



# Convolutional Neural Network Based Real Time Arabic Speech Recognition to Arabic Braille for Hearing and Visually Impaired

Surbhi Bhatia<sup>1\*</sup>, Ajantha Devi<sup>2\*</sup>, Razan Ibrahim Alsuwailem<sup>1</sup> and Arwa Mashat<sup>3</sup>

<sup>1</sup> Department of Information Systems, College of Computer Sciences and Information Technology, King Faisal University, Al Hasa, Saudi Arabia, <sup>2</sup> Research Head, AP3 Solutions, Chennai, India, <sup>3</sup> Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia

## OPEN ACCESS

### Edited by:

Celestine Iwendi,  
School of Creative Technologies  
University of Bolton, United Kingdom

### Reviewed by:

Pradeep Singh,  
KIET Group of Institutions, India  
Zaki Khan,  
University of Worcester,  
United Kingdom  
Praise Young,  
Nnamdi Azikiwe University, Nigeria  
Ayodeji Olateju,  
University of Bolton, United Kingdom

### \*Correspondence:

Surbhi Bhatia  
sbhatia@kfup.edu.sa  
Ajantha Devi  
research\_head@ap3-solutions.com

### Specialty section:

This article was submitted to  
Digital Public Health,  
a section of the journal  
Frontiers in Public Health

Received: 17 March 2022

Accepted: 25 April 2022

Published: 27 May 2022

### Citation:

Bhatia S, Devi A, Alsuwailem RI and  
Mashat A (2022) Convolutional Neural  
Network Based Real Time Arabic  
Speech Recognition to Arabic Braille  
for Hearing and Visually Impaired.  
Front. Public Health 10:898355.  
doi: 10.3389/fpubh.2022.898355

Natural Language Processing (NLP) is a group of theoretically inspired computer structures for analyzing and modeling clearly going on texts at one or extra degrees of linguistic evaluation to acquire human-like language processing for quite a few activities and applications. Hearing and visually impaired people are unable to see entirely or have very low vision, as well as being unable to hear completely or having a hard time hearing. It is difficult to get information since both hearing and vision, which are crucial organs for receiving information, are harmed. Hearing and visually impaired people are considered to have a substantial information deficit, as opposed to people who just have one handicap, such as blindness or deafness. Visually and hearing-impaired people who are unable to communicate with the outside world may experience emotional loneliness, which can lead to stress and, in extreme cases, serious mental illness. As a result, overcoming information handicap is a critical issue for visually and hearing-impaired people who want to live active, independent lives in society. The major objective of this study is to recognize Arabic speech in real time and convert it to Arabic text using Convolutional Neural Network-based algorithms before saving it to an SD card. The Arabic text is then translated into Arabic Braille characters, which are then used to control the Braille pattern via a Braille display with a solenoid drive. The Braille lettering triggered on the finger was deciphered by visually and hearing challenged participants who were proficient in Braille reading. The CNN, in combination with the ReLU model learning parameters, is fine-tuned for optimization, resulting in a model training accuracy of 90%. The tuned parameters model's testing results show that adding the ReLU activation function to the CNN model improves recognition accuracy by 84 % when speaking Arabic digits.

**Keywords:** braille, Speech to Text, speech to Braille, speech recognition, text to braille, Arabic digit

## INTRODUCTION

Louis Braille invented Braille, a worldwide method of reading and writing for visually impaired people, in 1821. The Braille cell is the system's smallest unit. Because of the prosody (intonation) of the system, visually and hearing-impaired individuals who are skilled in Finger Braille can communicate words and express a variety of emotions (1). Because there are so few people who

can read Finger Braille, visually and hearing-impaired people can only communicate through an interpreter. As a result, those who are visually and hearing impaired are severely limited in their participation. Finger Braille input techniques, including a wearable device, have been developed (2). Created a wearable input device employing accelerometers affixed on the tops of rings and invented finger Braille gloves with accelerometers integrated in the fingers (3). In addition (4), created a wristband with an eighteen-accelerometer Finger Braille input device.

Visually and hearing-impaired people must wear gloves, rings, or bracelets to input Finger Braille (5) into these devices. People who use these assistive devices (6) who are blind and deaf must not only wear the sensors, but also learn a new communication system. Children who have both hearing and vision impairments require different teaching strategies than children who only have one impairment. Although different countries define hearing and vision impairments differently, the negative effects of dual sensory impairments on access to environmental information are universally acknowledged, as is the fact that this unique disability necessitates specific teaching strategies to aid and support learning.

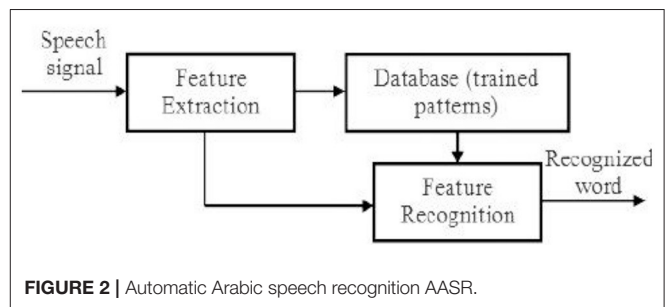
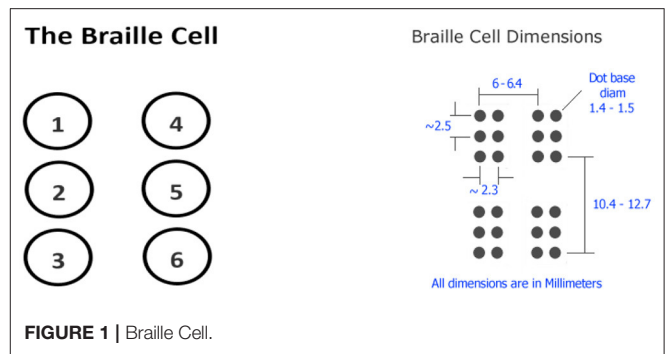
Because they cannot read text written on regular paper, visually impaired people use the Braille system in their writing and reading. However, Visual and hearing individuals in UAE are deprived of education and information due to a lack of basic reading materials, such as books and notes in Braille written format. As a result, we created an automatic Arabic speech to Braille translation system to address this issue by converting electronic Arabic books to Arabic Braille books.

