Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 1, 2022, 117-126.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Persian Phoneme and Syllable Recognition using Recurrent Neural Networks for Phonological Awareness Assessment

Maryam Khanzadi[1], Hadi Veisi[1]*, Roghaye Alinaghizade[1], and Zahra Soleymani[2]

1. Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran.
2. Department of Speech Therapy, School of Rehabilitation, Tehran University of Medical Sciences, Tehran, Iran.

Article Info

Article History:
Received 05 October 2021
Revised 16 December 2021
Accepted 31 January 2022

DOI:10.22044/JADM.2022.11185.2268

Keywords:
Speech Therapy; Phonological Awareness Assessment; Kid's Speech Recognition; Persian Phoneme and Syllable Recognition; Long Short-term Memory Neural Network.

**Corresponding author: h.veisi@ut.ac.ir (H. Veisi).*

Abstract

One of the main problems in the children with learning difficulties is the weakness of phonological awareness (PA) skills. In this regard, the PA tests are used to evaluate this skill. Currently, this assessment is paper-based for the Persian language. In order to accelerate the process of the assessments and make it engaging for the children, we propose a computer-based solution that is a comprehensive Persian phonological awareness assessment system, implementing the expressive and pointing tasks. For the expressive tasks, the solution is powered by the recurrent neural network-based speech recognition systems. To this end, various recognition modules are implemented including a phoneme recognition system for the phoneme segmentation task, a syllable recognition system for the syllable segmentation task, and a sub-word recognition system for the three types of phoneme deletion tasks including the initial, middle, and final phoneme deletions. The recognition systems use the bidirectional long short-term memory neural networks to construct the acoustic models. In order to implement the recognition systems, we design and collect the Persian kid's speech corpus that is the largest in Persian for the children's speech. The accuracy rate for the phoneme recognition is 85.5%, and for the syllable recognition, it is 89.4%. The accuracy rates for the initial, middle, and final phoneme deletions are 96.76%, 98.21%, and 95.9%, respectively.

1. Introduction

Generally, the phonological awareness (PA) test consisting of the three levels of syllabic, intra-syllabic, and phoneme awareness involves expressive or pointing tasks; the expressive tasks refer to the tasks in which children express the answers to the questions by speech; the pointing tasks refer to the tasks in which children only point to the answers. The PA test of Soleymani [1] can be implemented in both the paper-based and computer-based methods. Currently, this assessment is paper-based. Jafari Sadr [2] has implemented the pointing tasks and one of the expressive tasks (i.e. phoneme segmentation) in a computer-based method. For the first time, we performed all tasks of the PA test by computer-based and machine learning methods in order to accelerate the process of the assessments and make it engaging for children. The implemented system is a user-friendly web-based software that provides a whole range of online-test advantages, with the appropriate instructions, attractive images, and animations for the children to use the system easily. Besides, implementing a computer-based assessment is advantageous for the speech-language pathologists; recording of assessment results helps them in diagnosing the children's learning problems and predicting their success. Despite the pointing tasks, the implementation of the expressive tasks is challenging since it requires the speech recognition module. One of the biggest challenges in designing a Persian speech recognition system is the scarcity of the kid's speech corpus. Exceptionally, AVAFA [2] is available with a duration of 48 minutes, which is limited and does not include syllables and sub-words. Also it only contains 23 Persian phonemes out of 29. Also another kid's speech corpus is a raw one without annotation, which is not appropriate for speech recognition [3]. Moreover, none of the available speech datasets [4]–[6] is exclusively for children, and therefore, we designed and collected a Persian Kid's Speech corpus, which covered all the Persian phonemes and included the units larger than phoneme (i.e. syllable, sub-word, and word).

We have used long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) neural networks in implementing the recognition system. An energy-based voice activity detection (VAD) module is implemented to segment the input speech into isolated units. In summary, the contributions of this paper are as follow:

- A computer-based phonological awareness assessment for Persian equipped with the phoneme and syllable recognition system.
- Implementing a speech recognition system based on phoneme, syllable, and sub-word using the recurrent neural networks (RNNs).
- Designing and collecting the largest kid's speech corpus for Persian

The rest of this paper is organized as what follows. In Section 2, we review the background and related works. In Section 3, we explain the PA tasks. In Section 4, an overview of the architecture of the proposed solution is given, and the implementation of the expressive tasks including the steps of designing a speech recognition system based on an LSTM neural network is presented. In Section 5, we describe the evaluation results, and finally, the conclusion will be presented in Section 6.

2. Background and Related Works

In this section, first, we address dyslexia as a learning disorder and the importance of PA as a prerequisite for learning to read. Then we will review the benefits of the computer-based assessment. In the following, in order to model the expressive tasks using the machine learning method, we discuss the speech recognition basics, recurrent neural networks, and research on speech recognition based on phoneme and syllable using deep learning.

2.1 Learning Disorders

Learning disorders can cause problems and limitations in the personal and social life. Dyslexia is a reading disorder that is observed in children who are unable to read properly due to the weakness of PA skills despite normal intelligence. A PA skill is a prerequisite of reading that focuses on manipulating words' structure including phoneme and syllable [7]. Some studies in several languages show that there is a potentially considerable correlation between reading performance and PA skills [8]–[11]. Concerning the vital role of PA, the validated tests are required in order to evaluate this skill to be used in the research and clinical fields. Currently, in the Persian language, the assessment of Soleymani [1] and Asha-5 phonological awareness skills [12] are used to evaluate PA. We focus on the Soleymani test due to its high degree of validity and reliability. In this test, the words of each task are chosen according to the features of the words as follow: single-syllable words in the intra-syllabic task, one to four syllables words in the syllabic task, and one and two-syllable words in the phonological tasks. They presented the words through images in order to remove the child's auditory memory.

2.2 Computer-based Phonological Awareness

Although the paper-based phonological awareness tests are very popular, there are several attempts to implement the computer-based versions. Carson in his research works showed that using the computer-based method speeded up the assessment execution, and children followed the assessment eagerly in comparison with the traditional paper-based methods [13], [14]. Some papers have addressed the efficiency of the computer-based methods to help children with the learning disorder; however, these are focused on the PA training, and speech recognition is ignored in the expressive tasks [15]–[17].

AVAFA [2], especially relevant to the aims of our work, is a Persian computer-based assessment that includes the pointing tasks, and is limited to a phoneme segmentation task. It indicates the validity and reliability of the computer-based assessments.

