Speakers Identification Using Diarization Techniques

Vinod K. Pande(✉) and Vijay K. Kale

Dr. G. Y. Pathrikar College of Computer Science and Information Technology, MGM University, Aurangabad, Maharashtra, India
vinodkpande2014@gmail.com, vkale@mgmu.ac.in

Abstract. Research work analyses speaker voice identification and voice separation development methodologies and show an overview of the findings. Several speech recognition methods, such as Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ), Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), End-to-End Neural Diarization (EEND), Generative Adversarial Networks (GANs), Convolutional Neural Networks, and Audio Embeddiment, can be used for adaptive processing with multiple speakers identification in audio data. Additionally, we addressed the uses of speaker diarization, the potential for future development, and the databases used to evaluate diarization systems.

The speaker diarization method consists of seven steps, including input, front-end processing, speech activity detection, segmentation, speaker embedding, clustering post-processing, and output.

Speaker identification recognizes speakers during an audio conversion, a kind of speech recognition. Diarization of the speaker is a way of recognizing the speaker in a multi-speaker audio file. And The procedure of identifying who talks when in an audio recording is known as speaker diarization. The audio file includes information from conferences, broadcast news, and any other public gathering with many speakers.

Keywords: Speaker Diarization · End-to-End Neural Diarization(EEND) · Mel Frequency Cepstrum Coefficients (MFCC) · Generative Adversarial Networks (GANs) · Hidden Markov Model (HMM)

1 Introduction

Making a note or keeping a record of happenings in a diary is what it means to "diarize". Like maintaining a log, Speaker Diarization includes recording speaker-specific salient occurrences on multiparticipant (or many speakers) audio data. Throughout the Diarization process, the audio data would be divided and grouped into collections of speech segments with the same speaker identity/label. As a result, most notable occurrences, such as the transition between speech and non-speaking, detection of speaker turn changes, and speaker clustering, are

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 905–915, 2023.
https://doi.org/10.2991/978-94-6463-136-4_80

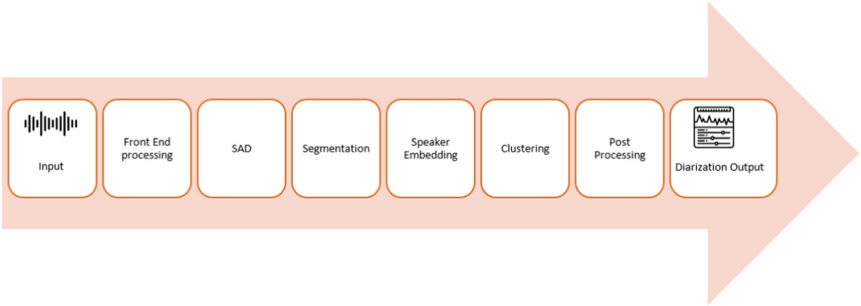

Fig. 1. Speaker Diarization Process

carried out automatically [1] The audio file containing voice data from multiple speakers in a meeting, broadcast news speech etc. The number of participants in the audio data or their identities is optional knowledge for the speech diarization process [1].

Speaker Diarization is identifying a speaker's start and end time in an audio file, together with the speaker's identity, i.e., who spoke when. Organising the audio stream into speaker turns and providing the speaker's real identity can boost the readability of an automatic speech transcription [2]. Speaker Diarization combines speaker segmentation with speaker clustering. The first seeks out speaker transitions in an audio stream. The second seeks to organise speech fragments according to speaker attributes [2].

Speaker diarization can be used effectively for indexing or analysing various audio data, including audio/video broadcasts from media stations, conference conversations, personal videos from online social media or handheld devices, court proceedings, business meetings, and earnings reports in a fintech company. This is because speaker diarization is innate to separate audio streams by speaker-specific events [1].

Speech recognition with speaker identification, speaker indexing, speaker retrieval, and meeting and lecture diarization are just a few of the many uses for diarization. Speaker diarization has drawn much interest from the speech community due to the rising volume of broadcasts, meeting recordings, and voicemails gathered yearly.

