Corpus Based Mapping of Urdu Characters for Cell Phones

M. Aamir Khan, M. Abid Khan, Asad Habib and M. Naveed Ali
Department of Computer Science, University of Peshawar, Pakistan
bitvox@yahoo.com, m.abid6@gmail.com, asadsan@gmail.com, naveed_asadecp@yahoo.com

Abstract

The use of cell phones has become prevalent in Pakistan. Several cell phone manufacturers have incorporated Urdu language keypads into their cell phone products. This paper analyzes the Urdu cell phone keypads and proposes a much better layout of Urdu character set on cell phone 12-buttons keypad.

1. Introduction

Cell phones use a standard telephone 12-key keypad. The standard numeric telephone keypad contains digits 0-9, * and # symbols. The cell phone keypad also contains characters on keys for entering text into cell phones. Several characters are mapped to the same key because of small number of buttons available on cell phone keypads. The multitap method is the simplest text entry method in such situation. In multitapping, the user presses each key one or more times to specify the desired input character [4]. Bilingual keyboards provide the ability to enter text in different languages [3]. Urdu-English twelve button keyboard is a bilingual keyboard for cell phones. Urdu language contains 45 characters compared to English language, which contains 26 characters. The large number of characters in Urdu language makes text entry very slow. Moreover, out of the 12 keys on mobile phones only 8 are used for entering text. All of the major brand cell phones use a standard mapping of Urdu characters given in table 1.

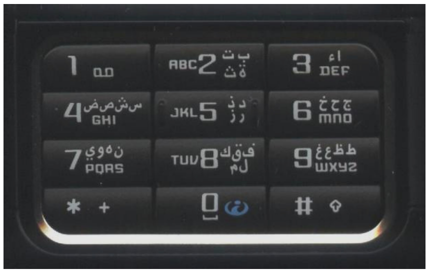

Figure 1: Nokia 3250 Arabic keypad

Table 1: Standard 12-button keyboard layout

	Order						
Key	I	II	III	IV	V	VI	VII
2	ب	پ	ت	ة	ٹ	ث	
3	ا	آ	ؤ	ۂ	ء	ئ	
4	س	ش	ص	ض			
5	د	ڈ	ز	ر	ڑ	ذ	ژ
6	ج	چ	ح	خ	ہ		
7	ن	و	ھ	ی	ے	ۓ	
8	ف	ق	ک	گ	ل	م	ں
9	ط	ظ	ع	غ			

The standard layout for Urdu language is derived from standard Arabic keypad implemented by handsets such as Nokia 3250 (www.nokia.com) (figure 1).

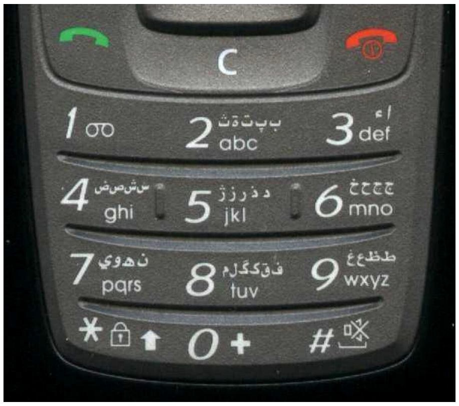

Figure 2: Samsung SGH-C140 Arabic/Urdu keypad

The extended keypad layout for Urdu is implemented by handsets such as Samsung SGH-C140 (www.samsung.com) (figure 2). This mapping is inefficient in terms of keystrokes per character (KSPC) and keystrokes per word (KSPW). The layout of characters for Urdu language on cell phone keypad can be improved based on the frequency analysis of Urdu alphabets.

2. Frequency Based Character Mapping

Frequency based cell phone keyboard layout has been studied for English language to make typing English text on cell phones easier and faster [1]. For Urdu language the optimized layout presented in Table 2 is based on character frequency analysis of 16,638,852 words raw corpus. The frequencies of individual characters in the corpus are shown in Table 3. The ordering of characters on each key was decided based on digraph frequencies. Figure 3 shows the optimized keypad based on mapping shown in table 2.

