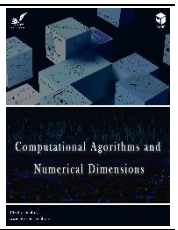
Computational Algorithms and Numerical Dimensions

www.journal-cand.com

Com. Alg. Num. Dim Vol. 1, No. 2 (2022) 72-83.



Paper Type: Original Article

Speech Recognition System Based on Machine Learning in Persian Language

Shahed Mohammadi[1], Niloufar Hemati[2,*], Neda Mohammadi[3]

[1] Department of Computer Since and Systems Engineering, Ayandegan Institute of Higher Education, Tonekabon, Iran; Mohammadi.ac@gmail.com.

[2] Department of Computer Science, Islamic Azad University Central Tehran Branch, Tehran, Iran; Niloofarhemati628@gmail.com.

[3] Department of Industrial Engineering, Sadra University, Tehran, Iran; Setare.mh66@gmail.com.

Citation:

Mohammadi, Sh., Hemati, N., & Mohammadi, N. (2022). Speech recognition system based on machine learning in Persian language. *Computational algorithms and numerical dimensions, 1*(2), 72-83.

Received: 06/11/2021 Reviewed: 22/01/2022 Revised: 02/03/2022 Accept: 27/04/2022

Abstract

In today's world, where speech recognition has become an integral part of our daily lives, the need for systems equipped with this technology has increased dramatically in the past few years. This research aims to locate the two selected Persian words in any given audio file. For this purpose, two standard and native datasets were prepared for this model one for train and the other for the test. Both datasets were converted into images of audio waveforms. Using the object detection technique, the model could extract different bounding boxes for each test audio, and then each box image goes through a CNN classifier and returns a corresponding label. Finally, a threshold is set so that only boxes with high accuracy are displayed as output. The results showed 93% accuracy for the CNN classifier and 50% accuracy for testing the model with object detection.

Keywords: Speech recognition, Signal processing, Object detection, Neural network, Deep learning.

1 | Introduction

The extraordinary human ability to process and understand audio signals has attracted the attention of researchers to simulate this human ability in intelligent electronic systems. Also, in today's life, we are faced with a substantial volume of data. Due to the growth of this data, most of which is audio, speech processing, and recognition has become an integral part of our lives and also increases the demand for building and developing systems equipped with a speech recognition engine. We can find examples of these systems almost everywhere [1] from the intelligent electronic assistants found on most mobile phones and other applications.

Most of these systems have a speech recognition engine. In general, speech recognition technology allows computers to receive and understand speech. The system receives speech from the user using sound waves recorded by a microphone or signals extracted from a video. After, this audio file converts into a set of electrical signals. Then the resulting signals can be decomposed using advanced

Corresponding Author: Niloofarhemati628@gmail.com

https://doi.org/10.22105/cand.2022.146462

signal processing technologies to process and understand the words spoken by the user. Finally, this set of algorithms can determine what the user is saying.

First, the model should extract the audio features similar to how the human ear can understand the words and every slight change in frequency. Our model must be sensitive to changes between different sounds and at the same time be able to understand the common features between them. For this purpose, in this research, computer vision algorithms have been used due to their high ability to extract features. The benefit of computer vision systems is more on image analysis capabilities, extracting useful information from them, and detecting objects. The object detection algorithm used in this research is one of the standard algorithms in computer vision. This technology can simultaneously detect one or more objects in an image with very high accuracy.

Object detection now has many applications in the industry and many models like [2] were developed, such as face recognition, which uses each person's unique features for authentication, and even in self-driving machines, which has become a popular topic in this field to separate objects from humans. The database used in this research is a native database consisting of two types of Persian words. Furthermore, these two words are recorded with different tones and different frequencies. The goal is to find out whether or not our native word is pronounced in one sentence. For this purpose, the object recognition technique with CNN layers is used, which looks inside the file for the wanted words. The process is similar to detecting cats and dogs in a photo, in which computer vision algorithms can distinguish between dogs and cats in the photo with great accuracy. The researcher converts audio files into images to use powerful computer vision algorithms to detect speech, as fully explained in the third section. The words expressed in each audio file play the role of objects in an image.

At first glance, the algorithm used in computer vision [3] and speech recognition [4] due to different applications in the industry may seem to have a different purpose. Still, the ultimate goal of all these algorithms is to learn the existing patterns inside any piece of information, and for this reason, the researcher has tried to combine these two methods for word detection in speech. In this paper, previous studies have been shown in the second part. Also, in the third part, the details of the researcher's work and the method used in this research are described as methodology. In the fourth part, different methods with answers are compared.

