ARQII Publication

APPLICATIONS OF
MODELLING AND SIMULATION

http://aPRQiipubl.com/ams eISSN 2600-8084 VOL 9, 2025, 22-36

Systematic Literature Review of Speaker Diarization Techniques: Toward Bridging Gaps in Low-resourced Languages using Machine Learning

Mohd Zulhafiz Rahim[1], Sarah Samson Juan[1,2*] and Syahrul Nizam Junaini[1]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]Data Science Centre, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

*Corresponding author: sjsflora@unimas.my

Submitted 21 September 2024, Revised 29 October 2024, Accepted 13 November 2024, Available online 02 January 2025.

Abstract: Speaker diarization, the process of segmenting audio into speaker-specific regions, plays a critical role in various speech technologies by determining "who spoke when" in a conversation. This technique is particularly valuable for enhancing automatic speech recognition (ASR) and conversational artificial intelligent systems. However, its application to low-resourced languages remains underexplored, limiting not only the performance of speaker diarization among low-resourced languages, but also stagnating the advancements of ASR to low-resourced languages. This is due to the fact that speaker diarization enables speaker adaptation in ASR, crucial for maximizing the performance of ASR itself. This lack of digital resources of speaker diarization to low-resourced languages, as well as the scarcity of its implementation presents a gap between low-resourced languages and popular languages in terms of the advancements of speech technologies involving the particular languages. This paper focuses on Sarawak Malay, a low-resourced language, and presents conversational data collected through a crowd-sourced approach, which needs speaker turns and transcripts. These missing annotations create challenges for building accurate acoustic models. To address this, we conducted a systematic review of recent speaker diarization research and related machine learning techniques. Using the PRISMA methodology, we reviewed 42 articles published between 2018 and 2023. Our findings identify key machine learning models, such as i-vectors and x-vectors, and open-source tools like Pyannote, which offer promising advancements in diarization performance. Besides that, these tools have shown potential to be implemented in developing speaker diarization models for low-resourced language. By highlighting the gaps in current research for low-resourced languages, we provide a pathway for improving speaker diarization models in these underrepresented languages through machine learning techniques.

Keywords: Deep neural network; Low-resourced; Machine learning; Speaker diarization; x-vectors.

1. INTRODUCTION

In the era of machine learning and artificial intelligence, speech technologies such as automatic speech recognition (ASR) ([1], [2], [3]), speaker verification ([4–7]), and conversational systems [8] have seen significant advancements. These technologies rely heavily on accurately identifying speakers within an audio stream, a task referred to as speaker diarization. Speaker diarization involves segmenting audio into speaker-specific regions to determine "who spoke when." This task is essential for applications requiring speaker-specific insights, such as in meeting transcription services, voice assistants, or forensic audio analysis [9].

Speaker diarization emerged in the 1990s, initially used to identify speakers in air traffic control communications [10] or broadcast news recordings [11,12]. The technology has since evolved, incorporating advanced statistical and machine learning techniques to improve accuracy. Traditional methods such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) laid the groundwork for speaker segmentation. Still, these techniques often needed to be improved with variability in speakers' voices and environmental noise. Over time, the introduction of i-vectors [1,13] and later x-vectors [14], which leverage deep neural networks (DNNs), marked significant improvements in diarization performance, particularly in large-scale, multi-speaker environments.

Despite these advancements, speaker diarization research has primarily focused on high-resourced languages such as English, French, and Spanish, where vast amounts of annotated audio data are available for training models. Low-resourced languages, however, face challenges due to the need for annotated datasets, phonetic resources, and linguistic tools necessary for building robust diarization systems. These languages often exhibit significant variability in dialects, accents, and speaker styles, complicating the development of accurate diarization models. To achieve substantial results while improving the performance of speaker diarization, at least 100 hours of labelled data would be required for development of model from scratch [15], and even optimization of existing models would require at least 10 hours of labelled data [16].

Sarawak Malay, a variant of the Malay language spoken by over one million people in Malaysia, exemplifies these challenges. As a low-resourced language, Sarawak Malay has limited speech and text data availability for training machine learning models. This lack of resources has hindered the development of accurate and reliable speech applications for the language. Although Sarawak Malay is a stable language used widely in conversational settings, it remains underrepresented in speech processing research. In recent years, machine learning techniques such as x-vectors, which are embeddings derived from deep neural networks, have shown great promise in improving speaker diarization performance. Open-source tools like Pyannote provide accessible frameworks for implementing these models, even for languages with limited resources. However, significant gaps remain in applying these state-of-the-art techniques to low-resourced languages like Sarawak Malay.

This paper aims to bridge that gap by reviewing recent advancements in speaker diarization techniques, specifically focusing on low-resourced languages. Using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, we systematically reviewed studies published between 2018 and 2023 to identify key machine learning models and tools that can be applied to speaker diarization for low-resourced languages. By highlighting the challenges and opportunities in this field, this paper provides a path for improving speaker diarization models for Sarawak Malay and other low-resourced languages using machine learning.

1.1 Speaker Diarization

Speaker diarization, by definition, is a task to segment conversational speech into corresponding sections based on the speakers' identities. Speaker diarization systems can annotate speech signals with time boundaries and speaker IDs to label speaker change. In simpler terms, speaker diarization determines “who spoke when?” in an audio conversation. Initially, the research objective of diarization technology was to benefit ASR by enabling speaker adaptations in speech transcriptions [17]. In the 1990s, the earliest forms of speaker diarization were for speaker identification for dialogues between an air traffic controller (ATC) and several pilots and speaker adaptation on news broadcasts [10–12]. In more recent research, speaker diarization has been implemented in multi-talker ASR pipelines, where speaker diarization is included as part of the foundational benchmark alongside speech recognition and speech separation [18].

Figure 1 shows a general process of speaker diarization. The initial step is speech detection, where audio-containing regions are identified. Afterward, the audio signal is partitioned into small segments based on the speaker boundaries, creating a timeline or “diarization” of “who spoke when?”. Next, feature extraction occurs, where the acoustic features are extracted for clustering. The speech segments are then grouped up into clusters based on the speaker’s acoustic features that were extracted prior. The speaker diarization process outputs a collection of segmented and clustered utterances, with each segment representing its respective speaker.

However, the implementation of speaker diarization poses challenges during the data preparation process because the statistical methods require large amounts of data to produce better models. A large dataset could increase training data diversity, allowing the diarization model to learn the different interaction styles and speaker features, leading to a higher performance during the diarization process. Researchers have addressed this issue by developing large open-source datasets with annotations such as MUSAN [7] and AMI Corpus [19], which have become popular and extensively used in speaker diarization tasks. Nevertheless, low-resourced languages still need to perform better due to a lack of annotated data or insufficient resources for training statistical models. A low-resourced or under-resourced language is described as a language that suffers from limited data for representing orthography systems, minimal presence in digital applications, a shortage of language experts, or a lack of automated resources for building natural language processing (NLP) systems such as speech recognition, speaker identification, machine translation and many more [20,21]. Malaysia, for example, has 140 living languages and dialects, most of which are unwritten. Thus, this country has many low-resourced languages, including Sarawak Malay, a dialect widely spoken in Sarawak, located northwest of the Borneo Island.

1.2 Sarawak Malay Language

Standard Malay is the national language of Malaysia and there are various variants and dialects of the Malay language in the context of regional differences [22] such as Sarawak Malay, Kelantan Malay and Sabah Malay. Sarawak Malay dialect, a variant of the Malay that belongs to the Austronesian language family [23,24] is widely spoken by approximately 1,000,000 people of all races and ethnicities, such as the Sarawakian Malays, the Ibans, Bidayuhs, and even the Chinese and the Indians. Furthermore, the dialect is commonly used in conversations within the Sarawak community, thus, making it a stable language [24]. However, despite many dialect users, it is still considered a low-resource NLP task, as limited resources (text and speech data) are available for building applications. Therefore, machine learning research on this target language is still lacking, and only a few studies can be found.