As seen in Fig. 1, Typical speaker diarization systems are made up of many separate sub-modules. First, several front-end processing methods are employed to reduce artefacts in auditory environments, including speech improvement, dereverberation, speech separation, and target speaker extraction. After that, speech from other sounds is separated using voice or speech activity detection (SAD).

The selected speech segment processes raw speech signals into acoustic characteristics or embedding vectors. The speech segments altered are categorised and labelled by speaker classes during the clustering stage, and the clustering

findings are enhanced further during the post-processing stage. In general, each of these sub-modules is optimised separately [1].

2 Techniques

The challenge of figuring out "who talked when" in an audio recording is addressed by speaker diarization. It is a key component of speech and speaker recognition pipelines and has a wide range of applications linked to speaker indexing data. The components of a general speaker diarization system are (a) a speech activity detection module that separates speech from non-speech sections and (b) speaker segmentation that separates the input audio into uniform speaker parts. In addition, allows for the extraction of discriminative speaker embeddings from those audio chunks, including speaker factors, i-vectors, x-vectors, convolutional neural network (CNN) and long short-term memory (LSTM) based embeddings, and d-vectors, and (c) speaker clustering, which establishes the number of speakers that make up an audio stream and assigns unique speaker labels to each segment (and possibly, identities).

Speaker diarization research has advanced due to numerous recent works on deep neural network-based embedding extraction, yielding noteworthy performance improvements. In addition, they have successfully replaced earlier i-vector-based diarization embedding techniques. Since they are more successful than conventional i-vectors, especially for short-duration speech, the commonly utilised x-vector embeddings have become the de-facto standard for speaker detection and diarization. However, most unsupervised algorithms have been used for speaker clustering over time. Some algorithms include the gaussian mixture model, AHC, mean shift, k-means spectral clustering, integrated linear programming, and links. These algorithms are based on similarity measures like the Bayesian information criterion, generalised log-likelihood ratio, and information bottleneck (IB). Several supervised speaker clustering techniques, including UIS-RNN and affinity propagation, have recently been introduced for diarization. However, despite the previous clustering techniques' effectiveness, speaker diarization remains a difficult challenge in many practical applications because of the audio's significant heterogeneity and fluctuation.

The recent effectiveness of generative adversarial networks (GANs) in capturing complex data distributions by storing rich latent structures has piqued interest. However, because of the problem of mode collapse, training GANs is difficult. Many GAN versions, such as the Wasserstein GANs (WGAN), multi-generator GANs, and mixture GANs, have been proposed to address this problem. The GANMM, a unique adversarial architecture containing a mixture of generators and discriminators and a classifier trained in an expectation maximisation (EM) approach, was recently introduced. This methodology has proven to be effective for picture and character data clustering.

Although deep learning approaches have substantially improved the performance of speech recognition and speaker verification systems, most existing clustering strategies for speaker diarization still need to take full advantage of them.

As a result, it's worth looking at the possibilities of neural network-based clustering for speaker diarization [8].

2.1 Bottom-Up

The bottom-up approach is by far the most common in the literature. Agglomerative hierarchical clustering (AHC or AGHC), another name for the bottom-up approach, involves training several clusters or models to merge them and eventually leaving just one set for each speaker. Many systems use a uniform initialization, which separates the audio stream into numerous equal-length abutted segments. While some initializations have looked at k-means clustering, there has been a study on various initializations. This more straightforward approach usually produces comparable results.

Every time, there are more parts of the audio stream than the maximum number of speakers anticipated. The number of sets is then decreased by one with each iteration of the bottom-up strategy by choosing closely related clusters to merge.

Gaussian mixture models (GMMs) are frequently used to model clusters, and when two sets of data are joined, one new GMM is trained using the previously provided data to the two independent groups. Standard distance measurements are used to determine which clusters are closest.

