RM, Popov AN (2009) Denoising speech signals by wavelet transform. Annual J Electron 6:2–5
45. Kumariss VSR, Devarakonda Dileep Kumar (2023) A Wavelet Based Denoising of Speech Signal. Int J Eng Trends Technol (IJETT) V5(2):107–115. ISSN:2231–5381
46. Kaladharan N (2014) Speech enhancement by spectral subtraction method. Int J Comput Applic 96(13):45–48. https://doi.org/10.5120/16858-6739
47. Karam M et al (2014) Noise removal in speech processing using spectral subtraction. J Signal Inf Process 5:32–41. https://doi.org/10.4236/jsip.2014.52006
48. Ahmad R, Zubair S, Alquhayz H, Ditta A (2019) Multimodal speaker diarization using a pre-trained audio-visual synchronization model. Sensors 19(23):5163. https://www.mdpi.com/1424-8220/19/23/5163
49. Ahmad R, Zubair S, Alquhayz H (2020) Speech enhancement for multimodal speaker diarization system. IEEE Access 8:126671–126680. https://doi.org/10.1109/ACCESS.2020.3007312
50. Gupta A, Purwar A (2022) Enhancing speaker diarization for audio-only systems using deep learning. In: Applications of artificial intelligence, big data and internet of things in sustainable development. CRC Press. pp 65–79
51. Das N, Chakraborty S, Chaki J, Dey N (2021) Fundamentals, present and future perspectives of speech enhancement. Int J Speech Technol 24(4):883–901. https://doi.org/10.1007/s10772-020-09674-2

52. Islam MR, Rahman MF, Khan MAG (2009) Improvement of speech enhancement techniques for robust speaker identification in noise. In: 2009 12th International conference on computers and information technology. IEEE, pp 255–260
53. Défossez A, Usunier N, Bottou L, Bach F (2019) Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254
54. Defossez A et al (2020) Real time speech enhancement in the waveform domain. Interspeech
55. Défossez A, Usunier N, Bottou L, Bach F (2019) Demucs: deep extractor for music sources with extra unlabeled data remixed. arXiv preprint arXiv:1909.01174
56. Stoller Daniel, Ewert Sebastian, Dixon Simon (2018) Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. URL https://arxiv.org/abs/1806.03185
57. https://www.kaggle.com/code/mauriciofigueiredo/methods-for-sound-noise-reduction/notebook.
58. Wang X, Qian B, Davidson I (2014) On constrained spectral clustering and its applications. Data Min Knowl Disc 28(1):1–30
59. Li J, Xia Y, Shan Z, Liu Y (2014) Scalable constrained spectral clustering. IEEE Trans Knowl Data Eng 27(2):589–593
60. Raj D, Huang Z, Khudanpur S (2021) Multi-class spectral clustering with overlaps for speaker diarization. In: 2021 IEEE Spoken Language Technology workshop (SLT). IEEE, pp 582–589
61. Huang Z, Zhou JT, Peng X, Zhang C, Zhu H, Lv J (2019) Multi-view spectral clustering network. IJCAI 2(3):4
62. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. Interspeech
63. Chung Joon Son, Nagrani Arsha, Zisserman Andrew (2018) Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622
64. Chung Joon Son et al (2020) Spot the conversation: speaker diarisation in the wild. arXiv preprint arXiv:2007.01216. Interspeech
65. (2017) Herve Bredin, pyannote. metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. Hypothesis 100(60):90
66. Kang Z, Huang Z, Lu C (2022) Speech Enhancement Using U-Net with Compressed Sensing. Appl Sci 12(9):4161. https://doi.org/10.3390/app1209416
67. Macartney Craig, Weyde Tillman (2018) Improved speech enhancement with the wave-u-net. arXiv preprint arXiv:1811.11307
68. Gupta A, Purwar A (2022) Analysis of clustering algorithms for Speaker Diarization using LSTM. 2022 1st International Conference on Informatics (ICI), Noida, India, pp. 19–24. https://doi.org/10.1109/ICI53355.2022.9786928