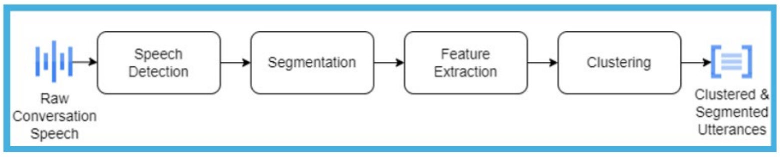

Figure 1. Speaker diarization process.

Knowing these challenges for Sarawak Malay and other languages in Malaysia, we tackled the prime issue in machine learning: data. We developed the first version of *Kalaka*: Language Map of Malaysia website in 2018 (https://kalakamap.unimas.my/kalaka). *Kalaka* is a website created to store language data and preserve and revitalize Indigenous languages in Malaysia. As of 2024, there are a total of 1244 recordings uploaded to the website among 144 languages, 461 registered system users and 781 registered speakers. The language with the most recordings uploaded to the website is the Sarawak Malay dialect, with 237 audio recordings containing 14 hours of conversations. We gathered the data through crowdsourcing by giving course assignments to Universiti Malaysia Sarawak students, who were given several topics for discussion, such as traditional stories and games, to record interview sessions with their speakers. Then, the students uploaded their data (audio files and speaker details) to the Kalaka website.

Annotating speaker turns and transcribing conversational data is a time-consuming and labour-intensive task. To address this challenge, we aim to design a computational approach to accelerate the process. While we have collected sufficient Sarawak Malay conversational data, it remains unlabeled. The next step involves conducting a detailed review of recent works on speaker diarization to identify the most effective methodologies, tools, and techniques for diarizing Sarawak Malay speech data. This paper, guided by the PRISMA methodology, systematically reviews existing research to uncover speaker diarization methods, highlight trends over the years, and explore their application to low-resourced languages. By leveraging state-of-the-art techniques, tools, and evaluation metrics identified in the latest research, we seek to improve the performance of speaker diarization models for low-resourced languages. The following section provides an overview of the PRISMA framework and its results.

1.3 PRISMA Protocol

PRISMA stands for Preferred Reporting Items for Systematic Reviews and Meta Analyses. It is a set of guidelines aiming to improve the quality, transparency and completeness of systematic reviews and meta-analyses reporting and to facilitate the critical appraisal and use of these studies by researchers [25]. The strategy consists of a 27-item checklist and a four-phase flow diagram that describes the essential information to include in a systematic review or meta-analysis report. The PRISMA methodology also guides the registration of a systematic review, the search and selection of studies, the assessment of the risk of bias, the synthesis and analysis of the data and the reporting of the results and implications.

Table 1 summarises three research articles reporting their PRISMA strategy for conducting a systematic literature review (SLR). Despite having different research aims, each research implemented similar steps for conducting a systematic review, where a specific search strategy was undertaken to obtain past research within a timeframe. Besides answering research questions, each research implemented its inclusion criteria to screen past articles. For example, Alharbi et al. [26] screened their articles by filtering them for relevance to ASR, focusing on the English language and passing quality assessments such as articles stating the research aim, providing new techniques or contributions to ASR and mentioning challenges related to ASR. On the other hand, Deka et al. [27] only included full articles, review papers and short papers proposing automated speech therapy tools using artificial intelligence techniques such as machine learning and deep learning. Furthermore, Jahan and Oussalah [28] screened their articles for eligibility by ensuring each article is a review or survey document strictly related to Hate Speech (HS) or Computer Science and Engineering (CSE) domains.

Table 1. Systematic literature reviews using PRISMA protocol.

Research Title	Cited Paper	Aim	Databases	Search Strategy	Publication Year	Number of Articles
Automatic speech recognition: systematic literature review	[26]	To study research trends in ASR and suggest new research directions.	IEE, ACM, Scopus, Web of Science, Science Direct	"artificial intelligence" AND ("speech recognition" OR "automatic speech recognition")	2015-2020	82
AI-based automated speech therapy tools for persons with speech sound disorders: a systematic literature review	[27]	To investigate AI-based automated speech therapy tools for speech sound disorders (SSD) and suggest future directions for this field.	IEEE, ACM, Scopus	("AI" OR "Artificial Intelligence" OR "automa*") AND ("speech" OR "language") AND ("disorder" OR "impairment") AND ("assessment" OR "therapy" OR "rehabilitation" OR "treatment")	2007-2022	24
A systematic review of hate speech automatic detection using natural language processing	[28]	To review the current state and challenges in hate speech detection using machine learning and deep learning.	Google Scholars, ACM	("Review" OR "survey") AND ("hate speech detection" OR "abusive language detection" OR "sexism detection" OR "cyberbullying detection")	2000-2021	7

2. MATERIAL AND METHODS

Figure 2 depicts our PRISMA flow for selecting relevant articles for review. The initial identification phase retrieved 1,614 records from four databases—IEEE, ACM, Scopus, and Web of Science after we applied the following keywords in our search: (“speaker diarization” OR “speaker recognition”) AND (“low-resourced” OR “under-resourced” OR “machine learning”). After removing 32 duplicates and excluding 1,101 records with fewer than five citations, 481 records remained for screening. At this stage, we assessed the titles and excluded 318 records, leaving 163 reports for abstract-based evaluation. Following this, 121 reports were excluded due to irrelevance based on their abstracts.

As a result, 42 studies were included in the final review and Table 2 reports the distribution of the articles found in respective databases. These studies were selected based on specific inclusion criteria: publication between 2018 and 2023, at least five citations, and relevance to speaker diarization and speaker recognition in low-resourced languages. Exclusion criteria involved removing duplicates, studies with fewer than five citations, and those irrelevant to the topic. Applying the PRISMA framework allowed us to select high-quality and relevant studies for the systematic review, rigorously.

2.1 PRISMA Research Questions on Speaker Diarization

The foundation of this systematic review was the development of specific research questions to guide the investigation. The primary objective of this study is to review recent advancements in speaker recognition, with a particular emphasis on speaker diarization in low-resource language settings. To achieve this, five key research questions were formulated to address critical aspects of the topic. The detailed research questions and their corresponding aims are outlined in Table 3.

