Int. J. Nonlinear Anal. Appl. 14 (2023) 1, 3111–3119
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.27851.3735

Hidden Markov model and Persian speech recognition

Masoume Shafieian

Department of Technology and Media Engineering IRIBU University, Tehran, Iran

(Communicated by Mohammad Bager Ghaemi)

Abstract

Nowadays, speech recognition, which simply refers to the process of converting an audio signal into its equivalent text, has become one of the most important research topics. Although many studies have been conducted in the field of speech recognition for many languages of the world, but can be said that no more study has been conducted in the Persian language and therefore it is necessary to conduct more studies in this field. Since Persian is a rich language that can create many new words by adding a suffix (prefix) to its main root, so it can be said that the success rate of voice recognition programs in this language has also increased with the increase in the number of phonemes and therefore can have a significant improvement. Therefore, in this study, a practical approach to Persian speech recognition based on syllables, which are a unit between phonemes and words, has been used and done by the hidden Markov model. After obtaining syllable utterances, multiple coefficients are calculated for all syllables. Finally, suitable models were created and the success rate was calculated by conducting tests for the systems. To measure the performance of the system, the error rate criterion was used. The results of this study show that the word error rate for the hidden Markov model was 18.3% and increased the system performance by approximately 16% after post-processing.

Keywords: hidden Markov model, persian language, speech recognition, syllable, syllable based speech recognition
2020 MSC: 62M05

1 Introduction

Phonemes are linguistic units that differentiate the words of the language from each other. Phonemes appear in the signal level as a static function per unit of time, and for this reason, in speech processing, the values of each representation parameter in the frequency domain in successive frames that belong to a phoneme or part of a phoneme do not have statisticaly significant differences. For this reason, phoneme as a structural fact in the speech signal has become extremely important in speech processing [23]. We know that every spoken language has a certain set of faces and researches of linguistic sciences show that the general characteristics of this set are generally different for various languages. For example, the set of English speech phonemes has 44 elements, while the set of Persian speech phonemes has 19 elements [26].

Nowadays, due to the expansion of the use of multimedia contents, the importance of searching for sound, image, text and video has become very important. The recognition of spoken phrases provides the possibility to retrieve a phrase from the text representation of that speech, and therefore the recognition of spoken phrases can be considered a preliminary step for the recovery of spoken documents [25]. Therefore, the main use of spoken phrases can be summarized in Audio Indexing (AI) and speech data mining, (SDM). The most important challenge in speech

Email address: shafieian@iribu.ac.ir (Masoume Shafieian)

Received: June 2022 *Accepted:* August 2022

recognition is audio signals, sound transmission and recording environments, which are different from one speech to another. Another challenging situation is the tone of voice caused by the speaker's emotional state [14]. One of the other challenges in this field is the wide variety of spoken documents. The existence of discrete or continuous speech, clean or noisy speech is also one of the other challenges in this field, which causes the speed and accuracy of the proposed solutions to be different. Therefore, these methods are not yet accurate and powerful whit compared to text recovery methods [13].

Speech recognition can generally be classified into two categories, "pattern-based" or "model-based", according to the characteristics of the used method. Dynamic Time Warping (DTW), Linear Time Alignment methods (LTA), are examples of pattern-based speech recognition. Multilayer Sensor, support vector machine (SVM) and hidden Markov model (HMM) are models based methods. In template-based methods, a template is created for each sound sample and compared with this template. On the other hand, in model-based methods, trained on sound samples and general features are extracted and the final model is created [22].

In another classification, we can divide speech search methods into the following two categories: a- Direct Phonetic Matching, b- Automatic Speech Recognition (ASR). In the direct phonetic matching approach, such as the dynamic time warping method, an attempt is made to directly match the phonetic features between the speech and the keyword. But in the second approach, first the speech is converted into text by the automatic speech recognition system, and later the search is performed on the ASR output text using multiple text recovery methods [27].

The first speech recognition studies began in the late 1940s. However, these studies have gained momentum in the last 30 years. Most of these studies have used phonemes and lexical units as basic components in speech recognition. However, it must be acknowledged that determining the boundaries between phoneme units is a very difficult process that must be carefully considered. In addition, systems based on lexical units, although they do not have the problems of systems using phonetic units, but they involve a lot of calculations and data processing [14]. In a simple classification, speech recognition systems can be classified by the number of words in small scale (1-100 words), medium (100-100 words) and large scale (more than 1000 words).

