



VOICE)))

USER INTERFACE DESIGN

Michael H. Cohen
James P. Giangola
Jennifer Balogh

Voice User Interface Design

Michael H. Cohen
James P. Giangola
Jennifer Balogh

◆◆Addison-Wesley

Boston • San Francisco • New York • Toronto • Montreal

London • Munich • Paris • Madrid

Capetown • Sydney • Tokyo • Singapore • Mexico City

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, please contact:

U.S. Corporate and Government Sales
(800) 382-3419
corpsales@pearsontechgroup.com

For sales outside of the U.S., please contact:

International Sales
(317) 581-3793
international@pearsontechgroup.com

Visit Addison-Wesley on the Web: www.awprofessional.com

Library of Congress Cataloging-in-Publication Data

Voice user interface design / Michael H. Cohen, James P. Giangola, and Jennifer Balogh.
p. cm.

ISBN 0-321-18576-5 (alk. paper)

1. Automatic speech recognition. 2. Human-computer interaction.
3. User interfaces (Computer systems) I. Cohen, Michael H. (Michael Harris), 1953- II. Giangola, James P. III. Balogh, Jennifer.

TK7895.S65.V735 2004
621.39'9—dc22

2003022900

Copyright © 2004 by Nuance Communications, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior consent of the publisher. Printed in the United States of America. Published simultaneously in Canada.

For information on obtaining permission for use of material from this work, please submit a written request to:

Pearson Education, Inc.
Rights and Contracts Department
75 Arlington Street, Suite 300
Boston, MA 02116
Fax: (617) 848-7047

ISBN 0-321-18576-5
Text printed on recycled paper
1 2 3 4 5 6 7 8 9 10—CRS—0807060504
First printing, January 2004

Brief Contents

About the Authors and Radio Rex. xvii

Preface xxi

Part I. Introduction

1 Introduction to Voice User Interfaces 1

2 Overview of Spoken Language Technology 15

3 Overview of the Methodology 33

Part II. Definition Phase: Requirements Gathering and High-Level Design

4 Requirements and High-Level Design Methodology 45

5 High-Level Design Elements 63

6 Creating Persona, by Design 75

7 Sample Application: Requirements and High-Level Design 85

Part III. Design Phase: Detailed Design

8 Detailed Design Methodology	103
9 Minimizing Cognitive Load	119
10 Designing Prompts	133
11 Planning Prosody	171
12 Maximizing Efficiency and Clarity	205
13 Optimizing Accuracy and Recovering from Errors	217
14 Sample Application: Detailed Design	229

Part IV. Realization Phase: Development, Testing, and Tuning

15 Development, Testing, and Tuning Methodology	245
16 Creating Grammars	257
17 Working with Voice Actors	275
18 Sample Application: Development, Testing, and Tuning.	297
19 Conclusion	311
Appendix	313
Bibliography	315
Index	325

About the A

Preface

Part I. Introduction

1 Introduction

- 1.1 What Is
 - 1.1.1 Aim
 - 1.1.2 Spe
- 1.2 Why Spe
- 1.3 Where E

2 Overview of

- 2.1 Architec
 - 2.1.1 Ele
 - 2.1.2 Rea
 - 2.1.3 Oti
- 2.2 The Imq
 - 2.2.1 Per
 - 2.2.2 Pru
 - 2.2.3 De
- 2.3 Conclus

Contents

About the Authors and Radio Rex	xvii
---------------------------------------	------

Preface	xxi
---------------	-----

Part I. Introduction

1 Introduction to Voice User Interfaces	3
--	----------

1.1 What Is a Voice User Interface?	5
1.1.1 Auditory Interfaces	6
1.1.2 Spoken Language Interfaces	7
1.2 Why Speech?	9
1.3 Where Do We Go from Here?	12