2 | Related Works

Acoustic Models (AMs) based on transformers are proposed and evaluated for hybrid speech recognition. This work suggests several options for constructing deep transformers, including embedding and iterated loss methods. Furthermore, they present an early study of incorporating a limited right context into transformer models, which enables streaming. The transformer-based AM outperforms the standard n-gram Language Model (LM) by 19% to 26% in Librispeech, a widely used benchmark. Additionally, the findings were also confirmed with an internal dataset that is several times larger [5].

Using an end-to-end neural acoustic model, the author proposes a new approach for Automatic Speech Recognition (ASR). In addition to the 1D convolutional layers with separable channels, batch normalization, and ReLU layers, a multi-block model is composed of residual connections between the blocks. It is then trained with the CTC loss function. The proposed network performed well with fewer parameters compared to other models using LibriSpeech and Wall Street Journal. Moreover, the authors showed that this model could be adaptable to new datasets [6].

AMI, Broadcast News, Common Voice, LibriSpeech, Switchboard/Fisher, Tedlium, and Wall Street Journal datasets train SpeechStew, a speech recognition model based on multi- datasets: AMI, Broadcast News, Common Voice, LibriSpeech, Switchboard/Fisher, Tedlium, and Wall Street Journal. It is not rebalanced or reweighted in any way when datasets are mixed in SpeechStew. Despite the lack of an external LM, SpeechStew produces SoTA or near SoTA results in a variety of tasks. Their results indicate that their

Word Error Rate (WER) on the AMI-IHM is 9.0%, the Switchboard is 4.7%, Call Home is 8.3%, as well as WSJ is 1.3%, which significantly outperforms previous work with such strong external LMs [7].

In this article, the authors present SpecAugment, a simple method of enhancing speech recognition data. A neural network is fed directly with the inputs from SpecAugment in this study. Feature warping, masking blocks of frequency channels, and masking blocks of time steps are the three components of the augmentation policy. End-to-end speech recognition tasks have been completed using SpecAugment on Listen, Attend, and Spell networks. Compared to previous work, they achieve outstanding performance on LibriSpeech 960h and Switchboard 300h tasks. Their test-other WER on LibriSpeech was 6.8% without a LM, and 5.8% with shallow fusion [8].

Despite the interest in Convolutional Neural Networks (CNNs), their performance remains behind other well-known methods despite promising results. ContextNet, the proposed CNN-RNN-transducer architecture, was investigated in this paper to overcome this problem. ContextNet adds squeeze-and-excitation modules to convolution layers to incorporate global context information. With only 10M parameters, ContextNet achieved a WER of 2.1%/4.6% on the clean/noisy LibriSpeech sets without external LM, 1.9%/4.1% with LM, and 2.9%/7.0% with only 10M parameters [9].

Most object detection algorithms assume that training and test data come from the same distribution, but this does not always hold in practice. To reduce domain discrepancy, they design two domain adaptation components, on image level and instance level: 1) domain style adjustment on the image level, and 2) domain adjustment on the instance level. Their approach is based on the latest state-of-the-art Faster R-CNN model and two domain adaptation components, on image level and instance level. A domain classifier is learned by applying adversarial training to the two domain adaptation components based on the H-divergence theory. A domain-invariant Region Proposal Network (RPN) in the Faster R-CNN model is learned by combining both domain classifiers on different levels with consistency regularization. Despite the interest in CNNs, their performance remains behind other well-known methods despite promising results. ContextNet, the proposed CNN-RNN-transducer architecture, was investigated in this paper to overcome this problem. ContextNet adds squeeze-and-excitation modules to convolution layers to incorporate global context information. With only 10M parameters, ContextNet achieved a WER of 2.1%/4.6% on the clean/noisy LibriSpeech sets without external LM, 1.9%/4.1% with LM, and 2.9%/7.0% with only 10M parameters [10].

With Deep Neural Network (DNN) based AMs, ASR has improved significantly. It is worth mentioning that it is not always possible to use DNN-based systems in reverberating environments. CNN AMs performed less than DNN AMs in distant speech recognition. The multiresolution CNN presented in this work would improve performance if two streams were provided: one for the wideband feature and one for the narrowband feature and window. WER increased by 8.79% for ASR tasks and 8.83% for simulated test results in the REVERB 2014 experiments [11].

Acoustic-lexicon-LMs such as Hidden Markov Models (HMM) form a part of conventional ASR built on HMM/DNN. In the context of ASR, there are two main types of end-to-end architectures: attention-based methods utilize attention mechanisms to align acoustics and symbols, and Connectionist Temporal Classification (CTC) uses a Markov-based approach to efficiently solve sequential problems. A hybrid CTC/attention end-to-end ASR methodology is presented in this paper that combines the advantages of both architectures in training and decoding. In order to improve robustness and achieve fast convergence, they use a multi-objective learning framework [12].

