Design of Urdu Virtual Keyboard

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

Abstract

This paper presents the first ever virtual keyboard layout based on character frequency analysis of Urdu Corpus. To optimize the keyboard layout Monte Carlo Simulation with simulated annealing is used. Furthermore, the proposed keyboard layout is augmented with word prediction list derived from Urdu corpus to speed up text entry. Performance analysis of keyboard layout is done for justification purposes.

1. Introduction

Virtual/soft keyboards allow users, to input text using touch screen and stylus. For English language, several virtual keyboard layouts have been proposed. These include MacKenzie's and Zhang's OPTI layout, improved OPTI layout in a 5x6 layout (OPTI II) with 38 wpm (words per minute), FITALY keyboard and Chubon keyboards [6]. Evaluation of the performance of virtual keyboards involves the use of Fitts' Law [1][2]. Keyboard input speed is measured in wpm (words per minute). Mean time (MT), to move to a key on virtual keyboard, is computed in terms of moving to a target key K of width W lying at distance A from the current position of pointing device [3]. The layout of keys on virtual keyboard should be such that to minimize the mean time for all digraph movements. The digraph frequencies are a natural feature of languages. Mackenzie and Zhang evaluated the performance of their virtual keyboard by computing 27x27 digraph frequencies from a corpus [5]. The distances (amplitudes) for all the 27 x 27 digraph movements in a given keyboard layout were computed, and for each movement the Fitts' Law was used to compute the MT [5]. The following equation was used to compute the MT [5].

$$MT = \frac{1}{4.9}\log_2\left(\frac{A}{W}+1\right)$$

Here, W is the width of the key and A is the distance to move to the target key K. Each mean time was weighted by the digraph probability. The wpm (words per minute) was calculated by multiplying MT with the average number of characters per word. The computed wpm is an "upper limit" on the text entry speed. The "*visual scan"* time to find a key was assumed to be zero. The keyboard layout can be done manually [5] or using optimization techniques such as Monte Carlo simulation [6].

2. Urdu virtual keyboard design

Urdu has 37 base characters. The character set used for designing the keyboard, proposed in this research paper, also contains Arabic characters. This facilitates keypad to be used for entering Arabic text as well, but it is not optimized for Arabic language. Table 1 shows the set of Urdu alphabets.

Table 1: Urdu language characters

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

The first step in designing a virtual keyboard is to determine the digraph frequencies. For Urdu language computing, digraph frequencies require a corpus. A raw corpus consisting of 16,638,852 words was collected. It contained collections of newspaper

articles, books and magazines. Table 2 shows the character frequencies of individual Urdu alphabets in descending order.

Table 2: 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	ۃ	2275	0.00413
6d3	ۓ	1479	0.00269
Total		*55032482*	*100.00000*

The 46x46 digraph frequency table was computed from the corpus. The digraph is shown in the form of color chart in Figure 1. The dark shaded cells represent higher frequency digraphs, whereas light shaded cells represent lower frequency digraphs.

In Figure 1, the character on the Y axis shows the first character while the one on the X axis shows the second character in a digraph. The order of characters in columns and rows of digraph's color chart is the same as in table 1. The last column and the last row represent the space character. The digraph frequency table was used for computing the wpm performance of the keyboard.

To compute the performance of a given keyboard layout, the following equation was used [5].

$$MT = \sum_{i=0}^{k}\sum_{j=0}^{k}\frac{1}{4.9}\log_2\left(\frac{A_{i,j}}{W}+1\right)\times d(i,j)$$

Here, $A_{i,j}$ is the distance from key i to key j. The $d(i,j)$ represents the digraph frequency of character i followed by character j. The variable k is the number of characters in a given language. The diagonal entries in digraph frequency table where $i=j$ denote repeated character where no movement of stylus is involved. For repeated characters, the repeat stylus tapping time was set as 0.127 seconds as in Zhai et.al [6].

