Deep learning algorithms for speech interaction analysis

Algoritmos de aprendizaje profundo para el análisis de interacciones de habla

Héctor Cornide-Reyes[1*] https://orcid.org/0000-0002-9130-184X
José Jorquera[1] https://orcid.org/0009-0008-1707-6163
Diego Monsalves[2] https://orcid.org/0000-0001-6780-3754

Recibido 25 de agosto de 2025, aceptado 23 de diciembre de 2025
Received: August 25, 2025 *Accepted: December 23, 2025*

ABSTRACT

This article offers a comparative assessment of deep learning algorithms for automatic analysis of verbal interactions, drawing upon evidence from n recent scientific literature. Four widely used algorithms –LSTM, CNN, GRU, and x-vectors– are analyzed across key dimensions relevant to speech analytics in collaborative contexts: speaker diarization accuracy (DER), ability to handle overlapping speech (SDR), robustness to noise as indicated by error recognition (WER), and computational efficiency assessed via inference time. Based on previously published empirical results, this synthesis evaluates the strengths, limitations, and optimal application scenarios for each algorithm, with a focus on educational and collaborative settings. The analysis shows that LSTMs excel in temporal modeling, CNNs achieve superior performance in voice separation, GRUs reach a favorable balance between accuracy and efficiency, and x-vectors enable highly scalable speaker identification. This study establishes a comprehensive framework for selecting and integrating deep learning techniques into speech analysis systems that demonstrate robustness and context-sensitivity across educational, professional, and research applications.

Keywords: Speaker diarization, deep learning, voice separation, collaborative environments, speech analysis.

RESUMEN

Este artículo presenta una evaluación comparativa de algoritmos de aprendizaje profundo aplicados al análisis automático de interacciones verbales, basada en la evidencia reportada en la literatura científica reciente. Se analizan cuatro algoritmos ampliamente utilizados –LSTM, CNN, GRU y x-vectors– considerando dimensiones clave para la analítica de habla en contextos colaborativos: precisión de la diarización del hablante (DER), capacidad para gestionar solapamientos (SDR), robustez frente al ruido medida mediante errores de reconocimiento (WER) y eficiencia computacional a través de tiempos de inferencia. A partir de resultados empíricos previamente publicados, se sintetizan las fortalezas, limitaciones y escenarios de aplicación más adecuados para cada algoritmo, especialmente en entornos educativos y de trabajo en equipo. El análisis muestra que los LSTM destacan en el modelado temporal, las CNN obtienen un rendimiento superior en separación de voces, las GRU logran un equilibrio favorable entre precisión y eficiencia, y los x-vectors ofrecen una identificación de hablantes altamente escalable. Este estudio proporciona un marco fundamental para seleccionar e integrar técnicas de aprendizaje profundo en sistemas de análisis del habla robustos y sensibles al contexto para aplicaciones educativas, profesionales y de investigación.

Palabras clave: Diarización de hablantes, aprendizaje profundo, separación de voces, entornos colaborativos, análisis del habla.

[1] Universidad de Atacama. Departamento de Ingeniería Informática y Ciencias de la Computación. Copiapó, Chile. E-mail: hector.cornide@uda.cl; jose.jorquera@uda.cl
[2] Universidad de Valparaíso. Escuela de Ingeniería Informática. Valparaíso, Chile. E-mail: diego.monsalves@uv.cl
* Autor de correspondencia: hector.cornide@uda.cl

DOI: 10.64966/ingeniare.v33.32.

INTRODUCTION

In an increasingly interconnected and collaborative world, teamwork serves as a fundamental pillar of productivity and success across diverse fields, such as business, education, and the science [1]. Team effectiveness is primarily determined by the ability to coordinate and pursue shared objectives, which is intrinsically linked to the quality of interactions among its members [2]. In this context, communication is a critical factor, functioning as the primary channel for collaboration, knowledge exchange, negotiate responsibilities, and collective decisions-making. Clear and effective verbal interaction aligns expectations and fosters an environment that supports idea exchange, innovation, and informed decision-making [3].

The analysis of interactions within collaborative teams encompasses not only verbal language but also non-verbal cues, such as facial expressions, body postures, and the spatial distribution of participants in the physical environment [4], [5]. Among these elements, speech interactions play a pivotal role, as they constitute the core of effective coordination. Speech enables participants to articulate ideas, respond to interlocutors, and coordinate actions to achieve group objectives [6]. A comprehensive understanding of these interactions is essential for assessing team performance and developing strategies to improve collaboration. However, collaborative environments pose significant challenges for speech analysis, particularly due to overlapping voices and variability in audio signal quality [7], [8]. Voice overlap, which occurs when multiple individuals speak simultaneously, complicates the accurate identification of speakers and the comprehension of each utterance's content. This issue introduces ambiguity regarding the speaker identity, content of speech, and timing, which complicates the interpretation of communication flow and internal group dynamics. Additionally, variations in volume, distance from recording devices, and background noise can degrade data quality, thereby limiting the potential for temporal and semantic analysis [9].

Voice overlap poses a significant challenge in environments with simultaneous interactions among multiple individuals, impeding communicative clarity and accurate assessment of individual contributions. Recent technologies, including ReSpeaker [7], [8] and the NAIRA platform [10], [11], have sought to address this challenge and have demonstrated promising results. However, these systems continue to face limitations in complex scenarios with high voice concurrency and adverse acoustic conditions. This context underscores the need to explore artificial intelligence-based solutions to achieve higher levels of voice separation and detection.

