Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
e-ISSN: 2149-3367
DOI: https://doi.org/10.35414/akufemubid.1565447
AKÜ FEMÜBİD **25** (2025) 055201 (1095-1105)
AKU J. Sci. Eng. **25** (2025) 055201 (1095-1105)

A Comparative Analysis of Traditional and Deep Learning Approaches for Adressing Challenges in Speaker Diarization

Makale Bilgisi / Article Info
Alındı/Received: 11.10.2024
Kabul/Accepted: 19.04.2025
Yayımlandı/Published: 01.10.2025

Konuşmacı Diarizasyonundaki Zorlukların Çözümünde Geleneksel Yöntemler ile Derin Öğrenme Yöntemlerinin Karşılaştırılması

Emsal ALTINAY*, Ecir Ugur KUCUKSILLE

Süleyman Demirel Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Isparta, Türkiye

Abstract

Speaker diarization is the task of distinguishing and segmenting speech from multiple speakers in an audio recording, a critical component for various applications such as meeting transcription, voice activated systems, and audio indexing. Traditional clustering-based methods have been widely adopted, but face challenges in real-world scenarios, including noisy environments, overlapping speech, speaker variability and variable recording conditions. This research addresses these limitations by examining deep learning-based approaches, which have demonstrated notable advancements in enhancing multi-speaker diarization accuracy. This study provides a comprehensive comparison between traditional clustering algorithms and contemporary deep learning approaches, including Time-Delay Neural Networks (TDNN), End-to-End Neural Diarization (EEND), and Fully Supervised UIS-RNN. By evaluating their performance on the CallHome dataset, the study highlights the limitations of traditional methods and the significant advancements offered by deep learning techniques. Results show that TDNN achieves slight improvements in non-overlapping speech, EEND demonstrates notable performance gains in overlapping speech scenarios, achieving a Diarization Error Rate (DER) of 12.6% compared to 23.7% for traditional methods. The UIS-RNN model outperforms all other techniques with a DER of 7.6%, showcasing its efficacy in handling complex acoustic conditions. This study underscores the transformative potential of deep learning in addressing the multifaceted challenges of speaker diarization and lays the foundation for future research in this evolving domain.

Keywords: *Speaker Diarization; Traditional Clustering Algorithm; Deep Learning; Overlapping Speech; Computational Complexity*

Öz

Konuşmacı günlüğü oluşturma, bir ses kaydında birden fazla konuşmacıdan gelen konuşmayı ayırt etme ve bölümlere ayırma görevidir ve toplantı transkripsiyonu, sesle etkinleştirilen sistemler ve ses indeksleme gibi çeşitli uygulamalar için çok kritik bir bileşendir. Geleneksel kümeleme tabanlı yöntemler yaygın olarak kullanılmaktadır, ancak gürültülü ortamlar, örtüşen konuşma, konuşmacı değişkenliği ve değişken kayıt koşulları gibi gerçek dünya senaryolarındaki zorluklarla karşılaşmaktadır. Bu araştırma, birden fazla konuşmacının bulunduğu senaryolarda diarizasyon doğruluğunu artırmada dikkate değer gelişmeler gösteren derin öğrenme tabanlı yaklaşımlara odaklanarak bu sınırlamaları ele almaktadır. Bu çalışma, geleneksel kümeleme algoritmaları ile Zaman Gecikmeli Sinir Ağları (TDNN), Uçtan Uca Sinirsel Diarizasyon (EEND) ve Tam Denetimli UIS-RNN gibi çağdaş derin öğrenme yaklaşımlarını kapsamlı bir şekilde karşılaştırmaktadır. Performanslarını CallHome veri seti üzerinde değerlendirerek, geleneksel yöntemlerin sınırlamalarını ve derin öğrenme tekniklerinin sağladığı önemli ilerlemeleri vurgulamaktadır. Sonuçlar, TDNN'nin örtüşmeyen konuşma durumlarında hafif iyileştirmeler sağladığını, EEND'in örtüşen konuşma senaryolarında dikkate değer performans kazançları göstererek, geleneksel yöntemlerin %23,7'lik oranına kıyasla %12,6'lık bir Diarizasyon Hata Oranı (DER) elde ettiğini ortaya koymaktadır. Tam Denetimli UIS-RNN modeli, %7,6'lık bir DER ile tüm teknikler arasında en iyi performansı sergileyerek karmaşık akustik koşulları yönetmedeki etkinliğini göstermektedir. Bu çalışma, konuşmacı diarizasyonunun çok yönlü zorluklarını ele almadaki derin öğrenmenin dönüştürücü potansiyelini vurgulamakta ve bu gelişen alandaki gelecekteki araştırmalar için bir temel oluşturmaktadır.

Anahtar Kelimeler: *Konuşmacı Diarizasyonu; Geleneksel Kümeleme Algoritması; Derin Öğrenme; Örtüşen Konuşma; Hesaplama Karmaşıklığı*

1. Introduction

Speaker Diarization (SD) refers to the process of segmenting a multi-speaker speech into homogeneous segments, allowing the identification and differentiation of speakers. This task is pivotal in many applications, such as meeting transcription, media playback, voice-activated systems, and audio indexing (Fiscus et al, 2006). By attributing speech to individual speakers, SD enhances speaker-specific voice search, improves the accuracy of automatic speech recognition (ASR), and facilitates transcript readability. Despite its growing importance, speaker diarization remains a challenging problem, especially in real-world scenarios characterized by

*Sorumlu Yazar/Corresponding Author: Emsal Altinay e-posta/e-mail: emsalaltinay@isparta.edu.tr

overlapping speech, speaker variability, and variable recording conditions (Raj, 2021).

