An Approach to Recognize and Pronounce Words with Alternative Pronunciations in Farsi

Iman Rasekh *Ehsan Rasekh* *Mohammad Eshghi*

Islamic Azad University, Arak, Iran
University of Western Ontario, London, Canada
Shahid Beheshti University, Tehran, Iran

ABSTRACT

In Farsi orthography some words have more than one pronunciation which corresponds to different meanings. For a good text to speech system, the words with alternative pronunciation should be determined. The proposed system in this paper is capable of recognizing and pronouncing the words with alternative pronunciations.

A new definition of parameter Vowel State (VS) is used to determine the phonemes of a word. A multi layer perceptron neural network with 48, 150 and 7 neurons in the input layer, the hidden layer and the output layer is chosen to extract the phonemes.

Comparing with other reported works which employ neural networks the proposed network shows efficient results according to the number of interconnections and performance.

The proposed network is tested over 2024 words and results show a performance index of 85% to 95% depending on the percentage of the training set.

Index Terms— Neural Network, Farsi, Alternative Pronunciation, Text-to-Phoneme

1. INTRODUCTION

In common Farsi orthography short vowels are not written. Therefore several pronunciations are possible for each word and the reader must guess the correct one. Still some words have alternative correct pronunciations. For a proper text to speech conversion words with alternative pronunciation should be determined. In this paper an Artificial Neural Network (ANN) [1] is used to determine if the word has an alternative pronunciation and extract the phonemes for the word.

Sejnowski-Rosenberg NETtalk [2] is one of the pioneer works in ANN based speech synthesis systems. NETtalk takes the letters as the input and gives the phonemes as the output. The network required for the NETtalk system had 203, 120 and 26 neurons in the input, hidden and output layers, respectively. Authors in [3] show that a hybrid neural network and simplified rule base system bring more efficiently in a text to phoneme system. In [4] authors report the implementation of a Staged Backpropagation Neural Networks (SBNNs) and a Self-organizing Maps (SOMs). They use a Staged Neural Network to determine the alternate phonemes. In the both Neural Networks, the first stage distinguishes between single and dual phoneme cases. In the second stage two different neural networks are used in parallel to deal with single and dual phoneme cases separately.

In Farsi the unwritten short vowels cause additional problem to the letter and phoneme complication.

The authors in [5],[6] use a rule based synthesis to construct Farsi phonemes. In [7] a word-phonemes dictionary is used to recognize phonemes in Farsi. The author in [8] applies some modifications to the pioneer work NETtalk [2] to derive phonemes from a Farsi text using neural networks. Moreover, variable called Vowel State (VS) defined in [9] helped increasing the neural networks performance considerably.

In this paper a new approach, using a new definition of parameter Vowel State [9], is introduced. This approach recognizes and pronounces words with alternative pronunciations.

In the next section a brief review of Farsi language is presented. Section 3 shows the details of the proposed design and ANN. Tests and results are presented in the Section 4. In the Section 5 the conclusion is presented.

2. FARSI LANGUAGE AT A GLANCE

Farsi alphabet consists of 32 alphabet letters. All of Farsi alphabet letters can be a consonant. Some of phonemes in Farsi have more than one letter representations. Considering repeated phonemes, there are 23 consonants in Farsi language. Besides consonants, total of 6 vowels are pronounced in Farsi language, including short vowels and long vowels. Long vowels are written in three shapes in Farsi orthography. These three letters can also be consonants.

The short vowels appear as diacritics and are not written in the adult texts [10]. Short vowels are Fathe (ـَ) /æ/, Kasre (ـِ) /e/, and Zame (ـُ) /o/.

978-1-4244-5377-1/10/$26.00 ©2010 IEEE

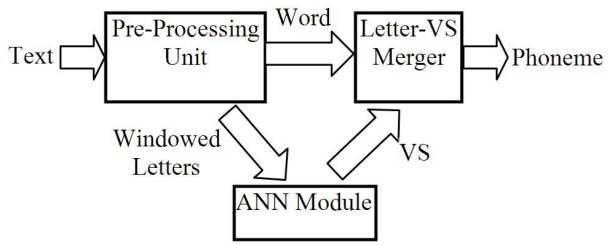

Figure 1. Presented Structure Block Diagram

The letters Aleph "ا", Ye "ی" and Vav "و" can be either a consonant or a long vowel. To determining if any of these three letters is a consonant or a long vowel is a big challenge in any Farsi text to phoneme system.