In comparison with the previous works, our system covers all the expressive and pointing tasks. Moreover, in this work, we used the neural network-based speech recognition, as the state-of-the-art technique, in order to recognize the kid's speech in expressive tasks.

2.3 Speech Recognition

Artificial Neural Network (ANN) has been used in automatic speech recognition (ASR) for a long time. LSTM is an RNN architecture for acoustic modeling, in which the memory blocks are embedded in the structure of hidden layer neurons, input, output, and forget gates in order to overcome the problem of vanishing gradients related to network training by controlling the input and output in the hidden layer [18]. Figure 1(a) shows the internal structure of the memory block, and the overall architecture of the LSTM network with two memory blocks in the hidden layer is shown in Figure 1(b).

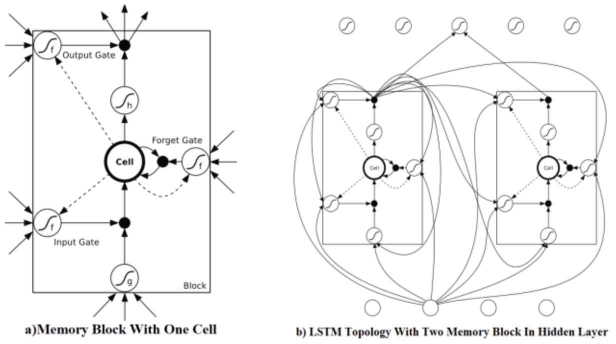

Output Gate
Forget Gate
Cell
Input Gate
Block
a)Memory Block With One Cell
b) LSTM Topology With Two Memory Block In Hidden Layer

Figure 1. General structure of LSTM (a), network and memory block with two memory block (b) [18].

Graves has introduced BiLSTM [19], which processes the input sequence in both directions with two separate hidden layers (the forward and backward layers). Due to the process of the full input context, the performance of the network improves in comparison to the unidirectional LSTM. Figure 2 shows the architecture of the BiLSTM model.

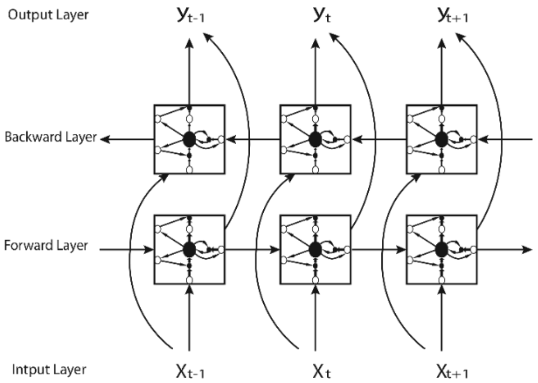


Figure 2. Architecture of BiLSTM.

Deep ANNs (DNN) outperforms the other machine learning methods in many areas as well as speech recognition [20]. A deep LSTM neural network has been used for speech recognition on the TIMIT data [21]. Different techniques of RNNs have been implemented in the Persian speech recognition systems. Nevisa [22] is one of the most important speech recognition systems, and Ansari and Seyyedsalehi [23] have used an ANN model for Persian continuous speech recognition. In other studies on the Farsdat, the LSTM network and deep BiLSTM have been used to recognize Persian speech [24] and also in [25], phoneme sequence recognition was improved on aforesaid data by using hidden semi Markov model and DNN.

Although the phonemes are the first and most widely used unit of speech recognition, the number of them in the languages is limited; also phonemes are context-dependent, and then the boundary between phonemes and coarticulation on the boundaries is not well-established [26]. Therefore, this feature may result in the accuracy reduction of the speech recognition system. Using a larger unit such as tri-phone and syllable- show that the syllable-based speech recognition system performs better than the phoneme-based recognition system [26]–[28].

To the best of our knowledge, no syllable-based speech recognition system for the Persian language has been performed on the kids' speech corpus so far.

3. Persian Phonological Awareness Assessment

Table 1 shows the pointing and expressive tasks related to every level of paper-based PA assessment [1]. Each task consists of 10 questions. The following sections explain the tasks and the process of a paper-based approach, which should be considered in a computer-based approach.

Table 1. PA Tasks.

Phonological Assessments	**Tasks**	**Type**
Syllabic awareness	Syllable segmentation	Expressive
Phonological assessment	Phoneme segmentation	Expressive
Phonological assessment	Naming and deleting final phoneme	Expressive
Phonological assessment	Naming and deleting middle phoneme	Expressive
Phonological assessment	Naming and deleting final phoneme	Expressive
Phonological assessment	Identifying different words with identical initial phoneme	Pointing
Phonological assessment	Identifying different words with identical final phoneme	Pointing
Phonological assessment	Phoneme blending	Pointing
Intra-syllabic awareness	Alliteration detection	Pointing
Intra-syllabic awareness	Rhyme detection	Pointing

3.1 Intra-syllabic Awareness

This level consists of two pointing tasks of alliteration and rhyme detection. Each one of these tasks utilizes three images per question. The examiner then asks the child to point to images of words that have the same initial (alliteration) or final (rhyme) sounds. For example, “میز” [miːz] (desk), and “میخ” [miːx] (nail) have the same initial sound “می” [miː] in Persian, among the words of “میز”' [miːz] (desk), “فیل” [fiːl] (elephant), and “میخ” [miːx] (nail). “بیل"” [biːl] (shovel) rhymes with “فیل”[fiːl] (elephant) among the three words of “فیل” [fiːl] (elephant), “بیل” [biːl] (shovel), and “میز” [miːz] (desk).

3.2. Syllabic Awareness

This level includes the expressive task of syllable segmentation. The examiner asks the child to express the syllables of the intended word. For example, the child should segment the syllables of “بادکنک” [bɒːdkonæk] (ballon) as “باد”, “کُ”, “نَک” [bɒːd, ko, næk] (/ba/, /llon/).