Acoustic modeling has benefited from the application of DNNs. The WER of CNNs is 4–12% higher in comparison with DNNs since they are a more advanced version of DNNs. CNNs are more effective at recognizing speech when there are spectral variations and local correlations in the signal. Acoustic data can be reinforced with higher-level representations when Bidirectional Long Short Term Memory (BLSTM) is used because they produce a higher recognition rate. It is essential for the speech signal to

have spatial and temporal properties for it to be recognized, so two different networks were combined. To make the continuous speech recognition task more effective, a hybrid architecture of CNN-BLSTM is presented. Moreover, they investigate how to achieve a high recognition rate with CNN by implementing different strategies such as weight sharing, the number of hidden units, and optimize pooling. Specifically, the focus of this paper is to identify the number of layers of BLSTM which are effective [13].

They further developed this architecture for noise-robust speech recognition based on their previous work on deep CNN. Researchers consider the best configuration for the sizes of filters, pooling, and input feature maps for the proposed deep CNN architecture. To accommodate the addition of additional layers of convolution, the filter and pooling sizes have been reduced, and the dimensions of input features have been extended. To make the very deep CNN more robust for speech recognition, we study the strategy of pooling, padding, and selecting input feature maps. Analyzing the architecture in depth revealed key characteristics, such as the compact model size, fast convergence speed, along robustness to noise. This study evaluates two tasks with additive noise and channel mismatch with the proposed new model: the Aurora4 task with multiple additive noise types and the AMI meeting transcription task with solid reverberation [14].

Long Short-Term Memory (LSTM) is a popular technique for speech recognition. In order to improve the accuracy of their predictions, machine learning researchers are building increasingly complex models. This type of model requires high computing power and a large amount of memory. Another problem is that Data centers built on these models must have a high Total Cost of Ownership (TCO). LSTM models may be compressed by 20× (10× from pruning and 2× from quantization) with negligible accuracy loss using the load balance-aware pruning method. The presented model is also convenient for parallel processing [15].

To learn cognitive emotions, researchers used semi-CNNs. In order to learn candidate features from unlabeled data, a convolutional neural network with reconstruction penalty is used in the first stage of training. Second, the features are analyzed for their saliency, orthogonality, and discrimination by using a semi-CNN. The researchers found that their approach outperformed SER features such as distortion of the speaker in complex scenes on benchmark datasets [16].

The method uses natural images to detect objects of a certain class, such as pedestrians and cars. According to their previous work, each object part was a probability vote for the centroid of a whole object and each detection hypothesis was a Hough image that cumulatively accounted for all the votes. Methods such as these use discrimination to identify objects. Their algorithm employs a class-specific Hough forest to locate objects based on their appearance [17].

Based on an HMM-based speech recognition system, linear transformations are investigated in this paper. Maximal likelihood-trained transformations are evaluated using adaptation data. It is not considered to perform a feature-space transform for linear transformations but only to perform a model-based linear transformation. There are two kinds of model-based transforms compared here: 1) constrained, where variance equals mean, and 2) unconstrained, where variance and mean can be mixed. All possible transformations are considered in the reestimation formulae. They also offer a new and efficient model-space transformation that uses full variance, plus the full-diagonal case of a constrained model-space transform [18].

The authors present an analytical method for compensating for the environment in speech recognition. Earlier attempts to solve the problem of noise in speech relied on overly simplified mathematical representations of the effects of noise, or large environment-specific adaptation sets. There are methods that utilize simultaneous degraded and clean speech recordings as adaptation data. Using the Vector Taylor Series(VTS) expansion, they investigate the impact of unknown additive noise and unknown linear filtering on speech statistics in a transmission channel [19].

Using the lattice-based framework described in this paper, this paper describes and evaluates the discriminative training of large vocabulary speech recognition systems using HMM with Gaussian mixtures. HMM systems have been trained using the Maximum Mutual Information Estimation (MMIE) criterion on up to 265 hours of training data. Using discriminative training in speech recognition at this scale has never been attempted before. This article discusses MMIE lattice-based implementation of a Baum-Welch algorithm, enabling the training of such extensive systems [20].

According to the authors, all HMM states should exhibit the same Gaussian Mixture Model (GMM) structure that includes the same number of Gaussians. Specifically, each state is represented by 50 vectors of varying dimensions; the vector space is mapped to the space of parameters of the GMM through a global mapping. As a result, the extra structure of the model appears to yield better results than a conventional model, while maintaining compatibility with most standard techniques [21].

It has not yet been determined how AM and FM contribute to speech recognition in terms of their relative contributions to communication. To close the gap between AM and FM, they synthesized slowly varying speech sounds and conducted listening tests on subjects with and without cochlear implants that used stimuli with different modulations. AM with a limited number of spectral bands may be sufficient for speech recognition in quiet, but FM significantly improves speech recognition in noise, as well as speaker and tone identification [22].