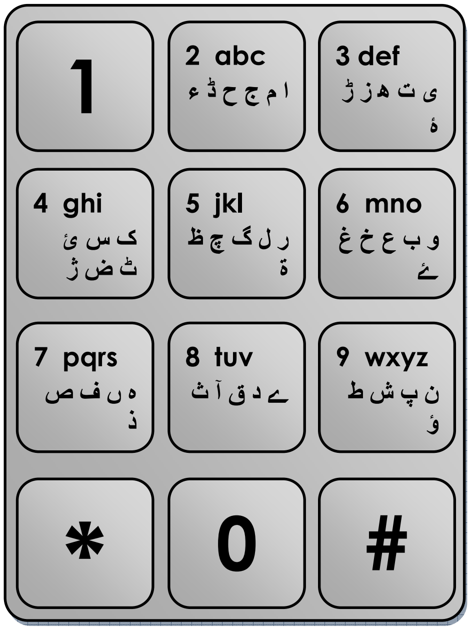

Figure 3: Optimized layout

Table 2: Optimized 12-button keyboard layout

	Order					
Key	I	II	III	IV	V	VI
2	ا	م	ج	ح	ڈ	ء
3	ی	ت	ھ	ز	ڑ	ۂ
4	ک	س	ئ	ٹ	ض	ژ
5	ر	ل	گ	چ	ظ	ۃ
6	و	ب	ع	خ	غ	ۓ
7	ہ	ں	ف	ص	ذ	
8	ے	د	ق	آ	ث	
9	ن	پ	ش	ط	ؤ	

Table 3: Urdu character frequencies

Unicode	Alphabet	Frequency	Percentage
627	ا	6733610	12.23570
6cc	ی	5752357	10.45266
6a9	ک	3911143	7.10697
631	ر	3669392	6.66768
648	و	3327481	6.04639
6c1	ہ	2994305	5.44098
6d2	ے	2857846	5.19302
646	ن	2773651	5.04003
645	م	2684946	4.87884
62a	ت	2117669	3.84803
633	س	1987451	3.61141
644	ل	1915841	3.48129
628	ب	1492997	2.71294
6ba	ں	1469466	2.67018
62f	د	1431230	2.60070
67e	پ	914273	1.66133
62c	ج	844670	1.53486
6be	ھ	800600	1.45478
626	ئ	664594	1.20764
6af	گ	643263	1.16888
639	ع	636166	1.15598
641	ف	546973	0.99391
642	ق	544460	0.98934
634	ش	532262	0.96718
62d	ح	501602	0.91147
632	ز	454158	0.82525
679	ٹ	420666	0.76440
686	چ	358159	0.65081
62e	خ	352729	0.64095
635	ص	327434	0.59498
622	آ	259879	0.47223
637	ط	220613	0.40088
688	ڈ	183081	0.33268
691	ڑ	143244	0.26029
636	ض	142813	0.25951
638	ظ	104163	0.18928

63a	غ	100331	0.18231
630	ذ	79372	0.14423
62b	ث	69641	0.12655
624	ؤ	32355	0.05879
621	ء	24930	0.04530
6c2	ۂ	4390	0.00798
698	ژ	2522	0.00458
629	ة	2275	0.00413
6d3	ۓ	1479	0.00269
Total		*55032482*	*100.00000*

The proposed optimized layout shown in Table 2 is based on the frequencies of Urdu characters. The layout has been constructed by mapping consecutive characters from rows of Table 3 to cells of Table 2.

3. Evaluation

The proposed keypad layout has been evaluated on character-set and words from the lexicon derived from Urdu corpus. Table 4 shows the comparison of keystrokes per character for individual Urdu alphabets on 'standard' layout and frequency based layout. The keystrokes per character for 'standard' layout are shown in column named KSPC-S (Keystrokes per Character on Standard layout). The column with KSPC-F (Keystrokes per Character on Frequency based layout) shows keystrokes per character on frequency based layout of the keypad.