The evolution of speaker diarization methodologies has witnessed significant transitions, from traditional statistical approaches to contemporary deep learning solutions. Traditional speaker diarization techniques have long been the foundation of this field. These methods often rely on statistical modeling approaches, such as Gaussian Mixture Models (GMMs) combined with Hidden Markov Models (HMMs). These models operate on manually crafted features and predefined statistical assumptions, which makes them effective in controlled environments (Anguera et al., 2012). Additionally, clustering-based techniques, including spectral clustering and agglomerative hierarchical clustering (AHC), have been widely adopted for segmenting speech. The incorporation of i-vector embeddings further refined the performance of these methods by offering a compact representation of speaker characteristics. However, these traditional approaches exhibit substantial limitations when confronting real-world scenarios characterized by overlapping speech, speaker variability, and inconsistent recording conditions (Park et al., 2022). To overcome these challenges, x-vector embeddings introduced by TDNNs marked a significant improvement in speaker representation (Snyder et al., 2018). Unlike previous handcrafted features, x-vectors leverage neural network architectures to extract robust speaker embeddings, improving the ability of clustering algorithms to accurately separate speakers. Nonetheless, these advancements remain insufficient when applied to overlapping speech scenarios or environments with high speaker variability.

Recent years have witnessed a paradigm shift in speaker diarization with the advent of deep learning-based methods. These models automate the feature extraction process, enabling a more nuanced understanding of speaker characteristics. EEND exemplifies this innovation by integrating feature extraction, speaker classification, and diarization into a unified framework. By utilizing mechanisms like self-attention and bidirectional long short-term memory (BLSTM) networks, EEND effectively addresses overlapping speech and achieves significant reductions in DER (Fujita et al., 2019a). Similarly, the Fully Supervised UIS-RNN model represents another milestone in the field. This model employs a recurrent neural network (RNN)-based architecture combined with supervised learning techniques to dynamically manage speaker transitions. The UIS-RNN leverages high-quality speaker annotations to learn speaker-specific characteristics, achieving state-of-the-art DER performance in benchmark datasets (Zhang et al., 2019). Unlike traditional methods that rely heavily on clustering, UIS-RNN directly models speaker sequences, making it particularly effective for scenarios involving overlapping speech and dynamic speaker changes.

The transition from traditional methodologies to deep learning frameworks in speaker diarization presents significant implementation challenges. While these advanced models demonstrate superior feature extraction capabilities and complex pattern recognition, they necessitate substantial computational infrastructure and comprehensive annotated datasets. Moreover, these systems exhibit limitations in out-of-domain scenarios, particularly when confronted with data sparsity and acoustic interference (Fujita et al., 2019b; Hamza et al., 2023). Such constraints emphasize the imperative for continued methodological advancement to enhance the accessibility, interpretability, and environmental adaptability of deep learning-based diarization systems, especially in scenarios involving multiple speakers and overlapping speech patterns.

This study aims to provide a comprehensive comparison between traditional methods and deep learning models in addressing the multifaceted challenges of speaker diarization, including speaker variability, overlapping speech, computational complexity, scalability, and adaptability in real-world scenarios. Through this comparative analysis, this research contributes to a broader discourse on the evolution of speaker diarisation techniques, specifically by highlighting the transformation from heuristic-driven frameworks to data-driven learning models. It provides a foundation for future research in this dynamic and evolving field.

The remainder of this paper is organized as follows: Section 2 reviews related work, examining both traditional and deep learning-based diarization methods. Section 3 details recent advancements in SD techniques using deep learning, with dataset specifications and evaluation metric. Section 4 discusses the results and provides an in-depth analysis of the performance of different approaches. Finally, Section 5 concludes with insights into future research directions aimed at addressing the current challenges in speaker diarization.

2. Related Work

This literature review explores the evolution of methodologies and techniques in multi-speaker diarization, with an emphasis on the transition from traditional statistical approaches to deep learning-based solutions. The developments in this domain reflect ongoing efforts to address critical challenges such as overlapping speech, speaker variability, and scalability in real-world applications.

Traditional speaker diarization techniques have primarily relied on statistical approaches and handcrafted features. Early methods, such as Gaussian Mixture Models (GMMs) paired with Hidden Markov Models (HMMs), were effective in controlled acoustic environments but struggled with overlapping speech and dynamic variability in speaker characteristics (Reynolds et al., 2000; Anguera et al., 2012). Other traditional methods include spectral clustering and agglomerative hierarchical clustering, which rely on distance metrics to separate speech

segments based on their spectral properties. These methods achieved moderate success when integrated with i-vector embeddings, which provided compact statistical representations of speaker characteristics. While i-vectors improved scalability, their performance was constrained in settings involving overlapping speech and noisy environments (Park et al., 2022).

Deep learning models has revolutionized speaker diarisation by automating feature extraction and learning directly from data. A significant milestone in this transition was the introduction of DNN-based features, known as D-vectors, which demonstrated superior performance over traditional i-vectors, particularly in noisy conditions (Variani et al., 2014). Subsequent work has continued to develop the new paradigm of using DNNs to extract speaker identity discriminative features. In particular, the invention of the x-vector extractor has led to a major improvement. This extractor uses a TDNN and a statistical pooling layer to obtain low-dimensional speaker identity representations (Snyder et al., 2018). This innovation enabled clustering algorithms to more effectively distinguish between speakers, even in challenging environments with moderate variability. Other studies have focused on refining training strategies and loss functions to improve the discriminative power of speaker embeddings. Techniques such as metric learning and angular-softmax loss were proposed to enhance the robustness of deep learning models in capturing speaker-specific features (Bredin, 2017; Chung et al., 2020; Li et al., 2018). Despite these improvements, the predominant reliance on clustering in these systems continued to limit their efficacy in scenarios involving overlapping speech or high speaker variability.

