2025 شباط العدد 16 المجلة العراقية للبحوث الإنسانية والإجتماعية والعلمية
No.16 Feb 2025 Iraqi Journal of Humanitarian, Social and Scientific Research
Print ISSN 2710-0952 Electronic ISSN 2790-1254

Review:Speech Recognition Using Deep Learning

Hadeel Luhaib Fouad
Husam Ali Abdulmohsin
[1,2] Computer science department, college of science, university of Baghdad, Iraq
Emails: Hadeel.Fouad2301@sc.uobaghdad.edu.iq; Husam.a@sc.uobaghdad.edu.iq

Abstract

speech recognition is translating a speech signal into a series of words using a computer program and its algorithms. The main goal of speech recognition is to allow machines to recognize and act on sounds. The ability of a computer to identify “receive and interpret” speech and translate it into readable form or text is known as automatic speech recognition. speech recognition is the ability of a computer to understand speech as well as execute an action based on the human's instructions. Recent advancements in speech-to-text (STT) and text-to-speech (TTS) recognition technology have revolutionized many sectors and applications. Spoken language can be translated into written text using STT, and written material can be translated into genuine voice using TTS. Present a thorough analysis of the most recent advancements in STT and TTS recognition systems in this study, including fundamental approaches, practical uses, difficulties, and potential directions. The primary elements of STT and TTS systems, such as speech synthesis methods and automated speech recognition (ASR), are covered first. Through the sources it consulted and earlier studies, where the researcher did an extensive evaluation of speech-to-text (STT) and voice recognition, this research study offers a review that will help future investigations. With an explanation of the basic concepts, in addition to the history of Speech recognition and the techniques that have been used over the years. Many artificial intelligence techniques have been used for this purpose, but the most accurate used algorithms are CNN and RNN. By comparing the results of previous research, notice that the highest accuracy reached is 99.7% using the CNN algorithm and 95% using the RNN algorithm.

Keyword: speech to text; speech recognition; deep learning; CNN; STT; ASR

مراجعة: التعرف على الكلام باستخدام التعلم العميق

هديل لهيب فؤاد
حسام علي عبدالمحسن
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، العراق

المُلخص

التعرف على الكلام هو طريقة ترجمة إشارة الكلام إلى سلسلة من الكلمات باستخدام برنامج كمبيوتر وخوارزمياته. الهدف الرئيسي للتعرف على الكلام هو تمكين الآلات من التعرف على الأصوات والتفاعل معها. قدرة الكمبيوتر لتحديد "استلام وتفسير" الكلام وترجمته إلى شكل مقروء أو نص يُعرف بالتعرف التلقائي على الكلام. أحدثت التطورات الأخيرة في تقنية التعرف على الكلام إلى نص (STT) والنص إلى

كلام (TTS) ثورة في عدد من القطاعات والتطبيقات. يمكن ترجمة اللغة المنطوقة إلى نص مكتوب باستخدام(STT) ، ويمكن ترجمة المواد المكتوبة إلى صوت حقيقي باستخدام(TTS). قُدم تحليلاً شاملاً لأحدث التطورات في أنظمة التعرف STT و TTSفي هذه الدراسة، بما في ذلك الأساليب الأساسية والاستخدامات العملية والصعوبات والاتجاهات المحتملة. من خلال المصادر التي تم الاشارة اليها والدراسات السابقة، حيث قام الباحث بتقييم موسع لتقنية تحويل الكلام إلى نص (STT) والتعرف على الصوت، حيث تقدم دراسة البحث هذه مراجعة من شأنها أن تساعد في البحوث المستقبلية. مع شرح المفاهيم الأساسية، بالإضافة إلى تاريخ التعرف على الكلام والتقنيات التي تم استخدامها على مر السنين. تم استخدام العديد من تقنيات الذكاء الاصطناعي لهذا الغرض، ولكن أكثر الخوارزميات المستخدمة دقة هي CNN وRNN. من خلال مقارنة نتائج الأبحاث السابقة، لاحظ أن أعلى دقة تم الوصول إليها هي 99.7٪ باستخدام خوارزمية CNN و95٪ باستخدام خوارزمية RNN.

الكلمات المفتاحية: تحويل الكلام الى نص، التعرف على الكلام، التعلم العميق، CNN ، STT ، ASR

1. Introduction

Automatic speech recognition is translating a speech signal into a series of words using a computer program and its algorithms. The main goal of speech recognition is to allow machines to recognize and act on sounds. The ability of a computer to identify “receive and interpret” speech and translate it into readable form or text is known as automatic speech recognition. Automatic speech recognition is the ability of a computer to understand speech as well as execute an action based on the human's instructions [1].

Nowadays, automatic speech recognition (ASR) has become a significant field in artificial intelligence (AI) research. ASR tasks are employed for translating speech waves, or signals, to word-sequence representation based on smart computations [1]. There is a wide area of Automatic speech recognition applications in many fields, such as voice applications, automatic language translation, human-computer interactions (HCI), and many via-voice systems [2].

ASR is a basic component of a virtual assistant. It works by processing a human voice and training a system to recognize vocabulary in that voice. ASR has many applications ranging from speech-based controls to online gaming to deliver commands to Iota devices. In the last five decades, ASR became an active research area. ASR is important for human and human-machine communication [1].

Speech recognition has a long history in technology and has seen several significant advancements. Recent developments in big data and deep learning have helped the sector. The abundance of scholarly works on the subject and, more significantly, the widespread industrial use of diverse deep learning techniques in the development and implementation of voice recognition systems serve as indicators of the field's advancement [3].

