Letter-to-Sound Conversion for Urdu Text-to-Speech System

Sarmad HUSSAIN
Center for Research in Urdu Language Processing,
National University of Computer and Emerging Sciences
B Block, Faisal Town
Lahore, Pakistan
sarmad.hussain@nu.edu.pk

Abstract

Urdu is spoken by more than 100 million people across a score countries and is the national language of Pakistan (http://www.ethnologue.com). There is a great need for developing a text-to-speech system for Urdu because this population has low literacy rate and therefore speech interface would greatly assist in providing them access to information. One of the significant parts of a text-to-speech system is a natural language processor which takes textual input and converts it into an annotated phonetic string. To enable this, it is necessary to develop models which map textual input onto phonetic content. These models may be very complex for various languages having unpredictable behaviour (e.g. English), but Urdu shows a relatively regular behaviour and thus Urdu pronunciation may be modelled from Urdu text by defining fairly regular rules. These rules have been identified and explained in this paper.

1 Introduction

Text-to-speech synthesis is logically divided into two stages. The first stage takes raw text input, processes it and converts it into precise phonetic string to be spoken, appropriately annotated with prosodic markers (e.g. stress and intonation). The second stage takes this phonetic representation of speech and generates the appropriate digital signal using a particular synthesis technique. These stages may be referred to as Natural Language Processing (NLP) and Speech Synthesis (SS) respectively (e.g. Dutoit 1997, p.14).

For SS, formant based techniques (e.g. Klatt 1980) or diphone based techniques (e.g. Dutoit 1997) are normally employed and are generally script independent (as they are only dependent on temporal and spectral acoustic properties of the language and take input in script-neutral form, e.g. in IPA). However, NLP is very dependent on cultural and linguistic specific usage of script.

NLP may also be divided into further parts. The first component is dedicated to pre-processing, 'cleaning' and normalizing input text. Once the input text is normalized, the second component does phonological processing to generate a more precise phonetic string to be spoken. One of the first tasks in the Phonological Processing Component is to convert the input text into a phonemic string using Letter-to-Sound (LTS) rules. This string is then eventually converted to precise phonetic transcription after application of sound change rules and other annotations, as explained later. This paper overviews Urdu writing system, phonemic inventory, NLP for TTS and gives details of the LTS rules for Urdu (also see Rafique et at. (2001) and Hussain (1997: Appendix A), for introductory work).

2 Urdu Writing System and Phonemic Inventory

Urdu is written in Arabic script in Nastaleeq style using an extended Arabic character set. Nastaleeq is a cursive, context-sensitive and highly complex writing system (Hussain 2003). The character set includes basic and secondary letters, aerab (or diacritical marks), punctuation marks and special symbols (Hussain and Afzal 2001, Afzal and Hussain 2001). Urdu is normally written with only the letters. However, the letters represent just the consonantal content of the string and in some cases (under-specified) vocalic content. The vocalic content can be (optionally) completely specified by using the aerab with the letters. Aerab are normally not written and are assumed to be known by the native speaker, thus making it very hard for a foreigner to read. Certain aerab are also used to specify additional consonants. Urdu letters and aerab are given in Table 1 below.

چ	ج	ث	ٹ	ت	پ	ب	ا
ز	ڑ	ر	ذ	ڈ	د	خ	ح
ع	ظ	ط	ض	ص	ش	س	ژ
ن	م	ل	گ	ک	ق	ف	غ
			ے	ی	ء	ہ	و

ھ	ۃ	ں	آ

ْ	ّ	ً	ٰ	ُ	ِ	َ

Table 1: Urdu basic (top) and secondary (middle) letters and aerab (bottom)

Combination of these characters realizes a rich inventory of 44 consonants, 8 long oral vowels, 7 long nasal vowels, 3 short vowels and numerous diphthongs (e.g. Saleem et al. 2002, Hussain 1997; set of Urdu diphthongs is still under analysis). This phonemic inventory is given in Table 2.