2 Overview of Spoken Language Technology	15
---	-----------

2.1 Architecture of a Spoken Language System	16
2.1.1 Elements of a Spoken Language System	16
2.1.2 Recognition	19
2.1.3 Other Speech Technologies	24
2.2 The Impact of Speech Technology on Design Decisions	26
2.2.1 Performance Challenges	27
2.2.2 Problem Solving	28
2.2.3 Definition Files	29
2.3 Conclusion	31

3 Overview of the Methodology	33
3.1 Methodological Principles	33
3.1.1 User Input	34
3.1.2 Integrated Business and User Needs	35
3.1.3 Thorough Early Work	35
3.1.4 Conversational Design	36
3.1.5 Context	37
3.2 Steps of the Methodology	38
3.2.1 Requirements Definition	38
3.2.2 High-Level Design	39
3.2.3 Detailed Design	39
3.2.4 Development	39
3.2.5 Testing	40
3.2.6 Tuning	40
3.3 Applying the Methodology to Real-World Applications	41
3.3.1 Coordination of Phases	41
3.3.2 Dealing with Real-World Budget and Time Constraints	42
3.4 Conclusion	42

Part II. Definition Phase: Requirements Gathering and High-Level Design

4 Requirements and High-Level Design Methodology	45
4.1 Requirements Definition	45
4.1.1 Understanding the Business	46
4.1.2 Understanding the Users	48
4.1.3 Understanding the Application	53
4.2 High-Level Design	55
4.2.1 Key Design Criteria	55
4.2.2 Dialog Strategy and Grammar Type	56
4.2.3 Pervasive Dialog Elements	56
4.2.4 Recurring Terminology	56
4.2.5 Metaphor	57
4.2.6 Persona	57
4.2.7 Nonverbal Audio	58
4.3 Conclusion	61

5 High-Level Design	
5.1 Dialog Strategy	
5.2 Pervasive Design	
5.2.1 Error Handling	
5.2.2 Universal Design	
5.2.3 Login	
5.3 Conclusion	
6 Creating Personas	
6.1 What Is a Persona	
6.2 Where Does a Persona Come From	
6.3 A Checklist for Creating a Persona	
6.3.1 Metaphor	
6.3.2 Brand	
6.3.3 End User	
6.3.4 Application	
6.4 Persona Definition	
6.5 Conclusion	

7 Sample Applications	
7.1 Lexington Business	
7.2 Requirements	
7.2.1 Understanding	
7.2.2 Understanding	
7.2.3 Understanding	
7.3 High-Level Design	
7.3.1 Key Design	
7.3.2 Dialog	
7.3.3 Pervasive	
7.3.4 Recurring	
7.3.5 Metaphor	
7.3.6 Persona	
7.3.7 Nonverbal	
7.4 Conclusion	

5 High-Level Design Elements	63
5.1 Dialog Strategy and Grammar Type	63
5.2 Pervasive Dialog Elements	67
5.2.1 Error Recovery Strategies	67
5.2.2 Universals	72
5.2.3 Login	72
5.3 Conclusion	73
 6 Creating Persona, by Design	75
6.1 What Is Persona?	77
6.2 Where Does Persona Come From?	78
6.3 A Checklist for Persona Design	79
6.3.1 Metaphor and Role	79
6.3.2 Brand and Image	80
6.3.3 End Users	80
6.3.4 Application	81
6.4 Persona Definition	81
6.5 Conclusion	83
 7 Sample Application: Requirements and High-Level Design	85
7.1 Lexington Brokerage	86
7.2 Requirements Definition	86
7.2.1 Understanding the Business Goals and Context	87
7.2.2 Understanding the Caller	89
7.2.3 Understanding the Application	92
7.3 High-Level Design	95
7.3.1 Key Design Criteria	95
7.3.2 Dialog Strategy and Grammar Type	95
7.3.3 Pervasive Dialog Elements	96
7.3.4 Recurring Terminology	97
7.3.5 Metaphor	97
7.3.6 Persona	98
7.3.7 Nonverbal Audio	100
7.4 Conclusion	100