Speech recognition is a process of converting speech signal to a sequence of word. Various approach has been used for speech recognition. We can see the classification of speech recognition in the below diagram:

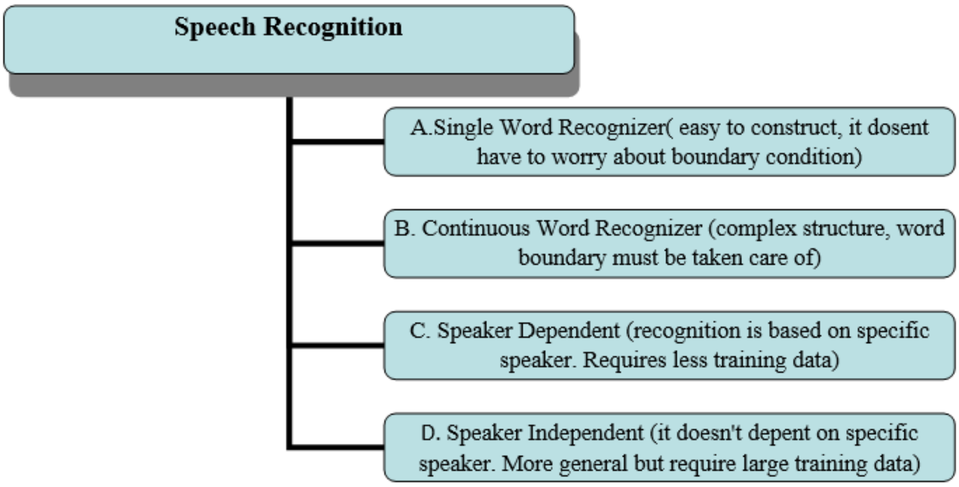
Speech Recognition
A.Single Word Recognizer(easy to construct, it dosent have to worry about boundary condition)
B. Continuous Word Recognizer (complex structure, word boundary must be taken care of)
C. Speaker Dependent (recognition is based on specific speaker. Requires less training data)
D. Speaker Independent (it doesn't depent on specific speaker. More general but require large training data)

Figure 1: Speech recognition clasification

Usually, the preferred features in speech recognition are Linear Prediction Coefficients (LPC), such as: [7, 17, 18, 22] and the coefficients of MFCC and Parcor. The most widely used methods in studies are: Dynamic Time Warping (DTW), such as [12, 15, 16], Artificial Neural Networks(ANN) and hidden Markov model (HMM) such as Studies: [14, 24, 22].

Extensive studies for speech recovery in languages such as English [5, 6] have been conducted on a large scale (more than 1000 words). In recent years, new researches have been conducted on languages such as Pashto, Vietnamese, Turkish, with a limited and medium scale, which have reached acceptable results [1, 26]. Although the studies in the Persian language are very limited, which have been done with both limited and large data, but it is necessary to carry out more studies in this field. Therefore, this study has been done to expand the scope of internal studies and pay more attention to applied resarche and theoretical findings in this field.

In the continuation of the article and in the second part of this research, a summary of the findings of previous internal studies is presented and then general information about the overall structure of the system is presented. In the rest of the article, it pointed out how to determine the boundaries of syllables and finally, how to extract the features used from the sound signals of syllables is explained. Then, Markov's hidden method, incremental algorithm, system tests are respectively explained and finally conclusion and summary are stated.

2 Research background

With regard to the previous study and related to the subject that were mentioned in the previous sections, in this section only a number of internal researches that are consistent with the method of this study will be mentioned.

Khanzadi et al [11] recognized Persian phonemes and syllables with neural networks and for the first time launched a comprehensive system for this purpose. 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 findings of this study show that 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.

Asadolahzade and Homayounpour [2] have investigate to improve phoneme sequence recognition with using hidden semi-Markov model (HSMM) and neural networks (HSMM-DNN). Furthermore, they investigate the performance of a post-processing method that corrects the phoneme sequence obtained from the neural network based on our knowledge about phonemes.The experimental results obtained using the Persian Fars Dat corpus show that using the extended Viterbi algorithm on HSMM achieves phoneme recognition accuracy improvements of 2.68% and 0.56% over the conventional methods using Gaussian mixture model-hidden Markov models (GMM-HMMs) and Viterbi on HMM, respectively. In addition, postprocessing method also increases the accuracy compared to before its application.