The present paper addresses these challenges by providing a comprehensive analysis of deep learning algorithms, relying exclusively on evidence from scientific literature. The objective is to examine the capabilities, limitations, and operational characteristics of these algorithms, specifically regarding their applicability to speaker recognition and voice separation in collaborative contexts. The selected algorithms will be characterized using key metrics reported in the literature, such as speaker identification accuracy, robustness under adverse acoustic conditions, and computational efficiency. Additional considerations will include the required amount of training data, algorithmic complexity, and scalability for future applications. This analysis will facilitate the classification of algorithms based on their suitability for addressing the voice overlap problem across different collaborative scenarios. The comparison will be based on theoretical analysis, contrasting evidence from previous studies. This process will help identify performance patterns, highlight more robust approaches, and justify the selection of algorithms with the highest potential for application. The final selection will be determined by the algorithm's capacity to separate and detect voices in complex conditions, and by its technical feasibility for future implementation.

This paper presents a theoretical analysis and the characterizes the algorithms identified in the scientific literature. The aim is to provide a solid reference framework to advance voice recognition research and for enhance understanding of interaction dynamics in collaborative environments. Through a conceptual approach, this study provides an in-depth analysis that establishes a foundation for subsequent research and future technological advancements.

DEEP LEARNING IN SPEECH RECOGNITION

Deep Learning (DL) is a subcategory of machine learning that employs artificial neural networks

to model and address complex problems. These networks, inspired by the structure of the human brain, are composed of multiple layers of nodes (neurons) that incrementally transform input data into more abstract and informative representations. Thanks to this capability, deep learning-based systems have achieved high levels of accuracy in tasks such as speech recognition, computer vision, and machine translation [12]. Although the fundamental concepts of artificial neural networks originated in the 1940s, significant advances in deep learning occurred in the 2000s, driven by the availability of large volumes of data, enhanced processing power, and novel training algorithms. Notable architectures include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) variants, which are widely applied to modeling spatial and sequential data [12].

Speech recognition represents one of the most advanced yet challenging applications of deep learning, primarily due to speech variability, different accents, and background noise. Techniques based on RNNs and CNNs have significantly enhanced the accuracy of these systems by learning and adapting to complex audio signals. These developments have expanded the applicability of speech recognition to dynamic and uncontrolled environments where acoustic conditions fluctuate significantly [13]. In speech recognition, speaker identification constitutes a specialized subarea focused on determining a speaker's identity. Models achieve specific acoustic features from the signal, such as pitch, frequency, and timbre, to build unique speaker profiles. Mel-Frequency Cepstral Coefficients (MFCCs) and Vector Quantization (VQ) are among the most widely used techniques for this purpose, enabling high-precision speaker differentiation, even under adverse conditions [14].

- MFCCs are a standard feature extraction technique in speech recognition. These coefficients capture the spectral characteristics of speech by applying a sequence of transformations, including the Fast Fourier Transform (FFT), Mel filter bank, and Discrete Cosine Transform (DCT). This process generates a compact set of coefficients that are robust to variations in pitch and speech rate [14].
- VQ reduces the amount of data needed to process the extracted features by clustering similar vectors into clusters and assigning a representative value to each cluster. This method facilitates comparison and enhances efficiency in speaker identification.

Speaker recognition systems can be classified into text-dependent and text-independent based on the recognition rule. Text-dependent systems require the speaker to articulate predetermined phrases during recognition. In contrast, text-independent systems operate without this restriction, making them more flexible and better suited for dynamic environments. Currently, text-dependent systems are considered largely obsolete because users frequently perceive this requirement as a functional limitation.

Modern voice recognition involves three main phases: signal recording and processing, feature extraction, and feature matching. Deep learning has significantly enhanced each phase:

- Signal recording and processing: This phase captures acoustic signals and prepares them for analysis. DL models have enhance noise and distortions filtering, increasing signal clarity and quality.
- Feature extraction: In this stage, relevant speech representations are identified. Deep neural networks learn complex beyond the reach of traditional methods. For example, CNNs detect patterns in audio spectrograms, while RNNs capture the temporal structure of the signal.
- Feature matching: This phase compares extracted features with pre-trained models to identify the speaker or content. Deep learning techniques have improve accuracy by adapting to speech variations and diverse acoustic conditions.

Advances in deep learning have led to the development of numerous tools and frameworks that support each stage of the speech recognition process. These resources improve accuracy and efficiency and enable the design of more robust and adaptable models.

METHODOLOGY OF LITERATURE REVIEW

Using a narrative literature review with a structured search, this study identifies the principal deep learning

approaches for speech interaction analysis.The aim was not to perform a formal systematic review or mapping study, but to gather and synthesize the most relevant scientific evidence to characterize the behavior of the selected algorithms (LSTM, CNN, GRU, and x-vectors) across key performance metrics.

The following protocol ensures baseline transparency and reproducibility in the review.

1. Databases consulted: IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar.
2. Search period: Publications from 2015 to 2024, reflecting the rise and establishment of modern deep learning methods for speech processing.
3. Keywords used in different combinations:
   - "speaker diarization deep learning"
   - "speech separation CNN LSTM"
   - "x-vectors speaker recognition"
   - "GRU speech recognition performance"
   - "Deep learning collaborative speech analysis"
4. Inclusion criteria:
   - Articles describing deep learning architectures applied to speaker identification, diarization, speech separation, or speech recognition.
   - Studies reporting performance metrics such as DER, SDR, WER, or inference time.
   - Works that provided qualitative or quantitative evidence relevant to the comparative dimensions defined in this paper.
5. Exclusion criteria:
   - Papers without technical descriptions of the model.
   - Works that lacked performance metrics or comparative evaluation.
   - Studies focused exclusively on non-deep-learning approaches.
6. Selection process: The authors reviewed titles, abstracts, and, when needed, full texts to determine relevance. Representative works were selected based on their clarity, methodological rigor, and influence on the field.
7. Nature of the comparison: No experimental benchmarking was conducted. All performance values and qualitative assessments derive exclusively from the reviewed literature. The comparison synthesizes commonly reported patterns rather than offering controlled empirical validation.