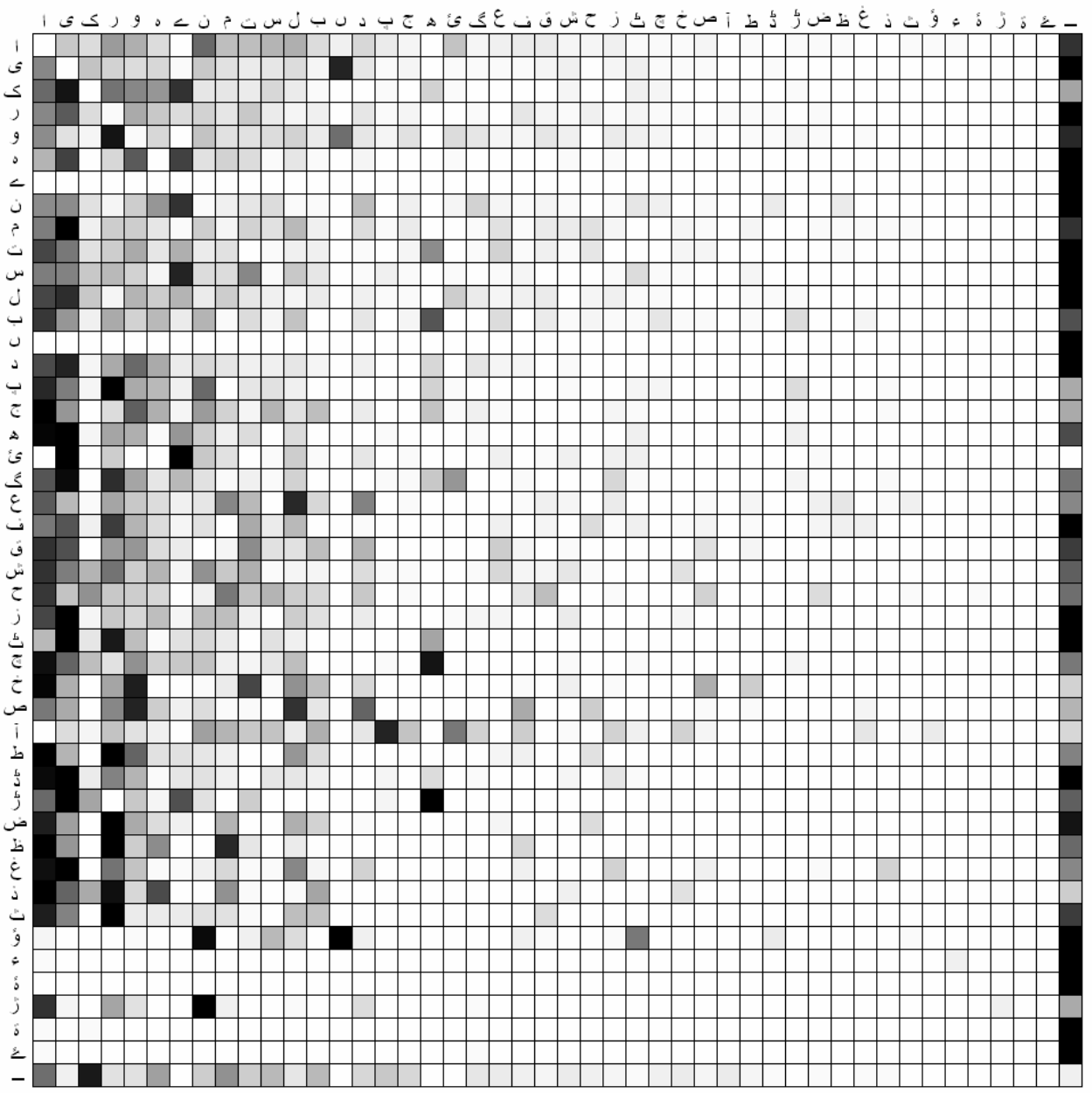

Figure 1: Urdu digraph color chart

Designing a keyboard layout is a combinatorial task and requires $O(n!)$ searches [6]. The layout should be arranged such that the MT is minimized. For optimizing the keyboard layout, 700 runs of Monte-Carlo Simulation with simulated annealing were executed. The best layout was found at 53rd simulation. In each run 100,000-200,000 random movements were tried each on keyboard layout. Annealing schedule was adjusted by trial and error. The width of each key was set at 50 pixels. The shape of keys is based on the work of Zhai et. al [6]. Figure 2 shows the optimized layout of the Urdu keyboard.