Zoughi and Homayounpour [28] in a study related to Persian language, by presenting a adaptive windows convolutional neural network(AWCNN), have investigated the speech recognition system in relation to the difference in expression between speakers and the difference in the expressions of a speaker.The obtaind results and analysis on FARSDAT and TIMIT datasets show that,for phone recognition task, the proposed structure achieves 1.2%, 1.1% absolute error reduction with respect to CNN models respectively, which is a considereble improvement in this problem. They conclude that the use of speaker information is very beneficial for recognition acuuracy.

Sheikh Zadegan [23] has studied the effectiveness of Persian speech phonemes in terms of speaker recognition. To estimate phoneme efficiency, he used a criterion defined as the ratio of “inter-speaker distance (IerSD)” of phonemes to “intera-speaker distance (IraSD)”. The results of the tests and calculations performed with the “FARSDAT” dataset showed that vowels and semi-vowels are in the first place in terms of efficiency in speaker recognition.

Homayounpour, Mousavi [9] have used the hidden Markov methods to model the parameters related to speech units to implement the synthesis system. In order to generate speech synthesis parameters by HMMs, an algorithm has been used in which the characteristics of Mel Frequency Cepstral coefficients and pitch frequency, as well as their first and second derivatives have been used. The results of this study show that the scores obtained for the training sentences (the sentences that were available in the used data set) according to the determined parameters set in 4.2, 4.4 and 4.1 respectively, and for the experimental sentences (sentences outside the used data set) are 4.3, 4.2 and 3.4 respectively.

Salehi [20] used hidden Markov models and artificial neural networks to create a speech recognition system capable of recognizing Persian digits. She used the CSLU toolbox to implement the combined ANN/HMM model for Persian speech recognition and collected 210 samples of the speech of a male person and after removing the noises, he manually labeled 47 of the samples. Then, the remaining training samples were automatically labeled and new ANN neural networks were created for the final recognition of the three-layer MLP. To extract the features, four methods including MEL (12 coefficients), MEL derivative (12 coefficients), energy (1 coefficient), and energy derivative (1 coefficient) were used and by applying recognition on the data, the success of the test was 99.4%, which Considering the small number of speech data, it is considered a very suitable result.

3 General structure of the model

As shown in Figure 2, our proposed developed system consists of four stages. These stages are respectively pre-processing stage, detection of syllable boundaries and features, hidden Markov model (speech recognition method) and finally post-processing. In the first step, audio signals of predetermined words are pre-processed. In the second step, the feature vectors corresponding to each syllable sound signal are determined. In the third stage, syllable recognition is done with Markov's hidden method [14] and in the last stage, post-processing method are performed to increase the success of speech recognition.

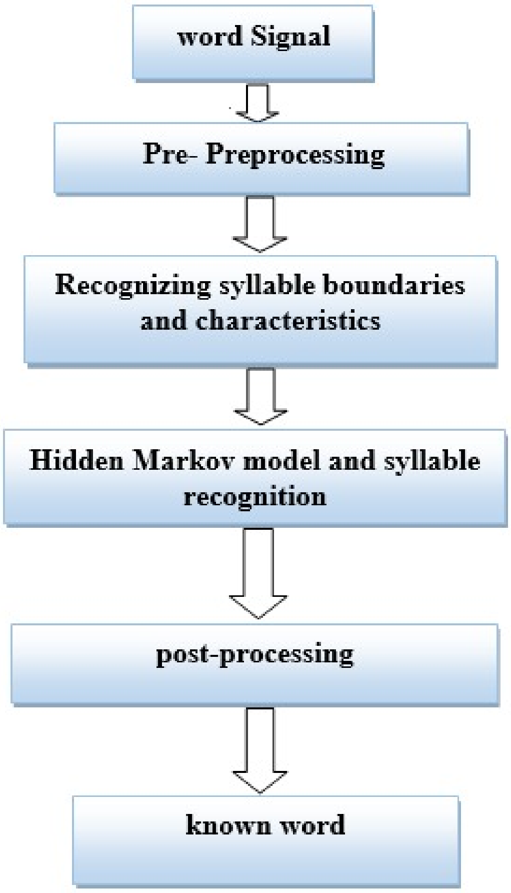
word Signal
Pre- Preprocessing
Recognizing syllable boundaries and characteristics
Hidden Markov model and syllable recognition
post-processing
known word

Figure 2: General structure of the system

In order to calculate the feature pattern of each syllable, the small Persian data base "FARSDAT" [3] is used. In this data, 304 speakers randomly read twenty sentences. The data of this database were recorded in a low-noise environment with an average signal to noise of about 31 db and a sampling rate of 22050 Hz. These data are segmented and labeled at the phoneme level. All 450 unique sentences used in this database have been used. From these data, which contain 1414 unique words whose phonetic equivalents are available in the dataset, samples were taken and pre-processing was performed using 16-bit pulse code modulation (PMC). In the pre-processing stage, the average audio signals were rearranged to become zero.

