Scientia Iranica, Vol. 11, No. 3, pp 225–233
© Sharif University of Technology, July 2004

SCIENTIA
IRANICA

Syllable Duration Prediction for Farsi Text-to-Speech Systems

B. Nazari*, K. Nayebi[1] and H. Sheikhzadeh[2]

In this paper, two different statistical approaches are used for duration prediction of the Farsi language. These two statistical models are Neural Networks (NN) and Classification And Regression Trees (CART). The first step in this work was to create a database and develop a flexible feature extraction and selection module. In the next step, the output of the feature selection module was used to train both models. The results of the trained models are further studied to determine the most important parameters affecting the syllable duration in Farsi. The model accuracy is evaluated by using separate training and test data. In the third step of this work, an automatic rule generator module was added to the CART model. These duration prediction rules can be easily applied in a rule-based speech synthesis system.

INTRODUCTION

Obtaining rules and parameters that effectively determine the duration in any language has been an important topic of research. One of the very basic tasks in the design of any Text-To-Speech (TTS) system is to develop a model for duration prediction. Development of such models, to accurately represent natural speech duration, is not a trivial task. This is mainly because there are many contextual factors that affect segmental duration.

Since text-to-speech systems must be able to handle all such contexts, generalization is a critical point in any prosody prediction module.

Factors determining the duration of a syllable in a word, phrase or sentence include the phonemes in the syllable, its neighboring syllables, position in the word, phrase and sentence, stress pattern, emphasis, syntactic and semantic functions, etc.

The goal of this research is to use statistical methods to develop models to predict duration for Farsi Text-to-Speech systems. In this work, classification trees and neural networks are used for duration modeling. The results of neural networks and classification trees are then analyzed to determine the most important factors affecting duration in Farsi. An automatic rule generator, based on the CART model, is also developed. The size of the database determines the tree depth and the number of rules, which show the model accuracy. So, as the database grows, more complex trees can be trained without over-tuning and can generate more complete rule-based systems. The paper is organized as follows:

First, a general background of the duration models in TTS systems is given, which is followed by linguistic studies about Farsi duration. The two following sections describe the database, the feature extraction and selection module and the preprocessing methods employed. Then, the duration models, based on neural networks and classification trees, are described. Furthermore, rule-based systems are explained that are automatically derived from CART models and the results are compared to the heuristic rules of Farsi duration. Finally, the test results are presented.

BACKGROUND

Several methods have been used to predict and model duration in different languages. The first group of methods are rule-based, which model duration by a set of equations, i.e., the Klatt model, the modified Klatt model, the models that use factor scales and sums of products models [1-7].

Increasing the complexity of such systems to model more complex interactions leads to systems with many rules whose overall behavior is intractable.

*. *Corresponding Author, Department of Electrical Engineering, Sharif University of Technology, Tehran, I.R. Iran.*

[1] *Department of Electrical Engineering, Sharif University of Technology, Tehran, I.R. Iran.*

[2] *Department of Electrical Engineering, Amir Kabir University of Technology, Tehran, I.R. Iran.*

Adding new rules to the system to improve the performance in all contexts becomes even more difficult as the number of rules increases.

Look-up table methods have been used in some languages, where the most important parameters are found in advance and the table is arranged based on these parameters [8]. Another widely used method is the multiplicative model [9]. This model assumes a base value for the duration and tries to include the effect of each factor in the model as a multiplicative factor.

Tree model is one of the most effective methods in predicting duration [10-12]. These systems derive a decision tree using a database. This model has the potential to be interpreted as a set of rules that can be applied in a rule-based system. This idea is discussed thoroughly in the following sections and it is shown how a complete rule-based system can be derived from tree models in the Farsi language. Another result of using tree models is being able to identify the most important parameters affecting duration for the language under study.

Another powerful method for duration prediction is to use neural networks [12-15]. Networks with different topologies have been used for different languages and are reported to be promising. Recurrent neural networks are among the appropriate neural networks considered for this purpose [16]. The idea behind using recurrent networks is that the network itself somehow saves the history of the previous sections of the sentence and eliminates the need for larger feature sets. Of course, this would be at the expense of problems in the convergence of the network. It is claimed, also, that the first layer of recurrent neural networks models the syntactic structure and the following layers model the prosody [16].