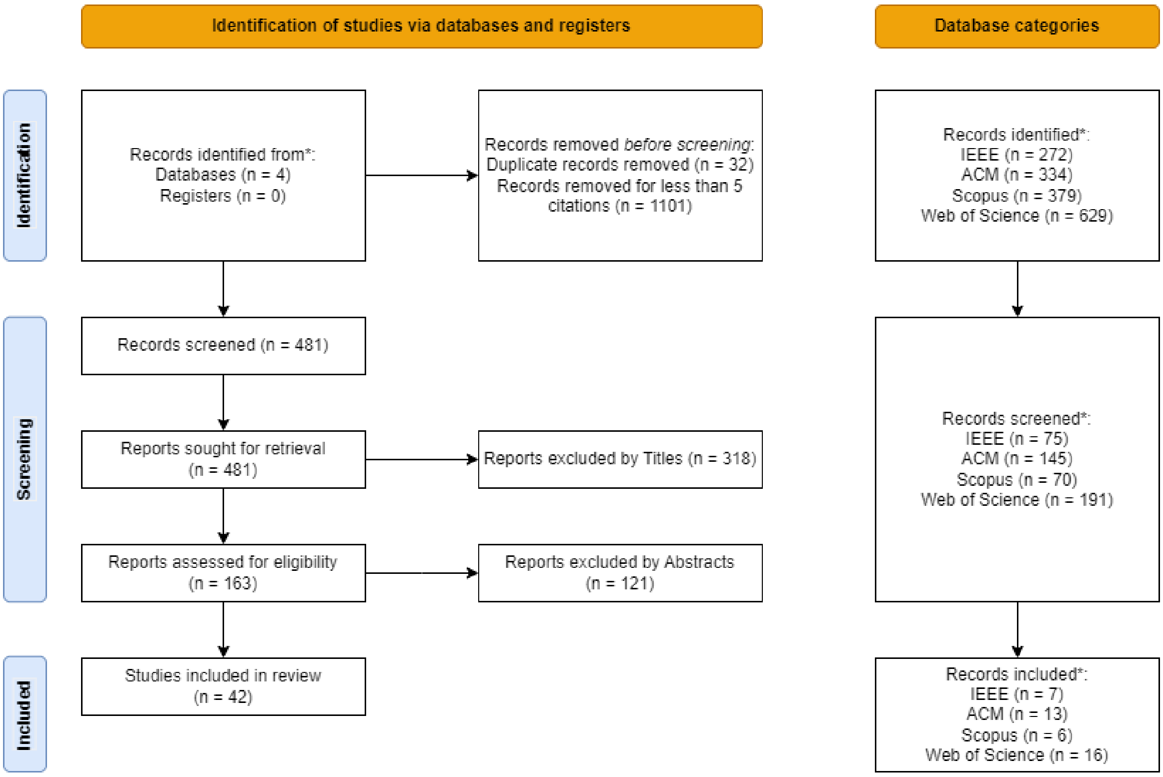
Identification of studies via databases and registers
Database categories
Identification
Records identified from*:
Databases (n = 4)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 32)
Records removed for less than 5 citations (n = 1101)
Records identified*:
IEEE (n = 272)
ACM (n = 334)
Scopus (n = 379)
Web of Science (n = 629)
Screening
Records screened (n = 481)
Reports sought for retrieval (n = 481)
Reports excluded by Titles (n = 318)
Reports assessed for eligibility (n = 163)
Reports excluded by Abstracts (n = 121)
Records screened*:
IEEE (n = 75)
ACM (n = 145)
Scopus (n = 70)
Web of Science (n = 191)
Included
Studies included in review (n = 42)
Records included*:
IEEE (n = 7)
ACM (n = 13)
Scopus (n = 6)
Web of Science (n = 16)

Figure 2. The PRISMA flow chart for selecting articles related to speaker diarization techniques.

Table 2. Number of articles selected after screening.

Database	Number of Articles Selected for Review
IEEE Xplore Digital Library	7
ACM Digital Library	13
Scopus	6
Web of Science	16
Total	**42**

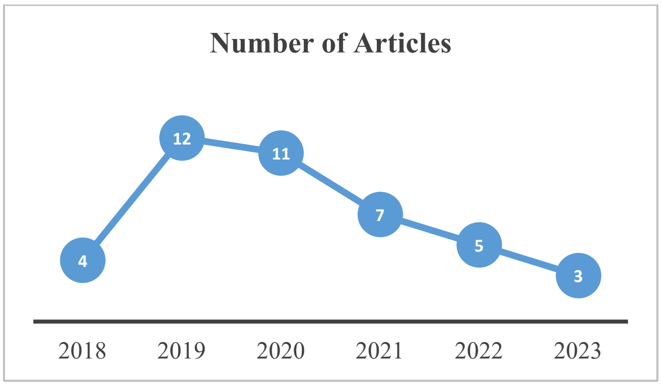

Figure 3. Statistics of the selected articles based on publication year.

Table 3. PRISMA research questions for investigating speaker diarization and its application to low-resourced languages.

No.	PRISMA Research Question (PRQ)	Aim
1	What are the research problems with speaker diarization?	To determine the general issues researchers tend to face while researching this speaker diarization.
2	How has speaker diarization evolved throughout the years?	To review the chronology of the advancements made in the speaker diarization approaches made by researchers to tackle the research problems that they experience
3	What applications and tools are used to conduct speaker diarization on conversation data?	To ascertain the different tools and applications that can be used to perform the task of speaker diarization.
4	How do we evaluate the performance of speaker diarization models based on conversation data?	To determine the relevant evaluation metric used to measure speaker diarization performance.
5	What are the significant challenges in speaker diarization for low-resourced languages?	To identify the significant challenges while performing speaker diarization for low-resourced languages.

3. DESCRIPTIVE ANALYSIS

3.1 Publication Timeline for Speaker Diarization Studies

Figure 3 presents the publication year statistics for the selected studies. According to the pie chart, most speaker diarization studies were published in 2019 and 2020. This surge can be attributed to the introduction of DNN embeddings in 2017, further advanced and reintroduced as x-vectors in 2018. This new state-of-the-art diarization technique sparked significant interest among researchers, leading to increased studies exploring and applying this method in the subsequent years.

3.2 Keyword Inclusion

Table 4 presents the number of studies that include specific keywords within the papers. Speech recognition appears in all the studies, reflecting the interconnection between speech recognition, speaker recognition and speaker diarization. In terms of techniques, "DNN" and "machine learning" are the most frequently mentioned keywords, surpassing "i-vectors," likely due to the growing trend of researchers adopting DNN and machine learning approaches over i-vectors. Few studies address low-resourced or under-resourced languages, largely because speaker recognition and diarization are often considered language-agnostic fields. This is due to their focus on speaker-specific features rather than transcriptions of speech input.

Table 4. Keywords found in the articles.

Keyword	Number of articles with the keywords
speech recognition	42
speaker recognition	38
dnn	35
speaker diarization	32
machine learning	32
i-vector	30
unsupervised	28
gmm	28
hmm	22
x-vector	20
low resource	11
self-supervised	10
under-resourced	3

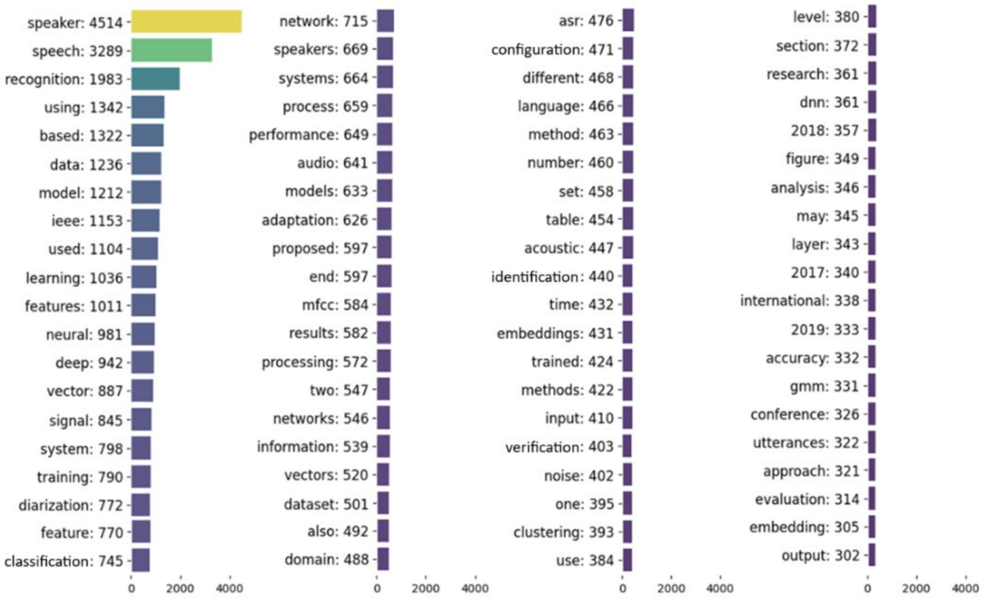

Figure 4. A word frequency graph was generated from the articles using matplotlib.