There has been substantial advancement in the past for each of the aforementioned three voice recognition technology components. In actuality, one

of the most significant turning points in the history of speech recognition is the creation of the fundamental statistical framework.

The statistical model, together with related models and methods that make the model work, belong to the first group of noteworthy historical advancements in speech recognition. The most significant paradigm shift for advances in speech recognition has been the shift from previous non-statistical approaches to statistical methods, particularly the introduction of stochastic processing using HMMs as an acoustic modeling component of speech recognition in the early 1970s. Thirty years later, this strategy is still in high demand. This framework effectively incorporates several models and methods. The two primary methods for effectively teaching HMMs from data are the expectation and maximization (EM) technique and the forward-backward, or Baum-lch, algorithm [4, 5].

Similar to this, N-gram language models and their variations that are trained using EM methods or simple counting techniques are very resilient and flexible in language modelling. Beyond these simple algorithms, statistical discriminant training approaches have been developed since the late 1980s. These methods concentrate on the lost relevant error or maximal mutual information needs rather than the likelihood of data matching. Segmental models, speech models, and structured language are additional developments beyond the basic HMM-like phonological models and the N-gram-like basic language models [6, 7].

Even if there is still a lot to learn in this field, no real progress has been accomplished thus far. To handle a wide range of changing conditions related to channel, noise, speaker, vocabulary, dialect, recognition domain, etc., adaptation is another crucial field of algorithm study. Effective adaptive algorithms are essential to the successful commercial usage of voice recognition technology because they facilitate the quick integration of applications. Maximum posterior likelihood (MAP) estimates and AQ7 maximum likelihood linear regression (MLLR) are the most often used adjustment approaches [8].

In addition to "one-off" or "unsupervised" learning during testing, training can also be conducted using small quantities of data from new tasks or domains of supplementary training material. Speaker Adaptive Training, or SAT, is a strategy that uses these adaptive strategies to train "general" models to improve their ability to reflect the complete statistics of the whole training data set [9].

The establishment of the computer infrastructure needed to support the above-mentioned breakthroughs in statistical models and algorithms falls under the second category of notable achievements in speech recognition. According to Moore's Law, which tracks long-term advancements in computer technology, every 12 to 18 months, memory prices will decrease and CPU will double for a given price. These have been crucial in helping researchers studying voice recognition create and assess intricate algorithms on sufficiently big jobs to yield

significant results. Furthermore, the capacity to create sophisticated systems with ever-increasing capabilities has been made possible by the availability of standard speech corpora for speech training, development, and assessment. Large corpora are required for successful speech modelling since speech is a highly changeable signal with various properties, making automated systems unable to perform the task. The European Language Resources Association (ELRA), the Linguistic Data Consortium (LDC), the National Institute of Standard and Technology (NIST), and other organizations have developed, annotated, and made these corpora available to the global community throughout time. Over time, recorded speech has evolved from limited, precious speech resources to copious volumes of increasingly naturalistic, spontaneous speech [10].

Organizations like as NIST and others have supported the formulation and implementation of stringent benchmark assessments and standards, which have been essential to the development of increasingly competent and powerful systems. Common research tools like Sphinx, HTK, CMU LM toolkit, and SRILM toolkit have helped many laboratories and researchers. Today's system developments would not be possible without the substantial research money, workshops, task descriptions, and system assessments given by organizations such as the U.S. Department of Defence Advanced Research Projects Agency (DARPA). The third category, which is referred to as knowledge representation, includes significant historical advancements in voice recognition. This includes normalizations for Cepstral Mean Subtraction (CMS) and vocal tract length (VTLN). Furthermore, perceptually driven speech signal representations are created, including PLP coefficients and Mel-Frequency Cepstral Coefficients (MFCC) [11, 12].

The most significant advancement in knowledge representation architecture to date has been the creation of searchable unified graph representations, which enable the integration of multiple information sources into a single probabilistic framework. Several sound streams, different speech recognition systems connected at the hypothesis level, ROVER, multiple pass systems with increasing constraints (bigram vs. four-gram, within word dependence vs. cross-word, etc.) are examples of non-compositional techniques [12].

Recently, several techniques have been applied both sequentially and simultaneously with success. Numerous feature-based transformations, including feature-space minimum phone error (fMPE), heteroscedastic linear discriminant analysis (HLDA), and neural net-based features, have also shown this to be the case. The last category includes the search and decoding techniques covered earlier in this section, together with notable historical advancements in voice recognition. The majority of these tactics have concentrated on stack decoding (A∗ search) and time-synchronous Viterbi algorithms. Continuous voice

recognition on a broad scale would not be possible without these useful decoding methods [13, 14].

In the 2000s, Global Autonomous Language Exploitation (GALE) and Effective Affordable Reusable voice-to-text (EARS) systems re two speech recognition projects funded by DARPA. IBM, a BBN-led team including LIMSI and the Universities of Pittsburgh and Cambridge, SRI, and the University of Washington, and ICSI made up the other two of the four teams that took part in the EARS initiative. The Switchboard telephone voice corpus was assembled with funding provided by EARS. It includes 260 hours of speeches from more than 500 speakers that have been recorded. The GALE project focuses on news broadcast speech in Mandarin and Arabic. In 2007, Google started experimenting with voice recognition after employing many Nuance specialists. The first offering was a phone-based directory service called GOOG-411. GOOG-411 recordings have produced good data that has aided Google's recognition algorithms. Google Voice Search is now accessible in more than 30 languages. At least since 2006, the National Security Agency in the United States has used voice recognition technologies for keyword detection. By using this method, analysts may search through vast amounts of chat records for keyword references. Analysts may use queries to search the database for interesting discussions, and recordings may be indexed. A few government research projects, such as IARPA's Babel program and DARPA's EARS program, concentrated on the use of speech recognition technology for intelligence [15, 16].