The italicized phonemes, whose existence is still not determined, are not considered any further (see Saleem et al. 2002 for further discussion). Mapping of this phonetic inventory to the characters given in Table 1 is discussed later.

(a)

p	b	p^h	b^h	m	*m*^{h}	
t̪	d̪	t̪^{h}	d̪^{h}	n	*n*^{h}	
ʈ	ɖ	ʈ^h	ɖ^h			
k	g	k^h	g^h	ŋ	*ŋ*^{h}	
tʃ	dʒ	tʃ^h	dʒ^h	q	ʔ	
f	v	s	z			
ʃ	ʒ	x	ɣ	h		
r	*r*^{h}	ɽ	*ɽ*^{h}	j	l	*l*^{h}

(b)

i	e	*ɛ*	æ
u	o	ɔ	ɑ
ɪ	ʊ	ə	
ĩ	ẽ	æ̃	
ũ	õ	ɔ̃	ɑ̃

Table 2: Urdu (a) Consonantal and (b) Vocalic phonemic inventory

3 NLP for Urdu TTS

As discussed earlier, to enable text-to-speech system for any language, a Natural Language Processing component is required. The NLP system may have differing requirement for different languages. However, it always takes raw text input and always outputs precise phonetic transcription for a language. The system can be divided into two parts, Text-Normalization Component and Phonological Processing Component. These components may be further divided. A simplified schematic is shown in Figure 1[1].

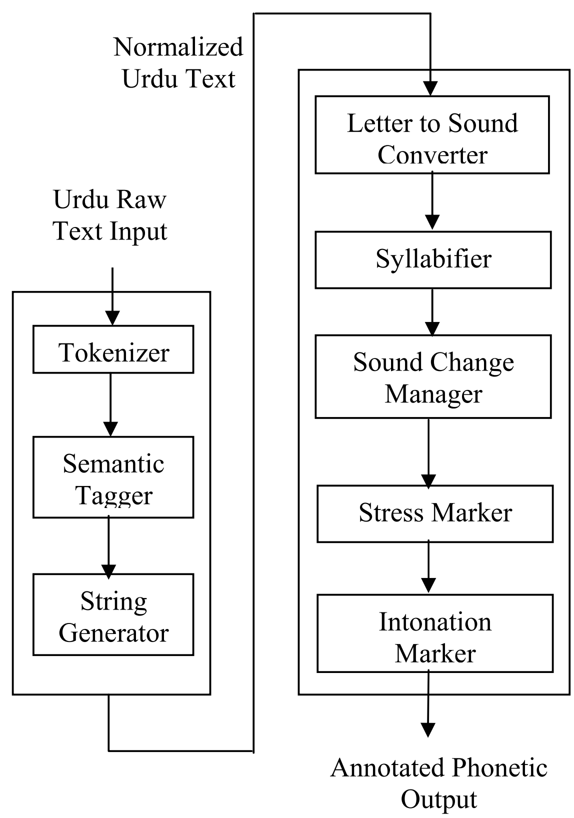

Figure 1: NLP architecture for Urdu TTS system

[1] This diagram is based on the architecture of Urdu Text to Speech system under development at Center for Research in Urdu Language Processing (www.crulp.org).

The Text Normalization component takes a character string as input and converts it into a string of letters. Within it, the Tokenizer uses the punctuation marks and space between words to mark token boundaries which are then stamped as words, punctuation, date, time and other relevant categories by the Semantic Tagger. The String Generator takes any non-letter based input (e.g. a number or a date containing digits) and converts it into a letter string.

After the input is converted into a string comprising only of letters, the Phonological Processing Component generates the corresponding phonetic transcription. This is done through a series of processes. The first process is to use Letter-to-Sound Converter (detailed below) to convert the normalized text input to a phonemic string. This process may also be referred to as grapheme-to-phoneme conversion. This is followed by Syllabifier, which marks syllable boundaries. The intermediate output is then forwarded to a module which applies Urdu sound change rules to generate the corresponding phonetic string. Following these modules, Stress Marker and Intonation Marker modules add stress and intonation to the string being processed. Re-syllabification is also performed after sound change rules are applied, in case phones are epenthesized or deleted and syllable boundaries require re-adjustment. Urdu shows a reasonably regular behavior and most of these tasks can be achieved through rule-based systems (e.g. see Hussain 1997 for stress assignment algorithm). This paper focuses on Letter-to-Sound rules for Urdu, the first in the series of modules in Phonological Processing Component.