3 | Methodology

In this paper, the main goal is to train a computer vision model and use it for a speech recognition task. The task that was tried to be overcome is the detection of two Persian keywords inside of any given audio file. The overall goal for the model is to output whether "Taghaza" or "Hoviat" (the two selected native words) are in a given sentence or not. If they have appeared in the sentence, the model should identify the class name of the word and the location of the word. That led to the development of this model that looks for these two native words in a sentence and can give the location of those two words as its output. This model uses an object detection algorithm with a CNN classifier to find the targeted words.

Now, before examining the different parts of the model, first, the native database prepared for this research is introduced in Section 3.1. Then the researcher goes to the various stages of the proposed method.

Before going into the details of the preparation of the dataset, the block diagram of the presented method is shown in *Fig. 1*.

As you can see in *Fig. 1*, this model is divided into three main steps: 1) preprocessing, 2) feature extraction and classification, 3) post-processing. The model works in such a way that when the audio will consider as input, it goes through each step, and then the output will appear that shows whether the selected words have been presented in that audio or not.

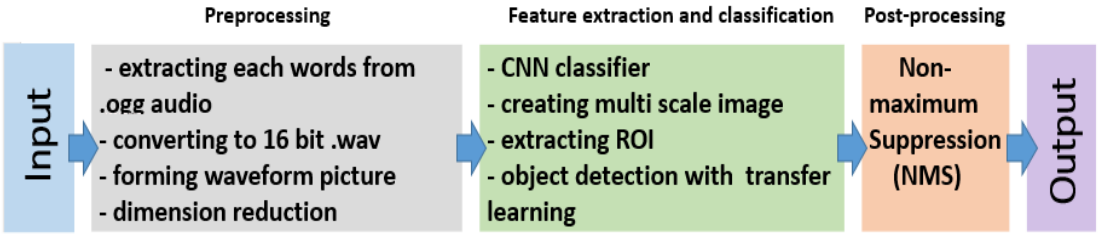


Fig. 1. Block diagram of the proposed model.

3.1 | Dataset

The dataset used by the model is a native dataset consisting of images of audio signals for the two selected Persian words, "Taghza" and "Hoviat". This dataset was recorded by 63 people in different age groups between 17 and 70 years old with different genders.The initial dataset consists of 63 audio files for the word "Hoviat". Each audio file was recorded by one person and contained 15 different pronunciations of "Hoviat". Similarly, 63 audio files were recorded for the word "Taghaza". From each audio, 15 smaller audios with varying accents were created.

As mentioned, the audio files in the primary dataset were recorded by many people of different ages and genders. The overall goal of creating such a rich dataset was to use this wide range of pronunciations for these two words to provide the model with a wide range of patterns of people's audio frequencies. Initially, 63 people were assigned to prepare this dataset, but in the preprocessing section, some files with high amounts of noise were removed from the dataset due to unwanted noise in some files.

The original audio data was in OGG format. After performing preprocessing, described in detail in Section 3.2, and the general form of the data, it was changed to 910 words of "Taghaza" expressed by 63 different people and 855 words of identity defined by 60 other people. Before feeding the data to the CNN classifier, a dataset consisting of images of waveform signals of these audio files was created.

Next, to test the performance of the model, a smaller dataset was prepared, which included 76 audio files, including sentences containing the selected words and 50 audio files without having those words. Like the training dataset, the test dataset was recorded by people of different age and gender groups. After the dataset description, each step in *Fig. 1* is explained step by step.

3.2 | Preprocessing

After preparing the dataset, the first part is to implement preprocessing operations on the prepared data. The overall goal of this section is to extract each word from the given audio file (which includes the average of 15 words per file), convert the data into the correct format, and remove unwanted noise that reduces the model's accuracy. As explained in 3.1, 63 audio files containing 15 times the pronunciation of the word "Hoviat" and 63 audio files containing 15 times of the word "Taghaza" were individually recorded. These files must be converted from the existing multi-word audios into single-word files.

In general, every two audio files, including 15 words for "Taghaza" and 15 words for "Hoviat", were taken from one selected announcer. Because the model must learn the patterns using the structure of each term, these audio files were broken down into words expressed within each audio file. And on average, 10 to 15 words were extracted from each audio file in the initial dataset and placed in the appropriate group. So 945 words for "Taghaza" and 945 words for "Hoviat" were extracted from the original audio files. These files are converted from OGG format to the appropriate format (16-bit WAV format).

Then 30 waveform signals of 16-bit mono-channel audio files with a sample rate of 16,000 were produced for each person. Then audio files that reduce the model's accuracy were removed from the dataset. These files usually included files with high background noise or in which people did not correctly pronounce the selected words. Then after deleting the noisy files, it remains 910 files for "Taghaza" and for the "Hoviat", 855 audio files and files with the appropriate format were placed in the dataset. A vector containing the waveform signal of the audio was extracted from each file. After extracting the audio signal's vector from each file, each vector was used to create an image of a waveform audio signal. Then the collection of those images was used in training the classifier.