STUDIES IN FARSI PROSODY

In this section, a brief description of Farsi syllable structure, prosody and earlier research results in Farsi Text-To-Speech systems is presented. Farsi is an Indo-Iranian subcategory of the Indo-European language family and is spoken in Iran, Tajikistan and Afghanistan, etc. This research is based on the formal Farsi spoken in Iran.

The study of Farsi prosody is somewhat new, as compared to other languages. Kamyar did the first formal study of Farsi prosody [17], which, mostly, focused on stress positions. According to his studies, the stress positions in Farsi are quite regular compared to other languages, such as English.

There are 6 vowels and 23 consonants in Farsi. (In some texts, a diphthong "ou" is added to Farsi vowels for words such as "Ferdousi" [18].) The vowels are grouped into short and long (a, e and o are short vowels and A, i and u are long vowels), as there is a significant difference in the duration of these two groups in Farsi.

The structure of syllables is very well defined in Farsi. Samareh [19] gives a good review of both syllabic structures and allophones that are used by native speakers of Farsi.

Let C be the symbol for a consonant and V for a vowel. In all Farsi syllables, there is one and only one vowel and one consonant before it. There may be 0, 1 or 2 consonants following the vowel. So, all syllables in Farsi are divided into three groups: CV, CVC and CVCC and the number of vowels is equal to the number of syllables in any given sentence [18-21]. Thus, in Farsi, the task of breaking any phoneme sequence into syllables is trivial. Furthermore, CVCC syllables with a long vowel are very rare in Farsi and usually exist in words that have been adopted from foreign languages.

The difference in the duration of short and long vowels can be easily observed in Farsi poems, which are quite rhythmic and whose rhythm is based on the number of phonemes in a syllable (syllable length) and the type of vowel (short/long). For example, if in a verse, a short/long vowel were changed with another short/long vowel, the rhythm would not change and a Farsi native reader would read the new verse without noticing any change in rhythm. Also, if a CV syllable with a short vowel were replaced by a CV syllable with a long vowel or a CVC syllable, the verse would lose its rhythm and would no longer by considered a rhythmic verse in Farsi.

Mohtashami has introduced few rules for Farsi duration. Although these rules are useful, they are quite limited in covering all cases in Farsi [18]. These rules are as follows:

1. The short vowels in CVCC syllables are lengthened, as compared to other syllables, such as "mard", which means "man";
2. If a vowel in a syllable is long and the consonant following is 'n', then the vowel can be shortened, such as in "ZamAn" which means "times";
3. If a CV syllable is the last syllable in the word and V is a short vowel, it can be lengthened, such as in "parde" which means "curtain";
4. If the vowel 'i' is followed by consonant 'y', it is shortened, such as in "siyAh", which means "black".

Although these rules are valid, they do not give any quantitative clue and it is not evident that these are the most important rules to model the duration. Kamyar discusses different theories about the role of stress in duration. Some theories claim that stress influences duration and some claim that they do not [17].

Samareh and Kamyar have claimed that duration is only a phonetic factor in Farsi and does not have a major semantic role [17,19]. Kamyar reasons that this is because stress does not have a basic influence in Farsi duration. But stress and pitch contour do affect the semantics of the word such as in "rAhat". When stress is placed on the first syllable, it would be a compound meaning "your way", but, when stress is placed on the second syllable, it would be an adjective and would mean "comfortable". The results of statistical models regarding how stress effects duration will be discussed later.

Considering the points mentioned about syllable duration in Farsi and because available Farsi synthesizers are mainly syllable based, this research was developed to model and predict duration at syllabic level.

There have been some studies to develop a Farsi TTS in recent years. Klatt's formant synthesizer has been modified for Farsi [20] and a Harmonic plus Noise Model (HNM) synthesizer and a Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) synthesizer have also been developed [21,22]. Both of these synthesizers use very simple prosody modeling. A text-to-phoneme for Farsi has also been presented as a rule-based approach [23].

