6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

IMPLEMENTATION OF A TEXT-TO-SPEECH SYSTEM FOR FARSI LANGUAGE

Hamid Reza Abutalebi (1,3) , *Mahmood Bijankhan* (2,3)

(1) Dept. of Electrical Eng., Amirkabir University of Technology
(2) Dept. of Linguistics, University of Tehran
(3) Research Center of Intelligent Signal Processing (RCISP)
Tehran , Iran

Email: *a7723906@cic.aku.ac.ir*

10.21437/ICSLP.2000-164

ABSTRACT

In this research, a Text-To-Speech system for Farsi language has been implemented. The proposed synthesizer concatenates Farsi syllables in a TD-PSOLA manner. This paper is mainly concentrated on investigation about pitch variations in Farsi sentences and presentation of some novel rules for modeling these variations. Based on the location of stressed syllable, we obtain a primary pitch curve for each word. Using prosodic grouping and sentence type effects, the final pitch contour can be determined. High intelligibility and acceptable naturalness of the synthesized speech have been confirmed by subjective listening tests.

1. INTRODUCTION

Farsi (Persian) language is comparatively little researched. In recent years, a few of native researchers have begun to work on various branches of Farsi language processing (Recognition, Synthesis, etc.).

In this paper, we introduce the first Text-To-Speech (TTS) system for this language, which is constructed based on concatenation method of speech synthesis. This synthesizer is an implementation of Time Domain Pitch Synchronous OverLap-Add (TD-PSOLA) methodology [1]. After selection of *Syllable* as the basic unit of concatenative speech synthesizer, a syllable database for Farsi language is gathered. According to the input text, we choose proper syllables from database and concatenate them to obtain the primary output. To achieve a natural output speech, we need to correct prosodic features of syllable-connected output. In Farsi language, pitch frequency variations have the most significant effect on prosodic structure of speech [2]. Therefore, in the proposed system, we concentrate on specifying a proper pitch frequency contour for every sentence of input text.

The next section, presents a general outlook of the TTS system; both text analysis and speech synthesis parts of the system will be briefly explained. In Sec. 3, we have described the proposed algorithm for determination of pitch frequency contour. Finally, intelligibility and quality of the resulted speech will be reported in Sec. 4.

2. SYSTEM OUTLOOK

In a general view, any TTS system consists of two important blocks: (1) Text analysis, and (2) Speech synthesis. In the following, the structure and function of these blocks will be discussed (according to the implemented system in this work).

2.1. Text Analysis

Giving Farsi text from input (keyboard), A lexicon of 5000 most frequent words of Farsi and some primary rules help us determine a phonemic string. Additionally, in the lexicon, there are supplementary information about each word, which will be used to specify a prosodic pattern [3]. (This will be further discussed in Sec. 3.)

2.2. Speech Synthesis

TD-PSOLA framework [1] is implemented as the speech synthesis block of the system. TD-PSOLA is a very popular method used in many concatenative synthesizers for different languages.

The first step in design of a concatenative speech synthesizer is the determination of basic unit. Various types of basic unit have been used in synthesizers of different languages. Based on the following reasons, *Syllable* has been selected as basic unit in our speech synthesizer [3]:

- Via the examination of some typical speech spectra, it is observed that the coarticulation effect is minimized in syllable boundaries; since the handling of coarticulation phenomenon is the most difficult problem in speech synthesis, selection of syllable as the basic unit can result in some simplicity and better output quality [4].

- The second reason is concerned with limited syllable structure of Farsi language; the syllable structure variety of Farsi is limited to only three types: *CV*, *CVC* and *CVCC*. Among all of syllables, which can be constructed from 23 consonants and 6 vowels of Farsi in the above three syllable types (structures), only about 4000 syllables are used in Farsi words and the others are not used [4].

- Prosodic patterns are mainly based on the syllables of the input text. In other words, type of the syllable and its stressed/unstressed situation are the significant parameters in determination of prosodic variations [2].
- Since the syllable, on the average, is relatively the largest linguistic unit in size among other units, its utilization as the basic unit of concatenation will reduce the necessary smoothing process.
- Because of minimum effect of coarticulation in syllable boundaries, the segmentation process of database will be comparatively simple.