## Braille System

The Braille system (7) is a reading and writing system used worldwide by visually impaired people. Passing fingers across Braille characters or Braille cells is how it's read. In a Braille cell, six dots are arranged in a rectangle grid of two dots horizontally in a row and three dots vertically in a column, yielding four different patterns. Each dot arrangement is known as a cell (1), and each cell will have at least one elevated dot and a maximum of six.

Each of the twenty-five rows of text on a standard Braille printed page has forty cells. The size of a standard Braille page is about 11 inches by 11 inches. A Braille cell's size is also specified, though it varies greatly depending on the country. The dimensions of a Braille cell as printed on an embosser are shown in **Figure 1**.

The inventor of the Braille system anticipated the need to convey information in a compact format so that a group of cells could convey a lot more information in the string of letters that make up the cells. The cell's six dots (7) can be combined to make 63 different dot patterns. Despite the fact that there are a total of 64 patterns, the final one is a cell with no dots, which serves as a gap. When viewed through the lens of text representation in a computer, a Braille cell is thus analogous to a 6-bit character code. Almost all information can now be created electronically, and computer use in daily activities (typing and printing) has increased significantly. As a result, more specialized technologies that are accessible to visually impaired individuals are critical,

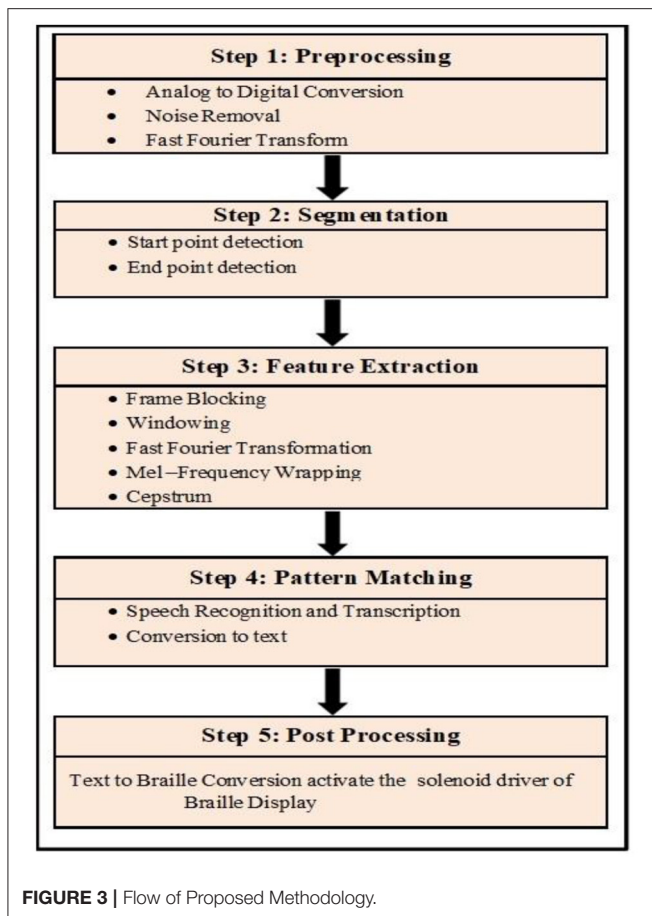


allowing them to use computers for everyday transactions and contribute as much to society and the evolution of IT as others.

## Automatic Arabic Speech Recognition—AASR

Automatic Arabic Speech Recognition (AASR) (8) is a system that converts human speech into text. It's a difficult task because human speech signals are highly unpredictable due to a variety of speaker characteristics, speaking styles, and imprecise background noises (9). Variable-length voice signals must also be translated into variable-length word or label sequences by AASR. An AASR system is made up of many components, as shown in **Figure 2**. A microphone captures the speech signal first and foremost, which is then saved in digital format (10). The voice signal is non-stationary and fluctuates a great deal. As a result, the signal is divided into small chunks known as frames, with each frame's signal becoming more stationary and its features represented by a fixed length feature vector (11, 12). The procedure is referred to as feature extraction (13, 14), and the component that performs it is referred to as the front end. The front end's output is the observation sequence  $X$ , with each observation being a feature vector representing a single frame.

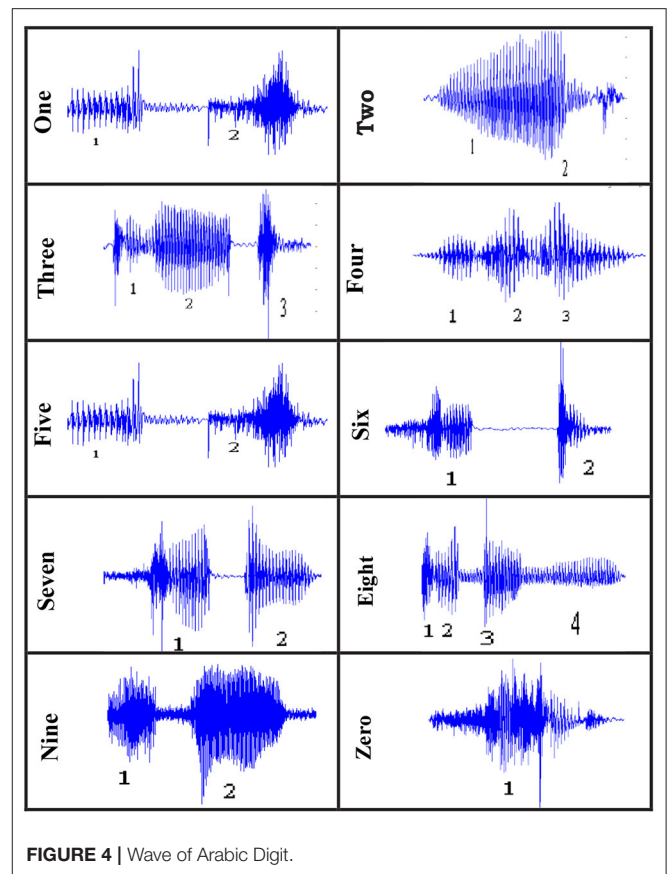
The acoustic model's goal is to simulate the acoustic characteristics of various speech units, as well as the relationship between the acoustic aspects of the entire voice utterance and the utterance label sequence or text. For spoken acoustic signals, the hidden Markov model (HMM) (15) is a popular and successful model. The HMM is very versatile when it comes to modeling variable length sequences, and it tolerates changes in speaking speeds.