Speech diarization technology has evolved over time with the development of automatic speech recognition (ASR) technologies. Tools like Amazon Transcribe illustrate how diarization can enhance transcription tasks such as call analytics, medical documentation, and media subtitling. However, traditional frameworks often treat ASR and speaker diarization as separate tasks, leading to inefficiencies and suboptimal performance in real-time applications. To address these challenges, researchers have proposed joint ASR and speaker diarization models. For example, Mao et al. (2020) demonstrated that merging these tasks improves performance by exploiting audio-verbal dependencies. This approach proved particularly effective when utterance boundaries were unknown, with attention-based decoding algorithms and data augmentation techniques further enhancing accuracy. These integrated models represent a shift towards holistic frameworks that streamline diarization and transcription processes.

Specifically, a method has been developed to manage speaker identity discriminative features from overlapping speech regions in the clustering phase (Raj et al., 2021b). Methods such as the VBx approach introduced overlap detection mechanisms to mask speaker posterior matrices, improving clustering accuracy in overlapping regions (Bullock et al., 2020). The effectiveness of these models is commonly evaluated using standardized metrics such as DER and character error rate (CER Recent advancements underscore the potential of end-to-end systems in addressing the complexities of real-world scenarios. For instance, a novel system designed for an in-vehicle multi-channel ASR competition achieved a 49.58% reduction in DER compared to baseline models, highlighting the effectiveness of integrated frameworks in challenging acoustic environments (Tian et al., 2024).

However, significant challenges persist in accurately segmenting speech and attributing it to the correct speakers, particularly in scenarios characterized by noisy or overlapping speech. Both approaches exhibit unique advantages and limitations, in addressing critical issues such as speaker variability, overlapping speech, computational complexity, scalability and adaptability to diverse real-world conditions. These challenges necessitate continued methodological advancements to enhance the robustness and applicability of speaker diarization systems.

3. Materials and Methods

Deep learning models can be computationally intensive, requiring significant resources for training and inference. This complexity may present a significant obstacle to the practical deployment of diarization systems, particularly in real-time applications. There are no standardized benchmarks and evaluation metrics in this area, leading to inconsistencies in performance evaluation between different studies. Establishing common data sets and metrics will facilitate better comparisons and progress in this field (Kshirod, 2020).

Advanced speech recognition and speaker separation technologies are leveraged to transcribe a multi-speaker audio file separately for each speaker. In order to perform this process, the following operations are required:

- Pre-processing of the audio file,
- Cleaning up the sound by reducing background noise,
- Normalization to ensure consistent sound levels throughout the recording.

3.1 Dataset

The CallHome dataset is as a standard for the evaluation of speaker diarization models. The dataset comprises multilingual telephone conversations gathered from natural everyday settings. The English subset is the most widely used (Canavan et al., 1997). The dataset is suitable for testing scenarios with overlapping speech or varying acoustic conditions, as it contains multiple speakers. Each recording is manually labelled with speaker labels and timestamps, providing a ground truth reference for assessing logging accuracy.

The CallHome dataset comprises approximately 120 audio recordings, with individual file durations ranging from 2 to 15 minutes. These recordings consist of telephone conversations collected in naturalistic settings, incorporating realistic acoustic challenges including background noise, varying speech dynamics, and overlapping speaker transitions. The dataset's key characteristics include:

- Total Duration: In excess of 100 hours of annotated audio content
- Speaker Population: Approximately 500 distinct speakers
- Demographic Distribution: Balanced representation with approximately 55% male and 45% female speakers
- Linguistic Diversity: Multilingual recordings, with the English subset serving as the primary corpus for analysis
- Acoustic Variability: Integration of both clean and noisy recordings, including ambient sounds and background conversations, ensuring evaluation under realistic and challenging conditions.

The diverse characteristics of the CallHome dataset provide a comprehensive evaluation framework for assessing the efficacy of traditional clustering methods and advanced deep learning approaches. The dataset's incorporation of realistic acoustic conditions challenge models to adapt to overlapping speech, speaker variability, and varying recording environments. These attributes make it particularly valuable for conducting comparative analyses between different diarization approaches.

3.2. Diarization Error Rate (DER)

DER is the primary metric for assessing both traditional and deep learning-based diarization methods. By analyzing DER values across overlapping and non-overlapping speech scenarios, the study compares the robustness and accuracy of approaches like End-to-End Neural Diarization (EEND) and Fully Supervised UIS-RNN with traditional clustering-based methods. A lower DER indicates better performance. For example, traditional clustering methods generally exhibit higher DER, particularly in overlapping speech conditions, due to their limited ability to handle speaker ambiguity. In contrast, advanced models like EEND achieve significantly lower DER by integrating neural architectures specifically designed to address overlap and speaker variability.

It quantifies the system's accuracy in attributing speech segments to the correct speakers within an audio recording. DER is a composite error metric, expressed as a percentage, and is calculated as in Formula 1.

$$DER = \frac{\text{Missed Speech+False Alarm Speech+Speaker Confusion}}{\text{Total Speech Duration}} \times 100 \quad (1)$$

where:

- Missed Speech: Segments of speech that the system fails to identify as spoken.
- False Alarm Speech: Non-speech segments or noise incorrectly identified as speech.
- Speaker Confusion: Segments of speech attributed to the wrong speaker.

3.3 Clustering Based Algorithms

Clustering-based methods have long been the backbone of speaker diarization. Segments belonging to the same speaker are grouped together and each speaker is assigned a unique identity. While effective, traditional clustering methods often struggle with overlapping speech and rapid speaker changes. GMM, mean shift, hierarchical, k-means and spectral algorithms are most common clustering approaches for speaker diaries. Recent research has demonstrated that deep learning methodologies have significantly transformed the field of multi-speaker diarization, offering substantial improvements in performance and accuracy compared to traditional approaches.