Therefore, we made a database of all Farsi syllables. In addition to speech waveform, the laryngogram of the speaker has been synchronously recorded [3]. This signal (laryngogram) is used for determination of pitch onset times (pitch marks or epochs) in database preparing stage. These onsets are necessary in pitch synchronous processing. After labeling the speech material (in syllabic level) and completing the database, a method based on *Difference Function* is used to specify pitch onset times from laryngogram [5].

TD-PSOLA synthesizer selects the entries of syllable database and concatenates them to make a primary output speech corresponding to the input text. Then, the synthesizer changes prosodic features (pitch frequency, duration and intensity) of that primary speech in a Pitch Synchronous and OverLap-Add manner in Time Domain and makes the final output speech. The proper values of pitch, duration and intensity are specified in a manner which will be explained in Sec. 3.

3. PITCH CONTOUR GENERATION

As mentioned above, in Farsi language, pitch frequency is the most significant prosodic feature. So, in the following, we will describe implemented algorithm for F0 contour generation in detail. Furthermore, in the TTS system, some primary rules for determination of two other prosodic features (duration and intensity) have been used [3].

The pitch frequency contour of the input text is resulted by applying the effects of these three parameters: pitch contour of the word, prosodic grouping and the sentence type.

3.1. Pitch Contour of the Word

To describe pitch problem, it is necessary to glimpse stress patterns in Farsi. In this language, primary stress pattern has a comparatively regular structure and one can determine *Strong* and *Weak* (or *Stressed* and *Unstressed*) syllables based on the word type. In this research, we have used some rules that specify if any syllable is strong or weak according to its place in the word and the word type. Briefly, in any polysyllable noun or adjective and most of the adverbs, the last syllable is strong; In simple past tense verbs, the last syllable of the verb root is strong; In progressive past tense verbs, simple present tense verbs and negative verbs, the first syllable has the primary stress.

From signal processing view, the strong syllables frequently have higher pitch frequencies, longer durations and greater intensities [2,4].

As a matter of fact, in this stage, we classify syllables of input text in five different groups. This classification is based on Strong/Weak situation and location of syllable in word and also on word type (Noun, Adjective, Preposition, Conjunction, Verb and the type of verb, etc.). Each element of the lexicon has some supplementary information, which used for assigning the syllable types.

After this classification, one of the predefined pitch frequency patterns will be assigned to each syllable. These pitch frequency patterns are illustrated in Fig. 1 (cited from [6]). Among these patterns, Type 3 corresponds to strong syllables and the others display various types of weak syllables. In Table 1, the most common occurrence cases of each pitch type have been explained.

For better description of the subject, actual pitch contours of four Farsi words have been shown in Figs. 2 to 5. (In this paper, phonetic symbol /A/ is used as the back-open-rounded-long Farsi vowel. Other symbols are similar to their English equivalents.) In Figs. 2,3 and 4, the first syllable of each word approximately has a Type 1 pitch pattern. Last syllable of nouns /?irAn/ and /?AzAdi/ and third syllable of /barAdare/ are the strong syllables and their pitch contours are similar to Type 3. Figs. 2 and 3 show that the second syllables of the nouns have Type 2 pitch pattern. Besides, the genitive morpheme, /e/, in /barAdare/ has the shape of a Type 5 pattern in pitch contour. At last, Fig. 5 illustrates that the sign of direct object, /rA/, follows Type 4 pattern in its pitch curve.

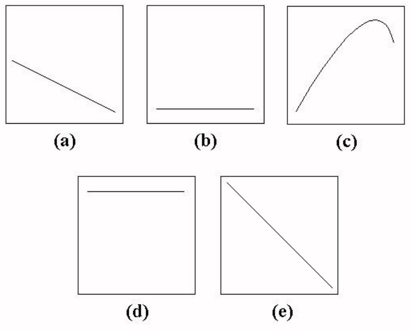

Figure 1: Pitch frequency patterns: (a) Type 1: Falling, (b) Type 2: Flat-Low, (c) Type 3: Rising, (d) Type 4: Flat-High, (e) Type 5: Quickly Falling**.**

Table 1: Occurrence cases of different pitch patterns

	Occurrence Cases
Type 1	First syllable of polysyllable nouns
Type 2	Weak syllables of polysyllable words (except first syllable of nouns)
Type 3	Strong syllables
Type 4	Monosyllable words, The sign of direct object: /rA/
Type 5	Genitive morpheme: /e/, Prepositions and conjunctions

By concatenating the above patterns according to the syllables of the input text, we can obtain microscopic structure of pitch contour. Then, prosodic grouping of the words and sentence type should be considered for specifying an intonational envelope (macroscopic structure) for pitch contour.

