Combined CNN-LSTM for Enhancing Clean and Noisy Speech Recognition

Noussaiba Djeffal*
Speech and Signal Processing Laboratory University of Sciences and Technology, USTHB Algiers, Algeria
Email: ndjeffal@usthb.dz

Djamel Addou
Speech and Signal Processing Laboratory University of Sciences and Technology, USTHB Algiers, Algeria
Email: daddou@usthb.dz

Hamza Kheddar
LSEA Laboratory, dept. Electrical engineering University of MEDEA Medea, Algeria
Email: kheddar.hamza@univ-medea.dz

Sid Ahmed Selouani
Research Laboratory in Human-System InteractionUniversity of Moncton, Shippagan Campus Shippagan, Canada
Email:sid-ahmed.selouani@umoncton.ca

Received: 18/11/2024 **Accepted:** 03/12/2024 **Published:**30/12/2024

Abstract:

This paper presents a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) approach for Automatic Speech Recognition (ASR) using deep learning techniques on the Aurora-2 dataset. The dataset includes both clean and multi-condition modes, encompassing four noise scenarios : subway, babble, car, and exhibition hall, each evaluated at different signal-to-noise ratios (SNRs), and clean condition, and the results are compared with those from the ASC-10 dataset and the ESC-10 dataset. The problem addressed is the need for robust ASR models that perform well in both clean and noisy environments. The aim of utilizing the CNN-LSTM architecture is to enhance the recognition performance by combining the strengths of CNNs and LSTMs, rather than relying on either CNNs or LSTMs alone. Experimental results demonstrate that the combined CNN-LSTM model achieves superior classification performance, in clean environments on the Aurora2 dataset, attaining an accuracy of 97.96%, surpassing the individual CNN and LSTM models, which achieved 97.21% and 96.06%, respectively. In noisy conditions, the hybrid model also outperforms the standalone models, with an accuracy of 90.72%, compared to 90.12% for CNN and 86.12% for LSTM. These findings indicate that the CNN-LSTM model is more effective in handling various noise conditions and improving overall ASR accuracy.

Keywords: ASR - CNN - LSTM - Clean speech - Noisy speech - CNN-LSTM - DNN - SNR.

***Corresponding author: Noussaiba Djeffal

دمج تقنية CNN-LSTM لتحسين التعرف على الكلام النظيف والصاخب

الملخص:

تخص هذه الورقة البحثية نهجًا هجينًا للشبكة العصبية التلافيفية والذاكرة طويلة المدى القصيرة (CNN-LSTM) للتعرف الآلي على الكلام (ASR) باستخدام تقنيات التعلم العميق على قاعدة بيانات Aurora-2 . تتضمن قاعدة البيانات كلًا من الوضعين النظيف ومتعدد الشروط، وتشمل أربعة سيناريوهات للضوضاء: مترو الأنفاق، والثرثرة، والسيارة، وقاعة المعرض، يتم تقييم كل منها عند نسب إشارة إلى ضوضاء (SNRs) مختلفة، وحالة نظيفة لا تحتوي على ضوضاء، ويتم مقارنة النتائج بتلك الموجودة في قاعدة البيانات ASC-10 وقاعدة البيانات ESC-10 . المشكلة التي تمت معالجتها في هذه الدراسة هي الحاجة إلى نماذج قوية للتعرف الآلي على الكلام التي تعمل بشكل جيد في كل من البيئات النظيفة والصاخبة (التي تحتوي على ضوضاء). والهدف من استخدام بنية CNN-LSTM هو تحسين أداء التعرف من خلال الجمع بين نقاط القوة في كل من النماذج CNNs وLSTMs، بدلاً من الاعتماد على CNNs أو LSTMs وحدها. تُظهِر النتائج التجريبية أن نموذج CNN-LSTM المدمج يحقق أداء تصنيف مرتفع جدا، في البيئات النظيفة على قاعدة البيانات Aurora2 ، حيث حقق نسبة 97.96% من الدقة، متجاوزًا نماذج CNN وLSTM عندما تأخذ فرديا، والتي حققت 97.21% و96.06% على التوالي. في الظروف الصاخبة، يتفوق النموذج الهجين أيضًا على النماذج المستقلة، بدقة 90.72%، مقارنة بـ 90.12% لـ CNN و86.12% لـ LSTM في النهاية، تشير هذه النتائج إلى أن نموذج CNN-LSTM أكثر فعالية في التعامل مع ظروف الضوضاء المختلفة وتحسين دقة التعرف على الكلام بشكل عام.

كلمات مفتاحية: التعرف الآلي على الكلام - CNN - LSTM - الكلام النظيف - الكلام الصاخب - CNN -LSTM - DNN - SNR.

Combinaison CNN-LSTM combiné pour améliorer la reconnaissance vocale propre et bruyante

Résumé :
Cet article présente une approche hybride de réseau neuronal convolutionnel et de mémoire à long terme (CNN-LSTM) pour la reconnaissance automatique de la parole (ASR) utilisant des techniques d'apprentissage profond sur la base de données Aurora-2. Cette base de données comprend des modes propres et multi-conditions, englobant quatre scénarios de bruit : métro, babillage, voiture et hall d'exposition, chacun évalué à différents rapports signal/bruit (SNR) et condition propre, et les résultats sont comparés à ceux de l'ensemble de données ASC-10 et de la base de données ESC-10. Le problème abordé est le besoin de modèles ASR robustes qui fonctionnent bien dans les environnements bruités et non bruités (propres). L'objectif de l'utilisation de l'architecture CNN-LSTM est d'améliorer les performances de reconnaissance en combinant les points forts des CNN et des LSTM, plutôt que de s'appuyer uniquement sur les CNN ou les LSTM pris en isolés. Les résultats expérimentaux démontrent que le modèle combiné CNN-LSTM atteint de hautes performances de classification, dans des environnements non bruités sur l'ensemble de données Aurora2, atteignant une précision de 97,96 %, surpassant les modèles CNN et LSTM pris individuellement, qui ont atteint respectivement 97,21 % et 96,06 %. Dans des conditions bruitées, le modèle hybride surpasse également les deux modèles cités, avec une précision de 90,72 %, contre 90,12 % pour CNN et 86,12 % pour LSTM. Ces résultats indiquent que le modèle hybride CNN-LSTM est plus efficace pour gérer diverses conditions de bruit et améliorer la précision globale du taux de reconnaissance de la parole.

Mots clés: ASR - CNN - LSTM - Parole propre - Parole bruitée - CNN-LSTM - DNN - SNR.