3.4 Deep Learning Models: The New Frontier

3.4.1 EEND: End-to-End Neural Diarization

The End-to-End Neural Diarization (EEND) model leverages deep learning to process the audio input directly and generate speaker labels requiring traditional pre-processing steps. This method can accommodate varying numbers of speakers and is particularly effective in complex acoustic environments (Al-Hadithy et al., 2022). EEND-vector clustering represents a groundbreaking methodology that integrates neural networks and clustering-based diarization techniques into a unified framework. By simultaneously learning speaker separation and classification from audio input,

EEND-based methods have achieved competitive performance across multiple speaker diarization benchmarks. Combining deep learning-based techniques with clustering algorithms, EEND facilitates efficient diarization and management of overlapping speech in extended recordings (Fiscus, J., et al., 2006).

EEND was initially proposed by Fujita. This initial version integrated voice diarization methods using a bidirectional long short-term memory (BLSTM) network. Subsequently, EEND has been extended over time with networks based on the self-attention mechanism. This extension has demonstrated state-of-the-art diarisation error rate (DER) results on two-speaker data. In particular, the methodology yielded successful results on two speaker samples from the Spontaneous Japanese Corpus and the CALLHOME dataset (Fujita, Y., et al., 2019a). The EEND's architecture consists of four self-attention encoder blocks, each containing 256 attention units with a dropout rate of 0.1 for regularization. The model processes 40-dimensional log-mel filterbank features extracted from audio segments of 25ms with a 10ms shift. Its training protocol includes data augmentation techniques such as speed perturbation (0.9x, 1.0x, 1.1x) and artificial creation of overlapping segments. The training utilizes an Adam optimizer with a learning rate of 1e-4, processing batches of 64 sequences over 100 epochs, with binary cross-entropy loss function for speaker activity detection.

EEND offers several distinct advantages. First, it effectively handles overlapping utterances, a critical challenge in speaker diarization. Second, the network's design prioritizes diarization accuracy, enabling high performance. Thirdly, the system can be retrained using real data, simply by incorporating a reference diarization label. Nevertheless, EEND is not without its limitations. The model architecture restricts the maximum number of speakers it can support, and its reliance on self-attention-based neural networks or BLSTM presents challenges for online processing. Additionally, experimental results indicate that EEND tends to overfit the training data distribution, limiting its generalizability in out-of-domain scenarios (Fujita et al., 2019b).

In conclusion, the EEND algorithm has emerged as a significant innovation in the field of speaker diarization, evolving over time to become increasingly effective and efficient. This evolution has been driven by the advances of deep learning techniques and neural networks in the field of audio processing.

In recent years, EEND has solidified its status as a powerful alternative to traditional clustering-based speaker diarisation methods, particularly for managing overlapping speech and simplifying the diarization pipeline. The EEND approach consistently demonstrates superior performance compared to traditional clustering-based methods, especially in scenarios involving overlapping speech and complex acoustic environments.

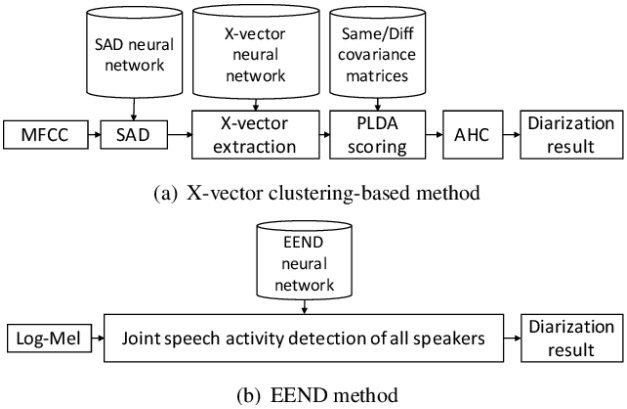

Figure 1. *Comparison of X-Vector Clustering-Based and EEND Methods.* (Fujita, Y., et al., 2019a).

3.4.2 Time-Delayed Neural Network (TDNN)

Time-delayed neural network (TDNN) constitute a specialized artificial neural network architecture designed specifically for sequential data processing, particularly in speech signal analysis. The feature of a TDNN is its ability to model temporal relationships between inputs using time-delayed connections. This makes it particularly useful in situations where the temporal sequence of inputs is of significance. In a TDNN, the input of each neuron can contain not only the data of the current time step, but also data from previous time steps. The TDNN employs weight sharing mechanisms similar to Convolutional Neural Networks (CNNs) to reduce parameter complexity while maintaining temporal pattern detection capabilities. The TDNN architecture consists of five sequential layers, with the input layer processing 40-dimensional MFCC features through a context window of 15 frames. The subsequent layers implement specific temporal contexts: the initial layer employs 512 units with context {-2,-1,0,1,2}, followed by three layers maintaining 512 units with varying temporal spans (-2,0,2; -3,0,3; -4,0,4), and culminating in a 1500-unit layer with context {0}. This hierarchical structure incorporates a statistical pooling layer for mean and standard deviation computation, followed by two 512-unit fully connected layers, generating 512-dimensional x-vector embeddings crucial for speaker discrimination in multi-speaker environments. The network's training utilizes a natural gradient stochastic gradient descent optimizer, implementing an initial learning rate of 0.001 with exponential decay, processing 64-sample batches

across three epochs under L2 regularization (weight: 0.001).

A pre-trained TDNN model processes audio frames to extract x-vectors, which serve as fixed-length embeddings representing speaker features in each segment. The model can use several layers with time-delayed connections to capture temporal information. A similarity matrix is calculated using the extracted x-vectors, where each element represents the similarity between two segments. A clustering algorithm, such as Agglomerative Hierarchical Clustering (AHC) or Spectral Clustering, is employed to group similar x-vectors together. Each cluster is associated with a distinct speaker. The system assigns a unique speaker identity to each segment. This facilitates the identification of instances where one speaker stops speaking and another begins(Snyder, D., 2018). While TDNNs are effective in generating robust speaker embeddings, they fail to address the challenge of overlapping speech, which persists as a limitation during the clustering phase.