Speaker independence, or the ability to differentiate speech recognition from speaker recognition, was considered a breakthrough in speech recognition (also known as voice recognition) in the early 2010s. Systems needed a "training" phase up until that point. The catchphrase in a 1987 ad with a doll that "children could train to respond to their voice" was "Finally, the doll that understands you." By using the extensively benchmarked Switchboard effort to transcribe conversational telephone audio in 2017, Microsoft researchers achieved a historical human parity milestone. Several deep learning models are applied to improve the accuracy of voice recognition. It was reported that word mistakes in speech recognition had been made by as few as four skilled human transcribers working together on the same benchmark—funded by the IBM Watson voice team [17].

However, most ASR works have been applied in the English language, and limited studies have utilized the Arabic language in the ASR field. Arabic is one of the complicated languages in computerizing works; hence, developing Arabic speech recognition systems has a tangible level of hardness for many reasons; the richness and sparseness nature of Arabic vocabulary data, lexical diversity, the non-discretized huge amount of available text, and the number of distinct alive verbal dialects in the world. Besides, the complexity of the Arabic language's

written word morphology. Although, it is very rich in vocabulary [43]. So, several challenges for speech recognition research have been presented due to the large vocabulary of the Arabic language [44].

Recently, there has been a growth of studies that have developed powerful Arabic automatic speech recognition (AASR) systems [44]. The written words in the Arabic language have diacritics (Arabic haricot); several distinct symbols that are positioned above the Arabic letters [43]. The Arabic diacritics denote sounds corresponding to English vowels and Chinese tones, they are important for recognizing the word's and sentence's sense, as another Arabic ASR challenge [44]. Furthermore, the dialectal Arabic ASR acoustic model building is a challenging task. The training task of these models entails the correct Arabic dialect. So, working with Arabic dialectal ASR has many challenges. To obtain a good and accurate model, a large dataset has to be gathered due to the lack of training data benchmarks [43].

2. Related Work

In many creative occupations today, computers have displaced a significant percentage of human workers. As a result, computer vision, robotics, machine learning, and natural language processing are the domains that make up artificial intelligence. Computers may also be used to predict voice recognition. Deep learning is employed in this study to classify speech. To train the model, a Google voice data warehouse was utilized. Accuracy was 66.22% achieved [18].

The objective of a study by researchers on deep convex neural networks for voice recognition was to create an accurate system for Arabic speech recognition using CNNs. Using a dataset of Arabic voices captured in various settings. Using signal processing methods to extract speech characteristics. Constructing a CNN model and using the training set of data to train it. Analyzed the model's output using test data. When the model was used to classify words from an unknown data sample, it obtained 97.06% accuracy. When testing the model on voice samples captured in various loud situations, it performed well. For Arabic, the model performed better than conventional voice recognition techniques [19].

The researchers carried out an exercise to determine how the Arabic letter should be pronounced precisely according to guidelines. The precise recognition of a high number of short voles is one of the key obstacles in accurately pronouncing the Arabic alphabet. This problem cannot be solved by conventional statistical sound processing approaches or machine learning models. As a result, created a deep neural network (DNN) model to categorize Arabic short vols. A fresh audio dataset was gathered, a neural network architecture was established, and the resultant model was improved and refined via several iterations to reach high classification accuracy. These steps are taken to create the model from scratch. Given a set of unseen audio samples of spoken short vols, our proposed model

reached a test accuracy of 95.77%. Can say that experts and researchers can use our results to build better intelligent learning support systems in Arabic speech processing. [20].

According to researchers, the necessity for systems with voice recognition technology has grown significantly over the past several years since it has become an essential part of our everyday life. The purpose of this study is to find the two chosen Persian words in any audio recording. Two standard and local datasets one for training and the other for testing recreated specifically for this model. Audio waveform images were recreated from both datasets. For every audio test, the model can extract distinct bounding boxes using object detection technologies. Each box picture is then sent to the CNN classifier, which outputs the associated label. Ultimately, a threshold is established to ensure that the output is limited to just high-resolution squares. The results showed 93% accuracy for the CNN classifier and 50% accuracy for testing the model with object detection [21].

It has recently been demonstrated that the conventional Gaussian mixture model (GMM)-HMM voice recognition performance is much inferior to that of a hybrid deep neural network (DNN) hidden Markov model (HMM). The capacity of DNNs to represent intricate correlations in speech characteristics contributes to enhanced performance. This research demonstrates that convolutional neural networks (CNNs) may be used to achieve even more error rate reduction. First, give a quick overview of the fundamental CNN and describe how voice recognition may be accomplished with it. Additionally, suggest a restricted light-sharing plan that improves speech feature modeling. A certain amount of invariance to slight alterations in speech parameters along the frequency axis is demonstrated by the unique structure of CNNs, including local connection, weight sharing, and clustering. This is crucial for adjusting for variations in speakers and environments. According to experimental data, in voice search, broad vocabulary recognition, and TIMIT phone identification tasks, CNNs outperform DNNs by 6%–10% in terms of error rate reduction [22].