Part III. Design Phase: Detailed Design

8 Detailed Design Methodology	103
8.1 Anatomy of a Dialog State	104
8.2 Call Flow Design	105
8.3 Prompt Design	107
8.3.1 <i>Conversational Design</i>	108
8.3.2 <i>Auditory Design</i>	110
8.4 User Testing	110
8.4.1 <i>Formal Usability Testing</i>	111
8.4.2 <i>Card Sorting</i>	116
8.5 Design Principles	116
8.6 Conclusion	118
9 Minimizing Cognitive Load	119
9.1 Conceptual Complexity	120
9.1.1 <i>Constancy</i>	121
9.1.2 <i>Consistency</i>	123
9.1.3 <i>Context Setting</i>	124
9.2 Memory Load	125
9.2.1 <i>Menu Size</i>	126
9.2.2 <i>Recency</i>	126
9.2.3 <i>Instruction</i>	126
9.3 Attention	129
9.4 Conclusion	131
10 Designing Prompts	133
10.1 Conversation as Discourse	135
10.2 Cohesion	137
10.2.1 <i>Pronouns and Time Adverbs</i>	138
10.2.2 <i>Discourse Markers</i>	139
10.3 Information Structure	147
10.4 Spoken Versus Written English	152
10.4.1 <i>Pointer Words</i>	153
10.4.2 <i>Contraction</i>	155
10.4.3 <i>Must and May</i>	156
10.4.4 <i>Will Versus Going To</i>	158
10.4.5 <i>"Romans Perspire, Anglo-Saxons Sweat"</i>	160

10.5 Register and Context	
10.6 Jargon	
10.7 The Cooperative Principle	
10.8 Conclusion	

11 Planning Prosody	
11.1 What Is Prosody	
11.2 Functions of Prosody	
11.3 Stress	
11.4 Intonation	
11.4.1 <i>Basic Intonation</i>	
11.4.2 <i>Contours</i>	
11.5 Concatenating	
11.5.1 <i>The Prosodic Unit</i>	
11.5.2 <i>Concatenation</i>	
11.5.3 <i>Concatenation</i>	
11.6 Minimizing Cognitive Load	
11.7 Pauses	
11.8 TTS Guideline	
11.8.1 <i>Analyze</i>	
11.8.2 <i>Choose</i>	
11.8.3 <i>When to Pause</i>	
11.8.4 <i>Make a List</i>	
11.8.5 <i>Use Appropriate</i>	
11.8.6 <i>Mark Up</i>	
11.9 Conclusion	

12 Maximizing Efficiency	
12.1 Efficiency	
12.1.1 <i>Don't Let</i>	
12.1.2 <i>Make a List</i>	
12.1.3 <i>Provide</i>	
12.1.4 <i>Use Caution</i>	
12.2 Clarity	
12.2.1 <i>Mental</i>	
12.2.2 <i>Navigating</i>	
12.3 Balancing Efficiency	
12.3.1 <i>Stress</i>	

10.5	Register and Consistency	161
10.6	Jargon	164
10.7	The Cooperative Principle	166
10.8	Conclusion	168
11	Planning Prosody	171
11.1	What Is Prosody?	172
11.2	Functions of Prosody	173
11.3	Stress	175
11.4	Intonation	180
	11.4.1 Basic Intonation Contours	180
	11.4.2 Contours in Context	183
11.5	Concatenating Phone Numbers	189
	11.5.1 The Prosodic Structure of Phone Numbers	189
	11.5.2 Concatenation Digit-by-Digit	190
	11.5.3 Concatenation by Groups	191
11.6	Minimizing Concatenation Splices	192
11.7	Pauses	196
11.8	TTS Guidelines	199
	11.8.1 Analyze Application Usage	201
	11.8.2 Choose an Appropriate Voice	201
	11.8.3 When Possible, Use Audio Recordings	201
	11.8.4 Make Content Easy to Understand	202
	11.8.5 Use Appropriate Formats	203
	11.8.6 Mark Up Text for Naturalness	203
11.9	Conclusion	204
12	Maximizing Efficiency and Clarity	205
12.1	Efficiency	206
	12.1.1 Don't Lose Work	207
	12.1.2 Make Frequent Tasks Efficient	207
	12.1.3 Provide Shortcuts	207
	12.1.4 Use Caller Modeling to Save Steps	208
12.2	Clarity	209
	12.2.1 Mental Models for Natural Language Understanding	210
	12.2.2 Navigational Clarity Through Landmarking	211
12.3	Balancing Efficiency and Clarity	212
	12.3.1 Stress Clarity in Individual Prompts	212