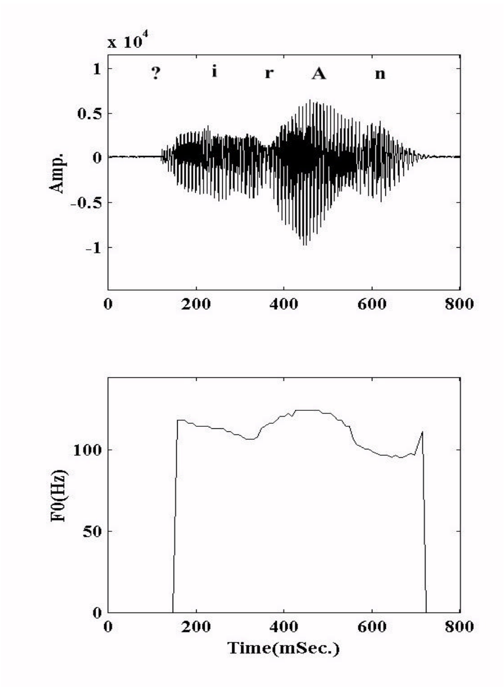

Figure 2: Waveform (above) and pitch contour (below) of the word /?irAn/

3.2. Prosodic Grouping

This block of the system uses some boundary markers, which determine begin/end of the groups. These markers include grammatical words (prepositions, conjunctions, etc.), verbs, the sign of direct object and punctuation marks. During each group, microscopic structure of pitch curve (resulted from the previous part) will be modulated by a normalized rising line [3].

3.3. Sentence Type

Speech intonation is very related to the corresponding sentence type. Our observations show that in the natural (not-synthesized) speech, pitch frequency contour falls at the ending part of declarative sentences; also, it have been shown that the pitch curve rises at the ending part of simple question sentences (with no question word in the sentence). By using these simplified intonation rules, we can obtain the final pitch frequency contour from the results of two previous parts [3].

4. DISCUSSION AND CONCLUSION

In this paper, a TTS system for Farsi language was introduced. For some aforementioned reasons, syllable has been selected as the basic unit of concatenative synthesizer. This selection has guaranteed high intelligibility of output speech.