3.4.3 Fully Supervised UIS-RNN Model

The UIS-RNN represents a significant advancement in fully supervised speaker diarization, offering a sophisticated framework for temporal data partitioning and clustering through sample-based learning. Operating within a fully supervised paradigm, the model leverages high-quality, time-stamped speaker annotations to establish precise relationships between acoustic features and speaker identities, resulting in demonstrably superior performance compared to traditional unsupervised approaches. This supervised methodology enables direct learning of speaker-specific characteristics, facilitating more effective pattern recognition and temporal segmentation.

The architectural implementation of UIS-RNN comprises three LSTM layers with 256 hidden units, incorporating 256-dimensional speaker embedding vectors for robust speaker representation. The model implements a distance-dependent Chinese restaurant process for modeling speaker transitions, enabling dynamic speaker modeling without requiring prior knowledge of the number of speakers. The training protocol utilizes the Adam optimizer with a learning rate of 1e-3, processing data across 50 epochs with 32-sample batches. Speaker transitions are managed through a 0.5 probability threshold, while speaker similarity assessments employ cosine similarity metrics for precise discrimination.

The system achieved impressive results on the NIST SRE 2000 CALLHOME dataset, outperforming state-of-the-art methods with a diarization error rate of 7.6% (Zhang, A., et al., 2019). Parameter sharing across RNNs enhances the model's ability to generalize to new speakers while improving computational efficiency. This architectural approach offers significant advantages, particularly in terms of dynamically modeling speaker transitions and effectively handling overlapping speech scenarios. The model's implementation of unbounded interleaved states allows it to adapt to varying numbers of speakers, making it particularly effective for real-world applications where the number of speakers is not known in advance.

The diarization system utilises the same segmentation module and embedding extraction module as in another study, but replaces the clustering module with an infinite interval state RNN (Wang, Q., et al., 2018). This modification enables real-time processing capabilities while maintaining the quality standards typically associated with offline methods. The Figure 2 illustrates the generative process of the UIS-RNN model. In the diagram, x_t represents the input sequences provided to the model at each time step, while h_t denotes the hidden states, which capture the temporal dependencies learned from the input sequences. m_t indicates the intermediate states or transitions of the model, determining which state is active at a given time step. y_t represents the predicted labels at each time step, with different labels (y_t =1,2,3,4) illustrated using distinct colors. The lines and arrows depict the flow of information between inputs and hidden states, as well as the transitions of hidden states over time. Solid lines represent deterministic connections, whereas dashed lines indicate probabilistic dependencies.

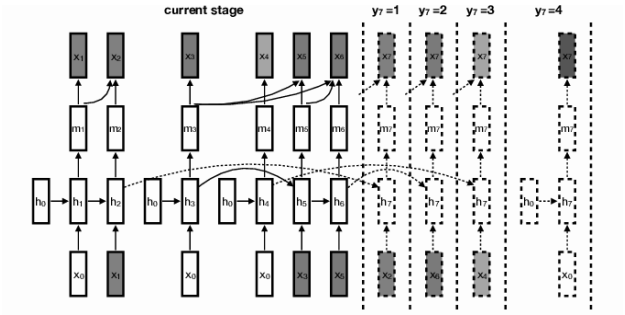

Figure 2. *The UIS-RNN generation process. Colours indicate labels for speaker segments (Zhang, A., et al., 2019).*

These systems leverage detailed temporal labels during the training phase, enabling the extraction of speaker-specific characteristics from unambiguous exemplars. This approach demonstrates particular efficacy in domains where annotated training data is readily accessible, facilitating the establishment of robust speaker identification frameworks. Empirical evidence indicates that supervised approaches achieve

substantially lower error rates in terms of data dispersion compared to their unsupervised counterparts, primarily due to their capacity to establish direct correlations between acoustic features and speaker identities. The architectural sophistication of fully supervised systems enables automatic determination of speaker cardinality while implementing advanced modeling techniques, specifically recurrent neural networks (RNNs), to address the inherent complexities of speaker diarization tasks. This integration of supervised learning paradigms with sophisticated neural architectures facilitates more precise speaker discrimination and temporal segmentation, establishing a robust framework for complex speaker diarization scenarios.

3.5 Experimental Setup

The experimental anlayses were conducted on a computer system equipped with an 11th Gen Intel(R) Core(TM) i7-1165G7 2.80GHz processor and 16GB RAM. The software environment and tools were configured as follows:

The system operated on Windows 10 Pro, with Python 3.8 serving as the primary programming language. PyTorch framework was employed for the implementation of deep learning models. Audio data processing and feature extraction were performed using the Librosa library. For data preprocessing and analysis phases, NumPy, Pandas and SciPy libraries.

The model training and evaluation processes leveraged GPU support provided by the Google Colab Pro platform. The training duration varied across models, with EEND requiring approximately 36 hours, UIS-RNN requiring 18 hours, and TDNN-based systems completing in 8 hours. Performance evaluations and comparative analyses were conducted on the local computer system.

Audio file preprocessing and segmentation were accomplished using PyDub and Wave libraries, while custom Python scripts were developed for calculating evaluation metrics. All code development processes were carried out in the Visual Studio Code integrated development environment.

This setup enabled efficient implementation and evaluation of the speaker diarization models while maintaining reproducibility of results. This detailed specification of the experimental infrastructure enhances the reproducibility of our research and provides readers with a thorough understanding of the computational resources utilised in this study.