DATABASE

A database with a large variety and number of sentences is one of the essential requirements in statistical models. The only database developed for Farsi is FARSDAT. This database is not designed for synthesis purposes and is more suited for recognition. This is because there are no more than 20 sentences from a single speaker in this database and it has many speakers with many dialects in it. As in all studies to design a practical TTS system, the sentences of the databases that are to be used in any TTS systems should be generated from a single speaker in a uniform style and speed.

So, the first step in this study was to create a basic database. In the first step towards building such a database, two hundred sentences were selected. Since a database with this number of sentences cannot be rich in complex structures and contexts, this database was limited to sentences that are not very complex and are indicative. The sentences were selected from three sources: A newspaper, a book and some sentences made by two native Farsi speakers. These sentences were read by a native Iranian male speaker in a formal style and recorded in 44 kHz, 16-bit sampling format.

The second step was to label and transcribe these recordings. Automatic segmentation and labeling should be the first candidate for this step, but there is no such utility for Farsi. The utilities developed for this purpose in other languages, such as Microsoft Wave Editor in Microsoft Speech SDK, could not be used because of the differences in Farsi and English phonetics. So, the entire database was hand-labeled and segmented to syllable level.

The data obtained from these segmentations were saved, along with some further information regarding each sentence. This information includes:

- Duration of each syllable,
- Pause duration,
- Syllable, word transcription,
- Phrasing of the sentences.

In the third step, a small dictionary was prepared to store the type of each word so that the feature extraction and selection module could determine the word type automatically. One of the major issues in Farsi is that, in many cases, special suffixes may be appended to nouns and adjectives. For example, an "e" is attached to the noun when a noun has a descriptive adjective in a noun phrase, e.g., "Abi" (Blue), "AsemAn"(Sky), "AsemAne Abi" (Blue Sky). Another example is in indefinite nouns. Nouns can be made indefinite by an "i" suffix, e.g., "ketAb"(book), "ketAbi"(a book).

The pitch contour of these sentences was also extracted and smoothed for future studies in pitch. These pitch contours were saved in numeric form and in a parameterized form using orthonormal polynomial expansion [16].

FEATURE EXTRACTION AND SELECTION MODULE

The main task of this module is to simplify the process of feature selection and representation. There are many features that may be used as the input to the models and these features can be represented in different formats. This module provides an automated mechanism to select a subset of the available features and present them in an input file to prediction models. A complete list of extracted features is shown in Table 1.

Each feature may be represented in three different formats. Each format may improve the results for a specific model. These three formats are: Simple number, 1 of n and additive 1 of n (temperature format). For example, if a parameter can take values from 1 to 5 and if the current value is 2, it may be represented as:

Simple number	2
1 of n	01000
Temperature format	11000

Comments can also be added in the output file of this module, both for readability and rule generation.

Table 1. List of the features.

Feature	Value
Syllable place in word 1 (Place of syllable in word)	Beginning, middle, end of word, mono-syllable
Syllable place in word 2 (Place of syllable in word)	$1\ldots5$
Syllable length	$1\cdots4$
Phoneme type*	$1\cdots11$
Sentence length in syllable	$1\cdots32$
Phrase length in syllable	$1\cdots16$
Word length 1 in syllables	$1\cdots6$
Word length 2 in phonemes	$1\ldots14$
Word type	$1\cdots16$
Stress	$0, 1$
Word place 1 (Place of word in phrase)	Beginning, middle end of phrase
Word place 2 (Place of word in phrase)	$1\cdots4$

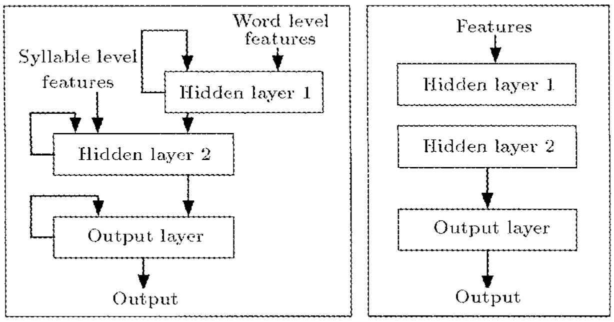
* Since considering each phoneme separately will lead to 29 classes of phonemes, they are grouped to decrease the number of phoneme classes and the feature space dimension. For example, short vowels {'a', 'e', 'o'} are grouped in one class and unvoiced plosives {'p', 't', 'k'} are considered to be in another class.