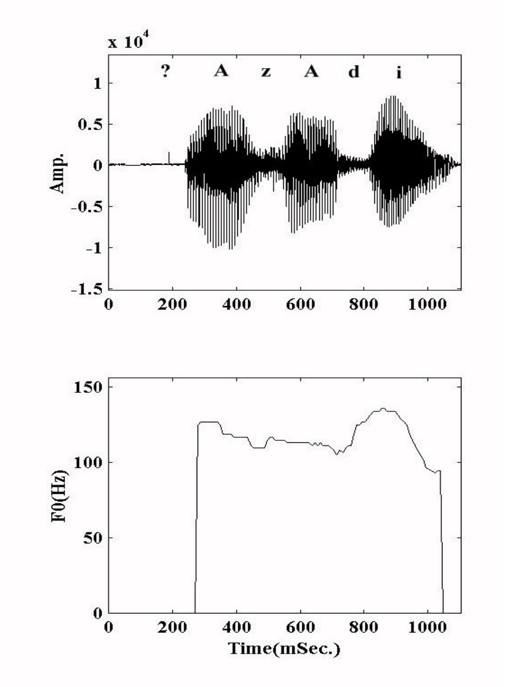

Figure 3: Waveform (above) and pitch contour (below) of the word /?AzAdi/

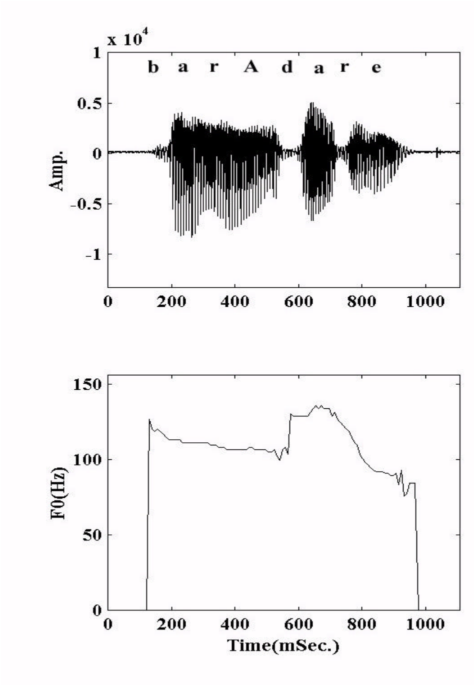

Figure 4: Waveform (above) and pitch contour (below) of the word /barAdare/

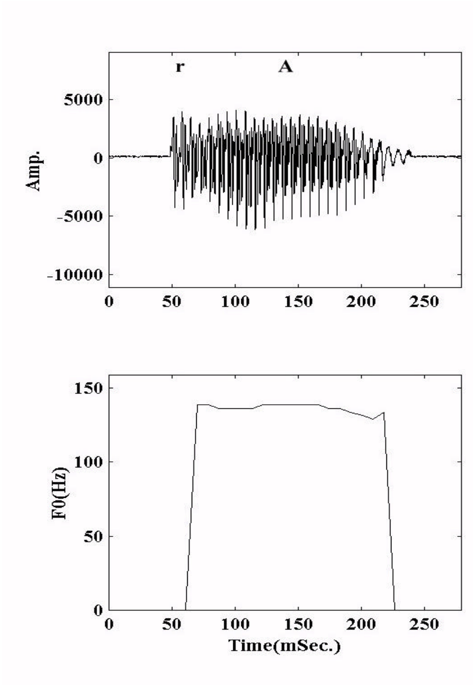

Figure 5: Waveform (above) and pitch contour (below) of the word /rA/

Considering significance of the pitch in Farsi prosody, we concentrated on making a proper pitch curve for the input text. The proposed algorithm generates the pitch contour in three stages: word, prosodic group and sentence. In spite of its simplicity, the implemented rules for pitch curve generation, can drastically improve the output quality and naturalness.

Furthermore, our methods for duration and intensity determination, were over-simplified. Better quality of output speech will be achieved by using some more linguistic rules for specifying duration and intensity.

Informal subjective listening tests showed that the proposed system could convert Farsi input text to a synthetic speech with high intelligibility and acceptable naturalness. This system is implemented on a Pentium-II PC and converts the typewritten texts into speech in real-time.

5. REFERENCES

1. Moulines, E., and Charpentier, F., " Pitch Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones," *Speech Communication*, Vol. 9, No. 5, pp. 453-467, 1990.

2. Vahidian Kamyar, T., *Prosody of Farsi Language*, Iran, Ahwaz, Jondi-Shahpour University Press, 1978. (in Farsi)

3. Abutalebi, H.R., " Studying on and Implementing a Proper Farsi Speech Synthesizer," *M.Sc. thesis*, Dept. of Elec. Eng., Sharif Univ. of Tech., Tehran, Iran, 1998.

4. Samareh, Y., *Phonetics of Farsi Language*, Iran, Tehran, Academic Press Center, 1995. (in Farsi)

5. Abutalebi, H.R., and Tebyani, M., " Pitch Onset Detection from the Laryngogram," in *Proc. of the 4th Annual Inter. CSI Computer Conf.*, CSICC'98, Tehan, Iran, pp. 182-187, Jan. 1999.

6. Almasganj, F., and Hashemi Golpayegani, S.M.R., " Lexical Segmentation of Farsi Sentences Using Prosodic Features of Continuous Speech Signal," in *Proc. of 3rd Electronics Conf.*, TEC'95, Shiraz, Iran, Vol. 1, pp. 167-175, 1995.