4. Results and Discussions

In this investigation, a comprehensive comparative analysis was conducted between traditional clustering methods and new deep learning models on the CallHome dataset. DER and key challenges metrics were used to compare the traditional and deep learning models for speaker diarization. The data presented in Tables 1, 2, 3 and 4 demonstrate the performance values for each deep learning model in comparison to traditional methods. Table 5 illustrates the comparative summary of all the discussed techniques on challenges of speaker diarization.

In comparison, EEND significantly reduces the DER, particularly in overlapping speech where it achieved 14.6%. The overall DER for EEND on CallHome is 12.6%, representing a substantial improvement over the 23.7% DER observed with traditional clustering methods. This enhancement in performance can be attributed to EEND's end-to-end architecture and sophisticated attention mechanisms, enabling more effective speaker separation in complex acoustic environments.

Table 1. Performance of traditional clustering-based methods and EEND on the CallHome dataset.

Method	DER (Oveall)	DER (Overlapping Speech)	DER (Non-overlapping Speech)
Traditional Clustering Methods (x-vector + Spectral Clustering)	23.7%	35%	13-15%
EEND (End-to-End Neural Diarization)	12.6%	14.6%	10-12%

Table 2. Performance of traditional clustering-based methods and TDNN on the CallHome dataset.

Method	DER (Oveall)	DER (Overlapping Speech)	DER (Non-overlapping Speech)
Traditional Clustering Methods (x-vector + Spectral Clustering+AHC)	23.7%	35%	13-15%
TDNN (Time-delayed neural network)	23%	35%	12-14%

While TDNN-based methods improve the quality of speaker embeddings (e.g., through x-vectors), they still encounter limitations due to the traditional clustering step. The overall DER remains approximately 23% for both traditional clustering-based methods and TDNN-based methods when applied to the CallHome dataset, with the

DER for overlapping speech hovering ranging 35%**.** Systems based on TDNNs demonstrated slight improvements in performance for non-overlapping speech scenarios, with DER around 12-14%**,** compared to 13-15% for traditional clustering-based systems. Consequently, the overall DER and overlapping speech DER figures are analogous to those of traditional x-vector systems. The Fully Supervised UIS-RNN model significantly outperformed traditional clustering-based diarization methods on the CallHome dataset, achieving a much lower DER of 7.6% overall. Its ability to handle overlapping speech and dynamically manage speaker transitions gives it a clear advantage, particularly in complex, real-world scenarios with multiple speakers. This result further demonstrates that fully supervised, sequence-aware models like UIS-RNN represent a substantial advancement in speaker diarisation technology.

Comparative analysis of performance across diverse acoustic conditions reveals distinct patterns in methodological efficacy. In non-overlapping speech scenarios, all deep learning approaches demonstrated measurable improvements over traditional methods, with the UIS-RNN architecture achieving particularly significant enhancements in accuracy. For overlapping speech conditions, both EEND and UIS-RNN architectures substantially outperformed traditional methodologies and TDNN-based approaches, demonstrating their superior capability in managing simultaneous speaker scenarios. The comparative performance across different methodologies and scenarios is illustrated in Figure 3, which presents the evolution of DER across various methods and acoustic conditions, clearly demonstrating the superior performance of deep learning approaches, particularly in challenging overlapping speech scenarios.

Table 3. Performance of traditional clustering-based methods vs. Fully Supervised UIS-RNN on the CallHome dataset.

Method	**DER (Oveall)**	**DER (Overlapping Speech)**	**DER (Non-overlapping Speech)**
Traditional Clustering Methods (x-vector + Spectral Clustering)	23.7%	35%	13-15%
Fully Supervised UIS-RNN	7.6%	10-12%	5-7%

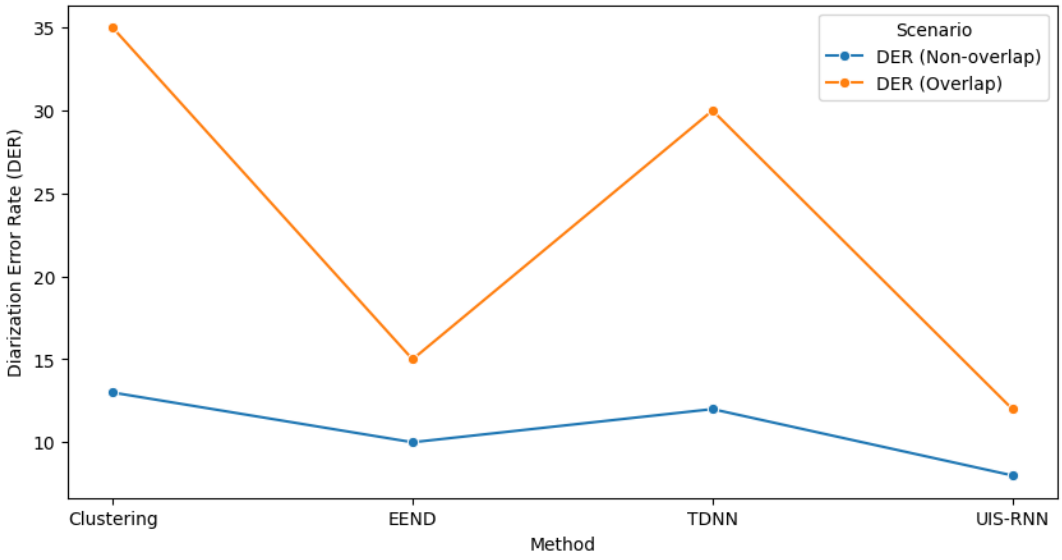

Figure 3. *Evolution of Diarization Error Rate (DER) by methods and scenario.*

Table 4. Performance of Traditional Clustering-Based Methods and Deep Learning Models on the CallHome Dataset.

