Preface

Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics provides a forum for today’s speech technology industry leaders – drawn from private enterprises and academic institutions all over the world – to discuss the challenges, advances, and aspirations of voice technology.

The collection of essays contained in this volume represents the research findings of over 30 speech experts, including speech engineers, system designers, linguists, and IT (information technology) and MIS (management information systems) specialists. The book’s 14 chapters are divided into three sections – mobile environments, call centers, and clinics. But given the practical ubiquity of mobile devices, this three-part division sometimes seems almost irrelevant. For example, one of the chapters in the “call centers” section provides a vivid discussion of how to provide today’s call centers with multimodal capabilities – to support text, graphic, voice, and touch – in self-service transactions, so that customers who contact the call center using their mobile phones (rather than a fixed line) can expect a sophisticated interface that lets them resolve their service issues in a way that uses the full capabilities of their handsets, and similarly call center agents using mobile devices that support multimodality can experience more efficient navigation and retrieval of information to complete a transaction for a caller. In the “clinics” section, for that matter, one of the chapters focuses on validating user satisfaction with a voice-activated medical tracking application, run on a compact mobile device for a “hands-free” method of data entry in a clinical setting.

In spite of this unavoidable overlap of sections, the authors’ earnest discussions of the manifold aspects of speech technology not only complement one another but also divide into several areas of specific interest. Each author brings to this roundtable his or her unique insights and new ideas, the fruits of much time spent formulating, developing and testing out their theories about what kinds of voice applications work best in mobile settings, call centers and clinics.

The book begins with an introduction to the role of speech technology in mobile applications written by Bill Meisel, President of TMA Associates. Meisel is also editor of Speech Strategy News and co-chair (with AVIOS) of the annual Mobile Voice conference in northern California. He opens his discussion by quoting the predictions published by the financial investment giant Morgan Stanley in its Mobile Internet Report, issued near the end of 2009. Meisel shows that in Morgan
Stanley’s 694-page report, Mobile Internet Computing was said to be “the technology driver of the next decade,” following the Desktop Internet Computing of the 1990s, the Personal Computing of the 1980s, the Mini-Computing of the 1970s and, finally, the Mainframe Computing of the 1960s. In his chapter, fittingly titled “Life on the Go – The Role of Speech Technology in Mobile Applications,” Meisel asserts that since “the mobile phone is becoming an indispensable personal communication assistant and multi-functional device… [such a] range of applications creates user interaction issues that can’t be fully solved by extending the Graphical User Interface and keyboard to these small devices.” “Speech recognition, text-to-speech synthesis, and other speech technologies,” Meisel continues, “are part of the solution, particularly since, unlike PCs, every mobile phone has a microphone and speech output.”

Advances in Speech Recognition – which is being published at the very beginning of this auspicious decade for mobile computing – examines the practical constraints of using voice in tandem with text. Following Meisel’s comprehensive overview of the role of speech technology in mobile applications, Scott Taylor, Vice President of Mobile Marketing and Solutions at Nuance Communications, Inc., offers a chapter titled “Striking a Healthy Balance – Speech Technology in the Mobile Ecosystem.” Here, Taylor cautions the reader about the need to “balance a variety of multimodal capabilities so as to optimally fit the user’s needs at any given time.” While there is “no doubt that speech technologies will continue to evolve and provide a richer user experience,” argues Taylor, it is critical for experts to remember that “the key to success of these technologies will be thoughtful integration of these core technologies into mobile device platforms and operating systems, to enable creative and consistent use of these technologies within mobile applications.” This is why speech developers, including Taylor himself, view speech capabilities on mobile devices not as a single entity but rather as part of an entire mobile ecosystem that must strive to maintain homeostasis so that consumers (as well as carriers and manufacturers) will get the best service from a given mobile application.

