ABSTRACT

Prosody is a suprasegmental feature of speech that has an undeniable role in human speech perception and generation. However, employing of prosodic features in CSR process mostly is difficult and we must not expect huge accuracy progress by using them. In this way, the main problem arises from high dependency of prosodic patterns to factors like speakers, psychological state of speakers and superposition effects of higher-level prosodic patterns on lower level of them. In our approach, the selected microprosodic feature case is the lexical word stress pattern and relative stresses of cross-word syllables. We aim to verify if we succeed to modify speech recognition process. We employed a proper neural network approach to the word and cross-word stress recognition task. Then we incorporated these features into a spontaneous Farsi speech recognition system called SHENAVA-1. We found 1.3% better word accuracy.

1. INTRODUCTION

Prosodic features perform an undeniable role in the human process of speech perception. These features are used widely in different stages of human recognition process, from the phoneme recognition stage to final stages that help understanding syntactic and semantic information underlay the utterances. Have in mind that prosodic features are mostly context and speaker dependent make us to result recognizing these features is difficult and usually the accuracy rate is not very high. In addition we can see a lot of variation in these features by changing the psychological state of speakers. In this manner, we must not expect a high recognition rate for such a high variable nature feature.

Anyhow, in the first phase of prosodic feature recognition, we need precise studying of a large amount of speech data to detect prosodic patterns of interest. Then we must develop proper models for recognition of different prosodic patterns. After all, by applying these models to the test data, we get an accuracy rate. If this rate were enough high, we could be able to use effectively the recognized features in speech recognition process. That is why we see a less serious advancement to handle prosodic features in current CSR systems. We believe that the main problem in using these features in CSR process arises from low confidence level of recognition them.

In this paper, the noticed prosodic feature is syllable stress. When we speak about the syllable stress, we are concerned to three major factors: pitch, energy and duration. In spite of this fact that the association rate between syllable stress and each of these factors changes by language, there is a strong accordance between them. There are many works that compose mentioned microprosodic features in CSR process in a direct or indirect manner. We believe a high correlation between approaches that employ syllable stress indirectly by using each of the three mentioned related features or directly as syllable stress in CSR process. If we extend the circle of our thoughts, we can consider many references related to the matter of discuss. There are many attempts at using syllable stress and related microprosodic features to modify speech signal front-end or to improve acoustic modeling of CSR systems [1,2,3,4,5]. Approaches within Straightforward application of word stress are fewer. We can see works on the recognition of stress patterns to reduce word candidates or to disambiguate stress- minimal pairs [6,7]. Another approach is to develop a word boundary detection methodology to enhance word recognition accuracy in continuous speech recognition [8,9]. A recent work is highly relevant to this paper, tests the approach of scoring the lexical stress patterns of recognizer hypotheses to improve automatic speech recognition performance [10]. In this paper we intent to present our syllable stress recognition approach in Farsi and applying it to a CSR system. By scoring the syllable stress pattern of hypothesis that are contained in the lexicon, we attempt to employ this extra in two distinct phase of recognition process, to improve overall accuracy of CSR system.

Next sections are arranged as follows: in section 2, we will explain the outline of SHENAVA-1 [11], a Farsi spontaneous speech recognition system, that is our approach test domain. Then we will present the FARSDAT speech database [12]. We have got all needed train and test data of word stress models from aforesaid database. Section 4 is allocated to our word stress modeling approach. In section 5, we will explain, how we succeed to use effectively word stress and cross-word syllable stress in CSR process. Sections 6, 7 and 8 are assigned to experimental results, discussion and references respectively.