As previously described, the CART module can automatically generate a rule-based system to predict duration. The feature selection and extraction module may add tags and comments that can be used by the CART module. Using these tags, the rules in the generated rule-based systems are quite readable and can be followed quite easily. All these features may be turned on/off.

DURATION MODELING WITH NEURAL NETWORK

This section describes different types of neural network with different input feature sets, which were applied to the duration analysis and prediction.

The first model is based on the recurrent neural network [16]. A block diagram of this type of neural network is shown in Figure 1a. The feedback nature of this network has the advantage that it can save the history of the previous syllables in the network. On the other hand, because of its feedback structure, stability problems may arise in this type of network. Although this configuration seems to be very promising, experimental results with simple feed-forward networks showed that non-recurrent networks perform better.



(a) Block diagram of RNN (b) Simple feed forward NN

Figure 1. A sample of recurrent neural network (a) and simple feed forward neural network (b).

As mentioned earlier, the goal of this research was not only to develop a model to predict syllable duration, but also to find the most important factors in durational analysis. So, in the first series of the experiments (Experiment 1), only a single parameter, plus the phoneme type, were applied to the neural network. The results are shown in Tables 2 and 3. Each row in these tables shows a test with a feature and the resulting errors with both training and test data.

As can be observed from the table, the number of phonemes in the syllable is the most dominant factor in Farsi syllable duration (sixth row in Table 3). Other important factors are the number of syllables and the syllable position in the word (beginning/middle/end/mono-syllable), respectively (fifth and eighth rows in Table 3). Next, comes the effect of the number of phonemes in the word and number of syllables in the phrase (forth and third rows in Table 3).

Later it is shown that these results are quite in agreement with the results of the CART model. It is also interesting to notice that stress is not a dominant factor in Farsi duration (seventh row in Table 2). This study is in agreement with Kamyar's assessment that stress does not have a significant role in Farsi duration [17].

In the second series of experiments (Experiment 2), the information about previous and next syllables was also supplied to the network. These experiments showed that syllable length is again the most effective factor. Other important factors are word place in the phrase (beginning/middle/end), phrase length in syllables and syllable place in the word (beginning/middle/end/mono-syllable), respectively.

In Experiment 2, prediction error in the test data was much higher than that of the training data, however, the error for the test and training data in Experiment 1 were comparable. This shows that the database size was acceptable for Experiment 1, but was inadequate for Experiment 2. In other words, in order to model the effect of the previous and, possibly,

Table 2. Result of the neural net model with training data (Experiment 1).

With Training Data			
Test		Root Mean Square Error (msec)	Abs Error (msec)
1	Only phoneme type	85.37	67.72
2	Phoneme type + sentence length in syllables	84.78	67.18
3	Phoneme type + phrase length in syllables	75.39	58.88
4	Phoneme type + word length in phonemes	79.37	63.89
5	Phoneme type + word length in syllables	78.62	61.33
6	Phoneme type + syllable length	54.91	43.11
7	Phoneme type + stress	81.97	64.31
8	Phoneme type + syllable place 1	79.12	62.63
9	Phoneme type + syllable place 2	85.04	67.15
10	Phoneme type + word place 1	81.40	63.90
11	Phoneme type + word place 2	85.07	67.54
12	Phoneme type + word type	80.58	64.32

Table 3. Result of the neural net model with test data (Experiment 1).

Test		Mean Square Error (msec)	Abs Error (msec)
1	Only phoneme type	82.5578	66.3492
2	Phoneme type + sentence length in syllables	81.3560	65.4966
3	Phoneme type + phrase length in syllables	79.8050	65.2880
4	Phoneme type + word length in phonemes	78.1224	63.9773
5	Phoneme type + word length in syllables	77.7415	62.6304
6	Phoneme type + syllable length	48.1406	38.9705
7	Phoneme type +stress	80.2993	65.5692
8	Phoneme type + syllable place 1	78.3492	63.0567
9	Phoneme type + syllable place 2	82.1905	66.0317
10	Phoneme type + word place 1	80.6757	65.3243
11	Phoneme type + word place 2	83.4014	66.9478
12	Phoneme type + word type	82.1859	66.5170

next syllables in the prediction of the desired syllable's duration, the database size must be much larger than the present one.