4 Urdu Letter to Sound Rules

Urdu shows a very regular mapping from graphemes to phonemes. However, to explain the behavior, the letters need to be further classified into the following categories:

a. Consonantal characters
b. Dual (consonantal and vocalic) behavior characters
c. Vowel modifier character
d. Consonant modifier character
e. Composite (consonantal and vocalic) character

Similarly, the aerab set can also be divided into the following categories:

f. Basic vowel specifier
g. Extended vowel specifier
h. Consonantal gemination specifier
i. Dual (vocalic and consonantal) insertor

Finally, there is a third category which may take shape of an letter and aerab:

j. Vowel-aerab placeholder

The Consonantal characters in (a) above always represent a consonant of Urdu. In Urdu, there is always a single consonant corresponding to a single character of this category, unlike some other languages e.g. English maps "ph" string to phoneme /f/. Most of the Urdu consonantal characters fall into this category. These characters and corresponding consonantal phonemes are given in Table 3 below. A simple mapping rule would generate the phoneme corresponding to these characters.

چ	ج	ث	ٹ	ت	پ	ب
tʃ	dʒ	s	ʈ	t̪	p	b
ڑ	ر	ذ	ڈ	د	خ	ح
ɽ	r	z	ɖ	d̪	x	h
ط	ض	ص	ش	س	ژ	ز
t̪	z	s	ʃ	s	ʒ	z
گ	ک	ق	ف	غ	ع	ظ
g	k	q	f	ɣ	ʔ	z
		ۃ	ہ	ن	م	ل
		t̪	h	n	m	l

Table 3: Consonantal characters and their corresponding phonemes

Three characters of Urdu show dual behavior, i.e. in certain contexts they transform into consonants, but in certain other contexts, they transform into vowels. These characters are Alef (ا), vao (و), and Yay (ی or ے). Alef acts exceptionally in this category and therefore it is discussed separately in (j) below. Vao changes to /v/ and Yay changes to the approximant /j/ when they occur in consonantal positions (in onset or coda of a syllable). However, when they occur as nucleus of a syllable, they form long vowels. As an example, Yay occurs as a consonant when it occurs in the onset of single syllable word یار

(/jar/, "friend") but is a vowel when it occurs word medially in بَیل (/bæl/, "ox"). These characters represent category (b) listed above.

There is only one character in category (c), the letter Noon Ghunna (ں), which does not add any additional sound to the string but only nasalizes the preceding vowel. This letter follows and combines with the category (b) characters (when occurring as vowels) to form the nasal long vowels, e.g. جا (/dʒɑ/, "go") vs. جاں (/dʒɑ̃/, "life"). Catergory (d) is the letter Do-Chashmey Hay (ھ), which combines with all the stops and affricates to form aspirated (breathy or voiceless) consonants but does not add an additional phoneme. It may also combine with nasal stops and approximants to form their aspirated versions, though these sounds are not clearly established phonetically. As an example, adding this character adds aspiration to the phoneme /p/: پل(/pəl/, "moment") vs. پھل (/p^{h}əl/, "fruit"). Finally, there is also a single character in category (e), the Alef Madda (آ). This character is a stylistic way of writing two Alefs and thus represents an Alef in consonantal position (see (j) below) and an Alef in vocalic position, forming /a/ vowel, e.g. اَب (/əb/, "now") vs. آب (/ɑb/, "water").