This methodological approach offers a transparent foundation for the analysis while aligning with the study's theoretical and comparative scope. While this review does not use systematic review methodology, it applies targeted search criteria and documents selection steps to ensure transparency and traceability.

RELATED WORK

The analysis of interactions within collaborative teams is a pivotal research area for understanding how effective dynamics are established in educational and professional contexts. Recent developments in technologies, such as multimodal analysis and sensor use, have expanded the possibilities for studying these interactions by integrating verbal, non-verbal, and gestural dimensions. These tools enable a more comprehensive understanding of group processes and facilitate the assessment of the influence of individual and collective characteristics on collaborative performance.

Various studies have addressed this issue from complementary technological and methodological perspectives. The following section reviews representative works that have employed multimodal analysis techniques and machine learning to explore collaboration dynamics. In [9], the Multimodal Learning Analytics (MMLA) approach was applied to analyze collaboration among software engineering student teams during the writing user stories. The study employed multidirectional microphones (ReSpeakers) and social network analysis metrics to assess indicators such as speaking time, intervention frequency, and each member's role in idea generation. The findings showed that the most collaborative teams produced more stories, although no direct relationship with quality was identified. Additionally, prior experience in requirements analysis was found to influence collaborative performance. The study's limitations included the short duration of the activity and the absence of a real client.

In [8], the authors introduced influence graphs as a means of visualizing interactions among student groups engaged in engineering tasks. Data collection via ReSpeakers enable the construction of graphical representations subsequently analyzed using metrics such as permanence, persistence, and idea promotion. The study identified several forms

of effective collaboration, including dynamics with distributed participation, centralized leadership, or minimal interaction. This diversity challenges the notion that symmetrical interactions exclusively indicate collaborative success. However, the study acknowledged limitations, such as the absence of non-verbal cues and the lack of consideration of individual participant profiles.

In [7], the use of low-cost sensors integrated into ReSpeakers devices was explored to analyze collaboration dynamics during simulations of the agile Scrum framework using the Lego4Scrum methodology. Verbal data were collected during retrospective sessions and combined with social network analysis and personality assessments. The results indicated that teams with higher levels of collaboration produced more accurate estimates and adopted more democratic leadership styles. Limitations of the study included its exclusive focus on verbal analysis and the influence of background noise.

Complementing this approach, [10] introduced NAIRA, a real-time feedback platform designed for remote collaborative activities. This tool integrates spoken interactions with visualizations, including influence graphs and silence alerts, allowing instructors to intervene promptly to balance participation. The study reported improvements in participation equity through the platform; however, technical limitations such as audio calibration and environmental noise were also identified.

In [15], an alternative approach for classifying collaborative dynamics using only gestures on tablets and speech patterns was introduced. A Random Forest-based machine learning model was developed to categorize interactions as collaboration, cooperation, or asymmetric contribution, achieving an accuracy rate of 92%. While the study demonstrates the potential of computational approaches, it also acknowledges limitations due to the absence of non-verbal data and technical challenges encountered during data collection.

Collectively, these studies demonstrate significant advances in applying technologies such as MMLA and machine learning to study collaboration in educational settings. Technologies such as ReSpeakers, social network analysis, and classification models have proven capable of capturing valuable information on participation, leadership, and interaction quality. Nevertheless, several persistent limitations have also been identified:

- Challenges persist in analyzing scenarios with overlapping voices and multiple simultaneous speakers.
- Accuracy in identifying individual speakers remains limited in noisy or dynamic environments.
- Most approaches are focused on controlled settings and demonstrate little generalization to complex, real-world environments.
- A lack of integration of non-verbal or gestural cues is evident in many studies.
- There is limited adoption of advanced deep learning-based models for voice separation and speaker recognition.

This literature review reveals a significant gap in current systems' ability to provide robust, accurate analysis of verbal interactions in real collaborative scenarios. Most existing solutions have prioritized the recording of interactions but have not integrated practical tools for voice separation or individual speaker identification, which are essential for understanding participation balance and leadership within teams. This research aims to address these limitations through a theoretical analysis of deep learning algorithms applied to voice recognition and separation. Integrating source separation techniques with multimodal tools is expected to enable more accurate capture of interactions in prolonged and complex collaborative environments, thereby providing a foundation for assessing and optimizing teamwork dynamics in educational and professional settings.

Table 1 enhances transparency and traceability in comparative analysis by providing a structured summary of the most relevant studies reviewed. It highlighting the algorithms used, the datasets referenced, the metrics reported, and the contributions derived from each work. This table addresses the reviewers' request for methodological clarity and consolidates the empirical evidence underpinning the analysis.

VARIABLES OF INTEREST AND ASSESSMENT METRICS

For the algorithm's evaluation, four key variables were defined to represent crucial dimensions of

Table 1. Summary of representative studies included in the literature analysis.

Reference	Algorithms involved	Datasets referenced	Metrics reported	Main contribution
[17], [18]	LSTM / BiLSTM	General diarization corporal	DER	Improve temporal stability and diarization accuracy
[19]	LSTM + x-vector backend	CHiME, DIRHA	DER, WER	Robust under noise; effective hybrid model performance.
[20]	CNN, LSTM, GRU	DIRHA, CHiME	DER, WER, Inference time	CNN as a strong frontend; GRU efficient; LSTM robust in noise.
[21]	LSTM, GRU, RNN	Not specified	DER, WER	LSTM is superior in diarization; GRU is stable but less precise.
[22]	Lightweight CNN	ICASSP multichannel	SDR	CNN improves multichannel speech separation.
[24]	CNN, GRU	VoxCeleb	DER	Strong embeddings for speaker identification.
[25]	CNN + GRU hybrid	Not specified	DER, SDR	CNN improves diarization by 14%; strong in overlap handling.
[26], [27]	CNN	Not specified	SDR	CNN-based architecture leads to voice separation performance.
[29]	GRU + attention	Not specified	–	Shows GRU's limited SDR role but potential in multimodal tasks.
[31], [32], [34]	x-vectors	DIHARD, VoxCeleb	DER	Highly effective diarization; weak in overlapped speech.
[35]	x-vectors + noise separation modules	Not specified	–	Confirms the need for preprocessing to improve noise resilience.

automatic speech analysis. These variables were selected to address prevalent challenges in real-world contexts, including classrooms, meetings, or collaborative environments, and to align with the specific objectives of this research. Each variable was associated with a recognized, standardized metric that is widely used in the scientific literature to evaluate the performance of voice processing systems. The following section describes each variable and the corresponding metric used to assess it.