It models the likelihood of getting the speech observation sequence given a set of latent (hidden) discrete states. A decoder is used to recognize speech based on the acoustic and language models to model the relationship between an HMM state and the speech observations that pertain to this state. The decoder component searches for the best label sequence for a given voice utterance that maximizes the scores computed by the acoustic (16) and linguistic models. Because the number of viable label sequences grows exponentially with the length of the input feature sequence, the efficiency and speed of the decoder are critical elements of a successful ASR system (17).

### Feature Extraction on ASR

The initial stage in speech recognition is the extraction (13, 14) of a sequence of feature vectors  $X$  that reflects the input voice signal. Feature vectors (18) are available in a wide range of sizes and shapes. In an acoustic model-based system, a good feature vector should include all relevant information for classification of the input signal (19). It's also a good idea to get rid of any data that isn't absolutely necessary. Mel frequency spectral coefficients (MFSC) and Mel-frequency cepstral coefficients (MFCC) are two feature vectors commonly used in academic and commercial ASR systems (20).



The first step is to take short overlapping windows of the speech signal, making it quasi-stationary and incorporating properties unique to single speech units. A Hamming window (or similar bell-shaped functions) is multiplied by this speech window to reduce the effect of discontinuity at the window's two edges. Because distinct speech units and sounds differ primarily in the distribution of energy along different frequencies, it is common practice to convert the speech time domain data into a frequency domain signal using Fast Fourier Transform (FFT) analysis (21). Filter bank analysis is then used to measure the energy in a small number of frequency ranges. These filters are dispersed along the Mel scale frequency to imitate human ear's sensitivity to frequency variances (22). Mel frequency filter bank analysis employs triangular filters. The energy in each filter is estimated to estimate the MFSC feature.

In practice, the log is used to calculate log-MFSC characteristics, which are closer to the sensitivity of the human ear and allow for better discrimination. In fact, neural network models, particularly convolutional neural networks, benefit from the log-MFSC feature. The log-MFSC features are closely correlated because the spectral envelop changes smoothly with frequency and neighboring filters overlap. The discrete cosine transform is used to minimize correlation in order to generate MFCC features (23). When using Gaussian Mixture Models (GMMs), this is especially important when using a diagonal covariance matrix.

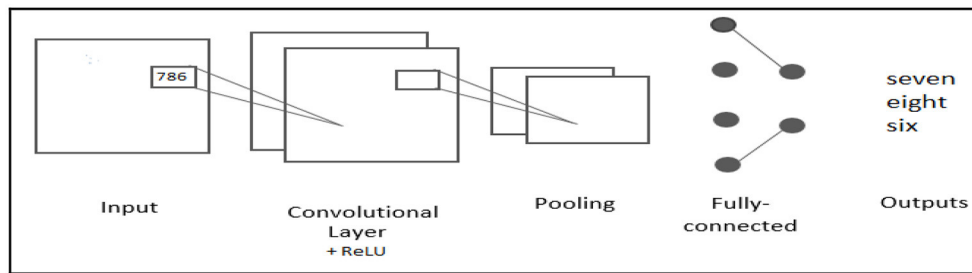


FIGURE 5 | CNN Architecture of Spoken Arabic digit to Text.

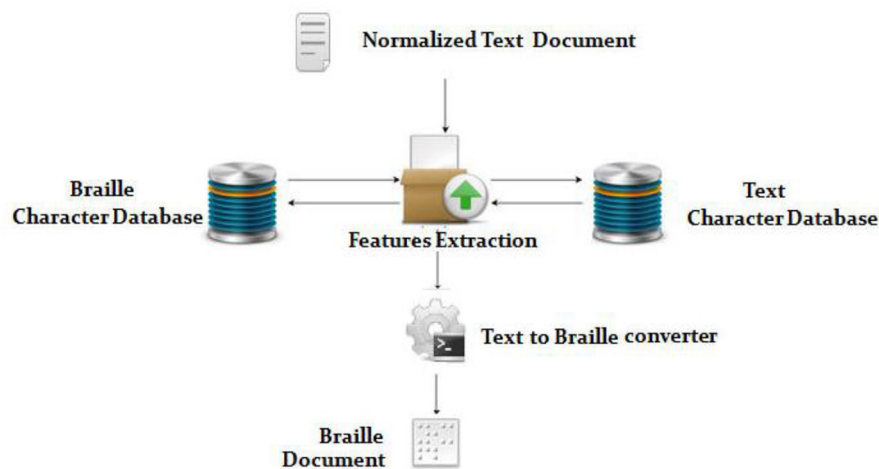


FIGURE 6 | Text to Braille Conversion.