The process is typically repeated iteratively until some stopping criterion is reached, at which point there should only be one cluster for each detected speaker. After each cluster merger, a reassignment of frames to groups, such as by Viterbi realignment, is commonly carried out. The Bayesian information criterion (BIC), kullback-Leibler (KL)-based metrics, the generalised likelihood ratio (GLR), or the recently proposed Ts metre are examples of threshold techniques that could be used as halting criteria. Bottom-up solutions have consistently fared admirably in NIST RT evaluations [10].

2.2 Top-Down

The process is typically repeated iteratively until some stopping criterion is reached, at which point there should only be one cluster for each detected speaker. After each cluster merger, a reassignment of frames to groups, such as by Viterbi realignment, is commonly carried out. The Bayesian information criterion (BIC), kullback-Leibler (KL)-based metrics, the generalised likelihood ratio (GLR), or the recently proposed Ts metre are examples of threshold techniques that could be used as halting criteria. Bottom-up solutions have consistently fared admirably in NIST RT evaluations..

To extract helpful training data from the unlabeled segments, new speaker models are incrementally introduced to the model with interspersed Viterbi realignment and adaptation. One of these new models can be given credit for labelled components. The process can be stopped using stopping criteria akin to those found in bottom-up systems, or it can go on indefinitely as long as there are still relevant unlabelled segments available for training new speaker models.

Top-down approaches are far less common than bottom-up approaches. Several examples include. The best bottom-up systems have consistently outperformed top-down methods, but they always have and honourably surpassed the larger field of other bottom-up entries. In addition to being exceptionally computationally effective, top-down techniques can benefit from cluster purification [10].

2.3 Generative Adversarial Network (GANs)

GANs are a type of unsupervised learning neural network. In 2014, Ian J. Goodfellow developed and introduced it. GANs are made up of two competing neural network models that can analyse, capture, and duplicate variations in a dataset. GANs are divided into three categories that is Generative Adversarial Networks [15].

Generative: learning a generative model, which specifies how data is generated in terms of a probabilistic model, is referred to as generative learning. Adversarial: a model is trained in an adversarial environment. Networks: deep neural networks are artificial intelligence (AI) systems that can be used for training [4].

2.4 Fully End-to-End Neural Diarization

Given a multi-talker recording, a bidirectional extended short-term memory network, presented as the EEND, directly outputs speaker diarization findings. Most speaker diarization methods employ speaker embedding clustering. For instance, i-vectors, d-vectors, and x-vectors are often used in speaker diarization tasks. These embeddings of brief segments are then divided into speaker groups using clustering techniques, including gaussian mixture models, agglomerative hierarchical clustering, mean shift clustering, k-means clustering, links, and spectral clustering. The efficacy of these clustering-based diarization algorithms has been demonstrated.

A self-attention-based neural network directly outputs the aggregated speech activity of all speakers for each time frame given an input of a multi-speaker audio recording. Our approach naturally handles speaker overlaps during training and inference time by utilising a multi-label classification architecture. Additionally, the neural network is trained end-to-end using a recently proposed permutation-free objective function, resulting in a low number of diarization errors [16].

A neural network-based source separation approach was recently presented to cope with speaker-overlapping speech. In one iteration, the model separates one speaker's time-frequency mask, and in another iteration, it separates another speaker's mask.

Speaker diarization can be achieved even in overlapping speech using the source separation technique. However, their source separation training goal does not always imply minimising diarization errors. It is preferable to utilise a

diarization error-oriented objective function when addressing the speaker diarization problem. Furthermore, as their model requires clear, non-overlapping reference speech, their method must be able to train on actual multi-speaker recordings.

For optimization based on diarization errors, a fully supervised diarization strategy has been provided. The speaker diarization problem is resolved by this method using a factored probabilistic model with modules for speaker change, assignment, and feature development. However, the speaker-change model used in their method presupposes one speaker for each segment, making it difficult to apply the concept to speaker-overlapping speech.