3.3 Word Frequency

Figure 4 shows a word frequency graph illustrating the most common terms extracted from the reviewed articles. The graph was generated using the Matplotlib library and highlights the frequency of specific keywords related to speaker diarization and recognition within low-resourced languages. The words "speaker," "speech," and "recognition" are the most frequently mentioned, as they represent the core focus of the research on speaker diarization and speaker recognition. Among the various words, those that stand out as technique-related are "deep" and "neural," which are frequently mentioned, highlighting the widespread use and implementation of Deep Neural Networks (DNN) in these studies.

4. RESULTS

4.1 What are the Research Problems with Speaker Diarization? (PRQ1)

Speaker diarization has long been an area of research, yet several challenges have been identified throughout its development. These challenges were revealed during the review of speaker diarization itself, particularly regarding the techniques used, recent advancements and real-world applications. One of the key issues in implementing speaker diarization is its strong dependency on data. As mentioned by Park et al. [17] speaker diarization techniques are data-driven, requiring a large amount of annotated data to train the models effectively. Additionally, the datasets must be structured and organized consistently, which demands considerable human effort in preparation and annotation.

Another challenge lies in the variability of speaker diarization performance, which is influenced by numerous factors, such as the speaker's style, accent, the environment in which the speech occurs and the quality of the raw audio [29]. To achieve consistent and reliable diarization results, models must be trained on diverse datasets [30] that includes a sufficient amount of data for each variety, as discussed, to ensure a better diarization performance of the models.

Furthermore, regarding variability, speech data involving low-resource languages tend to be very different compared to high-resource ones [31] and thus are more varied. Therefore, more resources are needed for the speaker diarization field if it is to be implemented in low-resource languages. The challenges in implementing speaker diarization for low-resourced languages will be further discussed in detail while answering the fifth research question for this chapter.

The channel usage during the application of speaker diarization also poses another challenge in this process. This is because real applications of speaker diarization typically involve numerous types of phones, such as landline phones, payphones, cordless phones and cell phones [32]. However, this difference in channel usage advances their influence on channel efficiency, as the high variability of cross-channel usage further affects speaker diarization performance.

4.2 How has Speaker Diarization Evolved Throughout the Years? (PRQ2)

The field of speaker diarization has seen continuous research and experimentation over the years, resulting in significant advancements in methods for analysing conversation data. Over time, researchers proposed various algorithms for speaker diarization tasks, ranging from conventional methods such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) to more modern techniques such as deep learning and machine learning. The exploration and experimentation of these algorithms have led to identifying strengths and weaknesses, fostering the development of more effective and accurate methods. The increasing demand for speaker diarization in real-world applications (transcription services, voice assistance, surveillance systems) has also motivated researchers to develop practical and scalable solutions.

In the early stages of speaker diarization, GMMs were often used to model the acoustic features of different speakers [30]. At the same time, HMMs were employed to capture the temporal dependencies in the data. This approach involved clustering the feature vectors into different Gaussian components, each representing a speaker and then were used model the transitions between these speaker states [30]. A notable implementation of this method was by Université du Maine (LIUM), which developed an open-source toolkit called LIUM_spkdiarization [33] utilising GMMs and HMMs for speaker diarization. However, LIUM researchers found that both GMM and HMM models are sensitive to environmental variability, making their performance inconsistent in different conditions. To mitigate this issue, researchers often optimise their models using more advanced techniques such as Probabilistic Linear Discriminant Analysis (PLDA) [13] or employ superior models like Recurrent Neural Networks (RNNs) [34] to capture transitions between speaker states better and improve diarization accuracy.

With the recent rise of deep learning methods, researchers began utilising Recurrent Neural Networks (RNNs) for speaker diarization modelling. The critical advantage of RNNs lies in their superior ability to model temporal dependencies, surpassing traditional HMMs in this context. RNNs are particularly effective in processing sequential input data, such as audio features over time and making predictions about speaker changes [35]. Hence, RNNs are often combined with GMMs in speaker diarization tasks, leveraging the strengths of both methods to improve performance in identifying and segmenting speakers.

In the late 2010s, i-vectors were introduced as a compact representation of speaker identity. These vectors function by capturing the variability of an individual's speech characteristics. The i-vector approach involved extracting a fixed-dimensional vector for each speech segment and using it for clustering or classification to identify speakers [13]. Because this model is trained discriminatively, classification errors can be minimised directly, making it less sensitive to environmental variability and more focused on the speaker's unique features. This not only outperforms traditional GMM-based methods in diarization accuracy but also reduces the need for extensive model optimisation, thus requiring less human effort. Even so, i-vectors face limitations when dealing with large datasets, particularly in scenarios where the diversity of the testing data is low, as their performance can degrade under such conditions [36].

With the advancements in deep neural networks, x-vectors were introduced to leverage these networks to extract speaker embeddings. This model captures hierarchical and complex features from the input signal, providing a more discriminative representation [14]. X-vectors are similar to i-vectors in that they have fixed-dimensional embeddings. Both models are suitable for integration into systems that require consistent input dimensions. However, x-vectors have shown superior performance, especially in large data scenarios, where they excel over i-vectors[37].

While i-vectors rely on simpler statistical models, x-vectors leverage deep neural networks, allowing for the capture of more complex and discriminative features. Although i-vectors are usually discriminative, they are still trained using a generative model approach within the Gaussian Mixture Model – Universal Background Model (GMM-UBM) [4]. In contrast, x-vectors are trained end-to-end using discriminative, as outlined by Snyder et al. [14]. Thus, makes x-vectors more proficient in capturing complex features and diversity of large datasets (different speaking styles, etc.), which can, in turn, improve speaker diarization performance.

Table 5 summarises the key techniques relevant to speaker diarization research. In summary, the GMM and HMM method is the most versatile, but faces the problem of inability to model complex pattern [39] and environment-sensitivity, where the less of the complex speaker features are being processed [31]. The RNN method is able to model complex patterns while still being versatile, but still insensitive to slightly differing speaker features [35], while demanding a high amount of computational power [41]. The i-vectors method was introduced to detect and model complex speaker features better, but the performance degrades when there is little difference between speaker features in large datasets [47], which poses a problem while optimizing speaker diarization of the same language.

Table 5. Summary of implemented speech processing techniques: reviewed articles.

Techniques	Researchers	Language (*low-resourced)	Results & Key Takeaways	Advantages & Disadvantages
GMM, HMM	Meignier & Merlin [33]	French, English	Introduces open-source diarization toolkit using GMM and HMM techniques	Advantages: • Simple and efficient • Probabilistic framework for modelling data • Versatile Disadvantages: • Environment-sensitive • May not capture complex data dependencies (i.e. speaker features) • Unscalable for large and complex datasets
	Anguera et al. [38]	American English	Reviews recent research (2012 and below), mostly implementing GMM and HMM techniques	
	Alsharhan & Ramsay [39]	Arabic	Showcases the importance of a more extensive dataset in model training	
	Sethuram et al. [31]	Telugu*	Highlights the environment variability sensitivity of GMM and HMM, requiring extra optimisations	
RNN	Kanwal et al. [34]	Urdu*	Discusses the challenges of speaker diarization for low-resourced language,	Advantages: • Can model complex patterns