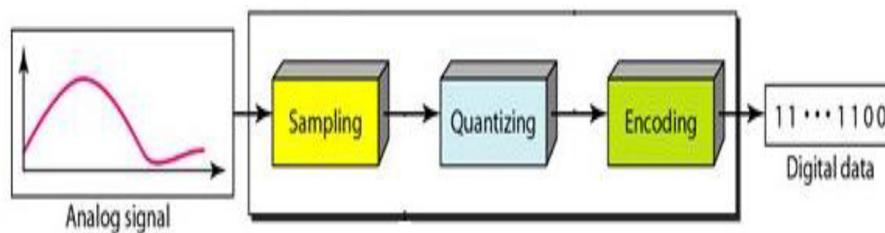


FIGURE 7 | Speech signal Capture.

### HMM Based ASR System

An HMM can be used in a variety of ways to recognize speech. Isolated speech recognition is the most basic type, in which the system’s goal is to classify individual words. In this scenario, each word is modeled using a different HMM model. The number of distinct acoustic portions of the word corresponds to the number of states in each HMM model. To classify the segment, the term associated with the model with the highest likelihood is used. The likelihood is calculated either exactly using the forward probability or roughly using the Viterbi algorithm’s likelihood of the optimum state sequence in this scenario.

Speech recognition that is connected (continuous) is a more advanced technology that allows the speaker to continuously utter a large number of words. In this scenario, different word models are combined to create a large composite HMM. The types of transitions that are allowed between words are defined by a grammar. Furthermore, a language model is frequently used to assign different weights to different word transitions. Speech recognition is accomplished by determining the most likely state sequence and mapping that sequence into a word sequence that represents the uttered sentence. The disadvantage of this model is that learning a good model for each word necessitates a large

amount of training data. Furthermore, the model is incapable of recognizing terms not found in the training data.

To overcome these difficulties, models that represent sub-word units such as phonemes can be used. In these models, each phoneme is represented by a three-state sequence that includes the phoneme’s transitional beginning and end, as well as the middle stationary state. To represent a word, a series of phoneme models are combined to form a word model. A dictionary is used to specify the phoneme models that represent a word. This method has the advantage of sharing each phoneme’s training data across multiple words, reducing the amount of data needed for training. It can also recognize unseen words if they have equivalent entries in the dictionary.

### Neural Network (NN) Based ASR

As a result of improved performance on a variety of speech recognition tasks, deep neural network models (DNNs) have recently resurfaced (26). Previous attempts to use NNs in acoustic modeling attempted to use the NN as a separate model for

detecting simple speech units. To replicate the Viterbi algorithm’s work in Lippmann (27), devised a recurrent neural architecture dubbed Viterbi net. As a result, it can recognize speech almost instantly. However, it lacked a training algorithm. The weights of the Recurrent Neural Network (RNN) were calculated using a trained HMM model. As a neural network architecture, the Time Delay Neural Network (TDNN) was proposed in Waibel et al. (28). The TDNN combines longer input contexts by using a series of consecutive input frames and hidden activations in the top layers. The model was used to classify a small number of phonemes. To successfully use NNs for ASR, a hybrid NN-HMM model was used. In one version of the model that has recently gained popularity, the NN replaces the GMM in scoring speech frames, while the HMM is used to model the speech signal’s sequential temporal behavior.

## RESEARCH METHODOLOGY

The Arabic language is the world’s largest and most experienced Semitic language (26), and it differs from other European dialects such as English in several ways. Arabic is the Arab world’s official language. Arabic is the world’s sixth most communicated language, based on the number of people who speak it as their first language. Furthermore, ~250 million people speak Arabic as their first language. Arabic sentences are written from right to left, and some letters may change shape depending on their position within a word (29).

Speech-to-Braille STB (30) is a branch of computer science that focuses on developing computer frameworks that can recognize expressed words from an amplifier and convert them to messages. Speech to Text is the process of converting or deciphering an audible signal obtained by a receiver or a phone into a string

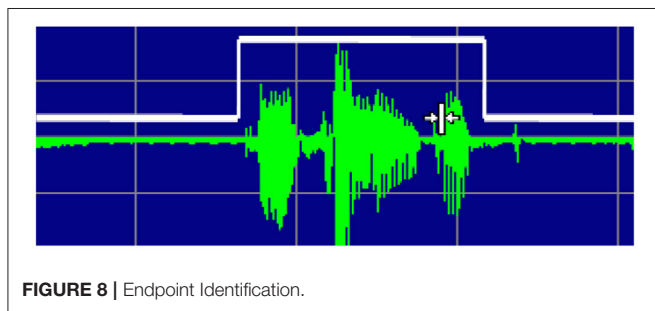


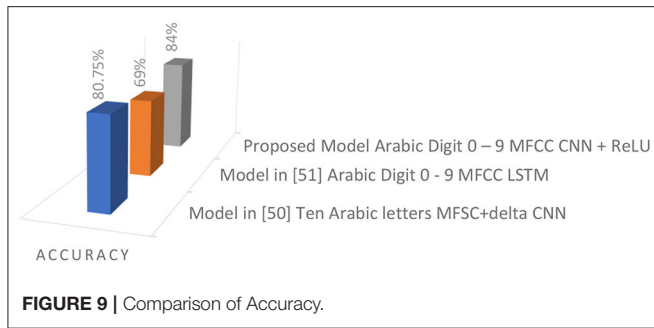
TABLE 1 | Performance Table.

Digit	Correct recognition	Wrong recognition	Accuracy of MFCC	Correct recognition (CNN and ReLU)	Accuracy After CNN and ReLU
1	15	5	75%	18	90%
2	13	7	65%	16	80%
3	13	7	65%	16	80%
4	15	5	75%	18	90%
5	13	7	65%	16	80%
6	12	8	60%	15	75%
7	14	6	70%	17	85%
8	13	7	65%	16	80%
9	14	6	70%	17	85%
0	16	4	80%	19	95%
Accuracy % (24)			69%	Accuracy % after CNN & ReLU	84%

TABLE 2 | Comparison of Accuracy.

Models	Dataset	Features	Recognizer	Accuracy
Model in Rajagede et al. (25)	Ten Arabic letters	MFSC+delta	CNN	80.75%
Model in Wazir and Chuah (24)	Arabic Digit 0–9	MFCC	LSTM	69%
Proposed Model	Arabic Digit 0–9	MFCC	CNN +ReLU	84%





of words. Extraction of highlights from a gathered signal and matching them to a voice test in an information base is used to perceive a word (31, 32).

The goal of the procedure is to create a Digit speech recognition refers to the recognition of utterances in which the user can speak normally without pausing between words, which is similar to how humans speak. Subword sounds, such as phonemes (alphabets in Indian languages) (29), are commonly used as the basic unit of recognition in continuous speech recognition systems. The sequence of phonetic models forms the larger linguistic unit, such as words/sentences, during recognition. Because the number of sub-word units is limited, collecting enough data to train each of these basic sounds in a variety of acoustic environments is not difficult. Adding a new word is simple because all that is required is a description of the new word in terms of previously learned speech sounds. Among the features are speech analysis, training tools, recognition tools, outcome analysis, and labeling tools. It can recognize both single and related words. It is capable of modeling both full words and sub-word units.

The system’s post-processing involves retrieving the exact data associated with the recognized speech. As shown in Figure 3, the fetched Arabic data corresponding to the Arabic Corpus dataset is converted into Arabic Braille script and delivered to the user in the form of Braille Display.

## CONVOLUTIONAL NEURAL NETWORKS FOR ASR

Speech signals can take many different forms that have little to do with the speech’s grammatical content. For example, different speakers may have different voice qualities. There can be a lot of differences between speakers, even if they are the same person. Variations in speaking speed, pitch, and formant frequencies due to changes in vocal tract length, as well as background noise, are examples.

A variety of strategies have been proposed to deal with these differences. In this case, there are two approaches: passive and active. The goal of passive strategies is to standardize speech variances. This can be accomplished by employing special features such as MFCCs, which filter out information that is unrelated to the speech’s linguistic content. The HMM

CHARACTER	BRAILLE	BRAILLE DOTS
number sign		3456
١ 1		3456 1
٢ 2		3456 12
٣ 3		3456 14
٤ 4		3456 145
٥ 5		3456 15
٦ 6		3456 124
٧ 7		3456 1245
٨ 8		3456 125
٩ 9		3456 24
٠ 0		3456 245

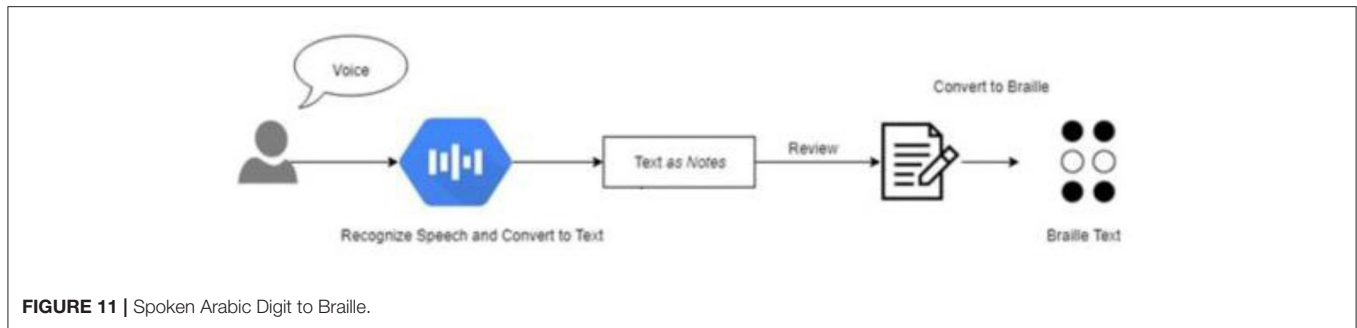
FIGURE 10 | Arabic Digit to Braille.

normalizes rate variations because it is tolerant of variations in speaking pace.

Active approaches aim to estimate some parameters that transform the speech signal or to fine-tune the trained speech models to perform better in new situations. Some speaker adaptation strategies, for example, change the HMM model parameters to reflect the new speaker’s characteristics. Other speaker adaptation strategies convert a new speaker’s speech signal into a canonical speaker that the acoustic model can learn from. The new model or transformation parameters are calculated from some speech adaption data collected from the new speaker in both situations before performing speech recognition on this speaker.

The Convolutional Neural Network (33) (CNN) is a type of artificial neural network that processes data in the form of patterns. CNN is an effective method for building deep learning models that have been used in object recognition applications. CNN (34) was proposed as a method of reducing the amount of data required for traditional artificial neural networks. Three layers make up a traditional Convolutional Neural Network: convolution, pooling, and fully connected layers (35). The pooling layer is used to reduce the number of values in feature maps, while the convolution layer is used to extract features. The best feature maps are passed into the classifier using a fully linked layer (35, 36). CNN is scalable in general and the training procedure takes less time.

The main advantage of the CNN is that domain knowledge can be used to create the CNN structure in order to manage



application-specific variances. In these applications, CNNs are used to take advantage of translation invariance and to tolerate small changes in visual patterns in multiple directions. Weight sharing, local connectedness, and pooling are some of the network topologies used by CNNs, as opposed to standard NNs. As a result, minor image deformations and noise, as well as small rotations and scaling, are less noticeable to the CNN. These invariances are difficult to learn automatically in typical NNs. In a similar way, CNN ideas have recently been applied to voice processing.