3.3. Phonological Awareness

This section consists of pointing or expressive tasks. The pointing tasks identify different words with identical initial phoneme, identifying different words with identical final phoneme and phoneme blending. The expressive tasks include phoneme segmentation and initial, middle, and final phoneme deletion of the word. In order to perform the task of identifying different words with identical initial or final phoneme, the examiner asks the child to point to two images of words, which have the same initial or final phoneme. For example, the two words “ماهی” [mɒ:hi] (fish) and “مداد” [medɒ:d] (pencil) have the same initial phoneme among the three words “جوراب” [dʒu:rɒ:b] (socks), “ماهی” [mɒ:hi] (fish), and “مداد” [medɒ:d] (pencil); also the two words “آهو” [ɒ:hu:] (deer) and “جارو” [dʒɒ:ru:] (broom) have an identical final phoneme among the three words “آهو” [ɒ:hu:] (deer), “جارو” [dʒɒ:ru:] (broom), and “هویج” [hævidʒ] (carrot).

In the task of phoneme blending, multiple images are presented to the child for each question. The examiner segments the word into the constituent phonemes within a two-second interval. For example, the examiner pronounces the word “کیف” [ki:f] (bag) as “ک“, ”ی“, ”ف” [k, i:, f], and then asks the child to point to the image that matches to the segmented syllables.

For the task of phoneme segmentation, the child is asked to segment the phonemes of the name of the presented image with a two-second interval. For example, the word “سیب” [si:b] (apple) is segmented as “س”, “ی”, “ب” ([s, i:, b]) in this task. In the phoneme deletion tasks, the examiner states the name of the image and the phoneme that must be deleted in the word, and then asks the child to delete the intended phoneme in the word and express the word without the intended phoneme. For example, in the task of the initial phoneme deletion of the word “درخت” [deræxt] (tree), the examiner askes the child to delete the first phoneme, “د” /d/, and pronounce it as “اِرخت” [eræxt]. In the task of the middle phoneme deletion, the child deletes the phoneme “ر” /r/ from the word “برگ” [bærg] (leaf), and then pronounces “بَگ” [bæg]. In the task of final phoneme deletion of such a word “خانه” [xɒ:ne] (home), the child should delete the final phoneme and pronounce “خان” [xɒ:n].

4. Proposed Phonological Awareness Assessment System

In this section, we discuss the implementation steps of the proposed system according to the paper-based assessment. The proposed system consists of the two general sections of admin and test. The admin section includes the examinee and question management; the test section performs the PA tasks. Figure 3 shows the overall architecture of this system. The admin section utilizes the general well-known software implementation techniques, and is not presented in detail here; however, the test section is the main contribution of this work, and is given in detail below.

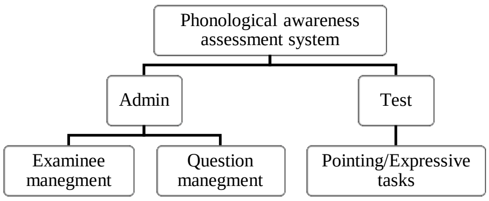

Phonological awareness assessment system
Admin
Test
Examinee manegment
Question manegment
Pointing/Expressive tasks

Figure 3. Architecture of the proposed system

The tasks in the test section are divided into the three phonological, syllabic, and intra-syllabic levels. The system presented requires guides to the student in performing the tasks. At the end of each task, the scores are stored in the results table, which is monitored by the examiner. All the pointing tasks, except the phoneme blending task, are implemented in such a way that the image is displayed at the top of the two other images. According to the type of the tasks, the child must select one of the images as a correct answer corresponding to the top image. If the child responds to each question correctly, the score is one; otherwise, the score is zero. In the phoneme blending task, audio is played phoneme by phoneme; then the child should blend them and select the image that matches the blended phonemes.

Expressive tasks, as discussed in Section 3, include the phoneme segmentation, initial phoneme deletion, middle phoneme deletion, final phoneme deletion, and syllable segmentation tasks; the child should segment the word into its constituent syllables and phonemes, and also delete a phoneme from the word, and then pronounce it without the intended phoneme. An image is displayed according to the task, and then the child sends the voice by clicking on the answer button. The speech signal is recorded and processed by the speech processing module, and finally, the result of the recognition system is stored on the database. The sample images of the proposed system can be found in Figures 4 to Figure 6. In the following section, we outline the design steps in detail.

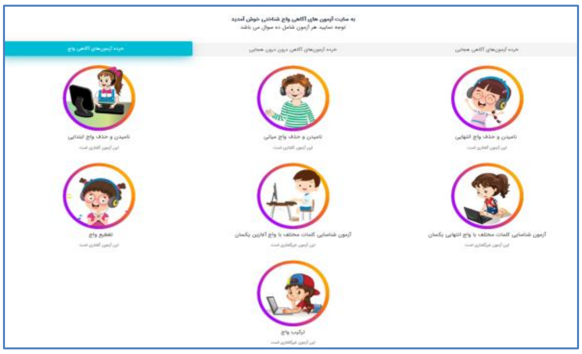
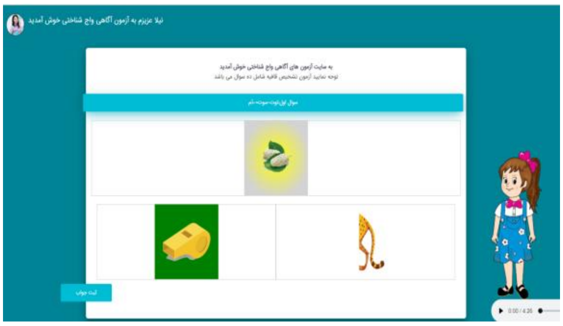

به سایت آزمون های آگاهی واج شناختی خوش آمدید
توجه نمایید هر آزمون شامل ده سوال می باشد

Figure 4. Task selection page.



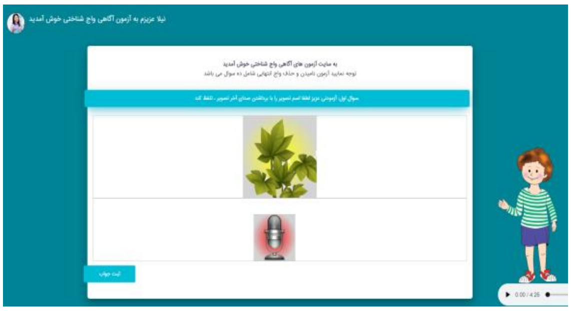
نیلا عزیزم به آزمون آگاهی واج شناختی خوش آمدید

به سایت آزمون های آگاهی واج شناختی خوش آمدید

توجه نمایید آزمون تشخیص قافیه شامل ده سوال می باشد

سوال اول:توت-سوت-نام

Figure 5. Rhyme detection task (child chooses image of "سوت" [sut](whistle) since it rhymes with "توت" [tut] (Berry).