Method	**Overall DER**	**DER (Overlapping Speech)**	**DER (Non- Overlapping Speech)**
Traditional Methods	Automatically adapts to speaker variations	Relies on x-vectors to handle speaker variability	Handles speaker variability well, especially for known speakers
TDNN-based systems	Computationally expensive training phase	Efficient but clustering is computationally expensive with larger datasets or many speakers	RNNs and supervision add high complexity
EEND	Excellent performance to handle overlaps	Struggles with overlaps**.**	More effective than TDNN, but not as effective as EEND
UIS-RNN	Training is resource-intensive	Scalable but clustering slows	RNNs harder to scale

Table 5. Comparison of traditional methods and new deep learning models on the key challenges of speaker diarization.

Metric	Traditional Clustering Methods	EEND	TDNN (x-vectors + Clustering)	Fully Supervised (UIS-RNN)
Speaker Variability	Prone to errors with speaker variability.	Automatically adapts to speaker variations	Relies on x-vectors to handle speaker variability	Handles speaker variability well, especially for known speakers
Computational Complexity	Clustering is moderate,but struggles with large datasets.	Computationally expensive training phase	Efficient but clustering is computationally expensive with larger datasets or many speakers	RNNs and supervision add high complexity
Overlapping Speech Handling	High DER in overlapping speech scenarios.	Excellent performance to handle overlaps	Struggles with overlaps.	More effective than TDNN, but not as effective as EEND
Scalability	Struggles with large datasets and a high number of speakers	Training is resource-intensive	Scalable but clustering slows	RNNs harder to scale
Real-Time Applicability	Relatively lightweight and applicable in real-time systems	High computational cost	Efficient but clustering delays	Real-time performance limited with longer conversations by RNNs

4.1. Computational Complexity and Real-time Implementation Considerations

The implementation of advanced deep learning architectures including EEND, TDNN, and UIS-RNN introduces significant computational considerations that warrant careful examination, particularly for real-time applications. While EEND demonstrates superior performance in managing overlapping speech with a DER of 12.6%, its neural network architecture and self-attention mechanisms require substantial computational resources. The model's performance advantages are accompanied by increasing processing demands, particularly in scenarios involving multiple speakers and extended conversations.

TDNN architectures, despite offering improved speaker embeddings through x-vector extraction, present substantial computational demands during both training and inference phases. The time-delayed layers, while effective for capturing temporal dependencies, require significant memory resources for maintaining multiple time-steps of speech features. The computational complexity increases linearly with the number of time-delay layers and the dimension of the feature representations, potentially impacting real-time processing capabilities.

Similarly, while the UIS-RNN model achieves optimal performance with a DER of 7.6%, this enhanced performance necessitates significant computational overhead. The model's implementation of parameter-sharing RNNs and distance-dependent Chinese restaurant processes for speaker clustering adds complexity to real-time processing capabilities. These computational demands become particularly evident in scenarios involving multiple speakers and variable acoustic conditions. For real-time applications, considerations emerge regarding the latency effects of processing overlapping speech segments, resource allocation for maintaining speaker embeddings, computational scalability with increasing numbers of speakers, and the balance between model processing efficiency.

5. Conclusions

Traditional speaker diarization methods are effective in controlled environments and scenarios when computational resources are limited. However, these methods often encounter challenges in terms of scalability and accuracy in more complex, real-world scenarios, particularly in the context of overlapping speech or variable acoustic conditions. While traditional techniques can be beneficial in constrained situations, their limitations become apparent in dynamic or noisy environments.

Conversely, deep learning methods have made notable advancements by utilising neural networks to automatically extract features from data, which enhances accuracy in these challenging scenarios. These approaches stand out for their flexibility and ability to generalize well across varied datasets. However, deep learning models require significant amounts of labelled data, considerable computational power, and extended training durations.

TDNNs provides better speaker representation, resulting in clearer speaker separation when speech segments do not overlap. However, the improvements achieved by TDNN-based systems are incremental, particularly when dealing with overlapping speech and multispeaker interactions. To overcome these challenges, more

advanced techniques such as End-to-End Neural Diarization (EEND) and UIS-RNN have proven to be more effective. These models have been developed to address the issue of overlapping speech and to manage dynamic speaker transitions more effectively. They have been demonstrated to perform significantly better than traditional and TDNN-based systems in complex environments.

Overlapping speech presents a significant and persistent challenge in speaker diarization, as it requires the accurate separation of simultaneously occurring vocal signals. Traditional clustering-based methods, which rely handcrafted features and statistical assumptions, leading to high Diarization Error Rates (DER). While Time Delay Neural Networks (TDNN) enhance speaker embeddings through x-vectors, the clustering phase remains inadequate for resolving the complexities of overlapping speech. In contrast, End-to-End Neural Diarization (EEND) signifies a notable advancement by employing self-attention mechanisms to directly model and differentiate overlapping signals. This approach reduces the DER for overlapping speech substantially. However, its scalability is limited, particularly in scenarios involving a greater number of speakers than the network is trained to accommodate. The Fully Supervised UIS-RNN further enhances performance by dynamically managing speaker transitions and modeling overlapping speech with high precision, achieving the lowest DER among the methods evaluated. Nevertheless, its computational intensity and reliance on detailed, time-stamped annotations constrain its applicability in large-scale or real-time contexts. These findings underscore the critical need for diarization approaches that effectively address overlapping speech while maintaining a balance between computational efficiency, scalability, and accuracy.

The future direction of speaker diarisation research necessitates investigation of several critical domains. The integration of sophisticated neural network architectures, particularly hybrid attention mechanisms and graph neural networks, emerges as fundamental to enhancing system performance in overlapping speech scenarios. These architectures demonstrate significant potential for modeling inter-speaker dynamics and facilitating real-time detection of overlapping speech patterns. Furthermore, the integration of multimodal approaches, specifically the utilization of visual cues and spatial audio information, presents promising avenues for advancing speaker identification and localization performance.

