VOWEL INSERTION GRAMMAR

MUHAMMAD KHURRAM RIAZ, MUHAMMAD MUSTAFA RAFIQUE, SYED RAZA SHAHID

ABSTRACT

In Urdu language the position of vowels within the syllable is highly considerable issue. This paper contains vowel insertion grammar for Urdu language. The grammar is presented in the form of an automaton. The presented automaton was tested successfully on many words of Urdu language. The tested words were supposed to have all the diacritics correctly marked on them. The following vowel insertion grammar is extremely useful for building speech synthesis for Urdu language.

1. INTRODUCTION

In Urdu text, diacritics are very important as they indicate the existence of vowels between two consonants. In an Urdu text-to-speech converter, it is extremely important to judge where to produce a vocalic sound, as there is no particular letter in Urdu language, which shows the existence of a vowel in Urdu text. Vowel insertion grammar serves the purpose in determining the existence of vowels due to the presence of diacritics in Urdu text.

2. LITERATURE REVIEW

Seventeen vowels have been identified during the analysis of Urdu vocalic sounds (Urdu Consonantal and Vocalic Sounds, 2002). According to Kachru (1990), there are seven long oral vowels, and three short oral vowels. Khan (1997) also agrees with the long and short vowel distribution of Kachru (1990).

Hussain (1997, p. 153) suggests that Urdu vowels are written with the help of three letters (phonemes) and three diacritics—namely, letters are 'ا', 'و' and 'ی', and diacritics are paish (ُ), zair (ِ) and zabar (َ). The long vowels are written using one of the three letters (phonemes) mentioned above with one of the diacritics on the preceding consonant. The short vowels are written with only a diacritic on the preceding consonant. However, these diacritics are not necessarily written in Urdu script (they can be ignored of the writer thinks no ambiguity can occur).

There are some languages that expand their vowel inventory by opening the velar port during some of their vowels' utterance.

Such a language uses nasalized vowels to distinguish within pairs of words that are otherwise same. It uses nasal vowels that are produced with shapes of oral configuration similar to the oral shapes of some of its non-nasal vowels (Picket, 1999, p. 71). Urdu is also one such language, which enhances its vowel inventory by adding nasal vowels in it. Each long vowel in Urdu language has its nasalized version, but there is no existence of any nasalized short vowel in Urdu language. Therefore, overall, there are seventeen vowels (including oral and nasal vowels) in Urdu language.

3. RESULTS

Vowel insertion grammar is represented in the form of an automaton shown in Figure 1. Since the automaton is represented in the form of a *Mealay* machine, so there is a corresponding output associated with each input in transition from one state to another. Circles in the figure represent states. Start and final states are mentioned explicitly in the given automaton. Each intermediate state remembers all the previous inputs, which lead the system to that particular state.

The grammar is also explained with examples in Table 1. First column in Table 1 shows the pattern for each of the possible input. The second column shows the next input character for each possible input. First and second column form the composite key for Table 1. Output vowel is shown in the column Output. Examples for each case are shown in IPA (International Phonetics Alphabet) symbols.