12.3.2 Taper Prompts	213
12.3.3 Use Barge-In	213
12.4 Conclusion	215
13 Optimizing Accuracy and Recovering from Errors	217
13.1 Measuring Accuracy	218
13.2 Dialog Design Guidelines for Maximizing Accuracy	219
13.3 Recovering from Errors	221
13.3.1 Confirmation Strategies	221
13.3.2 Recovering from Rejects and Timeouts	224
13.4 Conclusion	228
14 Sample Application: Detailed Design	229
14.1 Call Flow Design	229
14.1.1 The Login Subdialog	231
14.1.2 The Quotes Subdialog	233
14.1.3 The Trading Subdialog	234
14.2 Prompt Design	235
14.3 User Testing	237
14.4 Conclusion	241

Part IV. Realization Phase: Development, Testing, and Tuning

15 Development, Testing, and Tuning Methodology	245
15.1 Development	245
15.1.1 Application Development	246
15.1.2 Grammar Development	246
15.1.3 Audio Production	247
15.2 Testing	247
15.2.1 Application Testing	247
15.2.2 Recognition Testing	248
15.2.3 Evaluative Usability Testing	249
15.3 Tuning	249
15.3.1 Dialog Tuning	251
15.3.2 Recognition Tuning	253
15.4 Conclusion	256

16 Creating Grammar

16.1 Grammar	
16.1.1 De	
16.1.2 De	
16.1.3 De	
16.1.4 De	
16.2 Grammar	
16.2.1 Te	
16.2.2 Te	
16.2.3 Te	
16.2.4 Te	
16.3 Grammar	
16.3.1 Tu	
16.3.2 Tu	
16.3.3 Tu	
16.3.4 Tu	
16.4 Conclusi	

17 Working with

17.1 Scripting	
17.1.1 An	
17.1.2 Sc	
17.2 Choosing	
17.2.1 Pn	
17.2.2 Co	
17.2.3 Fi	
17.2.4 De	
17.3 Running	
17.3.1 Pn	
17.3.2 Vo	
17.4 Conclusi	

18 Sample Appli

18.1 Developr	
18.1.1 Ap	
18.1.2 Gr	
18.1.3 Au	

16 Creating Grammars	257
16.1 Grammar Development	259
16.1.1 <i>Developing Rule-Based Grammars</i>	259
16.1.2 <i>Developing Grammars for Statistical Language Models</i>	263
16.1.3 <i>Developing Robust Natural Language Grammars</i>	264
16.1.4 <i>Developing Statistical Natural Language Grammars</i>	266
16.2 Grammar Testing	266
16.2.1 <i>Testing Rule-Based Grammars</i>	267
16.2.2 <i>Testing Statistical Language Models</i>	270
16.2.3 <i>Testing Robust Natural Language Grammars</i>	270
16.2.4 <i>Testing Statistical Natural Language Grammars</i>	270
16.3 Grammar Tuning	271
16.3.1 <i>Tuning Rule-Based Grammars</i>	271
16.3.2 <i>Tuning Statistical Language Models</i>	272
16.3.3 <i>Tuning Robust Natural Language Grammars</i>	272
16.3.4 <i>Tuning Statistical Natural Language Grammars</i>	272
16.4 Conclusion	273
 17 Working with Voice Actors	 275
17.1 Scripting for Success	276
17.1.1 <i>An Introductory Case Study</i>	276
17.1.2 <i>Scripting Tips</i>	278
17.2 Choosing Your Voice Actor	289
17.2.1 <i>Professionalism and Experience</i>	289
17.2.2 <i>Coachability</i>	290
17.2.3 <i>Fit with Persona</i>	290
17.2.4 <i>Demo Tapes (or CDs) and Auditions</i>	290
17.3 Running a Recording Session	291
17.3.1 <i>Procedural Considerations</i>	291
17.3.2 <i>Voice Coaching</i>	293
17.4 Conclusion	295
 18 Sample Application: Development, Testing, and Tuning	 297
18.1 Development	297
18.1.1 <i>Application Development</i>	297
18.1.2 <i>Grammar Development</i>	298
18.1.3 <i>Audio Production</i>	300