2. "SHENAVA-1" SYSTEM FRAMEWORK

SHENAVA-1 is the first version of a Farsi spontaneous speech recognition system that has been developed in RCISP research center [11]. We have performed all the efficiency test of prosodic features application in CSR process, in the SHENAVA-1 domain. Fig. 1 shows the Block diagram of the system.
The system consists of three main parts. The first part transforms the speech signal into an acoustic phonetic representation. The second part makes use of phonetic sequence to perform syllabification and then selection of the best word candidates using searches in a vocabulary that includes the different pronunciation of each containing word. The third part determines a number of the best possible phrases via using an application of a semi-viterbi algorithm to the word lattice output of vocabulary search. In addition this part derives the best phrase by N-best rescoring application to the speech signal. The dashed parts in Fig. 1 are augmented to "SHENAVA-1" to perform the approach of this paper by using syllable stress in recognition process. These parts will be explained in section 5.

3. THE FARSDAT CORPUS

The FARSDAT corpus produced with acoustic-phonetic research goal, for ASR application projects [12]. It consists of 386 sentences read aloud by 300 Farsi native speakers. Each speaker read twenty sentences in two sessions. FARSDAT S/N is 32 db. All 6000 utterances are manually segmented and labeled by IPA characters. FARSDAT provides a vast amount of computer files of words, syllables, phone sequences and allophones. Speech files can be accessed using different fields pertaining to age, sex, dialect and educational level of speaker. All the data needed to this research is taken from FARSDAT database but it is a phonetically labeled corpus and we need some extra stress labels to develop our approach. For this reason, we have created manually a distinct text file that contains FARSDAT words and their stress patterns. In this manner, when we take the speech signal of a FARSDAT word, we can extract its stress pattern from this file.

4. WORD STRESS MODELING

Neural networks are used to model word stress location as a microprosodic feature. Syllable is the basic prosodic unit of the modeling process. In the proposed model, other than pure prosodic parameters like pitch and energy, we exploited segmental information like phones of word syllables. Fig. 2 shows the outline of our word stress modeling approach.

![Figure 1: Block diagram of the SHENAVA-1](image-url)

Some detail aspects of this approach are specifically for Farsi language but the main framework doesn’t differ by language.

4.1. Feature extraction

In this section we will explain how we extract the needed features for word stress modeling.

4.1.1. Syllable pitch contour extraction

We extract pitch contour by a modified method of cepstrum parameters. First, the speech signal is filtered by lowpass and highpass filters respectively. The cutoff frequencies of these filters are over and below than the pitch frequency range. By evaluating the signal amplitude, voiced and unvoiced parts of speech signal are being segregated. Then we apply cepstrum analysis to voiced segments to calculate pitch frequency. We usually see some pitch contour overshoots, especially in the adjacent of the syllable boundaries. To delete these overshoots, we impose a smoothing algorithm to extracted pitch contour.

4.1.2. Syllable pitch contour classifying

Properly selecting classes is very critical and must be done upon the FARSI speech specifications. If we choose a variety of classes less than necessity, we will have an improperly data reduction after classification and if we choose classes more than what exactly desired, stressed syllable recognition algorithms may be confused. We studied carefully the shape of Farsi syllable pitch variation and we found seven distinct shape classes: Flat, Slightly Rising, Sharply Rising, Slightly Falling, Sharply Falling and Flat-falling.

Syllable pitch contour overall shape is coded using a MLP neural network. To train this N.N. we used a training data set, which is consisted of about 150 syllable pitch contours (selected from "FARSDAT" database [12]) and classified by hand. Syllables pitch contour differ in length and pitch amplitude and must be normalized prior to feeding into N.N.

![Figure 2: word stress modeling outline](image-url)

Table 1: number of neurons of the MLP classifier layers

<table>
<thead>
<tr>
<th>Layer</th>
<th>Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input layer</td>
<td>10</td>
</tr>
<tr>
<td>Hidden layer 1</td>
<td>20</td>
</tr>
<tr>
<td>Hidden layer 2</td>
<td>20</td>
</tr>
<tr>
<td>Output layer</td>
<td>3</td>
</tr>
</tbody>
</table>

Input layer neurons 10
Hidden layer 1 neurons 20
Hidden layer 2 neurons 20
Output layer neurons 3

Table 1: number of neurons of the MLP classifier layers
Normalization accomplished by shifting the average value of the syllable pitch contour to zero and linearly warping the syllable pitch contour to a fixed length.