To achieve that goal, Mike Phillips, Chief Technology Officer at Boston-based Vlingo, together with members of the company has been at pains to design more effective and satisfying multimodal interfaces for mobile devices. In the chapter following Taylor’s, titled “Why Tap When You Can Talk – Designing Multimodal Interfaces for Mobile Devices that Are Effective, Adaptive and Satisfying to the User,” Phillips and his co-authors present findings from over 600 usability tests in addition to results from large-scale commercial deployments to augment their discussion of the opportunities and challenges presented in the mobile environment. Phillips and his co-writers stress how important it is to strive for user-satisfaction: “It is becoming clear that as mobile devices become more capable, the user interface is the last remaining barrier to the scope of applications and services that can be made available to the users of these devices. It is equally clear that speech has an important role to play in removing these user interface barriers.”

Johan Schalkwyk, Senior Staff Engineer at Google, along with some of his colleagues provide the book’s next chapter, aptly titled “Your Word is my
Command – Google Search by Voice: A Case Study.” In this chapter, Schalkwyk and his co-authors illuminate the technology employed by Google “to make search by voice a reality” – and follow this with a fascinating exploration of the user interface side of the problem, which includes detailed descriptions and analyses of the specifically tailored user studies that have been based on Google’s deployed applications.

In painstaking detail, Schalkwyk and his colleagues demystify the complicated technology behind 800-GOOG-411 (an automated system that uses speech recognition and web search to help people find and call businesses), GMM (Google Maps for Mobile) which – unlike GOOG-411 – applies a multimodal speech application (making use of graphics), and finally the Google Mobile application for the iPhone, which includes a search by voice feature. The coda to the chapter is its discussion of user studies based on analyses of live data, and how such studies reveal important facts about user behavior, facts that impact Google’s “decisions about the technology and user interfaces.” Here are the essential questions addressed in those user studies: “What are people actually looking for when they are mobile? What factors influence them to choose to search by voice or type? What factors contribute to user satisfaction? How do we maintain and grow our user base? How can speech make information access easier?”

The mobile environments section concludes with the presentation of a well-planned study on speech recognition in noisy mobile environments. Sid-Ahmed Selouani, Professor of Information Management at the Université de Moncton, Shippagan Campus, New Brunswick, Canada, in “Well Adjusted – Using Robust and Flexible Speech Recognition Capabilities in Clean to Noisy Mobile Environments,” presents study findings on a new speech-enabled framework that aims at providing a rich interactive experience for smartphone users – particularly in mobile environments that can benefit from hands-free and/or eyes-free operations.

Selouani introduces this framework by arguing that it is based on a conceptualization that divides the mapping between the speech acoustical microstructure and the spoken implicit macrostructure into two distinct levels, namely the signal level and linguistic level. At the signal level, a front-end processing that aims at improving the performance of Distributed Speech Recognition (DSR) in noisy mobile environments is performed.

The linguistic level, on the contrary, “involves a dialogue scheme to overcome the limitations of current human-computer interactive applications that are mostly using constrained grammars.” “For this purpose,” says Selouani, “conversational intelligent agents capable of learning from their past dialogue experiences are used.”

In conducting this research on speech recognition in clean to noisy mobile environments, Selouani utilized the Carnegie-Mellon Pocket Sphinx engine for speech recognition and the Artificial Intelligence Markup Language (AIML) for pattern matching. The evaluation results showed that including both the Genetic Algorithms (GA)-based front-end processing and the AIML-based conversational agents led to significant improvements in the effectiveness and performance of an interactive spoken dialog system in a mobile setting.

Matthew Yuschik, Senior User Experience Specialist at Cincinnati-based Convergys Corporation provides the perfect segue to the next section of Advances in Speech
Recognition. In “It’s the Best of all Possible Worlds – Leveraging Multimodality To Improve Call Center Productivity,” Yuschik makes a convincing argument for equipping today’s call centers with multimodal capabilities in self-service transactions – to support text, graphic, voice, and touch – so that customers who contact the call center using their mobile phones (rather than a fixed line) can expect an interface that “provides multiple ways for the caller to search for resolution of their [service] issue.” Given market research predictions that there will be over 4 billion wireless subscribers in 2010, Yuschik draws the sound conclusion that more and more callers will be using their mobile devices when availing themselves of customer support services at customer care and contact centers. After all, most customers who need to resolve product and service issues, or to order new products and services, squeeze in their calls “on the go” instead of taking up crucial time while working at their desks.