The speed of entering text, using the layout in figure 2, was computed using the following equation from Zhai et.al [6].

$$wpm = 60 / AWL \times MT$$

where AWL is the average word length in a language. MT is the mean time. In Urdu language, the average

word size was found to be 7 characters. The constant 60 is the number of seconds in a minute. When a space after each word is added it becomes 8 characters.

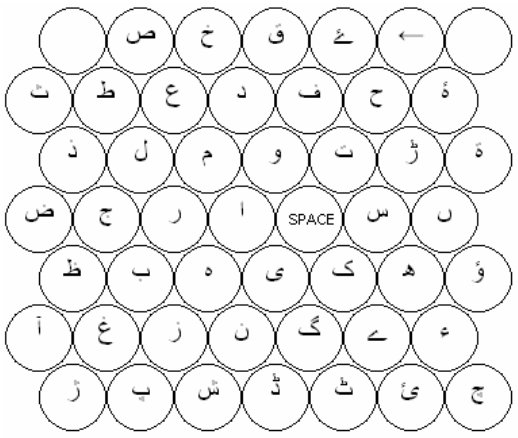

Figure 2: Corpus based optimized Urdu virtual keyboard layout arranged in 7x7 cells

For the optimized keyboard, *MT* was found to be

$$MT = 0.20609985$$

$$wpm = (60/8 \times 0.20609985) = 36.3901\ wpm$$

The predicted speed of the keyboard is thus 36.3901 words per minute. To utilize the space between circular keys, the shape of keys was changed to hexagonal. Figure 3 shows the improved design of optimized layout.

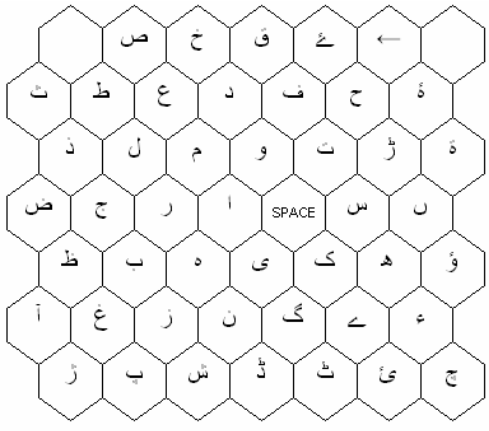

Figure 3: Improved Urdu virtual keyboard to utilize the empty gaps between keys

A prototype version of keyboard was implemented using Microsoft Visual C++ 6.0 for Microsoft windows. The program helped the user by highlighting the next probable keys and drawing rings around the keys. The darker color showed the higher probability of occurrence while the lighter color showed lower probability of occurrence. The most probable next character shows the ring in blinking mode.

Figure 4 shows the typing of the word مشاعرہ on the keyboard. After the first three characters have been entered, the next set of probable characters is highlighted in different shades of red color. To further improve the performance of user input speed, a prediction list was added. Figure 5 shows the use of prediction list.