Back propagation algorithm is used to train the MLP. In the recognition phase we feed a normalized syllable pitch contour to MLP input. The MLP output determines the class of the syllable.

4.1.3. Average gradient of syllable pitch contour
To estimate the average gradient of syllable pitch contour, we used the method of a polynomial fitness of degree one to syll. pitch contour. The coefficients of the polynomial are found based on a least square methodology. The gradient of this line is evaluated as the average gradient of syll. pitch contour.

4.2. Word stress pattern recognizer
MLP neural network is adapted to model word stress. We didn’t use HMM for this purpose, as is reported in previous work [9], because of two essential reasons: First, The prosodic unit for this kind of modeling is syllable and the model input has a fixed length for every N-syllable word. In such modeling situation we must cancel HMM state transitions that are not between adjacent states and we can’t exploit time warping capability of HMM model. Since neural network is more powerful than HMM in modeling static patterns, it is reasonable to use neural network instead of HMM. We have verified above reasoning by making experiment with both of these tools in this task. We found that neural network performs better recognition rate by more than 8% that is a very noticeable difference.

we chose separate networks to model the word stress within multi-syllable words. In the current model, each syllable is represented by four pure prosodic parameters and a syllable phonetic combination code. Pure prosodic parameters are: pitch pattern code, pitch average differences of adjacent syllables, pitch trajectory gradient and vowel normalized energy.

To explain the necessity of using the phonetic content for syllable representation, we must mention that legal phonemic pattern of Farsi syllable are three: CV (consonant-vowel), CVCC and CCVC. If the onset consonant belongs to unvoiced plosive group ($/p/, /q/$) or unvoiced consonants group ($/s/i/ and $/s/h/$), we must expect a large increase of pitch level at the beginning of syllable pitch contour and an intense variation in normal syllable pitch trajectory. Also vowels in identical conditions have different intrinsic pitch. $/A/$ and $/a/$ of Farsi vowels have a low, $/o/$ and $/e/$ have a moderate and $/i/$ and $/u/$ have a high intrinsic pitch. For this severe effect of the phones in the shape and gradient of syllable pitch trajectory, we considered the class code of the syllable phone combination into the syllable representation. We have chosen five distinct class in this field based upon onset consonant and vowel combinations of syllables. Table 2 shows these classes.

The final used syllable representation is consisted of four parameters: Average gradient of syllable pitch contour, pitch contour shape code, pitch averages difference of two adjacent syllables, phone combination class code, syllable vowel energy that is normalized to vowel duration. We have exploited three distinct MLP neural networks for modeling stressed syllable location of N-syllable words that are presented in table 3.

<table>
<thead>
<tr>
<th>MLP layers</th>
<th>2-syl. words</th>
<th>3-syl. words</th>
<th>4-syl. words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>11</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>Hidden 1</td>
<td>22</td>
<td>32</td>
<td>42</td>
</tr>
<tr>
<td>Hidden 2</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Output</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: neurons number of N-syl. word MLP model

The stressed syllable of the Farsi 4-syllable words is usually third or fourth syllable and rarely first syllable of the word. Corresponding to this matter we have chose three output neurons for 4-syllable words MLP model. The MLP’s train and test sets specifications are inserted in table 4.

<table>
<thead>
<tr>
<th>Training set words</th>
<th>Test set words</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-syllable words</td>
<td>800</td>
</tr>
<tr>
<td>3-syllable words</td>
<td>700</td>
</tr>
<tr>
<td>4-syllable words</td>
<td>250</td>
</tr>
</tbody>
</table>

Table 4: MLP’s training and test sets specifications

Both training and test sets are consisted of words uttered by more than 30 male and female speakers speaker in continuous speech manner. The data source is FARSDAT.

4.3. Word stress recognition result
The experimental results of word stress recognition using MLP models are shown in table 5.