1. **Speaker identification accuracy**: One of the first questions when analyzing a conversation is: Who said what? Transcribing the spoken content is not enough; in many contexts, identifying the speaker is just as relevant as knowing what was said. This capability is essential for tracking participation in group activities, managing speaking turns, and attributing responsibilities in educational or professional settings. This task, known as speaker diarization, involves segmenting an audio recording and assigning each segment to its corresponding speaker. The complexity increases when speaking turns are not clearly defined, interruptions occur, or participants' voices are similar. The Diarization Error Rate (DER) [16] is used to evaluate this dimension by calculating the percentage of audio segments misassigned to the wrong speaker. DER accounts for three types of errors: incorrect assignments, missed speakers, and false detections of non-existent speakers. Lower DER values indicate higher system precision and reliability.
2. **Handling of voice overlap**: In natural conversations, it is common for multiple individuals to speak simultaneously. Far from being an anomaly, this phenomenon constitutes an inherent part of conversational flow. However, voice overlap presents a significant challenge for automatic systems it can hinder voice separation, introduce ambiguity in speaker identification, and lead to information loss. The solution demands systems capable of separating mixed acoustic signals without degrading speech clarity or speaker's identity. The Signal-to-Distortion Ratio (SDR) is commonly employed to assess this A capability by measuring the quality of reconstructed signals after separation. Higher SDR values indicate reduced distortion and therefore reflect improved performance in recovering original voices.

3. **Computational efficiency**: system performance encompasses not only accuracy but also processing speed and efficient use of computational resources. This consideration is particularly relevant for applications that require real-time responses, such as virtual assistants or active learning platforms, as well as scenarios with restricted access to high-performance infrastructure. For this dimension, the relevant metric is inference time, defined as the time required to process one input unit, such as an audio file, and generates an interpretable output. A shorter inference time implies greater efficiency and better adaptability of the system for real-world operational conditions.
4. **Noise robustness**: Speech analysis in real-world environments must account for adverse acoustic conditions, including ambient noise, echoes, background conversations, and other factors. These Such factors can degrade signal quality and negatively affect the performance of automatic systems, which are typically trained in controlled settings. Robustness refers to a system's capacity to sustain acceptable performance levels even under unfavorable environmental conditions. The Word Error Rate (WER) is commonly used to assess this aspect under noisy conditions. WER quantifies the percentage of words that are misrecognized or mis-transcribed compared to a reference. A lower WER indicates greater resilience to noise and better preservation of the intelligibility of spoken content.

The selected variables and metrics meet technical requirements and are designed to assess algorithm performance from both a functional and contextual perspective. This research does not merely seek to identify which system achieves the best performance figures. It seeks to determine how and why specific solutions may be more suitable than others in real-world speech analysis scenarios, including educational, collaborative, or applied research contexts.

DEEP LEARNING ALGORITHM EVALUATION

This section provides a comparative evaluation of deep learning architectures commonly used for analyzing speech interactions, focusing on key metrics that assess their performance in real collaborative contexts. Each subsection examines a specific algorithm and analyzes its behavior across four evaluative dimensions: DER, SDR, Inference Time, and WER. This comparative approach establishes a solid and transparent technical foundation for understanding each architecture's capabilities, facilitating a direct link between their technical characterization and functional applicability in speech analysis systems.

Long short-Term memory

Long Short-Term Memory (LSTM) networks are a variant of recurrent neural networks (RNNs) specifically designed to model sequences with long-term dependencies. Their architecture, which includes input, forget, and output gates, has enabled extensive application in tasks such as speaker diarization, automatic speech recognition, and acoustic sequence modeling.

a) **DER**: Speaker diarization, defined as the segmentation of audio recordings to determine "who speaks when," benefits from the robust approach for capturing temporal patterns in audio data. Performance in this area is typically assessed using DER metrics. Empirical studies such as [17], [18] demonstrated that LSTMs, particularly their bidirectional variant (BiLSTM), significantly improve DER when integrated into d-vector-based systems. Within the ESPnet project [19], LSTMs are used as a backend for x-vector architectures, highlighting their effective modelling of temporal sequences effectively. Further evidence from [20], employing PyTorch-Kaldi, indicates enhanced temporal stability under noisy conditions. Comparative studies such as [21] confirm that LSTMs consistently outperform alternatives, such as GRUs and simple RNNs, in speaker segmentation tasks. LSTMs demonstrate superior performance in diarization tasks, particularly when integrated within hybrid architectures. The capacity of LSTMs to model long-term relationships justifies a high rating in this dimension.
b) **SDR**: Although LSTMs were not initially designed for source separation tasks, various studies report their usefulness in contexts involving partial voice overlap. In ESPnet, their integration into multichannel systems has yielded modest increases in the SDR metric [19]. In [22],

improvements were reported when combining LSTMs with lightweight CNNs to achieve acoustically coherent reconstructions. However, the sequential structure of LSTMs prevents them from handling multiple simultaneous sources, unlike convolutional or self-attentive architectures. While LSTMs can support voice separation in hybrid systems, they are not a specialized solution. Their performance depends heavily on the overall architectural design.