			limitations of technique used (RNN)	• Versatile Disadvantages: • Insensitive to speaker features variability • Resource intensive (computational power, processing time)
	Li et al. [40]	Multi-lingual	Achieves speaker error rate (SER) of 29.4%, which is not a state-of-the-art level	
	Nammous et al. [41]	Arabic, English, Polish	Describes the RNN challenges for Arabic language regarding variability (background noise, voice similarity, age span)	
	Jati et al. [35]	Multi-lingual	Displays RNN NPC embeddings underperforming compared to in-domain i-vector and x-vector methods	
i-vectors	Mane et al. [42]	Unstated	Reviews the advantages of i-vectors, compared to older GMM and HMM methods	Advantages: • Scalable for large and complex datasets • Sensitive to speaker features variability Disadvantages: • Performance degrades with high quantity of data with low diversity • Requires implementation of other techniques for extra optimisation
	Karadayi et al. [36]	Tsimane*, Moseten*	Showcases the poor performance of i-vectors in surroundings with a low prevalence of speech and similar types of voices (children's voices)	
	Kanda et al. [43]	Multi-lingual	Suggests the implementation of x-vectors instead of i-vectors for future work to optimise performance	
	Lin et al. [13]	English	Discusses further optimisations that can be done towards the parameters to improve i-vectors performance in the future	
	Kang & Kim [4]	Multi-lingual	Describes the degradation in performance for i-vectors for short-duration speech utterances, extra optimisation with GMM needed	
x-vectors/DNN embeddings	Snyder et al. [7]	English	The first implementation of DNN embeddings (x-vectors), advantages over i-vectors highlighted	Advantages: • Deep learning integration • Sensitive to speaker features variability • State-of-the-art performances in speaker recognition tasks Disadvantages: • Requires large amounts of labeled data for training • Resource intensive (computational power, processing time)
	Thanh et al. [6]	Vietnamese*	Discusses the optimisers that can be used with DNN embeddings (SGD and Adam optimisers) to improve verification performance on low-resource language further	
	Bai & Zhang [44]	Multi-lingual	Reviews of recent research (2021 and below), highlights x-vectors being the state-of-the-art models	

	Levow [45]	Lengthy list of low-resourced languages	Compares the performance of four different toolkits (LIUM, Kaldi, pyannote, VBx), emphasises on the high performance of toolkits using x-vectors (Kaldi, pyannote, VBx)	
	Bredin et al. [46]	Multi-lingual	Introduces open source diarization toolkit implementing x-vectors, showcases toolkit's user friendliness	

As identified from the reviewed articles, the table highlights that x-vector models consistently deliver state-of-the-art performance in speaker recognition tasks, including speaker verification, speaker identification and speaker diarization. During its first implementation, Snyder et al. [14] discovered that the x-vectors system consistently outperforms the i-vectors system, achieving an improvement of 44% in EER (equal error rate) and 29% in DCF (detection cost function), compared to i-vectors system's improvement of 32% in EER and 17% in DCF.

Consequently, x-vector models can be further exploited for future diarization modelling to see its capabilities and level of performance. As previously discussed, x-vectors are speaker embeddings—fixed-dimensional vector representations of speech utterances that capture speaker-specific characteristics extracted through deep neural networks (DNN). DNNs are a type of artificial neural network (ANN) distinguished by their depth, with multiple layers between the input and output layers[48].

Figure 5 shows the visual representations of the layers and the nodes inside the DNN. Each layer comprises interconnected nodes, neurons or units organised into input, hidden and output layers. Inside the DNN, the input layers function to receive the initial data or features. The hidden layers are intermediate between the input and output layers that process the input data through a series of weighted connections and activation functions. The output layer, on the other hand, produces the final output or prediction. The nodes inside the DNN are computational units that process information. Each node receives the input, performs a weighted sum, applies an activation function (complex patterns and relationships) [49] and produces an output for the other nodes.

As x-vectors are extracted from these networks, the input layer takes the speech features as vectors. It has them interconnected to generate a set of predefined speakers in the form of vectors inside the hidden layers. Each node inside the hidden layer processes all the different patterns and variability of speaker features before sending them to the next hidden layer or output layer. The output layer predicts the speaker identity from the predefined speakers set by the hidden layers, returning x-vectors. The result of this network is a robust model that is very discriminative of speaker features, resulting in a better speaker recognition performance and highly leverageable on large-scale training datasets compared to other techniques.

During its first implementation by Snyder et al. in 2017 [7], x-vectors were introduced to completely replace the i-vector extraction process in the speaker diarization pipeline, marking a significant advancement in state-of-the-art techniques. The shift occurred because using i-vector clustering for short speech segments was considered cumbersome and costly for front-end processing, as it involved a two-step generative process. This process required the extraction of i-vectors and applying a probabilistic linear discriminant analysis (PLDA) scoring function to determine whether two i-vectors originated from different speakers. In contrast, x-vectors, introduced as DNN embeddings, were designed to jointly learn a fixed-dimensional embedding and a scoring metric. This streamlined the process and resulted in superior diarization performance compared to previous i-vector methods, establishing x-vectors as the new state-of-the-art technique for speaker diarization.

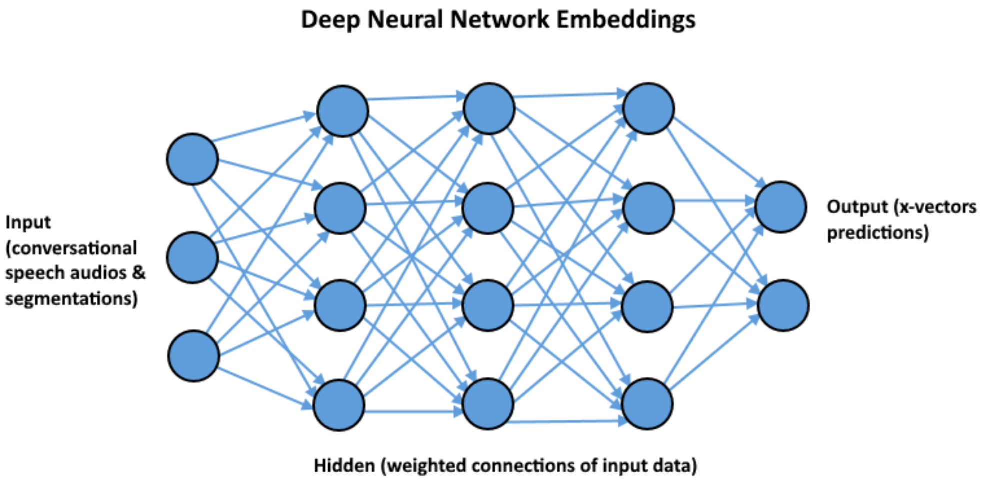
Deep Neural Network Embeddings
Input (conversational speech audios & segmentations)
Output (x-vectors predictions)
Hidden (weighted connections of input data)

Figure 5. Deep neural networks representation.

4.3 What Applications and Tools are Used to Conduct Speaker Diarization on Conversation Data? (PRQ3)

Kaldi is an open-source toolkit for developing automatic speech recognition (ASR) systems. It provides tools for feature extraction, acoustic modelling, decoding and supports state-of-the-art algorithms like Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs)and Deep Neural Networks (DNNs) [48]. Written in C++, Kaldi is widely used in transcription, speaker diarizationand spoken language understanding tasks. Its modular architecture allows for customisation, making it popular among researchers and engineers. However, Kaldi has a steeper learning curve, requiring a strong understanding of speech processing [50].

VBHMM x-vectors Diarization, often called VBx, is a state-of-the-art approach to speaker diarization, a task involving segmenting and labelling an audio stream based on different speakers [51]. This method utilises various techniques, including Variational Bayesian Hidden Markov Models (VBHMM) and x-vectors. "VBx" refers to applying Variational Bayesian methods in diarization tasks. In VBx diarization using x-vectors, the system leverages probabilistic models to estimate the most likely sequence of speakers in an audio stream [52]. By incorporating the variability captured in x-vectors, VBx aims to enhance the accuracy and robustness of speaker diarization, particularly in scenarios with multiple speakers and diverse acoustic conditions.