INTRODUCTION

Automatic speech recognition (ASR) systems convert spo- ken language into text, playing a crucial role in various applications. As the need for ASR services rises, it is essential to address the challenges of noise and distortion, which can greatly affect the performance of these systems (Djeffal et al., 2023). Deep learning (DL) technologies, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) net- works (Selouani & Yacoub, 2018), have made significant strides in enhancing ASR capabilities. These models excel in processing and interpreting human speech, offering improved performance over traditional methods. CNN effectively capture structural locality within the feature space and mitigate translational variance (Mazari & Kheddar, 2023). They manage disturbances and minor shifts in the feature space through pooling over local frequency regions, taking advantage of long-term dependencies among speech frames by utilizing prior knowledge of speech signals. However, it has been observed that CNNs encounter difficulties when dealing with semi-clean data in ASR systems, leading to performance degradation. In contrast, RNN capture long contexts and offer higher recognition accuracy, particularly for noise-robust tasks. However, RNNs face limitations due to the vanishing and exploding gradient problem, which impede their capability to learn temporal dependencies. To address these issues, long short-term memory (LSTM) networks were developed, featuring memory blocks that control the flow of information (Kamilya & Pappachen, 2019). LSTM-RNNs, however, tend to be sensitive to static data, leading to delays between inputs and corresponding outputs. For acoustic modeling, low latency is preferred, consequently, an architecture that processes input sequences in both directions. known as bidirectional LSTM (BLSTM), was introduced to make more informed decisions. Deep LSTM-RNNs are created by stacking multiple LSTM-RNN (Kheddar et al., 2024). This deep structure allows for optimal parameter utilization by distributing them across multiple layers. As a result, it have demonstrated impressive performance in large vocabulary speech recognition tasks (Passricha & Aggarwal, 2019) CNN-LSTM architectures combine the strengths of CNNs and LSTMs (Dar & Pushparaj, 2024). CNNs excel in extracting spatial features from input sequences, while LSTMs are adept at capturing contextual dependencies and temporal dynamics over time. This integration proves invaluable in tasks requiring a comprehensive understanding of both local features and long- term relationships, such as video analysis, medical signal processing, and ASR in noisy environments.

The collaboration between CNNs and LSTMs has driven significant advancements across diverse fields (Wang et al., 2020), enhancing the accuracy and robustness of predictions by effectively learn- ing from intricate sequential data. This introduction lays the groundwork for exploring the design, training, and application of CNN-LSTM architectures in various domains, addressing complex challenges in sequence learning tasks. This study aims to propose and evaluate CNN and LSTM models, as well as a hybrid CNN-LSTM architecture (Gueriani et al., 2024), in the aim to improve ASR performance in noisy environments. The experiments are conducted using the Aurora-2 dataset (Hirsh & Pearce., 2000), which includes various noise scenarios and signal-to-noise ratios (SNRs) (Naing et al., 2020) and are compared with the ASC-10 and ESC-10 datasets. Our contributions are structured as follows:

- Propose a CNN model in both clean and noisy modes.
- Propose an LSTM model in both clean and noisy modes.
- Compare the proposed CNN-LSTM hybrid model with individual CNN and LSTM models.
- Compare the Aurora-2 dataset with the ASC-10 and ESC-10 datasets.

1. BACKGROUND

Deep learning, a specialized area within machine learning, employs multi-layered neural networks to capture intricatepatterns in data. This approach has significantly contributedto progress in various domains, including image and speech recognition, natural language processing, and autonomous systems. Among the core architectures in deep learning are CNNs, Deep Neural Networks (DNNs) (Exter & Meyer, 2016), and LSTM networks. CNNs are a cornerstone of deep learning, particularly excellingin tasks such as image and audio recognition, where data hasa structured form. Unlike traditional fully connected networks,CNNs are built with convolutional and pooling layers, which are stacked sequentially (Kheddar et al., 2024). The convolutional layers apply filters to the input, each acting as a feature detector that iden- tifies patterns like edges, textures, and shapes. This process produces feature maps, capturing the spatial hierarchies in the data. By using shared weights for these filters, CNNs reduce the number of parameters, improving computational efficiencyand mitigating overfitting. The shared-weight structure also allows the network to recognize patterns across different loca-tions, making it more robust. Pooling layers, which typically follow convolutional layers, reduce the dimensionality of the feature maps through down-sampling. Techniques like max pooling and average pooling are common, helping retain im- portant information while discarding less relevant details. Thisdimensionality reduction serves several purposes, it lowerscomputational complexity, enhances the model's ability togeneralize to new data, and reduces overfitting by emphasizingkey features. Together, the convolutional and pooling layers allow CNNs to progressively capture spatial features, from basic elements like edges to complex structures, making them especially effective in domains like image and speech recognition (Djeffal et al., 2023).

In contrast, LSTM networks are designed for sequential data, excelling in capturing temporal relationships (Li et al., 2015). LSTMs introduce three types of input gates, output gates, and forget gates that regulate the flow of information, allowingthe network to decide which information to remember or discard. This mechanism enables LSTMs to capture both short-term and long-term dependencies, which is essential for tasks like speech and language processing, where the meaning of words or sounds depends on previous context (Greg et al., 2020). LSTMs can handle sequences of varying lengths while maintaining a consistent input-output size, making them highly versatile for time-dependent tasks. However, traditional LSTMs are uni- directional, meaning they process data from past inputs only and ignore future context. This can limit their effectiveness for tasks that require understanding of context in both directions, such as certain pattern recognition tasks where bidirectional context is valuable. The unidirectional nature of LSTMs can also increase their susceptibility to overfitting, especially in complex or noisy environments where subtle temporal patternsneed to be captured and maintained (Passricha & Aggarwal, 2019).

To overcome these limitations, CNNs and LSTMs are oftencombined to leverage the strengths of both. CNNs excel at extracting spatial features from raw data, which is particularly useful for tasks involving images and sound, while LSTMsare adept at modeling temporal dependencies in sequences. The CNN layers process the input to generate feature maps that highlight spatial patterns, which are then passed to the LSTM layers to model temporal dependencies. This hybrid CNN-LSTM architecture (Passricha & Aggarwal, 2019) is effective at capturing intricate spatiotemporal patterns that individual models struggle to capture on their own. This approach has been successful in fields such as video analysis, speech recognition, and sensor data processing,

where both spatial and temporal information are crucial for accurate predictions.

In our study, we developed a CNN-LSTM model tailoredto improve recognition performance in noisy environments. The CNN component serves as a feature extractor, identifying local patterns within each audio frame. These features are thenfed into the LSTM component, which captures the temporal relationships between frames and models how audio features evolve over time. By combining CNNs ability to extract patterns with LSTMs capacity for sequential data processing, this architecture is particularly well-suited for tasks like ASR, where both spatial (spectral features) and temporal (phonetic context) factors are critical. Furthermore, integrating CNNand LSTM layers (Wu al., 2018) allows the model to retain long-term dependencies while maintaining short-term pattern recognition. This hybrid approach enhances both the accuracy and robustness of the model, particularly in noisy environments where the model must filter out irrelevant information. CNN- LSTM hybrid models (Daouad et al., 2023) have been shown to achieve significant improvements in various domains, effectively balancing feature extraction with sequence modeling. As a result, this architecture enhances the model's ability to generalize well to different environments, a crucial trait for real-world ASR applications (Xie et al., 2020). The development of hybrid models such as CNN-LSTM represents a significant step forward in tasks thatrequire spatiotemporal understanding. By combining CNNs ability to detect local patterns with LSTMs ability to manage temporal dependencies, this model offers a comprehensivesolution for ASR in noisy settings. This study highlights the potential of hybrid models to improve recognition accuracy while adapting to challenging, real-world conditions, providing a foundation for further advances in noise-robust ASR systems.