c) **Inference time**: LSTMs face significant challenges due to their poor inference performance in terms of time. Their sequential design impedes efficient parallelization, resulting in significant latencies, particularly in tasks that require real-time processing. Studies such as [20], [21] report that lighter variants, such as GRU or Li-GRU, offer a better balance between speed and accuracy. Due to their low efficiency, LSTMs are not well-suited for latency-sensitive or resource-constrained environments. This limitation justifies a minimal rating in this dimension.

d) **WER**: In adverse acoustic conditions, LSTMs effectively preserve phonetic content and maintain low WER. In corpora such as CHiME and DIRHA, both ESPnet [19] and PyTorch-Kaldi [20] demonstrate that their memory mechanisms improve resilience to noise. Evidence in [23] further confirms that LSTMs preserve phonetic coherence even under degraded conditions. LSTMs perform reliably in noisy environments, validating their suitability for real-world scenarios with acoustic disturbances.

Convolutional neural networks

CNNs are widely used in speech processing due to their ability to extract hierarchical patterns and relevant acoustic features. Their parallel architecture and efficiency in extracting discriminative representations make them well-suited for segmentation and source separation tasks.

a) **DER**: while originally developed for computer vision, CNNs have delivered competitive results in diarization systems. In models such as Deep Speaker [24], CNN architectures trained with triplet loss generate robust embeddings that can be used for speaker verification or identification. In [25], a CNN+GRU architecture is proposed for multi-speaker diarization tasks, achieving a 14% reduction in DER compared to models without convolutional components. In PyTorch-Kaldi [20] uses CNNs as the acoustic frontend to enhance early discrimination of features relevant to speaker identification. CNNs perform well in speaker identification, particularly as feature extractors. However, their effectiveness improves when combined with sequential models.

b) **SDR**: CNNs have shown outstanding performance in source separation. Models like Conv-TasNet [26], which operate directly in the time domain, have demonstrated substantial improvements in SDR and have served as a benchmark in multiple subsequent studies [27]. Integrating deep CNNs with attention mechanisms enables precise handling of regions with overlapping speech. In [25], this capability is validated by showing improvements in acoustic discrimination when CNNs are used in hybrid architectures. Deep CNNs configurations offer excellent capabilities for handling voice overlap. Their performance in complex scenarios justifies the highest rating in this dimension.

c) **Inference time**: One of the main advantages of CNNs is their highly parallelizable structure. Unlike recurrent models, CNNs allow for the simultaneous processing of multiple regions of the acoustic spectrum, thereby significantly reducing inference time. In [24], their applicability to embedded devices is highlighted, while [25] reports notable latency improvements in hybrid systems. [27] confirms that CNNs maintain low response times even in complex architectures. Additionally, CNNs excel in computational efficiency. CNNs' low inference cost makes them an ideal solution for real-time applications.

d) **WER**: Although not explicitly designed for noisy environments, CNNs have shown stable behavior under noise, particularly when used as initial acoustic extraction modules. In PyTorch-Kaldi [20], reductions in WER are reported on corpora such as DIRHA and CHiME when using CNNs as a frontend. Additionally, [27] indicates that when attention mechanisms are incorporated, CNNs preserve intelligibility under moderately degraded acoustic conditions. However, other studies [25] suggest that their noise resilience is lower than that of recurrent models such as LSTMs. CNNs deliver solid performance in noisy environments, especially when integrated

into hybrid systems. While their resilience is not exceptional, CNNs achieve acceptable levels that justify a positive evaluation.

Gated recurrent units

GRUs emerged as a more computationally efficient alternative to LSTM networks, while maintaining the ability to model temporal dependencies. These networks have been explored in various speech processing applications, including diarization and recognition in acoustically challenging conditions.

a) **DER**: In speaker identification tasks, GRUs have been used as backends in systems that generate speaker embeddings. For example, in the Deep Speaker model [24], GRUs are employed to represent speaker identity, demonstrating good discrimination on corpora such as VoxCeleb. Similarly, [20] reports that GRUs achieve performance comparable to LSTMs in multichannel environments, particularly within the PyTorch-Kaldi framework. Comparative studies such as [21], [28] indicate that although GRUs do not reach the precision levels of LSTMs, they exhibit stable behavior when properly trained. Overall, GRUs provide acceptable performance in speaker identification. While they do not outperform more complex models, their balance between simplicity and performance makes them viable for resource-constrained scenarios.

b) **SDR**: The use of GRUs for source separation remains relatively less common and yields limited results. In PyTorch-Kaldi [20], GRUs are documented as generally applicable for acoustic modeling; however, they do not provide substantial improvements in the SDR. In [29], GRUs are integrated into multimodal tasks with attention mechanisms, but these implementations do not directly address voice overlap. Likewise, [21] does not report significant benefits from GRUs in this metric. Unlike architecture designed explicitly for source separation, GRUs lack the structural mechanisms needed to disentangle multiple voices simultaneously. GRUs are not optimized for handling voice overlaps and their limited performance warrants a low rating in this metric.

c) **Inference Time**: A primary advantage of GRUs is their computational efficiency. Due to their simpler architecture with fewer gates and parameters than LSTMs, GRUs significantly reduce inference time. Studies such as [21], [30] highlight their suitability for low-latency applications. In PyTorch-Kaldi [20], it is shown that lightweight variants, such as Li-GRU, reduce per-frame processing time while maintaining competitive accuracy. The low computational cost of GRUs makes them a practical option for embedded systems or real-time applications.