نیلا عزیزم به آزمون آگاهی واج شناختی خوش آمدید

به سایت آزمون های آگاهی واج شناختی خوش آمدید

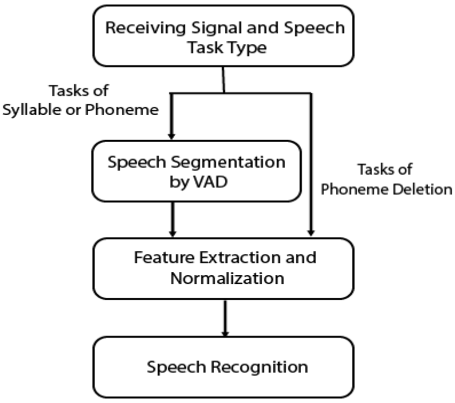
توجه نمایید آزمون نامیدن و حذف واج انتهایی شامل ده سوال می باشد

سوال اول: آزمودنی عزیز لطفا اسم تصویر را با برداشتن صدای آخر تصویر ، تلفظ کند

Figure 6. Naming and final deletion task (voice of "برگ" [barg] (leaf) is played and child deletes "گ" /g/ and pronounces "بگ" /bag/).

4.1. Implementation of Expressive Tasks using Speech Recognition

In order to implement the expressive tasks, we proposed a speech recognition system. In order to create the acoustic model for the speech recognition system in the training step, we require a suitable speech corpus. In addition to designing and collecting a kid's speech corpus, the data pre-processing and feature extraction are the prerequisites of building the acoustic model, although we can use the phoneme-based speech recognition to implement a syllable or sub-word speech recognition system; however, it is challenging; it means that the phonemes are context-dependent; besides, the frames of phonemes are short, and the phoneme boundaries are not accurately identifiable. Therefore, the accuracy of the system will be decreased [26]. Thus we consider syllables and sub-words as the acoustic units, and model them individually. As shown in Figure 7, in the application phase, after receiving a speech signal, the signal is segmented using a VAD module (phoneme and syllable tasks). Then feature extraction is performed by Mel-Frequency Cepstral Coefficient (MFCC), and finally, the child's speech is recognized by the acoustic model according to the type of the tasks.



Receiving Signal and Speech Task Type
Tasks of Syllable or Phoneme
Speech Segmentation by VAD
Tasks of Phoneme Deletion
Feature Extraction and Normalization
Speech Recognition

Figure 7. Speech processing steps in proposed system.

4.1.1 Creating kid's Speech Corpus

The kid's speech dataset is required to train the kid's speech recognition systems at the level of phoneme, syllable, and sub-word. Since the available corpora do not cover all of the acoustic units of the expressive tasks in Table 1 and concerning all the limitations have been referred to Persian Kids as a speech corpus is designed, collected, and prepared.

In order to design a comprehensive corpus, we added some additional words to the vocabulary of paper-based assessment; therefore, the intended corpus covers all the Persian phonemes and some syllables and sub-words.

The data is recorded in the school environment, so some audio files contain the real environment noises. Kids' speech signals are recorded by an Andreas Mic Anti-Noise microphone and a Premium Speechmike headphone. We utilized the CoolEdit Pro2.1 software to record the speech at 16 kHz, single-channel, 16bit resolution, and saved it in the WAV files. The speech signals from 286 children (141 girls, 145 boys), ages 6 to 9-year-old, were recorded. This recorded data is manually checked and labeled. Finally, a corpus containing 162,395 samples with a duration of 33 hours and 44 minutes is created. The prepared speech corpus is particularly applicable to speech recognition and linguistics studies. It comprehensively contains all the 29 Persian phonemes, 118 syllables, 56 sub-words, and 711 words. A summary of the corpus is given in Table 2 [29].

Table 2. Specification of Persian Kids' Speech Corpus[29]

Title	**Persian Kids Speech Version 1**		
Authors	Maryam Khanzadi, Hadi Veisi		
Release date	October 22, 2018		
DCMI type	Speech		
Data sources	Andreas Mic Anti-Noise, Speechmike Premium		
Sample type	1-Channel		
Sample rate	16 KHz		
Resolution	16-Bit		
File format	Wav		
Total amount of data time	2024 minute (33 hours and 44 minutes)		
Application(s)	Speech recognition		
Language(s)	Persian		
Total number of samples	162,395		
	Words	29,057 samples	(478 min)
	Sub-words	17,429 samples	(260 min)
	Syllables	43,838 samples	(485 min)
	Phonemes	70,078 samples	(765 min)
	Extra vocabulary	1,993 samples	(36 min)

We selected a part of the collected corpus as the train, validation, and test sets. Table 3 shows the number and duration (in the parentheses) of the samples in these sets according to the acoustic model.

Table 3. Number of samples and duration of data used in train, validation, and test sets according to acoustic units.

Phonetic model	No. of class	Total set	Train set	Validation set	Test set
Phoneme	30	73,819 (788:39)	51,660 (551:35)	11,076 (118:27)	11,083 (118:37)
Syllable	26	21,320 (234:13)	14,916 (164:19)	3,196 (34:53)	3,208 (35:10)
Deleting initial phoneme	10	3,688 (58:30)	2,578 (40:05)	554 (08:42)	556 (08:47)
Deleting middle phoneme	10	4,855 (82:00)	3,394 (57:53)	731 (12:46)	730 (12:21)
Deleting final phoneme	10	7,648 (100:34)	5,342 (70:21)	1,152 (14:51)	1,154 (15:25)

4.1.2 Pre-processing of Signal

In the evaluation/application phase, firstly, the kid's speech signal and the type of the task are received, and then speech pre-processing is required depending on which task the received speech belongs to. As Figure 7 shows, all the received signals, except signals of deletion, need to be segmented into phonemes or syllables according to the type of task. To do this, we implemented an energy-based VAD module in order to find the silence segment in the input signal. Several methods such as energy calculation, Zero Crossing Rate (ZCR), cepstral features, and HMM (Hidden Markov Model) can be used for detecting speech from non-speech [30]. In the current research work, we used the power of a fixed-length analyzing window of the signal. By moving this window over the speech, if the incoming frame's energy exceeds a pre-defined threshold value, the frame is judged as a speech frame; otherwise, the frame is considered as a non-speech frame. The border of the speech segments is then defined in the middle of consecutive non-speech frames. The most important parameter of this algorithm is the threshold that is adjusted manually concerning the recording circumstances.