2. RELATED LITERATURE

The combination of CNN and LSTM has attracted significant interest from researchers. In (Passricha & Aggarwal, 2019) the authors proposea hybrid CNN-BLSTM architecture for acoustic modeling in ASR to leverage both spectral and temporal properties of speech signals, improving continuous speech recognition. They explore methods such as weight sharing, optimizing the number of hidden units, pooling strategies, and the effectiveness of BLSTM layers. Additionally, they address the limitation of CNNs in modeling speaker-adapted features and examine various non-linearities with dropout. Their experiments demonstrate a 5.8% and 10% relative decrease in WER over CNN and DNN systems, respectively, when incorporating speaker-adapted features and maxout non-linearity with dropout. However, our study contributes to the discussionby specifically tackling the challenges presented by noisy environments in ASR. We employ a hybrid CNN-LSTM modeldesigned for various noise scenarios, combined with robust performance metrics in both clean and noisy settings. This approach provides a more thorough understanding of ASR. While the authors concentrate on optimizing acoustic modeling architectures, our research expands on this by exploring thereal-world implications of techniques like weight sharing and pooling strategies. This highlights the necessity of adapting models to different acoustic conditions. Additionally, whereas the authors demonstrate improvements in word error rate (WER) through the use of speaker-adapted features, our resultsreveal substantial performance gains in both clean and noisy environments, emphasizing the robustness of our model. This broader perspective not only showcases the effectiveness of our approach but also underscores the importance of resiliencein ASR systems, distinguishing our study in the field.

Our study and the research presented in (Alsayadi et al., 2021) both employ CNN-LSTM architectures to enhance ASR performance, yet they differ significantly in focus and approach.

The study in (Alsayadi et al., 2021) targets Arabic speech recognition, concentrating on the impact of diacritics on accuracy. Utilizing CNN-LSTM and attention-based techniques within the Espresso framework, it achieves a 13.52% reduction in WER and improves language model performance when trained on non-diacritized data fromthe SASSC corpus. In contrast, our work prioritizes robustnessto environmental noise instead of linguistic features. We rigorously evaluate our CNN-LSTM model on the Aurora-2 dataset, which includes clean speech as well as multiple noisy scenarios, such as subway, babble, car, and exhibition hall noises at various signal-to-noise ratios. Our results demonstrate that our hybrid model excels in noisy conditions, high- lighting its effectiveness in real-world applications compared to models designed for clean or language-specific contexts. This comparison underscores the unique contributions of each study, while (Alsayadi et al., 2021) focuses on resolving linguistic challenges in Arabic ASR, our research emphasizes the model's resilienceto diverse environmental noise.

In (Dat et al., 2020), the authors propose a hybrid CNN-BLSTM model with an attention mechanism for Vietnamese speech recognition in noisy operating room environments. This architecture combines CTC and attention loss functions to enhance alignment accuracy and accelerate label sequence estimation during both training and inference. Evaluation on real surgery room data shows a notable 13.05% reduction in WER indicatessubstantial improvements in ASR system performance. Our study demonstrates several significant improvements over the research presented in (Dat et al., 2020). While both studies utilize hybrid neural network architectures to enhance ASR performancein noisy environments, our work evaluates a broader rangeof noise conditions. The study in (Dat et al., 2020) specifically targets Vietnamese speech recognition in noisy operating room environments. In contrast, our research showcases enhanced adaptability and resilience in ASR systems by addressing various noise scenarios. Overall, our findings indicate a more comprehensive approach to tackling noise challenges com-pared to the more specialized context explored in (Dat et al., 2020).

3. PROPOSED APPROACH

The proposed approach for ASR is combines CNNs and LSTM networks, leveraging their respective strengths to process and classify speech data effectively. The given neural network architecture starts with a sequential model, indicatingthat layers will be added one by one in sequence. The model begins with a convolutional layer, which applies 64 filters witha kernel size of 3 to the input data. This layer uses the ReLU activation function and expects input data shaped according tothe dimensions of the training set.

Following the first convolutional layer is a BatchNormalization layer that standardizes the inputs to a layer for each mini- batch, stabilizing and speeding up the learning process. Nextis a MaxPooling layer with a pool size of 2, which reducesthe dimensionality of the data and helps to capture dominant features. A dropout layer with a dropout rate of 0.3 is then applied to prevent overfitting by randomly setting 30% of the input units to zero during training. This pattern is repeated with a second convolution layer that applies 128 filters, followed by another BatchNormalization, MaxPooling, and dropout layer with the same configuration. A third Convolution layer then applies 256 filters, followed by similar normalization, pooling,and dropout layers.

After that, the model uses a Flatten operation to each time step in the sequence independently. This is followed by an LSTM layer with 64 units that returns sequences, allowingthe model to capture temporal dependencies in the data. A dropout layer with a 0.5 rate is applied to this

LSTM layerto further prevent overfitting. Next, another LSTM layer with 64 units is added, which does not return sequences, followed by another dropout layer with a 0.5 rate. The model thenincludes a dense layer with 128 units and a ReLU activation function, adding fully connected layers to further process the features. Another dropout layer with a 0.5 rate is applied.Finally, the model ends with a dense output layer with anumber of units equal to the number of classes and uses the softmax activation function to produce probability distributions over the classes. This configuration allows the model to make multiclass predictions based on the input data.

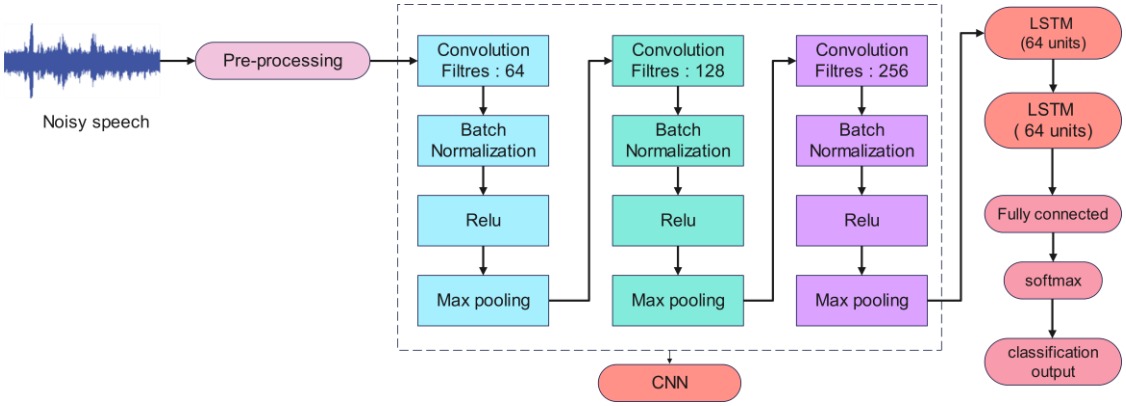

Figure 1. The proposed CNN-LSTM architecture for efficient ASR applied to noisy dataset.