In “It’s the Best of all Possible Worlds,” Yuschik explains how leveraging multimodality to improve call center productivity is achieved by striking a healthy balance between satisfying the caller’s goal and maximizing the agent’s productivity in the call center. He points out that “a multimodal interface can voice-enable all features of a GUI.” Yet, he cautions “this is a technologically robust solution, but does not necessarily take into account the caller’s goal.” Conceding that “voice activating all parts of the underlying GUI of the application enables the agent to solve every problem by following the step-by-step sequence imposed by the GUI screens,” Yuschik states that “a more efficient approach...is to follow the way agents and callers carry on their dialog to reach the desired goal.” He shows that “this scenario-based (use-case) flow – with voice-activated tasks and subtasks – with tasks and subtasks voice activated – provides a streamlined approach in which an agent follows the caller-initiated dialog, using the MMUI [multimodal user interface] to enter data and control the existing GUI in any possible sequence of steps. This goal-focused view,” as explained by Yuschik, “enables callers to complete their transactions as fast as possible.”

Yuschik’s chapter details a set of Convergys trials that “follow a specific sequence where multimodal building-blocks are identified, investigated, and then combined into support tasks that handle call center transactions.” Crucial to those trials were the Convergys call center agents who “tested the Multimodal User Interface for ease of use, and efficiency in completing caller transactions.” The results of the Convergys trials showed that “multimodal transactions are faster to complete than only using a Graphical User Interface.” Yuschik concludes that “the overarching goal of a multimodal approach should be to create a framework that supports many solutions. Then,” he writes, “tasks within any specific transaction are leveraged across multiple applications.”

Every new technology deserves an accurate method of evaluating its performance and effectiveness; otherwise, the technology will not fully serve its intended purpose. David Suendermann, Principal Speech Scientist at the New York-based SpeechCycle, Inc., and his colleagues Roberto Pieraccini and Jackson Liscombe, are joined by Keelan Evanini of Educational Testing Services in Princeton, New Jersey, for the presentation of an enlightening discussion of a new framework to measure accurately the performance of automated customer care contact centers.
In “‘How am I Doing?’ – A New Framework To Effectively Measure the Performance of Automated Customer Care Contact Centers,” the authors carefully dissect conventional methods of measuring how satisfied customers are with automated customer care and contact centers, pointing out why such methods can produce woefully misleading results. They point to a problem that is ever-present when evaluating callers’ satisfaction with any of these self-service contact centers. Namely: quantifying how effectively interactive voice response (IVR) systems satisfy callers’ goals and expectations “has historically proven to be a most difficult task.” Suendermann and his co-authors convincingly show that such difficulties in assessing automated customer care contact centers can be traced to two assumptions [albeit misguided] made by most stakeholders in the call center industry:

1. Performance can be effectively measured by deriving statistics from call logs; and
2. The overall performance of an IVR can be expressed by a single numeric value.

The authors introduce an IVR assessment framework that confronts these misguided assumptions head on, demonstrating how they can be overcome. The authors show how their “new framework for measuring the performance of IVR-driven call centers incorporates objective and subjective measures.” Using the concepts of hidden and observable measures, the authors demonstrate how to produce metrics that are reliable and meaningful so that they can better provide accurate system design insights into multiple aspects of IVR performance in call centers.