The EEND has a number of advantages over traditional approaches [9].

1. By simply supplying it as input during training and inference, the EEND can handle overlapping speech explicitly.
2. The detection of speech activity, speaker identification, source separation, and clustering does not require separate modules when using the EEND.
3. The EEND does not require clear, non-overlapping speech to train synthetic conversational blends, in contrast to the source separation method.
4. This allows domain adaptation to be used with genuine overlapping speech dialogues.

2.5 Speech Segmentation

Speaker segmentation divides an input audio stream into acoustically homogeneous chunks depending on the speaker's identity. Using acoustic characteristics, a typical speaker segmentation system detects probable speaker transition points.

Speech segmentation is performed by our deployed system using Automated Speech Recognition (ASR). There are two advantages of employing ASR [10].

1. ASR can precisely remove the non-speech portion.
2. The ASR output is perfectly aligned with the speaker diarization output

The speaker semester's objective is to analyse continuous audio segments and identify potential speaker turn points. The technique is as follows: at each word boundary within the contiguous segment, the segment is split into two sub-segments (left to the boundary, right to the boundary), using the conventional cepstral acoustic characteristics and the ASR output as inputs. The features from the relevant sub-segments are fitted to two Gaussians, which are then compared [12].

2.6 Audio Embedding Extraction

The i-vector architecture has become the dominant model for speaker verification over the last decade. The i-vector framework can predict speaker characteristics and adjust for channel fluctuation using factor analysis and backend classifiers.

The main idea behind the framework is to express a variable-length utterance using a fixed-length low-dimensional vector. Speaker diarization and speech synthesis are two other uses for these vectors. Additionally, there has been a recent rise in interest in using neural networks to extract speaker-discriminant vectors, also known as speaker embeddings. The d-vector is an early success of speaker embedding; it was primarily developed for text-dependent speaker verification but has since been discovered to perform well in text-independent applications. To better capture speaker characteristics, RNNs, CNNs, and other neural network architectures are used.

Currently, speaker embedding extraction just takes into account speaker labels and ignores all other data. Contrarily, speech signals are intricate and subject to a range of influences. Therefore, to reflect what is said and who is speaking, the speech's two main components, phonetic content and speaker traits, are combined. Background noise and channel effects are also included. Speech and speaker recognition are both hampered as a result of this. Speaker adaption techniques are used in ASR to reduce the effects of different speakers and to improve accuracy. In speaker recognition, phonetic independence is also desirable. In statistical models like joint factor analysis (JFA) and i-vector, deep neural networks (DNNs) can use phonetic vectors as indicators to locate more speaker-dependent information. In contrast, phonetic vectors can be employed in neural models as markers to guide DNNs in finding more speaker-dependent data.

But even though the phonetic and speaker characteristics are different, they have some things in common. For identifying voices and speakers, for instance, the spectral energy distribution and pitch trajectory are helpful. Based on this discovery, multi-task learning has been promoted in both domains. In multi-task learning, the speaker and phonetic discriminant networks share specific hidden layers and simultaneously predict speaker and phonetic labels. However, there are some flaws in earlier techniques based on phonetic vectors and multi-task learning: 1. An independently trained ASR model is used to extract the phonetic vector, which is then used to train a speaker discriminant network. 2. The limited frame-by-frame operation of existing multi-task learning techniques indicates that speaker and phonetic classification are only used at the frame level. But training the model is difficult since speaker characteristics are typically chaotic in short time frames [11].

2.7 Clustering

Cluttering is a speech and communication disorder marked by a quick velocity of speech, unpredictable rhythm, and poor syntax or grammar, all of which make speech difficult to understand. Cluttering is a fluency condition in which the speaker's rate is considered to be unnaturally fast, irregular, or both (although measured syllable rates may not exceed normal limits). These rate irregularities express itself in one or more of the following symptoms: (a) an abnormally high number of disfluencies, the vast majority of which are not characteristic of stutterers;(b) excessive (generally inappropriate) degrees of coarticulation among

sounds, especially in multisyllabic words; and (c) frequent placement of pauses and usage of prosodic patterns that do not correspond to syntactic and semantic constraints [1].