In the end, due to the ease of image processing in the convolutional layers and dimension reduction, the images produced by the waveform signals of the audios are converted to black and white photos. After

training the model using these images, the model can accurately distinguish the word "Taghaza" from "Hoviat", which is explained in the next section.

3.3 | Feature Extraction and Classification

The various steps of the proposed method are shown in the *Fig. 2*. At first, there are two types of input: training inputs for the classifier and testing inputs for calculating the model's performance. Training inputs are audio files containing several words, are then split into each word with a proper format and after than is converted into the images of waveform signals in the preprocessing section. The test audio is also converted into a waveform image in the preprocessing section as well. Then the training images are fed into the CNN classifier after that the object detection model uses the test image to find the targeted words. The picture, as shown above, is used to create an image pyramid, and each scale goes throw the CNN classifier, and after the output of this section is the waveform image with different scales bounding boxes. A threshold was applied to bounding boxes in the post-processing section, so only one with higher accuracy remains.

The model's output is the bounding box and the label of the word. After understanding the general idea of the model shown in *Fig. 2* and the knowledge from the previous sections, it's time to go into details of the feature extraction and classification part of the proposed work.

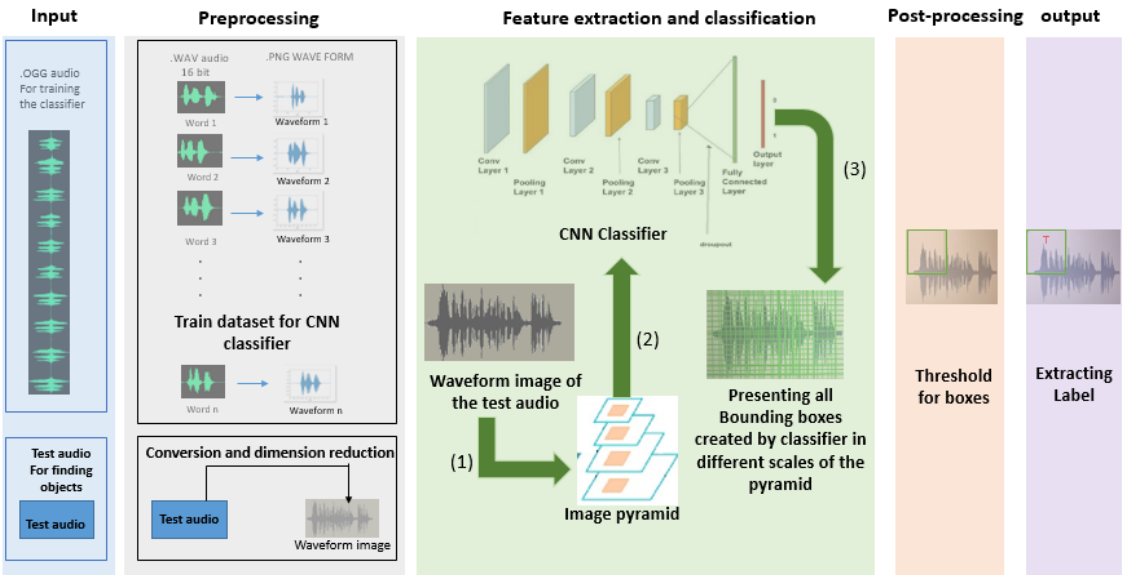


Fig. 2. The general scheme of the proposed method is shown in this figure.

In *Fig. 2*, there are two types of inputs, one for training the CNN classifier and the other for testing the object detection. All inputs go through the preprocessing section, and after extracting waveform images from them, the training inputs are used for the classifier. The test image is converted to a multi-scale image pyramid after applying bounding boxes for each scale of the pyramid, a threshold is applied, and the output is a bounding box with a label for the founded word

The waveform images of the audio files produced in Section 3.3 were given directly to the convolutional classifier for training, resulting in 99.7% accuracy. This classifier consists of three convolution layers, and we have max-pooling between each. The main purpose of max-pooling is to subsample the input image to reduce the computational costs, memory usage, and the number of parameters, which reduces the risk of overfitting. Reducing the size of the input image also reduces the neural network's sensitivity to image displacement. After these three convolution layers, we use a hidden Dense layer and one dense layer for the output, with one neuron. The activation function of the last layer is the sigmoid function, and the other layers are the Relu functions. A flatteng layer is also used to represent neurons' data as a vector, and then it is passed to a fully connected layer.

