

SINHALA T9 TEXT ENTRY SYSTEM

M.H. Dewapura

This dissertation was submitted in requirements for the Master of Engineering degree Master of Science in computer science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

2007

93366

Abstract

T9 Text Input is an input technology used in mobile devices. It lets words be formed by a single key press for each letter, as opposed to the multiple key press approach used in the older generation of mobile phones. It works via active reference to a dictionary of commonly used words.

T9 databases are currently available in 15 different character scripts in 62 languages including specialized language engines for Alphabetic, Chinese and Japanese languages. However it is not available for Sinhala. Development of T9 is more valuable for Sinhala than English as the number of letters assigned to a key in the Sinhala keypad is more than that of the English keypad.

We developed a system for predictive keypad text entry in Sinhala. Predictive keypad text entry allows the user to type words efficiently just pressing a key one time for each letter.


The major objectives achieved in this project are the building of a Sinhala word database, identification of common words and development of the algorithm for predictive keypad text entry and an application for the T9 PC simulator. We used the Sinhala keypad layout used in the Nokia mobile phones for our project.

A large number of Sinhala words were collected, several tools to process the words were developed and a database, mapping key sequences to prioritized lists of Sinhala words, was built. A PC application to simulate keypad text entry, update the database with new words and to change word priorities was developed.

Finally we compared the required number of key presses for Sinhala text input using T9 and using multi-tap text entry, and showed that our system enables users to enter Sinhala text easily, quickly and efficiently.

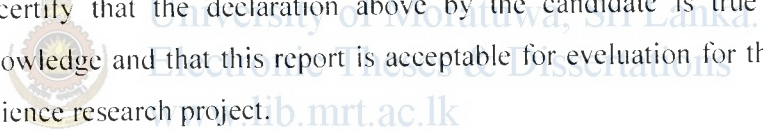
Declaration

"I certify that this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due references made in the text."


.....
M.H. Dewapura.

20/04/2009
.....
Date

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the MSc in Computer Science research project.


.....
Prof. Gilhan V. Dias.
Professor,
Department of Computer Science and Engineering,
University of Moratuwa.

.....
.....
Date

Acknowledgements

Project supervisor, Prof. Gihan V. Dias has shown lot of flexibility in consulting him either personally, through emails or even via phone as it saved lot of time and effort and led us in the correct direction. Also, we referred research papers published by him. His intervention helped to find research materials, collect the Sinhala Corpus etc. I would really like to appreciate his support extended to date from the day the project has initiated.

Mr. Kasun Karunaratna, who is a M.Sc. student in the Computer Science and Engineering department of the University of Moratuwa, gave us a tool to convert non-Unicode text into Unicode text. It was really helped us to collect words in Unicode format.

Dr. Ruwan Weerasingha, Language Technology Research Lab of the University of Colombo was able to provide us the beta version of the Sinhala Corpus and without it, building of the Sinhala word database had not been successful.

Also, I would be thankful to all the lectures in the Dept. of Computer Science and Engineering, University of Moratuwa, who attended to my progress presentations. Their feedback led us in the correct path. Special thank is deserved by Dr. Sanath Jayasena for his relentless efforts on scheduling progress meetings and guiding us on a methodical project progress.

Table of Contents

	Page
Declaration	iii
Abstract	iv
Acknowledgement	v
List of Figures	viii
List of Tables.....	ix
1.0 Chapter 1 - Introduction	1
2.0 Chapter 2 - Literature Review	4
2.1 Text Entry Methods.....	4
2.1.1 Keyboards and allocations of letters.....	4
2.1.2 Chorded Keyboards.....	8
2.1.3 Virtual Keyboards.....	8
2.1.4 Touch-screens.....	8
2.1.5 Keypads.....	9
2.2 Unicode Representation	16
2.3 T9 Solutions.....	20
2.3.1 T9 Text Input.....	20
2.3.2 T9 Text output.....	23
2.3.3 XT9 Mobile Interface.....	23
2.3.4 T9 Coding Systems.....	24
2.3.5 T9 Vs Multi-tap text entry.....	29
2.4 Dictionary Building.....	29
2.4.1 Building of a Corpus.....	30
2.4.2 Corpus Annotation	31
2.4.3 Documentation of the UCSC/LTRL Sinhala Corpus.....	32
3.0 Chapter 3 - Methodology of Design.....	36
3.1 Methodology Overview.....	36
3.2 Implementation of Sinhala T9 text entry system.....	38

3.2.1	Building of a Sinhala Dictionary.....	38
3.2.2	Development of an algorithm.....	42
3.2.3	Development of an application for Sinhala T9 text entry system.....	46
4.0	Chapter 4 - Analysis, conclusion and future work	51
4.1	T9 Sinhala Dictionary.....	51
4.2	T9 - PC Simulator	51
4.3	Additional Features	51
4.3.1	Ability to add user's own words.....	51
4.3.2	Ability to completing words.....	52
4.3.3	Ability to adjust the order of the words based on user preference.....	52
4.4	Future work.....	52
4.4.1	Ability to adjust the order of the words based on user prior usage....	52
4.4.2	Next word prediction.....	52
4.4.3	Enhanced word completion	52
5.0	Chapter 5 – Summary	53
6.0	References.....	54



University of Moratuwa, Sri Lanka.
 Electronic Theses & Dissertations
www.lib.mrt.ac.lk

List of Figures

	Page
Figure 1: Allocation of letters in the US English keyboard.....	5
Figure 2: Allocation of letters in the French (France) keyboard.....	6
Figure 3: The Divehi Phonetic keyboard layout.....	6
Figure 4: Wijesekara Keyboard.....	7
Figure 5: Standard Sinhala computer keyboard layout.....	7
Figure 6: An ergonomic chorded keyboard.....	8
Figure 7: A standard mobile phone keypad	11
Figure 8: Three different English keypad layouts.....	12
Figure 9: Keypad design for Tamil in alphabetical arrangement.....	13
Figure 10: The Sinhala keypad layout	15
Figure 11: XT9 Mobile Interface.....	23
Figure 12: Allocation of English letter to keypad.....	24
Figure 13: Classification of the articles in the corpus.....	33
Figure 14: Functionality of the T9 Text entry system.....	43
Figure 15: T9 Text Entry System Flowchart	44
Figure 16: Find list of complete words	45
Figure 17: Find list of partial words	45
Figure 18: Typing word - බලනවා - after key press 8.....	47
Figure 19: Typing word - බලනවා - after key press 89.....	47
Figure 20: Typing word - බලනවා - after key press 897.....	48
Figure 21: Typing word - බලනවා - after key press 8970.....	48
Figure 22: Typing word - බලනවා - after key press 89702.....	49
Figure 23: Sample sentence – “මම අද ඉක්මනට ගෙදර එනවා”.....	49
Figure 24: T9-OFF mode.....	50

List of Tables

	Page
Table 1 : The best eight-key constrained and unconstrained keypad designs....	12
Table 2 : The assignment of Sinhala characters to keypad.....	15
Table 3 : The Unicode characters assigned to Sinhala letters and signs.....	18
Table 4 : Sample words from the database and their frequencies.....	41
Table 5 : Sample words and their key combinations.....	42
Table 6 : Ordering words with same key sequences.....	50



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk