SPEAKER DIARIZATION WITH DEEP LEARNING TECHNIQUES

Kshirod Sarmah[1], Swapnanil Gogoi[2]

[1]Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College) Amjonga, Goalpara, Assam, Pin 783124, INDIA, kshirodsarmah@gmail.com
[2]Gauhati University IDOL, Guwahati, Assam Pin 781014, INDIA, swapnanil@gauhati.ac.in

ABSTRACT

Speaker diarization is a task to identify the speaker when different speakers spoke in an audio or video recording environment. Artificial intelligence (AI) fields have effectively used Deep Learning (DL) to solve a variety of real-world application challenges. With effective applications in a wide range of subdomains, such as natural language processing, image processing, computer vision, speech and speaker recognition, and emotion recognition, cyber security, and many others, DL, a very innovative field of Machine Learning (ML), that is quickly emerging as the most potent machine learning technique. DL techniques have outperformed conventional approaches recently in speaker diarization as well as speaker recognition. The technique of assigning classes to speech recordings that correspond to the speaker's identity is known as speaker diarization, and it allows one to determine who spoke when. A crucial step in speech processing is speaker diarization, which divides an audio recording into different speaker areas. In-depth analysis of speaker diarization utilizing a variety of deep learning algorithms that are presented in this research paper. NIST-2000 CALLHOME and our in-house database ALSD-DB are the two voice corpora we used for this study's tests. TDNN-based embeddings with x-vectors, LSTM-based embeddings with d-vectors, and lastly embeddings fusion of both x-vector and d-vector are used in the tests for the basic system. For the NIST-2000 CALLHOME database, LSTM based embeddings with d-vector and embeddings integrating both x-vector and d-vector exhibit improved performance with DER of 8.25% and 7.65%, respectively, and of 10.45% and 9.65% for the local ALSD-DB database.

Keywords: Speaker Diarization; Machine Learning; MFCC; Deep Learning

1. INTRODUCTION

Speaker Diarization (SD) is the process of determining "who spoke when" during a multi-party conversation. Without any prior information of the speakers or the number of speakers in the dialogue, SD seeks to identify all the utterances made by each speaker. Speech analytics and transcription systems heavily rely on the process of segmenting and categorizing an audio recording into different speaker segments. Applications like audio indexing, transcription, and sentiment analysis all depend on SD. In SD, the sub-tasks of speaker segmentation and speaker clustering may be seen. While speaker clustering refers to gathering all speaker turns that are associated with one speaker, speaker segmentation looks for the speaker in relation to limits.

A few years ago, speaker adaptive processing was made possible by the development of SD algorithms for voice identification on multi-speaker audio recordings. A variety of audio data broadcast from different media stations, conference conversations, private videos from online social media, business meetings, court sessions, etc. may all be efficiently indexed or analyzed using SD. In essence, an SD system is made up of numerous separate sub-modules. Several front-end processing approaches, such as voice augmentation, speech separation, or target speaker extraction are utilized to minimize any artifacts in auditory settings. The selected speech segment's raw speech signals are converted to their comparable acoustic characteristics or embedding vectors. The altered speech segments are classified and speaker courses' label during the clustering phase, and the clustering findings are further refined during the post-processing phase. The most cutting-edge improvements in SD have been accomplished right now. On the other hand human perception is almost entirely capable of performing speaker diarization without knowing the fundamentals of language, and the process can be completed on the spot without specialized knowledge of the languages spoken by various speakers. In this study, we investigate new improvements made feasible by cutting-edge deep learning techniques as well as the evolution of speaker diarization across time.

Deep learning (DL), one of the most popular branches of machine learning (ML), has achieved outstanding success in practically all application domains. We are aware that Deep Learning (DL) uses a multi-layer structure, commonly referred to as Deep Neural Networks (DNN), to automatically learn features from enormous amounts of raw data. By utilizing layer wise processing techniques, task-related information may be strengthened and kept, and task-irrelevant variations can be reduced and eliminated. Since there are numerous layers between the input and output layers in deep learning (DL), many non-linear information processing phases can be used for feature learning and pattern categorization [1][2]. Low-level conceptions can define high-level concepts, and vice versa, according to current literature views on DL-based representation learning. Deep learning techniques have gained prominence in recent research due to their ability to automatically learn complex speaker features. In deep speaker embeddings approach, deep learning models are used to learn powerful speaker embeddings, such as x-vectors or d-vectors, which capture unique speaker characteristics. These embeddings enable better representation and separation of speakers in diarization tasks. In several study evaluations, DL has been characterized as a general learning method that can tackle almost all real-world problems in a number of application domains. In light of this, we can say that DL is not task-specific [3].