In the model implementation, the Softmax function is used in the last layer to recognize selected groups of images. With the Softmax function, the output of the neurons of the previous layer is such that the value of each is a number between 0 and 1, and the sum of all will be equal to one. This function is used for the last layer of classification problems. In this problem, one neuron with the sigmoid function is used instead of using two neurons with the Softmax function. This function returns a number between 0 and 1. Because there are only two classes in the dataset, the value of one of the neurons is equal to the difference of the other with the number 1. So the number 0 corresponds to one of the classes (for example, "Hoviat"), and the number 1 corresponds to another class ("Taghaza"). In this case, instead of two neurons in the output layer, the same thing is done with one neuron.

In the following, the object detection section is prepared.

Object detection aims to process and detect where the word "Taghaza" and "Hoviat" is used every time a waveform image is received from audio with an average length of ten seconds for this goal. The model uses the classifier described above. In a way, the object detection model detects targeted words by looking at the different parts of the waveform images and gives these fragments to the classifier. Classifier, which was trained with many words for "Hoviat" and "Taghaza", get this box provided by the object detection and predict which word it is. If the accuracy for either of the terms isn't acceptable, it's neither of them. When it accurately recognizes the word's location, it displays it as the output.

First, from the image that contains the waveform signals for "Taghaza" and "Hoviat", different scales with different sizes are prepared. These different scaling called image pyramids. This pyramid representation is an image and signal processing technique to represent a single image using a set of cascading images. At the bottom of the pyramid is the original image with its original size, and in each layer of the pyramid, if upwards, the size of the images changes and shrinks. Such a pyramid of images makes it possible to find objects of different sizes at once. The image is progressively subsampled until some stopping criterion is met, which is usually when a minimum size has been reached, and no further subsampling needs to occur.

In the next part, a sliding bounding box for object detection is defined to look inside the images generated by the image pyramid. And check for each of the words. This bounding box is in the form of a rectangle. The dimensions of this rectangle should be the same when searching inside all the pictures in a level of the pyramid, but before that, suitable size can be defined for it. Otherwise, the default size for the box is selected, which is a rectangle with a length of 200, And the width is 150. As mentioned, it can be changed.

Now, for each scale of the pyramid, we execute the sliding box for the multi-scale images which was generated by the image pyramid. Each movement of the box Region of Interests(ROI) of the rectangular bounding boxes is extracted from each scale of the image pyramid. The rectangular region of interest is generated from the pyramid, and the output of the sliding window as its label goes through the classifier. So to this point, each produced box should go through the classifier to determine whether the selected word is present inside the box or not, and for that, the researcher used the prediction of the model.

Due to prevent training the model over and over for the detection process in each audio, transfer learning is used to predict the classifier at each step of the bounding box. After using the classifier's pre-trained weights, the probability of the presence of selected words is calculated for each box. For each box, the word with the higher probability was chosen as the label of that box. So for the given waveform image, after implementing preprocessing on the test audio, which was explained in Section 3.2, there are many boxes and a label assigned to it with a probability of the given label.

3.4 | Post-Processing

In the previous section, we used the images created in preprocessing to train the CNN classifier. The test image was taken, and after implementing preprocessing in Section 3.2, the object detector began its work. Then using image pyramids, multi-scale images were created from the original test image. Then a sliding

bounding box was applied in each level of the pyramid to find multiple objects of different sizes in the original waveform picture. Each box then goes through the classifier to determine the probability of the presence of each word in those boxes. A threshold should be applied in this section to minimize the number of these boxes as many as possible. A threshold was applied using the probability of the desired words inside the boxes, which was given by the classifier. Then threshold (NMS) sets the minimum accuracy to 95% to increase the model's performance. Instead of having multiple boxes for each waveform signal image, there is only one box with the highest probability. Otherwise, the model for each photo might show various boxes of different sizes and probability instead of the actual location of the wanted words. The solution to the problem is to apply Non-Maxima Suppression (NMS), which removes the low probability overlapping bounding boxes in favor of the more confident ones. The output of the model may even result in finding only one word. Or, if it does not find a word in the output, the output does not show a picture.So far, the details of the method have been presented. Below, the results of the method are discussed.

4 | Results

In the third section, the different stages of the model were fully explained. This section deals with the results of the model. In general, due to the rich dataset containing a wide range of different pronunciations and accents of the two selected words, results on the test data shown in *Figs. 4* and *5* were acceptable. CNN Classifier resulted in 99.7% and the object detection precision resulted in 50.0%(which is compared in Table 1 with other models). So, It could be established the model can be used on different audio files spoken by people of different ages and genders or with various accents. It can find these two three-syllable words in that audio files with good precision.