The overall results are summarized in Table 4 and Figure 2. As can be seen in Figure 2, the resulting error for the training data in Experiment 2 is much less than the error in Experiment 1, but the resulting error for the test data in Experiment 2 is higher than that of Experiment 1. This is mainly because of over-tuning which is a direct result of the small database used.

In summary, besides phoneme type, the number of phonemes in the syllable, the length of the word in syllables and the syllable's place in the word (beginning/middle/end/mono-syllable) are the most important parameters in Farsi duration.

CLASSIFICATION AND REGRESSION TREES (CART)

As a second approach, classification trees are employed in the duration modeling of Farsi syllables. CART uses the extracted features described in the feature extraction and selection module.

A simple tree is shown in Figure 3. As can be

Table 4. Important factors in Farsi duration.

Experiment 1		Experiment 2	
1	Phoneme type + syllable length	1	Phoneme type + syllable length
2	Phoneme type + word length	2	Phoneme type + word place 1
3	Phoneme type + syllable place 1	3	Phoneme type + phrase length in syllables
4	Phoneme type + word length in phonemes	4	Phoneme type + syllable place 1
5	Phoneme type+phrase length in syllables		

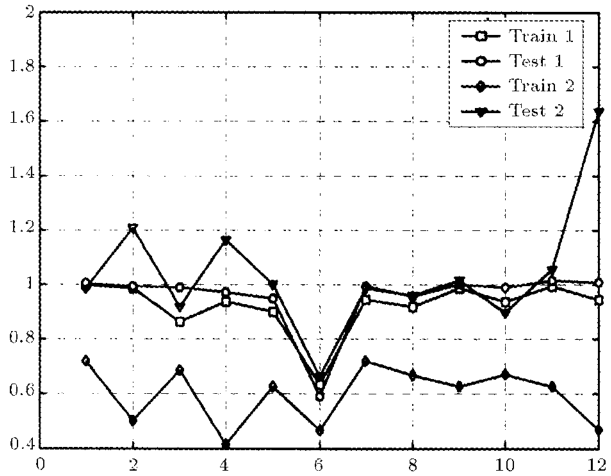


Figure 2. Neural net training errors and test errors with 12 types of input features (Experiments 1 and 2). Horizontal axis shows the input type in accordance with Table 2.

seen, at each nonterminal node of the tree, a question is asked and, depending on the answer to this question, the left or right branch is selected. This is repeated until a terminal node is reached. The prediction for the duration of that syllable is the value in this terminal node. Each question is about one of the features in this database. At each node of the tree, a value has been predicted for the syllable duration. The deeper one travels in the tree, the finer predictions can be obtained.

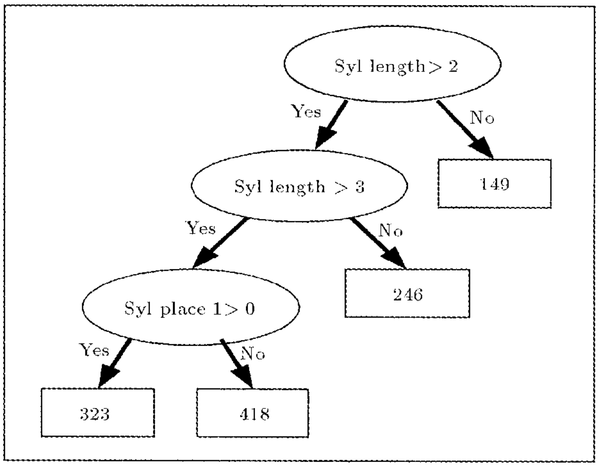


Figure 3. A sample of small CART.

For example, in Figure 3, the first feature to be determined is the length of the syllable. Should the number of phonemes in the syllable (syllable length) be two, the right branch would be selected and the syllable duration would be set to be 149 ms, otherwise, should the syllable have more than 2 phonemes, the left branch would be selected and so on.