In addition, in Farsi orthography there are two special diacritics called TASHDID and TANVIN. "TASHDID" (ّ) is a diacritic that means that the letter is doubled. "TANVIN" (ً) is another diacritic that comes often over a letter "Aleph" means that the letter "Aleph" should be replaced by letter "Nun" and the letter before has "Fathe" as short vowel [10].

The unwritten short vowels and unresolved long vowels are the main difficulties in Farsi text to speech conversions.

3. STRUCTURE FOR RECOGNIZING ALTERNATIVE PRONOUNCIATIONS

In this section a new approach to pronounce Farsi orthography with the ability of recognizing the words with alternative pronunciation is introduced. The presented work is base on previous work, reported in [9]. In [9] a new variable called Vowel State (VS) is introduced to efficiently produce phonemes using an ANN.

The new structure comprise of 3 parts, as follows:

- Pre-Processing Unit
- ANN Module
- Letter-VS Merger

This structure is shown in Figure 1 and the parts are described in following subsections.

3.1. Pre-Processing Unit

In the pre-processing unit some normalization processes on the inputted raw text is done. The output of this process is a normalized text which is inputted into the ANN Module and Letter-VS Merger. Main jobs in this unit are as follows:

– Cutting input string into words.
– Eliminating TASHDID and TANVIN.
– Converting digits and signs to letters.
– Eliminating unpronounced letters.
– Determining the vowel states of common and special cases in Farsi orthography.

Table I
Vowel States and Their Description [9]

Vowel State	Description
No Vowel	The letter is a consonant.
Long Vowel	The letter is a long vowel.
Fathe (ـَ)	The letter is a consonant and short vowel /æ/ must be added after it.
Kasre (ـِ)	The letter is a consonant and short vowel /e/ must be added after it.
Zame (ـُ)	The letter is a consonant and short vowel /o/ must be added after it.

Table II
Vowel States For Alternative Pronunciation

Assigned Code	First Vowel State	Alternative Vowel State
0000001	Fathe	No Alternate
0000010	Zame	No Alternate
0000100	Kasre	No Alternate
0001000	Long Vowel	No Alternate
0010000	No Vowel	No Alternate
0100011	Fathe	Kasre
1100011	Kasre	Fathe
0100101	Fathe	Zame
1100101	Zame	Fathe
0101001	Fathe	No Vowel
1101001	No Vowel	Fathe
0100110	Kasre	Zame
1100110	Zame	Kasre
0101010	Kasre	No Vowel
1101010	No vowel	Kasre
0101100	Zame	No Vowel
1101100	No Vowel	Zame
0111000	Long Vowel	No Vowel
1111000	No Vowel	Long Vowel

3.2. ANN Module

Each letter is coded with a 6-bit string. Same categorized coding as [9] is considered to improve the results. To extract the phonemes, a new shape of the Vowel State (VS) is introduced. The combination of the letters of a word in Farsi orthography, and their regarding VSs determine phonemes of the word efficiently. Initial VSs and their definitions are shown in Table I; also the new vowel states for alternative pronunciations are introduced in Table II. The presented new VS has the ability to show the alternative pronunciations.

Some investigation shows 14 vowel states with alternative pronunciations exist in Farsi orthography. Considering 5 states for initial vowel states, a 19-state coding is developed. A 5-bit VS coding, as a common solution is suggested and the neural network is trained but the results were not acceptable. Therefore, different codes assigned and tested. Finally best results are achieved using a 7-bit coding, as shown in Table II.

The most right five bits are dedicated to five VSs so it is possible to show different Vowel States together. Sixth bit of each code shows if the alternative VS exists for the letter or not. If the sixth bit is '0' then the word has just one VS

Table III
VS for Word "روی" and its two possible pronunciation

possible phonemes	letter	Output VS	First VS	Alternate VS
/r/	ر	0101001	Fathe	No Vowel
/v/ & /u:/	و	1111000	No Vowel	Long Vowel
/y/ & /i:/	ی	0111000	Long Vowel	No Vowel
System Outputs:			/rævi:/	/ru:y/

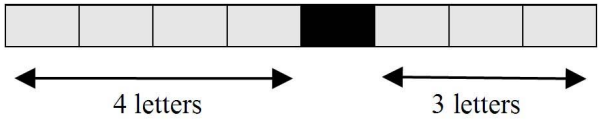

Figure 2. Proposed Windowing, (Farsi orthography is written from

else there is an alternative VS. The most left bit shows the priority of the Vowel States; first VS shows most common VS. If there is more than one letter with alternative VS in a word, first Vowel States must be considered together and alternate Vowel States together.