2. STATE-OF-ART RESEARCH IN SPEAKER DIARIZATION

Significant progress has been achieved in the development of SD technology starting in the 1990s and early 2000s. An important accomplishment in speaker recognition has been achieved by SD systems using the i-vector, a speaker-specific representation in a total variability space derived from a condensed JFA, as a feature representation for short audio segments that were segmented unsupervisedly [4]. In speaker diarization, Speaker embeddings based on neural networks, sometimes referred to as d-vectors, have recently gained popularity and outperform earlier state-of-the-art methods based on i-vectors. The emphasis was mainly on using SD to address the issues associated with speech [5]. As a result of the rapid improvements in deep learning techniques that addressed numerous technical problems across several machine learning areas, SD systems have undergone many noteworthy changes. To improve clustering performance in SD, the mel-frequency cepstral coefficient (MFCC) has been successfully substituted by the i-vector or speaker factors [6]. Mean shift, probabilistic linear discriminant analysis (PLDA), variational Bayesian Gaussian mixture model (VB-GMM), and PCA were all combined with it [7][8][9][10]. The Agglomerative Hierarchical Clustering (AHC) technique is the approach that SD practitioners employ the most frequently. In this method, a voice Different methods can be used to construct this first segmentation directly dividing the speech stream into homogenous segments, or [11][12][13][14][15]. AHC involves arranging these segments interchangeably up until and unless each segment is paired with its appropriate speaker. Additionally, to improve the speaker turn boundaries, two clusters are connected and a new segmentation is built Bayesian Information Criterion (BIC) Tranter and Reynolds' [11][12] technique is one of the conventional ways to determine which cluster pair should be merged. On the other side, Viterbi Decoding was one of the most widely used algorithms for speaker re-segmentation.

Since the emergence of DL in the 2010s, There has been a substantial amount of study on how to best utilize the deep neural networks (DNN) potent modeling capabilities for speaker diarization. For speaker diarization, lots of DL techniques have currently been successfully used for better performances [16][17][18][19]. DL uses several techniques for both clustering and segmentation problems. Speaker turn boundaries have been effectively identified using audio signals using recurrent neural networks (RNNs), often referred to as Long Short-Term Memory Networks (LSTMs). LSTMs are designed to capture long-range relationships in sequential data [20][21]. After combining linguistic and auditory characteristics, it shows a remarkable result in SD system [22]. The performance of smoothly detecting speaker turn boundaries was also improved by [23][24]. Above all, the adoption of the same approach for clustering tasks has been motivated by the success of speaker embeddings for speaker verification [25][26][27]. Another excellent example of how neural networks can produce superior outcomes is the speaker embeddings' extraction [18][28][29]. Similar to the x-vector developed by Snyder [30], which, depending on the output of a DNN trained for a voice recognition task, is one of the most widely utilized embedding vector representations and receiving a pertinent milestone [20][21].The performance and ease of training with large amounts of data as well as robustness against speaker variability and acoustic environments were significantly improved by the switching issues from i-vector to these neural embeddings [31][32]. End-to-end neural diarization (EEND), which replaces individual sub-modules in the conventional SD system with a single neural network, has received increasing attention recently and has produced positive results [33][34]. Another cutting-edge research project combined acoustic data with linguistic contents and produced speaker diarization systems that performed better when using LSTM neural networks for word level DER .Transfer learning techniques, such as using pre-trained models from related tasks like automatic speech recognition (ASR) or speaker verification, have been studied in an effort to enhance speaker diarization models' performance, particularly when labeled diarization data is

limited. More recently, there is ongoing research into developing and rapid advancement of speaker diarization models that can learn from weakly labeled or semi-supervised data, where only partial annotations are available. In multimodal approaches combining audio and visual data enhances diarization accuracy. Deep learning models simultaneously process both modalities for improved speaker separation and tracking.

This is particularly useful when obtaining fully labeled data is costly or time-consuming. At present researchers are focusing on making speaker diarization systems more robust to real-world challenges, such as noisy environments, overlapping speech, and varying recording conditions. These improvements are crucial for practical applications like meeting transcription, voice assistants, and surveillance systems.

3. CLASSIFICATION OF SPEAKER DIARIZATION

Attempting to classify the most different speaker diarization methods currently in use, including in the realm of earlier deep learning-era modularized speaker diarization systems and more modern neural network-based systems. There are a total of four categories based on the main category we mentioned, which is based on two criteria. An objective function focused on speaker discrimination was used to train the model, or not, is the first criterion. The term "diarization objective" refers to any trainable methods for learning speaker relationships and optimizing models in multispeaker environments. If many modules are jointly optimized toward an objective function, that is the second condition. "Single-module optimization" refers to a technique wherein only one sub-module is transformed into a trainable one. Conversely, the "Joint optimization" class comprises joint segmentation and clustering modeling by Zhang [31], a fully end-to-end neural diarization system [34], and joint speaker diarization and speech separation modeling by Von Neumann [35].

4. METRICS OF SPEAKER DIARIZATION

Diarization Error Rate (DER) is a method for evaluating the accuracy of speaker diarization systems [36]. The combination of these three separate error types is known as DER, which stands for false alarm of speech (FA), missing speech detection (MSD), and confusion between speaker labels. So, the following is a definition of DER:

$$\mathrm{DER} = \frac{(\mathrm{FA}+\mathrm{Missed}+\mathrm{Speaker\_Confusion})}{\mathrm{Total\ Duration\ of\ Time}} \quad (1)$$

The speaker diarization studies have been the ones that have employed this evaluation scheme the most.