ALIZE-Speaker-Recognition is a software toolkit designed for speaker recognition tasks. It provides tools and algorithms for building speaker recognition systems, which involve identifying or verifying individuals based on their voice characteristics. ALIZE, "A Lattice Implementation of the Z/n Lattice," refers to the lattice-based algorithms used in the toolkit [53]. ALIZE-Speaker-Recognition typically employs Gaussian Mixture Models (GMMs) as one of the primary modelling techniques. More advanced approaches, such as deep neural networks (DNNs) and neural embeddings like x-vectors, are also commonly used in modern speaker recognition systems [53]. However, it is easier to provide precise information about all the models involved with specific details about the version or components of ALIZE-Speaker-Recognition being used.

Pyannote is an open-source Python library designed for speaker diarization and speaker embedding tasks in speech processing ([45,46]). Developed by the Pyannote-Audio team, it provides a comprehensive set of tools and algorithms for various tasks related to speaker analysis in audio data. The library supports tasks such as speaker diarization (segmenting and labelling speakers in an audio stream), speaker embedding (extracting speaker representations)and speaker change detection. The significant advantage of this toolkit is that it is written in Python, a programming language with a narrow learning curve. Other than that, Pyannote is designed to be user-friendly and provides a higher-level API (providing many models for different functions), making it more straightforward to use even for users who need to become more adept in deep learning or signal processing.

Table 6. Advantages and disadvantages of applications/tools.

Application/Tool	Advantages	Disadvantages	Cited Papers
Kaldi	▪ Supports various algorithms & techniques ▪ Highly customisable ▪ Can be combined with other models ▪ Strong community support	▪ Steep learning curve	[49,50,54]
VBHMM X-Vectors Diarization (VBx)	▪ Employs multiple techniques (HMM, x-vectors, AHC)	▪ Requires initialisation upon use ▪ High complexity	[51,52]
ALIZE	▪ Supports various techniques ▪ Provides a set of tools and algorithms related to speaker recognition	▪ Complicated version handling	[53]
Pyannote	▪ Narrow learning curve ▪ User friendly ▪ Highly customisable ▪ Provides a set of tools and algorithms related to speaker analysis ▪ Strong community support	▪ Data-hungry model training	[19,46]

4.4 How Do We Evaluate the Performance of Speaker Diarization Models Based on Conversation Data? (PRQ4)

The primary evaluation metric for assessing speaker diarization performance is the Diarization Error Rate (DER), which accounts for three types of errors: (1) Missed Speech, where speaker segments are not detected; (2) False Alarm, where speaker labels are incorrectly assigned to empty segments; and (3) Confusion, where speaker segments are mismatched with the reference ground truth [55]. Another commonly used metric is the Word Error Rate (WER), widely applied to evaluate ASR systems. WER measures errors in transcription, including (1) Substitution errors, where words are replaced incorrectly; (2) Deletion errors, where words are omitted; and (3) Insertion errors, where extra words are introduced [56]. The Equal Error Rate (EER) is frequently used for speaker verification and identification systems. EER measures the balance between two error types: 1) False Acceptance Rate (FAR), where an impostor is mistakenly accepted and (2) False Rejection Rate (FRR), where a genuine user is wrongly rejected [7]. Lastly, the Jaccard Error Rate (JER), which evaluates the dissimilarity between two sets, is used in clustering or partitioning algorithms. It is computed by dividing misclassified pairs by the total number of pairs and was newly introduced in DIHARD II, though it is less commonly used as a secondary metric [54].

Table 7. Speaker recognition evaluation metrics.

Evaluation Metric	**Specific Field**	**Calculations**	**Cited Papers**
Diarization Error Rate (DER)	Speaker Diarization	(False Alarm + Missed Detection + Speaker Confusion) / Ground Truth Duration	[39]
Word Error Rate (WER)	Automatic Speech Recognition (ASR)	(Substitutions + Insertions + Deletions) / Number of Words Spoken	[56]
Equal Error Rate (EER)	Speaker Verification & Speaker Identification	(False Acceptance Rate + False Rejection Rate) / 2	[7]
Jaccard Error Rate (JER)	Speaker Diarization (rarely used)	(False Alarm + Missed Speech) / Total Speaker Segments Duration	[54]

4.5 What are the Significant Challenges in Speaker Diarization for Low-Resourced Languages? (PRQ5)

The challenges in speaker diarization lead to more unique ones when low-resource languages are involved. This is due to the high variability of the process of speaker diarization itself. The common understanding is that speaker diarization is considered to be language-agnostic. However, as more research is conducted on this topic, speaker diarization has advanced to the point that the state-of-the-art speaker diarization techniques are robust, thus more discriminative to different speaker features [45]. The reason for this advancement being the implementation of modern machine learning techniques, such as deep learning, where the focus of the model is less towards the environment, and more towards the speakers themselves. This advancement makes the speaker diarization models sensitive to even the smallest variability in the speech inputs, marking more obvious challenges when considering low-resource languages.

One of the challenges of performing speaker diarization for low-resource languages is that these languages typically need more extensive labelled datasets crucial for training robust speaker diarization models. The scarcity of diverse and representative data hampers the ability to develop accurate and generalisable systems, especially while trying to implement newer techniques such as x-vectors [56]. Also, low-resource languages often exhibit significant variations in accents, dialects and linguistic nuances. Developing models that can effectively handle this variability is challenging, especially when insufficient training data captures the full spectrum of linguistic variation [6].

Low-resource languages often need more comprehensive phonetic resources, making it easier to build accurate acoustic models. Without detailed phonetic variation data, developing a reliable speaker diarization system becomes a significant challenge [31]. Additionally, speakers of low-resource languages frequently engage in code-switching and multilingual conversations, further complicating speaker diarization as the system must adapt to transitions between languages and dialects. Furthermore, these languages often lack the necessary infrastructure for developing and deploying advanced speech technologies, such as speech recognition systems, language models and tools for creating annotated datasets, all essential for effective speaker diarization [36]. Because of these challenges, the task of optimizing speaker dairization for low-resourced languages significantly requires more human effort, as there are minimal available resources of low-resourced languages, requiring an abundant amount of speech data manual annotation to get substantial results.

Table 8 summarizes the articles that helped identify key techniques, trends and challenges in speaker diarization research and contributed to answering our PRISMA research questions.

Table 8. Literature review summary.

Research Question	Related Studies
What are the research problems with speaker diarization? (PRQ1)	[17,29,31,32]
How has speaker diarization evolved throughout the years? (PRQ2)	[4,6,7,13,31,33–37,39–44,46,48]
What applications and tools are used to conduct speaker diarization on conversation data? (PRQ3)	[19,46,49–54]
How do we evaluate the performance of speaker diarization models based on conversation data? (PRQ4)	[7,54–56]
What are the significant challenges in speaker diarization for low-resourced languages? (PRQ5)	[8,45]

5. DISCUSSION

5.1 Discussions after Reviewing Selected Work Guided by PRISMA

This study systematically reviewed the current state of speaker diarization techniques, with a particular emphasis on their application to low-resourced languages such as Sarawak Malay. The review revealed that while significant advancements have been made using techniques like x-vectors and DNNs, these methods are predominantly developed and optimised for well-resourced languages. This gap highlights the need for tailored approaches that address the unique challenges of low-resourced languages, including limited annotated datasets and significant linguistic variability.