There are two types of data clustering algorithms: hierarchical and partitional. Partitional algorithms identify all clusters at once, whereas hierarchical algorithms find consecutive clusters utilising previously established clusters. Hierarchical algorithms can be agglomerative (from the bottom up) or divisive (from the top down) (top-down). Agglomerative algorithms start with each element as a single cluster and merge it into larger clusters as time goes on. Divisive algorithms start with the entire set and divide it into smaller and smaller groups [17].

2.8 Re-Segmentation

Despite the limits imposed by the telephone channel, such as bandwidth, transducer, noise, and echo, this is a difficult task. Fast speaking speeds, poor coarticulation at word boundaries, a diverse variety of dialects, speaking styles, and accents, and a wide range of word pronunciations all pose distinct challenges for detection of spontaneous speech. Furthermore, dysfluencies such as ungrammatical pauses, stutters, laughs, repeats, and self-repairs abound in these talks. The vocabulary is extensive, with monosyllabic terms predominating, making it difficult to distinguish them. As a result, acoustic modelling for recognition is weak, and there is a lot of mismatches between training and test data. The vocabulary is extensive, with monosyllabic terms predominating, making it difficult to distinguish them. As a result, acoustic modelling for recognition is weak, and there is a lot of mismatches between training and test data.

Predicting a large number of typical variant pronunciations and using them as additional routes into the acoustic models is one way to lessen this acoustic-level mismatch. Such pronunciation modelling techniques have had limited success due to challenges with intelligent integration of language model and acoustic model scores. Instead, the acoustic models can be recalculated to account for such differences in pronunciation. This necessitates high-quality transcriptions and database segmentation [18].

2.9 Mel Frequency Cepstrum Coefficient (MFCC)

The Mel-Frequency Cepstrum (MFC) is a representation of a sound wave's short-period power spectrum, and the MFCC (Mel frequency cepstrum coefficient) is a collection of MFC coefficients based on human auditory characteristics. According to psychological studies, a human can only recognise sounds below 1000 Hz, implying that the human ear's essential bandwidth is restricted to 1000 Hz. Because it is linearly spaced below 1000 Hz and logarithmically above 1 kHz, MFCC offers a fairly comparable response [20].

2.10 Body Linear Predictive Codes (LPC)

Signal compression is desirable for efficient transmission and storage. For optimum channel use on wireless media, digital signals are compressed before transmission. LPC is the most generally used medium or low bit rate coder. The signal's power spectrum is calculated by the LPC. It's a tool for analysing each of several prominent bands of frequency that determine the phonetic quality of a vowel. LPC is a prominent formant estimation approach that is one of the most powerful speech analysis techniques [21].

3 Databases

2000 NIST Database Introduction. The LDC and NIST developed the 2000 NIST Speaker Recognition Evaluation. Where LDC means The Linguistic Data Consortium and the NIST means National Institute of Standards and Technology. It has English conversational telephone approximately 150 h of speech collected by LDC. It is used as training and testing data in the NIST-sponsored 2000 Speaker Recognition Evaluation . Data This publication consists of 10,328 single-channel SPHERE files en- coded in 8-bit mu-law containing approximately 4.31 GB of data, totalling 148.9 h [3].

CALLHOME American English Speech Introduction. The LDC created American English Speech. It has 120 spontaneous phone calls to native English speakers. The length of all calls is 30 min. 30 calls are recorded from North America. 90 calls from different countries outside of North America. The majority of participants dialled relatives or close friends. The database has documentation describing the format and content.

VoxConverse Speaker Diarisation Dataset Introduction. It is an audio-visual diary dataset made up of more than 50 h' worth of multi-speaker human voice samples that were taken from YouTube videos.