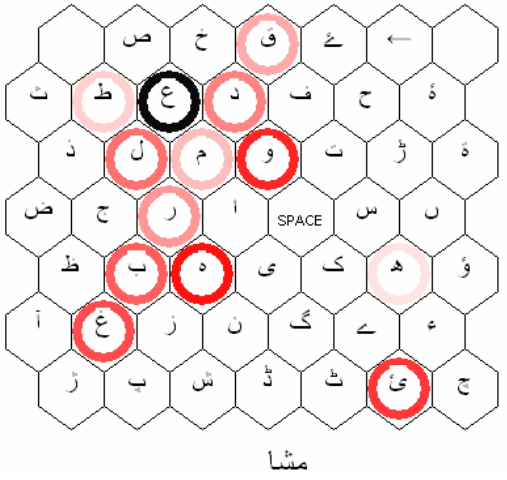

Figure 4: Predictive input of the word مشاعرہ

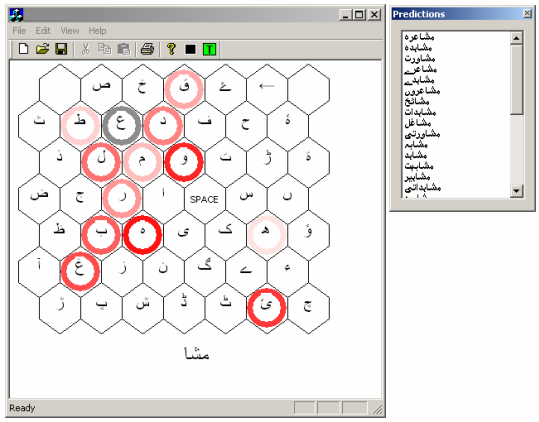

Figure 5: Input with prediction list assistance

For analysis, the performance of the keyboard was determined on various words. Table 3 shows the computed distances of typing 38 most frequently occurring words [2] in the corpus along with a space character after each word. The distances are computed in terms of keys to traverse a given particular word. The distance covered depends on the position of keys and characters in a word.

Table 3: Distances for 38 frequent words along with a space character

Word	Characters	Frequency	Distance
کے	ک ے	618958	3
میں	م ی ں	510330	6
کی	ک ی	495344	3
ہے	ہ ے	417230	4
اور	ا و ر	352897	5
سے	س ے	319683	4
کا	ک ا	268072	4
کو	ک و	239480	4
اس	ا س	221585	5
نے	ن ے	200405	4
ہیں	ہ ی ں	196799	6
کہ	ک ہ	184643	3
پر	پ ر	173181	5
بھی	ب ھ ی	127457	5
یہ	ی ہ	120063	2
ایک	ا ی ک	116695	4
کر	ک ر	111749	5
نہیں	ن ہ ی ں	103967	7
ان	ا ن	97549	3
ہو	ہ و	90129	4
کیا	ک ی ا	89452	4
تو	ت و	82484	3
وہ	و ہ	75497	3
لئے	ل ئ ے	60458	8
تھا	ت ھ ا	55527	5
پاکستان	پ ا ک س ت ا ن	55404	15
کرنے	ک ر ن ے	52084	8
جو	ج و	51059	4
ہی	ہ ی	45321	2
و	و	43449	2
نہ	ن ہ	42718	3
اپنے	ا پ ن ے	41801	10
کہا	ک ہ ا	41399	5
آپ	آ پ	40080	6
گیا	گ ی ا	39963	5
جس	ج س	38483	5
تھے	ت ھ ے	37790	6
تک	ت ک	37160	4

3. Evaluation

The proposed layout presented in Figure 2 was evaluated on 20 students of computer science program. The average text entry speed was found to be 13.47 wpm based on an initial two hour training prior to the evaluation. The maximum speed achieved was 22.5 wpm. When compared to virtual keyboards for English language, the text entry speed is comparable to OPTI and QWERTY layout [5]. The predicted speed of OPTI layout is 58.2 words. Actual speed of OPTI has been found to be 44.3 wpm after 20 sessions of text entry, each for 45 minutes [5]. With extended training of the user the text entry speed of Urdu virtual keyboard can be improved.

4. Conclusion

The design of the virtual keyboard presented in this paper is based on character analysis of Urdu corpus. The virtual keyboard is particularly useful for occasional users, who do not want or do not have time to learn the hardware based keyboard layout. Being the first virtual keyboard for Urdu language, comparative study of performance with other virtual keyboards is not possible. The predicted speed of text entry using this keyboard layout is 36.3901 wpm. The keyboard is also usable for entering Arabic text, but it is not optimized for Arabic.

5. References

[1] P..M. Fitts, *"The information capacity of the human motor system in controlling the amplitude of movement", Journal of Experimental Psychology*, 1954, pp. 381-391.

[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, Agust 7-11, 2007, pp. 85-94.

[3] I. S. MacKenzie, A. Sellen and W. Buxton, "*A comparison of input devices in elemental pointing and dragging tasks"*, in proc. *CHI*, 1991.

[4] I. S. MacKenzie, "*A note on the information theoretic basis for Fitts' law", Journal of Motor Behavior*, 1989, pp. 323-330.

[5] I. S. MacKenzie and S. X. Zhang, "*The Design and evaluation of a High performance Soft Keyboard"*, in proc. *CHI 91*, 15-20 May 1999.

[6] S. Zhai, M. Hunter and B. A. Smith, "*The Metropolis Keyboard -- An Exploration of Quantitative Techniques for Virtual Keyboard Design",* in proc. *UIST'2000 - the 13th Annual ACM Symposium on User Interface Software and Technology,* 2000.