To get the new sound signal (y_n), we can use equation (3.1) where x_n is the sound signal; m is the average sound signal:

$$y_n = x_n - m, \quad m = \frac{\left(\sum_{i=1}^{k} x_i\right)}{k} \tag{3.1}$$

Before the syllable boundaries are defined and the feature extraction process is performed, the sample sounds are pre-emphasized. Then syllable boundaries are defined. The sound samples of each syllable are divided into 20 ms frames and Hamming windowing is applied on the frames. The overlap between the frames is considered 10 milliseconds and then for each frame of the syllable, 8 feature values are obtained from the LPC, parcor and MFCC feature vectors.

With the HMM speech recognition method, the HMM syllable model is created for each syllable in the training phase. Then, by calculating the similarity between syllable sound signals and syllable patterns, the recognized syllables in the word are determined. Post processing is done at the end of proses to realize better recognition. In this study, applications are coded with Matlab software version 2019.

4 Setting syllable boundaries

The method of determining syllable boundaries consists of two stages. The first step is the process of determining the starting and ending points of words and audio signals. For this, the parts without noise are removed until the part where the word is pronounced and again from the place where the pronunciation of the word ends until the end.The second step is the process of determining the boundaries of the syllables in the word. The algorithm for determining syllable boundaries is given below.

4.1 Algorithm for determining the boundaries of syllables in a word

1. After determining the beginning and end indices (SB and SS) of the sound, syllable boundaries are determined with the following algorithm:

$$n = (n_1, n_2, \dots, n_k) = (\tilde{x}_{SB}, \tilde{x}_{SB+1}, \dots, \tilde{x}_{SS}) \tag{4.1}$$

2. The n vector is divided into windows with k number of non-overlapping samples. The $\bar{n}$ vector is also the average of each window with L samples:

$$\bar{n} = (\bar{n}_1, \bar{n}_2, \dots, \bar{n}_p), \qquad P = \frac{K}{L} \tag{4.2}$$

$$\bar{n}_i = \left(\sum_{n=m^*L}^{(i+1)^*L-1} n_m \right) / L, \quad i = 1, 2, \dots, p \tag{4.3}$$

3. The slopes between successive values of the N vector are calculated and the training vector is formed. For $i = 1, 2, \dots, p-1$,

$$\bar{n}_E = (\bar{n}_{E_1}, \bar{n}_{E_2}, \dots, \bar{n}_{E_{p-1}}) \quad \text{ve} \quad \bar{n}_{E_i} = \bar{n}_{i+1}/\bar{n}_i \tag{4.4}$$

4. A new vector $a = (a_1, a_2, \dots, a_{p-1})$, consisting of +1 and -1, is calculated from the slope vector. In this way, the increasing and decreasing vector is calculated. Therefore, we will have:

```
For $k = 1$   To $p - 1$
              If $a_{k-1} = 1$ and $a_k = -1$
              Otherwise $a_k = -1$
              End                                   (4.5)
```

5. H: Number of syllables in the word

```
H = 0
For $k = 2$ To $p - 1$
if $a_{k-1} = 1$ and $a_k = -1$
if $H = H + 1$
End                                                 (4.6)
```

6. Index groups containing -1 values in vector a are values with main syllable boundaries. The margin of syllables will be $H - 1$. The $s = (s_1, s_2, \dots, s_{H-1})$ vector is calculated for syllable boundaries. The S_i values are the values that hold the indices of the $\tilde{x}$ vector:
For $k = 1$ To $H - 1$
If the index in the middle of the indexes with kth consecutive -1 values in vector a is W :
$Sk = SB + L * W$
END

7. So far, the starting value of SB and the ending value of SS of the sound have been determined exactly in the $\tilde{x}$ vector. The vector S is the vector of approximate boundary indices between syllables. To find more precise boundaries, the following procedure is performed and its $\tilde{s} = (\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_{H+1})$ vector is obtained. Here, assuming $\tilde{s}_1 = SB$ and $\tilde{s}_{H+1} = SS$:

For $i = 1$ To $H - 1$
In the interval $S_i - 500$ and $S_i + 500$, windows with the number of 20 samples are formed and after calculating the average of these windows, a window with the smallest average is selected and the index between this window, will be equal to $\tilde{s}_{i+1} = q$.