This hybrid CNN-LSTM approach is designed to improve ASR accuracy by effectively combining spatial feature ex- traction capabilities from CNNs with LSTM's proficiency in handling sequential data dependencies. This architecture is particularly suited for real-world applications where speech recognition must perform reliably in diverse and noisy environments. CNN-LSTM architecture is illustrated in Figure 1.

4. EXPERIMENTS

A. Dataset exploration

- ***AURORA-2 database:*** Is developed by the European telecommunications standards institute (ETSI), tailored for assessing robust feature extraction within a distributed recognition framework. This dataset builds on the TIDig-its dataset, originally in English and accessible through the linguistic data consortium (LDC). It includes artificially added noise signals to the clean speech data to evaluate performance under various noisy conditions (Hirsh & Pearce, 2000), AURORA-2 (Soe Naing et al., 2020) supports two training modes. Summary of detail data usage was described in Table I.
- ***ASC dataset:*** This dataset is an Arabic Command Speech multiple examples of every command. This dataset captures a diverse array of samples, reflecting various ages, genders, and slight accent differences typical of Syrian Arabic. Importantly, all pronunciations follow Standard Arabic rather than regional dialects, providing consistent pronunciation across recordings. With a size of approximately 384 MB, the dataset is compact yet rich in diversity, offering ASR models ample variability to learnreliable representations of Arabic commands.

This datasetis valuable for researchers and developers focused onArabic voice recognition and command-based applications, especially those requiring models that generalize well across speaker age and accent variations.

- ***ESC-50 dataset:*** The dataset is divided into five main categories: animal sounds, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noises. The ESC-50 dataset
(Takazawa et al., 2024) includes 2,000 audio files across 50 classes, with each class containing 40 audio samples (Karam et al., 2023). The ESC-10 dataset is a subset of the ESC-50, comprising only 10 classes. Both ESC-50 and ESC-10 datasets are recordedat a sampling rate of 44.1 kHz, and each audio file is 5 seconds long (Demir et al., 2020). These datasets provide a comprehensive resource for studying sound classification, especiallyin noisy environments, due to the diversity of sounds, including urban and outdoor noises.

B. Performances metrics

- ***WER:*** measures the proportion of incorrect words inrelation to the total number of words processed (Kheddar et al., 2023). Is defined as follows:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{H + S + D} \quad (1)$$

The calculation is based on the counts of insertions (I), deletions (D), substitutions (S), hits (H), and total input words (N).

Accuracy, recall, precision: To evaluate the effectiveness of a proposed method in the field of ASR, assessment criteria such as classification accuracy, recall (sensitivity) and precision (positive predictive value), are commonly used (Habchi et al., 2023). The aforementioned metrics can be expressed as follow:

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN}.100 \quad (2)$$

$$Recall(\%) = \frac{TP}{TP + FN}.100 \quad (3)$$

$$Precision(\%) = \frac{TP}{TP + FP}.100 \quad (4)$$

TABLE I. THE SUMMARY OF AURORA-2 DATABASE

Category	Description	
Vocabulary	continuous digits sequences (0-9, ’oh’)	
Sampling	44.1 KHZ, 16 bits, mono	
Male	111 speakers	21-70 ages
Female	114 speakers	17-59 ages
Training	8,440 utterances	Multi-condition
	subway, babble, car, exhibition hall	

	20db, 15db, 10db, 5db and clean
Testing	1,001 utterances
	subway, babble, car, exhibition hall
	10db, 5db, 0db, -5db

- ***F1-score :*** is a preferred metric over accuracy when FN and FP are of significant importance. Additionally, in thepresence of imbalanced class distributions, the F1 scoreis a more appropriate metric for evaluating ASR models (Kheddar et al., 2023). F1-score can be calculated using equation 5:

$$\text{F1- score (\%)} = \frac{2.\text{Precision.Recall}}{\text{Precision + Recall}}.100 \quad (5)$$

Analysis of results

Experiment 1*:* This study utilizes the Aurora-2 dataset, which comprises 4,824 isolated digit recordings, equally di- vided into 2,412 files for clean and noisy modes. The dataset covers 11 classes, ranging from 0 to 'oh'. Approximately40% of the data was reserved for testing, using extractedMFCC features with 39 coefficients and the original audio sampling rate, label encoding was performed. The multi- condition mode features four noise scenarios (subway, babble,car, and exhibition hall) across five SNRs: 20dB, 15dB, 10dB,5dB, and clean conditions.

Experiments were conducted using a model compiled with the Adam optimizer, sparse categorical cross-entropy loss, and accuracy as the evaluation metric. Training was performed for 50 epochs with a batch size of 32. These settings yielded the recognition rates presented in Table II. The results high- light that the hybrid CNN-LSTM model significantly outper- forms the individual CNN, LSTM, and BiLSTM architectures. Specifically, the CNN-LSTM model achieved a recognition rate of 97.96% in clean mode, surpassing CNN, LSTM, DNN, and BiLSTM. In noisy environments, the CNN-LSTM model speech also demonstrated superior performance with a recognition rate of 90.72%, compared to other models. Table III shows the classification report for the CNN-LSTM model, while Figure 3 presents the confusion matrix for multiclass classification in clean speech. In noisy speech, the classification report using the CNN-LSTM model is in Table IV, and the corresponding confusion matrix is illustrated in Figure 4.

The WER recognition rates in clean and noisy environ- ments, depicted in Figure 2, emphasize the low WER of the hybrid CNN-LSTM model compared to the CNN, LSTM, and BiLSTM models in both conditions. In clean condition the CNN-LSTM model achieved a recognition rate with a 2.1% reduction in WER compared to the CNN, LSTM, and BiLSTM models. In noisy environments, the WER for the CNN-LSTM model is 9.3%, which is a reduction compared to the CNN, LSTM, and BiLSTM models. Figures 5 and 6 illustrate the model's accuracy and loss for both training and validation sets over 50 epochs, using clean speech data from the Aurora2 dataset. In Figure 5 (model accuracy), the training accuracy starts around 0.6 (60%) and increases rapidly. After approximately 10 epochs, the training and validation accuracy curves converge, both reaching a near-perfect accuracy close to 1.0 (100%). This steady alignment between training and validation accuracy indicates effective learning and general- ization, with no signs of overfitting. In Figure 6 (model loss), the loss for both training and validation begins above 2.0 and decreases sharply as the model trains. Around the 10th epoch, the curves converge and stabilize around 0.1 for both training and validation, reflecting strong convergence. The low, stable loss shows that the

model effectively reduces prediction errors, even on the validation data. These results demonstrate that the CNN-LSTM model achieves nearly 100% accuracy and main- tains a low loss (around 0.1), indicating robust performance in recognizing clean speech on the Aurora2 dataset.