Fig. 3 shows the different steps of word recognition in an audio file, which was discussed in detail in the third section. After being converted to the appropriate format, the audio file must be converted to a waveform image and then through the image pyramid to create multi-scale images. A sliding box is applied to each image resulting from the pyramid. The model looks for objects in different sizes. Finally, the threshold is set after creating different boxes with different sizes on the image surface. The box where the word is most likely to be present is displayed as an output next to the label corresponding to that word. The advantage of this model is that in the case of expanding the dataset, in other words, a more powerful classifier can be created to cover a more diverse range of terms by having the appropriate data for different words.

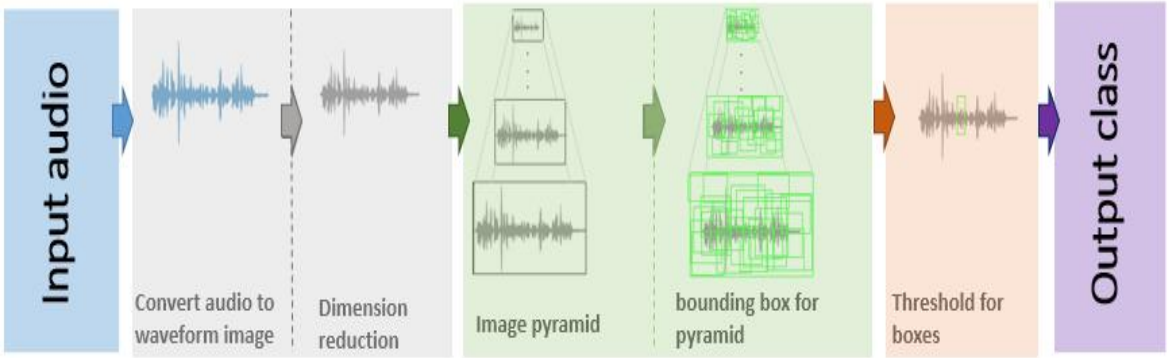


Fig. 3. This figure shows the process of finding the selected words inside an audio signal using the discussed method in previous sections.

The process of finding the targeted words inside an audio signal is shown in *Fig. 3*. as it was discussed in previous sections, the audio for testing the model will be changed into the proper format and then it will be converted to a waveform image of that input. Then it goes through the image pyramid to get different image scales. Then for each scale, bounding boxes will be implemented, and after setting a threshold for boxes, the output, including the box containing the selected word with the label corresponding to that word, will appear. And all of these steps are gathered in *Fig. 3*.

After this, results from other object detection methods are gathered inside *Table 1*. using data inside this table; the proposed model can be compared to the different standard models. In *Table 1*, the model proposed in this article is compared with other standard models using the standard dataset in [3]. From the results, it could be established that the proposed model is a standard model comparable to other common models in object detection and it can be used in speech recognition systems.

Table 1. The proposed model is compared using other object detection models with the help of a standard dataset.

N	Method	AP
1	CoupleNet [23]	34.4
2	cascade r-cnn [24]	42.8
3	nas-fpn [25]	48.0
4	Discussed Method	50.0

Due to the linguistic differences in the structure of the two words "Taghaza" and "Hoviat", finding the two words is also different. Apparently, the model finds the word "Hoviat" better than "Taghaza" in a spoken sentence. In general, one problem with this model is that the model may, in some cases, misdiagnose the combination of three words in a sentence with a three-syllable word.

Also, it should be mentioned that because of the difficulty of finding annotations for audio data, this model can even find word annotations inside audio and the location of words inside the audio with the corresponding label in the output.

5 | Conclusion

According to the content presented in the previous sections, despite the high accuracy of the classifier with 99.7%, the object detection part of the model led to only 50% accuracy in finding two Persian words on the test data. Moreover, some adjustments can be made to increase the model's accuracy.

If more attention is paid to the box design, it may lead to higher accuracy.

One idea for better accuracy is to use spectrogram data instead of waveform signals because spectrograms have shown outstanding performance in speech recognition systems. Another idea is to resolve the undetected words problem using an algorithm to determine the size of the bounding boxes dependent on the size of the targeted word.

In general, this model was able to determine the presence or absence of the keyword using the combination of speech recognition and computer vision algorithms with accurate accuracy.

6 | Feature Works

The proposed method, which was described in this paper, is a machine learning-based speech recognition system. For future works, we plan to design models similar to the described method, using more features and tools in machine learning and deep learning algorithms, hoping to achieve better performance and a lower computational cost.

we also plan to increase the number of classes inside the native dataset, which was proposed in this paper, so that it could be used to detect much more words inside an audio signal.

The final goal is to grow this research into a complete, well-designed speech recognition system for the Persian language.

For that goal, other machine learning and deep learning algorithms with the help of the signal processing methods will be used; hoping one day this goal will be achieved.