Just as it is possible to jettison poor methods of evaluating caller satisfaction with IVR performance in favor of more accurate ones, it is equally possible to meet (or even exceed) user expectations with the design of a speech-only interface that builds on what users have come to expect from self-service delivery in general, whether at the neighborhood pharmacy or at the international airport. Stephen Springer, Senior Director of User Interface Design at Nuance Communications, Inc., shows how to do this in his chapter (aptly) titled “Great Expectations – Making Use of Callers’ Experiences from Everyday Life To Design a Satisfying Speech-Only Interface for the Call Center.” According to Springer, “the thoughtful use of user modeling achieved by employing ideas and concepts related to transparency, choice, and expert advice, all of which most, if not all, callers are already familiar with from their own everyday experiences” better meets the users’ expectations than systems whose workings are foreign to what such users encounter in day-to-day life.

Springer carefully examines a wide variety of expectations that callers bring to self-service phone calls, ranging from broad expectations about self-service in general to the more specific expectations of human-to-human conversation about consumer issues. As a specialist in user interface design, Springer recommends to the system designer several indispensable steps to produce more successful interaction between callers and speech interfaces. The irony is that the secrets for meeting greater expectations for caller satisfaction with speech-only interfaces in the call center are not really secrets: they can be found uncannily close to home, by
extrapolating from callers’ everyday self-service experiences and from their quotidian dialog with human agents at customer care contact centers.

Next, two German academics and SpeechCycle’s CTO, Roberto Pieraccini, tackle the inscrutable and often elusive emotions of callers to ascertain when task completion and user satisfaction with the automated call center may be at risk. Alexander Schmitt of Ulm University, and his two co-authors, in their chapter titled “‘For Heaven’s Sake, Gimme a Live Person!’ – Designing Emotion-Detection Customer Care Voice Applications in Automated Call Centers,” show how their voice application can robustly detect angry user turns by considering acoustic, linguistic, and interaction parameter-based information – all of which can be collected and exploited for anger detection. They introduce, in addition, a valuable subcomponent that is able to estimate the emotional state of the caller based on the caller’s previous emotional state, supporting the theory that anger displayed in calls to automated call centers, rather than being an isolated occurrence, is more likely an incremental build-up of emotion. Using a corpus of 1,911 calls from an Interactive Voice Response system, the authors demonstrate the various aspects of speech displayed by angry callers.

The call center section of Advances in Speech Recognition is rounded off by a fascinating chapter on advanced speech analytic solutions aimed at learning why customers call help-line desks and how effectively they are served by the human agent. Yes, that is correct: a human agent, a specimen of call center technology that still exists notwithstanding the push for heavily automated self-service centers. In “The Truth Is Out There – Using Advanced Speech Analytics To Learn Why Customers Call Help-Line Desks and How Effectively They’re Being Served by the Call Center Agent,” Marsal Gavalda, Vice President of Incubation and Principal Language Scientist at Nexidia, and Jeff Schlueter (the company’s Vice President of Marketing & Business Development) describe their novel work in phonetic-based indexing and search, which is designed for extremely fast searching through vast amounts of media.

The authors of “The Truth is Out There” explain the nuts and bolts of their method, showing how they “search for words, phrases, jargon, slang and other terminology that are not readily found in a speech-to-text dictionary.” They demonstrate how “the most advanced phonetic-based speech analytics solutions,” such as theirs, “are those that are robust to noisy channel conditions and dialectal variations; those that can extract information beyond words and phrases; and those that do not require the creation or maintenance of lexicons or language models.” The authors assert that “such well performing speech analytic programs offer unprecedented levels of accuracy, scale, ease of deployment, and an overall effectiveness in the mining of live and recorded calls.” Given that speech analytics has become indispensable to understanding how to achieve a high rate of customer satisfaction and cost containment, Gavalda and his co-author demonstrate in their chapter how their data mining technology is used to produce sophisticated analyses and reports (including visualizations of call category trends and correlations or statistical metrics), while preserving “the ability at any time to drill down to individual calls and listen to the specific evidence that supports the particular categorization or data
point in question, all of which allows for a deep and fact-based understanding of contact center dynamics.”