4 Conclusion

In the research work, long and careful consideration is given to the technique used in speech recognition for various tasks, such as the extraction of spectral features, which is measured using MFCC and matching the feature using VQ. Furthermore, the speaker identification technique that was investigated for speaker diarization, i.e. the most commonly used techniques, is LSTM with a DER of 5%, segmentation, speaker extraction, and clustering. And recently researchers begun working on GANs.

References

1. Wang, Quan, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. "Speaker diarization with LSTM." In 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 5239-5243. IEEE, 2018.
2. Aishwary Joshi, "Speech Diarization" "https://www.iitg.ac.in/cse/robotics/?page_id=2442", [Online; accessed July-2021], 2021.
3. 2000 NIST Speaker Recognition Evaluation - Linguistic Data Consortium."https://catalog.ldc.upenn.edu/LDC2001S97
4. Geeks for Geeks Org, "Generative Adversarial Network (GAN)", https://www.geeksforgeeks.org/generative-adversarial-network-gan/%7D,[/,[Online;accessed July-2021],2021.
5. Bullock, Latan, Hervé Bredin, and Leibny Paola Garcia-Perera. "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), pp. 7114-7118. IEEE, 2020.
6. Fujita, Yusuke, Shinji Watanabe, Shota Horiguchi, Yawen Xue, Jing Shi, and Kenji Nagamatsu. "Neural speaker diarization with speaker-wise chain rule." arXiv preprint arXiv:2006.01796 (2020).
7. Zhang, Aonan, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. "Fully supervised speaker diarization." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6301-6305. IEEE, 2019.
8. Pal, Monisankha, Manoj Kumar, Raghuveer Peri, and Shrikanth Narayanan. "A study of semi-supervised speaker diarization system using gan mixture model." arXiv preprint arXiv:1910.11416 (2019).
9. Fujita, Yusuke, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. "End-to-end neural speaker diarization with permutation-free objectives." arXiv preprint arXiv:1909.05952 (2019).
10. Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. "Speaker diarization: A review of recent research." IEEE Transactions on audio, speech, and language processing 20, no. 2 (2012): 356-370.
11. Liu, Yi, Liang He, Jia Liu, and Michael T. Johnson. "Speaker embedding extraction with phonetic information." arXiv preprint arXiv:1804.04862 (2018).
12. Dimitriadis, Dimitrios, and Petr Fousek. "Developing On-Line Speaker Diarization System." In INTERSPEECH, pp. 2739-2743. 2017.
13. Tranter, Sue E., and Douglas A. Reynolds. "An overview of automatic speaker diarization systems." IEEE Transactions on audio, speech, and language processing 14, no. 5 (2006): 1557-1565.
14. Huang, Zili, Shinji Watanabe, Yusuke Fujita, Paola García, Yiwen Shao, Daniel Povey, and Sanjeev Khudanpur. "Speaker diarization with region proposal network." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6514-6518. IEEE, 2020.
15. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." Communications of the ACM 63, no. 11 (2020): 139-144.
16. Fujita, Yusuke, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. "End-to-end neural speaker diarization with self-attention." In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 296-303. IEEE, 2019.

17. Madhulatha, T. Soni. "An overview on clustering methods." arXiv preprint arXiv:1205.1117 (2012).
18. Deshmukh, Neeraj, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. "Resegmentation of SWITCHBOARD." In ICSLP. 1998.
19. Kwon, Soonil, and Shrikanth Narayanan. "A method for on-line speaker indexing using generic reference models." In Eighth European Conference on Speech Communication and Technology. 2003.
20. Bharti, Roma, and Priyanka Bansal. "Real time speaker recognition system using MFCC and vector quantization technique." International Journal of Computer Applications 117, no. 1 (2015).
21. Dave, Namrata. "Feature extraction methods LPC, PLP and MFCC in speech recognition." International journal for advance research in engineering and technology 1, no. 6 (2013): 1-4.