4.1.3 Feature Extraction

Like many other speech recognition systems, we used the MFCC algorithm in order to extract the features from each frame of the signal and implement it by Speechpy Python Library. After windowing and framing, 13 features were extracted from each frame. Due to the dynamic nature of the speech signal, and to consider the effect of the neighborhoods for each frame, we added the first and second derivatives of MFCC, resulting in a 39-dimensional vector for each frame. Table 4 summarizes the parameters used to extract the MFCC features. These feature vectors are then normalized using Cepstral Mean Normalization (CMN) to reduce the channel noise.

Table 4. Parameters of feature extraction.

Parameter	**Value**
Sampling rate	16 KHz
Window function	Hamming
Frame length	20 ms
Frame overlapping	10 ms
No. of filter bank	24
No. of MFCC feature	39

4.1.4 Training Acoustic Models using LSTM

In this paper, LSTM and BiLSTM were used to construct the acoustic models, and then these models are obtained to map the input speech to their corresponding labels (i.e. classification). In order to implement the speech recognition system, we used the Keras TensorFlow's high-level deep learning framework. In addition to the single hidden layer LSTM and BiLSTM networks, we also evaluated the deeper versions of them by increasing the number of hidden layers. According to our evaluations, the performance of BiLSTM is consistently higher than or equal to the LSTM one in all tasks. In the implementations, various hyper-parameters of the networks including the number of epochs, learning rate, number of the hidden layers, number of memory blocks in each hidden layer, and the dropout parameter are considered to be optimized. The number of input layer neurons equals the number of the extracted feature vector by MFCC, i.e. 39. Due to the limitations in the number of training samples of the corpus, our evaluations on the deep ANNs do not show the improvement for the depth higher than two. Therefore, as the results in Section 5 show, the evaluation results are given for the networks with one and two hidden layers. For each one of them, the number of memory blocks in the hidden layer(s) are evaluated for several numbers. The numbers of the output layer neurons for phonemes, syllables, and sub-words are 30, 26, and 10, respectively, due to the required assessment tasks in Table 1. After specifying the architecture of the neural network, other parameters including the loss function, the optimization algorithm, and the network evaluation metrics are selected. The loss function measures the performance of the network on the train set, and it is designed to minimize the distance between the network's output and the intended output. Due to the classification model, cross-entropy is used in our experiments. We used the ADAM optimizer in the neural network training. If the mean squares error on the validation set is increased or the maximum number of epochs is reached, the network training ends.

In the training phase, for each one of the acoustic units (phoneme, syllable or sub-word), separate networks are created, and in the evaluation/application phase, the appropriate model is selected according to the assessment task type in order to recognize the input speech signal.

5. Evaluation Results

In this section, the results of our experiments are presented. The evaluations are done on the dataset, as given in Table 3. Although the VAD module is not the focus of this research work, the performance of this module is evaluated, and the results are given in the following sub-section. After that, the evaluation results of the recognition modules are presented.

5.1 VAD Evaluation

In order to evaluate the VAD algorithm, a reference dataset is required, and the speech frames are labeled manually as speech and non-speech. Table 5 presents the characteristics of the dataset and also the parameters used.

Table 5. Parameters of VAD and evaluation dataset.

Parameter	Value
Threshold	10^{-6}
Frame overlapping	10 ms
Frame length	20 ms
Sampling rate	16 KHz
No. of non-speech frames	1,437
No. of speech frames	1,038
No. of total frames	2,475

In the evaluations, we use the recall (sensitivity), precision, F-measure, and specificity criteria [30]. In order to compute the criteria, the output of the VAD algorithm is compared with the reference labels, and the results of this comparison in terms of the confusion matrix are presented in Table 6.

Table 6. Confusion matrix of VAD module.

		Output of VAD	
		Negative	Positive
Actual value	True (speech)	FN = 182	TP = 856
	False (non-speech)	TN = 1,292	FP = 145

The recall (or sensitivity) is defined as the percentage of correctly identified speech frames, i.e. TP/(TP+FN), which equals 86%. Precision is the percentage of the identified frames as speech, which are correctly real speech, i.e. TP/(TP+FP), which is 82%. A high precision rate means that the speech frames are more accurately labeled. The high rate is significant since we need to make sure that the segmentation is accurate. Specificity is the recall for the negative, i.e. non-speech, samples that calculate the percentage of correctly identified non-speech frames, i.e. TN/(FP+TN). The specificity of the VAD module is 90%. Also F-measure, which is the harmonic mean of precision and recall, equals 84%. Table 7 summarizes the final results of VAD.

Table 7. VAD evaluation result.

Precision	Recall	Specificity	F-measure
0.8246	0.8551	0.90	0.84

5.2 Speech Recognition Evaluation

As Table 1 shows, five expressive tasks including the syllable segmentation, phoneme segmentation, naming and deleting final phoneme, deleting middle phoneme, and naming and deleting the final phoneme are implemented in the proposed system, and and for each of them, a speech recognition module is trained. Therefore, the results in this section are given for these five different ANNs separately. All the results of this section are performed on the test sets for each task. As there are many experiments, in order to reduce the paper length, we only present the BiLSTM results. In our evaluations, the performance rates of the BiLSTM networks are consistently higher than or at least equal to the equivalent LSTM networks. Our evaluations also indicate that the performance rates of ANNs with the number of hidden layers more than two are lower than the network with the depth of two, probably due to the size of the training sets. Therefore, only the results for a single hidden layer and two hidden layers are given.

In all experiments in this section, accuracy is used as the evaluation criteria, the learning rate and the dropout parameters are set to 0.0001 and 0.5, respectively. In the following, we first present the recognition results of BiLSTMs for one and two hidden layers according to various hyper-parameters in each task. After that, the results given by the optimized networks are present.

5.2.1 One Hidden Layer BiLSTM Network

For the one hidden layer BiLSTM, the effect of the number of memory blocks in the hidden layer and the number of epochs were investigated. The values obtained based on our experiments and the results are given in Table 8 to Table 12. The results showed that the increase in the number of neural network memory blocks improved the speech recognition accuracy in the five models. The best result for each task is highlighted in boldface in these tables.