TABLE II. ACCURACY FOR MODELS TESTED IN CLEAN AND MULTI-CONDITION TRAINING AURORA2 DATASET.

Models tested	Clean speech	Noisy speech
CNN	97.21%	90.12%
LSTM	96.06%	86.12%
BiLSTM	94.33%	83.43%
DNN	80.04%	50.66%
Concatenate CNN-LSTM	**97.96%**	**90.72%**

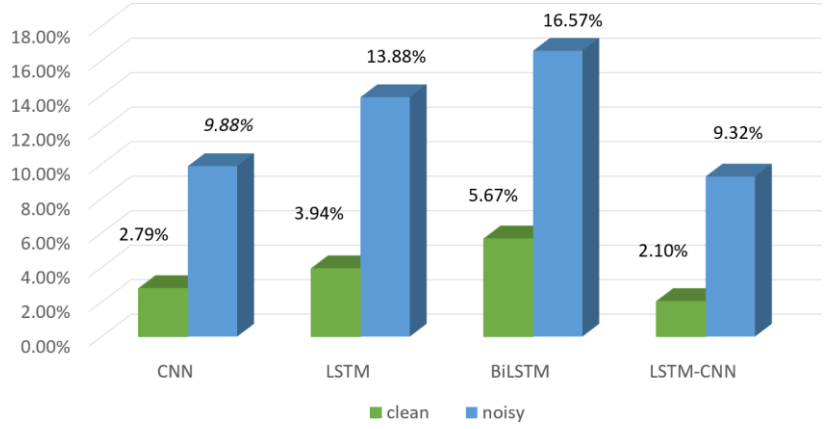

Figure. 2. WER (%) recognition performance achieved by CNN, LSTM, BiLSTM, and CNN-LSTM models in both clean and noisy environments Aurora2 dataset.

Figure. 3. Confusion matrix of multi class classification using a combined CNN- LSTM model in clean speech (Aurora2 dataset).

TABLE III. CLASSIFICATION REPORT OF CLEAN SPEECH USING CNN- LSTM (AURORA2 DATASET)

	Precision	Recall	F1-score	Support
eight	95%	98%	96%	42
five	100%	100%	100%	50
four	98%	98%	98%	52
nine	97%	97%	97%	38
oh	98%	98%	98%	46
one	93%	98%	95%	43
seven	100%	94%	97%	49
six	100%	98%	99%	43
three	96%	98%	97%	47
two	97%	95%	96%	39
zero	97%	100%	99%	34
Accuracy			98%	483
Macro avg	97%	98%	97%	483
Weighted avg	98%	98%	98%	483
Average WER	0.021			

Experiment 2: In this study, we use a specialized subset of the ASC dataset (Lichouri et al., 2023), known as ASC-10, which is tailored for essential command recognition tasks. The ASC-10 dataset has been simplified to include only 10 distinct command classes, making it streamlined and efficient for ASR research. With a total of 3,000 audio files 300 unique samples per command this

balanced structure ensures no command is over- or under-represented, providing a solid foundation for training models with consistent accuracy across all classes. The 10 commands in ASC-10 backward, cancel, close, left, move, next, no, ok, start, and stop were selected for their frequent use in voice-activated applications like virtual assistants, navigation systems, and device controls. These keywords, while simple, cover a broad range of actions necessary for interactive voice-response systems. The ASC-10 dataset is not only compact and computationally efficient but also rich in command variety, making it a valuable resource for developing ASR models that need reliable performance with essential command inputs. Its structured design offers an effective basis for evaluating and refining command recognition, particularly in applications requiring fast, accurate responses to a core set of commonly used commands. The recognition rates are provided in Table V, which clearly show that the hybrid CNN-LSTM model significantly outperforms the individual CNN, LSTM, and BiLSTM architectures. In particular, the CNN-LSTM model achieved a 98.02% recognition rate in clean conditions, surpassing the performance of CNN, LSTM, DNN, and BiLSTM models. A detailed classification report for the CNN-LSTM model found in Table VI, while Figure 8 displays the confusion matrix for multiclass classification under clean speech conditions.

Figure 4. Confusion matrix of multiclass classification using a combined CNN-LSTM model in noisy speech (Aurora2 dataset).

TABLE IV. CLASSIFICATION REPORT OF NOISY SPEECH USING CNN-LSTM (AURORA2 DATASET).

	Precision	Recall	F1-score	Support
eight	87%	93%	90%	42
five	92%	94%	93%	50
four	87%	88%	88%	52
nine	94%	84%	89%	38
oh	93%	91%	92%	46
one	86%	86%	86%	43
seven	88%	90%	89%	49
six	95%	91%	93%	43
three	92%	94%	93%	47
two	97%	95%	96%	39
zero	89%	91%	90%	34
Accuracy			91%	483
Macro avg	91%	91%	91%	483
Weighted avg	91%	91%	91%	483
Average WER	0.093			

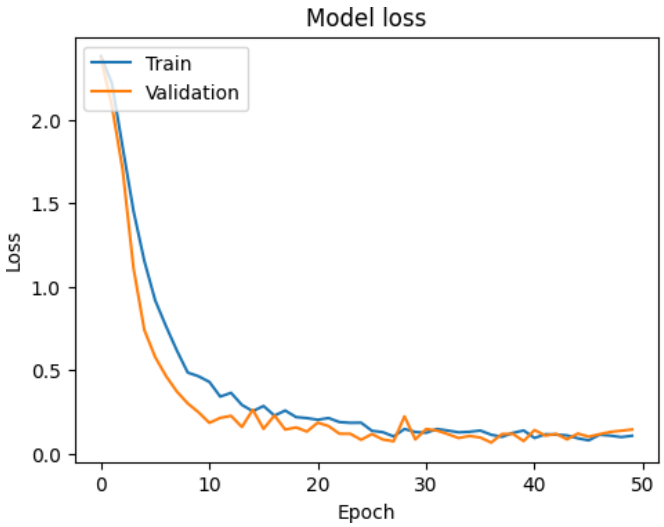

Figure 5. Model accuracy in clean speech (Aurora2 dataset).