Table 4: Keystroke per character comparison

Alpha	%	KSPC-S	KSPC-F	S-Exp	F-Exp
ا	12.24	1	1	12.24	12.24
ی	10.45	4	1	41.81	10.45
ک	7.11	3	1	21.32	7.11
ر	6.67	4	1	26.67	6.67
و	6.05	2	1	12.09	6.05
ہ	5.44	5	1	27.20	5.44
ے	5.19	5	1	25.97	5.19
ن	5.04	1	1	5.04	5.04
م	4.88	6	2	29.27	9.76
ت	3.85	3	2	11.54	7.70
س	3.61	1	2	3.61	7.22
ل	3.48	5	2	17.41	6.96
ب	2.71	1	2	2.71	5.43
ں	2.67	7	2	18.69	5.34
د	2.60	1	2	2.60	5.20
پ	1.66	2	2	3.32	3.32
ج	1.53	1	3	1.53	4.60
ھ	1.45	3	3	4.36	4.36
ئ	1.21	6	3	7.25	3.62
گ	1.17	4	3	4.68	3.51
ع	1.16	3	3	3.47	3.47
ف	0.99	1	3	0.99	2.98
ق	0.99	2	3	1.98	2.97
ش	0.97	2	3	1.93	2.90
ح	0.91	3	4	2.73	3.65
ز	0.83	3	4	2.48	3.30
ٹ	0.76	5	4	3.82	3.06
چ	0.65	2	4	1.30	2.60
خ	0.64	4	4	2.56	2.56
ص	0.59	3	4	1.78	2.38
آ	0.47	2	4	0.94	1.89
ط	0.40	1	4	0.40	1.60
ڈ	0.33	2	5	0.67	1.66
ڑ	0.26	5	5	1.30	1.30
ض	0.26	4	5	1.04	1.30
ظ	0.19	2	5	0.38	0.95
غ	0.18	4	5	0.73	0.91
ذ	0.14	6	5	0.87	0.72
ث	0.13	6	5	0.76	0.63
ؤ	0.06	3	5	0.18	0.29
ء	0.05	5	6	0.23	0.27
ۂ	0.01	4	6	0.03	0.05
ژ	0.00	7	6	0.03	0.03
ة	0.00	4	6	0.02	0.02
ۓ	0.00	6	6	0.02	0.02
Total	*100.00*	*154*	*150*	*309.96*	*166.73*

The last two columns of Table 4 show the expected values of keystrokes for 'standard' (S-Exp: Standard layout Expectancy) and frequency based (F-Exp: Frequency based layout Expectancy) layouts respectively. The expectancy values for each character have been computed by multiplying percentage by the number of keystrokes required by each layout. Looking at the last row of table 4, it is evident that a total of the expectancy value of all the characters for the 'standard' layout i.e. 309.96 is much larger than 166.73 for frequency based layout. As a result, most frequently occurring characters are typed quickly compared to the least occurring characters. The layout has also been evaluated on 100 most frequent Urdu words [2]. The number of keystrokes per word (KSPW) was reduced by 50.16357688% in frequency based layout (FBL) KSPW as compared to current standard KSPW. The comparison of keystrokes between current standard layout and frequency based layouts is given in table 5. Moreover, the number of keystrokes required for a lexicon of 51218 words (excluding the probability of each word) reduced by 36.73491806% KSPW in FBL-KSPW which is a significant improvement over the current standard layout KSPW.