John Shagoury, Executive Vice President of the Healthcare & Imaging Division of Nuance Communications, Inc., opens Advances in Speech Recognition's last section with a cogent discussion of “the benefits of incorporating speech recognition as part of the everyday clinical documentation workflow.” In his chapter – fittingly titled “Dr. Multi-Task – Using Speech To Build up Electronic Medical Records While Caring for Patients” – Shagoury shows how speech technology yields a significant improvement in the quality of patient care by increasing the speed of the medical documentation process, so that patients’ health records are quickly made available to healthcare providers. This means they can deliver timely and efficient medical care. Using some fascinating, and on point, real-world examples, Shagoury richly demonstrates how the use of speech recognition technology directly affects improved productivity in hospitals, significant cost reductions, and overall quality improvements in the physician’s ability to deliver optimal healthcare. But Shagoury does not stop there. He goes on to demonstrate that “beyond the core application of speech technologies to hospitals and primary care practitioners, speech recognition is a core tool within the diagnostics field of healthcare, with broad adoption levels within the radiology department.”

Next, James Rodger, Professor of Management Information Systems and Decision Sciences at Indiana University of Pennsylvania, Eberly College of Business and Information Technology – with his co-author, James A. George, senior consultant at Sam, Inc. – provides the reader with a rare inside look at the authors’ “decade long odyssey” in testing and validating end-user acceptance of speech in the clinical setting aboard US Navy ships. In their chapter, titled “Hands Free – Adapting the Task-Technology-Fit Model and Smart Data To Validate End-User Acceptance of the Voice Activated Medical Tracking Application (VAMTA) in the United States Military,” the authors show how their extensive work on validating user acceptance of VAMTA – which is run on a compact mobile device that enables a “hands-free” method of data entry in the clinical setting – was broken down into two phases: 1) a pilot to establish validity of an instrument for obtaining user evaluations of VAMTA and 2) an in-depth study to measure the adaptation of users to a voice-activated medical tracking system in preventive health care. For the latter phase, they adapted a task-technology-fit (TTF) model (from a smart data strategy) to VAMTA, demonstrating that “the perceptions of end-users can be measured and, furthermore, that an evaluation of the system from a conceptual viewpoint can be sufficiently documented.” In this chapter, they report on both the pilot and the in-depth study.

Rodger and his co-author applied the Statistical Package for the Social Sciences (SPSS) data analysis tool to analyze the survey results from the in-depth study to determine whether TTF, along with individual characteristics, will have an impact on user evaluations of VAMTA. In conducting this in-depth study, the authors modified the original TTF model to allow adequate domain coverage of patient care applications. What is most interesting about their study – and perhaps a testament to the vision of those at the forefront of speech applications in the clinical setting – is that,
according to Rodger and his co-author, their work “provides the underpinnings for a subsequent, higher level study of nationwide medical personnel.” In fact, they intend “follow-on studies [to] be conducted to investigate performance and user perceptions of VAMTA under actual medical field conditions.”

Julia Hirschberg and Noémie Elhadad, distinguished faculty members at Columbia University, in concert with Anna Hjalmarsson, a bright and talented Swedish graduate student studying at KTH (Royal Institute of Technology), make a strong argument that if “language cues” – primarily acoustic signal and lexical and semantic features – “can be identified and quantified automatically, this information can be used to support diagnosis and treatment of medical conditions in clinical settings [as well as] to further fundamental research in understanding cognition.” In “You’re As Sick As You Sound – Using Computational Approaches for Modeling Speaker State To Gauge Illness and Recovery,” Hirschberg and her co-authors perform an exhaustive medical literature review of studies “that explore the possibility of finding speech-based correlates of various medical conditions using automatic, computational methods.” Among the studies they review are computational approaches that explore communicative patterns of patients who suffer from medical conditions such as depression, autism spectrum disorders, schizophrenia, and cancer.