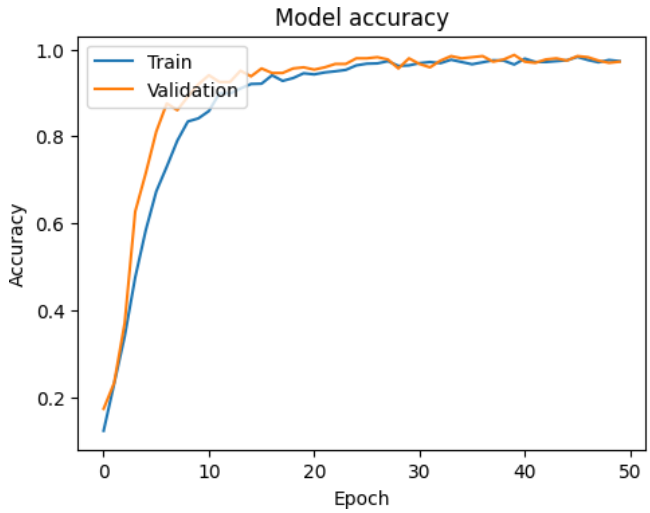

Figure 6. Model loss in clean speech (Aurora2 dataset).

TABLE V. ACCURACY FOR MODELS TESTED IN CLEAN SPEECH ASC-10 DATASET

Models tested	Clean speech
CNN	97.32%
LSTM	96.49%
BiLSTM	97.20%
DNN	73.16%
Concatenate CNN-LSTM	**98.02%**

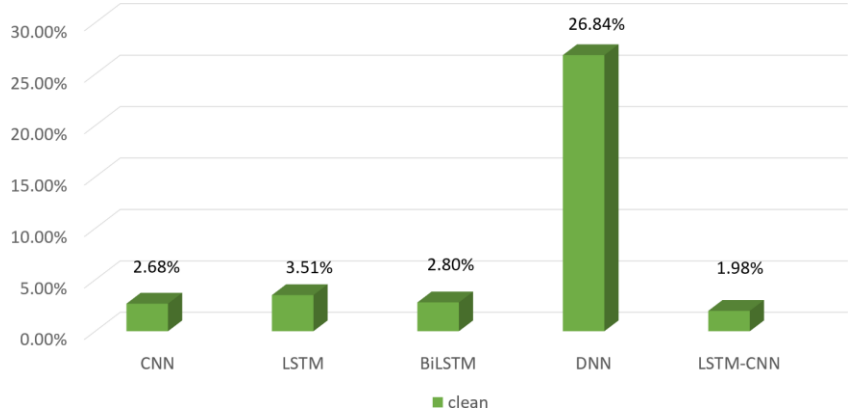

Figure 7. WER (%) recognition performance achieved by CNN, LSTM, BiLSTM, DNN and CNN-LSTM models in clean environments ASC-10 dataset.

The WER rates in clean environments, illustrated in Figure7, highlight the advantage of the hybrid CNN-LSTM model, which exhibits a lower WER compared to the CNN, LSTM, BiLSTM, and DNN models. Under clean conditions, the CNN-LSTM model achieved a 1.98% reduction in WER relativeto the other models. Figures 9 and 10 illustrate the model's performance in terms of accuracy and loss across the training and validation sets over 50 epochs on the ASC-10 dataset for clean speech.

In Figure 9 (model accuracy), the training accuracy starts around 20% and rises rapidly, reaching close to 100% after approximately 10 epochs. The validation accuracy follows a similar pattern, converging with the training accuracy around the 10th epoch and stabilizing near 100%. This alignmentbetween training and validation accuracy suggests that themodel is learning effectively and generalizing well, indicating that the model is not overfitting. In Figure 10 (model loss),the initial loss for both training and validation begins above 2.0 and decreases sharply during the first 10 epochs. After thissteep decline, the loss stabilizes around 0.1 for both sets. The low, stable loss values indicate that the model is successfully minimizing errors and maintaining strong performance on the validation data. These results show that the CNN-LSTM modelperforms well on the ASC-10 dataset, achieving near-perfect accuracy and a consistently low loss. This suggests that the model is robust and can generalize effectively to clean speech data.

Experiment 3: The dataset used is derived from the ESC-50 dataset (Piczak et al., 2015), reduced to 10 classes and containing 400 audio files in total, with 40 audio files per class. The classes in the ESC-10 dataset include dog bark, airplane, clapping, crow,door knock, chainsaw, can opening, fireworks, thunderstorm, and vacuum cleaner. As shown in Table VII, in noisy envi- ronments on the ESC-10 dataset, the results demonstrate that the hybrid CNN-LSTM model significantly outperforms the standalone CNN, LSTM, and BiLSTM architectures. Specifi- cally, the CNN-LSTM model achieved an 80.81% recognition rate in noisy conditions, surpassing the CNN, LSTM, DNN, and BiLSTM models. Table VIII provides the classification report for the CNN-LSTM model, while Figure 12 displays theconfusion matrix for multiclass classification in noisy speech.

Figure 8. Confusion matrix of multiclass classification using a combined CNN- LSTM model in clean speech (ASC-10 dataset).

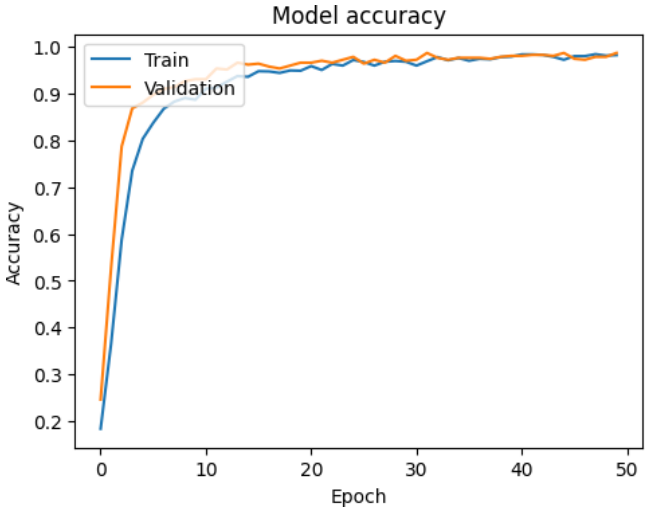

Figure 9. Model accuracy in clean speech (ASC-10 dataset)

TABLE VI. CLASSIFICATION REPORT OF CLEAN SPEECH USING CNN-LSTM (ASC-10 DATASET)

	Precision	Recall	F1-score	Support
backward	96%	100%	98%	72
cancel	97%	97%	97%	69
close	95%	100%	97%	57
left	98%	98%	98%	56
move	100%	95%	79%	59
next	100%	100%	100%	63
no	100%	96%	98%	56
ok	98%	98%	98%	55
start	98%	96%	97%	57
stop	98%	98%	98%	56
Accuracy			98%	600
Macro avg	98%	98%	98%	600
Weighted avg	98%	98%	98%	600
Average WER	0.02			