<table>
<thead>
<tr>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-syllable word</td>
</tr>
<tr>
<td>3-syllable word</td>
</tr>
<tr>
<td>4-syllable word</td>
</tr>
</tbody>
</table>

Table 5: word stress recognition accuracy using MLP

There are two main remarks to discuss here. First of all, we consider the Accuracy rate of Farsi bisyllabic word stress recognition is less than the trisyllabic or quadssyllabic words. As we are faced to the words uttered in continuous speech manner, it seems due to the influence of the sentence intonation within word syllables pitch trajectory. thus, as the number of word syllables increases, redundant parts of word stress pattern increase too and comparing and detection of word syllables stress becomes easier.

Second, Since Farsi like most Indo-European languages is a pitch-accent language, it seems to us that the pitch dependent features we used in the model, are mostly effective to detect stressed syllable. In addition, experiment showed that vowel relative energy, but not duration, is an effective feature that helps to detect stressed syllable. Among the pitch dependent features, the average gradient of syllable pitch contour and pitch
average difference of adjacent syllables have the most important role in stressed syllable recognition. By using only these two features we achieved about 90% of reported accuracy level.

5. EMPLOYING WORD STRESS IN CSR

We have shared word and cross-word stress in two different phases of SHENAVA recognition process. Fig. 1 shows the modification to the original SHENAVA-1 system to handle using word stress in CSR process.

In the first phase, applying the word stress is carried out in vocabulary search. The output of vocabulary search block is a lattice of words and their related scores. The scores of searched words are calculated by evaluating acoustic similarities. In our approach we modify the word scores via sharing word stress parameter. Although, we didn’t impose a constant sharing rate but a rate variable with confidence level of word stress model recognition. For example if the output of N.N. model in recognition phase is more than .85, we will apply higher sharing rate.

In the second phase, applying the cross-word stress is carried out in semi-viterbi search. The output of this block is a group of sentences or phrases that earned higher scores (resulted through their contained word scores). We have modified this part by checking up the relative stresses of the intersection word syllables. We share this parameter only at two specified conditions: one of the cross-word syllables must be stressed and the other must be unstressed. We can verify these two condition by using 2-syllable word stress models and we don’t need extra special model. Like the first phase, the sharing rate of this parameter depends on N.N. output confidence level.

We don’t consider monosyllable words stress in this approach. As we have discussed in section 4.3, higher prosodic patterns like intonation, affect word stress features. In the case of monosyllable words these effects cause a high stress recognition error rate that we couldn’t use it effectively in CSR process.

We used a test set of 200 sentences uttered by 10 male speakers to evaluate the performance of the approach. The majority of the speakers and the applied words of test set are not common to FARSDAT corpus.

6. EXPERIMENTAL RESULTS

Table 6 indicates the experimental results of augmenting the mentioned approach to SHENAVA system.

<table>
<thead>
<tr>
<th>SHENAVA-1 situation</th>
<th>Word accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>63.5%</td>
</tr>
<tr>
<td>+Word stress</td>
<td>64.4%</td>
</tr>
<tr>
<td>+Word and cross-word stress</td>
<td>64.8%</td>
</tr>
</tbody>
</table>

Table 6: SHENAVA word accuracy rate

7. DISCUSSION

In this paper, we presented our approach to word stress modeling and sharing it to spontaneous speech recognition process. We trained distinct models for 2-syllable, 3-syllable and 4-syllable words stress. We got the results of 83%, 89% and 90% words stress recognition accuracy rates for 2-syllable, 3-syllable and 4-syllable words, respectively. We verified that syllable is a proper unit for Farsi (and similar languages) word stress modeling and for this special task, using MLP is more efficient than HMM.

More than, we contributed a proper framework for interposing word stress in two phase of a spontaneous speech recognition system and we achieved 1.3% better word recognition accuracy (in our primary attempt) that is noticeable. We believe that we can access better results in the future, by making efforts to modify the approach of sharing word stress in recognition process. However, we must observe, because of the nature of prosodic features, applying prosodic approaches to speech recognition have a limited performance and we mustn’t expect more than achievable rate.

8. REFERENCES