Unsupervised learning of acoustic features was achieved using the CNN structure (37). Convolution over time is used in these studies to manage small temporal shifts for better speaker, gender, and phonetic classification representations, but not for audio recognition.

The CNN structure allows us to acquire frequency invariant features when applied along the frequency axis (38, 39). A CNN can also accommodate temporal changes and thus better handle speaking speed variances when applied along the time axis (40). However, the HMM can handle these variances inside the hybrid NN-HMM model, reducing the benefit of applying convolution and pooling over time.

## Description of Arabic Digits

Due to background noise interference, speech is harmed in voice applications. As a result, we can't tell whether a signal contains valid data just by looking at it (41). Many efforts have been made in speaker system independent, automatic Arabic speech recognition for an isolated digit or continuous digit speech for many languages. There is still a scarcity of research on speech recognition in the Arabic language. All speech recognition applications, including those described here, are, however, available in English. Despite the fact that our study focused on Arabic speech recognition in digit contexts (Figure 4).

## Feature Extraction and Classification Using CNN

The feature extraction technique for audio data is depicted in Figure 5 (42–46). It shows the spectrum of audio data before and after feature extraction using the Python's 'sklearn' module. The data was split 80:20, with 80% going into the training dataset and 20% into the test dataset, with 20% of the training dataset being used as a validation set for this Speech recognition

processing (47). The model processes the training, test, and validation datasets, which will be used in the trained model's test dataset prediction analysis.

- In the convolution layers, we created two separate models, each with its own set of parameters.
- In both models, the first 2D-convolutional layer (conv2D) was used to analyze train data, and it has a 128-bit resolution (Model-1).
- A RELU activation function with a filter size of 64 (model-2).
- The output of the first Conv2D layer was sent to the first batch normalization layer.
- For each model, this procedure is repeated twice using the second and third conv2D.
- The first model employs 48 filters with a normalization pool size of (3, 3), whereas the second employs 32 filters with the remaining two parameters unchanged.

After the feature learning stage, the classification technique starts by flattening the resulting vector into a one-dimensional vector. The data was flattened because categorization is easier with a one-dimensional vector. The 1D vector of the flatten layer is then passed through a series of layers.

Finally, the output vector is sent to a 20-unit dense layer, which corresponds to the number of key words in the Arabic dataset. After that, the final result was obtained. The optimizer in this model is 'Adam,' and the loss function is 'Sparse-Categorical Cross-entropy.' Figure 5 depicts the basic architecture of our proposed model.

## TEXT TO BRAILLE

Figure 6 depicts the steps involved in converting text to braille.

- Enter Arabic Text: Entering Arabic text in the provided area or by browsing and attaching a text file (48).
- Validation and data verification of Arabic text: The primary function of this block is to validate and verify the Arabic content extracted from the speech. Validating a meaningful text and verifying the converted word in a database (Text Database) (48) is part of the validation procedure.
- Braille conversion process: The Braille Process converts each character into an equivalent Braille character. The symbols for each and every Arabic digit character can be found in the currently available database (Braille Database) (49–51).

- Arabic Braille Output: After the conversion has been replaced appropriately, the final output, as shown in **Figure 6**, can be handled through the output interface to reach the end user.

## EXPERIMENTAL ANALYSIS

The Methodology of Speech recognition and Conversion to Braille system (30) has following steps.

**Step 1:** The first step in this research is to gather information from a variety of Arabic speakers. Speakers must record the Arabic numerals 0 to 9 and say each Arabic number separately in a clear voice with minimal background noise, as shown in **Figure 7**. As a result, data collection was required in order to develop the speech recognition system; this data was collected by various countries around the world, each with its own dialect for pronouncing words (24, 25). Geographical areas influence how speech is delivered, but they aren't always the source of variation; for example, dialects may be influenced by family history. The voice signal is obtained by recording or capturing the entire audible frequency range of a human voice. The sampling frequency is defined as the number of samples obtained in a second.

**Step 2:** To define where a single utterance begins and ends, the zero-crossing rate (ZCR) and average amplitude thresholds are used. It can be classified as a speech signal segment if the energy in the frame exceeds the energy threshold. However, because unvoiced speech signals are sometimes merged or combined with ambient noises, the marking of start and stop locations may not be precise. Unvoiced speech has a higher ZCR than noise, so it can be used to distinguish between noise and a weak speech signal, allowing for better start and stop point marking, as shown in **Figure 8**.

**Step 3:** Using CNN to recognize speech: The classification stage is followed by a fully connected layer and a softmax layer, and the convolutional layers and pooling layers are used to learn features. Except for the fully-connected layer, all convolutional and pooling layers used the ReLU activation function. To extract features, have used the following settings: 16,000 sample rate, 512 and 160 sample FFT lengths and hop sizes, as well as the Hamming window function. The first CNN layer conducts a convolutional over the input feature vector using 32 ReLU kernels with  $3 \times 3$  receptive fields and a stride of 1 in both dimensions, as shown in **Table 1**.

The proposed model was tested using an Arabic speech corpus of isolated words. The used corpus is subjected to a variety of modifications, including pitch, speed, dynamic range, noise, and time shifting forward and backward. As shown in **Table 2** and **Figure 9**, the maximum accuracy achieved when using the ReLU activation function with CNN was 84%. The ReLU activation function, a non-linear activation function, has gained prominence in this proposed methodology. When comparing the results of this study to previous research, it is clear that the proposed methodology performed better in the AASR.

**Step 4:** Text to Braille Conversion: The result, as shown in **Figure 10**, will be processed through the Arabic Text database, which will identify the Arabic text.

**Step 5:** Postprocessing: The Arabic letter is embossed as Arabic Braille script on the Braille display using solenoid drive (52) as

shown in **Figure 11**. The evaluation is based on how it is handled and prioritized including some other performance parameters. The primary reason for employing the device would be decided by determining if the device's benefit outweighs the negative encountered. Following the prototype's implementation, it is evaluated with a small group of visually challenged people, as well as people who are both visually and hearing impaired.