Table III shows the word "روی" as an example and its two possible pronunciations: /ru:y/ and /rævi:/. The first column shows the possible phoneme(s) of the letter in the second column. Output vowel states are shown in the third column. The first and alternative vowel states are shown in the forth and the fifth columns.

Each letter enters the network with its neighboring letters to determine the phonemes. Different windowing is suggested and tested. In a modern Farsi text, a good result is achieved by entering each letter along with its 3 past and 4 next letters. Therefore, an 8-character windowing is used, shown in Figure 2. When there is no letter before or after a letter, a Null code is entered, instead.

Each inputted letter is replaced by its 6-bit code. Replacing all the letters with their respecting codes, makes a 48-bit string. Each bit is counted as an input of neurons.

Seven neurons are considered to produce the 19 vowel states at the output. Each neuron in output layer represents one bit of the assigned coding as shown in Table II.

The Vowel State (VS) of each letter in a word is determined by a Multi Layer Perceptron ANN. Hyperbolic Tangent function is chosen as the transfer function in the all layers. Before the training, the values of weights in the all layers are chosen randomly between -0.3 and +0.3.

Different networks with different number of neurons in the hidden layer are implemented. A good result is obtained with 150 neurons in the hidden layer.

Error Back Propagation learning rules used to improve the weights. Also momentum method [1] is applied to the ANN to obtain better results. Finally, a multi layer perceptron network is chosen with 48, 150 and 7 neurons in the input layer, the hidden layer and the output layer, respectively. Figure 3. shows the proposed ANN in this design.

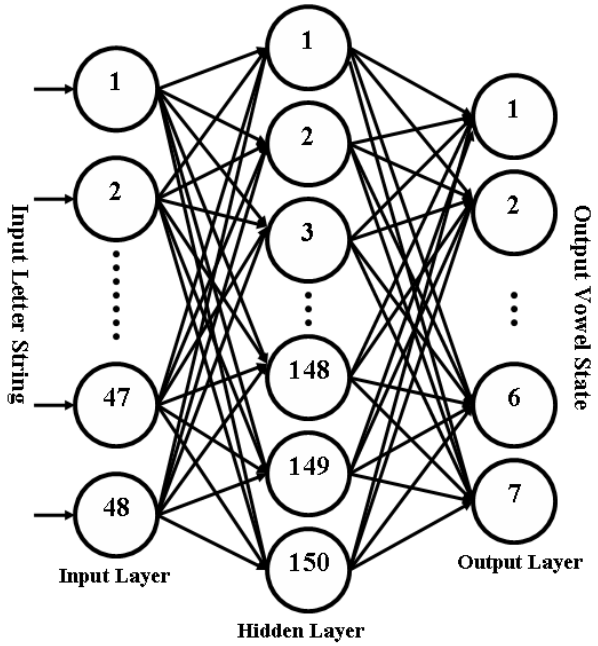

Figure 3. Applied Artificial Neural Network

3.3. Letter-VS Merger

This unit merges ANN module Inputs and outputs to make the phonemes. Where there are alternative pronunciations two phoneme strings for each word is generated. Some general rules and exceptional rules are considered and applied. Exceptional rules are applied according to some of pre-processing rules to improve the results. General rules are applied according to Table I to generate output phonemes. In the case of alternative outputs all the rules applied to each possible ANN outputs.

4. TEST AND RESULTS

The same selection of 2024 common Farsi words as in [9] is used to test the designed network. After preparing phonemes for each word, VSs are added to database using Table II.

Different sets of 65, 75, 85, 95 and 100 percent of the database are used to train the network. The entire database is used to test the network. The letters of each word are windowed as mentioned in Section 3.2.

The results of the proposed network are shown in Table IV. The percentage of correct pronounced words is chosen as the performance index. The correct answers increase by increasing the volume of the database used in the training.

NETtalk system [2] had 203, 120 and 26 neurons in the input, hidden and output layers. Considering this neuron dimension NETtalk involve more than 27000 interconnections. A performance index of 97% is reported for training and testing over a 2000-word corpus [2].