The features are divided into two groups [10]. The first group is 'categorical' features, which take on one of a fixed set of unordered values and the second group is 'continuous' features that may take any number for which order is important.

An example of categorical features is phoneme types and classes and an example of continuous features is the number of syllables in a word.

The node question for the first group is:

$$x \in A \quad \text{vs.} \quad x \in X - A \qquad \forall A.$$

And the node question for the second group is:

$$x \leq k \quad \text{vs.} \quad x > k \qquad \forall k.$$

In each node, all such questions about all of the features are considered. Adding two branches to a node results in a smaller prediction error, because finer predictions are made. The question that minimizes the new duration prediction error in that node is selected.

Each terminal node in the tree is called a cluster and the depth of the tree is increased until the desired number of clusters is obtained.

Another problem in classification trees is to have a criteria to stop the growing of the tree. A series of test results are shown in Table 5. Each row in this table shows the result with a cluster number. The first part of the database is used in the tree design (training) and the second part in the generalization test. The prediction errors for both training and test data are shown in Table 5.

As expected, increasing the number of clusters reduces the error in both training and test data. After 26 clusters, error in the test data increases due to over-tuning in training data, which is a usual event in statistical learning. It is expected that with a larger

Table 5. CART simulation results.

Cluster Number	Error in Test Data (msec)	Error in Training Data (msec)
100	45.29	23.71
50	38.06	28.6
40	38.88	31.18
35	33.49	31.91
32	33.53	32.24
30	33.56	32.32
29	33.33	32.46
28	33.19	32.57
27	33.19	32.61
26	33.19	32.70
25	33.42	33.04
24	33.93	33.38
23	33.93	33.44
20	34.07	34.16
10	36.35	40.28
9	46.65	40.62
3	59.42	44.22

database, the optimum cluster number would be larger than 26 and that would result in better prediction accuracy.

The next section explains how CART can be used to derive complete rule-based solutions and discusses the results of these trees with the heuristic consideration in Farsi.

RULE DERIVATION AND ANALYSIS WITH CART MODELS

One of the advantages of the CART model is that it can be used to generate duration rules for any rule-based TTS system. In this section, a module is introduced that automatically generates a rule-based system for the given tree without any manual processing. This module uses the comments and tags provided by the preprocessing module. A small rule-based sample obtained from a simple tree is shown in Figure 4.

For further investigation of duration in Farsi, a number of trees and the corresponding set of rules were derived. These derived rules give an excellent insight into the most important parameters affecting duration in Farsi.

One begins by using the set of rules shown in the example of Figure 4. As can be seen, the first rule separates two-phoneme syllables from three- and four-phoneme syllables. This is a well-known phenomenon used in Farsi poetry. Poems in Farsi have a built-in rhythm. The difference between two-phoneme syllables and syllables with more than two phonemes are widely used by Farsi poets. The second rule separates three-phoneme syllables from four-phoneme syllables. This is also a rule used in Farsi poetry.

```
if(SylLen::5>2)
 {
   if(SylLen::5>3)
   {
       if(SylPlace1::3>0)
        {
              Duration = 7118.13(N = 32)
        }
       else {
              Duration = 9221.40 (N = 48)
         }
     }
    else {
            Duration = 5432.48(N = 361)
      }
 }
else {
       Duration = 3275.13(N = 716)
 }
```

Figure 4. Sample of rule-based system derived from a tree structure.

Finally, one can see the effect of a syllable's place in a word in the 3rd line of Figure 4. It should be noticed that this parameter is assigned a value of 0 for monosyllables, 1 for the first syllable in a polysyllable word, 3 for the last syllable in a polysyllable word and, otherwise, 2.

In the third rule of Figure 4, the syllables for which a syllables place in a word is zero are separated from the others. This means that being a monosyllable is the most important feature in this node for four-phoneme syllables. This can be seen from the path, which was passed to reach this node. In the previous node, the four-phoneme syllables had been separated from the others and in this node, four-phoneme syllables that are in monosyllable words are separated from the others.