## CONCLUSIONS

Braille is a text-based communication system that is solely dependent on touch. People with vision and hearing impairments would use it. We propose a Speech to Braille conversion procedure to provide the visually and hearing-impaired people can easily make notes in Braille using the provided methods (53) by simply speaking or playing a tape. This app will help people with both visual and hearing impairments communicate in real time, and it will be a cost-effective way. This research produces output that is comparable to traditional high-cost braille printers, as well as assisting visually and hearing-impaired people in recognizing the Speech and converting into Braille output. Speech recognition is a bonus feature that allows visually and hearing-impaired people to print braille for personal use without a computer or a mediator by simply speaking the phrases to be printed into a microphone.

Arabic digits six are particularly difficult to recognize when spoken. The complexity of the voiced signals of the spoken digits is the reason. Deep learning requires a large amount of data to achieve better accuracy, the accuracy of CNN model could be greatly improved by using a large amount of training data. More digit voice samples with noise, for example, could give us a more robust classifier. It's worth noting that as the amount of training data grows, so does the amount of time spent training.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

AD and SB: conceptualization, investigation, and methodology. SB and RA: funding acquisition and resources. AM: project administration and visualization. AD: writing of the original draft. AM and SB: writing of the review and editing. SB: validation. RA: data curation. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Project No. AN000602).



## REFERENCES

- Fukushima S. *Person With Deafblind and Normalization*. Tokyo: Akashi Shoten (1997).
- Uehara N, Aoki M, Nagashima Y, Miyoshi K. Bidirectional telecommunication using wearable i/o terminal for finger braille. In: *Proceedings of the Human Interface Symposium* (San Diego, CA). (2000). pp. 37–40.
- Hoshino T, Otake T, Yonezawa Y. A study on a finger-braille input system based on acceleration of finger movements. In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. 85*. (2002). p. 731.
- Ochi T, Kozuki T, Suga H. Bracelet type braille interface. *Correspond Hum Interf.* (2003) 51:25–7. doi: 10.5772/51472
- Matsuda Y, Isomura T, Sakuma I, Kobayashi E, Jimbo Y, Arafune T. Finger Braille teaching system for people who communicate with deafblind people. In: *2007 International Conference on Mechatronics and Automation*. Heilongjiang: IEEE (2007). pp. 3202–7.
- Matsuda Y, Sakuma I, Jimbo Y, Kobayashi E, Arafune T, Isomura T. Development of finger Braille teaching system-teaching of dotting finger and position using speech recognition. *J Life Support Eng.* (2007) 19:105–16. doi: 10.1109/ICMA.2007.4304074
- Baboo CD, Devi VA. Computer based Tamil Braille system—a review. In: *11th International Tamil Internet Conference, Annamalai University, Chidambaram, India, Jointly Conducted by International Forum for Information Technology (INFITT) and Annamalai University (Tamil Nadu)*. (2012). pp. 57–61
- Khilari P, Bhoje VP. A review on speech to text conversion methods. *Int J Adv Res Comput Eng Technol.* (2015) 4:3067–72. Available online at: <https://www.irjet.net/archives/V7/i5/IRJET-V7I5854.pdf>
- Devi VA, Suganya MV. An analysis on types of speech recognition and algorithms. *Int J Comp Sci Trends Technol.* (2016) 4:350–5. Available online at: <http://www.ijcstjournal.org/volume-4/issue-2/IJCST-V4I2P62.pdf>
- Ping Z, Li-Zhen T, Dong-Feng X. Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network. *Inform Technol J.* (2009) 8:796–800. doi: 10.3923/ij.2009.796.800
- O'Shaughnessy D. Interacting with computers by voice: automatic speech recognition and synthesis. *Proc IEEE.* (2003) 91:1272–305. doi: 10.1109/JPROC.2003.817117
- Anusuya MA, Katti SK. Front end analysis of speech recognition: a review. *Int J Speech Technol.* (2011) 14:99–145. doi: 10.1007/s10772-010-9088-7
- Jung S, Son J, Bae K. Feature extraction based on wavelet domain hidden Markov tree model for robust speech recognition. In: *Australasian Joint Conference on Artificial Intelligence*. Berlin; Heidelberg: Springer (2004). pp. 1154–9.
- Sárosi G, Mozsáry M, Mihajlik P, Fegyó T. Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment. In: *2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. Brasor: IEEE (2011). pp. 1–8.
- Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE.* (1989) 77:257–86. doi: 10.1109/5.18626
- Walker W, Lamere P, Kwok P, Raj B, Singh R, Gouvea E, et al. Sphinx-4: A flexible open source framework for speech recognition.
- Almeman K, Lee M, Almiman AA. Multi dialect Arabic speech parallel corpora. In: *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. Sharjah: IEEE (2013). pp. 1–6.
- Amin A, Mari JF. Machine recognition and correction of printed Arabic text. *IEEE Trans Syst Man Cybernet.* (1989) 19:1300–6. doi: 10.1109/21.44052
- Sharma HV, Hasegawa-Johnson M. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Comp Speech Lang.* (2013) 27:1147–62. doi: 10.1016/j.csl.2012.10.002
- Awaidah SM, Mahmoud SA. A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. *Signal Process.* (2009) 89:1176–84. doi: 10.1016/j.sigpro.2008.12.022
- Tiwari V. MFCC and its applications in speaker recognition. *Int J Emerg Technol.* (2010) 1:19–22. Available online at: <https://www.semanticscholar.org/paper/Finger-Braille-Teaching-System-173-Finger-Braille-Matsuda-Isomura/8192b9983f231fcb6bf971cf860b8cebb5a3f3f0>
- Farooq O, Datta S. Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Proc Lett.* (2001) 8:196–8. doi: 10.1109/97.928676
- Chen C, Bunescu RC, Xu L, Liu C. Tone classification in Mandarin Chinese using convolutional neural networks. In: *Interspeech* (San Francisco, CA). (2016). pp. 2150–4.
- Wazir AS, Chuah JH. Spoken arabic digits recognition using deep learning. In: *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. Selangor: IEEE (2019). pp. 339–44.
- Rajagede RA, Dewa CK. Recognizing Arabic letter utterance using convolutional neural network. In: *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. Kanazawa: IEEE (2017). pp. 181–6.
- Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc Mag.* (2012) 29:82–97. doi: 10.1109/MSP.2012.2205597
- Lippmann RP. Neural-net classifiers useful for speech recognition. In: *Proceedings International Conference on Neural Networks, Vol. 4* (Piscataway, NJ). (1987). pp. 417–25.
- Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KJ. Phoneme recognition using time-delay neural networks. *IEEE Transac Acoustics Speech Signal Process.* (1989) 37:328–39. doi: 10.1109/29.21701
- Elmahdy M, Gruhn R, Minker W. *Novel Techniques for Dialectal Arabic Speech Recognition*. Springer Science & Business Media (2012).
- Devi VA. Conversion of speech to Braille: interaction device for visual and hearing impaired. In: *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*. Tamil Nadu: IEEE (2017). pp. 1–6.
- Ajami Alotaibi Y, Alghamdi M, Alotaiby F. Speech recognition system of Arabic alphabet based on a telephony Arabic corpus. In: *International Conference on Image and Signal Processing*. Berlin; Heidelberg: Springer (2010). pp. 122–9.
- Alotaibi YA. Investigating spoken Arabic digits in speech recognition setting. *Inform Sci.* (2005) 173:115–39. doi: 10.1016/j.ins.2004.07.008
- Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. *IEEE ACM Trans Audio Speech Lang Process.* (2014) 22:1533–45. doi: 10.1109/TASLP.2014.2339736
- Nazih W, Hifny Y, S Elkilani W, Mostafa T. Fast detection of distributed denial of service attacks in VoIP networks using convolutional neural networks. *Int J Intellig Comput Inform Sci.* (2020) 20:125–38. doi: 10.21608/ijicis.2021.51555.1046
- Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* (2018) 9:611–29. doi: 10.1007/s13244-018-0639-9
- Abeßer J. A review of deep learning based methods for acoustic scene classification. *Applied Sciences.* (2020) 10:2020. doi: 10.3390/app10062020
- Lee H, Pham P, Largman Y, Ng A. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Adv Neural Inform Process Syst.* (2009) 22:1–5. doi: 10.1109/ICASSP.2013.6638952
- Abdel-Hamid O, Mohamed AR, Jiang H, Penn G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto: IEEE (2012). pp. 4277–80.
- Deng L, Abdel-Hamid O, Yu D. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC: IEEE (2013). pp. 6669–73.
- Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: *Interspeech, Vol. 2013* Lyon. (2013). pp. 1173–5.
- Veena KV, Mathew D. Speaker identification and verification of noisy speech using multitaper MFCC and Gaussian Mixture models. In: *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*. Thrissur: IEEE (2015). pp. 1–4.
- Hammami N, Bedda M. Improved tree model for Arabic speech recognition. In: *2010 3rd International Conference on Computer*