d) **WER**: In acoustically complex environments, GRUs have shown acceptable robustness. Ravanelli [20] demonstrates that these networks maintain stable WER levels across corpora such as DIRHA and CHiME. [21] also reports consistent behavior under noisy conditions. GRU-based speech recognition models, such as those in [28] and [30], have shown noise resilience when combined with regularization techniques and training on diverse datasets. GRUs maintain reliable performance under degraded acoustic conditions. Although GRUs do not match the robustness of more complex architectures, their dependability justifies a positive evaluation.

x-vectors

X-vectors are among the most advanced and widely adopted embedding techniques for speaker identification and verification. Originally proposed by [31], these embeddings are generated using Time Delay Neural Networks (TDNNs) and are characterized by their ability to efficiently capture both phonetic features and speaker-specific traits.

a) **DER**: x-vectors in diarization tasks has proven to be highly effective. Within the DIHARD Challenge framework [32], significant reductions in DER were achieve by combining x-vectors, agglomerative clustering, and probabilistic models such as PLDA. Likewise, [33] shows that these embeddings enable speaker discrimination without requiring prior acoustic segmentation. Other studies, including [31] and [34], have demonstrated their robustness in multichannel environments and across diverse speech conditions. X-vectors provide outstanding accuracy in speaker identification. Performance in challenging benchmarks supports the use of X-vectors as a reliable and efficient solution.

b) **SDR**: Although x-vectors demonstrate strong performance in classification tasks, they present

significant limitations when dealing with voice overlap. These representations are intended for use with clean or pre-filtered audio segments and lack explicit source separation mechanisms. Both [31] and [32] report that simultaneous speech increases DER, thereby affecting diarization quality. In [35], it is suggested that identity and noise representations should be explicitly separated, confirming the need for specialized preprocessing. X-vectors are not designed to solve voice overlap problems. Their performance depends on external modules, which limit their direct applicability in this dimension.

c) **Inference time**: One of the most notable attributes of x-vectors is their efficiency during inference. Building on this efficiency, TDNNs and the processing of fixed-length temporal frames, enable quickly parallel embedding generation. This efficiency is emphasized in [31] and [36], where its quality is highlighted as key attribute for implementation in large-scale applications. Furthermore, [34] confirms that despite intensive training requirements, inference is exceptionally lightweight. This makes x-vectors particularly suitable for real-time monitoring systems. X-vectors stand out for their low computational cost and scalability, which render them highly suitable for production environments and online deployments.

d) **WER**: Although x-vectors provide stable speaker representations, their robustness in adverse acoustic conditions is not inherent. In [31], data augmentation techniques are implemented to improve performance in noisy environments; however, effectiveness remains highly dependent on the quality of preprocessing. In [35], architecture is proposed that explicitly separates identity and noise components, highlighting that noise resilience is not a native property of x-vectors but is achieved through complementary processes. X-vectors demonstrate tolerance to moderate levels of noise; however, their performance significantly depends on external techniques. This dependency justifies an intermediate rating.

Because the studies reviewed report performance metrics under heterogeneous datasets, preprocessing pipelines, and evaluation protocols, numerical values cannot be directly compared. Consequently, Table 2 summarizes quantitative trends from the literature using relative performance indications rather than absolute values.

Synthesis of the analysis

This analysis provides a qualitative, evidence-supported understanding of how different deep learning algorithms behave when addressing the challenges of automatic speech analysis in real-world collaborative contexts. Four approaches were assessed, LSTM, CNN, GRU, and x-vectors, based on key metrics such as DER, SDR, Inference time, and WER.

Analysis of the results indicate that no single algorithm demonstrate superiority across all aspects. Each algorithm has distinctive strengths that make it better suited to specific tasks different and usage contexts. For instance, when the main objective is accurate speaker identification and temporal segmentation, LSTM architectures and x-vectors show outstanding results. LSTMs architectures are effective at modeling temporal continuity in speech, whereas x-vectors offer a compact, highly efficient solution for scalable or real-time systems. In scenarios involving multiple individuals speaking simultaneously, CNNs are the best option. Their structure allows for more precise voice separation, even in complex scenarios with heavy overlap.

Regarding computational efficiency, both GRUs and x-vectors demonstrate notable advantages. GRUs, due to their lighter design, offer faster response times than LSTMs while maintaining comparable accuracy. In contrast, X-vectors combine rapid processing and low resource consumption with highly competitive performance. When environments are characterized by noisy conditions, as may occur in classrooms, laboratories, or open environments, LSTMs prove to be particularly robust. Thanks to their memory mechanism, LSTMs can maintain high recognition quality even under suboptimal audio conditions. Thus, for real-world scenarios that demand both speaker identification accuracy and noise robustness, LSTMs deliver excellent performance. When efficient processing and scalability are prioritized, x-vectors represent a strong alternative. When the challenge lies in voice overlap, CNNs provide the most effective solution.

A five-level qualitative scale (Excellent, Good, Medium, Low, and Very Low) was defined to structure

Table 2. Quantitative trends reported in the literature for the selected algorithms.

Algorithm	DER (relative)	SDR (relative)	Inference time (relative)	WER (relative)
LSTM	Significant DER reduction in [17]-[21].	Modest SDR improvements in hybrid systems.	High latency; slower than GRU and CNN.	Very low WER; strong robustness in [19]-[23].
CNN	Competitive DER, improved in [24]-[25].	Strong SDR improvements; state-of-the-art.	Very low inference time; highly parallel.	Acceptable WER; moderate resilience.
GRU	Comparable DER to LSTM in [20]-[21].	Limited SDR performance in [20]-[29].	Lowest inference time among RNNs.	Stable WER under noise in [20]-[30].
x-vectors	Strong DER reduction in DIHARD [31]-[34].	Very low SDR; limited in overlapped speech.	Very low inference time; highly scalable.	Moderate WER; depends on preprocessing [31]-[35].

the theoretical evaluation of the selected algorithms, allowing for a synthesized performance across each metric. This scale is not intended to be absolute or to cover all possible variants; instead, it serves as an analytical tool to facilitate conceptual comparison of architectures based on empirical evidence from scientific literature. The corresponding qualitative evaluation of the algorithms is shown in Table 3.