For example, consider the word "zahr"(poison) and the word "pAdzahr" (antidote). The duration of the syllable "zahr" in the monosyllable word is longer than the two-syllable word "pAdzahr". According to Figure 4, the duration of four-phoneme syllables, which are in monosyllable words, is 9221.40 msec and the duration for other four-phoneme syllables is 7118.13 msec. It means that four-phoneme syllables in monosyllable words are lengthened about 30 percent, as compared to four-phoneme syllables in other words.

Although viewing a complete tree gives us the most important parameters, it does not give a clear hint as to the effect of each parameter. Adding the nodes one by one gives an insight, in order to proceed in this way. The results are shown in Table 6. Indeed, the index shows the number of non-terminal clusters in the tree and the number of rules and the second column shows the last feature that appeared in the tree. Again, it is seen that stress is not one of the parameters in Table 6. This means stress does not have any major effect on the duration in Farsi. Figure 5 is an example of a more complete set of rules.

RESULTS AND CONCLUSION

Statistical models have been used to predict duration in the Farsi language. The methods employ neural networks and classification and regression trees.

The quantitative results of this research were presented in previous sections, which showed the error to be quite comparable with the results reported for other languages [8,16].

The resources were quite limited for performing acceptable subjective tests. Some experiments were attempted with three different tools: Cool Edit Software, Microsoft English TTS (Whistler in Microsoft speech SDK 4.0) and a HNM synthesizer for Farsi.

Degradation of quality, due to piecewise stretches in Cool Edit, phonetic differences between Farsi and English in Microsoft English TTS and the lack of a pitch prediction module for Microsoft English TTS and HNM Farsi synthesizer, prevents performance of a meaningful subjective test for the time being. However, these limited tests showed that the predicted duration was significantly better than the intrinsic duration syllable of the HNM synthesizer.

In addition, the models were further studied to find the most dominant features that determine syllable duration in the Farsi language. These parameters are: Phoneme type, the number of phonemes in the syllable (being mono or polysyllable) and the number of syllables in the word and phrase.

Table 6. List of features that appear in the tree and in the rule-based system.

Index	Parameter
1	Syllable length
2	Syllable length
3	Syllable place 1
4	Phoneme type of 1st phoneme in syllable
5	Phoneme type of 1st phoneme in syllable (1 of n format)
6	Word type 1 (1 of n format)
7	Phrase length
8	Phoneme type of 2nd phoneme in syllable
9	Phoneme type of 3rd phoneme in syllable

```
if(SylLen::5 > 2)
 {
     if(SylLen::5>3)
      {
          if(SylPlace1::3>0)
           {
               if(Phoneme Type1::16>8)
               {
                        Duration = 9916.50(N = 2)
               }
              else {
                        if(WordType::BN::3>0)
                        {
                                  Duration = 4900.33(N = 3)
                        }
                        else {
                                  Duration = 7157.26(N = 27)
                        }
               }
          }
        else {
                 if(PhraseLen::16>2)
                  {
                    if(Phoneme Type2::16>1)
                     {
                                  Duration = 6678.00(N = 3)
                     }
                    else {
                                  Duration = 8879.29(N = 17)
                      }
           }
          else {
                         Duration = 9701.61(N = 28)
            }
          }
     }
     else {
        if(Phoneme Type1::BN::7>0)
        {
               Duration = 7212.19(N = 47)
          }
         else {
           if(Phoneme Type3::16>3)
            {
                         Duration = 5412.31(N = 170)
            }
           else {
                         Duration = 4875.43(N = 144)
               }
          }
     }
 }
else {
        Duration = 3275.13(N = 716)
 }
```

Figure 5. Sample of a larger derived rule-based system (BN shows format "1 of n" for the parameter).

Finally, the CART model was further developed to generate complete equivalent rule-based systems. These rules give good insight into the features that effect Farsi syllable duration.

REFERENCES

1. Santen, J.V. "Deriving text-to-speech durations from natural speech", in *Talking Machines, Theories, Models & Designs*, G. Bailly, C. Benoit, Eds., Elsevier Publishers, pp 275-285 (1992).

2. Coker, et al. "Automatic synthesis from ordinary English text", *IEEE Transactions on Audio and Electro Acoustics*, **AU-21**(3), pp 293-298 (1973).