Gaussian mixing is used by Hidden Markov models (HMM) to create a spectral explanatory model of the wave. Speech recognition using HMM has shown limited time domain efficiency and low accuracy. To improve accuracy and efficiency, 1D convolutional neural networks (CNN) can be used in place of HMM in the suggested system. Due to their low efficiency, HMMs are being replaced by deep neural networks (DNN). For voice analysis, DNN techniques can handle nonlinear data and lor error rates. One-dimensional speech may occasionally be processed using sliding windows supplied into the network, in which case the speech recognition model chooses the optimal disambiguation of the speech signal by extracting the feature of the audio signal within the temporal

domain. Conv1D handles speech signals by providing a full-frequency feature vector at each instant that describes the entire sample [23].

To achieve voice recognition, this article makes use of convolution neural networks or CNNs. It is a different kind of neural network that can simulate spectral correlations in signals and lessen spectral variance. In addition, the neural network in the article is trained via backpropagation. The study uses a set of voice recordings made on our own as training data for the entire experiment and then uses the remaining recordings to evaluate the neural network. According to experimental data, CNNs are capable of doing solitary word recognition effectively [24].

This research presents a convolutional neural network-based speech recognition system that has industrial applications. The findings of the experiments demonstrate that the convolutional neural network training approach may considerably increase the voice recognition accuracy of command words when compared to the conventional deep neural network training method. In this study, a novel approach to large-scale picture recognition using convolutional neural networks is described. Convolutional neural networks are superior to deep and cyclic neural networks in command voice recognition, and this has been demonstrated by contrast experiments, making them suitable for industrial applications. Nevertheless, our experience still has several shortcomings, such as an inadequate spoken corpus of command words and a poorly constructed neural network, among other things. Need some people with high ideals to continue to explore command word speech recognition, to achieve the best [25].

The goal of this project is to develop a voice command recognition system that can anticipate speech commands that have been predefined. Convolutional neural networks, recurrent neural networks, and other neural network models are implemented using the dataset that Google's Tensor Flow and AIY teams have given. Convolutional and recurrent neural networks operate together to produce 96.66% accuracy for 20 labels, an 8% improvement over convolutional neural networks alone [26].

Using English speech as the research topic, a deep learning speech recognition system that blends speech characteristics and speech features is suggested in this work. Using a deep neural network supervised learning technique, high-level speech characteristics are initially retrieved to train the GMM-HMM acoustic model with the new speech features. Next, the newly constructed network's new voice feature is selected from the fixed hidden layer output. The neural network based on the linear feature fusion technique then fuses the speech features and speech feature features inside the same CNN framework. Finally, a deep neural network-based speech feature extractor is trained on a variety of speech features, and the deep neural network classifies the extracted speech features as audio. Based on the results of experiments, the proposed English speech recognition

algorithm can directly and efficiently combine the two approaches by combining the speaker's speech features and the speech features of the deep neural network in the input layer of the deep neural network. This may improve the system's ability to recognize speech. Primarily in English [27].

It has been suggested to use recurrent neural networks (RNNs) in conjunction with connective temporal classification (CTC) to identify unsegmented sequences. This allows for the training of an end-to-end speech recognition system as opposed to one that operates in mixed environments. RNNs are costly to compute and can be challenging to train, though. This research, combines hierarchical CNNs with direct CTC without recurrent connections to present a comprehensive speech framework for sequence tagging, motivated by the benefits of both CNNs and CTC techniques. Demonstrate that the suggested model is competitive with current platforms and computationally efficient by analyzing its approach to the TIMIT speech recognition challenge. Moreover, see that CNNs can model temporal correlations with suitable context knowledge [28].

Table 1: Studies in which researchers relied on (CNN) in speech recognition

REF	Year	Algorithm	ACC. % or WER
[18]	2018	CNN	66.22%
[19]	2021	CNN	97.06٪
[20]	2022	CNN	95.77٪
[21]	2022	CNN Object detection	99.7٪ 50٪
[22]	2014	CNN	6% to 10% error rate
[23]	2020	CNN	70%
[24]	2016	CNN	99%
[25]	2020	CNN RNN DNN	92% 89% 82%
[26]	2020	CNN-RNN	96%
[27]	2019	CNN-RBM-ASAT	0.082
[28]	2017	CNN-CTC	Test PER= 18.2%

Using the unlabelled audio of the Libri-Light dataset, the researchers used recent advancements in semi-supervised learning for automatic speech recognition to achieve state-of-the-art results on Libri Speech. More precisely, employ Spec Augment to do training of noisy students using huge Conformer models that have been pre-trained using wav2vec 2.0 pre-training. By doing this, get Word Error

Rates (WERs) of 1.4%-2.6% in the Libri Speech test and other test suites, which is lor than the current state-of-the-art Rs of 1.7%-3.3% [29].

For this experiment, the researchers employed feature extraction and deep learning approaches. The machine is trained on 5–10 isolated words by the suggested approach. Since speaker sound is the foundation of the system, the words are first saved in a wav (wave audio) file for training. For every word, several examples are trained and kept. These saved words will be grouped with the word that you submit via audio. The voice signal is digitally formatted as part of the pre-processing step. In order to smooth the signals and increase the signal por at a higher frequency, this digital signal is processed through first order filters. A methodical approach to feature extraction is called MFCC. This step is finished, and the vertical frequency slope coefficients are acquired. The MFCC uses human sense sensitivity to scan frequencies. The HMM is used for speech identification and training following pre-processing, segmentation, and feature extraction processes. An LSTM, as well-liked kind of neural network, was used in the implementation of the ANN-based voice recognition system [30].