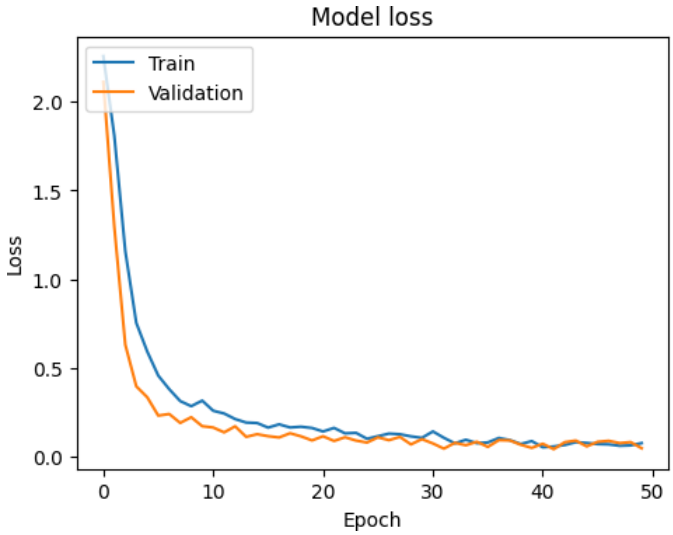

Figure 10. Model loss in clean speech (ASC-10 dataset).

The WER results in noisy environments, illustrated in Figure 11, highlight the low WER of the CNN-LSTM model compared to the CNN, LSTM, BiLSTM, and DNN models. Under noisy conditions, the CNN-LSTM model achieved a 19.19% reduction in WER relative to the other models.

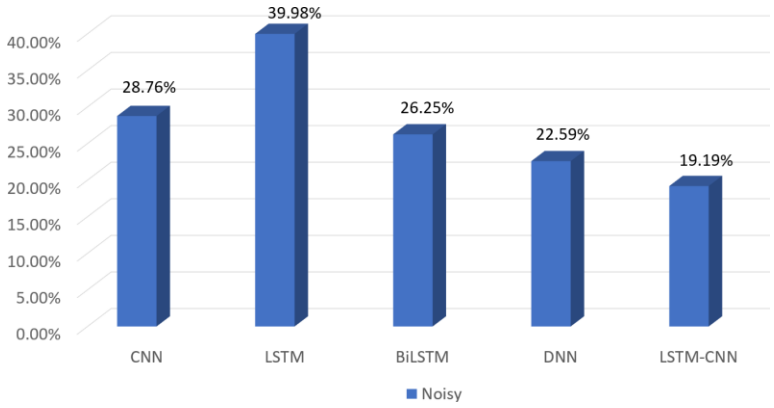

Figure 11. WER (%) recognition performance achieved by CNN, LSTM, BiLSTM, DNN and CNN-LSTM models in noisy environments ESC-10 dataset.

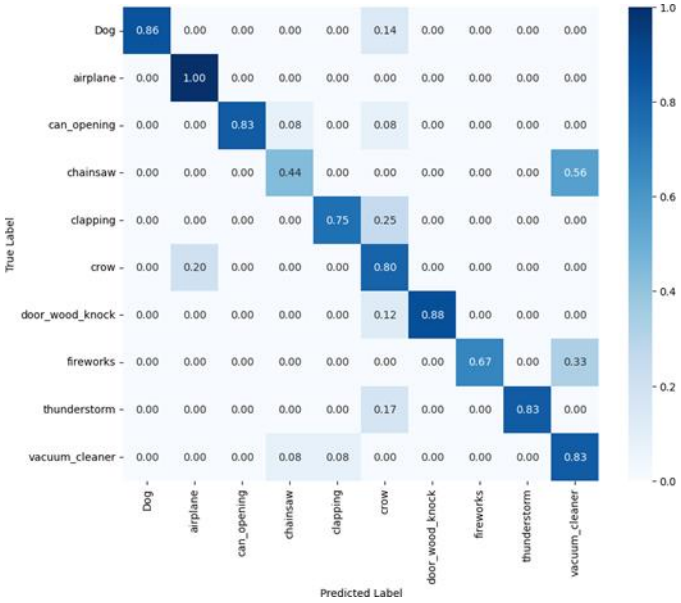

Figure 12. Confusion matrix of multiclass classification using a combined CNN-LSTM model in noisy speech (ESC-10 dataset).

TABLE VII. ACCURACY FOR MODELS TESTED IN NOISY SPEECH ESC-10 DATASET

Models tested	Noisy speech
CNN	71.24%
LSTM	60.02%
BiLSTM	73.75%
DNN	77.41%
Concatenate CNN-LSTM	**80.81**%

Figure13 presents a comparison of various model architectures CNN, LSTM, BiLSTM, DNN, and CNN-LSTM on the Aurora2 dataset in both clean and noisy environments, as well as on the ASC-10 clean and ESC-10 noisy datasets. The blue line represents the Aurora2 dataset in a clean environment, where all models achieve high accuracy, approaching 100%. In contrast, the orange line shows performance on the Aurora2 dataset under noisy conditions, where accuracy decreases across all models but remains relatively higher compared to the ESC-10 noisy dataset, demonstrating the models' robustness on the Aurora2 dataset even in challenging conditions. The ASC-10 dataset, represented by the gray line, is evaluated in a clean environment and achieves slightly lower accuracy than the Aurora2 clean data but remains relatively high, above 80%. In comparison, the ESC-10 dataset in a noisy environment, shown by the yellow line, is the most challenging, with all models, especially LSTM and BiLSTM, showing a notable drop in accuracy. Among the models, CNN-LSTM demonstrates particularly strong performance in both clean and noisy conditions. It performs well in clean conditions and shows resilience to noise, outperforming methods like LSTM and BiLSTM, which tend to struggle in noisy environments.

This makes CNN-LSTM a promising choice for applicationsrequiring consistent effectiveness across varying noise levels.We chose the Aurora2 dataset for this analysis because itperforms well in both clean and noisy conditions comparedto the ASC-10 clean and ESC-10 noisy datasets, making it avaluable dataset for assessing the robustness of models across different noise levels.

TABLE VIII. CLASSIFICATION REPORT OF NOISY SPEECH USING CNN-LSTM (ESC-10 DATASET).