3. Klatt, D.H. "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *Journal of the Acoustical Society of America*, **59**, pp 1208-1221 (1976).

4. Klatt, D.H. "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America*, **82**(3), pp 737-793 (1987).

5. Umeda, N. "Vowel duration in American English", *Journal of the Acoustical Society of America*, **58**, pp 434-445 (1975).

6. Umeda, N. "Consonant duration in American English", *Journal of the Acoustical Society of America*, **61**, pp 846-858 (1977).

7. Santen, J.V., et al. "Multi-lingual duration modeling", *5th European Conf. on Speech, Communication and Technology (Eurospeech)*, Rhodes, Greece, **5**, pp 2651-2654 (1997).

8. Xavier, F.S. and Banga, E.R."Segmental duration modelling in a text-to-speech system for the Galician language", *6th European Conf. on Speech, Communication and Technology (Eurospeech)*, Budapest, Hungary, pp 1635-1638 (1999).

9. Shinan, L.U., et al. "Prosodic control in Chinese TTS system", *6th Int. Conf. of Spoken Language Processing (ICSLP)*, Beijing, China (2000).

10. Riley, M.D. "Tree-based modeling of segmental durations", in *Talking Machines, Theories, Models & Designs*, G. Bailly, C. Benoit, Eds., Elsevier Publishers, pp 265-273 (1992).

11. Deans, P., et al. "CART-based duration modeling using a novel method of extracting prosodic features", *6th European Conf. on Speech, Communication and Technology (Eurospeech)*, **4**, Budapest, Hungary, pp 1823-1826 (1999).

12. Fackrell, J.W.A., et al. "Multilingual prosody modeling using cascades of regression trees and neural networks", *6th European Conf. on Speech, Communication and Technology (Eurospeech)*, **4**, Budapest, Hungary, pp 1835-1838 (1999).

13. Campbell, W.N. "Syllable-based segmental duration", in *Talking Machines, Theories, Models & Designs*, G. Bailly, C. Benoit, Eds., Elsevier Publishers, pp 211-224 (1992).

14. Córdoba, R. and Vallejo, J.A. "Automatic modeling of duration in a Spanish text-to-speech system using neural networks", *6th European Conf. on Speech, Communication and Technology (Eurospeech)*, **4**, Budapest, Hungary, pp 1619-1622 (1999).

15. Corrigan, G., et al. "Generating segment durations in a text-to-speech system: A hybrid rule-based/neural network approach", *5th European Conf. on Speech, Communication and Technology (Eurospeech)*, **5**, Rhodes, Greece, pp 2675-2678 (1997).

16. Chen, S.H. and Hwang, S.H. "An RNN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE Trans. on Speech and Audio Processing*, **6**(3), pp 226-239 (May 1998).

17. Kamyar, T.V., *Prosody of Farsi Language (in Farsi)*, 1st Ed., Jondi-Shahpour University Press, Ahvaz, Iran, pp 15-20, 23-72 (1978).

18. Mohtashami, B., *Complete Farsi Grammar (in Farsi)*, 1st Ed., Eshragh Press, Tehran, pp 558-600 (1991).

19. Samareh, Y., *Farsi Phonetics, Phonemes and Syllable Structure (in Farsi)*, 1st Ed., Nashr-e-Daneshgahi Press, Tehran, pp 102, 127-130 (1985).

20. Sheikhzadeh, H., et al. "Farsi language prosodic structure, research and implementation using a speech synthesizer", *6th European Conf. on Speech, Communication and Technology (Eurospeech)*, **4**, Budapest, Hungary, pp 1647-1650 (1999).

21. Abutalebi, H.R. and Bijankhan, M. "Implementation of a text-to-speech system for Farsi language", *6th Int. Conf. of Spoken Language Processing (ICSLP)*, Beijing, China (2000).

22. Mazdisni, K. "Speech synthesis for Farsi language using HNM", M.Sc. Thesis, Amir Kabir University of Technology, Tehran, Iran (1999).

23. Sadigh, M.R. and Sheikhzadeh, H. "A rule-based approach to Farsi language text-to-phoneme conversion", *6th Int. Conf. of Spoken Language Processing (ICSLP)*, Beijing, China (2000).