18.2 Testing	301
18.2.1 Evaluative Usability Testing	302
18.3 Tuning	303
18.3.1 Dialog Tuning	303
18.3.2 Recognition Tuning	306
18.3.3 Grammar Tuning	307
18.3.4 User Survey	308
19 Conclusion.....	311
Appendix.....	313
Bibliography.....	315
Index.....	325

Michael Cohen has a variety of roles in a variety of roles. He is a Nuance Professional, a deploy applications technology. He has a user interface (including Say) such as Voyage. Before for years. At SRI I gy development and natural language understanding. Michael speech and V trade shows. I the board of received his P In his sp his band, the

1.2 Why Speech?

Given all the challenges we have cited, why bother with speech interfaces?

To answer that question, let's begin by looking at the Schwab deployment (the application referred to in the opening story). The Schwab application was a big success, from the points of view of the business and the end users. Schwab's original motivation for deploying a speech system for stock quotes was that a large percentage of callers to Schwab's touchtone system immediately pressed 0 on their touchtone keypad to get a live agent. Those callers were the primary target for the new speech system.

Most of those callers ended up very happy using an automated speech application, despite their reluctance to use the touchtone system. The primary reason was that they no longer had to enter stock symbols using complex two-keystroke sequences for each letter, as they did with the touchtone system. Now, they simply said the company name and got their quote. Many customers actually expressed a preference for talking to the speech system rather than a live agent. They were no longer hesitant to ask for a large number of quotes, or even to ask for the same company twice in the same call if they had forgotten the result or were curious to see whether there had been movement in the price.

Charles Schwab & Company was also pleased with the results. The firm saved a lot of money and was able to reassign live agents to deal with more complex user requests, and caller satisfaction was high. Schwab reinforced its reputation as a leader in providing innovations that bring value to their customers.

Since the Schwab deployment, there have been thousands more.⁵ You can now talk to automated systems to trade stocks, check airline schedules and book flights, access your bank account, track packages you have shipped, rent a car, check on weather and traffic conditions, get driving directions, check bus schedules, find out what movies are playing nearby, and make restaurant reservations. You can have your own automated personal agent (discussed shortly) to handle your phone calls, calendar, e-mail, and voice mail. You can access **voice portals** that package a variety of voice services and, in some cases, personalize them to your needs. The list goes on and on, with new ways of providing information and transactions and facilitating human-to-human communications appearing every day.

5. Counting systems from all the major vendors.

The companies deploying these systems are motivated by a number of factors:

- **Saving money:** Speech systems typically save companies significant amounts of money, usually paying for themselves within a few months. When companies replace touchtone services, abandonment rates typically go down substantially, and automation rates often rise dramatically. Decreased call durations have also been a factor, saving companies toll call fees.
- **Improving reach:** Companies want to be available to their customers everywhere (at home and when mobile) at all times (24×7×365). In some cases, a system is deployed to complement the self-service already provided over the Web. In this way, companies can reach customers who either don't have Web access or desire the same self-service access when they are away from their desk or in their car.
- **Extending a brand:** When you engage in spoken interaction, you get more than "just the facts." Speech communicates at many levels. As we listen to someone speak, we make judgments about numerous attributes: the speaker's emotional state, social class, trustworthiness, friendliness, and so on (Soukup 2000; Giles and Powesland 1975). If we design a speech application with these things in mind, we can connect with customers in new ways. The system becomes an extension of the company's brand and image. In effect, we can design the "ideal employee"—with the desired voice, the desired personality traits, the desired mood, and the desired way of handling customer needs and problems. We can present that image reliably, phone call after phone call.
- **Solving new problems:** Many kinds of problems can be solved, or services offered, by speech applications that were simply impossible in the past. For examples, see the sidebar.
- **Increasing customer satisfaction:** Numerous surveys and deployment studies have shown high user satisfaction with speech systems.