This paper examines the use of cutting-edge comprehensive deep learning techniques to construct a reliable partitioned Arabic ASR system. These techniques rely on Mel-Scale Filter Bank register energies as acoustic characteristics and Mel-Frequency Cepstral coefficients. To our knowledge, the task of automated segmented Arabic voice recognition has not been approached using a thorough deep learning technique. This study closes this gap by presenting CNN-LSTM, a novel CTC-based ASR system, and a thorough attention-based method for enhancing Arabic-modulated ASR. To get better outcomes, a word-based language model is also applied. The extensive techniques used in this study are predicated on the cutting-edge frameworks, Espresso and ESP net. These frameworks are trained and tested using the Classical Arabic Speaker Corpus (SASSC), which comprises seven hours of Modern Standard Arabic speech. The findings of the experiment indicate that in the Arabic voice recognition test, CNN LSTM with attention framework performs better than both regular ASR and the combined CTC-attention ASR framework. The word error rate that CNN-LSTM with attention framework can get is superior to both combined CTC ASR and conventional ASR by 2.62% and 5.24%, respectively [31].

Highlight the significance of "speech signal processing" in the suggested approach, which uses vertical frequency slope coefficients (MFCCs) as the best recovered characteristics to diagnose Arabic mispronunciations. Additionally, the categorization procedure made advantage of long short-term memory (LSTM). Moreover, a two-level classification was carried out by combining the gender recognition model with the analytical framework. Our findings demonstrate that in addition to gender recognition, the LSTM network considerably improves pronunciation fault detection. In the suggested system, LSTM models attained an

average accuracy of 81.52%, demonstrating strong performance in comparison to earlier pronunciation mistake detection methods [32].

This research describes the use of deep neural network (DNN) transfer learning for the creation of an automatic speech recognition (ASR) system for the Central Kurdish (CKB) language. An AsoSoft CKB speech dataset is utilized to generate an Acoustic Model (AM) using a mix of Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction from speech signals and Long Short-Term Memory (LSTM) with a Contact Temporal Classification (CTC) output layer. Additionally, applied the N-gram language model to the extensive text dataset (about 300 million tokens) that was gathered. In addition, a dynamic lexicon model with approximately 2.5 million CKB terms is extracted from the text corpus. The outcomes demonstrated that, as compared to traditional statistics units, employing DNN yields better results. The proposed method achieves a word error rate of 0.22% by combining transfer learning and language model adaptation. This result outperforms the best result reported for CKB [33].

Table 2: Studies in which researchers relied on (LSTM) in speech recognition

REF	Year	Algorithm	ACC. or WER
[29]	2022	LSTM-conformer	Rs= 1.4%-2.6%
[30]	2021	LSTM	95٪
[31]	2020	CTC-attention، CNN-LSTM	CTC-attention = 31.10% CNN-LSTM = 28.48%.
[32]	2023	LSTM	81.52٪
[33]	2022	LSTM-CTC	0.22

Address voice recognition in Modern Standard Arabic (MSA) in this article. Discuss the difficulties associated with Arabic, including its complicated morphology and the lack of short vols in written language, which leads to several alternative vols per letter, many of which are at odds with one another. Using the Kaldi toolbox, create an ASR system for MSA. Numerous language and sound models are trained. For the baseline system, achieved a Word Error Ratio (R) of 14.42, re-recording the network and rewriting the output with the correct Z-hamza above or below @Alif resulted in a 12.2 relative improvement [34].

This investigator selected to play the preferred sounds from a large file. Deep learning is employed in this study to classify speech. To train the model, Google Collection was utilized. Accuracy was 66.22% achieved [35].

Research on finding novel techniques to multilingual speech recognition and contrasting the efficacy of current approaches with suggested novel ones are included in the article. The multilingual audio and text collection include multilingual phrases that incorporate vocabulary from both Kannada and English.

With the new approach, our voice recognition model achieves 71% accuracy. When compared to other multilingual translation methods now in use, this is a really decent outcome [36].

In this study, an artificial neural network classifier is used to detect pronunciation problems using a novel approach based on Acoustic Discriminative Feature (APF). Arabic consonants are divided into two groups according to their phonetic similarities using domain knowledge. Consonants with comparable final sounds make up the first group, and consonants with entirely distinct final sounds make up the second. To train the classifier in our suggested method, discriminative audio characteristics are needed. Different components of Arabic consonants re discovered in order to extract these properties. The dataset is gathered from male and female, indigenous and non-indigenous, and children of various ages as a test case. There are 5,600 distinct Arabic consonants in this collection. When the system was evaluated using basic audio parameters, its average accuracy was reported to be 73.57%. Using discriminative audio characteristics raised the accuracy average to 82.27% on average. Bad Phonemes are consonant pairings that sound very close phonetically yet have undesirable outcomes. Additionally, a subjective study was carried out to confirm the suggested system's efficacy [37].

This study looks at how challenging it is to assess English speech recognition and pronunciation quality using deep learning. This research selects three separate indicators—intonation, speed, and rhythm—to assess the quality of English pronunciation. The findings of our assessment and the results of the manual evaluation clearly demonstrate the better reliability of the deep learning model of English voice recognition and pronunciation quality described in this research. Only thirty-two of the two hundred and forty samples that re analysed shod a one-degree difference [38].

In this work, provide an automatic speech recognition (ASR) system for the Lithuanian language that uses deep learning techniques to recognize uttered words only by their phonetic patterns. The ASR challenge is solved using two different encoder-decoder models: a conventional model and a model that incorporates an attention mechanism. These models are tested on two different tasks: one for isolated voice recognition (accuracy: 0.993) and another for extended phrase recognition (accuracy: 0.992) [39].