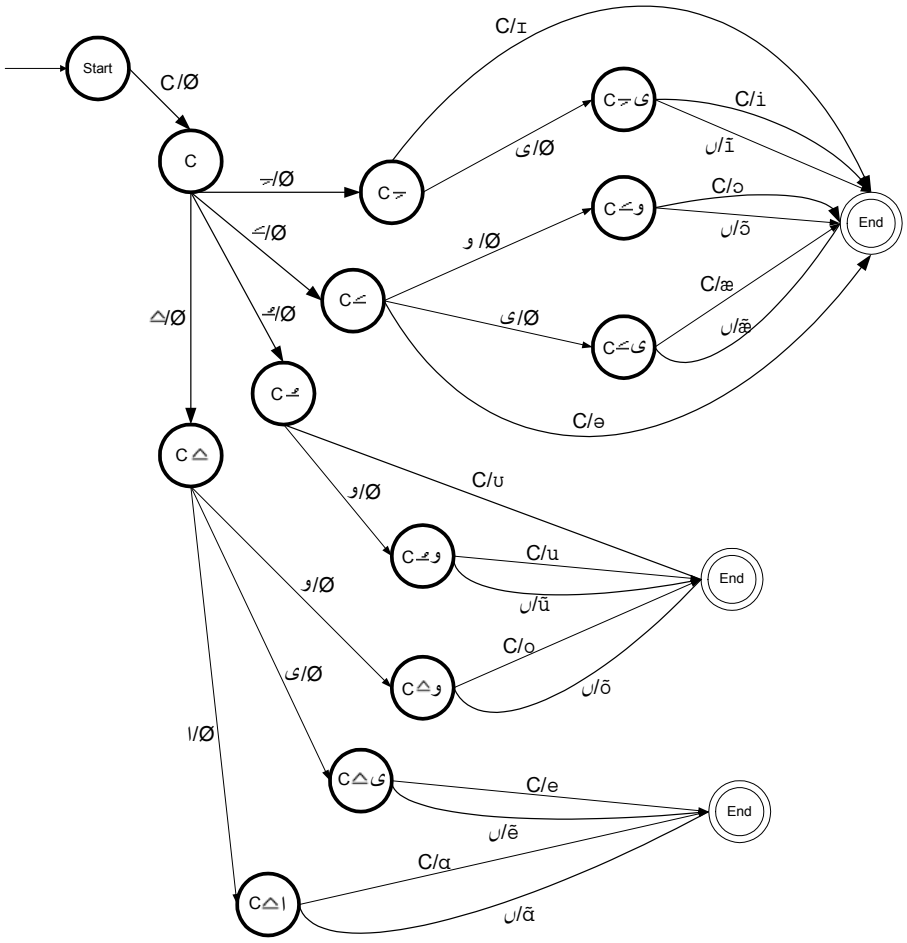

FIGURE 1 Automaton for vowel insertion grammar

4. DISCUSSION

The existence of vowels in Urdu language is partially similar to English language. But unlike English language, vowels in Urdu language are very much predictable because of the presence of punctuation marks (diacritics) in Urdu script. There are three punctuation marks in Urdu that help in the identification of vowels. These punctuation marks are paish (ـُ), zair (ـِ) and zabar (ـٰ).

The results reveal that long vowels in Urdu language are inserted if a letter has any of the diacritics paish (ـُ), zair (ـِ) and zabar (ـٰ), followed by any of these three Urdu letters 'ا', 'و' or 'ی'. If any other letter apart from the above-mentioned letters follows these diacritics, then a short vowel is inserted. Similarly if any of the Urdu letter 'ا', 'و' or 'ی' has a diacritic preceding and letter 'ں' following it, then a nasalized long vowel is inserted.

Apart from the punctuation marks—paish (ـُ), zair (ـِ) and zabar (ـٰ), there is another punctuation mark jazm ('ـْ') that causes no insertion of vowel in Urdu language if the letter following the jazm ('ـْ') is any other than the letters 'ا', 'و' or 'ی'. So, only long vowels can exist if there is a jazm ('ـْ') on any letter. Jazm ('ـْ') is considered as default punctuation mark if a letter has no punctuation mark on it.

It was also observed that the letter 'ا' can exists in Urdu script only if there is a jazm ('◌ْ') on the previous letter.

TABLE 1 Possible diacritics combination. 'C' indicates any consonent

Pattern	*Next Level*	*Output*	*Example*
C◌ِ	C	ɪ	bɪtʃʰnɑ
C◌ِی	C	i	ʈin, ɑmin
C◌ِی	ں	ĩ	vəhĩ
C◌ُ	C	ə	ʈərʃ, ʈərɑdʒɪm
C◌ُ و	C	ɔ	pɔɖɑ, sɔdɖ
C◌ٗی	C	æ	tʃæn, bæl
C◌ٗی	ں	æ̃	hæ̃
C◌ٗ	C	ʊ	ɖʊlahn, sʊkh
C◌ٗ و	C	u	χun, ʈu
C◌ٗو	ں	ũ	hũ
C ◌ٰی	C	e	ʈevr, bekɑr
C ◌ٰی	ں	ẽ	ɪəhẽ
C ◌ٰ و	C	o	ʈoɽ, sonɑ
C ◌ٰ و	ں	õ	hõ
C ◌ٰا	C	ɑ	ʈalɪ
C ◌ٰا	ں	ɑ̃	hɑ̃

The automaton shown in Figure 1 has been tested on many words and successful results were noted. Since the automaton/grammar requires diacritics to decide between the vowels, so it becomes extremely important to mention all the diacritics to get the correct vowel as output. Generally, in Urdu text these diacritics are ignored and are considered to be understood by default. So, for using the automaton shown in Figure 1, extra effort is required to mention all the diacritics on the input text.

5. REFERENCES

Hussain, S. 1997. "Phonetic correlates of lexical stress in Urdu"; Unpublished Ph.D. dissertation, North Western University, IL, USA.

Kachru, Yamuna 1990. Hindi-Urdu in The Major Languages of South Asia, The Middle East and Africa, edited by Bernard Comrie.

Khan, Mahboob Alam 1997. Urdu ka Soti Nizaam.

Pickett, J. M. 1999. The Acoustics Of Speech Communication Fundamentals, Speech Perception Theory, And Technology. USA

Urdu Consonantal and Vocalic Sounds; Speech Synthesis Group, Center for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences, Lahore.