The scalability dimension presents another crucial area for investigation, necessitating the development of computationally efficient models capable of processing multiple speakers while maintaining adaptability across diverse acoustic environments. In this context, the optimization of computational resources through model compression techniques and enhanced inference methodologies has emerged as a critical consideration. Moreover, the development of unified frameworks that integrate complementary tasks, such as automatic speech recognition and diarization, demonstrates potential for both performance enhancement and resource optimization, representing a significant advancement in system architecture.

Finally, the establishment of standardised evaluation frameworks, encompassing novel metrics for assessing overlapping speech performance and implementing consistent testing protocols, constitutes a fundamental requirement for rigorous assessment of technological advancements in this domain. These research trajectories collectively aim to enhance the robustness and practical applicability of speaker diarization systems in real-world implementations while addressing current technological limitations. The systematic progression in these areas is anticipated to facilitate the development of more sophisticated and adaptable speaker diarization systems, capable of effectively managing increasingly complex acoustic scenarios in real-world applications.

Declaration of Ethical Standards
The authors declare that they comply with all ethical standards

Credit Authorship Contribution Statement Author
Author 1: Research, Methodology, Experiment, Writing
Author 2: Research, Methodology, Experiment, Writing

Declaration of Competing Interest
The authors have no conflict of interest to declare regarding the content of this article.

Data Availability
All data generated or analyzed during this study are included in this published article.

6. References

Al-Hadithy, T.M., Frikha, M., & Maseer, Z.K., 2022. Speaker Diarization based on Deep Learning Techniques: A Review. *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 856-871. https://doi.org/10.1109/ISMSIT56059.2022.9932710

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(2), 356-370. https://doi.org/10.1109/TASL.2011.2125954

Bredin H.,2017, TristouNet: Triplet loss for speaker turn embedding," 2017 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5430-5434. https://doi.org/10.1109/ICASSP.2017.7953194

Bullock, L., Bredin, H., & Garcia-Perera, L. P., 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 7114-7118. https://doi.org/10.1109/ICASSP40776.2020.9053096

Canavan, A., Graff, D., & Zipperlen, G., CALLHOME American English Speech (LDC97S42), https://catalog.ldc.upenn.edu/LDC97S42, (11.04.2025)

Chung, J. S., Nagrani, A., & Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*. https://doi.org/10.48550/arXiv.1806.05622

Çelik, H., & Ekşi, H., 2013. SÖYLEM ANALİZİ. Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, **27**(27), 99-117.

Fiscus, J.G., Ajot, J., Michel, M., & Garofolo, J.S., 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S., Fiscus, J.G. (eds) Machine Learning for Multimodal Interaction. MLMI 2006. Lecture Notes in Computer Science, vol 4299. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11965152_28

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. (2019b). End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*. https://doi.org/10.48550/arXiv.1909.05952

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S., 2019. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Singapore,* 296-303).

https://doi.org/10.1109/ASRU46091.2019.9003959

Hamza, H., Gafoor, F., Sithara, F., Anil, G., & Anoop, V. S., 2023. EmoDiarize: Speaker Diarization and Emotion Identification from Speech Signals using Convolutional Neural Networks, *arXiv preprint arXiv:2310.12851*. https://doi.org/10.48550/arXiv.2310.12851

Kshirod, K.S., 2020. Speaker Diarization with Deep Learning Techniques. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11, 3, 2570–2582.* https://doi.org/10.61841/turcomat.v11i3.14309

Li, Y., Gao, F., Ou, Z., & Sun, J., 2018. Angular softmax loss for end-to-end speaker verification. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taiwan, IEEE,* 190-194.

https://doi.org/10.1109/ISCSLP.2018.8706570.

Mao, H. H., Li, S., McAuley, J., & Cottrell, G., 2020. Speech recognition and multi-speaker diarization of long conversations. *arXiv preprint arXiv:2005.08072*. https://doi.org/10.48550/arXiv.2005.08072

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S., 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317. https://doi.org/10.1016/j.csl.2021.101317

Raj, D., Huang, Z., & Khudanpur, S. 2021. Multi-class spectral clustering with overlaps for speaker diarization. In *2021 IEEE Spoken Language Technology Workshop (SLT), China, IEEE,* 582-589 https://doi.org/10.1109/SLT48900.2021.9383602

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B., 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, ***10*****(1-3)**, 19-41. https://doi.org/10.1006/dspr.1999.0361

Serafini, L., Cornell, S., Morrone, G., Zovato, E., Brutti, A., & Squartini, S., 2023. An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings. *Computer Speech & Language*, *82*, 101534. https://doi.org/10.1016/j.csl.2023.101534

Shum, S. H., Dehak, N., Dehak, R., & Glass, J. R., 2013. Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Transactions on Audio, Speech, and Language Processing*, ***21*****(10)**, 2015-2028. https://doi.org/10.1109/TASL.2013.2264673

Tian, J., Ye, S., Chen, S., Xiang, Y., Yin, Z., Hu, X., & Xu, X., 2024. The RoyalFlush Automatic Speech Diarization and Recognition System for In-Car Multi-Channel Automatic Speech Recognition Challenge. *arXiv preprint arXiv:2405.05498*. https://doi.org/10.48550/arXiv.2405.05498

Variani, E., Lei, X., McDermott, E., Moreno, I., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2014**, 4052-4056. https://doi.org/10.1109/ICASSP.2014.6854363

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L., 2018. Speaker diarization with LSTM. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), Canada,* 5239-5243. IEEE. https://doi.org/10.1109/ICASSP.2018.8462628

Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C., 2019. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), UK,* 6301-6305). IEEE. https://doi.org/10.1109/ICASSP.2019.8683892