JER, or Jaccard Error Rate, was initially applied in the DIHARD II assessment, is another widely used metric for assessing speaker diarization. JER aims to give each speaker the same weight when evaluating them. JER is calculated by first computing per-speaker error rates, as opposed to DER, which is estimated for the entire utterance. JER is specifically calculated using the following formula:

$$\mathrm{JER} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}}^{\mathrm{N_{ref}}} \frac{\mathrm{FA_i}+\mathrm{Miss_i}}{\mathrm{Total_i}} \quad (2)$$

The i^{th} speaker's speaking time in the reference transcript and the i^{th} speaker's speaking time in the hypotheses are combined in Eq. (2). is how many speakers there are in the reference script. Keep in mind that the JER computation partially accounts for the Speaker-Confusion in DER.

The third diarization statistic is the Word-level DER (WDER) metric. As is well known, the DER is based on how long each speaker speaks. WDER, on the other hand, is designed to quantify lexical side of transcription error. Since DER depends on speaking time that isn't always in line with word boundaries, there is a disparity between that output and the correctness of the final transcript that drives WDER. Word-breakage shares a similar principle with WDER suggested the idea of word-breakage ratio [37]. Word-breakage ratio, as opposed to WDER, counts the instances of speaker change within a word boundary. The term "WDER"—which stands for "evaluating the diarization output with ground-truth transcription"—was proposed by Park and Georgiou's research [38]. Shafey recently looked at the WDER format's combined ASR and speaker diarization mechanism [39]. In this study, we exclusively utilize DER as a metric to assess the SD system.

5. TRADITIONAL SPEAKER DIARIZATION

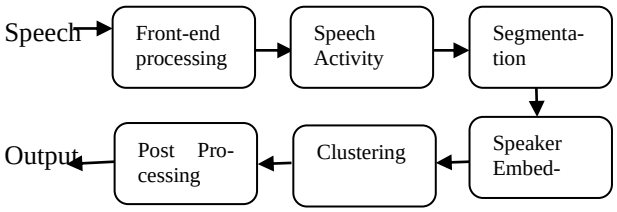

Figure.1. Common speaker diarization system

Conventional speaker diarization systems, as shown in Figure 1, are composed of multiple separate sub modules. Different front-end processing methods are employed to lessen any artifacts in auditory environments, including speech separation, target speaker extraction, speech augmentation, and dereverberation. Next, speech from non-speech events is distinguished using voice or speech activity detection (SAD). The selected speech segment's raw speech signals are converted into embedding vectors, or acoustic characteristics. During the clustering stage, speaker classes recognize and label the modified speech fragments and the clustering findings are further refined during the post-processing stage. Each of these sub-modules is typically tuned independently. This section mostly discusses front-end processing methods used in the speaker diarization pipeline for voice augmentation, dereverberation, speech separation, and speech extraction.

suppressing the loudness element of chattering speech, which has considerably improved as a result of deep learning, is the main objective of speech improvement approaches. Multichannel processing can boost the efficacy of voice enhancement techniques, including minimal variance distortionless response (MVDR) beam formation [40].

The primary dereverberation technologies are based on statistical signal processing techniques, which sets them apart from other front-end solutions. The weighted prediction error (WPE) based dereverberation is one of the methods that is most frequently employed [41][42][43]. Since SAD is a preprocessing step that can cause mistakes to spread across the entire pipeline, In addition to speaker diarization, speaker recognition and voice recognition systems also depend on it. Two main components make up the majority of a SAD system. Elevated level statistics inside the residual domain of linear predictive coding, or MFCCs, are typically utilized in the first, along with acoustic factors like zero crossing rate and signal energy. The effectiveness of SAD significantly influences the speaker diarization system's overall effectiveness due to the possibility of a large number of missed speech segments or false positive salient events [47].

When speech areas overlap significantly, speech division represents a potential family of techniques. Several studies have proven the efficacy of beam formation-based multichannel speech separation [44][45][46]. Speech segmentation, as used in speaker diarization, is the division of the input audio stream into multiple speaker-uniform segments. Thus, a segmentation procedure is used to select the speaker diarization system's output unit. The two primary categories into which speech segmentation approaches for speaker diarization fall are segmentation via speaker-change point detection and uniform segmentation. The speaker-change point detection method of voice segmentation frequently results in uneven segment durations. Consequently, Following the emergence of the i-vector [4] and DNN-based embeddings [28][30], speaker-change point detection-based segmentation was mostly superseded by uniform segmentation. For speaker diarization systems to determine how similar two speech segments are, speaker representation is essential. Due to their close relationship, this section will also discuss the similarity measure in addition to such speaker representation.

6. NEURAL NETWORK BASED SPEAKER REPRESENTATION

The popularity of deep learning techniques has had a significant impact on speaker representations for speaker diarization. For face recognition tasks, DNN-based representation learning was initially proposed [50]. The DNNs' multiple layers' ability to do nonlinear modeling serves as the foundation for the process of adapting the speaker embedding to the input signal. Unlike classic factor analysis models based on decomposable components, the DNN training technique enables the neural networks to learn the mapping without giving any components or factors. Additionally, existing probabilistic models (such GMM-UBM) do not adequately describe the input acoustic information for DNN-based speaker representation learning. Due to the fact that factor-analysis based methods use a solution that requires a compu-

tationally intensive matrix inversion operation [49]. DNN-based speaker representation consequently performs more effectively during the inference stage.