Acknowledgments

Finally, we would like to thank the members of our research team and our classmates at the Islamic Azad University Central Tehran Branch, as well as the research colleagues in the Informatics unit of the Ministry of Interior, especially Dr. Abbasi and Mr. Mohammadi, who help us in collecting videos and pictures of faces, in addition of being the guide and helper of researchers in conducting the entire research process. We would also like to thank Mr. Yousefi for carrying out this study and Dr. Davtalab, a capable consultant for this research project, who assisted in advancing and finalizing it. We also want to thank Mr. Milad Garmabi, one of the brightest psychology students at ATU, for his exceptional knowledge and guidance.

Conflict of Interest

The authors have no conflict of interest.

References

[1] Rudnicky, A. I., Hauptmann, A. G., & Lee, K. F. (1994). Survey of current speech technology. *Communications of the ACM*, *37*(3), 52-57.

[2] Guo, J., & Gould, S. (2015). *Deep CNN ensemble with data augmentation for object detection*. Retrieved from https://doi.org/10.48550/arXiv.1506.07224

[3] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., & Qu, R. (2019). A survey of deep learning-based object detection. *IEEE access*, *7*, 128837-128868.

[4] Vadwala, A. Y., Suthar, K. A., Karmakar, Y. A., Pandya, N., & Patel, B. (2017). Survey paper on different speech recognition algorithm: challenges and techniques. *Int J comput appl*, *175*(1), 31-36.

[5] Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., ... & Seltzer, M. L. (2020, May). Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6874-6878). IEEE.

[6] Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., ... & Zhang, Y. (2020, May). Quartznet: deep automatic speech recognition with 1d time-channel separable convolutions. *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6124-6128). IEEE.

[7] Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., & Norouzi, M. (2021). *Speechstew: simply mix all available speech recognition data to train one large neural network*. Retrieved from https://doi.org/10.48550/arXiv.2104.02133

[8] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). *Specaugment: A simple data augmentation method for automatic speech recognition*. Retrieved from https://doi.org/10.48550/arXiv.1904.08779

[9] Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C. C., Qin, J., ... & Wu, Y. (2020). *Contextnet: Improving convolutional neural networks for automatic speech recognition with global context*. Retrieved from https://doi.org/10.48550/arXiv.2005.03191

[10] Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster R-CNN for object detection in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3339-3348). IEEE.

[11] Park, S., Jeong, Y., & Kim, H. S. (2017). Multiresolution CNN for reverberant speech recognition. 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA) (pp. 1-4). IEEE. DOI: 10.1109/ICSDA.2017.8384470

[12] Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE journal of selected topics in signal processing, 11(8), 1240-1253. DOI: 10.1109/JSTSP.2017.2763455

[13] Passricha, V., & Aggarwal, R. K. (2020). A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. Journal of intelligent systems, 29(1), 1261-1274.

[14] Qian, Y., Bi, M., Tan, T., & Yu, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM transactions on audio, speech, and language processing*, *24*(12), 2263-2276.

[15] Han, S., Kang, J., Mao, H., Hu, Y., Li, X., Li, Y., ... & Dally, W. B. J. (2017, February). Ese: Efficient speech recognition engine with sparse LSTM on FPGA. *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays* (pp. 75-84). https://doi.org/10.1145/3020078.3021745

[16] Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014, November). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 801-804). https://doi.org/10.1145/2647868.2654984

[17] Gall, J., & Lempitsky, V. (2013). Class-specific Hough forests for object detection. In *Decision forests for computer vision and medical image analysis* (pp. 143-157). Springer, London.

[18] Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, *12*(2), 75-98. https://doi.org/10.1006/csla.1998.0043

[19] Moreno, P. J., Raj, B., & Stern, R. M. (1996, May). A vector Taylor series approach for environment-independent speech recognition. *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings* (Vol. 2, pp. 733-736). IEEE. DOI: 10.1109/ICASSP.1996.543225

[20] Woodland, P. C., & Povey, D. (2002). Large scale discriminative training of hidden Markov models for speech recognition. *Computer speech & language*, *16*(1), 25-47. https://doi.org/10.1006/csla.2001.0182

[21] Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., ... & Thomas, S. (2011). The subspace Gaussian mixture model—A structured model for speech recognition. *Computer speech & language*, *25*(2), 404-439. https://doi.org/10.1016/j.csl.2010.06.003

[22] Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., ... & Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the national academy of sciences*, *102*(7), 2293-2298.

[23] Zhu, Y., Zhao, C., Wang, J., Zhao, X., Wu, Y., & Lu, H. (2017). Couplenet: Coupling global structure with local parts for object detection. *Proceedings of the IEEE international conference on computer vision* (pp. 4126-4134). IEEE.

[24] Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: delving into high quality object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6154-6162). IEEE.

[25] Ghiasi, G., Lin, T. Y., & Le, Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7036-7045). IEEE.