8. Syllable boundary indices in the $\tilde{x}$ sound vector are found in the form of the $\tilde{s}$ vector. The beginning of the kth syllable will be from $\tilde{s}_k$ and the end of that syllable will be index $\tilde{s}_{k+1}$. In each word, will be H number of syllables.

5 Identifying the features of LPC, Parcor and MFCC

Before calculating LPC, parcor and MFCC features, vectors of syllable audio signals are filtered by preprocessing. Then divided into 20 millisecond frames. After a 10 ms overlap, a Hamming windowing is applied to each frame. In this section, the correlation vector and the autocorrection vector [1, 18] are calculated. In the following, with the method proposed by Rabiner, Juang [18], Predictive linear coding and Parcor feature extraction. At the end, the value of eight characteristics of LPC, parcor and MFCC are obtained for each frame. Finally, these generated feature vectors for each syllable are saved with file name, syllable name and file extension "fetN" for later use. The letter N indicates that the syllable in the word is which syllable.

6 Hidden Markov model

HMM is a method that statistically models audio signals. This method is one of the most successful speech recognition methods and has the ability to mathematically describe audio signals in a very convenient way. The inputs in this method are a representation of time-dependent discrete data that is displayed as a vector. An HMM consists of finite states, and each of them is connected by probability distributions. Transitions between states are determined by probability values called "transition probabilities" [4]. An observation or outcome in a state is obtained from the probability distributions that depend on it. Since the states are not visible to outside observers, then word "hidden" is used in this method. To define the HMM method, the following variables are needed:

N: the number of modes in the model.

M: number of viewing symbols in alphabets. If the observations are continuous, M will be infinite.

A: Transfer probabilities as seen in equation (6.1).

$$\begin{aligned} &A = \{a_{ij}\} \\ &a_{ij} = p\{q_{t+1} = j | q_t = i\}, \quad 1 \le i, j \le N \end{aligned} \tag{6.1}$$

q_i represents the current state. The transition probabilities provide the normal probability constraints in equations (6.2) and (6.3):

$$a_{ij} \ge 0, \quad 1 \le i, j \le N \tag{6.2}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad 1 \le i \le N \tag{6.3}$$

The probability distribution of states can be seen as shown in equation (6.4).

$$\begin{aligned} &B = \{b_j(k)\} \\ &b_j(k) = p\{0_t = v_k | q_t = j\}, \quad 1 \le j \le N \quad 1 \le i \le M \end{aligned} \tag{6.4}$$

v_k, in alphabetical order, represents the observation symbol k. ot is the current parameter vector. The possible constraints in equations (6.5) and (6.6) must be satisfied.

$$b_j(k) \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M \tag{6.5}$$

$$\sum_{k=1}^{M} b_j(k) = 1, \quad 1 \le j \le N \tag{6.6}$$

If the observations are continuous, we should use probability density function instead of discrete probability. In this case, we must specify the parameters of the probability density function. In general, as seen in equation (6.5), the probability density M of the Gaussian distribution approaches the sum of their approximate weights. Possible restrictions must be met in the following relationship:

$$b_j(O_t) = \sum_{m=1}^{M} C_{jm}\Omega\left(\mu_{jm}, \sum_{jm}, O_t\right)$$

$$C_{jm}: \text{ weighting coefficients} \tag{6.7}$$

$$\mu_{jm}: \text{ Mean vectors}$$

$$\sum_{jm}: \text{ Joint exchange matrices}$$

$$C_{jm} \geq o, \quad ,1 \leq J \leq N, \quad 1 \leq m \leq M \tag{6.8}$$

$$\sum_{m=1}^{M} C_{jm} = 1, \quad 1 \leq j \leq N \tag{6.9}$$

The initial state distributions are given in the following equations.