The PRISMA protocol has helped us to determine the state-of-the-art techniques and tools for speaker diarization ideal for use in low-resourced languages. Based on the documented results, x-vectors consistently achieve state-of-the-art performances with the highest accuracies while performing speaker diarization, especially on large amounts of conversational speech data. Pyannote is the ideal toolkit for speaker diarization, especially for new researchers, because of its narrow learning curve, user-friendliness, customizability, and provided speech analysis algorithms (pre-trained models). Our analysis emphasises the critical role of machine learning, particularly the deep learning method, in improving speaker diarization accuracy. Advanced techniques like x-vectors leverage the robust, discriminative capabilities of DNNs to manage the complexities of speaker segmentation and identification. These methods have shown promise in handling the variability inherent in different languages, making them suitable candidates for adaptation to low-resourced contexts. However, applying these techniques to low-resourced languages is still emerging, requiring further research and development. Low-resourced languages like Sarawak Malay dialect face unique challenges that impact the effectiveness of speaker diarization. These languages often need more extensive labelled datasets, which is crucial for training robust diarization models. Additionally, the significant phonetic and syntactic variability within these languages requires models to adapt to diverse linguistic features. For instance, the complex multilingualism in Sarawak itself causes different accents and speaking intonations across the state, as the language and culture affliations shift and change [23] . Thus, x-vectors could solve these Sarawak Malay dialect speaker diarization challenges because of the availability of pre-trained x-vector speaker diarization models in the open-source PyAnnote tool. Hence, this research gap presents a need to research and exploit these x-vectors using approaches such as transfer learning to leverage Sarawak Malay data on speaker diarization.

Using the PRISMA to review the literature in research has made reviewing recent studies more convenient. This is due to the structured framework for reporting systematic reviews and meta-analyses, where authors are encouraged to utilise credible databases, develop the best search strategy based on keywords and describe the study selection, inclusion criteria and extraction steps. Furthermore, the PRISMA offers a standardised checklist and flow diagram, ensuring consistency across systematic reviews. Other researchers and reviewers will also find it convenient to conduct their research by studying the literature review using the PRISMA from other reviewers as the design is very standardised and easily understandable. Adopting the PRISMA has facilitated a structured review process that helped us select literature and prepare research questions.

5.2 Limitations

The search for articles revealed a need for more studies focusing specifically on low-resourced languages in the context of speaker diarization or recognition. These topics are widely considered language-agnostic, forcing the inclusion of the term "machine learning" in the search strategy to capture more relevant studies. Advanced techniques like x-vectors and DNNs are known for their robustness in handling linguistic variability, making this adjustment essential. However, the first inclusion criterion limited the search to articles published between 2018-2023, inadvertently excluding foundational studies employing traditional methods like GMM and HMM.

Additionally, the vast number of studies retrieved from the databases posed a challenge, as many were only loosely related to the topic despite using precise keywords. This resulted in a large volume of literature that still needed to be reviewed based on titles, as illustrated by the 481 articles in Figure 2. Furthermore, the PRISMA strategy, while systematic, does not inherently guarantee the quality of the included literature. We must still exercise significant effort to identify the most insightful and relevant studies within the context of their specific research focus.

6. CONCLUSION AND FUTURE WORK

In conclusion, this systematic literature review has highlighted the importance of advancing speaker diarization for low-resourced languages, focusing on a dialect from Malaysia, the Sarawak Malay. The review reveals that while significant progress has been made in speaker diarization for well-resourced languages using state-of-the-art techniques such as x-vectors and DNNs, there remains a substantial gap in applying these methods to low-resourced languages. Sarawak Malay's unique phonetic and syntactic features and the scarcity of extensive labelled datasets pose considerable challenges that must be addressed to improve diarization accuracy and reliability in these contexts.

Our findings underscore the necessity of developing tailored methodologies that can effectively handle the linguistic variability and limited resources inherent to low-resourced languages. Leveraging advancements in machine learning, particularly the robust and discriminative capabilities of x-vectors and DNNs, offers a promising path forward. However, further research is needed to adapt these techniques to the specific needs of languages like Sarawak Malay, ensuring they can capture the unique speaker characteristics and achieve high performance despite the constraints.

This investigation has identified several critical areas for future research. Future work should focus on strategies to annotate the unlabelled Sarawak Malay conversation data. We could work with speakers to annotate the data (which could be costly) with speaker labels or turns or adopt machine learning models such as x-vectors and DNNs as a starting point for automatically labelling the turns (mislabelled data could be high). Both approaches have their advantages and disadvantages. Nevertheless, infusing knowledge from speakers into machine learning models through transfer learning or fine-tuning can be advantageous during model training. We can see this opportunity through Pyannote’s pre-trained x-vectors that are ready to be used to diarize speech or used for transfer learning for customising speaker diarization models to target language. For instance, transfer learning can be conducted by implementing pseudo-labels of existing raw Sarawak Malay conversational speech audio files generated via diarization by Pyannote’s pre-trained x-vectors speaker diarization model. These pseudo-labels are to be used further to train the Pyannote’s pre-trained x-vectors speaker diarization model, optimizing this model to be more catered towards the Sarawak Malay dialect via transfer learning.

Lastly, investigating multilingual and cross-lingual speaker diarization approaches could provide valuable insights in this area of research. By leveraging data from multiple languages and exploring the transferability of diarization models across linguistically similar languages, researchers can enhance the robustness and generalizability of their techniques.

ACKNOWLEDGMENT AND FUNDING

The authors wish to thank Universiti Malaysia Sarawak and Yayasan Sarawak for the financial support of this research.

DECLARATION OF CONFLICTING INTERESTS

The authors declare no potential conflicts of interest with respect to the research and publication of this article.

REFERENCES

[1] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, Speaker adaptation of neural network acoustic models using i-vectors, *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, 55-59.

[2] Y. Miao, H. Zhang and F. Metze, Speaker adaptive training of deep neural network acoustic models using i-vectors, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23, 2015, 1938-1949.

[3] L. Sarı, N. Moritz, T. Hori and J. Le Roux, Unsupervised speaker adaptation using attention-based speaker memory for end-to-end ASR, *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 7384-7388.

[4] W. H. Kang and N. S. Kim, Unsupervised learning of total variability embedding for speaker with random digit strings, *Applied Sciences*, 9, 2019, 1-16.

[5] A. Nagrani, J. S. Chung, W. Xie and A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, *Computer Speech and Language*, 60, 2020, 101027.

[6] D. V. Thanh, T. P. Viet and T. N. T. Thu, Deep speaker verification model for low-resource languages and Vietnamese dataset, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, 2021, 442-451.

[7] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, Deep neural network embeddings for text-independent speaker verification, *International Speech Communication Association (INTERSPEECH)*, 2017, 999-1003.

[8] S. Baghel, S. Ramoji, S. Sidharth, H. Ranjana, P. Singh, S. Jain, P. R. Chowdhuri, K. Kulkarni, S. Padhi, D. Vijayasenan, and S. Ganapathy, *DISPLACE Challenge: Diarization of Speaker and Language in Conversational Environments*, *ArXiv.Org*, 2023.

[9] Z. Wang and J. H. L. Hansen, Multi-source domain adaptation for text-independent forensic speaker recognition, *IEEE/ACM Transactions on Audio Speech and Language Processing*, 30, 2022, 60-75.

[10] H. Gish, M. H. Siu and R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2, 1991, 873-876.

[11] S. S. Chen and P. S. Gobalakrishnan, Speaker, environment and channel change detection and clustering via the bayesian information criterion, *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, 127-132.

[12] U. Jain, M. A. Siegler, S. -J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj and R. M. Stern, Recognition of continuous broadcast news with multiple unknown speakers and environments, *DARPA Speech Recognition Workshop*, 1996, 61.

[13] W. W. Lin, M. W. Mak and J. T. Chien, Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders, *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26, 2018, 2412-2422.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, X-Vectors: robust DNN embeddings for speaker recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, 5329-5333.

[15] P. Cabañas-Molero, M. Lucena, J. M. Fuertes, P. Vera-Candeas and N. Ruiz-Reyes, Multimodal speaker diarization for meetings using volume-evaluated srp-phat and video analysis, *Multimedia Tools and Applications*, 77, 2018, 27685-27707.