Table IV
Trainings and Results for Different Sets of Training Data

Training Percent	Training Set	Testing Set	ANN Results	Total Results
%65	1316	2024	%82.73	85.31%
%75	1518	2024	%85.29	87.21%
%85	1720	2024	%87.03	88.87%
%95	1922	2024	%92.50	93.12%
%100	2024	2024	%93.48	94.93%

Table V
Comparison of the Present Method With Others Using Neural Networks

	NETtalk [2]	Hosaini [7]	Rasekh [8]	Presented Method
Input Layer	203	48	42	48
Hidden Layer	120	300	110	150
Output Layer	26	40	3	7
Interconnections	27480	26400	4950	8250

The authors in [8] implement ANN in Farsi orthography. The performance index of training and testing over a 12,000 word corpus is 94 %. The mentioned network has 48, 300 and 40 neurons in the input, hidden and output layers, respectively. It involves 26400 interconnections.

The other reported work which employs VS, [9], with a smaller network shows better results, however it cannot recognize the words with alternative pronunciations. The performance index of this system over same 2024 word corpus is 97 %.

No similar work with the ability of determining alternative phonemes in Farsi is reported. Proposed network in this paper has 48, 150 and 7 neurons in the input, hidden and output layers respectively. These dimensions lead to 8250 interconnections. As compared to other networks in [2] and [8], the proposed network is the most efficient. Moreover, the imposed additional interconnections make a good performance in recognizing the words with alternative pronunciations, as compared to that in [9]. The network in [9] also shows less performance index in lower training percentages.

The comparison of these 4 networks is shown in Table V.

5. CONCLUSION

In this paper a new approach for text-to-speech synthesis using neural networks is presented. The presented system has the ability of recognizing the words with alternative pronunciations.

A new definition of Vowel State (VS) is used to determine the pronunciation of a word. Nineteen vowel states are considered to show the phonemes or alternative phonemes of a letter. A network with 48, 150 and 7 neurons at the input, hidden and output layers of the ANN are considered to produce the 19 VSs.

Comparing with other ANN reported in [2] and [8] with more than 27000 and 26000 interconnections, the proposed network has 8250 interconnections. This means 70% reduction in the interconnections of the network moreover it can recognize the alternative pronunciations.

As compared to that proposed in [9], presented work shows about 60% increment in the interconnections of the system, which make it capable of recognizing and pronouncing the words with alternative pronunciations.

The proposed network is tested over 2024 words and the results show about 85% to 95% performance index, depending to the value of the training set. This approach can be applied to other languages such as Urdu, Pashto and Arabic with small changes.

6. REFERENCES

[1] P.J. Braspenning, F. Thuijsman, A.J.M.M. Weijters, *Artificial Neural Networks: An Introduction To ANN Theory And Practice*, Springer, Berlin, 1995.

[2] T.J. Sejnowski, C.R. Rosenberg, "*NETtalk: Parallel Networks That Learn to Pronounce English Text*" .Complex Systems, vol 1, pp 145-168, 1987.

[3] P.R. Gubbins, K.M. Curtis, J.D. Burniston, "*A Hybrid Neural Network/Rule Based Architecture Used as a Text to Phoneme Transcriber*", Proc. International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, pp 113-116, 1994.

[4] M.J. Embrechts , F. Arciniegas, "*Neural Networks for Text-to-Speech Phoneme Recognition*"- PROC IEEE INT CONF SYST MAN CYBERN, vol 5, pp 3582-3587, 2000.

[5] H. Aboutalebi, "*An Implementation of A Robust Farsi Speech Synthesize*r", M.S. Thesis, Sharif University of Technology, Tehran, IRAN, 1998.

[6] F. Daneshfar, B.Z. Azami, W. Barkhoda, "*Implementation of a Text-to-Speech System for Kurdish Language*", ICDT apos;09. vol 1, pp.117 – 120, 2009.

[7] P.G. Georgiou, H. Shirani-Mehr, and S.S. Narayanan, "*Context Dependent Statistical Augmentation of Persian Transcripts*", INTERSPEECH 2004, Jeju Island, Korea, vol 1, pp 853-856, 2004 .

[8] M. Hosaini, M. Homayonpour, "*Farsi Text-to-Phoneme Conversion Applying Neural Networks*" (In Farsi), 8th Iranian Conference on Electrical Engineering, vol 1, pp 195-163, 1999.

[9] E. Rasekh and M. Eshghi, "*An efficient hybrid solution for pronouncing Farsi text"* Springer International Journal of Speech Technology, vol 10, no 2-3, pp 153-161, 2007.

[10] Y. Samare, *Farsi Language Phonology* (In Farsi), Tehran, Iran, Tehran University Press, 1984.