From the point of view of the end user, there are many appealing advantages of using speech systems compared with other access modes:

- **Intuitive and efficient:** Spoken language systems draw on the user's innate language skills. Many tasks can be made simpler and more efficient than with touchtones. For example, in a travel application a caller may say things such as, "I wanna leave on June fifth" rather than enter an awkward and unintuitive

New Ways to

Speech applications are changing the way we interact with technology. Many of the most convenient and powerful applications are those that use a set of interactive services that are available to users without the need for a computer or even a phone. These services have not been possible until now. Earlier generations of user interfaces (e.g., command words) have been limited to a few words. Then, with the advent of touchtone services, users could recently deploy services to say things like "Yes, I want to speak with my voice agent." (These services are available on many phones.)

Another new way to interact with technology is through the use of a personal agent. This agent is not rich enough to be able to accompany a user, but it can read us a list of white pages or a list of personal agent information from the user's phone.

touchtone services. Even so, the spelling of words is often a problem.

- **Ubiquitous access:** Users can access services from their desktop or even from a mobile phone.

- **Enjoyment:** Users can engage in a more enjoyable and engaging interaction.

- **Hands-free:** Users can interact with a system without the need for a phone. Further, the system can be used to complete tasks that are difficult to recognize or perform.

New Ways to Solve Old Problems

Speech applications let organizations solve problems in new ways. One example is call routing. Many companies want to combine their various access phone numbers into one convenient point of entry to serve all the needs of their customers. But the resulting set of items from which a caller must choose (such as the names of the various available services) is often huge. Furthermore, the set of choices often does not lend itself to an intuitive menu hierarchy that callers can easily navigate. As a result, it has not been possible to create such systems with touchtone technology (or even with earlier generations of speech technology). Recent advances in **natural language understanding** (technology that figures out the intended meaning of a string of words) have made it possible to let callers state their needs in their own words and then be automatically forwarded to the appropriate service. For example, a system recently deployed by Sprint PCS, with 40 different "routes" or services, allows callers to say things such as, "I just had a call dropped, and I'd like to get a credit for it," "Yes, I want that store in Rockville, on Rockville Pike," and "I'm having a problem with my voice mail—it's locked up on my phone, and I can't get into it to get my messages." (These were transcribed from actual calls.)

Another new application of voice technology is the personal agent. Most of us are not rich enough to hire someone to sit beside us in our car, office, and home and accompany us as we walk down the street, all the while handling our phone calls, reading us our e-mail, managing our calendar, and carrying the entire world's white pages and yellow pages. A number of companies now offer automated **personal agents**, which use speech technology to provide such capabilities on the user's phone.

touchtone sequence (such as 0605) after hearing a long-winded instruction. Even simple things such as saying the name of your destination city, rather than spelling it on a touchtone keypad, bring significant convenience to end users.

- **Ubiquitous:** Telephones are ubiquitous. Many people carry cell phones. Spoken access makes systems available from anywhere, even when the user is far from the desktop.
- **Enjoyable:** A well-designed speech system can provide a user experience that is engaging and enjoyable, at the same time efficiently meeting the user's needs.
- **Hands-free, eyes-free:** Activities such as driving occupy the user's hands and eyes. Speech is an ideal solution for accessing services while engaged in such tasks. Furthermore, mobile devices generally have very poor user interfaces. Entering complex information is awkward with keypads, pointing devices, or handwriting recognition. Speech is the ideal solution.