In this work, provide three methods for automated voice recognition in Arabic. Suggest using decision trees to generate pronunciation variants for modelling purposes. Suggest the Hybrid technique for acoustic modelling, which uses a different native acoustic model to modify the original one. In terms of the language model, leverage processed text to enhance the model. According to the trial data, the suggested pronunciation model technique reduces R by about 1%. R is decreased by 1.2% in the acoustic modelling and 1.9% in the modified language modelling [40].

A hybrid technique for headphones and voice recognition has been proposed by researchers. Five solitary words re captured in this study with two male and one female speaker. In order to enhance the retrieved features and produce an appropriate feature set, employed the Mel Frequency Cepstral Coefficient (MFCC) feature extraction approach in conjunction with a genetic algorithm. MFCC is used in the first step for feature extraction, genetic algorithms are used for feature optimization, and deep neural networks are used for training in the third and final stages. Lastly, a genuine environment is specified in order to assess and validate the suggested business model. Computed measures like accuracy, precision rate, recall rate, sensitivity, and specificity to confirm the efficacy of the suggested work [41].

This work examines the internal representations that have been learned in a complete ASR model. Compare phonemes and grammatical units, as well as various pronunciation parameters, in order to assess the quality of representation across a number of categorization tasks. Examined three datasets and two languages, Arabic and English, and discovered a startling degree of consistency in the representation of various properties throughout the deep neural network's layers [42].

Table 3: Studies in which researchers relied on different speech recognition techniques

REF	year	Algorithm	ACC. Or WER
[34]	2017	DNN	17.65
[35]	2023	ResNet-34	66.22٪
[36]	2022	DNN	71%
[37]	2019	ANN	82.27٪
[38]	2022	DL	86.7%
[39]	2019	ASR (encoder-decoder), attention mechanism	0.993 0.96
[40]	2017	PM AM LM	1 1.2 1.9
[41]	2017	DNN	97.18٪
[42]	2019	End-to-end (E2E) deep neural network	87%

3. Result and Discussion

In this review, the body of research on speech recognition technologies is examined to assess the effectiveness of various machine learning and deep

learning techniques across diverse languages and datasets. The studies demonstrate the significant role of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models in advancing voice recognition accuracy and functionality. Below is a summary of the main findings organized by the different approaches used in the studies:

1. **Convolutional Neural Networks (CNNs)**:
- CNNs were widely used due to their ability to recognize speech features with high accuracy across different settings and languages. For instance, in Arabic speech recognition, CNNs achieved up to 99% accuracy by using signal processing methods to enhance speech feature extraction [19].
- CNNs also demonstrated an advantage over traditional Gaussian Mixture Models (GMM)-Hidden Markov Models (HMM) with significant error rate reductions (6%–10%) when recognizing broad vocabulary and voice search tasks [22].
- However, while CNNs achieved high accuracy in controlled environments, challenges remain with real-world noisy data and a lack of generalized datasets for command-based applications [25].

2. **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)**:
- Studies incorporating RNNs or LSTMs highlighted the strengths of these networks in handling temporal dependencies, especially when combined with techniques like CTC (Connective Temporal Classification). In one Arabic ASR (Automatic Speech Recognition) study, a CNN-LSTM model achieved 28.48% Word Error Rate (WER), outperforming standard ASR models [31].
- RNNs with attention mechanisms have also shown significant promise, particularly in languages with complex morphology such as Arabic and Kurdish, achieving WER reductions by leveraging linguistic models and CTC [33, 34].

3. **Hybrid Approaches and Feature Extraction Techniques**:
- Several studies employed hybrid approaches that combine CNNs, RNNs, and DNNs with advanced feature extraction methods like MFCC (Mel-Frequency Cepstral Coefficients) and Spec Augment. For example, a study that utilized MFCC and LSTM together achieved 95% accuracy for isolated words [30].
- Other innovative techniques include hierarchical CNNs with direct CTC for sequence tagging and hybrid frameworks incorporating attention mechanisms. These models have yielded competitive results with more conventional speech recognition systems, particularly for unsegmented audio and multilingual applications [28, 36].

4. **Limitations and Areas for Improvement**:
- Although high accuracies were achieved across several studies, limitations were noted. In real-world applications, model performance tends to decrease due to factors like background noise and dialectal variation [25].

- Many researchers highlighted a shortage of language-specific datasets, which hinders the development of robust models for underrepresented languages. For instance, Arabic ASR faces challenges due to limited corpora for diverse dialects, as well as pronunciation differences resulting from short vowels [34].

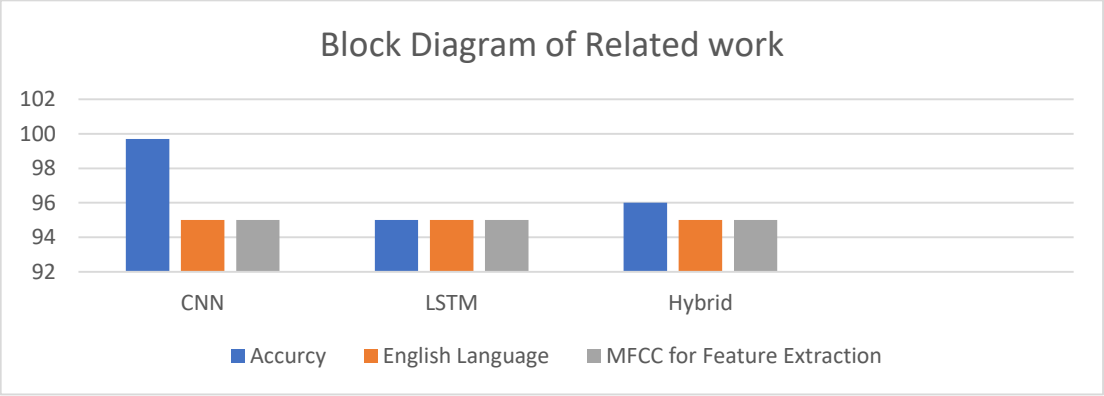
5. **Industrial Applications**:

- The reviewed studies have practical implications, especially in command-based voice recognition systems and real-time applications. For instance, CNN-based models in industrial applications demonstrated improved accuracy for command word recognition, which holds promise for integration into smart devices and AI-driven assistants [24].