Table 8. Results of one hidden layer BiLSTM for *phoneme segmentation* task.

Epochs	No. of memory blocks	Train set accuracy	Validation set accuracy	Test set accuracy
100	64	0.8766	0.7550	0.7550
100	128	0.8506	0.8091	0.8093
50	256	0.886	0.8216	**0.8251**
80	256	0.9125	0.8134	0.8167

Table 9. Results of one hidden layer BiLSTM for *syllable segmentation* task.

Epochs	No. of memory blocks	Train set accuracy	Validation set accuracy	Test set accuracy
100	64	0.7709	0.7946	0.7768
150	64	0.8153	0.7910	0.7983
200	64	0.8538	0.8141	0.8195
250	64	0.9338	0.8128	0.8168
50	128	0.8246	0.7982	0.8033
70	128	0.8687	0.8188	0.8235
100	128	0.9021	0.8348	0.8460
50	256	0.9291	0.8592	0.8636
100	256	0.9617	0.8642	**0.8660**
120	256	0.9899	0.8509	0.8630

Table 10. Results of one hidden layer BiLSTM for *initial phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks	Train set accuracy	Validation set accuracy	Test set accuracy
200	64	0.9414	0.8989	0.8866
50	64	0.9531	0.9007	0.8956
100	64	0.9829	0.8880	0.8796
200	128	0.9942	0.9404	0.9298
200	256	0.9988	0.9513	**0.9604**
220	256	0.9994	0.9487	0.9591

Table 11. Results of one hidden layer BiLSTM for *middle phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks	Train set accuracy	Validation set accuracy	Test set accuracy
100	64	0.9337	0.9097	0.9150
150	64	0.9758	0.9166	0.9438
200	64	0.9845	0.9121	0.9411
250	64	0.9974	0.9120	0.9378
150	128	0.9973	0.9494	0.9602
50	256	0.9876	0.9535	0.9690
100	256	0.9968	0.9644	**0.9712**
120	256	0.9991	0.9518	0.9630

Table 12. Results of one hidden layer BiLSTM for *final phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks	Train set accuracy	Validation set accuracy	Test set accuracy
200	64	0.9744	0.9361	0.9350
250	64	0.9794	0.9387	0.9471
300	64	0.9929	0.9238	0.9188
200	128	0.9966	0.9492	**0.9497**
250	128	0.9981	0.9466	0.9436

5.2.2 Two Hidden Layer BiLSTM Network

In the following experiments, we study the effect of increasing the number of hidden layers. In these evaluations, the number of memory blocks for each hidden layer and the number of epochs are investigated. As shown in Table 13, the results of the two hidden layer BiLSTMs with different parameters are evaluated for the phoneme segmentation task. As shown, the best performance is achieved in the 128-256 memory blocks. By comparing the results of Table 13 and Table 8, it can be concluded that the performance of the two hidden layer BiLSTM is higher than the one hidden layer network.

Table 13. Results of two hidden layer BiLSTM for *phoneme segmentation* task.

Epochs	No. of memory blocks for first hidden layer	No. of memory blocks for second hidden layer	Train set accuracy	Validation set accuracy	Test set accuracy
150	64	128	0.8938	0.7947	0.7952
200	64	128	0.9129	0.7903	0.7944
50	64	128	0.9096	0.8156	0.8224
50	128	256	0.8989	0.8365	0.8448
80	128	256	0.9361	0.8441	0.8525
50	256	256	0.9332	0.8455	0.8465
50	128	256	0.9747	0.8601	**0.8654**

The results of Tables 14 to Table 17 show the results of two hidden layer BLSTM for the other tasks. As these results show, increasing the number of memory blocks in the hidden layer increases the accuracy of the network. Speech recognition based on syllable and sub-word is more accurate since the pronunciation changes in speech can be modeled better in comparison to the phoneme-based model. Moreover, a syllable unit spans a longer time frame, and the classes of recognition are limited.

Table 14. Results of two hidden layer BiLSTM for syllable segmentation task.

Epochs	No. of memory blocks for first hidden layer	No. of memory blocks for second hidden layer	Train set accuracy	Validation set accuracy	Test set accuracy
50	64	128	0.8534	0.8360	0.8450
100	64	128	0.9145	0.8626	0.8647
120	64	128	0.9456	0.8534	0.8590
100	128	128	0.9615	0.8789	0.8887
100	256	128	0.9838	0.8877	**0.8949**
120	256	128	0.9929	0.8845	0.8924

Table 15. Results of two hidden layer BiLSTM for *initial phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks for first hidden layer	No. of memory blocks for second hidden layer	Train set accuracy	Validation set accuracy	Test set accuracy
100	64	128	0.9791	0.9495	0.9514
120	64	128	0.9876	0.9549	0.9568
150	64	128	0.9965	0.9711	0.9568
170	64	128	0.9961	0.9603	0.9532
50	128	128	0.9500	0.9386	0.9082
150	128	128	0.9926	0.9729	**0.9676**
170	128	128	0.9926	0.9513	0.9514

Table 16. Results of two hidden layer BiLSTM for e *middle phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks for first hidden layer	No. of memory blocks for second hidden layer	Train set accuracy	Validation set accuracy	Test set accuracy
65	64	128	0.9767	0.9535	0.9585
100	64	128	0.9900	0.9617	0.9684
120	64	128	0.9938	0.9631	0.9698
150	64	128	0.9976	0.9590	0.9712
170	64	128	0.9985	0.9470	0.9620
100	128	128	0.9971	0.9590	0.9726
170	128	128	1.0	0.9595	**0.9794**

Table 17. Results of two hidden layer BiLSTM for *final phoneme deletion (sub-word)* task.