- Science and Information Technology*. Vol. 5. IEEE (2010). pp. 521–6.
43. Mahmood A, Alsulaiman M, Muhammad G. Automatic speaker recognition using multi-directional local features (MDLF). *Arabian J Sci Eng*. (2014) 39:3799–811. doi: 10.1007/s13369-014-1048-0
  44. Elkourad A. Arabic isolated word speaker dependent recognition system. In: *Source Framework for Speech Recognition*. (2004).
  45. Alsulaiman M, Mahmood A, Muhammad G. Speaker recognition based on Arabic phonemes. *Speech Commun*. (2017) 86:42–51. doi: 10.1016/j.specom.2016.11.004
  46. Khelifa MO, Elhadj YM, Abdellah Y, Belkasm M. Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. *Int J Speech Technol*. (2017) 20:937–49. doi: 10.1007/s10772-017-9456-7
  47. Bhalla K, Koundal D, Bhatia S, Khalid M, Rahmani I, Tahir M. Fusion of infrared and visible images using fuzzy based siamese convolutional network. *Comput. Mater. Con*. (2022) 70:5503–18. doi: 10.32604/cmc.2022.021125
  48. Cong L, Asghar S. Robust speech recognition using fuzzy matrix/vector quantisation, neural networks and hidden Markov models. In: *International Conference on Advances in Pattern Recognition*. London: Springer (1999). pp. 255–64.
  49. Merazka F. VQ Codebook design using genetic algorithms for speech line spectral frequencies. In: *International Symposium on Intelligence Computation and Applications*. Berlin; Heidelberg: Springer (2012). pp. 557–66.
  50. Samal A, Parida D, Satapathy MR, Mohanty MN. On the use of MFCC feature vector clustering for efficient text dependent speaker recognition. In: *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*. Cham: Springer (2013) pp. 305–12.
  51. Ananthi S, Dhanalakshmi P. SVM and HMM modeling techniques for speech recognition using LPCC and MFCC features. In: *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*. Cham: Springer (2014). pp. 519–26.
  52. Baboo CD, Devi VA. Tamil braille system: A conversion methodology of tamil into contracted braille script (grade2). In: *UACEE International Journal of Advances in Computer Science and its Applications - IJCSIA*, Vol. 3. (2013).
  53. Devi A. *Portable Device Converting Braille and Text Document to Speech Output Using Raspberry Pi*. IJCTA (2016). pp. 4809–18.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bhatia, Devi, Alsuwailem and Mashat. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.