[16] K. Akesbi and S. Gandhi, Diarizers: a repository for fine-tuning speaker diarization models, https://github.com/huggingface/diarizers, 2024 (accessed 23.10.2024).

[17] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe and S. Narayanan, A review of speaker diarization: recent advances with deep learning, *Computer Speech and Language*, 72, 2022, 101317.

[18] Z. Jin, Y. Yang, M. Shi, W. Kang, X. Yang, Z. Yao, F. Kuang, L. Guo, L. Meng, L. Lin, Y. Xu, S.-X. Zhang and D. Povey, LibriheavyMix: a 20,000-hour dataset for single-channel reverberant multi-talker speech separation, ASR and speaker diarization, *International Speech Communication Association (INTERSPEECH)*, 2024, 702-706.

[19] V. Khoma, Y. Khoma, V. Brydinskyi and A. Konovalov, Development of supervised speaker diarization system based on the pyannote audio processing library, *Sensors*, 23, 2023, 2082.

[20] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, Automatic speech recognition for under-resourced languages: a survey, *Speech Communication*, 56, 2014, 85-100.

[21] S. S. Juan, *Exploiting Resources from Closely-Related Languages for Automatic Speech Recognition in Low-Resource Languages from Malaysia*, Ph.D. dissertation, Grenoble-Alpes University, France, 2015.

[22] I. Aman and R. Mustaffa, Social variation of malay language in Kuching, Sarawak, Malaysia: a study on accent, identity and integration, *GEMA Online Journal of Language Studies*, 9, 2009, 63-76.

[23] J. T. Collins, The study of sarawak malay in context, In *Between Worlds: Linguistic Papers in Memory of David John Prentice*, Australia: Pacific Linguistics, 2002, 65-75.

[24] D. M. Eberhard, G. F. Simons and C. D. Fennig, *Ethnologue: Languages of the World.*, Twenty-seventh Online version, Dallas, Texas: SIL International, 2024.

[25] L. Shamseer, D. Moher, M. Clarke, et al., Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation, *BMJ (Clinical Research Ed.)*, 350, 2015, 1-25.

[26] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri and M. Almojil, Automatic speech recognition: systematic literature review, *IEEE Access*, 9, 2021, 131858-131876.

[27] C. Deka, A. Shrivastava, A. K. Abraham, S. Nautiyal and P. Chauhan, AI-based automated speech therapy tools for persons with speech sound disorder: a systematic literature review, *Speech, Language and Hearing*, 2024, 1-22.

[28] M. S. Jahan and M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, *Neurocomputing*, 546, 2023, 126232.

[29] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, A survey of speaker recognition: fundamental theories, recognition methods and opportunities, *IEEE Access*, 9, 2021, 79236-79263.

[30] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, Speaker diarization: a review of recent research, *IEEE Transactions on Audio, Speech and Language Processing*, 20, 2012, 356-370.

[31] V. Sethuram, A. Prasad and R. R. Rao, Optimal trained artificial neural network for telugu speaker diarization, *Evolutionary Intelligence*, 13, 2020, 631-648.

[32] A. Q. Ohi, M. F. Mridha, M. A. Hamid and M. M. Monowar, Deep speaker recognition: process, progress, and challenges, *IEEE Access*, 9, 2021, 89619-89643.

[33] S. Meignier and T. Merlin, LIUM SPKDIARIZATION: An open source toolkit for diarization, *CMU SPUD Workshop,* 2020, 1-6.

[34] S. Kanwal, K. Malik, K. Shahzad and Z. A. F. and Nawaz, Urdu named entity recognition: corpus generation and deep learning, *ACM Transactions on Asian and Low-Resource Language Information*, 19, 2020, 1-13.

[35] A. Jati and P. Georgiou, Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics, *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27, 2019, 1577-1589.

[36] J. Karadayi, C. Scaff, J. Stieglitz and A. Cristia, Diarization in maximally ecological recordings: data from tsimane children, *6th Workshop on Spoken Language Technologies for Under-Resourced Languages* (*SLTU 2018*), 2018, 30-35.

[37] H. Taherian, Z. -Q. Wang and W. DeLiang, Deep learning based multi-channel speaker recognition in noisy and reverberant environments, *Proceedings of the Annual Conference of the International Speech Communication Association* (*INTERSPEECH*), 2019, 4070-4074.

[38] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, Speaker diarization: a review of recent research, *IEEE Transactions on Audio, Speech and Language Processing*, 20, 2012, 356-370.

[39] E. Alsharhan and A. Ramsay, Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition, *Language Resources And Evaluation*, 54, 2020, 975-998.

[40] Q. Li, F. L. Kreyssig, C. Zhang and P. C. Woodland, Discriminative neural clustering for speaker diarisation, *IEEE Spoken Language Technology Workshop* (*SLT 2021*), 2021, 574-581.

[41] M. K. Nammous, K. Saeed and P. Kobojek, Using a small amount of text-independent speech data for a bilstm-scale speaker identification approach, *Journal of King Saud University - Computer and Information Sciences*, 34, 2022, 764-770.

[42] A. Mane, J. Bhopale, R. Motghare and P. Chimurkar, An overview of speaker recognition and implementation of speaker diarization with transcription, *International Journal of Computer Applications*, 175, 2020, 1-6.

[43] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu and S. Watanabe, Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2019)*, 2019, 31-38.

[44] Z. Bai and X. -L. Zhang, Speaker recognition based on deep learning: an overview, *Neural Networks*, 140, 2021, 65-99.

[45] G. A. Levow, Investigating speaker diarization of endangered language data, *COMPUTEL 2023 - 6th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2023, 38-44.

[46] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, Di. Fustes, H. Titeux, W. Bouaziz and M.-P. Gill, Pyannote.Audio: neural building blocks for speaker diarization, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, 7124-7128.

[47] J. Karadayi, C. Scaff, J. Stieglitz and A. Cristia, Diarization in maximally ecological recordings: data from tsimane children, *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 2018, 30-35.

[48] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar and I. Ali, Text-independent speaker identification through feature fusion and deep neural network, *IEEE Access*, 8, 2020, 32187-32202.

[49] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, Speech recognition using deep neural networks: a systematic review, *IEEE Access*, 7, 2019, 19143-19165.

[50] T. J. Park, K. J. Han, J. Huang, X. He, B. Zhou, P. Georgiou and S. Narayanan, Speaker diarization with lexical information, *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, 391-395.

[51] M. Diez, L. Burget, S. Wang, J. Rohdin and H. Černocký, Bayesian HMM based x-vector clustering for speaker diarization, *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, 346-350.

[52] F. Landini, S. Wang, M. Diez, L. Burget, P. Matejka, K. Zmolikova, L. Mosner, A. Silnova, O. Plchot, O. Novotny, H. Zeinali and J. Rohdin, But system for the second dihard speech diarization challenge, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, 6529-6533.

[53] A. A. Joshy and R. Rajan, Automated dysarthria severity classification: a study on acoustic and deep learning techniques, *IEEE Transactions on Neural Systems And Rehabilitation Engineering*, 30, 2022, 1147-1157.

[54] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy and M. Liberman, The second dihard diarization challenge: dataset, task, and baselines, *Interspeech*, 2019, 978-982.

[55] M. Diez, L. Burget and P. Matějka, Speaker diarization based on bayesian hmm with eigenvoice priors, *Speaker and Language Recognition Workshop (ODYSSEY 2018)*, 2018, 147-154.

[56] E. Prud'hommeaux, R. Jimerson, R. Hatcher and K. Michelson, Automatic speech recognition for supporting endangered language, *Language Documentation & Conservation*, 15, 2021, 491-513.