6. **Multilingual Speech Recognition**:

- Studies on multilingual ASR illustrate the application of mixed-language datasets for improved recognition of languages with limited resources. For instance, a model combining Kannada and English achieved a 71% accuracy rate by leveraging multilingual audio collections, showing potential for scalable voice recognition across different linguistic contexts [36].

The studies reviewed here indicate that advanced neural networks like CNNs, RNNs, and their hybrid configurations provide substantial improvements in speech recognition accuracy. However, the results emphasize the need for ongoing research to address challenges associated with real-world application, language-specific datasets, and dialectal variability. The integration of advanced feature extraction methods, such as MFCC and attention mechanisms, has also contributed to performance improvements across various languages and use cases. Further development of these techniques will likely continue to drive advancements in the field of automatic speech recognition.



4. conclusion

Speech recognition has become a crucial area in artificial intelligence (AI) research, where automatic speech recognition (ASR) tasks convert spoken language into text representations through advanced computational methods. ASR has broad applications across diverse fields, including voice-activated

systems, automatic language translation, human-computer interaction (HCI), and voice-controlled devices. According to recent studies, convolutional neural networks (CNNs) achieve the highest accuracy in ASR tasks, particularly when combined with Mel Frequency Cepstral Coefficients (MFCC) for feature extraction. English remains the most widely studied language in this field.

Implications for Future Research

The findings from these studies highlight several key implications for future work in speech recognition:

- **Language-Specific Challenges**: The complexity of languages like Arabic necessitates tailored approaches that address unique phonetic and morphological characteristics.
- **Hybrid Models**: There is a growing need for hybrid models that combine different neural architectures, allowing for improved accuracy and robustness in recognizing diverse speech patterns.
- **Data Utilization**: The success of using large, varied datasets indicates that future research should continue to focus on data diversity and quality to enhance model performance.

5. **Acknowledgement**

Our sincere to all researchers in the field of speech recognition for their work in this area.

6. References

[1]. Abdulmohsin, Husam Ali, et al. "Automatic illness prediction system through speech." *Computers and Electrical Engineering* 102 (2022): 108224.
[2]. Abdulmohsin, Husam Ali. "A new proposed statistical feature extraction method in speech emotion recognition." *Computers & Electrical Engineering* 93 (2021): 107172.
[3]. Reynolds, Douglas; Rose, Richard (1995). "Robust text-independent speaker identification using Gaussian mixture speaker models" (PDF). IEEE Transactions on Speech and Audio Processing. 3 (1): pp 72–83.
[4]. Baker, J. (1975). Stochastic modeling for automatic speech recognition, in D. R. Reddy, (ed.), Speech Recognition, Academic Press, New York. p 5 - 20.
[5]. Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 39(1), p 1–21.
[6]. He, X., L. Deng, C. Wu (2008). Discriminative learning in sequential pattern recognition, IEEE Signal Processing Magazine, 25(5), 14–36.

[7]. Deng, L., D. Yu, and A. Acero (2006). Structured speech modeling, IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription), 14(5), 1492–1504.
[8]. Leggetter C. and P. Woodland (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, 9, 171–185.
[9]. Anastasakos, T., J.McDonough, and J. Makhoul (1997). Speaker adaptive training:Amaximum likelihood approach to speaker normalization, in Proceedings of the IEEE InternationalConference onAcoustics, Speech, and Signal Processing, pp. 1043–1046, Munich, Germany.
[10]. Baker, J., L.Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy (2009). Updated MINDS report on speech recognition and understanding part I, IEEE Signal Processing Magazine, 26(3), pp 75–80.
[11]. Hermansky H. and N. Morgan (1994). RASTA processing of speech, IEEE Transactions on Speech and Audio Processing, 2(4), 578–589.
[12]. Rosenberg, A., C. H. Lee, and F. K. Soong (1994). Cepstral channel normalization techniques for HMMbased speaker verification, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 1835–1838, Adelaide, SA.
[13]. Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), IEEE Automatic Speech Recognition and Understanding Workshop, pp. 3477–3482, Santa Barbara, CA.
[14]. Morgan,N., Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain,H.Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos (2005). Pushing the envelope—Aside, IEEE Signal Processing Magazine, 22, 81–88.
[15]. Povey, B., Kingsbury, L.Mangu, G. Saon, H. Soltau, and G. Zig (2005). FMPE: Discriminatively trained features for speech recognition, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA.
[16]. Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". Neural Networks. 61: 85–117.
[17]. Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. Proceedings of ICML'06, pp. 369–376.
[18]. Phoemporn Lakkhanawannakun & Chaluemwut Noyunsan, 2018, Speech Recognition using Deep Learning, Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen, 40000, Thailand.