	Precision	Recall	F1-score	Support
Dog	100%	86%	92%	7
aiplane	91%	100%	95%	10
can-opening	100%	83%	91%	12
chainsaw	67%	44%	53%	9
clapping	86%	75%	80%	8
crow	40%	80%	53%	5
door-wood-knock	100%	88%	93%	8
fireworks	100%	67%	80%	3
thunderstorm	100%	83%	91%	6
vacuum-cleaner	62%	83%	71%	12
Accuracy			80%	483
Macro avg	85%	79%	80%	80
Weighted avg	84%	80%	81%	80
Average WER	0.2			

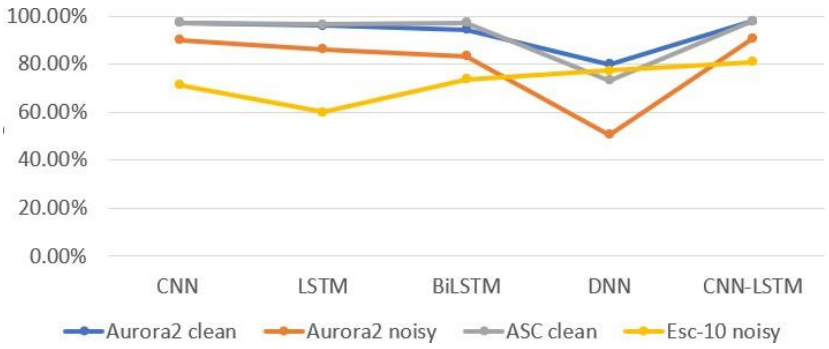

Figure 13. Performance Comparison of model architectures on Aurora2, ASC-10, and ESC-10 datasets in clean and noisy conditions.

5. CONCLUSION

In conclusion, our investigation into the combined CNN- LSTM model for speech recognition in both clean and noisy environments has demonstrated its superior performance over standalone CNN and LSTM models. The hybrid architecture effectively leverages the strengths of both CNNs and LSTMs, resulting in enhanced accuracy and robustness. These findings underscore the

potential of combining different neural net- work architectures to address complex challenges in speech recognition. For future work, several avenues can be explored to further enhance the performance and applicability of our model. Firstly, experimenting with different configurations of the CNN and LSTM components, such as varying the number of layers or units, could yield further improvements. Additionally, integrating other advanced techniques like attention mechanisms or Transformer models might provide better handling of long-range dependencies and contextual information. Moreover, applying the model to biomedical field such as cochlear implant (Essaid et al., 2024) would reduce the profound hearing loss disease. Finally, real-time implementation and optimization of the model for deployment in practical ASR systems would be an essential step towards its application in real-world scenarios.

References

Alsayadi, H.A. et al. (2021). Non-diacritized Arabic speech recognition based on CNN-LSTM and attention-based models. *Journal of Intelligent & Fuzzy Systems*, *41*(6), 6207-6219.

Daouad, M., Allah, F. A. & Dadi, E. W. (2023). An automatic speech recognition system for isolated Amazigh word using 1D & 2D CNN-LSTM architecture. *International Journal of Speech Technology*, *26*(3), 775-787.

Dar M.A. & J. Pushparaj, J. (2024). Hybrid architecture cnn-blstm for automatic speech recognition. *3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT)*. pp. 1–4.

Dat, T.T. et al. (2021). Convolutional recurrent neural network with attention for Vietnamese speech to text problem in the operating room. *International Journal of Intelligent Information and Database Systems*, *14*(3), 294-314.

Demir, F., Turkoglu, M., Aslan, M. & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. *Applied Acoustics*, *170*, 107520.

Djeffal, N. et al. (2023). Automatic speech recognition with BERT and CTC transformers: A review. In *2nd International Conference on Electronics, Energy and Measurement (IC2EM)* (Vol. 1, pp. 1-8).

Djeffal, N. et al. (2023). Noise-robust speech recognition: A comparative analysis of LSTM and CNN approaches. In *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM)* (Vol. 1, pp. 1-6). IEEE.

Greg V.H., Carlos M. & Gonzalo, N. (2020). A review on the long short- term memory model. *Artificial Intelligence Review*. Springer, 2020, pp. 5929–5955.

Gueriani, A., Kheddar, H. & Mazari, A. C. (2024). Enhancing IoT Security with CNN and LSTM-Based Intrusion Detection Systems. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (pp. 1-7). IEEE.

Essaid, B. et al. (2024). Artificial Intelligence for Cochlear Implants: Review of Strategies, Challenges, and Perspectives. *IEEE Access*.

Exter, M. & Meyer, B. T. (2016). DNN-Based Automatic Speech Recognition as a Model for Human Phoneme Perception. In *INTERSPEECH* (pp. 615-619).

Ghandoura, A., Hjabo, F. & Al Dakkak, O. (2021). Building and benchmarking an Arabic Speech Commands dataset for small-footprint keyword spotting. *Engineering Applications of Artificial Intelligence*, *102*, 104267.

Habchi, Y. et al. (2023). Ai in thyroid cancer diagnosis: Techniques, trends, and future directions. *Systems*, *11*(10), 519.

Hirsch, H. G. & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW*H.

Karam, S. et al. (2023). Episodic memory based continual learning without catastrophic forgetting for environmental sound classification. *Journal of Ambient Intelligence and Humanized Computing*, *14*(4), 4439-4449.

Kheddar, H. et al. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems*, *277*, 110851.

Kheddar, H. et al. (2024). Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions. *Neurocomputing*, 127528.

Kheddar, H., Hemis, M. & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 102422.

Li, J., Mohamed, A., Zweig, G. & Gong, Y. (2015). LSTM time and frequency recurrence for automatic speech recognition. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)* (pp. 187-191). IEEE.

Lichouri, M., Lounnas, K. & Bakri, A. (2023). Toward building another arabic voice command dataset for multiple speech processing tasks. In *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)* (pp. 1-5). IEEE.

Mazari, A. C. & Kheddar, H. (2023). Deep learning-based analysis of Algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital Systems*..

Naing, H.M.S., Hidayat, R., Hartanto, R. & Miyanaga, Y. (2020, November). A front-end technique for automatic noisy speech recognition. *23rd conference of the oriental COCOSDA international committee for the co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA)* (pp. 49-54).

Passricha, V. & Aggarwal, R. K. (2019). A hybrid of deep CNN and bidirectional LSTM for automatic speechrecognition. *Journal of Intelligent Systems*, *29*(1), 1261-1274.

Piczak, K. J. (2015). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).

Selouani, S. A. & Yacoub, M. S. (2018). Long short-term memory neural networks for artificial dialogue generation. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 761-768). IEEE.

Soe Naing, H.M. at al. (2020). Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System. *International Journal of Intelligent Engineering & Systems*, *13*(2).

Takazawa, S. K. et al. (2024). Explosion Detection Using Smartphones: Ensemble Learning with the Smartphone High-Explosive Audio Recordings Dataset and the ESC-50 Dataset. *Sensors*, *24*(20), 6688.

Wu, Y., Zheng, B. & Zhao, Y. (2018). Dynamic gesture recognition based on LSTM-CNN. In *2018 Chinese Automation Congress (CAC)* (pp. 2446-2450). IEEE.

Wang, W., Yang, X. & Yang, H. (2020). End-to-End low-resource speech recognition with a deep CNN-LSTM encoder. In *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)* (pp. 158-162). IEEE

Xie, J., Fang, J., Liu, C. & Li, X. (2020). Deep learning-based spectrum sensing in cognitive radio: A CNN-LSTM approach. *IEEE Communications Letters*, *24*(10), 2196-2200.