To represent both short-term and long-term features, the benefits of recurrent networks and deep feedforward networks are combined in DNN-RNN hybrids. As a result, Compared to the conventional factor-analysis based approaches, the inference speed has increased and the representation learning process has become more straightforward. Among the several neural-network based speaker representations, d-vector represents one of the most prominent frameworks for extracting speaker representations [28]. The d-vector technique was employed by Wang and Zhang in their speaker diarization research [31][35]. Using an x-vector improves the performance of DNN-based speaker representations [30]. The statistics pooling layer forwards the findings to the layer above after aggregating the frame-level outputs from the layer above and calculating their mean and standard deviation. It might therefore make it possible to extract x-vectors from an input with varying length. This is useful for speaker verification as well as speaker diarization, since speaker diarization systems must handle segments that, when reduced at the end of each utterance, are less than the predetermined uniform segment length.

The speech segments are grouped using a clustering technique based on the speaker representation and similarity metric. The most widely employed speaker diarization clustering methods are shown here. AHC, KL [52], and PLDA [53][54] are three examples of the clustering technique that has been regularly adopted by several speaker diarization systems. AHC is a procedure that repeatedly joins already-existing clusters until a prerequisite is satisfied by the clustering process. The AHC procedure is started by calculating the similarity between N singleton clusters. The most comparable pair of clusters from each phase is combined. Resegmentation is the post-processing procedure that involves fine-tuning the speaker boundary, which was approximately computed using the clustering approach. Kenny introduced the Baum-Welch algorithm-based Viterbi resegmentation technique [55].

This technique alternately applies the estimation of the speaker-specific Gaussian mixture model and Segmentation based on Viterbi algorithms and the estimated speaker GMM. Diez claims that variational Bayesian hidden Markov model (VB-HMM)-based variational resegmentation was found to be superior to Viterbi resegmentation [56].

Many studies have been conducted on the fusion strategy of different diarization findings in an effort to improve the accuracy of diarization as another post-processing path. Although it is observed that combining different systems results in superior results, combining different diarization hypotheses provides a number of particular challenges.

First of all, speaker labeling varies amongst different diarization systems. Second, the predicted speaker count may vary between diarization systems. Additionally, different diarization methods may have different predicted time bounds. These issues must be addressed by speaker diarization system combining procedures during the fusing of several theories. A strategy for choosing the optimal diarization outcome from a variety of diarization systems . This method views each diary system's complete series of diary entries for a recording as a single object that needs to be clustered. The outcomes of the diarization and the symmetric DER from the two clusters is used to compute the separation between the two clusters after the set of diarization results is exposed to AHC. The process of merging AHC iteratively is carried out till two clusters are formed.

7. DEEP LEARNING BASED SPEAKER DIARIZATION

Several speaker diarization methodologies have been proposed to enhance the deep learning-based clustering process. Xie suggested a novel approach called deep embedded clustering (DEC) with the aim of enhancing clustering [67]. DEC's ultimate goal was to alter the speaker embeddings, or input attributes, so that they could be more easily distinguished among speakers or clusters. In order to achieve cluster differentiation, every embedding is provided with a chance to belong to each of the potential speaker clusters.

The original DEC technique had various problems, thus Dimitriadis presented a further proposal for an enhanced and improved DEC (IDEC) technique with increased accuracy for speaker diarization [66]. Initially, there was a possibility that the neural network may compress to a primitive solution leading to skewed embeddings. To mitigate this issue, Guo suggested putting a reconstruction loss between the input feature and the autoencoder's output in order to explicitly retain the local structure of the data [68]. The problem was further addressed by including the loss function to guarantee that the distribution of speaker turns is uniform among all speakers or that each speaker contributes equally to the session [66]. Although this isn't usually the case for real recording. Lastly, to make the algorithm's behavior closer to k-means, Dimitriadis developed an additional loss term that penalizes the distance from the centroid μ_i. [66].

The issue of embedding speech fragments is a major one. Regardless of what they are discussing, different speakers should ideally be embedded in various locations throughout the embedding space. Long short-term memory (LSTM)-depend d-vectors and time-delay neural networks (TDNN)-based x-vectors have both been utilized to successfully incorporate data in recent years.

For capturing the temporal patterns in speech diarization, LSTM networks are well suited. TDDN are particularly good in capturing temporal connections and patterns in the data. TDNNs is capable of extracting high-level characteristics from speaker voice segments in the context of speaker diarization. To produce better results, we combine the two embedding techniques. For clustering methods like spectral clustering, the capacity to rate speaker similarity between speech segments is essential. Scores of speaker similarity between groups are necessary for other clustering techniques, such as agglomerative hierarchical clustering.

The x-vector is one of the deep speaker embeddings that is most frequently utilized in speaker diarization, which statistically pools frame-level representations to embed segment-level speaker characteristics. End-to-End-Neural-Diarization-vector clustering (EEND-vector clustering) is a revolutionary speaker diarization methodology that combines daarization techniques based on neural networks and clustering into a unified framework. Since EEND-based methods have the advantage of jointly learning speaker separation and classification from the audio input, they have shown competitive performance in a number of speakers diarization benchmarks. It combines the best features of both frameworks: robust clustering-based algorithms for handling long recordings with multiple speakers, and effective EEND-based diarization and overlapped speech management [36]. EEND is a recently developed framework that uses a single neural network to carry out all speaker diarization methods [34]. Fujita first suggested EEND using a bidirectional long short-term memory (BLSTM) network. It was then quickly extended to the self-attention-based network by demonstrating the state-of-the-art DER for two-speaker data, including the dialogue audio in the Spontaneous Japanese corpus and the two-speaker excerpt from the CALLHOME dataset (LDC2001S97)[34]. EEND provides several benefits. To begin with, Overlapping utterance can be handled sensibly using EEND. Second, we should expect a high accuracy because the network is specifically designed to increase diarization accuracy. Third, by simply inputting a reference diarization label, it may be retrained using real data. However, EEND is not without its limits. First of all, the model's design places a limit on the most speakers it can support. Second, because EEND uses self-attention based neural networks, or BLSTM, online processing is challenging. Third, empirical evidence indicates that EEND tends to overfit the training data distribution [33].