$$\pi = \{\pi_i\}$$

$$\pi_i = p\{q_1 = i\}, \quad 1 \leq i \leq N \tag{6.10}$$

If we want to use a more compact notation, we can express the probability distribution of this method using a continuous density as seen in Equations (6.11) and (6.12):

$$\lambda = (A, B, \pi) \tag{6.11}$$

$$\lambda = (A, c_{jm}, \mu_{jm}, \Sigma_{jm}, \pi) \tag{6.12}$$

7 Post-processing algorithm

After completing the syllable recognition process with the HMM method, the syllables are combined and the known word is identified.However, this found word may be a non-Persian word as a result of misrecognition. In order to increase recognition success, each syllable is sorted according to the first 10 orders. Therefore, Persian words are searched by combining syllables based on the highest order, and if a Persian word is found, the identification process ends.

N: the number of syllables related to the word retrieved from the test database.

$Hk(S)$: The kth syllable of the tested word is the most similar to the sth ranked syllable.

1. $i = 1, 2, 3, \ldots, 10$ and Si is one of the ten syllables that is most similar to the i-th syllable.
   Syllables are combined as $H1(S1)H2(S2)\ldots H10(S10)$ and a new word is formed. A total of 10^N words are obtained.
2. A level determined for each word. The sum of the rows of syllables that make up the word in step 1 is calculated and this sum will be the level of that word.
3. The word level starts with the smallest one and if this word exists in the word database, the word is found and the process ends regardless of other words. If there are no words in the database, the system cannot find a word.

8 Implementation of tests and research findings

After determining the syllable boundaries of audio files, LPC, Parcor and MFCC features of each syllable were calculated. 10 syllables similar to each syllable of the words in the test database can be obtained using the hidden Markov model method. Syllables with the least distance value will be the most similar syllables.

If the syllables with the smallest distance from the syllables of the word in the test database are combined, the closest text word is obtained. The detection rates of the system depending on the features used and whether post-processing was used or not are given in Table 1. Accordingly, the recognition success using post-processing increased by about 16%. The greatest success was achieved using post-processing in the MFCC feature, with a success rate of 81.3%.

Table 1: System word error rate

Speech Recognition Method	Characteristics		
	LPC	MFCC	PARCOR
Hidden Markov	41.1	36.3	32.5
Hidden Markov (post-processing)	23.9	19.5	18.7

9 Conclusion

In this paper, speech recognition systems for discrete and speaker-dependent Persian words based on syllables were developed using hidden Markov model method. As main features, linear predictive coding (LPC), parcor and MFCC features were selected and the programs were implemented and compared. The results of the tests showed that the post-processing method included in the system has greatly increased the performance of the system. The most successful feature was related to the MFCC system and the word error rate was determined to be 18.7%. After MFCC, the order of feature success belonged to parcor and LPC.

References

[1] A. Asliyan, K. Günel and T. Yakhno,*Syllable Based Speech Recognition Using Dynamic Time Warping*, Academic Informatics, Canakkale Onsekiz Mart University, Canakkale, 2008.

[2] M. Asadolahzade Kermanshahi and M.M. Homayounpour, *Improving phoneme sequence recognition using phoneme duration*, J. AI and Data Min. **7** (2018), no. 1, 137–147.

[3] M. Bijankhan, J. Shcikhzadegan, M.R. Rohani, Y. Samareh, C. Lucas and M. Tebyani, *FARSDAT- The speech database of Farsi spoken language*, Proc. Aust. Conf. Speech Sci. Technol. **2** (1994), 826–831.

[4] M. Farsinejad, B. Zamani Dehkordi and A. Akbari, *Proposing a two-stage sound detector method based on the hidden Markov model*, The fourteenth Ann. Nat. Conf. Iran. Comput. Assoc., Amirkabir University of Technology, 2007.

[5] J.G. Fiscus, J. Ajot, J.S. Garofolo and G. Doddingtion, *Results of the 2006 spoken tcrm detection evaluation*, Proc. ACM SIGIR Work, 2006, pp. 51–55.

[6] J.S. Garofolo, C.G.P. Auzance and E.M. Voorhees, *The TREC spoken document retrieval track: A success story*, Proc. TREC-8 **8940** (1999), no. 500-246, 109–130.

[7] A. Harma, *A comparison of warped and conventional linear predictive coding*, IEEE Trans. Speech Audio Process. **9** (2001), no. 5, 579–588.

