FARSI LANGUAGE PROSODIC STRUCTURE, RESEARCH AND IMPLEMENTATION USING A SPEECH SYNTHESIZER

H. Sheikhzadeh, A. Eshkevari, M. Khayatian, R. Sadigh, S. M. Ahadi
EE Dept., AmirKabir Univ. of Tech., Tehran, Iran
emails: hsheikh@cic.aku.ac.ir, sma@cic.aku.ac.ir

ABSTRACT
In this research, we have investigated about prosodic features of Farsi (Persian) language and quantified major stress rules and some intonation rules for speech synthesis purpose. The research is mostly concentrated on pitch variations and then on durational changes. We have implemented the proposed simplified prosodic rules using a Klatt formant synthesizer, specially modified for Farsi phonemes. In order to achieve to a better speech quality, we have exploited different allophonic forms for some consonants, leading to a total of 207 Farsi diphones synthesized by the speech synthesizer. Subjective listening tests show that the addition of the prosodic features drastically increases both the intelligibility and naturalness of the synthesized speech. The synthesizer is software-implemented on a Pentium PC and operates in real-time.

1. INTRODUCTION
The existing literature on Farsi stress rules for even single words is old, insufficient, qualitative and not extensive at all. One cannot find a quantitative or computer implementation of such rules for this language. In this work, using the existing literature, we both quantified and further developed the prosodic rules to make them suitable for a software implementation. We have chosen an engineering approach in this research and have utilized prosodic features to improve the synthesized speech quality.

In Farsi language, primary stress is characterized mostly by pitch and duration [1,2,3]. We have proposed a smooth analytic curve with varying parameters to fit the pitch contours of different classes of words (such as nouns, verbs, adjectives, pronouns,...). We have also quantified and implemented some of the known duration rules of syllables, and finally, have proposed a new set of stress rulers for reading different numbers with correct prosody.

Since the stress rules in Farsi language are mostly syllable-based, we have developed an algorithm to distinguish syllables given the phonetic sequence of a word.

Research on intonation features of Farsi language has been very limited. As a result, we have only managed to propose and implement a few major empirical rules which will be reported in this research.

In order to practically verify the proposed prosodic structure, we have modified a version of Klatt formant synthesizer [4,5] for Farsi speech synthesis. Farsi phoneme characteristics differ from their English similars drastically. There are phonemes in Farsi that do not exist in English and vice versa. For this reason, we had to re-design most of the Klatt parameters for Farsi language. In order to get a better result, we have used different allophonic forms of the phonemes when it was necessary to get a better sound quality, in terms of intelligibility and naturalness.

In the following sections, we first discuss about stress in Farsi and then explain the few intonation patterns occurring in Farsi sentences. In both sections, we will give some examples of natural and synthetic words and sentences. Finally, we discuss about the modifications in Klatt formant synthesizer to adapt it to Farsi language.

2. STRESS PATTERNS IN FARSI
In Farsi language, prosody in general and stress specially, play mostly an oppositional rather than a contrastive role [1,2]. So, there are many Farsi words which are distinguishable only based on their different stress patterns. As for contrastive role of prosody, while it is true that in many Farsi words the last syllable is stressed when uttered in isolation, in naturally spoken sentences there are many exceptions to this rule[1]. Most Farsi linguists believe that stress pattern of Farsi words are mostly governed by pitch rises and then by duration and finally by loudness variations [1,2].

Also some researchers have suggested that duration is highly correlated with pitch, and that stress increases syllable duration [1]. Disputing this suggestion, others believe that it is the syllable phonetic structure that determines its duration [2]. Our observations in this research mostly confirm the second idea and thus we have utilized the rules suggested in literature (or extracted through our investigations) for syllable durations in our implementation. To simplify the matters, we have only implemented the primary stresses in this work.
2.1. Word syllabication algorithm

In Farsi, only three types of syllables occur: CV, CVC and CVCC (C stands for consonant and V for vowel). We proposed and implemented the following simple algorithm for word syllabication.

1. Write the phonetic string in a CV string form.
2. If the CV string does not start with a CV sequence, set the error flag. Go to step 6.
3. Search the CV string for the next occurrence of a CV. If found, go to step 4, else go to step 5.
4. Cluster the sequence between the last CV and the current CV. Compare the cluster with one of the three Farsi syllable types. If a match found, a syllable is distinguished, else set the error flag. Go to step 6.
5. The sequence between current CV and the end of string is the last cluster. Set the termination flag.
6. If termination or error flags are set, end the process, otherwise go to step 3.

2.2. The basic pitch curve

We selected the Hanning function to simulate natural pitch frequency variations, because it resembles the pitch contour in natural utterances. Due to slow variations of pitch frequency, we can consider a syllable pitch contour to follow such a continuous curve, even in unvoiced segments. The function has a general form of

\[ f(x) = \frac{A}{2} \left[ 1 - \cos \left( \frac{2\pi x}{N} \right) \right]. \]

A time-scaled and shifted version of this curve is used to implement the basic prosody rules. Fig. 1 illustrates an example of the curve fit for the stressed CVC syllable /\'vaj/\'. The amount of time shift was experimentally determined for each of the stress and intonation rule cases.