8. DATA SET IN SPEAKER DIARIZATION SYSTEM

8.1 CALLHOME

SPE 2000 (Disk-8) and NIST SRE 2000 (LDC2001S97), dataset that is most frequently utilized in modern research for speaker diarization is called CALLHOME. There are four hundred multilingual telephone speaking sessions in this dataset. There are two to seven presenters per session, with two of them monopolizing the conversation.

8.2 ASSAMESE LANGUAGE SPEAKER DIARIZATION DATABASE (ALSD-DB)

The Assamese language of Northeast India's Assamese speaker diariazation database is described in this section. Assamese Language Speaker Diarization Database (ALSD-DB) is the name of the database. Assam, a state in northeastern India, is among the most linguistically varied and rich areas of the continent. The Tibeto-Burman language family includes the vast majority of the native tongues spoken in modern-day Assam. ALSD-DB is gathered to analyze the speaker diarization problem in a multilingual setting. Every speaker has two recordings, each lasting four to five minutes and available in Assamese and Hindi. Table -1 displays the two recording devices that were utilized to record voice data concurrently.

The speakers are recorded having a conversation. The server, air conditioner, and other equipment were all turned on during the speech data collection, which took place in a lab environment. The speech data was provided by ten men and ten women, ages 20 to 50, who were chosen as informants. Every informant participates in two recording sessions, separated by at least one week

Table 1. Recording Specifications with devices

Device Sl.No	Device Type	Sampling Rate	File Format
1	Headset microphone	16 kHz	.wav
2	Laptop microphone	16 kHz	.wav

9. EXPERIMENTS AND RESULTS

The pyannote.metrics library serves as the foundation for our diarization evaluation system [35]. Using the widely utilized experimental methodology with equally spaced overlapping short segments, Oracle voice activity detection, and 5-fold cross validation, we assessed the suggested SD systems using the publicly accessible Speech dataset NIST-2000 CALLHOME as well as our local database ALSD-DB. The Diarization Error Rate (DER) is used to represent the experiment's results, along with all of its features and diarization techniques. We use spectral clustering and K-means offline clustering techniques to evaluate every possible combination of both i-vector and d-vector models. A three-layer LSTM network with a final linear layer makes up the d-vector model. With a projection of 256 nodes, each LSTM layer has 768 nodes [21]. The ideal sliding window size and step for d-vector systems are 240 ms and 120 ms, respectively.

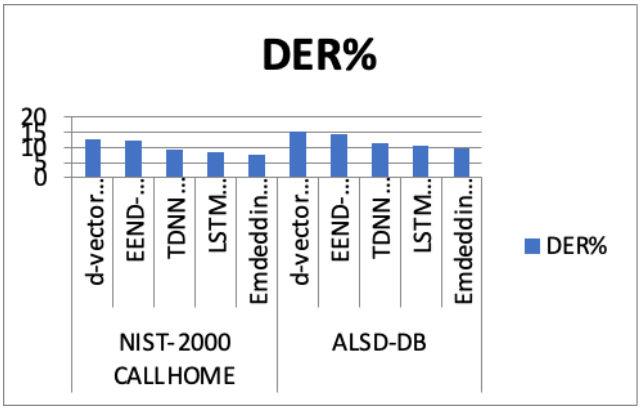

Figure 2. DER results for the experiments on NIST-2000 CALLHOME and ALSD-DB.

Table 2. DER results for the experiments on NIST-2000 CALLHOME and ALSD-DB

Dataset	Diarization Methodology	DER %
NIST-2000 CALL-HOME	d-vector + spectral	12.75
	EEND-vector clustering	12.25
	TDNN based embeddings (x-vector)	9.12
	LSTM based embeddings (d-vector)	8.25
	Emdeddings Combination (x-vector+d-vector)	7.65
ALSD-DB	d-vector + spectral	15.25
	EEND-vector clustering	14.50
	TDNN based embeddings (x-vector)	11.24
	LSTM based embeddings (d-vector)	10.45
	Emdeddings Combination (x-vector+d-vector)	9.65

10. THE MAJOR OBSTACLES AND FUTURE POTENTIAL OF SPEAKER DIARIZATION SYSTEM

This research has provided a thorough analysis of speaker diarization techniques, emphasizing the advancements in both state-of-the-art classical processes and deep learning-based diarization approaches. We point several possible avenues for future research, including the study of hybrid models that include many deep learning architectures and the use of domain adaptation and transfer learning methods to speaker diarization. Deep learning technology has led to advancements in speaker diarization that range from a fully EEND method to a system that substitutes a single module with a deep-learning-based approach. Furthermore, a tendency toward tightly integrating speaker diarization with ASR systems—such as using ASR output to increase speaker diarization accuracy—has emerged as speech recognition technology becomes more widely available. In an effort to enhance overall speaker diarization performance, integrated modeling for voice recognition and speaker diarization has been studied recently.