Epochs	No. of memory blocks for first hidden layer	No. of memory blocks for second hidden layer	Train set accuracy	Validation set accuracy	Test set accuracy
50	64	128	0.9581	0.9483	0.9490
120	64	128	0.9899	0.9536	0.9540
150	64	128	0.9925	0.9483	0.9566
170	64	128	0.9999	0.9470	0.9520
150	128	128	0.9940	0.9518	**0.9592**

5.2.3 Comparative Results

In Figure 8, the best results for the accuracy of the test data in terms of different tasks for the BiLSTM network with one hidden layer and two hidden layers are illustrated. The number of the memory blocks in the one hidden layer BiLSTM is 256 for all tasks, except for the final phoneme deletion task, which is 128. For the two hidden layer BiLSTMs, the best configurations are 128-256, 256-128, and 128-128 for the phoneme segmentation, the syllable segmentation, and for the other three tasks, respectively. As Figure 8 shows, changing the number of hidden layers of the BiLSTM network from one layer to two hidden layers gives 4% improvement for the phoneme recognition models and 2.9% for the syllable recognition. Also in the sub-word model including the deletion of initial, middle, and final phonemes, the accuracy is improved by 0.8%, 1.1%, and 0.9%, respectively.

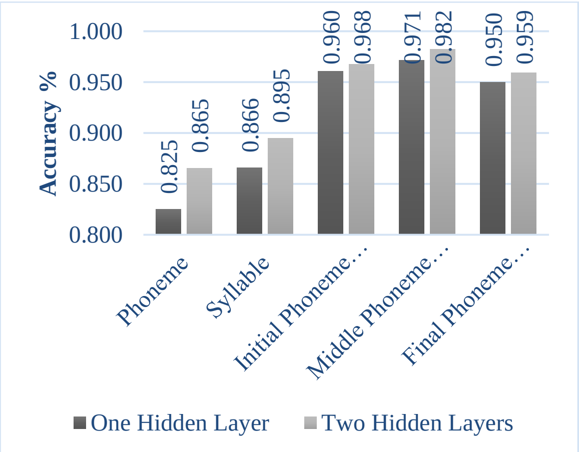


Figure 8. Comparative results of BiLSTM networks with one and two hidden layers for different tasks.

There is not a related work that could be compared properly. Although there is a similar research work [2], it is not comparable with our results because in that work, only one of the five expressive tasks (i.e. phoneme segmentation) was implemented using the HMM algorithm with a limited dataset.

6. Conclusions and Future Works

In this paper, we presented a computer-based phonological awareness assessment solution for Persian, covering all the assessment tasks. Implementation of the computer-based assessment using a machine learning method leads to a time-efficient implementation and improvement of the PA skills. In order to implement the five expressive tasks, we proposed an RNN-based speech recognition system. To do the recognition of the kids' speech, we designed and collected the Persian Kids Speech Corpus. In the ASR system, MFCC was used in order to extract the features, and CMN was applied to remove the convolutional noises, and then BiLSTM networks were utilized to perform the classification. The evaluations showed the superiority of BiLSTM with two hidden layers over the one hidden layer.

There are several directions to improve our work in the future; the Kids' Speech Corpus can be enlarged more, Also improving the corpus varieties benefits deeper ANNs in the modeling. Another interesting approach to do the automatic PA assessment in the expressive tasks is to use the speech verification instead of the speech recognition. The proposed VAD module is required to be improved using more accurate and more robust VAD algorithms.

Acknowledgment

We thank all the elementary schools and kindergartens in Districts 5 and 6 of Tehran, who helped us in gathering our speech corpus.

References

[1] Z. Soleymani, "Phonological awareness and effect of reading in 5.5 and 6.5 years old Persian children" *Arch. Rehabil.*, vol. 1, no. 2, pp. 27–35, 2000.

[2] E. Jafari Sadr, "Implementing Computer-Based Phonological Awareness Assessment in Persian,", M.S. thesis, Dept. Sci. Technol, Tehran Univ, Tehran , 2017.

[3] N. Family, J. Chandlee, M. Franchini, S. Lord, and G. Rheiner, "Lighten up: the acquisition of light verb constructions in Persian" in *Proceedings of the 33rd annual Boston University Conference on Language Development*, vol. 1, pp. 139–150, 2009.

[4] M. Eslami, J. Sheikhzadegan, Z. Ahmadinia, and R. Bahrami, "Developing Syllable And Diphone Speech Databases For Persian Text-To-Speech Synthesis System" *Signal and Data Processing*, vol. -, no. 2, pp. 3-12, 2009.

[5] M. Bijankhan, J. Sheikhzadegan, and M. R. Roohani, "Farsdat-The speech database of Farsi spoken language" *Proccedings Australian Conference On Speech Science And Technology*, vol. 2, pp. 826-830, 1994.

[6] M. Bijankhan, J. Sheykhzadegan, M. R. Roohani, R. Zarrintare, S. Z. Ghasemi, and M. E. Ghasedi, “Tfarsdat-the telephone Farsi speech database,” in *speech communication and technology., Geneva of Conf.*, Europ, 2003, pp. 1525-1528.

[7] M. Dastjerdi and Z. Soleymani, "What is Phonological Awareness?" *J. Except. Child.*, vol. 6, no. 4, pp. 931–954, 2007.

[8] M. Pérez-Pereira, Z. Martínez-López, and L. Maneiro, "Longitudinal relationships between reading abilities, phonological awareness, language abilities and executive functions: Comparison of low risk preterm and full-term children" *Front. Psychol.*, vol. 11, p. 468, 2020.

[9] H. Ahadi, R. Nadarkhani, and M. Ghayoomi, "A Study of Word Reading in Persian-speaking Children With Dyslexia and Normal Ones" *J. Mod. Rehabil.*, vol. 14, no. 4, pp. 207–216, 2020.

[10] C. Ergül, G. Akoğlu, M. Ç. Ö. Akçamuş, E. Demir, B. K. Tülü, and Z. B. Kudret, "Longitudinal Results on Phonological Awareness and Reading Performance of Turkish-Speaking Children by Socioeconomic Status" *Egit. ve Bilim*, vol. 46, no. 205, 2021.

[11] C. Míguez-Álvarez, M. Cuevas-Alonso, and Á. Saavedra, "Relationships Between Phonological Awareness and Reading in Spanish: A Meta-Analysis" *Language Learnimg*, pp. 1-46, October 2021.

[12] Z. Arani Kashani and A. Ghorbani, ”Auditory test of phonological awareness skills (ASHA-5) for 5-6 years old Persian speaking children, ” in *Setayeshe Hasti*, 1 ed. Tehran, Iran, 2010, ch. 1000.

[13] K. L. Carson, “Efficient and effective classroom phonological awareness practices to improve reading achievement”, Ph.D. dissertation, Dept. Philosophy., Canterbury Univ., New Zealand, April 2012.