[19]. Ayad Alsobhani & Hanaa M A ALabboodi & Haider Mahdi, 2021, Speech Recognition using Convolution Deep Neural Networks, IICESAT Conference, College of Material Engineering, University of Babylon, Iraq.
[20]. Amna Asif & Hamid Mukhtar & Fatimah Alqadheeb & Hafiz Farooq Ahmad & Abdulaziz Alhumam, 2022, An Approach for Pronunciation Classification of Classical Arabic Phonemes Using Deep Learning, Licensee MDPI, Basel, Switzerland, Appl. Sci. 2022, 12, 238. https://doi.org/10.3390/app12010238.
[21]. Shahed Mohammadi & Niloufar Hemati & Neda Mohammadi, 2022, Speech Recognition System Based on Machine Learning in Persian Language, Computational Algorithms and Numerical Dimensions Com. Alg. Num. Dim Vol. 1, No. 2.
[22]. Ossama Abdel-Hamid & Abdel-rahman Mohamed & Hui Jiang & Li Deng & Gerald Penn & Dong Yu, 2014, Convolutional Neural Networks for Speech Recognition, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 10.
[23]. A. Bhavani, 2020, SPEECH RECOGNITION USING THE NN, International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 12.
[24]. Du Guiming & Wang Xia & Wang Guangyan & Zhang Yan & Li Dan,2016, Speech Recognition Based on Convolutional Neural Networks.
[25]. Yang Xuebin & Jia Lei, 2020, Speech recognition of command words based on convolutional neural network.
[26]. Sushan Poudel & Dr. R Anuradha, 2020, Speech Command Recognition using Artificial Neural Networks.
[27]. Zhaojuan Song, 2019, English speech recognition based on deep learning with multiple features
[28]. Ying Zhang & Mohammad Pezeshki & Phil´emon Brakel & Saizheng Zhang & C´esar Laurent Yoshua Bengio & Aaron Courville, 2017, Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks, Departement d'informatique et de recherche op´erationnelle Universit´e de Montr´eal, Montr´eal, QC H3C 3J7 CIFAR Senior Fellow, 2 CIFAR Fellow.
[29]. Yu Zhang & James Qin & Daniel S. Park & i Han & Chung-Cheng & Chiu & Ruoming Pang & Quoc V. Le & Yonghui Wu, 2022, Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition, Google Research, Brain Team, arXiv:2010.10504v2 [eess.AS], Self-Supervised Learning for Speech and Audio Processing Workshop @ NeurIPS 2020.
[30]. Arun HP & Jithin Kunjumon & Sambhunath R3 & Ancy S Ansalem, 2021, Malayalam Speech to Text Conversion Using Deep Learning,IOSR Journal of Engineering (IOSRJEN) ,ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 11, Issue 7.
[31]. Hamzah A. Alsayadi & Abdelaziz A. Abdelhamid & Islam Hegazy & Zaki T. Fayed, 2020, Arabic speech recognition using end-to-end deep learning.

[32]. Abdelfatah Ahmed & Mohamed Bader & Ismail Shahin & Ali Bou Nassif & Naoufel rghi & Mohammad Basel, 2023, ArabicMispronunciation Recognition SystemUsing LSTMNetwork,Licensee MDPI, Basel, Switzerland.
[33]. Abdulhady Abas Abdullah & Hadi Veisi, 2022, Central Kurdish Automatic Speech Recognition using Deep Learning, Journal of University of Anbar for Pure Science (JUAPS), P- ISSN 1991-8941 E-ISSN 2706-6703.
[34]. Mohamed Amine Menacer & Odile Mella & Dominique Fohr & Denis Jouvet & David Langlois & Kamel Smaïli, 2017, An enhanced automatic speech recognition system for Arabic, Loria, Campus Scientifique, BP 239, 54506 Vandoeuvre L`es-Nancy, France.
[35]. Phoemporn Lakkhanawannakun & Chaluemwut Noyunsan, 2023, Speech Recognition using Deep Learning, Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen, 40000, Thailand.
[36]. P Deepak Reddy & Chirag Rudresh & Adithya A S, 2022, MULTILINGUAL SPEECH TO TEXT USING DEEP LEARNING BASED ON MFCC FEATURES, Machine Learning and Applications: An International Journal (MLAIJ) Vol.9, No.2.
[37]. Muazzam Maqsood & Adnan Habib & Tabassam Nawaz, 2019, An Efficient Mispronunciation Detection System Using Discriminative Acoustic Phonetic Features for Arabic Consonants, Department of Software Engineering, University of Engineering and Technology Taxila, Pakistan, The International Arab Journal of Information Technology, Vol. 16, No. 2.
[38]. Yushu Xu, 2022, English Speech Recognition and Evaluation of Pronunciation Quality Using Deep Learning, Foreign Language Department, Zhejiang College of Shanghai Finance and Economics, Jinhua, Zhejiang 321013, China.
[39]. Laurynas Pipiras & Rytis Maskeli & Robertas Damaševi˘cius, 2019, Lithuanian Speech Recognition Using Purely Phonetic Deep Learning, Computers 2019, 8, 76; doi:10.3390/computers8040076, MDPI.
[40]. Basem H. A. Ahmed & Ayman S.Ghabayen, 2017 , Arabic Automatic Speech Recognition Enhancement, Palestinian International Conference on Information and Communication Technology.
[41]. Gurpreet Kaur & Mohit Srivastava & Amod Kumar, 2017, Speaker and Speech Recognition using Deep Neural Network,International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-8).
[42]. Yonatan Belinkov & Ahmed Ali & James Glass, 2019, Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition.
[43]. Mohammed, Zahra K., and Nada AZ Abdullah. "Survey For Arabic Part of Speech Tagging based on Machine Learning." Iraqi Journal of Science (2022): 2676-2685.

[44]. Hussien, Anwar Abdul-Razzaq, and Nada AZ Abdullah. "A Review for Arabic Sentiment Analysis Using Deep Learning." Iraqi Journal of Science 64.12 (2023).