ONLINE SPEAKER DIARIZATION PROCESSING: Due to these outstanding findings, Speaker diarization systems are already in use for many different purposes, including audio indexing, meeting transcription, conversational AI systems, and conversational interaction analysis. However, there's still a lot of room for development. Lastly, we conclude this work by summarizing the outstanding issues for speaker diarization in terms of further investigation and advancement.

DOMAIN MISMATCH: To implement speaker diarization, most approaches presuppose that a whole recording can be watched. However, many applications, such meeting transcription systems or smart agents, require very low latency when selecting the speaker.. The subject of online speaker diarization has been tackled in a number of ways, including clustering-based systems [60] and neural network-based systems [31], but it is still challenging.

SPEAKER OVERLAP: When trained on data from a different domain, a model frequently exhibits poor performance. As an illustration, experimental research has shown that the EEND model overfits to the training data's speaker overlap distribution [33]. A domain mismatch issue will arise with any training-based approach. Evaluating trainable speaker diarization systems' capacity to handle a variety of inputs will become more and more important as interest in these systems grows. International speaker diarization evaluation initiatives, like the DIHARD challenge, are crucial in this regard [57][58][59]. Conversations inevitably overlap because of multiple speakers. For instance, speaker overlap was shown to be between 12% to 15% on average in meeting recordings [62][64], and it may even be higher in casual discussions [63][64]. Despite a lengthy history of research on the topic, controlling speaker overlaps to enhance speaker diarization is becoming more and more popular.

INTEGRATION WITH ASR: ASR results are required for various applications in addition to speaker diarization results. Certain systems find a SD system before ASR, whereas other systems [59][64] find a SD system after ASR in the modular coupling of speaker diarization with ASR. It is necessary to identify the optimal system design for the speaker diary and ASR tasks. For a given task, both kinds of systems performed admirably. ASR and speaker diarization have also been combined in another line of research [63][65]. The interdependency between speaker diarization and ASR could be taken advantage of by the joint modeling technique to more effectively complete both tasks. The question of whether such collaborative frameworks outperform tailored modular systems has not yet been extensively analyzed. Overall, one of the most hotly debated issues that many researchers are presently delving into is the combination of speaker diarization and ASR.

ROBUSTNESS TO NOISE: Background noise is a common feature of real-world audio data, and it can seriously impair speaker diarization systems' performance. It is crucial to create models that can withstand different kinds and intensities of noise.
Scalability: Due to their high computing demands, many diarization techniques are not as scalable for processing huge amounts of data in real-time or almost real-time applications. Efficiency and scalability are crucial factors to take into account.

LACK OF ANNOTATED DATA: Large volumes of labeled data are usually needed for training deep learning-based diarization models, and obtaining this data can be costly and time-consuming. It is imperative to develop efficient techniques for domain adaptation and poorly guided learning.

DOMAIN ADAPTATION: Speaker diarization systems often struggle to perform well in domains different from the ones they were trained on. Adapting models to new domains while maintaining high accuracy is a challenge.

EVALUATION METRICS: Developing comprehensive evaluation metrics that capture the complexities of diarization tasks, such as speaker overlap and speaker change, is an ongoing challenge. Existing metrics may not fully reflect real-world performance.

To address these challenges, ongoing research in speaker diarization focuses on developing more robust and efficient algorithms, creating larger and more diverse datasets, exploring weakly supervised and unsupervised learning approaches, and advancing the field of evaluation metrics. Researchers are also working on adapting diarization systems to emerging technologies and applications, such as voice assistants, virtual meetings, and automated transcription services.

11. CONCLUSIONS

This research shows, in summary, how several deep learning algorithms can enhance speaker diarization accuracy and robustness. The NIST-2000 CALLHOME and ALSD-DB speech corpora have been used to evaluate the SD system. TDNN-based embeddings with x-vectors, The tests for the basic system use LSTM-based embeddings with d-vectors, and finally embeddings fusion of both x- and d-vectors. For the NIST-2000 CALLHOME database, LSTM based embeddings with d-vector and embeddings integrating both x-vector and d-vector exhibit improved performance with DER of 8.25% and 7.65%, respectively, and of 10.45% and 9.65% for the local ALSD-DB database. The baseline system,

which combines x- and d-vectors, has been found to perform best across both datasets; however NIST-2000 CALL-HOME performs better than our local corpus ALSD-DB.