[8] A. Harma, *Linear predictive coding with modified filter structures*, IEEE Trans. Speech Audio Process. **9** (2001), no. 8, 769–777.

[9] M. M. Homayunpour and S. M. Mousavi, *Generation of Persian speech synthesis parameters using hidden Markov and decision tree models*, J. Comput. Sci. Engin. **2** (2007), no. 1–3.

[10] R.J. Jones, S. Downey and J.S. Mason, *Continuous speech recognition using syllables*, Proc. Eurospeech **3** (1997), 1171–1174.

[11] M. Khanzadi, H. Veisi, R. Alinaghizade and Z. Soleymani, *Persian phoneme and syllable recognition using recurrent neural networks for phonological awareness assessment*, J. Artif. Intell. Data Min. **10** (2022), no. 1, 117–126.

[12] J. Kruskall and M. Liberman, *The symmetric time warping problem: From continuous to discrete. In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley Publishing Co., 1983.

[13] L. Lee, J. Glass, H. Lee and C. Chan,*Spoken content retrival beyond cascading speeech rcognition whit text retrival*, IEEE/ACM trans. Audio Speech Lang. Process. **23** (2015), no. 9, 1389–1420.

[14] E. Mengusoglu and O. Derro, *Turkish LVCSR: Database preparation and language modeling for an agglutinative language*, ICASSP'2001, Student Forum, May, Salt-Lake City, 2001.

[15] C.S. Myers, L.R. Rabiner and A.E. Rosenberg, *Performance tradeoffs in dynamic time warping algorithms for isolated word recognition*, IEEE Trans. Acous. Speech Sig. Process. **ASPP-28** (1980), no. 6, 623–635.

[16] K.K. Paliwal, A. Agarwal and S.S. Sinha, *A modification over Sakoe and Chiba's dynamic time warping algorithm for isolated word recognition*, Signal Process. **4** (1982), no. 4, 329–333.

[17] J.G. Proakis, and D.G. Manolakis, *Digital Signal Processing: Principles and Application*, Prentice-Hall, Upper Saddle River, NJ, 1996.

[18] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prenctice-Hall, Englewood Cliffs, NJ, 1993.

[19] A.E. Rosenberg, L.R. Rabiner, S.E. Levinson and J.G. Wilpon, *A preliminary study on the use of demisyllables in automatic speech recognition*, Conf. Rec. Int. Conf. Acous. Speech Sig. Process. GA, 1981, pp. 967–970.

[20] F. Salehi, *Speech recognition using methods of hidden Markov models and artificial neural networks and hybrid speech recognition systems*, Nat. Conf. Engin. Sci. New Ideas, 2013.

[21] Y. Samere, *Phonology of the Persian language*, University Publishing Center, Second Edition, 1368.

[22] I. Shafran, *Clustering wide context and HMM topologies for spontaneous speech recognition*, Ph.D. Thesis, University of Washington, 2001.

[23] J. Sheikh Zadegan, *Ranking of persian speech phonemes from the point of view of efficiency in speaker recognition*, J. Languge Res. **7** (2015), no. 1, 77–96.

[24] T. Svendsen, K.K. Paliwal, E. Harborg and P.O. Husoy, *A modified acoustic sub-word unit based speech recognizer*, Proc. IEEE Int. Conf. Acoustics Speech Signal Process. 1989, pp. 108–111.

[25] J. Tejedor, D.T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Hernaez, A. Coucheiro-Limeres, J. Ferreiros, J. Olcoz and J. Llombart, *AlBAYZIN 2016 spoken term detection evaluation: An international open competitive evaluation in Spanish*, EURASIA J, Audio. Speech, Music Process. **2017** (2017), no. 1, 1–23.

[26] J. Trmal, M. Wiesner, V. Peddinti, X. Zhang, P. Ghahremani, Y. Wang, V. Manohar, H. Xu, D. Povey and S. Khudanpur, *The Kaldi open KWS system: improving low resource keyword search*, Interspeech, 2017, pp. 3597–3601.

[27] H. Veisey, S.A. Qureshi and A. Bastan Fard, *Recognition of speech phrases for Farsi news of the Islamic Republic of Iran*, Signal Data Process. Quart. **4** (2019), no. 46.

[28] T. Zoghi and M.M. Homayounpour, *Adaptive windows convolutional neural network for speech recognition*, Signal Data Process. Quart. **3** (2017), no. 37.