Table 5: Keystroke count for 100 most frequent words

S#	Word	Frequency	KSPW-S	KSPW-F
1	کے	618958	8	2
2	میں	510330	17	5
3	کی	495344	7	2
4	ہے	417230	10	2
5	اور	352897	7	3
6	سے	319683	6	3
7	کا	268072	4	2
8	کو	239480	5	2
9	اس	221585	2	3
10	نے	200405	6	2
11	ہیں	196799	16	4
12	کہ	184643	8	2
13	پر	173181	6	3
14	بھی	127457	8	6
15	یہ	120063	9	2
16	ایک	116695	8	3
17	کر	111749	7	2
18	نہیں	103967	17	5
19	ان	97549	2	2
20	ہو	90129	7	2
21	کیا	89452	8	3
22	تو	82484	5	3
23	وہ	75497	7	2
24	لئے	60458	16	6
25	تھا	55527	7	6
26	پاکستان	55404	12	10
27	کرنے	52084	13	4
28	جو	51059	3	4
29	ہی	45321	9	2
30	و	43449	2	1
31	نہ	42718	6	2
32	اپنے	41801	9	5
33	کہا	41399	9	3
34	آپ	40080	4	6
35	گیا	39963	9	5
36	جس	38483	2	5
37	تھے	37790	11	6
38	تک	37160	6	3
39	جائے	35325	13	8
40	لیکن	35145	13	5
41	بعد	34798	5	7
42	ساتھ	34483	8	8
43	ہوئے	33759	18	6
44	کوئی	33042	15	6
45	کریں	31839	18	5
46	دیا	31410	6	4
47	گا	31300	5	4
48	بہت	31106	9	5
49	اپنی	30280	8	5
50	رہے	30275	14	3
51	زیادہ	29315	14	9
52	کسی	29142	8	4
53	یا	28754	5	2
54	کرتے	28696	15	5
55	ہم	28667	11	3
56	ہونے	28484	13	4
57	تھی	28288	10	6
58	انہوں	27802	16	6
59	والے	27670	13	5
60	طرح	26850	8	9
61	بات	25951	5	5
62	جب	25853	2	5
63	اگر	25853	9	5
64	اب	25481	2	3
65	ہوں	25341	14	4
66	گئے	24980	9	4
67	ملک	24930	14	5
68	ہوتا	24642	11	5
69	وقت	24503	7	6
70	گی	24501	8	4
71	ہوا	23885	8	3

72	کام	23084	10	4
73	حاصل	23054	12	11
74	گئے	22708	15	7
75	دو	22705	3	3
76	حکومت	22680	17	10
77	کیلئے	22399	23	8
78	گئی	22145	14	7
79	رہا	21722	10	3
80	جاتا	21473	6	7
81	کرنا	21360	9	4
82	سب	20903	2	4
83	ہر	20774	9	2
84	کچھ	20083	8	8
85	صرف	19370	8	8
86	پہلے	19214	17	6
87	طور	19075	7	6
88	سی	18940	5	3
89	ہوتی	18865	14	5
90	دی	18834	5	3
91	کم	18739	9	3
92	جا	18571	2	4
93	وجہ	18099	8	5
94	رہی	18009	13	3
95	اسے	17907	7	4
96	انہیں	17751	18	6
97	اسی	17345	6	4
98	محمد	16988	16	10
99	سال	16655	7	5
100	بی	16584	5	3
		Total	*917*	*457*

The total number of keystrokes for 100 most occurring words in contemporary 'standard' layout is 917 where as in the proposed layout it is 457 which is an improvement of 50.16357688%.

4. Conclusion

The frequency based Urdu characters layout on 12-button phone keypad reduces the keystrokes per word significantly compared to the standard layout. The probabilistic analysis of 51218 words from the Urdu corpus shows that the proposed frequency based layout reduces keystrokes by 46% compared to the standard keyboard layout. Keeping in view the large number of character in Urdu language compared to the number of keys available on the mobile phone, memorizing the layout is worthwhile and practical.

5. References

[1] M. A. K. Azad, R. Sharmeen, S. Ahmad and R. Mahmud, *"Frequency and Flexibility Based Cell Phone Keypad Layout*", *TENCON* 2005 IEEE Region 10, Nov. 2005, pp. 1-5.

[2] M. Ijaz and S. Hussain. *"Corpus Based Urdu Lexicon Development"*, in proc. *the conference on Language & Technology (CLT07)*, Bara Gali Summer Campus, University of Peshawar, August 7-11, 2007, pp. 85-94.

[3] D. S. Katre, "*Position Paper On Cross-cultural Usability Issues of Bilingual (Hindi & English) Mobile Phones"*, in proc. *Indo-Danish HCI Research Symposium*, Hosted by Department of Design, Indian Institute of Technology, Guwahati, India., 2006.

[4] S. MacKenzie and S. R. William. *"Text Entry for Mobile Computing: Models and Methods, Theory and Practice". Human Computer Interaction Volume 17*, Lawrence Erlbaum Associates Inc, 2002, pp. 147–198.