REFERENCES

[1] J. Schmidhuber, Deep Learning in Neural Networks: An Overview. Neural Network, 61, 2015 pp.85–117
[2]Y. LeCun,G. Hinton, Deep Learning. Nature, 521,2015, pp.436–444.
[3] Y. Bengio, Learning deep architectures for AI. Found. Trends Mach. Learn. 2, 2009, pp.1–127.
[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-End Factor Analysis for Speaker Verification. Vol. 19. No. 4. IEEE,2011.
[5]X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G.Friedland,O. Vinyals, Speaker Diarization: a Review of Recent Research. Vol. 2. No. 2. IEEE, 2012, pp.356–370.
[6] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, Stream-based speaker segmentation using speaker factors and eigenvoices. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2008, pp. 4133–4136.
[7] S.Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, J. Glass, Exploiting intra-conversation variability for speaker diarization. In: Proceedings of the Annual Conference of the International Speech Communication Association,2011.
[8] S.Shum, N. Dehak,R. Dehak,J.R. Glass, Unsupervised Methods for Speaker Diarization: an Integrated and Iterative Approach. Vol. 21. No. 10. IEEE, 2013,pp. 2015–2028.
[9] S.Shum, N. Dehak, J. Glass, On the use of spectral and iterative methods for speaker diarization. In: Proceedings of the Annual Conference of the International Speech Communication Association. 2012, pp. 482–485.
[10] M.Senoussaoui, P. Kenny, T. Stafylakis, P. Dumouchel, A study of the cosine distance-based mean shift for telephone speech diarization. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (1), 2013b,pp. 217–227.
[11] S.E.Tranter,D.A. Reynolds, Speaker diarisation for broadcast news. In: Odyssey. 2004, pp. 337–344.
[12] S.E.Tranter,D.A. Reynolds ,An Overview of Automatic Speaker Diarization Systems. Vol. 14. No. 5. 2006, pp. 1557–1565.
[13] V. Gupta, Speaker change point detection using deep neural nets. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE,2015, pp. 4420–4424.
[14] R.Yin, H. Bredin, C. Barras, Speaker change detection in broadcast TV using bidirectional long short-term memory networks. In: Proc. Interspeech, 2017.pp. 3827–3831.
[15] A.Jati, P. Georgiou, Speaker2Vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation. In: Proc. Interspeech.2017, pp. 3567–3571.
[16] Z. Zajıc, M. Hrúz, L. Müller, Speaker diarization using convolutional neural network for statistics accumulation refinement. In: Proc. Interspeech, 2017, pp. 3562–3566.
[17] G. Le Lan, D.Charlet, A. Larcher, S. Meignier, A triplet ranking-based neural network for speaker diarization and linking. In: Proc. Interspeech ,2017, pp. 3572–3576.
[18] D.Wang, J.Chen, Supervised Speech Separation Based on Deep Learning: an Overview. Vol. 26. No. 10. IEEE, 2018,pp. 1702–1726.
[19] G. Sun, C. Zhang, P.C.Woodland, Speaker diarisation using 2D self-attentive combination of embeddings. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, 2019,pp. 5801–5805.
[20] H. Bredin, Tristounet: triplet loss for speaker turn embedding. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, 2017, pp. 5430–5434.
[21] Q.Lin, R. Yin, M. Li, H. Bredin, C. Barras, LSTM based similarity measurement with spectral clustering for speaker diarization. In: Proc. Interspeech, 2019. pp. 366–370.
[22] M.À. India Massana, J.A. Rodríguez Fonollosa, F.J.Hernando Pericás, LSTM neural network-based speaker segmentation using acoustic and language modelling. In: Proc. Interspeech, 2017. pp. 2834–2838.
[23] T.J. Park, P. Georgiou, Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks. In:Proceedings of the Annual Conference of the International Speech Communication Association.2018, pp. 1373–1377.
[24] T.J. Park, K.J.Han, J. Huang, X. He, B. Zhou, P. Georgiou, S. Narayanan, Speaker diarization with lexical information. In: Proceedings of the Annual Conference of the International Speech Communication Association. 2019, pp. 391–395.

[25] D.Garcia-Romero, D.Snyder, G.Sell, D.Povey, A. McCree, Speaker diarization using deep neural network embeddings. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, 2017,pp. 4930–4934.
[26] M.Diez,L. Burget, S. Wang, J. Rohdin, J. Cernock`y, Bayesian HMM based x-vector clustering for speaker diarization. In: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 346–350.
[27] L.Wan, Q. Wang, A. Papir, I.L. Moreno, Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, 2018, pp. 4879–4883.
[28] E.Variani, X. Lei, E.McDermott, I.L. Moreno, J. G-Dominguez, Deep neural networks for small footprint text-dependent speaker verification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2014,pp. 4052–4056.
[29] G.Heigold, I.Moreno, S. Bengio, N. Shazeer, End-to-end text-dependent speaker verification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.2016, pp. 5115–5119.
[30] D.Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.
[31] A. Zhang, Q.Wang, Z.Zhu, J. Paisley, C.Wang, Fully supervised speaker diarization. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2019, pp. 6301–6305.
[32] R.Yin, H. Bredin, C. Barras, Neural speech turn segmentation and affinity propagation for speaker diarization. In: Proc. Interspeech, 2018. pp. 1393–1397.
[33] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with permutation-free objectives. In: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 4300–4304.
[34] Y.Fujita,N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2019b, pp. 296–303.
[35] T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, All-neural online source separation, counting, and diarization for meeting analysis, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 91–95.
[36] J.G. Fiscus, J. Ajot, M. Michel, J.S. Garofolo, The rich transcription 2006 spring meeting recognition evaluation, in: Proceedings of International Workshop on Machine Learning and Multimodal Interaction,NIST, 2006, pp. 309–322.
[37] J. Silovsky,J. Zdansky, J. Nouza, P. Cerva,J. Prazak, Incorporation of the asr output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams, in: International Workshop on Multimedia Signal Processing, IEEE, 2012,pp. 118–123.
[38] T.J.Park, P. Georgiou ,Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks, in: Proceedings of the Annual Conference of the InternationalSpeech Communication Association, 2018, pp. 1373–1377.
[39] L.E. Shafey, H. Soltau, I. Shafran, Joint Speech Recognition and Speaker Diarization via Sequence Transduction, in: Proceedings of the Annual Conference of the International Speech Communication Association,ISCA,2019, pp. 396–400.
[40] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Ho_meister, M. L. Seltzer, H. Zen, M. Souden, Speech processing fordigital home assistants: Combining signal processing with deep-learning techniques, IEEE Signal Processing Magazine 36,2019, pp. 111–124.
[41] T. Nakatani,T.Yoshioka,K. Kinoshita,M. Miyoshi,B.H. Juang, Speech dereverberation based on variance-normalized delayed linear prediction, IEEE Transactions on Audio, Speech, and Language Processing 18,2010, pp. 1717–1731.
[42] T. Yoshioka,T. Nakatani, Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening, IEEE Transactions on Audio, Speech, and Language Processing 20,2012, pp.2707– 2720.
[43] L. Drude, J. Heymann, C. Boeddeker, R. Haeb-Umbach, NARAWPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and o_ine processing, in: SpeechCommunication; 13th ITG-Symposium, VDE, 2018, pp. 1–5.
[44] X. Anguera, C. Wooters, J. Hernando, Acoustic beamforming for speaker diarization of meetings, IEEE Transactions on Audio, Speech,and Language Processing 15,2007,pp.2011–2023.
[45] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, F. Alleva, Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 3038–3042.