These findings facilitate the selection of one model over another and inform the integration of their strengths into hybrid architectures suitable for diverse educational, collaborative, or professional contexts. Therefore, this work not only compares algorithms but also establishes a robust foundation for making technically sound decisions based on real-world evidence.

In addition to the qualitative assessment presented in Table 4, a detailed technical-functional comparison of the four architectures is summarized in Table 4. This table synthesizes architectural characteristics, key strengths, main limitations, and the optimal application scenarios for each algorithm, thereby providing a clearer basis for model selection in real-world speech interaction analysis.

CONCLUSIONS

This study provides a comparative evaluation of deep learning algorithms for analyzing speech interactions in collaborative contexts, focusing on LSTM, CNN, GRU, and x-vector architectures. The review is structured around four key dimensions: speaker identification accuracy (DER), management of voice overlap (SDR), computational efficiency, and noise robustness (WER). The analysis results indicate that no single model outperforms others across all metrics; however, each demonstrates distinct strengths that informs its suitability for particular scenarios:

- **LSTM** offers robust performance in diarization tasks and maintains resilience under adverse acoustic conditions. It is particularly suitable when speaker identification accuracy is the prioritized in noisy environments, although its computational efficiency remains limited.

Table 3. Theoretical evaluation of the selected algorithms.

Algorithms	DER	SDR	Inference time	WER
CNN	Good	Excellent	Excellent	Good
LSTM	Excellent	Medium	Medium	Excellent
GRU	Medium	Low	Excellent	Good
x-vectors	Excellent	Very low	Excellent	Medium

Table 4. Technical-functional comparison of the analyzed algorithms.

Algorithms	Architectural Nature	Strengths (from literature)	Limitations (from literature)	Most suitable scenarios
LSTM	Recurrent model with long-term memory gates.	Excellent temporal modeling; Robust under noise; High DER accuracy; Good phonetic coherence.	High inference time; Difficult parallelization; Computationally heavy.	Noisy classrooms; Multi-turn dialogue modeling; Long audio segments.
CNN	Hierarchical feature extractors with high parallelism.	Best SDR performance; Very low inference time; Strong feature extraction for diarization; Good scalability.	Moderate noise robustness; Requires hybrid models for best DER.	Overlapping speech scenarios; Real-time separation; Embedded systems.
GRU	Lightweight recurrent model with fewer parameters.	Fastest inference among RNNs; Stable WER under noise; Competitive DER in some benchmarks.	Weak SDR performance; Less expressive temporal modeling vs. LSTM.	Low-latency environments; Mobile/ embedded deployments.
x-vectors	TDNN-based embeddings for speaker identity.	Outstanding DER; Very efficient inference; Highly scalable; Strong in speaker verification tasks.	Very poor SDR; Sensitive to concurrent speech; Noise handling depends on preprocessing.	Large-scale speaker identification; Monitoring systems; Real-time diarization with clean segments.

- **CNN**, in contrast, stands out as the best option for handling voice overlap, excelling at source separation and computational efficiency due to its parallelizable architecture.
- **GRU** is an efficient alternative in terms of speed and resource consumption, making it suitable for embedded systems or applications that require low latency. However, its performance in source separation tasks is comparatively modest.
- **X-vectors** excel in speaker identification accuracy and inference efficiency, making them well suited for scalable applications. However, they require additional preprocessing when there is significant voice overlap or intense background noise.

LSTM architecture delivers the most balanced performance when consistent, robust functioning is required, particularly in real-world scenarios with high acoustic variability. Conversely, CNNs and x-vectors offer notable advantages in efficiency and specialization, highlighting the potential of hybrid approaches that combine the strengths of both architectures.

Drawing on evidence from scientific literature, this study establishes a theoretical and technical foundation for the informed selection of algorithms in the design of automatic speech analysis systems for collaborative contexts. Future research may expand this comparison by conducting empirical evaluations with real data obtained from educational, workplace, or clinical settings.

REFERENCES

[1] M. Laal and M. Laal, "Collaborative learning: what is it?," *Procedia - Social and Behavioral Sciences*, vol. 31, pp. 491-495, 2012, doi: 10.1016/j.sbspro.2011.12.092.

[2] E. Salas, N. J. Cooke, and M. A. Rosen, "On teams, teamwork, and team performance: discoveries and developments," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 540-547, 2008, doi: 10.1518/001872008x288457.

[3] T.J. Wiltshire, K. Van Eijndhoven, E. Halgas, and J.M.P. Gevers, "Prospects for augmenting team interactions with real-time coordination-based measures in human-autonomy teams," *Topics in Cognitive Science*, vol. 16, no. 3, pp. 391-429, 2024, doi: 10.1111/tops.12606.

[4] P. Blikstein and M. Worsley, "Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks," *Journal of Learning Analytics*, vol. 3, no. 2, pp. 220-238, 2016, doi: 10.18608/jla.2016.32.11.

[5] F. Riquelme *et al.*, "Where are You? Exploring Micro-Location in Indoor Learning

Environments," *IEEE Access*, vol. 8, pp. 125776-125785, 2020, doi: 10.1109/access.2020.3008327.

[6] A. Weinberger and F. Fischer, "A framework to analyze argumentative knowledge construction in computer-supported collaborative learning," *Computers & Education*, vol. 46, no. 1, pp. 71-95, 2006, doi: 10.1016/j.compedu.2005.04.003.