The authors see a ripe opportunity here for future medical applications. They point out that the emerging research into speaker state for medical diagnostic and treatment purposes – an outgrowth of “related work on computational modeling of emotional state” for studying callers’ interactions with call center agents and Interactive Voice Response (IVR) applications “for which there is interest in distinguishing angry and frustrated callers from the rest” – equips the physician with a whole new set of diagnostic and treatment tools. “Such tools can have economic and public health benefits, in that a wider population – particularly individuals who live far from major medical centers – can be efficiently screened for a broader spectrum of neurological disorders,” they write. “Fundamental research on mental disorders, like post-partum depression and post traumatic stress disorder, and coping mechanisms for patients with chronic conditions, like cancer and degenerative arthritis, can likewise benefit from computational models of speaker state.”

Hemant Patil, Assistant Professor at the Dhirubhai Ambani Institute of Information and Communication Technology, DA-IICT, in Gandhinagar, India, echoes the beliefs of Shagoury, Rodger and George, and of Hirschberg, Hjalmarsson and Elhadad, all of whom maintain that advances in speech technology have untold economic, social, and public health benefits. In “‘Cry Baby’ – Using Spectrographic Analysis To Assess Neonatal Health Status from an Infant’s Cry,” Patil demonstrates that the rich body of research on spectrographic analysis, predominantly used for performance of speaker recognition, may also be used to assess the neonate’s health status, by comparing a normal to an abnormal cry.

Spectrographic analysis is seen by Patil and his colleagues – who are just as passionately involved in this highly specialized area of infant cry research – as useful in improving and complementing “the clinical diagnostic skills of pediatricians and neonatologists, by helping them to detect early warning signs of pathology,
developmental lags, and so forth.” Patil points out to the reader that such technology “is especially helpful in today’s healthcare environment, in which newborns do not have the luxury of being solely attended by one physician, and are, instead, monitored remotely by a centralized computer control system.”

In explaining cry analysis – a multidisciplinary area of research integrating pediatrics, neurology, physiology, engineering, developmental linguistics, and psychology – Patil demonstrates in “Cry Baby” his application of spectrographic analysis to the vocal sounds of an infant, comparing normal with abnormal infant crying. In his study, ten distinct cry modes, viz., hyperphonation, dysphonation, inhalation, double harmonic break, trailing, vibration, weak vibration, flat, rising, and falling, have been identified for normal infant crying, and their respective spectrographic patterns were observed. This analysis was then extended to the abnormal infant cry. Patil observed that

the double harmonic break is more dominant for abnormal infant cry in cases of myalgia (muscular pain). The inhalation pattern is distinct for infants suffering from asthma or other respiratory ailments such as a cough or cold. For example, for the infant whose larynx is not well developed, the pitch harmonics are nearly absent. As such, there are no voicing or glottal vibrations in the cry signal. And for infants with Hypoxic Ischemic Encephalopathy (HIE), there is an initial tendency of pitch harmonics to rise and then to be followed by a blurring of such harmonics.

As part of this study, Patil also performed infant cry analysis by observing the nature of the optimal warping path in the Dynamic Time Warping (DTW) algorithm, which is found to be “near diagonal” in healthy infants, in contrast to that in unhealthy infants whose warping paths reveal significant deviations from the diagonal across most, though not all, cry modes.

Looking further into broader sociologic implications of cry analysis, Patil shows how this novel field of research can redress the social and economic inequities of healthcare delivery. “Motivated by a need to equalize the level of neonatal healthcare (not every neonate has the luxury of being monitored at a teaching hospital equipped with a high level neonatal intensive care unit), I propose for the next phase of research a quantifiable measurement of the added clinical advantage to the clinician (and ancillary healthcare workers) of a baseline comparison of normal versus abnormal cry.”

Now it is up to the reader, after assimilating the substance of this book, to envision how speech applications in mobile environments, call centers, and clinics will improve the lives of consumers, corporations, carriers, manufacturers, and healthcare providers – to say nothing of the overall improvements that such technology provides for the byzantine social architecture known as modern-day living.

Fort Lee, NJ

Amy Neustein, Ph.D
Advances in Speech Recognition
Mobile Environments, Call Centers and Clinics
Neustein, A. (Ed.)
2010, XXVI, 369 p., Hardcover
ISBN: 978-1-4419-5950-8