There are three Basic vowel aerab used in Urdu called Zabar (Arabic Fatha), Zer (Arabic Kasra) and Pesh (Arabic Damma). In addition, absence of these aerab also define certain vowels and thus this absence is referred to as Null aerab. They combine with characters to form vowels according to the following principles:

(i) *Short vowels*, when they occur with category (a) and (b) consonants not followed by category (b) letters.
(ii) *Long vowels*, when they occur with category (a) and (b) consonants followed and combined by category (b) characters.
(iii) *Long nasal vowels*, when they combine with category (a) and (b) consonants followed by category (b) characters followed by category (c) Noon Ghunna.

Different combination of these aerab with category (b) characters generate the various vowels, as indicated in Table 4 (all vowels shown in combination with ب (phoneme /b/) as a consonant character is required as a placeholder for the aerab).

Bay + Zabar	بَ	ə
Bay + Zer	بِ	ɪ
Bay + Pesh	بُ	ʊ
Bay + NULL + Alef	با	ɑ
Bay + NULL + Vao	بو	o
Bay + Zabar + Vao	بَو	ɔ
Bay + Pesh + Vao	بُو	u
Bay + NULL + Yay	بے	e
Bay + Zabar + Yay	بَے	æ
Bay + (NULL | Zer)[2] + Yay	بی	i
Bay + NULL + Alef + Noon Ghunna	باں	ɑ̃
Bay + NULL + Vao + Noon Ghunna	بوں	õ
Bay + Zabar + Vao + Noon Ghunna	بَوں	ɔ̃
Bay + Pesh + Vao + Noon Ghunna	بُوں	ũ
Bay + NULL + Yay + Noon Ghunna	بیں	ẽ
Bay + Zabar + Yay + Noon Ghunna	بَیں	æ̃

[2] NULL or Zer. It is controversial whether Zer is present for the representation of vowel /i/. One solution is to process both cases till the diction controversy is solved.

Bay + (Null | Zer) + Yay + Noon Ghunna (see Footnote 2)	بِیں	ĩ

Table 4: Letter and aerab combinations and corresponding vowels

Existence of the remaining vocalic phoneme /ɛ/ is controversial in Urdu as there is no way of expressing it using the Urdu writing system and because it is schwa conditioned by the following /h/ phoneme and only occurs in this context. However, it may exist phonetically e.g. in the word شہر (/ʃɛhɛr/, "city") (see discussion in Qureshi, 1992; also see some supporting acoustic evidence in Fatima et. al, 2003, e.g. duration of /ɛ/ is 136 ms compared with 235 ms for /æ/).

The next category (g) consists of Khari Zabar. This represents the vowel Alef and, whenever occurs on top of a Vao or Yay, replaces these sounds with the Alef vowel sound /a/ as in words زکٰوۃ (/zakat/,"zakat") and اعلیٰ (/əʔlɑ/, special"). Sporadically Khari Zer and Ulta Pesh are referred to in Urdu as well but they generally do not occur on Urdu words. These are not considered here.

The gemination mark of category (h) is called Shad in Urdu and occurs on consonantal characters (of categories (a, b) except Alef). Shad geminates the consonant on which it occurs, which is normally word medially and inter-vocalically. As a result of gemination, the duplicate consonant acts as coda of previous syllable and onset of following syllable. For example, گدا (/gə.ḍɑ/, "a poor person") vs. گدّا (/gəḍ.ḍɑ/, "mattress").