[46] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, R. Haeb-Umbach, Front-end processing for the CHiME-5 dinner party scenario, in: Proceedings of CHiME 2018 Workshop on SpeechProcessing in Everyday Environments, 2018, pp. 35–40.
[47] D. Haws, D. Dimitriadis, G. Saon, S. Thomas, M. Picheny, On the importance of event detection for asr, in: Proceedings of IEEE InternationalConference on Acoustics, Speech and Signal Processing, 2016.
[48] T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, All-neural online source separation, counting, and diarization for meeting analysis, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 91–95.
[49] Y. Jiang, K. A. Lee, L. Wang, Plda in the i-supervector space for textindependent speaker verification, EURASIP Journal on Audio, Speech, and Music Processing ,2014,pp. 1–13.
[50] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting10,000 classes, in: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, 2014, pp. 1891–1898.
[51] K. J. Han, S. S. Narayanan, A robust stopping criterion for agglomerativehierarchical clustering in a speaker diarization system, in: Proceedingsof the Annual Conference of the International Speech CommunicationAssociation, 2007.
[52] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, J. Martinez, Fast incrementalclustering of gaussian mixture speaker models for scaling upretrieval in on-line broadcast, in: Proceedings of IEEE InternationalConference on Acoustics, Speech and Signal Processing, volume 5,IEEE, 2006.
[53] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, A. Avdeeva, A. Gorlanov, A. Kozlov, Speaker diarization with deep speaker embeddings for dihard challenge ii., in: Proceedings of the Annual Conference of the International Speech Communication Association,2019, pp. 1003–1007.
[54] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, S. Khudanpur, Diarization is hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge, in: Proceedings of theAnnual Conference of the International Speech Communication Association, 2018, pp. 2808–2812
[55] P. Kenny, D. Reynolds, F. Castaldo, Diarization of telephone conversations using factor analysis, IEEE Journal of Selected Topics in Signal Processing 4, 2010,pp. 1059–1070.
[56] M. Diez, L. Burget, F. Landini, J. Cˇ ernocky`, Analysis of speaker diarizationbased on BayesianHMMwith eigenvoice priors, IEEE/ACM Transactionson Audio, Speech, and Language Processing 28 ,2019, pp. 355–368.
[57] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second DIHARD diarization challenge: Dataset, task, and baselines, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 978–982
[58] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The first dihard speech diarization challenge, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018.
[59] T. J. Park, K. J. Han, M. Kumar, S. Narayanan, Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap, IEEE Signal Processing Letters 27 (2019) 381–385.
[60] D. Dimitriadis, P. Fousek, Developing on-line speaker diarization system, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 2739–2743.
[61] K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, Overlapped speech detection for improved speaker diarization in multiparty meetings, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 4353–4356.
[62] O. Cetin, E. Shriberg, Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, IEEE, 2006, pp. 357–360.
[63] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, R. Haeb-Umbach, Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 1248– 1252
[64] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, H. Xuedong, Meeting Transcription Using Asynchronous Distant Microphones, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 2968–2972.
[65] L. E. Shafey, H. Soltau, I. Shafran, Joint Speech Recognition and Speaker Diarization via Sequence Transduction, in: Proceedings of the Annual Conference of the International Speech Communication Association, ISCA, 2019, pp. 396–400.
[66] D. Dimitriadis, Enhancements for Audio-only Diarization Systems, arXiv preprint arXiv:1909.00082 (2019).

[67] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings of International Conference on Machine Learning, 2016, pp. 478–487.
[68] X. Guo, L. Gao, X. Liu, J. Yin, Improved deep embedded clustering with local structure preservation, in: Proceedings of International Joint Conference on Artificial Intelligence, 2017, pp. 1753–1759.