[14] K. Carson, T. Boustead, and G. Gillon, "Predicting reading outcomes in the classroom using a computer-based phonological awareness screening and monitoring assessment (Com-PASMA)" *Int. J. Speech. Lang. Pathol.*, vol. 16, no. 6, pp. 552–561, 2014.

[15] P. Patel, M. Torppa, M. Aro, U. Richardson, and H. Lyytinen, "Assessing the effectiveness of a game-based phonics intervention for first and second grade English language learners in India: A randomized controlled trial" *J. Comput. Assist. Learn*, vol. 38, no. 1, pp. 76-89, February 2021.

[16] F. Fadaei, H. Kalantari Dehaghi, and M. Abdollahzadeh Rafi, "The effect of computer-based method of» sequential display of letters «on quick naming, phonological awareness, accurate and fluid reading of dyslexic elementary students" *Technol. Educ. J.*, vol. 16, no. 1, pp. 59-70.2021.

[17] T. Winn, J. Miller, and W. van Steenbrugge, "The efficacy of a computer program for increasing phonemic awareness and decoding skills in a primary school setting for children with reading difficulties" *Aust. J. Teach. Educ.*, vol. 45, no. 12, pp. 1–23, 2020.

[18] F. Gers, “Long short-term memory in recurrent neural networks”, PhD dissertation., Verlag nicht ermittelbar, 2001.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures" *Neural networks*, vol. 18, no. 5–6, pp. 602–610, 2005.

[20] D. Yu and L. Deng, Automatic Speech Recognition, 1 ed. Springer, London, 2016.

[21] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 4580–4584.

[22] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, and K. Hosseinzadeh, “Nevisa, a persian continuous speech recognition system,” in *Computer Conf.*, Iran, 2008, pp. 485–492.

[23] Z. Ansari and A. Seyyedsalehi, "Deep Modular Neural Networks with Double Spatio-temporal Association Structure for Persian Continuous Speech Recognition" *Signal Data Process.*, vol. 13, no. 1, pp. 39-56., 2016.

[24] M. Daneshvar and H. Veisi, “Persian phoneme recognition using long short-term memory neural network,” in *2016 Eighth International Conference on Information and Knowledge Technology (IKT)*, Iran, 2016, pp. 111–115.

[25] M. Asadolahzade Kermanshahi and M. M. Homayounpour, "Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM" *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 137–147, 2019.

[26] S. Bhatt, A. Jain, and A. Dev, "Syllable based Hindi speech recognition" *J. Inf. Optim. Sci.*, vol. 41, no. 6, pp. 1333–1351, 2020.

[27] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition" *IEEE Trans. speech audio Process.*, vol. 9, no. 4, pp. 358–366, 2001.

[28] M. M. Azmi, H. Tolba, S. Mahdy, and M. Fashal, “Syllable-based automatic Arabic speech recognition,” in *Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation*, 2008, pp. 246–250.

[29] M. Khanzadi and H. Veisi, “Creating Kid’s Speech Corpus for Phonological Awareness Assessment,” *24th Natl. CSI Comput. Conf.*, Iran, 2019.

[30] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement" *IET signal Process.*, vol. 6, no. 1, pp. 54–63, 2012.

بازشناسی واج و هجای فارسی با شبکه عصبی بازگشتی با کاربرد در آزمون آگاهی واج‌شناختی

مریم خانزادی [۱]، هادی ویسی [۱،*]، رقیه علی‌نقی‌زاده[۱] و زهراسلیمانی[۲]

[۱] دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران.
[۲] دانشکده توانبخشی، دانشگاه تهران، تهران، ایران.

ارسال ۲۰۲۱/۱۰/۰۵؛ بازنگری ۲۰۲۱/۱۲/۱۶؛ پذیرش ۲۰۲۲/۰۱/۳۱

چکیده:

از مشکلات عمده در کودکان مبتلا به اختلالات یادگیری، عدم مهارت کافی در آگاهی واج‌شناختی است. برای تشخیص و ارزیابی این مهارت، آزمون آگاهی واج‌شناختی شامل خرده‌آزمون غیرگفتاری و گفتاری انجام می‌شود. در حال حاضر این آزمون برای زبان فارسی با شیوه‌ی مبتنی بر کاغذ برگزار می‌شود. در این پژوهش جهت ایجاد رغبت در کودکان، سرعت و جذابیت بخشیدن به اجرای آزمون و با کمک به آسیب‌شناسان گفتار و زبان، برای نخستین بار سامانه‌ی جامع تحت وب آگاهی واج‌شناختی فارسی با امکان بازشناسی طراحی و پیاده‌سازی شد. چالش اصلی، پیاده‌سازی خرده‌آزمون‌های گفتاری با استفاده از سیستم بازشناسی گفتار مبتنی بر واج، هجا و زیرواژه‌های فارسی (کلمه‌های با حذف واج در ابتدا، وسط و انتهای کلمه) است. از مهم‌ترین نیازمندی‌های طراحی و ساخت ماژول بازشناسی گفتار کودک، دادگان گفتاری کودکان با واژگان خاص در نظر گرفته شده در خرده‌آزمون‌ها است که به همین منظور، پیکره گفتاری کودکان به نام Persian kids Speech با حجم ۳۳ ساعت و ۴۴ دقیقه، جمع‌آوری شد. در سیستم تشخیص واج از پیکره‌ی صوتی کودکان با ۳۰ واج (۲۹ واج فارسی و سکوت به‌عنوان یک واج)، برای سیستم تشخیص هجایی از ۲۶ هجا (۲۵ هجا و سکوت به‌عنوان یک هجا) و برای هرکدام از سیستم‌های زیرواژه ۱۰ زیرواژه استفاده شد. برای استخراج ویژگی از ضرایب MFCC و از شبکه عصبی BiLSTM، جهت ساخت مدل‌های آوایی استفاده شد. میزان دقت برای تشخیص واج ۸۵,۵ درصد و برای تشخیص هجا ۸۹,۴ درصد است. میزان دقت حذف واج اولیه، میانی و نهایی به ترتیب ۹۶,۷۶، ۹۸,۲۱ و ۹۵,۹ درصد است.

کلمات کلیدی: گفتاردرمانی، آزمون آگاهی واج‌شناختی، بازشناسی گفتار کودکان، بازشناسی واج و هجای فارسی، شبکه عصبی حافظه کوتاه‌مدت ماندگار.