The category (i) aerab, called Do-Zabar only occurs on Alef (in vocalic position) and converts the long vowel /a/ to short schwa followed by consonant /n/, e.g. in word فوراً (/fɔrən/, "immediately"). Do-Zer and Do-Pesh are similarly referred to in Urdu but are not generatively used and are mostly in foreign words especially of Arabic and are not considered further here. If considered, they would present a similar analysis. Finally, (j) is a very interesting category as it represents allo-graphs Alef and Hamza (former a character and latter (arguably) an aerab and character[3]). Both of them are default markers and occur in complimentary distribution, Alef always word initially and Hamza always otherwise. As discussed earlier, aerab in Urdu always need a *Kursi* ("seat"). If a short vowel occurs word initially without a consonant (i.e. in a syllable which has no onset), there is no placeholder for aerab. A default place holder is necessary and Alef is used. Word medially, if there is an onset-less syllable, Urdu faces the same problem. In these cases, Hamza (instead of Alef) is used as a placeholder for aerab. There are two further possible sub-cases. In one, the preceding syllable is open and ends with a vowel. This case is very frequent and Hamza is introduced inter.-vocalically (e.g. فائِدہ /fa.ɪdɛh/, "advantage"). In the second less productive sub-case, the preceding syllable is closed by a coda consonant. In this case, Hamza is (optionally) used with Alef (e.g. both forms are correct: جرأت / جرات /dʒʊr.ət/, "courage"). Hindi which employs a different mechanism by defining different shapes for vowels word-initially and word-medially (Matras). The Matras are anchored onto the consonants, e.g. in आने वाला , "about to come" vowel /a/ is written as आ word initially, but is written as ा word medially).

These rules have been implemented in an on-going project (see Footnote 1 above) and are successfully generating the desired phonemic output. This phonemic output is passed through sound change rule module to generate the desired phonetic form.

5 Conclusion

This paper briefly discusses the architecture of Natural Language Processing portion of an Urdu Text-to-Speech system. It explains the details of Urdu consonantal and vocalic system and Urdu letters. Urdu shows regular behavior and thus the phonemic forms are predictable from the textual input. The letter-to-sound rules define this

[3] Hamza sometimes requires a *Kursi* or seat (قائل and not قأل) and sometimes does not (پلاؤ and not پلائو) indicating it behaves both like a character and an aerab. It is still unclear on how this behavior is distributed and whether it is predictable. As it is a script centric issue, it is not discussed further here.

mapping and are thus essential for developing Urdu TTS.

6 Acknowledgements

This work has been partially supported by the grant for "Urdu Localization Project: MT, TTS and Lexicon" by E-Government Directorate of Ministry of IT and Telecommunications, Government of Pakistan.

The author also wishes to thank anonymous reviewers for comments, especially on glottal stop and Hamza and Tahira Khizar and Qasim Vaince for eventual discussion on the role of Hamza in Urdu script.

References

M. Afzal and S. Hussain. 2001. Urdu Computing Standards: Development of Urdu Zabta Takhti (UZT 1.01). *Proceedings of IEEE International Multi-topic Conference,* Lahore, Pakistan.

T. Dutoit. 1997. *An Introduction to Text-to-Speech Síntesis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

N. Fatima and R. Aden. Vowel Structure of Urdu. 2003. *CRULP Annual Student Report* published in *Akhbar-e-Urdu,* April-May, National Language Authority, Islamabad, Pakistan.

S. Hussain. 2003. www.LICT4D.aisa/Fonts/ Nafees_Nastalique. *Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society*, Asian Media Information Center, Singapore.

S. Hussain. 1997. *Phonetic Correlates of Lexical Stress in Urdu.* Unpublished Doctoral Dissertation, Northwestern University, Evanston, USA.

S. Hussain, and M. Afzal. 2001. Urdu Computing Standards: Urdu Zabta Takhti (UZT 1.01). *Proceedings of IEEE International Multi-topic Conference,* Lahore, Pakistan.

D. H. Klatt. 1980. Software for Cascade/Parallel Formant SynthesiZer. *JASA 67*: 971-995.

M. M. Rafique, M. K. Riaz, and S.R. Shahid. 2002. Vowel Insertion Grammar. *CRULP Annual Student Report* published in *Akhbar-e-Urdu,* April-May, National Language Authority, Islamabad, Pakistan.

B. A. Qureshi. 1992. *Standard Twentieth Centuary Dictionary: Urdu to English.* Educational Publishing House, New Dehli, India.

A. M. Saleem, H. Kabir, M.K. Riaz, M.M. Rafique, N. Khalid, and S.R. Shahid. 2002. Urdu Consonantal and Vocalic Sounds. *CRULP Annual Student Report* published in *Akhbar-e-Urdu,* April-May, National Language Authority, Islamabad, Pakistan.