[7] H. Cornide-Reyes *et al.*, "Introducing low-cost sensors into the classroom settings: Improving the assessment in agile practices with multimodal learning analytics," *Sensors*, vol. 19, no. 15, Art. no. 3291, 2019, doi: 10.3390/s19153291.

[8] F. Riquelme, R. Munoz, R. Mac Lean, R. Villarroel, T.S. Barcelos, and V.H.C. De Albuquerque, "Using multimodal learning analytics to study collaboration on discussion groups: A social network approach," *Universal Access in the Information Society*, vol. 18, no. 3, pp. 633-643, 2019, doi: 10.1007/s10209-019-00683-w.

[9] R. Noel *et al.*, "Exploring collaborative writing of user stories with multimodal learning analytics: A case study on a software engineering course," *IEEE Access*, vol. 6, pp. 67783-67798, 2018, doi: 10.1109/access.2018.2876801.

[10] H. Cornide-Reyes *et al.*, "A multimodal real-time feedback platform based on spoken interactions for remote active learning support," *Sensors*, vol. 20, no. 21, Art. no. 6337, 2020, doi: 10.3390/s20216337.

[11] D. Monsalves, H. Cornide-Reyes, and F. Riquelme, "Relationships between social interactions and Belbin role types in collaborative agile teams," *IEEE Access*, vol. 11, pp. 17002-17020, 2023, doi: 10.1109/access.2023.3245325.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, Art. no. 7553, pp. 436-444, 2015, doi: 10.1038/nature14539.

[13] A.Q. Ohi, M.F. Mridha, M.A. Hamid, and M.M. Monowar, "Deep speaker recognition: process, progress, and challenges," *IEEE Access*, vol. 9, pp. 89619-89643, 2021, doi: 10.1109/access.2021.3090109.

[14] B. Alkhatib and M.M.W. Kamal Eddin, "Voice identification using MFCC and vector quantization," *Baghdad Science Journal*, vol. 17, no. 3, 2020, doi: 10.21123/bsj.2020.17.3(suppl.).1019.

[15] S.A. Viswanathan and K. VanLehn, "Using the tablet gestures and speech of pairs of students to classify their collaboration," *IEEE Transactions on Learning Technologies*, vol. 11, no. 2, pp. 230-242, 2018, doi: 10.1109/tlt.2017.2704099.

[16] 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 & Language*, vol. 72, Art. no. 101317, 2022, doi: 10.1016/j.csl.2021.101317.

[17] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Information Processing & Management*, vol. 59, no. 6, Art. no. 103085, 2022, doi: 10.1016/j.ipm.2022.103085.

[18] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM based similarity measurement with spectral clustering for speaker diarization," in *Interspeech 2019*, Graz, Austria, Sep.15-19, 2019, doi: 10.21437/interspeech. 2019-1388.

[19] S. Watanabe *et al.*, "ESPnet: end-to-end speech processing toolkit," 2018, *arXiv*: 1804.00015.

[20] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," 2018, *arXiv* :1811.07453.

[21] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235-245, 2019, doi: 10.2478/jaiscr-2019-0006.

[22] M. Olivieri *et al.*, "Real-time multichannel speech separation and enhancement using a beamspace-domain-based lightweight CNN," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 04-10, 2023, doi: 10.1109/icassp49357.2023.10096891.

[23] L. Chen, Y. Dai, B. La, M. Sun, and Z. Xiong, "The key technology of speech interaction based on deep learning," *Journal of Physics: Conference Series*,

vol. 1087, no. 6, Art. no. 062042, 2018, doi: 10.1088/1742-6596/1087/6/062042.

[24] C. Li *et al.*, "Deep speaker: an end-to-end neural speaker embedding system," 2017, *arXiv*: 1705.02304.

[25] F. Bahmaninezhad, C. Zhang, and J.H. L. Hansen, "An investigation of domain adaptation in speaker embedding space for speaker recognition," *Speech Communication*, vol. 129, pp. 7-16, 2021, doi: 10.1016/j.specom.2021.01.001.

[26] Y. Luo, Z. Chen, J.R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 05-09, 2017, pp. 61-65, doi: 10.1109/icassp.2017.7952118.

[27] C.M. Yuan, X.M. Sun, and H. Zhao, "Speech separation using convolutional neural network and attention mechanism," *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1-10, 2020, doi: 10.1155/2020/2196893.

[28] C.C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 15-20, 2018, pp. 4774-4778, doi: 10.1109/icassp.2018.8462105.

[29] X. Xu and N. Zhang, "Characterization and prediction of performance interference on mediated passthrough GPUs for interference-aware scheduler," in *11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19)*, Renton, WA, USA, Jul. 2019. [Online]. Available: https://www.usenix.org/conference/hotcloud19/presentation/xu-xin

[30] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," 2016, *arXiv*: 1508.04395.

[31] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 15-20, 2018, pp. 5329-5333, doi: 10.1109/icassp.2018.8461375.

[32] G. Sell *et al.*, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Interspeech 2018*, Hyderabad, India, Sep. 02-06, 2018, doi: 10.21437/interspeech.2018-1893.

[33] B. Garcia-Ortega, M.Á. López-Navarro, and J. Galan-Cubillo, "Top management support in the implementation of industry 4.0 and business digitization: The case of companies in the main european stock indices," *IEEE Access*, vol. 9, pp. 139994-140007, 2021, doi: 10.1109/ACCESS.2021.3118988.

[34] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to VoxCeleb speaker recognition challenge 2019," 2019, *arXiv:* 1910.12592.

[35] B. Nguyen, T. Tran, T. Nguyen, and G. Nguyen, "An improved sea lion optimization for workload elasticity prediction with neural networks," *International Journal of Computational Intelligence Systems*, vol. 15, Art. no. 90, 2022, doi: 10.1007/s44196-022-00156-8.

[36] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333-339, 2019, doi: 10.1016/j.jbusres.2019.07.039.