2.3. Major primary stress rules for words

1. In most Farsi words the last syllable is stressed. Names, most adverbs, adjectives, most pronouns, main numbers, and the first part in multi-word combinational verbs are in this group.
2. In some Farsi words the first syllable is stressed. All negative verbs, imperative verbs, simple present verbs and some adverbs and some pronouns are so.
3. In some Farsi words, like simple past verbs, the syllable before the last one is stressed.
4. Some words are never stressed such as most single-word prepositions and conjunctions.
5. We chose to combine the above rules for combinational words in order to avoid adding more rules. For example in the word /dær su\'ræt i ke/\', rules 4,1,4, and 4 are applied sequentially.
6. There are some exceptions which do not follow the above main rules, however, their number is not so high and we used tables to specify their stress rules.

To illustrate some of the rules, consider Fig. 2 to Fig. 5. In each figure, the bar graph shows the naturally occurring pitch, and the solid line curve represents the fitted pitch contour in the speech synthesizer which utilizes the above rules.
2.4. Stress rules in numbers
We first define some terms about Farsi number system.
• Main numbers: The numbers in the set \{0,1,2,\ldots,9,10,20,30,\ldots,90,100,200,300,\ldots,900\}.
• Base numbers: Powers of 1000.
• Group: Starting from the right side, every three digits make a group. The left-most group can have less than three digits.
All numbers in Farsi can be uttered using main and base numbers with a concatenating phone /o/ between consequent segments. For example the number 8192 is uttered as /h\text{æ}\text{s}t \ h\text{e}\text{z}\text{a}r \ o \ s\text{æ}\text{d} \ o \ n\text{æ}\text{s}d \ o \ d\text{o}/ which makes two groups: a main+base+o and a main+o+main+o+main. The stressed syllable in each group is as follows:
1. The last syllable of the first main number in a group of main+o+main+o+main.
2. The last syllable of base.
3. The last syllable of the last main number in a group of main+o+main+o+main.

3. DURATION RULES
The duration of Farsi syllables mostly depends on the vowel part. Vowel duration itself depends on phonetic structure of the syllable [2]. Here is a few examples of the known duration rules:

1. The vowels in CVCC syllables have longer durations than their similars in other syllable types.
2. The vowel duration in CVC and CVCC syllables is affected by the consonant following it. For example, the consonants /h/ and /\text{\textacute{\v{e}}}/ increase the vowel duration while the nasal /n/ decreases the vowel duration strongly.

4. INTONATION
There is not much quantitative information available on Farsi intonation in literature. Thus we extracted a few major rules through careful listenings and observations. We classified Farsi sentences to major groups of: declaratives, simple questions, questions with question words, exclamations, and imperatives. Through experimentation, using the results of a grammatical analysis, we distinguished the major intonation carrying word in each type of sentence. We then used the same smooth curve characterized in Section 2.2 to fit the intonation pitch patterns. Fig. 6 to Fig. 8 demonstrate a few examples of curve-fitting of intonation patterns. For clarity, we have chosen not to apply the stress rules in the figures. Finally, in the speech synthesizer, we added up the effects of pitch variations due to stress and intonation.

Table 1: The phonemes of the Farsi language and examples of their usage.
5. A FARSI KLATT SYNTHESIZER

To practically verify the proposed rules and the effects of their interactions, we had to implement them in a speech synthesizer. We chose a version of Klatt formant synthesizer [4] for this purpose since it is parametric and offers a good platform for prosodic research. While the acoustics and phonetics of Farsi have some similarities with English language, technically, there are major differences both in variety of phones and in the way similar phones are uttered in the two languages. Table 1 shows a list of Farsi phonemes. Starting with limited available literature on acoustic-phonetics of Farsi [2,3], we studied the issue from an engineering point of view, enabling us to adjust Klatt parameters for different Farsi phonemes. In order to achieve a better performance, we used allophonic forms for stop consonants, nasals and some fricatives parts of the CV forms. The C parts in CV diphones had between one to six different parameter sets when necessary. As a result, we got a total of 207 diphones in the speech synthesizer [6,7,8]. We used both careful listening to synthesized speech and comparing the spectrograms of natural and synthetic speech in our Klatt parameter optimizations.

6. CONCLUSION

In this research we chose an engineering approach to practically verify the not well investigated scientific research done on Farsi prosodic structure. Our specific application has been speech synthesis. While this work is an initial step and be no means is complete yet, it offers a solid platform to subjectively accept or reject the sometimes conflicting theorems about Farsi prosody. In this work, we have only considered the primary stress, ignoring weaker stresses, which of course do exist in Farsi and are of some importance. Also, as many linguists believe, we associated primary stress with only pitch variations. More scientific research has to be done to be able to clearly associate (or disassociate) stress with other factors such as duration and loudness changes. Also, the duration rules proposed in the literature and implemented in this work are not perfect and complete. Finally, it is encouraging that by implementation of this over-simplified prosody, we could increase both the intelligibility and naturalness of the synthetic speech, as judged by native Farsi speakers. Our future work is aimed towards scientific research together with practical usage of prosodic models of stress and intonation. We have already developed speech synthesizers with higher speech qualities based on MBR-PSOLA and HNM methods and are in the process of utilizing the prosody models in such synthesizers.

Acknowledgment: This work was supported by National Research Council of Iran.

7. REFERENCES